申报	系列：教师系列 教学科研并重型
	专业：机械工程及自动化
	职称：教授

业绩成果材料

（申报人的业绩成果材料包括论文、科研项目、获奖以及其他成果等）

单　位（二级单位）工程学院

姓　名　吕盛坪

材料核对人：　　单位盖章：

核对时间：

华南农业大学制

目　录

一、教研业绩

二、科研项目

三、科研成果

一、教研业绩

结题证明

编号：JXJT20085

我校教师**吕盛坪**主持的校级教学改革项目《**面向新工科人才培养的车辆工程专业实践教育教学体系构建研究**》（项目编号：JG18104），经学校组织验收，已于2020年12月结题。

特此证明。

华南农业大学本科生院

2020年12月

项目编号： JG19 145

华南农业大学教育教学研究和改革项目

申 报 书

项目名称 面向新能源汽车产业需求的车辆工程本科生专业知识渗透与提升研究

项目负责人 郭嘉明

职 称 副教授

工作单位 工程学院 （盖章）

移动电话 [redacted]

电子邮箱 jmguo@scau.edu.cn

申报日期 2019 年 6 月 21 日

华南农业大学 教务处 制

2019 年 6 月

申请者的承诺与成果使用授权

本人自愿申报华南农业大学教育教学改革项目，承诺对所填写的《申报书》所涉及各项内容的真实性负责，保证没有知识产权争议。课题申请如获准立项，在研究工作中，接受华南农业大学教务处及本人所在单位的管理，并对以下约定信守承诺：

1. 遵守相关法律法规。遵守我国著作权法和专利法等相关法律法规；遵守我国政府签署加入的相关国际知识产权规定。

2. 遵循学术研究的基本规范，恪守学术道德，维护学术尊严。研究过程真实，不以任何方式抄袭、剽窃或侵吞他人学术成果，杜绝伪注、伪造、篡改文献和数据等学术不端行为；成果真实，不重复发表研究成果；维护社会公共利益，不以项目名义牟取不当利益。

3. 遵守华南农业大学教育教学改革项目有关管理规定以及华南农业大学财务规章制度。

4. 凡因项目内容、成果或研究过程引起的法律、学术、产权或经费使用问题引起的纠纷，责任由相应的项目研究人员承担。

5. 项目获批后务必按项目计划要求及时开展研究工作，确保研究工作如期完成。

6. 同意华南农业大学或其授权（委托）单位有权基于公益需要公布、使用、宣传《项目申请·评审书》内容及相关成果。

项目负责人（签章）：郭嘉明

2019 年　　月　　日

一、项目及项目负责人、项目组简况

<table>
<tr><td rowspan="3">项目简况</td><td>项目名称</td><td colspan="6">面向新能源汽车产业需求的车辆工程本科生专业知识渗透与提升研究</td></tr>
<tr><td>项目类别</td><td colspan="6">□1. 重点项目 ☑2. 一般项目 □3. 青年项目</td></tr>
<tr><td>起止年月</td><td colspan="6">2019.07-2020.12</td></tr>
<tr><td rowspan="11">项目申请人</td><td>姓名</td><td>郭嘉明</td><td>性别</td><td>男</td><td>出生年月</td><td colspan="2">1987 年 9 月</td></tr>
<tr><td colspan="2">专业技术职务/
行政职务</td><td colspan="2">副教授/无</td><td>最终学位/授予国家</td><td colspan="2">博士/中国</td></tr>
<tr><td rowspan="2">所在单位及联系方式</td><td>单位名称</td><td colspan="3">工程学院 车辆工程系</td><td>手机号码</td><td colspan="2">13760729050</td></tr>
<tr><td>电子邮箱</td><td colspan="5">jmguo@scau.edu.cn</td></tr>
<tr><td rowspan="6">主要教学工作简历</td><td>时间</td><td>课程名称</td><td colspan="2">授课对象</td><td>学时</td><td colspan="2">所在单位</td></tr>
<tr><td>2017.9-2017.12</td><td>汽车电器</td><td colspan="2">本科生</td><td>16</td><td colspan="2">工程学院</td></tr>
<tr><td>2017.9-2017.12</td><td>汽车构造 I</td><td colspan="2">本科生</td><td>32</td><td colspan="2">工程学院</td></tr>
<tr><td>2018.3-2018.5</td><td>汽车理论</td><td colspan="2">本科生</td><td>32</td><td colspan="2">工程学院</td></tr>
<tr><td>2018.9-2018.12</td><td>汽车构造 I</td><td colspan="2">本科生</td><td>32</td><td colspan="2">工程学院</td></tr>
<tr><td>2019.3-2019.5</td><td>汽车理论</td><td colspan="2">本科生</td><td>32</td><td colspan="2">工程学院</td></tr>
<tr><td rowspan="8">项目组</td><td rowspan="2">总人数</td><td colspan="3">职称</td><td colspan="3">学位</td><td></td></tr>
<tr><td>高级</td><td>中级</td><td>初级</td><td>博士后</td><td>博士</td><td>硕士</td><td>参加单位数</td></tr>
<tr><td>4</td><td>3</td><td>1</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="5">主要成员（不含申请者）</td><td>姓名</td><td>性别</td><td>出生年月</td><td>职称</td><td>工作单位</td><td>分工</td><td colspan="2">签名</td></tr>
<tr><td>吕恩利</td><td>男</td><td>年 月</td><td>青年教授</td><td>车辆工程系</td><td>整体指导</td><td colspan="2"></td></tr>
<tr><td>吕盛坪</td><td>男</td><td>年 月</td><td>副教授</td><td>车辆工程系</td><td>实施学教改革内容</td><td colspan="2"></td></tr>
<tr><td>曾志雄</td><td>男</td><td>年 月</td><td>实验师</td><td>工业设计系</td><td>调研、数据整理</td><td colspan="2"></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="2"></td></tr>
</table>

六、单位、评审小组及学校意见

所在单位意见：

同意

（公章）

单位负责人签字：

2019 年 6 月 21 日

评审小组意见：

同意立项

评审小组长签字：

2019 年 月 日

学校主管部门意见：

同意立项

签章：

2019 年 月 日

中国高等教育学会实验室管理工作分会会刊
中文核心期刊
RCCSE中国权威学术期刊

ISSN 1002-4956
CN 11-2034/T
CODEN SJYGAR

实验技术与管理

8
2021
第38卷 第8期
Vol. 38 No.8
月刊

Shiyan Jishu yu Guanli

Experimental Technology and Management

ISSN 1002-4956
08>

中华人民共和国教育部主管 清华大学主办

实验技术与管理
Shiyan Jishu yu Guanli
（1963年创刊　月刊）

主管：中华人民共和国教育部
主办：清华大学
中国高等教育学会实验室管理工作分会会刊
中文核心期刊
RCCSE中国权威学术期刊
中国高校优秀科技期刊
《CAJ-CD规范》执行优秀期刊

主编：黄开胜
编辑：《实验技术与管理》编辑部
地址：北京市海淀区清华大学科技服务楼
邮编：100084
网址·在线投稿：
http://syjl.cbpt.cnki.net
http://syjl.chinajournal.net.cn
编辑部电话：010-62783005
邮箱：sjg@tsinghua.edu.cn
广告电话：010-62797828
订刊发行电话：010-62792635
邮箱：syjsygl@tsinghua.edu.cn
出版与发行：清华大学出版社有限公司
印刷：北京卓诚恒信彩色印刷有限公司
发行范围：国内外公开发行
国际标准连续出版物号：ISSN 1002-4956
国内统一连续出版物号：CN 11-2034/T
国际期刊编码：CODEN SJYGAR
广告经营许可证：京海工商广字第0081号
出版日期：8月20日
定价：23.50元/期　全年12期　282.00元

收录本刊内容的国内外数据库与媒体：
- 中国学术期刊（光盘版）
- 中国核心期刊（遴选）数据库
- 万方数据资源系统数字化期刊群
- 中国期刊网
- 中国学术期刊综合评价数据库
- 中国期刊全文数据库
- 中文科技期刊数据库
- 中文电子期刊服务
- 中国学术期刊文摘（中文版）
- 中国科技论文在线
- 美国《剑桥科学文摘》（CSA）
- 美国《化学文摘（网络版）》（CA）
- 美国《乌利希期刊指南（网络版）》（Ulrichsweb）
- 英国《世界陶瓷文摘（网络版）》（WCA）
- 日本《日本科学技术振兴机构数据库（中国）》（JST China）
- 美国《艾博思科数据库》（EBSCOhost）

目　次

第38卷　第8期（总第300期）　2021年8月

期刊基本参数：CN11-2034/T*1963*m*A4*272*zh*P*¥23.50*6500*57*2021-08

《实验技术与管理》
微信公众号

实验技术与管理

第十一届编辑委员会名单

主　任：王希勤

顾　问：（按姓氏笔画排序）
朱　静　邱爱慈　陈小明　周　玉　周远清
高　松　席葆树　程建平　潘际銮　翟振元

副主任：（按姓氏笔画排序）
王小力　方东红　朱　臻　刘克新　孙　骞
孙小平　杨佩青　应　敏　张云怀　张社荣
罗正祥　罗立胜　宗俊峰　荣　昶　敖天其
唐睿康　黄开胜　符宁平　梁　宏　董　林
蒋兴浩　雷敬炎　熊宏齐

编　委：（按姓氏笔画排序）
马　强　马国玉　王　建　王　浩　王　辉
王　强　王　勤　王秀梅　王树彬　王海东
毛继泽　尹珍宝　石宏伟　史天贵　付洪利
冯建刚　朱再明　朱付青　向坚持　庄志鸿
刘　仁　刘　锋　刘庆刚　刘拥军　刘幽燕
许四杰　农春仕　孙　福　孙文磊　孙学军
孙春阳　孙胜春　孙桂五　芦　燕　李　莉
李　崧　李　清　李　建　李天书　李文中
李文涛　李方伟　李声威　李格升　李晓辉
李震彪　杨　波　杨　琦　杨旭静　杨建新
杨培飞　杨德嵩　吴　卫　吴国新　吴祝武
吴福根　邱化寅　何一萍　何都良　汪必琴
汪盛科　沈　江　沈如群　张　帆　张　旭
张　林(川)　张　林(苏)　张　莉　张宏玉　张若舒
张建成　张界新　张洪清　张冠鹏　张海峰
张维平　陈　越　陈　涌　陈小鸿　陈心浩
陈灵泉　罗茂斌　金仁东　周立超　赵　明
荆　莹　胡惠君　钟　冲　钟华勇　施芝元
贺占魁　袁　若　袁友亮　袁洪学　贾果欣
徐四平　徐秀吉　徐美勇　高　欣　高　鸿
高志华　郭　平　郭　庆　郭松斌　郭建中
唐俊峰　黄春麟　黄富贵　崔宏伟　崔锦峰
康传红　康智勇　梁　勇　韩英霞　傅志刚
曾　莉　楚丹琪　管图华　廖梦圆　颜忠诚
薛凌云　魏永前

特邀编委：（按姓氏笔画排序）
王　杰　王　健　冯建跃　兰中文　吕厚均
严　薇　李鸿飞　吴　兵　邹永松　张文桂
张家林　赵建新　胡今鸿　姜文凤　黄　刚

编辑部成员

主　　编：黄开胜
副 主 编：彭远红
编辑部主任：彭远红（兼）
编　　辑：罗立胜　张文杰　彭远红　张利芳　孙　浩
编　　务：杨荫荫　陈　昕
发　　行：段　然　吴　岩
广　　告：孙　浩　张利芳　陈　红（兼）
封面题字：刘仙洲院士、清华大学原第一副校长，1963 年题

ISSN 1002-4956
CN11-2034/T

实 验 技 术 与 管 理
Experimental Technology and Management

第 38 卷 第 8 期 2021 年 8 月
Vol.38 No.8 Aug. 2021

DOI: 10.16791/j.cnki.sjg.2021.08.034

数据挖掘教学示教与学生实践系统设计与开发

吕盛坪，王海林，李 君

（华南农业大学 工程学院，广东 广州 510642）

摘 要：为了更好地支持数据挖掘教学示教，辅助学生理解数据挖掘流程与原理、掌握算法的开发实现及其应用，构建了一个开放可扩展的数据挖掘原型系统。分析了数据挖掘流程与任务，划分了系统功能模块，优选了各模块挖掘机制，梳理了关键原理与调控参数，开发了教学示教与学生实践平台，设计了基于该系统的教师示教与学生实践范式，并根据企业具体需求和具体数据给出部分执行页面。

关键词：数据挖掘；使用范式；系统开发

中图分类号：G643；TP391.9 **文献标识码：**A **文章编号：**1002-4956(2021)08-0163-06

Design and development of data mining teaching demonstration and student practice system

LYU Shengping, WANG Hailin, LI Jun

(School of Engineering, South China Agricultural University, Guangzhou 510642, China)

Abstract: In order to better support the demonstration teaching of data mining, assist students to understand the process and principle of data mining, and master the development, implementation and application of algorithms, an open and extensible data mining prototype system is constructed. This paper analyzes the data mining process and tasks, divides the system function modules, optimizes the mining mechanism of each module, summarizes the key principles and control parameters, develops the teaching demonstration and student practice platform, designs the teacher demonstration and student practice paradigm based on the system, and presents some execution pages according to the specific needs and data of the enterprise.

Key words: data mining; usage paradigm; system development

强化学生数据科学相关知识与能力，提升学生数据科学思维是高校“新工科”建设人才培养新目标[1-2]，开设数据挖掘或大数据课程，并保证学生进行充分实践，是培养学生数据科学思维的重要手段[3-4]。近年来，相关高校针对数据挖掘课程教学改革开展了深入研究。佳木斯大学从数据意识培养、教学方法创新等方面对数据挖掘课程改革进行了探索[5]；韩山师范学院对数据挖掘课程的实践教学内容和方式进行了改革[6]；浙江工业大学从产学合作、项目驱动、多师授课、翻转课堂和综合评价等方面进行改革，以进一步提升学生解决复杂工程问题的能力[7]；浙江理工大学对数据挖掘教学内容、方法、考核方式等方面进行了分析和探讨[8]；同济大学对新工科背景下数据挖掘课程综合性实验进行了设计[9]；新疆财经大学以数据挖掘课程为例开展了以“课堂为主、线上为辅”的“混合式”教学实践[10]。

当前数据挖掘实践课程常基于两种方式进行。一种是基于已有软件，但这些软件一般只提供封装接口或页面，算法原理介绍与具体实现脱节，学生无法据此进行扩展开发。另一种是针对不同算法分别开发小程序，这种模式具有较好的开放性，可支持学生查阅源代码、进行扩展开发，并帮助学生理解实现机制，

收稿日期：2020-12-18

基金项目：2017 年广东省教育厅教学改革研究项目（粤教高函〔2017〕214 号）；2018 年华南农业大学教育教学改革项目（JG2018104）

作者简介：吕盛坪（1982—），男，湖南新邵，博士，副教授，主要从事工/农业数据挖掘研究，lvshengping@scau.edu.cn。

引文格式：吕盛坪，王海林，李君. 数据挖掘教学示教与学生实践系统设计与开发[J]. 实验技术与管理, 2021, 38(8): 163-168.

Cite this article: LYU S P, WANG H L, LI J. Design and development of data mining teaching demonstration and student practice system[J]. Experimental Technology and Management, 2021, 38(8): 163-168. (in Chinese)

但这种模式集成性和系统性差，难以支持贯穿于数据挖掘全流程课程的示教和综合性挖掘实践。

本研究在梳理数据挖掘流程与任务、划分系统模块基础上，优选了各模块挖掘机制，梳理了关键原理与调控参数，开发了相应的系统，设计了支持系统的使用范式，并以企业具体需求和工业软件采集累计的数据为背景，给出了教师示教和学生在此平台开展相应挖掘实践的部分页面。

1 挖掘流程和任务

分析数据挖掘流程和各环节主要任务，建立以流程为核心的数据挖掘任务框架，具体如图 1 所示。主要包括数据准备、预处理、（狭义）挖掘分析、评价解释以及开发实现等环节。

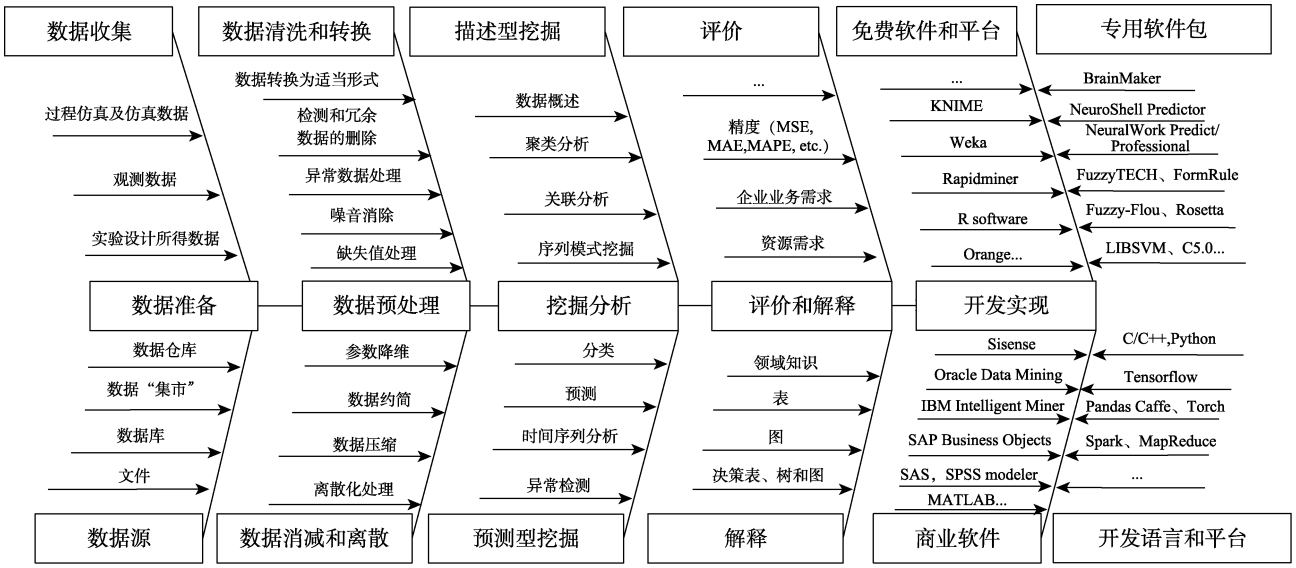


图 1 数据挖掘流程与任务

数据准备的主要任务是确定数据源及其收集方法。数据预处理主要包括数据清洗、转换、消减和离散等。数据清洗包括平滑噪声、离群值筛选、冗余数据检测等；数据转换是对数据形式的变换，如数据归一；数据离散化是通过用高层次概念替换低层次原始数据；数据消减主要包括参数降维、数据约简和压缩。参数降维的目的是检测和移除不相关、弱相关或冗余的属性；数值约简是通过较少的数据替换原始数据，如抽样；数据压缩是通过转换数据原始属性以获得对原始数据的压缩，如主成分分析。数据归一、离群值筛选、参数降维、数据抽样、压缩和离散化是常见的数据预处理技术[11]。

（狭义）挖掘分析主要指数据挖掘模型和算法。挖掘模型可以划分为描述型和预测型两类[12]。描述型挖掘的目的是表征目标数据特性，主要包括数据概述、聚类、关联分析（包括序列模式）等；预测型挖掘是基于当前数据建立模型并用于后续预测，主要包括分类、预测、时序分析等。聚类、分类、预测等是工业大数据中常用的挖掘模型。

对挖掘结果需要进行正确的评价和解释。不同挖掘模型具有不同的评价指标，例如对分类模型的评价一般是基于混淆矩阵计算精确率、准确率等。同时，还需要结合业务需求和合适的表现形式予以展现，以便于理解，图、表、决策树/图等是常见的可视化方式[13]。

上述框架明确了系统核心模块，任务划分细化了各模块待开发实现的功能，这将用于指导系统使用范式设计、教师示教及学生实践。

2 系统模块与实现

基于图 1 的指导框架，优选部分挖掘实现机制，将系统划分为数据基本统计、预处理、分类、预测和聚类等几大模块，并进一步优选相应的实现机制。图 2 给出了原型系统模块划分及各模块将要实现的主要

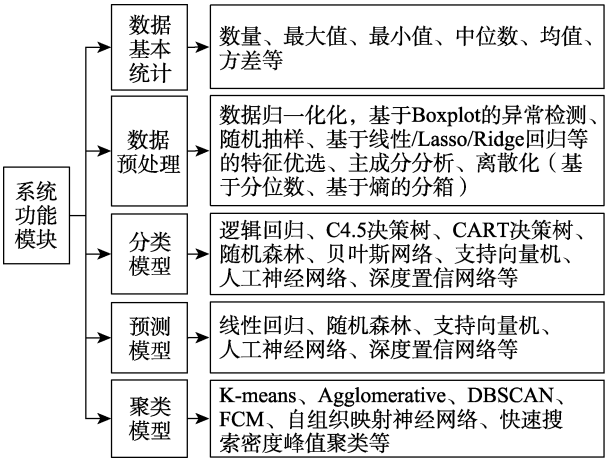


图 2 系统模块与主要功能

表 1　部分挖掘机制关键原理和调控参数

挖掘机制	关键原理	调控参数
线性回归	损失函数、过拟合、回归系数更新迭代	最大迭代次数、正则系数
逻辑回归	交叉熵损失函数、Sigmoid 函数、过拟合、逻辑回归系数更新迭代、条件期望损失	
C4.5 决策树	信道模型、先（后）验不确定性、先（后）验熵、后验熵期望、信息增益(率)、最佳分组变量和最佳分割点的确定、截枝	信息增益阈值
CART 决策树	Gini 系数、方差、异质性、最佳分组变量和最佳分割点的确定、裁剪损失函数	叶子节点最小规模
随机森林	Bagging、集成学习、Bootstrap 抽样、平均精度下降、平均 Gini 下降	决策树数量
贝叶斯网络	联合/条件/先验/后验概率、乘法定理、全概率/贝叶斯公式、最大似然和最大后验概率估计、有向无环图、条件概率分布表、结构/参数学习、Metropolis-Hastings 抽样等	–
支持向量机	结构风险最小、超平面、间隔最大化、支持向量、线性可分、核函数、松弛变量、对偶问题、Karush-Kuhn-Tucher 条件、序列最小优化算法	核函数、阶数、gamma 和 *r*、惩罚因子
人工神经网络	正则化损失函数、激活函数、前向传播、误差反向传播、小批量梯度下降法、权重和偏置更新、学习率、动量项	层数、各层节点数、各层激活函数、批次大小、迭代次数、学习率、动量系数等
CFSFDP	局部密度、Cut-off 和高斯核、聚类决策图、聚类核和聚类光环、聚类迭代	百分比、核函数（Cut-off、Gaussian 核）、距离类型

注：CFSFDP: Clustering by fast search and find of density peaks[14]

功能。除了决策树、随机森林、支持向量机、人工神经网络（ANN）、K-means、Agglomerative、DBSCAN、FCM 聚类等经典算法，还需考虑近年来的一些高水平研究成果，如 Science 发表的快速搜索密度峰值聚类算法等[14]。

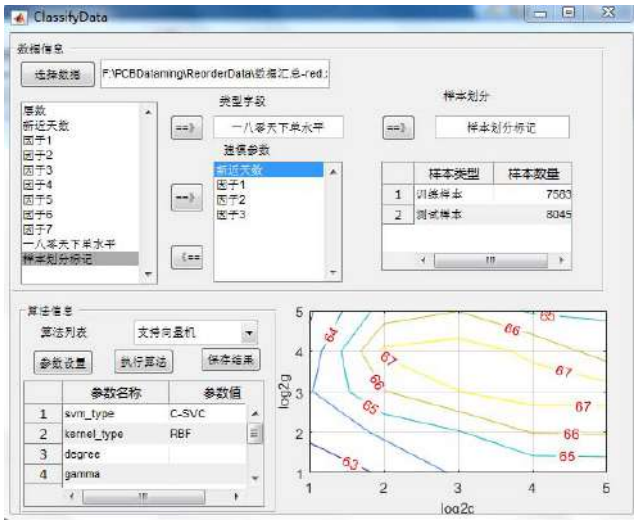
基于模块划分和待实现核心功能，梳理确定各算法核心原理及其关键调控参数，以指导开发时的页面封装、参数传递和代码编写，同时指导学生理解算法原理、查看源代码、掌握具体实现、配置优化参数等。表 1 给出了分类、预测模块的部分算法及其核心原理和调控参数，结合图 1 所示框架可帮助学生系统地理解挖掘流程、挖掘原理及其具体实现。

在图 2 功能模块基础上，系统菜单整体被划分为基本统计、数据预处理、分类模型、预测模型、聚类模型。数据预处理菜单包括数据转换、异常值筛选和数据消减等子菜单和相应的二级子菜单。每个子菜单对应一个页面，可独立完成相应的预处理功能。分类、预测和聚类模型菜单下分别对应一个执行页面，具体实现机制可在对应执行页面中的算法列表中选择调用。各页面可完成数据（或配置文件）选择、参数配置（弹出参数设置页面）、结果（包括中间迭代结果）可视化及最终结果输出操作等。学生可参考总体框架扩展开发新的实现机制或关联分析、时序分析等模块。

系统采用 MATLAB 和 Python 联合开发，主体框架、核心算法和可视化采用 MATLAB 及其开源包实现，部分算法则采用 Python 实现，如特征选择（属性重要性量化分析）算法需调用机器学习库“sklearn”来实现。

图 3 给出了支持向量机（SVM）参数设置及其执行页面。算法利用 MATLAB 版 LibSVM 实现。参数设置页面对 LibSVM 库中参数进行封装，包括 SVM 模型类型，核函数类型，gamma（多项式核函数、径向基核函数、Sigmoid 均需选择该参数），阶数（多项式核函数），是否输出分类概率（默认为 0、1），权重，系数（核函数中的 coef0 设置，针对多项式/sigmoid 核函数中的 *r*，默认为 0），惩罚因子（设置 C-SVC，e-SVR 和 v-SVR 的参数，损失函数 C，默认为 1），*n* 折验证（n-fold 交互检验模式，*n* 为 fold 的

(a) SVM 参数设置页面



(b) 基于 SVM 分类执行页面

图 3　SVM 参数设置与算法执行

个数，必须大于等于 2），并基于交叉验证利用执行页面的梯度图优选损失函数 C 和 gamma。

3 系统使用范式设计

为指导教师示教和学生实践，构建以教师为引导、以学生为主体的使用范式，具体如图 4 所示。

教师将结合具体业务需求对数据准备、挖掘流程、挖掘机制与核心原理以及开发实现等进行整体性讲授，并基于该平台演示综合性案例，具象化挖掘全流程、算法原理及其结果。同时将基于案例提出问题，布置、分配任务，促使学生自主学习，并指导学生使用此系统开展实践。

学生可依据图 1 所示流程框架与任务划分分解待挖掘任务；基于任务开展分工合作，完成各阶段不同实现机制的选择；通过源代码查阅，理解实现机制的原理、参数配置与调优；并根据真实挖掘需求开展全流程的综合性挖掘实践。也可在此平台上进行扩展开发，尝试实现新的机制，最后通过分享、讨论等方式强化学习效果。在师生互动中，教师主要起点评、启发和引导作用。

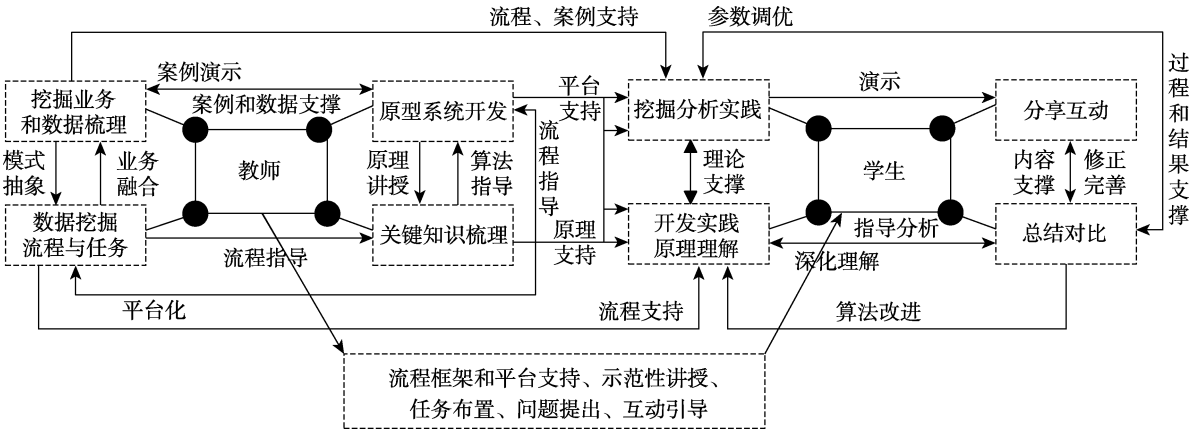


图 4 系统使用范式

4 教师示教与学生实践

基于该系统针对机械设计制造与自动化、电子、自动化和车辆工程等专业学生开设数据挖掘及其应用选修课程。采用问题驱动教学模式[15]，结合某印制电路板（printed circuit board，PCB）企业具体需求及工业软件所采集与累计的结构化数据（工程技术数据、资源数据、生产计划数据、制造结果数据、交易服务数据）开展教师示教和学生挖掘实践。教师要结合 PCB 样板订单和生产工艺特点，梳理出报废率预测和重复订单分类、聚类等业务需求，联合企业准备相应数据，指导学生按照挖掘流程开展挖掘分析，包括基本统计分析、预处理、挖掘和评价解释等环节。

以 PCB 报废率预测（本质是工艺水平预测和优化问题）为例，教师要先系统地介绍 PCB 生产工艺流程，使学生明确 PCB 生产制造工艺所涉及的对象数据，优化对象（工艺过程、工序、参数优化等），改进目标（质量、成本、时间等），挖掘分析方法及实现与测试方法等。再基于具体业务进一步演示部分算法，如基于图 5 菜单介绍系统模块和挖掘流程，结合 PCB 相关属性介绍关联业务，并开展基本统计。还要结合图 6 页面讲授基于 Boxplot 的异常值筛选，基于图 7 页面介绍基于线性回归、相关性、Lasso 回归、Ridge 回归、随机森林、互信息熵、特征消减等算法的关键特性优选机制。

在数据预处理基础上，以 3—5 人为一组分配挖掘任务，学生可基于所开发的系统尝试不同的挖掘算法，并理解、解释执行结果。图 8 给出了学生基于预测页面开展 PCB 报废率预测的执行效果，相应页面可设置迭代次数和 L2 正则化系数等参数。在该页面算法列表中，学生还可选择 ANN、支持向量机、随机森林等机制。迭代过程通过图形形象化展示，还可保存执行过程和最终结果数据以便进一步进行扩展分析。

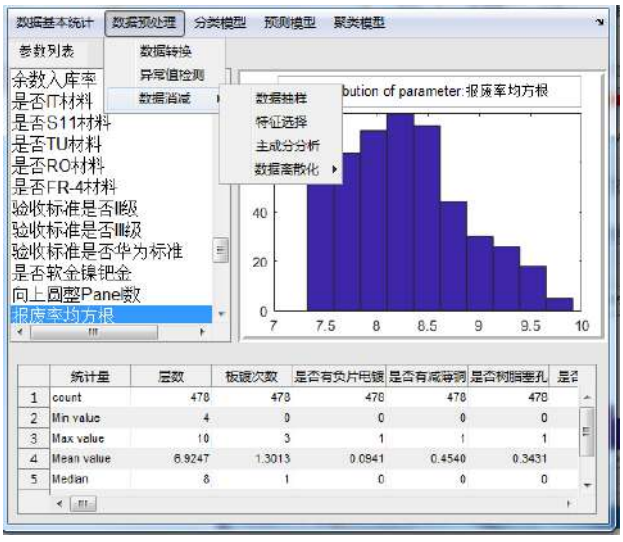


图 5 系统菜单和基本统计分析

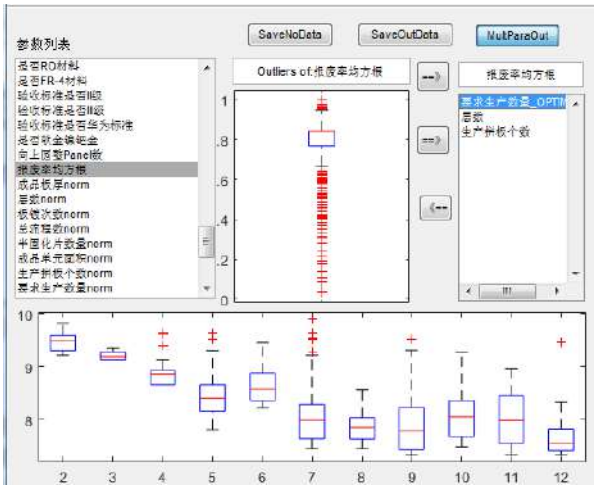


图 6　异常值筛选

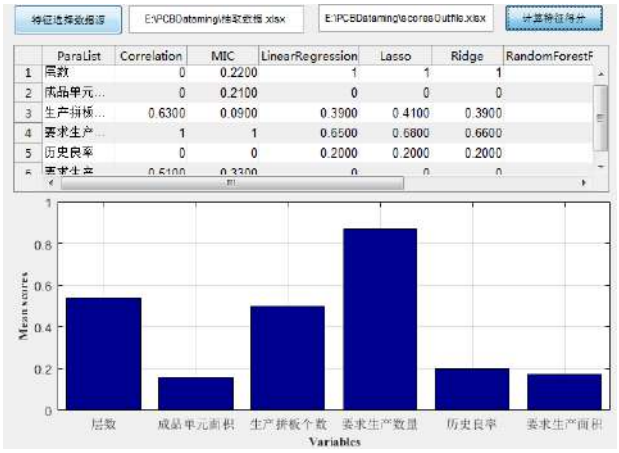


图 7　属性重要性定量分析

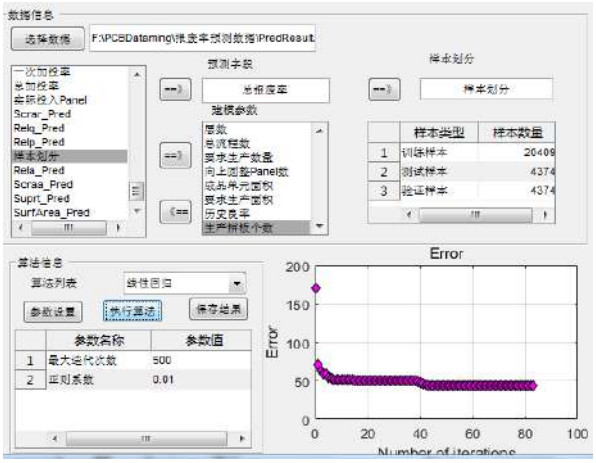


图 8　预测页面

以重复订单分类为例，学生开展了数据基本统计分析、预处理，并尝试了不同的挖掘分析机制。图 9 给出了学生基于主成分分析对 PCB 重复订单分类相关属性进行降维处理的页面，图 10—13 给出了学生尝试 ANN、CART、随机森林和贝叶斯网络等不同实现机制的执行页面。

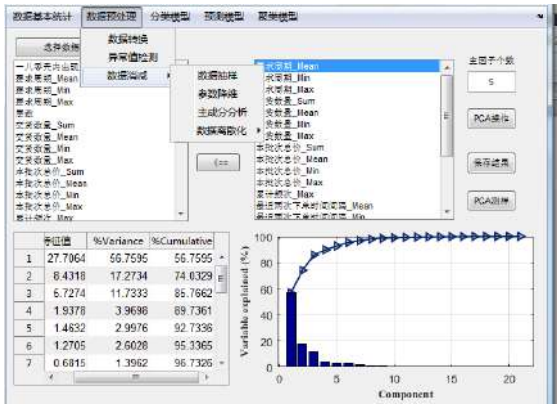


图 9　主成分分析

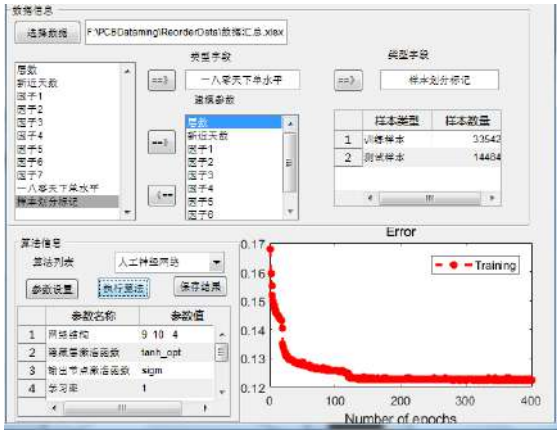


图 10　基于 ANN 分类

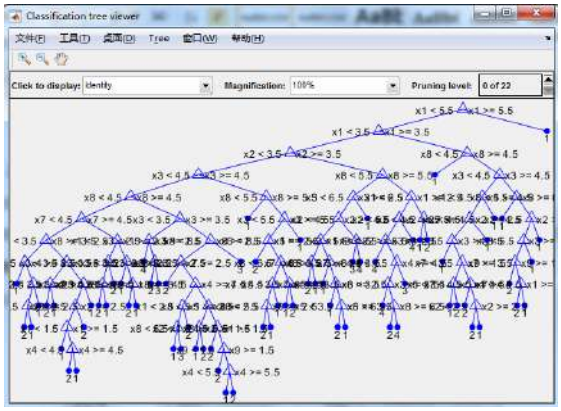


图 11　CART 分类生成的决策树页面

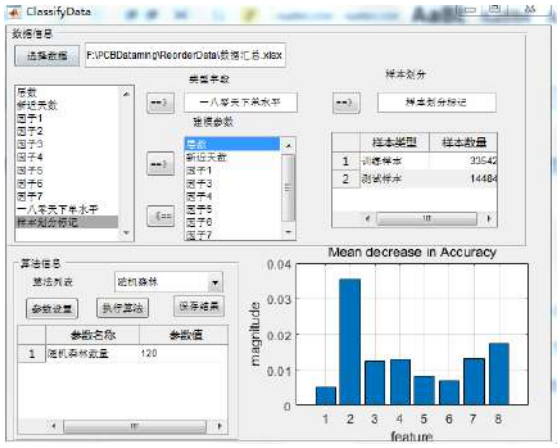


图 12　随机森林分类执行页面

图 13　贝叶斯分类执行页面

学生在教师的指导下还尝试了 K-means、DBSCAN、FCM、Agglomerative、waveCluster 等聚类机制，部分学生合作扩展开发了 CFSFDP、自组织映射等聚类机制。所有任务完成后，在最后一次课上按组进行成果分享，教师将进行点评，指出所存在的问题。

我们针对 2019、2020 年在此平台开展数据挖掘实践的 52 名学生进行了问卷调查。调查结果显示，88.46%的学生认为按照系统使用范式的示教和实践活动有利于系统地理解数据挖掘流程、挖掘任务、挖掘原理和实现机制，94.2%的学生认为所开发的平台有利于快速查看源代码、掌握算法原理、进行扩展开发及进行综合性挖掘实践。

5　结语

所建立的以流程为核心的挖掘任务框架可指导系统模块划分，结合优选的挖掘机制及其关键原理与调控参数梳理可帮助学生系统地理解挖掘流程、挖掘原理及具体实现，并可指导教师的示教和学生的综合性实践。

以挖掘流程和所设计的系统范式为指导，学生在包括基本统计分析、预处理、狭义挖掘算法、评价解释等功能的可扩展平台上，可查阅源代码、进行扩展开发和综合性挖掘分析，解决了单一算法小程序难以支持贯穿于数据挖掘全流程课程的综合性挖掘实践问题，也避免了封闭软件所带来原理和实现相互脱节问题。后续还将进一步扩展新模块和新功能。

参考文献(References)

[1] 吴爱华，侯永峰，杨秋波，等. 加快发展和建设新工科主动适应和引领新经济[J]. 高等工程教育研究，2017(1): 1–9.

[2] 李培根. 工科何以而新[J]. 高等工程教育研究，2017(4): 1–4, 15.

[3] 林健. 第四次工业革命浪潮下的传统工科专业转型升级[J]. 高等工程教育研究，2018(4): 1–10, 54.

[4] 吴贺俊，饶洋辉. 面向新工科的大数据专业课程建设[J]. 中国大学教学，2019(4): 34–37.

[5] 李春江. 大数据环境下的数据挖掘课程教学探索[J]. 黑龙江教育（理论与实践），2016(4): 54–55.

[6] 刘波，蔡燕斯，钟少丹. 大数据背景下数据挖掘课程实践教学的探索[J]. 高教学刊，2019(18): 124–125, 128.

[7] 范玉雷，杨良怀，高楠，等. 面向解决复杂工程问题的“大数据与数据挖掘”教学研究[J]. 中国信息技术教育，2019(10): 106–109.

[8] 郭传好. 需求驱动的数据挖掘课程教学改革研究[J]. 中国教育信息化，2019(21): 88–90.

[9] 卫志华，孔思尹，丁志军，等. 新工科背景下数据挖掘课程综合性实验设计[J]. 计算机教育，2020(3): 127–130, 135.

[10] 孙瑞娜. 基于网络教学平台的“混合式”教学模式研究：以数据挖掘课程为例[J]. 教育现代化，2020, 7(6): 67–69.

[11] HAN J W, KAMBER M, PEI J. Data mining: Concepts and techniques[M]. 3rd ed. Waltham, Massachusetts: Morgan Kaufmann, 2011.

[12] ZHANG Y, REN S, LIU Y, et al. A framework for big data driven product lifecycle management[J]. Journal of Cleaner Production, 2017(159): 229–240.

[13] LV S P, KIM H, ZHENG B B, et al. A review of data mining with big data towards its applications in the electronics industry[J]. Applied Sciences, 2018, 8(4): 582–615.

[14] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492–1496.

[15] 朱少民. 软件测试课程的问题驱动教学模式探索[J]. 中国大学教学，2018(10): 32–36.

SCAULIB202626300

检索证明

根据委托人提供的论文材料，委托人华南农业大学工程学院 吕盛坪(学科类型:自然科学) 1 篇论文收录情况如下表。

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	数据挖掘教学示教与学生实践系统设计与开发	实验技术与管理 出版年：2021 出版日期：2021-08-19 16:48 卷期：38 08 页码：- 文献号： 文献类型：期刊论文	第一作者	C 类	华南农业大学工程学院	北大核心	无	无

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明:如未盖章,报告无效

检索员：尹银怀

华南农业大学图书馆

2026-03-16

华南农业大学文件

华南农教〔2021〕28号

关于公布2021年华南农业大学教学成果获奖名单的通知

各学院、部处、各单位:

根据《关于做好2021年校级教学成果奖励工作的通知》(华南农教〔2021〕6号)要求，经组织申报、单位推荐、专家评审、校内公示等环节，确定“基于‘三全育人’的知农爱农新型人才培养创新与实践”等32项成果(含2项研究生教学成果)获2021年校级教学成果一等奖；“基于‘听说读写行考’的《概论》立体化建设与实践”等54项成果(含3项研究生教学成果)获2021年校级教学成果二等奖，现予以公布(详见附件)。

开展教学成果奖励活动是对学校人才培养工作和教育教学改革成果的检阅和展示。本次获奖的项目是全校教师员工在教学及管理工作岗位上，经过多年艰苦努力获得的创造性劳动成果，

充分体现了近年来我校在教育教学改革方面所取得的重大进展。各学院、各单位要结合实际，认真做好获奖成果的学习、借鉴与应用，进一步加强教学工作，深化教学改革，不断提高我校的整体办学水平和办学效益。

特此通知。

附件：2021年华南农业大学教学成果获奖名单

华南农业大学

2021年5月20日

（联系人：曹广祥、李艳丽，电　话：85280052）

附件

2021年华南农业大学教学成果获奖名单

校级一等奖名单

成果编号	成果名称	成果主持人姓名	成果主要完成人姓名（不含主持人）	备注
JXCG21001	基于“三全育人”的知农爱农新型人才培养创新与实践	王斌伟	叶晖有、朱蕾、赵凤、王海林、项赠、杨玉浩、张运红	
JXCG21002	“三融合、三自主”农业特色生物学卓越创新人才培养模式的探索与实践	邓诣群	朱国辉、陈乐天、文继开、曹广祥、郝刚、王汝干、陈超	
JXCG21003	农业高校在线教育内部质量保障体系的创新与实践	欧阳俊	张运红、朱蕾、郑向玲、李艳丽、徐汹涛、陈国菊	
JXCG21004	农业高校“双链”联动“五融合”新工科人才培养模式研究与实践	王海林	闫国琦、林伟波、朱鸿运、李君、高锐涛、王红军、郭涵	
JXCG21005	基于“三引领”理念的高素质应用型农科人才培养模式探索与实践	陈少雄	陈永晴、梁廷君、傅梅芳、代啟贵、蓝学明、朱鸿运、杨正喜、朱斌、顾方愉	
JXCG21006	精准招生 靶向培养 溯源监控：助力乡村振兴的农业硕士培养模式创新与实践	刘雅红	孟成民、陈翱、庄楚雄、彭新湘、王曙光、王忠、侯辉萍、陈华全、徐江	研究生教学成果
JXCG21007	契合新农科理念的跨学科基础实践教学一体化改革与实践	库夭梅	陈建军、谢虎、劳媚媚、曹广祥、文晟、刘小波、徐军、许奕进、陈海波	
JXCG21008	基于核心能力构建的新农科教师发展平台建设的创新与实践	朱　蕾	欧阳俊、张运红、孙京臣、邵家声、苏弟华、马勇江、李云锋	
JXCG21009	守望互助、产教融合培养家具产业高素质专业人才模式创建与实践	胡传双	郭琼、孙理超、宋杰、易欣、陈建、涂登云、欧荣贤、王清文	
JXCG21010	“三三三”制农科高素质创新人才培养模式构建与实践	曾曙才	黄文勇、潘军、马启彬、谭莹、朱国辉、曹广祥、刘月秀、何晓芳	
JXCG21011	《畜产食品工艺学》教材-基于国家级慕课的纸数融合教材建设与应用	蒋爱民	周佺、钟青萍、郭善广、肖南、郑华、黄继青、黄文勇	
JXCG21012	契合新农科建设需求的动物生产类专业课“金课”建设的创新与实践	张永亮	陈婷、孙加节、罗君谊、习欠云、邓铭	
JXCG21013	迈向新农科 构建国家级基础实验平台“一核多维”育人模式的研究与实践	陈建军	羊海军、库夭梅、方颖、郭海滨、李楠、李淮源、詹福建、俞新华	
JXCG21014	“两联动两融合”水产类新型人才培养的探索与实践	杨慧荣	黄嘉琪、秦启伟、辛其兴、但学明、王俊、赵会宏、孙际佳	

JXCG21069	公共管理专业乡村振兴人才“学-赛-研-练”链式培养模式的创新与实践	杨正喜	张运红、朱汉平、邹静琴、贾海薇、刘辉	
JXCG21070	面向“新三农”提升经管类人才运用本土化理论解决实际问题的高阶能力	米运生	万俊毅、谭莹、李鎏、王丽萍、陈风波、刘仁和、杨学儒、陈艳艳、李巧璇	
JXCG21071	生物化学线上下混合式教学的改革与创新	巫光宏	朱国辉、初志战、吴骏、陈庆梅、赵利锋	
JXCG21072	“两山”理论视域下大学生低碳素质教育长效机制的构建与创新	吕辉雄	陈少华、郑芊、梁瑜海、陈杨梅、陈火君	
JXCG21073	“四位一体”混合教学保障体系的建构与实践	张　芸	周恩浩、龚正想、黄志宏、李玉玲	
JXCG21074	面向绿色、健康与可持续大湾区优质生活圈的规划人才培养与教学实践	杨文越	王婷、王凌、彭昌操、章家恩、叶昌东、赵晓铭	
JXCG21075	面向创新人才培养的风景园林专业建筑类课程群的教学实践研究	潘建非	高伟、李梦然、江帆影	
JXCG21076	深度融合现代信息技术的高校数学类公共基础课程思政建设的探索与创新	肖　莉	周燕、杨志程、丁仕虹、杨德贵、岑冠军、夏强、谢韶锋	
JXCG21077	面向新工科电子科学与技术创新人才培养的实验实践教学模式研究与探索	罗　霞	罗阔、李震、刘洪山、王建、谢家兴	
JXCG21078	三全育人视域下农业院校资助育人体系实践与创新	赵　凤	殷舒、韩丽、周志荣、邵家声、田立、史锐	
JXCG21079	课程思政背景下“四位一体”情商培养模式的探索与实践	吴　琪	蔡传钦、欧阳俊、何凯、林媛、苏冠贤	
JXCG21080	构建多维融合的信息素养教育模式的创新与实践	刘　锋	张琴、吴贤奇、刘熙东、邓智心、何效平、欧群	
JXCG21081	“专项教学+体质监测+阳光体育”三位一体有效推进学校体育课堂教育教学改革发展	陈华东	钞飞侠、李嘉鹏、赵东升、潘林权	
JXCG21082	面向新型工科专业学生数据科学思维培养的数据挖掘课程建设与改革	吕盛坪	李庆、王海林、李君	
JXCG21083	蚕病学全英教学和管理模式应用	孙京臣	冯敏、何小敏、欧阳俊、任菲菲、张以农、王雄	
JXCG21084	粤港澳大湾区水产养殖专业高水平科研创新人才培养的探索与实践	王　俊	刘文生、杨慧荣、秦启伟、辛其兴、陈广龙	研究生教学成果
JXCG21085	四链协同驱动农业院校化工材料类研究生拔尖人才培养模式建设与实践	杨卓鸿	倪春林、袁腾、胡洋、张超群、杨宇	研究生教学成果
JXCG21086	研究生“水产养殖技术”课程的“学研产”教学模式	孙红岩	于宗赫、周胜、但学明、李言伟	研究生教学成果

南农业大学

科教学成果奖

奖证书

获奖成果： 面向新工科建设的涉农高校《汽车理论》课程教学改革与实践

获 奖 者： 郭嘉明、李　君、李　庆、吕盛坪、王　昱、吕恩利、武　涛、曾志雄

获奖等级： 二等奖

证书编号： JXCG24082

2025年9月9日

二、科研项目---主持项目清单

项目批准号	52275487
申请代码	E0510
归口管理部门	
依托单位代码	51064208A0499-0932

国家自然科学基金

资助项目计划书

（预算制项目）

资助类别：面上项目

亚类说明：

附注说明：

项目名称：多耦合影响广义作业车间调度模型构建与优化

直接费用：54万元　　执行年限：2023.01-2026.12

负 责 人：吕盛坪

通讯地址：广东省广州市天河区五山路483号工程学院北210B

邮政编码：510642　　电　　话：

电子邮件：

依托单位：华南农业大学

联 系 人：唐家林　　电　　话：

填表日期：2022年09月13日

国家自然科学基金委员会制

Version：1.004.537

国家自然科学基金资助项目计划书填报说明

（预算制项目）

一、项目负责人收到《国家自然科学基金资助项目批准通知》（以下简称《批准通知》）后，请认真阅读本填报说明，参照国家自然科学基金相关项目管理办法和新修订的《国家自然科学基金资助项目资金管理办法》（以下简称《资金管理办法》，请查阅国家自然科学基金委员会官方网站首页“政策法规”栏目），按《批准通知》的要求认真填写和提交《国家自然科学基金资助项目计划书》（以下简称《计划书》）。

二、填写《计划书》时要科学严谨、实事求是、表述清晰、准确。《计划书》经国家自然科学基金委员会相关项目管理部门审核批准后，将作为项目研究计划执行、检查和验收的依据。

三、《计划书》各部分填写要求如下：

（一）简表：由系统自动生成。

（二）摘要及关键词：各类获资助项目都应当填写中、英文摘要及关键词。

（三）项目组主要成员：计划书中列出姓名的项目组主要成员由系统自动生成，与申请书原成员保持一致，不可随意调整。如果《批准通知》所附“项目评审意见及修改意见表”中“修改意见”栏目有调整项目组成员相关要求的，待项目开始执行后，按照项目成员变更程序另行办理。

（四）资金预算表：根据批准的项目资助额度，按规定调整项目预算，并按照《国家自然科学基金项目计划书预算表编制说明》填报资金预算表和预算说明书。

（五）正文：

1. 面上项目、地区科学基金项目：如果《批准通知》所附“项目评审意见及修改意见表”中“修改意见”栏目没有修改要求的，只需选择“研究内容和研究目标按照申请书执行”即可；如果《批准通知》中上述栏目明确要求调整研究期限或研究内容等的，须选择“根据研究方案修改意见更改”并填报相关修改内容。

2. 重点项目、重点国际（地区）合作研究项目、重大项目、国家重大科研仪器研制项目、原创探索计划项目：须选择“根据研究方案修改意见更改”，根据《批准通知》的要求填写研究（研制）内容，不得自行降低、更改研究目标（或仪器研制的技术性能与主要技术指标、验收技术指标等）或缩减研究（研制）内容。此外，还要突出以下几点：

（1）研究的难点和在实施过程中可能遇到的问题（或仪器研制风险），拟采用的研究（研制）方案和技术路线；

（2）项目主要参与者分工，合作研究单位（如有）之间的关系与分工，重大项目还需说明课题之间的关联；

（3）详细的年度研究（研制）计划。

3. 创新研究群体项目：须选择“根据研究方案修改意见更改”，按下列提纲撰写：

（1）研究方向；

（2）结合国内外研究现状，说明研究工作的学术思想和科学意义（限两个页面）；

（3）研究内容、研究方案及预期目标（限两个页面）；

（4）年度研究计划；

（5）研究队伍的组成情况。

4. 基础科学中心项目：须选择“根据研究方案修改意见更改”，根据《批准通知》的要求和现场考察专家组的意见和建议，进一步完善并细化研究计划，按下列提纲撰写：

（1）五年拟开展的研究工作（包括主要研究方向、关键科学问题与研究内容）；

（2）研究方案（包括骨干成员之间的分工及合作方式、学科交叉融合研究计划等）；

（3）年度研究计划；

（4）五年预期目标和可能取得的重大突破等；

（5）研究队伍的组成情况。

5. 对于其他类型项目，参照面上项目的方式进行选择和填写。

简表

项目负责人信息	姓　　名	吕盛坪	性 别	男	出生年月		民 族	汉族
	学　　位	博士			职称	副教授		
	是否在站博士后	否			电子邮件	lvshengping@scau.edu.cn		
	电　　话	020-85280752			个人网页			
	工 作 单 位	华南农业大学						
	所 在 院 系 所	工程学院						
依托单位信息	名　　称	华南农业大学					代码	51064208A0499
	联 系 人	唐家林			电子邮件	kycjhk@scau.edu.cn		
	电　　话	020-85280070			网站地址	http://kjc.scau.edu.cn/		
合作单位信息	单 位 名 称							
项目基本信息	项 目 名 称	多耦合影响广义作业车间调度模型构建与优化						
	资 助 类 别	面上项目			亚 类 说 明			
	附 注 说 明							
	申 请 代 码	E0510:制造系统与智能化						
	基 地 类 别							
	执 行 年 限	2023.01-2026.12						
	直 接 费 用	54万元						

项目摘要

中文摘要：

制造系统各种耦合影响和柔性工艺约束给制造系统运行优化带来巨大挑战，多耦合影响广义作业车间调度（GJSPMC）就是其有待解决的关键问题之一。项目针对广泛存在于各离散制造业的GJSPMC问题开展深入研究。首先，分析GJSPMC问题特性和多耦合影响因素，基于多色集和混合整数规划方法构建约束关系模型和调度优化模型。然后，综合耦合影响构建广义邻域结构，提出融合广义邻域结构的局部搜索方法和全局搜索算法结合的复合优化机制与多目标平衡策略。接着，考虑工艺柔性影响调整GJSPMC模型和广义邻域结构，设计高效复合优化算法。最后，结合电子产品检测车间实际完成原型系统开发与应用验证。将在GJSPMC模型构建、领域知识挖掘和优化求解等方面取得系列创新成果，为多耦合影响的制造系统运行优化提供新的基础理论方法和关键技术支撑，具有重要的理论研究意义和工程应用价值。

Abstract：

Various coupling effects and flexible process plan constraints of manufacturing systems pose great challenges to the optimization of manufacturing systems. Generalized job shop scheduling problem considering multiple coupling effects (GJSPMC) is one of critical unsolved problems. In this project, the deep researches on GJSPMC which widely exists in various discrete manufacturing industries will be conducted. Firstly, the characteristics and multi-coupling influencing factors of GJSPMC will be analyzed, and a constraint relation model and a scheduling optimization model will be constructed based on polychromatic sets and mixed integer programming methods. Secondly, a generalized neighborhood structure (GNS) will be constructed by synthesizing the coupling effects. On this basis, hybrid optimization mechanisms and multi-objective balance strategies are proposed by combining the GNS-based local search methods with global search algorithms. Thirdly, the GJSPMC model and generalized neighborhood structure are adjusted considering flexible process plans. Finally, the prototype system development and application verification are conducted based on the above model and algorithm, and the effectiveness of the proposed method and developed software will be verified in an electronic product testing workshop. A series of original achievements will be obtained in the construction of GJSPMC model, domain knowledge mining and optimization mechanisms, which will provide new theory and key technologies support for the operation optimization of manufacturing systems influenced by multiple couplings effects. Therefore, this project has important theoretical significance and engineering application potential.

关键词(用分号分开)：车间调度；车间作业调度；柔性作业车间调度；广义作业车间调度；广义邻域结构

Keywords(用分号分开)： Shop scheduling; Job shop scheduling; Flexible job shop scheduling; Generalized job shop scheduling; Generalized neighborhood structure

项目组主要成员

编号	姓名	出生年月	性别	职称	学位	单位名称	电话	证件号码	项目分工	每年工作时间（月）
1	吕盛坪		男	副教授	博士	华南农业大学	020-85280752		项目负责人	10
2	金鸿		男	讲师	博士	华南农业大学			约束和优化模型构建	6
3	姜焰鸣		男	讲师	博士	华南农业大学	020-85280222		广义邻域结构设计	8

总人数	高级	中级	初级	博士后	博士生	硕士生
9	1	2			1	5

国家自然科学基金预算制项目预算表

项目批准号：52275487　　项目负责人：吕盛坪　　金额单位：万元

序号	科目名称	金额
1	一、基金资助项目直接费用合计	54.0000
2	1、 设备费	7.0000
3	其中：设备购置费	7.0000
4	2、 业务费	30.0000
5	3、 劳务费	17.0000
6	二、 其他来源资金	0.0000
7	三、 合计	54.0000

注：请按照项目研究实际需要合理填写各科目预算金额。

国家自然科学基金项目负责人、依托单位承诺书

国家自然科学基金项目负责人承诺书

本人郑重承诺：我接受国家自然科学基金的资助，严格遵守中共中央办公厅、国务院办公厅《关于进一步加强科研诚信建设的若干意见》《关于进一步弘扬科学家精神加强作风和学风建设的意见》《关于加强科技伦理治理的意见》等规定，及国家自然科学基金委员会关于资助项目管理、项目资金管理等各项规章，在《计划书》填写及项目执行过程中：

（一）按照《批准通知》《国家自然科学基金资助项目计划书填报说明》的要求填写《计划书》，未自行降低、更改目标任务或约定要求，或缩减研究（研制）内容；

（二）树立“红线”意识，严格履行科研合同义务，按照《计划书》负责实施本项目（批准号：52275487），切实保证研究工作时间，按时报送有关材料，及时报告重大情况变动，不违规将科研任务转包、分包他人，不以项目实施周期外或不相关成果充抵交差；

（三）遵守科研诚信、科技伦理规范和学术道德，认真开展研究工作，对资助项目发表的论著和取得的研究成果按规定进行标注，不在非本项目资助的成果或其他无关成果上标注本项目批准号，反对无实质学术贡献者“挂名”，不在成果署名、知识产权归属等方面侵占他人合法权益，并如实报告本人及项目组成员发生的违背科研诚信要求的任何行为；

（四）尊重科研规律，弘扬科学家精神，严谨求实，追求卓越，反对浮夸浮躁、投机取巧，不人为夸大学术或技术价值，不传播未经科学验证的现象和观点；

（五）将项目资金全部用于与本项目研究工作相关的支出，并结合科研活动需要，科学合理安排项目资金支出进度；

（六）做好项目组成员的教育和管理，确保遵守以上相关要求。

如违背上述承诺，本人愿接受国家自然科学基金委员会和相关部门做出的各项处理决定。

项目负责人（签字）：

2022年 9月27日

依托单位科研管理部门：	依托单位财务管理部门：
负责人（签章）： 2022年 9月30日	负责人（签章）： 2022年 9月30日

国家自然科学基金项目依托单位承诺书

我单位同意承担上述国家自然科学基金项目，将保证项目负责人及其研究团队的稳定和研究项目实施所需的条件，严格遵守国家自然科学基金委员会有关资助项目管理、项目资金管理、科研诚信管理和科技伦理管理等各项规定，并督促实施。

依托单位（公章）

2022年 9月30日

国家自然科学基金资助项目签批审核表

本栏目由自然科学基金委填写

科学处审查意见：

同意按计划执行

负责人（签章）叶鑫

年 月 日

2022年 12月21日

科学部审查意见：

同意科学处意见

负责人（签章）：

年 月 日

2022年 12月21日

关于国家自然科学基金资助项目批准及有关事项的通知

吕盛坪　先生/女士：

根据《国家自然科学基金条例》的规定和专家评审意见，国家自然科学基金委员会（以下简称自然科学基金委）决定批准资助您的申请项目。项目批准号：51605169，项目名称：面向模具生产的工艺与车间调度紧耦合集成规划，直接费用：20.00万元，项目起止年月：2017年01月至2019年12月，有关项目的评审意见及修改意见附后。

请尽早登录科学基金网络信息系统（https://isisn.nsfc.gov.cn），获取《国家自然科学基金资助项目计划书》（以下简称计划书）并按要求填写。对于有修改意见的项目，请按修改意见及时调整计划书相关内容；如对修改意见有异议，须在计划书电子版报送截止日期前提出。**注意：请严格按照《国家自然科学基金资助项目资金管理办法》填写计划书的资金预算表，其中，劳务费、专家咨询费科目所列金额与申请书相比不得调增。**

计划书电子版通过科学基金网络信息系统（https://isisn.nsfc.gov.cn）上传，由依托单位审核后提交至自然科学基金委进行审核。审核未通过者，返回修改后再行提交；审核通过者，打印为计划书纸质版（一式两份，双面打印），由依托单位审核并加盖单位公章后报送至自然科学基金委项目材料接收工作组。计划书电子版和纸质版内容应当保证一致。

向自然科学基金委提交和报送计划书截止时间节点如下：

1、提交计划书电子版截止时间为**2016年9月11日16点**（视为计划书正式提交时间）；

2、提交计划书电子修改版截止时间为**2016年9月18日16点**；

3、报送计划书纸质版截止时间为**2016年9月26日16点**。

请按照以上规定及时提交计划书电子版，并报送计划书纸质版，未说明理由且逾期不报计划书者，视为自动放弃接受资助。

附件：项目评审意见及修改意见

国家自然科学基金委员会
工程与材料科学部
2016年8月17日

项目批准号	51605169
申请代码	E051005
归口管理部门	
依托单位代码	51064208A0499-0932

国家自然科学基金委员会

资助项目计划书

资助类别：青年科学基金项目

亚类说明：

附注说明：

项目名称：面向模具生产的工艺与车间调度紧耦合集成规划

直接费用：20万元　　执行年限：2017.01-2019.12

负 责 人：吕盛坪

通讯地址：广东省广州市天河区五山路483号工程学院北402C

邮政编码：510642　　电　　话：

电子邮件：

依托单位：华南农业大学

联 系 人：全锋　　电　　话：

填表日期：2016年08月20日

国家自然科学基金委员会制

Version：1.008.627

国家自然科学基金委员会资助项目计划书填报说明

一、项目负责人收到《关于国家自然科学基金资助项目批准及有关事项的通知》（以下简称《批准通知》）后，请认真阅读本填报说明，参照国家自然科学基金相关项目管理办法及《国家自然科学基金资助项目资金管理办法》（请查阅国家自然科学基金委员会官方网站首页“政策法规”-“管理办法”栏目），按《批准通知》的要求认真填写和提交《国家自然科学基金委员会资助项目计划书》（以下简称《计划书》）。

二、填写《计划书》时要求科学严谨、实事求是、表述清晰、准确。《计划书》经国家自然科学基金委员会相关项目管理部门审核批准后，将作为项目研究计划执行和检查、验收的依据。

三、《计划书》各部分填写要求如下：

（一）简表：由系统自动生成。

（二）摘要及关键词：各类获资助项目都必须填写中、英文摘要及关键词。

（三）项目组主要成员：计划书中列出姓名的项目组主要成员由系统自动生成，与申请书原成员保持一致，不可随意调整。如果批准通知中“项目评审意见及修改意见表”中“对研究方案的修改意见”栏目有调整项目组成员相关要求的，待项目开始执行后，按照项目成员变更程序另行办理。

（四）资金预算表：按批准资助的直接费用填报资金预算表和预算说明书，其中的劳务费、专家咨询费金额不应高于申请书中相应金额。国家重大科研仪器研制项目、重大项目还应按照预算评审后批复的直接费用各科目金额填报资金预算表、预算说明书及相应的预算明细表。

（五）正文：

1. 面上项目、青年科学基金项目、地区科学基金项目：如果《批准通知》中没有修改要求的，只需选择“研究内容和研究目标按照申请书执行”即可；如果《批准通知》中“项目评审意见及修改意见表”中“对研究方案的修改意见”栏目明确要求调整研究期限和研究内容等的，须选择“根据研究方案修改意见更改”并填报相关修改内容。

2. 重点项目、重点国际（地区）合作研究项目、重大项目、国家重大科研仪器研制项目：须选择“根据研究方案修改意见更改”，根据《批准通知》的要求填写研究（研制）内容，不得自行降低、更改研究目标（或仪器研制的技术性能与主要技术指标以及验收技术指标）或缩减研究（研制）内容。此外，还要突出以下几点：

（1）研究的难点和在实施过程中可能遇到的问题（或仪器研制风险），拟采用的研究（研制）方案和技术路线；

（2）项目主要参与者分工，合作研究单位之间的关系与分工，重大项目还需说明课题之间的关联；

（3）详细的年度研究（研制）计划。

3. 国家杰出青年科学基金、优秀青年科学基金和海外及港澳学者合作研究基金项目：须选择“根据研究方案修改意见更改”，按下列提纲撰写：

（1）研究方向；

（2）结合国内外研究现状，说明研究工作的学术思想和科学意义（限两个页面）；

（3）研究内容、研究方案及预期目标（限两个页面）；

（4）年度研究计划；

（5）研究队伍的组成情况。

4. 对于其他类型项目，参照面上项目的方式进行选择和填写。

简表

申请者信息	姓　　名	吕盛坪	性 别	男	出生年月		民 族	汉族
	学　　位	博士			职称	副教授		
	电　　话			电子邮件		lvshengping@scau.edu.cn		
	传　　真			个人网页				
	工 作 单 位	华南农业大学						
	所 在 院 系 所	工程学院						
依托单位信息	名　　称	华南农业大学					代码	51064208A0499
	联 系 人	全锋		电子邮件		kycjhk@scau.edu.cn		
	电　　话			网站地址		http://web.scau.edu.cn/kjc/		
合作单位信息	单　位　名　称							代　码
项目基本信息	项 目 名 称	面向模具生产的工艺与车间调度紧耦合集成规划						
	资 助 类 别	青年科学基金项目			亚 类 说 明			
	附 注 说 明							
	申 请 代 码	E051005:制造系统调度、规划与管理			E051002:数字化制造与智能制造			
	基 地 类 别							
	执 行 年 限	2017.01-2019.12						
	直 接 费 用	20万元						

项目摘要

中文摘要(500字以内)：

工艺与车间调度集成规划是整体上提高制造系统效率的潜在机制。本项目在深入研究模具工艺规划和车间调度问题结构特性基础上，提出了面向模具生产的集成规划新模式。针对现有相关理论方法不足和难以满足模具生产集成规划的特殊需求，基于对象化、多色集和混合整数规划建模理论，建立以特征对象为核心的集成规划约束与优化模型；设计易融入集成算法和扩展到动态场景的可行集成方案解析生成机制；基于保劣性选择父个体和衰退种群策略开发遗传算法与禁忌搜索结合的复合机制；引入Pareto独立标量适应度函数和向量评估方法开发多目标优化复合算法；调整集成规划模型和解析机制，开发针对不同动态场景的优化策略并以部分模具零件进行验证。目标是协同优化确定模具零件工艺与调度方案，消减目标冲突，快速响应动态场景，整体上提高制造系统效率。本研究对于理解模具生产规划问题和拓宽求解思路是一种新的尝试；为集成规划建模和静动态优化也提供了系统性创新机制。

关键词：模具制造；工艺与车间调度集成；集成规划模型；静态集成优化；动态集成优化

Abstract(limited to 4000 words)：

Integrated process planning and job shop scheduling (IPPS) is a potential mechanism to improve the whole efficiency of manufacturing systems. The mould-oriented IPPS will be established based on the structure property of mould manufacturing process planning and scheduling. However, current theory and methods for IPPS cannot completely meet the special requirement of mould manufacturing. The theory investigation and practice exploration for the key scientific problem of mould-oriented IPPS will be carried out in this research. An integrated constraint and optimization model with feature as its core based on Object -Oriented, Polychromatic Sets (PS) and Mixed Integer Programming modeling theory will be established; and an analytical mechanism for the generation of feasible integrated plan will be designed and embedded in static/dynamic optimization algorithm. On this basis, a hybrid method by combining Genetic Algorithm with Tabu Search will be developed by introducing the strategies of inferior parents selection and population degeneration; meanwhile, a hybrid multi-objective optimization algorithm for IPPS based on Pareto-based scale-independent fitness function and vector evaluated method will be proposed; dynamic optimization strategies for different situations will also be provided with adjusted integration model and analytical mechanism. Finally, the above proposed mechanisms will be verified by an integrated instance with many mould parts. As a result, process and scheduling plan will be optimized and determined collaboratively, and the confliction between different optimization objectives can be reduced; meanwhile, the dynamic situations can be dealt with quick response and the whole efficiency of manufacturing system can be improved. This research facilitates the understanding of mould production planning problem and broadens new planning attempt for the problem; and it also provides innovative systematic mechanism for the development of integrated model and static/ dynamic optimization of IPPS.

Keywords: Mould Manufacturing; Integrated Process Planning and Job Shop Scheduling; Integrated Planning Model; Static Integrated Optimization; Dynamic Integrated Optimization

项目组主要成员

编号	姓名	出生年月	性别	职称	学位	单位名称	电话	证件号码	项目分工	每年工作时间（月）
1	吕盛坪		男	副教授	博士	华南农业大学	020-85282860		项目负责人	10
2	金鸿		男	讲师	博士	华南农业大学	020-85282860		集成规划业务对象和约束模型构建	6
3	王昱		女	讲师	博士	华南农业大学	020-85282860		集成规划优化模型构建	6
4	杨径		男	硕士生	学士	华南农业大学	020-85282860		静态集成规划单目标优化机制设计与实现	6
5	王飞仁		男	硕士生	学士	华南农业大学	020-85282860		集成多目标优化机制研究与开发实现	6
6	徐岩		女	硕士生	学士	华南农业大学	020-85282860		动态集成规划机制研究与实现	6
7	李鹏飞		男	硕士生	学士	华南农业大学	020-85282860		动态集成优化机制研究和系统开发	6
8	范思宇		男	硕士生	学士	华南农业大学	020-85282860		系统框架和数据维护模块开发	6

总人数	高级	中级	初级	博士后	博士生	硕士生
8	1	2				5

国家自然科学基金资助项目签批审核表

<table>
<tr>
<td colspan="2">我接受国家自然科学基金的资助，将按照申请书、项目批准意见和计划书负责实施本项目（批准号：51605169），严格遵守国家自然科学基金委员会关于资助项目管理、财务等各项规定，切实保证研究工作时间，认真开展研究工作，按时报送有关材料，及时报告重大情况变动，对资助项目发表的论著和取得的研究成果按规定进行标注。

项目负责人（签章）：
年　　月　　日</td>
<td>我单位同意承担上述国家自然科学基金项目，将保证项目负责人及其研究队伍的稳定和研究项目实施所需的条件，严格遵守国家自然科学基金委员会有关资助项目管理、财务等各项规定，并督促实施。

依托单位（公章）
年　　月　　日</td>
</tr>
<tr>
<td rowspan="3">本栏目由基金委填写</td>
<td colspan="2">科学处审查意见：

建议年度拨款计划（本栏目为自动生成，单位：万元）：
<table>
<tr><td>年度</td><td>总额</td><td>第一年</td><td>第二年</td><td>第三年</td><td>第四年</td><td>第五年</td></tr>
<tr><td>金额</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table>
负责人（签章）：
年　　月　　日</td>
</tr>
<tr>
<td colspan="2">科学部审查意见：

负责人（签章）：
年　　月　　日</td>
</tr>
<tr>
<td colspan="2">相关局室审核意见：

负责人（签章）：
年　　月　　日</td>
</tr>
<tr>
<td>本栏目主要用于重大项目等</td>
<td colspan="2">委领导审批意见：

委领导（签章）：
年　　月　　日</td>
</tr>
</table>

51064201-0932

国家自然科学基金

资助项目准予结题通知

吕盛坪　同志：

您承担的国家自然科学基金项目：(面向模具生产的工艺与车间调度紧耦合集成规划)，批准号：(51605169)按有关规定已审核完毕，准予结题。

与本项目资助有关的后续成果，请您继续及时报送。

祝您在研究工作中取得更好的成绩！

国家自然科学基金委员会

工程与材料科学部

2020年4月24日

项目编号：	2021A1515012395
资助类别：	广东省自然科学基金-面上项目
文件编号：	粤基金字〔2021〕4号

广东省基础与应用基础研究基金项目验收书

项目名称：	定制化印制电路板生产缺陷关键影响特性识别与关联分析方法		
项目负责人：	吕盛坪	财政经费：	10（万元）
计划完成时间：	2021-01-01 至 2023-12-31		
实际完成时间：	2021-01-01 至 2023-12-31		
依托单位：	华南农业大学		
参与单位：			
验收形式：	材料验收		
联系人：	倪慧群	联系电话：	
填表日期：	2023-12-31		

广东省基础与应用基础研究基金委员会
二〇二〇年制

（广东科技微信公众号）　（查看验收书信息）　（受理纸质材料二维码）

一、项目人员信息表

项目负责人：				
姓名	证件号码	职称	承担任务	所在单位
吕盛坪		副教授	项目负责人	华南农业大学
主要研究人员：				
姓名	证件号码	职称	承担任务	所在单位
金鸿		讲师	企业需求分析，集成挖掘分析数据	华南农业大学
廖鑫婷		未取得	缺陷关联数据梳理和基于多色集的数据视图构建	华南农业大学
罗勇		未取得	基于交叉熵的关键质量特性识别机制研究	华南农业大学
朱紫纯		未取得	FARM+FL关联分析机制研究	华南农业大学
江城		未取得	FARM+FL关联分析机制研究	华南农业大学

二、项目摘要

中文摘要：

精益管控多样定制化PCB样板生产质量以减少缺陷是行业迫切需求。在此，构建了PCB样板生产缺陷关联分析整体框架，梳理了报废关联相关数据，开展了数据预处理；利用特征识别方法，分析确定了影响报废的主要因素；利用关联分析方法确定相关工序报废主要关联因素。研究成果为PCB样板生产缺陷关联分析数据抽取、预处理、缺陷关键影响因素分析、关联分析等提供了系统实现机制。
引入基于改进谓词感知注意力机制的采样和考虑邻边方向的分数聚合对GENI知识图谱进行改进，提出了GENI-SD模型，采用GENI-SD对PCB样板工序重要性进行了评估对比实验验证了GENI-SD 的可行性和优越性；开展了超参数敏感性实验和消融实验，优化了模型超参数，验证了两个改进对模型性能均有正向作用。构建以PCB工序为核心的知识图谱为PCB相关知识的整合、连接和理解提供了新的实现机制。
基于各类缺陷形成原因、位置、形态和功能影响等因素将PCB常见表面缺陷划分9类。构建一个包括7908张图像和15748个缺陷标签的PCB表面缺陷数据集DsPCBSD+。通过不同深度学习模型验证了所构建缺陷图像集的可行性。DsPCBSD+将为PCB缺陷智能检测领域的研究发展提供了新的数据基准，且有助于对这些缺陷进行追溯分析。

关键词：

数据驱动 ；关键质量特性识别；关联分析；表面缺陷

三、报告正文

报告正文：

1、梳理了PCB报废关联相关数据，开展了数据预处理；利用特征识别方法和关联分析方法，分析确定了影响报废的主要因素。基于相关数据同时开展了时序分析，发表核心论文一篇。为PCB样板生产缺陷关联分析数据抽取、预处理、缺陷关键影响因素分析、关联分析等提供了系统实现机制。提出了GENI-SD知识图谱模型，采用GENI-SD对PCB样板工序重要性进行了评估，对比实验验证了GENI-SD的可行性和优越性；发表核心论文一篇。将PCB常见表面缺陷划分9类，构建一个包括7908张图像和15748个缺陷标签的PCB表面缺陷数据集DsPCBSD+；通过不同深度学习模型验证了所构建缺陷图像集的可行性。基于相关数据集和模型，发表SCI论文1篇、EI检索论文1篇，在审SCI论文和核心论文各1篇。
2、本项目主要工作计划包括研究数据采集集成和面向缺陷关联分析的数据预处理方法、研究样板生产高频缺陷与关键影响因素之间关联规律的挖掘。本研究完成了项目研究计划。
3、PCB样板生产缺陷关键影响因素提取和关联分析方法为PCB样板生产缺陷关联分析数据抽取、预处理、缺陷关键影响因素分析、关联分析等提供了系统实现机制。构建以PCB工序为核心的知识图谱为PCB相关知识的整合、连接和理解提供新的实现机制，对该PCB缺陷领域知识挖掘研究具有重要指导意义和相应的应用前景。所构建的PCB表面缺陷的分类体系和所构建的数据集DsPCBSD+为PCB缺陷智能检测领域的研究和发展提供了新的数据基准，在基于深度学习的AOI检测设备研发中具有非常好的应用前景；对PCB表面缺陷的精益分析也具有直接指导意义。
4、累计毕业硕士研究生5人，其中2人直接进入PCB企业从事PCB生产管理和大数据挖掘分析工作。
5、采用本研究成果，更精益管理定制化PCB车间品质将降低报废，减少原材料投入以及超投/补投等，降低车间综合成本1%以上。
6、国内外相关研究当前主要集中在基于深度学习的PCB表面缺陷的识别，本研究拟发布的PCB缺陷表面数据集将为该领域研究提供新数据基准，后续将研究新的检测算法、开发基于深度学习的AOI设备。

实际参加研究人数	高级职称	中级职称	初级职称	博士后	博士生	硕士生	其他人员
	1	1				6	
计划执行情况	时间方面	按原计划					
	内容方面	内容不变					

四、研究成果目录

序号	成果类型	成果或论文名称	主要完成者	成果说明	标注状况
1	论文	YOLOv4-MN3 for PCB Surface Defect Detection	廖鑫婷，吕盛坪，李灯辉等	PCB缺陷分析SCI论文	标注本基金号2021A1515012395
2	论文	融合浅层特征和注意力机制的PCB缺陷检测方法	廖鑫婷，张洁，吕盛坪	PCB缺陷分析EI论文	标注本基金号22021A1515012395
3	论文	基于GENI-SD的定制化印制电路板工序重要性评估	劳景春，金鸿，吕盛坪等	基于知识图谱精准评估影响定制化印制电路板质量的关键工序	标注本基金号22021A1515012395
4	论文	基于时间加权改进的LDTW 算法	朱紫纯，吕盛坪，廖鑫婷等	时间序列分析	标注本基金号22021A1515012395
5	论文	A hybrid teaching-learning-based optimization algorithm for QoS-awaremanufacturing cloud service composition	金鸿，江城，吕盛坪等	制造云服务组合	标注本基金号22021A1515012395
6	专利	一种分离编带的元器件的收纳装置	吕盛坪，李鑫等	电子产品元器件收纳和自动计数装置	22021A1515012395支助成果
7	项目后续资助	多合影响广义作业车间调度模型构建与优化	吕盛坪，金鸿等	国家自然基金面上项目计划书	支助编号:52275487

五、科技成果统计表

<table>
<tr><td rowspan="6">期刊论文（含已录用）</td><td colspan="7">论文数量及检索系统收录（篇）</td></tr>
<tr><td>总数</td><td>中文期刊</td><td>SCI</td><td>EI</td><td>SSCI</td><td>ISTP</td><td>其他</td></tr>
<tr><td>3.00</td><td>0</td><td>2</td><td>1</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td colspan="7">JCR大类分区（篇）（中科院期刊分区）</td></tr>
<tr><td colspan="2">一区</td><td colspan="2">二区</td><td colspan="2">三区</td><td>其他</td></tr>
<tr><td colspan="2">0</td><td colspan="2">0</td><td colspan="2">1</td><td>0</td></tr>
<tr><td colspan="7">代表性论文（单篇最高影响因子）</td></tr>
<tr><td colspan="5">发表期刊名称</td><td colspan="2">影响因子</td></tr>
<tr><td colspan="5">无</td><td colspan="2">0</td></tr>
<tr><td rowspan="3">著作（不含论文汇编、成果汇编、文化艺术作品等）</td><td colspan="8">数量（本）</td></tr>
<tr><td colspan="2">合计</td><td colspan="2">专著</td><td colspan="2">编著</td><td colspan="2">译著</td></tr>
<tr><td colspan="2">0</td><td colspan="2">0</td><td colspan="2">0</td><td colspan="2">0</td></tr>
<tr><td colspan="2">代表性著作名称</td><td colspan="6">无</td></tr>
<tr><td rowspan="5">专利（件）</td><td colspan="8">申请</td></tr>
<tr><td>合计</td><td colspan="2">发明专利</td><td colspan="2">实用新型专利</td><td colspan="2">外观设计专利</td><td>PCT国际专利</td></tr>
<tr><td>1</td><td colspan="2">1</td><td colspan="2">0</td><td colspan="2">0</td><td>0</td></tr>
<tr><td colspan="8">授权</td></tr>
<tr><td>合计</td><td colspan="2">发明专利</td><td colspan="2">实用新型专利</td><td colspan="2">外观设计专利</td><td>PCT国际专利</td></tr>
<tr><td>1</td><td colspan="2">1</td><td colspan="2">0</td><td colspan="2">0</td><td>0</td></tr>
<tr><td rowspan="2">标准（项）</td><td>总数</td><td colspan="2" rowspan="2">其中：</td><td colspan="2">国标</td><td colspan="2">行标</td><td>地标</td></tr>
<tr><td>0</td><td colspan="2">0</td><td colspan="2">0</td><td>0</td></tr>
<tr><td>软件著作权（项）</td><td colspan="8">0</td></tr>
<tr><td rowspan="2">人才培养</td><td colspan="4">人数</td><td colspan="4">按职称</td><td colspan="3">人才（团队）计划（人次）</td></tr>
<tr><td>合计</td><td>博士后</td><td>博士</td><td>硕士</td><td>合计</td><td>正高</td><td>副高</td><td>中级</td><td>国家级</td><td>省部级</td><td>荣誉名称</td></tr>
<tr><td>4.00</td><td>0</td><td>0</td><td>4</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>无</td></tr>
<tr><td rowspan="3">项目后续资助</td><td colspan="8">国家级项目后续资助（经费单位：万元）</td></tr>
<tr><td colspan="4">国家自然科学基金</td><td colspan="4">科技部项目（基础研究类）</td></tr>
<tr><td colspan="2">项数</td><td colspan="2">经费</td><td colspan="2">项数</td><td colspan="2">经费</td></tr>
<tr><td colspan="2">1</td><td colspan="2">54.00</td><td colspan="2">0</td><td colspan="2">0</td></tr>
</table>

六、经费决算表

经费下达总额：（大写）	壹拾万圆整	（小写）	10
项目编号：	2021A1515012395	项目类型：	广东省自然科学基金-面上项目
项目名称：	定制化印制电路板生产缺陷关键影响特性识别与关联分析方法	项目负责人：	吕盛坪
是否数学等纯理论基础研究项目： 否			
支出科目	经费支出(万元)		备注（说明）
1.设备费	0.00		无
2.业务费	3.43		版面费、专利申请费等
3.直接人力资源成本	3.98		硕士研究生助研经费
4.绩效支出	0.00		0
5.管理费用	0.50		学校管理费
6.其他费用	0.00		0
合计	7.91		0

七、其他财务信息

1.经费使用说明表

项目经费支出科研劳务费3.98万元，论文版面费2.15万元；专利申请1万元、科研差旅、交通、实验材料等0.27万元，学校管理支出0.5万元，累计支出7.91万元。

2.会计师事务所信息

会计师事务所名称	无
签字注册会计师	无
防伪报备编号	无

八、专家意见表

专家1评议表					
1.项目基本信息					
负责人	吕盛坪			项目名称	定制化印制电路板生产缺陷关键影响特性识别与关联分析方法
项目编号	2021A1515012395			项目类别	广东省自然科学基金-面上项目
项目金额	10（万元）			自筹金额	0（万元）
项目承担单位	华南农业大学			项目参与单位	
2.验收专家信息					
姓名	单位			职务职称	专家类别
专家1	******			副教授	技术专家
3.任务书指标完成情况					
成果内容			任务书指标	负责人填写的完成数	专家核实完成数	完成率
论文及专著情况	国家统计源刊物以上刊物发表论文（篇）		3	5	5	166%
	被SCI/EI/ISTP收录论文数（篇）		1	3	3	300%
	专著（册）		0	0	0	100%
	科技报告（篇）		1	1	1	100%
培养人才（人）			2	4	4	200%
引进人才（人）			0	0	0	100%
专利情况(项)	发明专利（件）	申请	1	1	1	100%
		授权	0	1	1	超额完成
	实用新型专利（件）	申请	0	0	0	100%
		授权	0	0	0	100%
	外观设计专利（件）	申请	0	0	0	100%
		授权	0	0	0	100%
	国外专利（件）	申请	0	0	0	100%
		授权	0	0	0	100%
其他	无		完成	完成	完成	
项目评价						
财务意见	经核查，本项目结余资金为2.10万元，违规使用的经费为0万元。					
验收结论	通过					
验收日期	2024-05-11					

专家2评议表

1.项目基本信息

负责人	吕盛坪	项目名称	定制化印制电路板生产缺陷关键影响特性识别与关联分析方法
项目编号	2021A1515012395	项目类别	广东省自然科学基金-面上项目
项目金额	10（万元）	自筹金额	0（万元）
项目承担单位	华南农业大学	项目参与单位	

2.验收专家信息

姓名	单位	职务职称	专家类别
专家2	******	教授	技术专家

3.任务书指标完成情况

成果内容			任务书指标	负责人填写的完成数	专家核实完成数	完成率
论文及专著情况	国家统计源刊物以上刊物发表论文（篇）		3	5	5	166%
	被SCI/EI/ISTP收录论文数（篇）		1	3	3	300%
	专著（册）		0	0	0	100%
	科技报告（篇）		1	1	1	100%
培养人才（人）			2	4	4	200%
引进人才（人）			0	0	0	100%
专利情况(项)	发明专利（件）	申请	1	1	1	100%
		授权	0	1	1	超额完成
	实用新型专利（件）	申请	0	0	0	100%
		授权	0	0	0	100%
	外观设计专利（件）	申请	0	0	0	100%
		授权	0	0	0	100%
	国外专利（件）	申请	0	0	0	100%
		授权	0	0	0	100%
其他	无		完成	完成	完成	

项目评价	
财务意见	经核查，本项目结余资金为2.09万元，违规使用的经费为0万元。
验收结论	通过
验收日期	2024-05-12

专家3评议表

1.项目基本信息

负责人	吕盛坪	项目名称	定制化印制电路板生产缺陷关键影响特性识别与关联分析方法
项目编号	2021A1515012395	项目类别	广东省自然科学基金-面上项目
项目金额	10（万元）	自筹金额	0（万元）
项目承担单位	华南农业大学	项目参与单位	

2.验收专家信息

姓名	单位	职务职称	专家类别
专家3	******	副教授	技术专家

3.任务书指标完成情况

成果内容			任务书指标	负责人填写的完成数	专家核实完成数	完成率
论文及专著情况	国家统计源刊物以上刊物发表论文（篇）		3	5	5	166%
	被SCI/EI/ISTP收录论文数（篇）		1	3	3	300%
	专著（册）		0	0	0	100%
	科技报告（篇）		1	1	1	100%
培养人才（人）			2	4	5	250%
引进人才（人）			0	0	0	100%
专利情况(项)	发明专利（件）	申请	1	1	1	100%
		授权	0	1	1	超额完成
	实用新型专利（件）	申请	0	0	0	100%
		授权	0	0	0	100%
	外观设计专利（件）	申请	0	0	0	100%
		授权	0	0	0	100%
	国外专利（件）	申请	0	0	0	100%
		授权	0	0	0	100%
其他	无		完成	完成	完成	

项目评价	
财务意见	经核查，本项目结余资金为2.09万元，违规使用的经费为0万元。
验收结论	通过
验收日期	2024-05-16

九、验收结论表

验收结论：

该项目专家验收结论为通过。

该验收结论经公示无异议。

广东省基础与应用基础研究基金委员会意见：

定制化印制电路板生产缺陷关键影响特性识别与关联分析方法项目（项目编号：2021A1515012395），经专家评审及公示后，最终的验收结论为通过。

广东省基础与应用基础研究基金委员会（盖章）

科技项目验收专用章

2024 年 07 月 22 日

项目编号：	2014A030310345
资助类别：	广东省自然科学基金-博士启动
文件编号：	粤科规财字[2015]18号

（广东科技微信公众号）

（受理纸质材料二维码）

广东省自然科学基金资助项目结题报告

项目名称：	面向车间调度的工艺规划与静动态集成优化		
项目负责人：	吕盛坪	资助总经费：	10（万元）
计划完成时间：	2015-01-01 至 2018-01-01		
实际完成时间：	2015-01-01 至 2018-01-01		
依托单位：	华南农业大学		
联系人：	石睿	联系电话：	
填表日期：	2018-01-02		

广东省自然科学基金管理委员会
二〇一五年制

填表说明

1.本报告是广东省自然科学基金项目实施完成后的全面回顾与总结，是评价研究工作和今后评审新上课题的依据。课题负责人须认真填写，填写内容必须真实、准确、齐全。

2.本报告要求一式两份，其中一份与《广东省自然科学基金项目管理办法》要求的其他各项材料及合同书，报广东省自然科学基金管理委员会办公室，一份留所在依托单位或项目负责人保存。

3.研究团队核心成员请在“项目人员信息表”中“承担任务”一栏中注明。

4.表内人才培养数中的博士、硕士人数应该是以本研究课题为研究方向的博士、硕士实际毕业人数，并在书面材料附其毕业论文首页的复印件。

一、项目人员信息表

项目负责人：					
姓名	证件号码	职称	承担任务	所在单位	签名
吕盛坪		副教授	整体负责	华南农业大学	
主要研究人员：					
姓名	证件号码	职称	承担任务	所在单位	签名
姜焰鸣		讲师	动态集成决策机制与优化方法研究	华南农业大学	
吕石磊		讲师	集成优化数学模型构建、对应算法设计与实现	华南农业大学	
岑康华		未取得	动态集成优化方法实现	华南农业大学	
曾志雄		未取得	面向调度集成的零件柔性工艺生成研究	华南农业大学	
方思贞		未取得	集成系统开发	华南农业大学	
刘杰坤		未取得	集成系统开发	华南农业大学	
詹志勋		未取得	集成系统开发	华南农业大学	
承担单位（盖章）： 参与单位1（盖章）： 参与单位2（盖章）：					

二、项目摘要

中文摘要：

上提高制造系统效率的重要机制。本项目以现有制造模式与管理方式对工艺与车间调度进行集成规划的需求为应用背景，以两者紧耦合集成内容为前提，从面向调度集成的零件柔性工艺规划、静动态集成优化以及系统实现等角度入手，对工艺与车间调度集成中关键科学问题进行理论研究和实践探索。基于多色集建模和逻辑运算理论，研究面向车间调度集成的柔性工艺形式化生成方法。建立工艺规划与车间调度集成优化数学模型；提出交叉熵和Pareto理论相融合的单/多目标集成优化机制；扩展研究面向不同动态场景对应的决策机制和优化方法。最后，利用服务技术开发支持网络互操作的集成原型系统，并以某机械加工厂部分零件的集成规划为例进行应用验证。本项目的研究成果将为面向车间调度的工艺规划与集成优化提供新思路，具有重要的理论意义和应用价值。

关键词：

柔性工艺规划； 车间调度； 集成优化

三、报告正文

报告正文：

1. 完成的主要研究内容，取得的主要成果，达到的目标、水平及创新之处。
（1）基于多色集建立了各零件特征与车间资源关联关系，从资源的特征加工能力等角度分析零件的制造性。申请了题为“一种应用于柔性工艺过程规划的约束关系描述与可行工艺方案解析生成方法”的发明专利一件。创新之处在于：将柔性工艺过程规划复杂约束统一转换为以特征为核心的约束关系进行描述；避免不可行工艺的产生，减少工艺优选不必要搜寻。
(2) 开展工艺与车间调度集成优化，从车间调度角度协同优化工艺与调度方案
提出了基于交叉熵的集成优化机制，通过相应实例进一步验证了所提出的优化机制的有效性和优越性，为集成研究提供了全新的优化机制。发表SCI检索论文1篇。
（3）开发支持工艺规划与车间调度耦合集成的原型系统
构建了支持集成系统开发的工件特征、资源信息模型及基于网络服务的实现技术框架，开发了支持工艺规划与车间调度集成的原型系统。
所开展的研究在工艺模型构建、工艺与调度集成模型构建、柔性工艺路线推理生成、集成优化算法的设计以及服务化的集成系统建模与开发上都具有较强的创新性。

2. 对照研究工作计划，是否完成预定的研究工作。
本项目整体工作目标包括如下三个：提供面向调度集成的柔性工艺生成的实现机制；优化生产效率、减少目标冲突、增强系统响应能力；建立支持工艺与调度集成规划的网络化平台。完成了项目设定的三个目标。

3. 研究成果的应用前景。
所研究的算法和相应的调度集成系统可以结合具体实施车间具体要求进行完善，具有较好市场前景。

4. 在人才培养方面的绩效，青年科技人员在项目研究中所起的作用。
系统性地培养了参与人员在模型构建、算法设计与实现、系统开发等方面的技能。项目申请人获批题为《面向模具生产的工艺与车间调度紧耦合集成规划》国家自然基金1项。
5. 成果推广及经济效益。
尚未进行成果推广。
6. 在此项目研究期间，国内外同类研究工作取得的新进展，以及对这方面研究工作的进一步设想
主要可分为三类：基于进化算法、帝制竞争、蚁群算法、文化基因算法等的单一策略；基于Agent的机制；结合进化算法和主动学习、禁忌搜索、多Agent、邻域搜索、人工神经网络等的复合策略。
后续将联合企业研究，以进一步将研究成果应用到具体车间。

实际参加研究人数	高级职称	中级职称	初级职称	博士后	博士生	硕士生	其他人员
	1	2		0	0	2	0
计划执行情况	时间方面	按原计划					
	内容方面	内容不变					

项目负责人：（签章）

年　　月　　日

四、研究成果目录

序号	成果类型	成果或论文名称	主要完成者	成果说明	标注状况
1	论文	A cross-entropy-based approach for joint process plan selection and scheduling optimization	Shengping Lv, Wei Liu	SCI检索论文	检索
2	专利	应用于柔性工艺过程规划的约束关系描述与可行工艺方案解析生成方法	吕盛坪，方思贞，杨径，王飞仁，徐岩	专利	公开
3	论文	基于教-学算法的制造云服务组合优化	金鸿，姚锡凡，杨洲，吕盛坪	EI源期刊	待刊
4	论文	基于数据挖掘的印制电路样板投料优化	吕盛坪，乐强生，刘涛	一级期刊源论文	待刊

五、成果统计数据表

<table>
<tr><td rowspan="8">获奖（项）</td><td colspan="9">国家级</td></tr>
<tr><td colspan="3">自然科学奖</td><td colspan="3">科技进步奖</td><td colspan="3">发明奖</td></tr>
<tr><td>一等</td><td>二等</td><td>三等</td><td>一等</td><td>二等</td><td>三等</td><td>一等</td><td>二等</td><td>三等</td></tr>
<tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td colspan="6">省部级</td><td colspan="2" rowspan="2">国际</td><td colspan="3" rowspan="2">其他省部级以上</td></tr>
<tr><td colspan="3">自然科学奖</td><td colspan="3">科技进步奖</td></tr>
<tr><td>一等</td><td>二等</td><td>三等</td><td>一等</td><td>二等</td><td>三等</td><td colspan="2" rowspan="2">0</td><td colspan="3" rowspan="2">0</td></tr>
<tr><td>0</td><td>0</td><td>0</td><td>0</td><td></td><td>0</td></tr>
<tr><td rowspan="6">专著/论文（篇）</td><td colspan="9">发表论文数（含已录用稿件数）</td></tr>
<tr><td colspan="3">国际会议</td><td colspan="3">全国会议</td><td colspan="2">省级会议</td><td colspan="2">其他刊物</td></tr>
<tr><td colspan="3">0</td><td colspan="3">0</td><td colspan="2">0</td><td colspan="2">1</td></tr>
<tr><td colspan="9">四大检索系统收录</td></tr>
<tr><td colspan="3">SCI</td><td colspan="3">EI</td><td colspan="2">ISTP</td><td colspan="2">ISR</td></tr>
<tr><td colspan="3">1</td><td colspan="3">1</td><td colspan="2">0</td><td colspan="2">0</td></tr>
<tr><td rowspan="3">专利、人才及学术交流</td><td colspan="9">专著</td></tr>
<tr><td colspan="5">中文</td><td colspan="4">外文</td></tr>
<tr><td colspan="5">0</td><td colspan="4">0</td></tr>
<tr><td rowspan="8"></td><td colspan="9">专利（项）</td></tr>
<tr><td colspan="3">合计</td><td colspan="6">发明专利</td></tr>
<tr><td colspan="2">申请</td><td colspan="2">授权</td><td colspan="2">申请</td><td colspan="3">授权</td></tr>
<tr><td colspan="2">1</td><td colspan="2">0</td><td colspan="2">0</td><td colspan="3">0</td></tr>
<tr><td colspan="9">人才培养（人）</td></tr>
<tr><td>博士后</td><td colspan="2">博士</td><td colspan="2">硕士</td><td colspan="2">新增正高</td><td>新增副高</td></tr>
<tr><td>0</td><td colspan="2">0</td><td colspan="2">0</td><td colspan="2">0</td><td>1</td></tr>
<tr><td colspan="9">学术会议</td></tr>
<tr><td colspan="3">国内</td><td colspan="6">国外</td></tr>
<tr><td colspan="2">主办</td><td colspan="2">参加</td><td colspan="2">主办</td><td colspan="3">参加</td></tr>
<tr><td colspan="2">0</td><td colspan="2">0</td><td colspan="2">0</td><td colspan="3">0</td></tr>
</table>

<table>
<tr><td rowspan="7">获后续资助及效益</td><td colspan="3">国家级后续资助（万元）</td></tr>
<tr><td>自然科学基金</td><td>其他基础研究类资助</td><td>其他非基础研究类资助</td></tr>
<tr><td>20.00</td><td>0</td><td>0</td></tr>
<tr><td colspan="3">省部级后续资助（万元）</td></tr>
<tr><td>自然科学基金</td><td>其他基础研究类资助</td><td>其他非基础研究类资助</td></tr>
<tr><td>0</td><td>0</td><td>0</td></tr>
<tr><td colspan="3">成果推广产值（万元）</td></tr>
<tr><td colspan="3">0</td></tr>
</table>

六、经费决算表

甲方经费下达总额：（大写）	壹拾万圆整	（小写）	10
项目编号：	2014A030310345	项目类型：	广东省自然科学基金-博士启动
项目名称：	面向车间调度的工艺规划与静动态集成优化	项目负责人：	吕盛坪
支出科目	预算经费（万元）	经费支出（万元）	备注（说明）
科研业务费	3.5	1.27	论文版面费和软件著作权申请费等2.2万元有待支出
实验材料费	0	0	未有预算
仪器设备费	1.5	1.03	大规模集成算例移动处理终端;0.47万有待支付给对方
实验室改装费	0.5	0	实验室购置桌椅、插座等；0.3万有待报销
协作费	0.5	0	软件开发测试费，待支付给对方
人员费	3	2.40	学生的劳务费，剩余部分后续为10-12月学生劳务费
专家咨询费	0	0	无专家咨询费
国际合作交流费	0.5	0	国外访学一年相应交流费由留学基金委支出
管理费	0.5	0.50	学校管理费用
合计		5.20	总支出
甲方拨付经费结余（万元）			
与本项目相关的其他经费来源	预算经费（万元）		经费支出（万元）
其他计划资助经费	0		0
本单位配套经费	0		0
其他经费资助	0		0
其他经费来源合计	0.00		0.00
项目负责人：吕盛坪 （签章） 年　月　日	财务部门负责人： （公章） 年　月　日		科研部门负责人： （公章） 年　月　日

七、附：经费使用说明表

1、科研业务费预算经费预算3.5万元，主要用于参加国内学术会议费、国内调研或工厂实践差旅费、油费。项目总支出1.27，会议和交通运输费用10000、论文版面费、申请专利和图书购置等2700；2.2万元有待支出。
2、仪器设备费预算1.5万元，主要用于大规模集成算例移动处理终端和数据存储设备购置，实际支出1.03万元，另外0.47万元移动设备费已经购置，经费有待支付给对方。
3、实验室改装预算0.5万元，实际未支出，0.3万元待报账。
4、协作费预算0.5万元，待支付给对方。
5、人员费3.0 万元，实际支出参与项目的硕士生劳务费2.4万元。
6、国际合作与交流费预算0.5万元，实际未支出。
7、管理费预算0.5万元，学校按照5%扣除支出0.5万元。
本项目总预算10万元，实际支出5.2万元，未使用完的项目经费中科研业务费预算相应论文版面费和软件著作权有待支出、仪器设备费购置任务已经完成，相应款项有待转账给对方；协作费预算对方已经开具发票，正待转账给对方账户；剩余部分后续为10-12月学生劳务费；其它各项经费完全按照实际预算进行开支。

项目负责人：吕盛坪 （签章） 年　　月　　日	财务部门负责人： （公章） 年　　月　　日	科研部门负责人： （公章） 年　　月　　日

八、专家意见表

项目编号：	2014A030310345	项目类型：	广东省自然科学基金-博士启动
项目名称：	面向车间调度的工艺规划与静动态集成优化		
项目负责人：	吕盛坪	承担单位：	华南农业大学

专家信息表

序号	姓名	职称	所在单位	签名
1	姚锡凡	教授	华南理工大学	
2	邹湘军	教授	华南农业大学	
3	王永华	高级工程师	广东工业大学	

专家组意见

受广东省教育厅委托，专家组对华南农业大学吕盛坪承担的广东省自然基金项目“面向车间调度的工艺规划与静动态集成优化”（编号2014A030310345）进行结题评论，专家组评阅了结题材料，经邮件和电话咨询和充分讨论后，形成如下验收意见：

1. 项目组提供的验收资料齐备，符合验收基本条件。

2. 本项目基于多色集建立了各零件特征与车间资源关联关系，从资源的特征加工能力等角度分析了零件的制造性。基于多色集合建立了柔性工艺过程规划的约束关系描述与可行工艺方案解析生成方法。基于交叉熵，对工艺与车间调度进行了集成优化，对比结果显示优化结果具有明显优势。在此基础上，基于服务化封装技术开发支持工艺规划与车间调度耦合集成的原型系统。

3. 项目执行期间，课题组人员申请发明专利1件，发表SCI检索论文1篇，待刊相关论文2篇。项目组人才配备合理，科研款项专款专用，使用规范。

验收专家一致认为，项目完成了合同书中的考核指标，经费使用合理，同意通过结题验收。

是否同意结题： 是	专家组组长签名： 年 月 日

九、本申请项目所附附件清单

会计师事务所名称：	
签字注册会计师：	
防伪报备编号：	

序号	附件类型	数量
1	项目下达文件	1
2	项目合同书	1
3	经费决算表	1
4	经费使用情况说明表	1
5	项目结题审计报告	0
6	人员信息表	1
7	项目完成报告	1
8	专家意见表	1
9	项目所获成果、专利一览表（含成果登记号、专利申请号、专利号等）	4
10	其他有关材料	0

十、项目负责人签字及审核意见表

项目负责人承诺：

我所承担的项目（编号：2014A030310345　名称：面向车间调度的工艺规划与静动态集成优化）结题报告内容填写实事求是，数据详实。在今后的研究工作中，如有与本项目相关的成果，将标注“广东省自然科学基金资助”，并报送广东省自然科学基金委员会。

负责人（签章）：

2018 年 1 月 20 日

项目依托单位审查意见：

同意

经办人（签章）：

单位公章：

2018年 1月 24日

广东省自然科学基金委员会管理办公室（省科技厅管理部门）意见：

负责人（签章）：

单位公章：

年　　月　　日

合同编号：

产学研技术服务合同

项目名称：基于大数据的投料优化

委 托 方：深圳明阳电路科技股份有限公司（甲 方）

受 托 方：华南农业大学（乙 方）

受 托 方：深圳市澳昇科技有限公司（丙 方）

签订时间：2019 年 12 月 26 日

签订地点：深圳明阳电路科技股份有限公司

有效期限：2019/12/01—2020/05/26

填 写 说 明

一、本合同为中华人民共和国科学技术部印制的技术服务合同示范文本，各技术合同认定登记机构可推介技术合同当事人参照使用。

二、本合同书适用于一方当事人（受托方）以技术知识为另一方（委托方）解决特定技术问题所订立的合同。

三、签约一方为多个当事人的，可按各自在合同关系中的作用等，在“委托方”、“受托方”项下（增页）分别排列为共同委托人或共同受托人。

四、本合同书未尽事项，可由当事人附页另行约定，并作为本合同的组成部分。

五、当事人使用本合同书时约定无需填写的条款，应在该条款处注明“无”等字样。

产学研技术服务合同

委托方（甲方）：深圳明阳电路科技股份有限公司

住　所　地：深圳市宝安区新桥街道上星第二工业区

法定代表人：张佩珂

项目联系人：秦小虎

联系方式：[redacted]

通讯地址：深圳市宝安区新桥街道上星第二工业区

电　　话：[redacted] 传真：[redacted]

电子信箱：[redacted]

受托方（乙方）：华南农业大学

住　所　地：广东省广州市天河区五山路 483 号

法定代表人：刘雅红

项目联系人：吕盛坪

联系方式：[redacted]

通讯地址：广东省广州市天河区五山路 483 号工程学院北 210B

电　　话：020-85280752　传真：

电子信箱：lvshengping@scau.edu.cn

受托方（丙方）：深圳市澳昇科技有限公司

住　所　地：深圳市龙华区龙华街道三联社区锦华发工业园 3 栋硅谷大院 T1 栋 B306

法定代表人：解孟军

项目联系人：刘威

联系方式：[redacted]

通讯地址：深圳市龙华区龙华街道三联社区锦华发工业园 3 栋硅谷大院 T1 栋 B306

电　　话：[redacted]　　传真：[redacted]

电子信箱：danny.lau@awesometech.cn

本合同甲方委托乙方和丙方就“基于大数据的投料优化项目”（以下简称本项目）进行的专项产学研合作与技术服务，并支付相应的技术服务报酬，同时开拓未来智能制造与大数据领域的产学研合作前景。三方经过平等协商，在真实、充分地表达各自意愿的基础上，根据《中华人民共和国合同法》的规定，达成如下协议，并由三方共同恪守。

第一条：甲方委托乙方和丙方进行技术服务的内容如下：

1. 技术服务的目标：依照由三方签字盖章后的《基于大数据的投料优化--工作说明书》（以下简称工作说明书或 SOW）中第 1.1 节“项目目标”所限定目标执行。

2. 技术服务的内容及范围：依照由三方签字盖章后的 SOW 中第 1.2 节“项目的范围”所限定内容及其范围执行。

第二条：乙方应按下列要求完成技术服务工作：

1. 技术服务地点：深圳市宝安区新桥街道上星第二工业区

2. 技术服务期限：4 个月　其中 3 个月为项目研究开发和初步验证期，后面 1 个月为甲方验证期限。

3. 技术服务进度：依照由三方签字盖章后的 SOW 中第 1.3 节“项目计划”执行。

第三条：为保证乙方有效进行技术服务工作，甲方应当向乙方提供下列工作条件和协作事项：

1、依照由三方签字盖章后的 SOW 中第 4 节“三方的责任和义务”所规定的甲方的责任要求执行。

第四条：甲方向乙方丙方支付技术服务报酬及支付方式为：

1. 技术服务费总额为：30 万人民币（乙方开具 3%的增值税专用发票，丙方开具 6%的增值税票，下同）。

2. 技术服务费由甲方分期支付乙方和丙方。 具体支付方式和时间如下：

（1）项目合同签订乙方和丙方开票后，10 个工作日由甲方给乙方和丙方分别支付技术服务费总额的 25 %作为预付款。即分别向乙方和丙方支付人民币 7.5 万元，大写人民 柒万伍仟 圆整。

（2）项目验收完成，乙方开票后 10 个工作日内，甲方给乙方和丙方分别支付技术服务费总额的 25 %，即人民币 7.5 万元，大写人民币 柒万伍仟 圆整。

乙方开户名称、开户银行名称、地址和帐号为：

开户名称：华南农业大学

开户银行：中国工商银行广州五山支行

地址：广州市天河区五山路 483 号

帐号：[redacted]

丙方开户名称、开户银行名称、地址和帐号为：

开户名称：深圳市澳昇科技有限公司

开户银行：广发银行股份有限公司深圳侨香支行

地址：深圳市福田区侨香路裕和大厦一层 106-111 号

帐号：[redacted]

第五条：三方确定因履行本合同应遵守的保密义务如下：

按照三方签字盖章的保密协议执行；

第六条：三方确定以下列标准和方式对乙方的技术服务工作成果进行验收：

1．技术服务工作成果的验收方法：甲方项目组领导成员组织验收，按照里程碑结点分期验收并推进行项目。

2．验收的时间和地点：项目完成后 10 个工作日内，项目实施地点验收。项目验收标准：按照工作说明书的验收标准部分确定。

第七条：三方确定：

1．在本合同有效期内，与乙方、丙方工作及服务内容相关的知识产权（包括但不限于软件著作权、数据、专利等）归甲方所有。

第八条：三方确定，按以下约定承担各自的违约责任：

1．如甲方无故拒绝验收，则甲方应当支付当期应付的合同款。如乙方和丙方所交付产品不符合本 SOW 的需求约定，导致甲方无法实现合同目的，乙方和丙方应在收到甲方通知后 15 日内将甲方所支付的所有款项全额退回，但由甲方原因导致的除外。

2. 实施服务过程中，三方均按合同约定进行工作，不得随意变更合同服务内容。如甲方要求修改或变更已按合同约定的交付服务，须书面提出要求并经乙方和丙方同意后方可进行修改，除交货期相应顺延外，由此引起的费用由甲方承担。甲乙丙三方无论以何种原因违反本合同约定，违约方承担的违约责任不超过本合同总金额的 100%。

3. 在上述违约期的计算中，应扣除第十条中不可抗力因素所造成的延迟。

第九条：三方确定，在本合同有效期内，甲方指定 秦小虎 为甲方项目联系人，乙方指定 吕盛坪 为乙方项目联系人，丙方指定 刘威 为丙方项目联系人。项目联系人承担以下责任：

1. 组织三方参与人员开展项目交流、协商、实施和验收；

2. 一方变更项目联系人，应当及时以书面形式通知另一方，未及时通知并影响本合同履行或造成损失的，应承担相应的责任。

第十条：三方确定，出现发生不可抗力致使本合同的履行成为不必要或不可能的，可以解除本合同：不可抗力必须是指一方不可控制的并不可预见的事件，包括但不限于：

1. 自然灾害、地震、洪水、雷击、火灾、磁电串入等；

2. 战争或准战争状态、恐怖活动、戒严、骚乱、罢工、行业纠纷等。

3. 由于上述不可抗力因素致使乙方和丙方无法按合同规定的时间提供软件，则此类延误将被视为不可抗力，乙方和丙方不承担违约责任，但必须设法及时通知甲方。

4. 在不可抗力事件结束后十五日内，受不可抗力影响一方应以挂号

或传真的方式将有关部门出具的证明送达至对方，否则对方可不予承认，并要求承担相应的违约责任。

5. 如不可抗力事故的影响连续 120 天以上时，三方应通过友好协商解决本合同履行问题，并尽快达成协议。

第十一条：三方因履行本合同而发生的争议，应协商、调解解决。协商、调解不成的，确定按以下第 1 种方式处理：

1. 提交 深圳市宝安区人民法院 仲裁委员会仲裁；

2. 依法向人民法院起诉。

第十二条：基于大数据的投料优化的工作说明书为本合同不可分割的一部分，具有同等法律效力。

第十三条：本合同一式 6 份，甲乙丙三方各一式两份。

第十三条： 本合同经三方签字盖章后生效。

甲方：_____（盖章）

法定代表人／委托代理人：_____（签名）

2019年 12月 26日

乙方：_____（盖章）

法定代表人／委托代理人：_____（签名）

年　　月　　日

丙方：_____（盖章）

法定代表人／委托代理人：_____（签名）

2019年 12月 26日

HXIGHT20219#3

合同编号：SJ-0-04-22103

技术服务合同

项目名称： 印制电路板表面缺陷图像处理及模型构建

委 托 方（甲 方）： 工业和信息化部电子第五研究所

受 托 方（乙 方）： 华南农业大学

签订时间： 2022 年 8 月 10 日

签订地点： 工业和信息化部电子第五研究所

有效期限： 2022 年 8 月 10 日——2022 年 12 月 31 日

中华人民共和国科学技术部印制

填 写 说 明

一、本合同为中华人民共和国科学技术部印制的技术服务合同示范文本，各技术合同认定登记机构可推介技术合同当事人参照使用。

二、本合同书适用于一方当事人（受托方）以技术知识为另一方（委托方）解决特定技术问题所订立的合同。

三、签约一方为多个当事人的，可按各自在合同关系中的作用等，在“委托方”、“受托方”项下（增页）分别排列为共同委托人或共同受托人。

四、本合同书未尽事项，可由当事人附页另行约定，并作为本合同的组成部分。

五、当事人使用本合同书时约定无需填写的条款，应在该条款处注明“无”等字样。

技术服务合同

委托方（甲方）：工业和信息化部电子第五研究所

住　所　地：广东省广州市增城区朱村街朱村大道西 78 号

法定代表人：陈立辉

项目联系人：蒋诗新

联系方式 ：[illegible]

通讯地址：广东省广州市增城区朱村街朱村大道西 78 号

电　　话：020-87237237　　传真：

电子信箱：jiangshixin@ceprei.com

受托方（乙方）：华南农业大学

住　所　地：广东省广州市天河区五山路 483 号

法定代表人：刘雅红

项目联系人：吕盛坪

联系方式 ：[illegible]

通讯地址：广东省广州市天河区五山路 483 号工程学院北 210B

电　　话：020-85280752　　传真：

电子信箱：lvshengping@scau.edu.cn

本合同甲方委托乙方就 印制电路板表面缺陷图像处理及模型构建 项目进行的专项技术服务，并支付相应的技术服务报酬。双方经过平等协商，在真实、充分地表达各自意愿的基础上，根据《中华人民共和国合同法》的规定，达成如下协议，并由双方共同恪守。

第一条：甲方委托乙方进行技术服务的内容如下：

1．技术服务的目标：针对印制电路板（PCB）表面缺陷检测需求，结合甲方提供的印制电路板表面缺陷图像样本，开展缺陷图像的标注、样本增强等处理，并基于深度学习算法构建 PCB 表面缺陷检测模型，实现 PCB 常见表面缺陷的在线检测。

2．技术服务的内容：1）建立 PCB 常见表面缺陷分类及判据体系；2）开展 PCB 表面缺陷样本图像分类、标注与数据增强工作；3）依托缺陷样本图片，基于深度学习算法构建 PCB 表面缺陷检测模型。

3．技术服务的方式：开展 PCB 表面缺陷调研，建立印制电路板缺陷分类及判据体系。按照甲方及其合作企业所提供表面缺陷数据集和行业标准要求收集、整理、清洗相关图像数据集。标注各类表面缺陷，利用数据增强技术扩增数据集，构建 PCB 表面缺陷样本库。基于缺陷图像样本训练深度学习模型，构建 PCB 缺陷检测模型，解决 PCB 表面缺陷识别（分类+定位）问题，支撑 PCB 表面缺陷检测软件的开发与应用 。

第二条：乙方应按下列要求完成技术服务工作：

1．技术服务地点：广东省广州市增城区朱村街朱村大道西 78 号。

2．技术服务期限：5 个月（研究开发和验证）。

3．技术服务进度：

（1）乙方应在本合同生效后 10 日内向甲方提交研究开发计划。研究开发计划应包括以下主要内容：项目实施途径、项目预期成果与周期、项目验收方式。

（2）乙方应按下列进度完成项目的设计及开发工作：（可能根据实际情况调整）

工作内容	完成时间	工作成果
印制电路板缺陷分类及模型构建技术方案	合同签订后 30 日内	技术方案文档 1 份
印制电路板缺陷检测模型构建	合同签订后 100 日内	模型源代码、模型封装与部署文档

以上各阶段完成时间到期前 7 个工作日内乙方应提供本阶段所产生的技术成果。甲方仅对乙方提交的最终成果验收，在乙方按期提交最终成果后 10 个工作日内进行验收，并书面确认。

第三条：为保证乙方有效进行技术服务工作，甲方应当向乙方提供下列工作条件和协作事项：

1．提供技术资料：提供项目实施必要的技术文档，包括但不限于缺陷图像数据、模型技术要求等。

2．提供工作条件：合同签订后 5 个工作日内。

3．其他：对乙方在研发过程中提出的问题，甲方应及时给予反馈。

4．本合同履行完毕后，上述技术资料按以下方式处理：乙方返还甲方光盘或纸质文档 。

第四条：甲方向乙方支付技术服务报酬及支付方式为：

1.技术服务费总额为:（大写）拾陆万 元整，计人民币（小写）¥160000元。

2．技术服务费由甲方 分期 支付乙方。

具体支付方式和时间如下：

1）合同签订后15个工作日内甲方支付乙方50%的合同金额；

2）项目通过验收后15个工作日内甲方支付乙方剩余50%合同金额。

合同履约过程中所有价格变动的风险均由乙方承担，甲方不需再向乙方支付任何费用。在双方协商一致的前提下，因项目的内容、计划、验收标准或双方职责等发生变更而导致合同总价款的变更的，须另行签订补充协议。

在中国境内、外发生的与本合同执行有关的一切税费均由乙方负担。甲方按本条第 2 项约定，在每次支付费用前，乙方均应开具相应金额的增值税专用 发票。如因乙方延迟提供发票的，甲方的支付时间顺延，不因此承担逾期支付的违约责任。

乙方开户银行名称、地址和帐号为：

开户银行：广州工行五山支行

地址： 广州市天河区五山路483号

帐号：

【注意】：如乙方采用邮寄方式提供发票的，按如下信息邮寄，如因填写信息有误或非按如下内容填写，导致发票丢失，损失由乙方承担，并视乙方未提供发票：

收件人： 蒋诗新

联系电话：

收件地址：广东省广州市增城区朱村街朱村大道西78号

第五条：双方确定因履行本合同应遵守的保密义务如下：

1.保密内容（包括技术信息和经营信息）：涉及本合同的技术文件、资料等秘密。

2．涉密人员范围: 直接或间接涉及本合同项目的有关人员。

3．保密期限：本合同双方的保密义务不因本合同的解除或终止而免除。

4．泄密责任：除法律法规或政府主管部门、司法部门要求披露的以外，乙方不得公开或向第三方提供或披露保密内容，若违反本协议规定，需承担全部责任且赔偿因此给甲方造成的损失，该损失计量不超过本合同的总金额。

第六条：本合同的变更必须由双方协商一致，并以书面形式确定。

第七条：双方确定以下列标准和方式对乙方的技术服务工作成果进行验收：

1．技术服务工作成果的验收标准：按本技术协议进行逐项验收。

2．技术服务工作成果的验收方法：印制电路板缺陷分类及模型构建技术方案 1 份；印制电路板缺陷检测模型，包括模型源代码一套、模型封装与部署文档 1 份。

3．验收的时间和地点：交付时间：2022 年 12 月 31 日前；交付地点：广东省广州市增城区朱村街朱村大道西 78 号。

第八条 乙方保证所提供的服务无侵害任何第三方的知识产权或其他合法权益。如有第三方指控乙方提供给甲方的服务侵犯了该方的知识产权或其他权利，乙方自行承担给甲方及第三方造成的一切损失（此损失包括但不限于甲方已经支付给第三方的或经法院生效判决、仲裁裁决确定的款

项及甲方因此而产生的诉讼费、鉴定费、公证费、调查费、律师费等）。

第九条 双方确定，本合同执行期内所产生的研究开发成果的知识产权归甲方所有。只有经甲方书面同意，方可将该研究开发成果向第三方转让或实施，对于实施、转让产生的收益由甲方享有。

甲方可对履行本合同所产生的研究开发成果进行后续改进，乙方不得将本合同产生交付的研究开发成果及由此后续改进产生的成果自行实施、转让给第三人。如有违反，需承担全部责任并且赔偿因此给甲方造成的损失，该损失计量不超过本合同的总金额。

第十条 乙方应根据本项目成果的性质、特征，以及相关法律法规和甲方要求，对所提供的服务进行全过程的质量检查和检验，必要时接受甲方的监督和评审，以确保所提供的成果，符合甲方及相关法律法规的质量保证要求和保障要求。

第十一条 双方确定，项目质保期为 6 个月 ，自项目通过甲方验收之日起计算，质保期内，乙方有义务无偿为本项目的成果提供咨询解释工作，并根据甲方的请求，为甲方指定的人员提供技术指导和培训，或提供与使用该研究开发成果相关的技术服务。质保期到期后，根据甲方的请求，乙方可为甲方指定的人员提供技术指导和培训，或提供与使用该研究开发成果相关的技术服务，相关内容、费用双方另行协商，并签订服务合同，但该服务费用不得高于其收取的最低标准且不得高于当前市场普通水平。

第十二条：双方确定，在本合同有效期内，甲方指定 蒋诗新（ ）为甲方项目联系人，乙方指定 吕盛坪（ ）为乙方项目联系人。项目联系人承担以下责任：

1．组织三方参与人员开展项目交流、协商、实施和验收；

2．一方变更项目联系人，应当及时以书面形式通知另一方，未及时通知并影响本合同履行或造成损失的，应承担相应的责任。

第十三条：双方确定，出现下列情形，致使本合同的履行成为不必要

或不可能的，可以解除本合同：

1．发生不可抗力；

2．合同依据国家、政府计划被取消

第十四条：双方因履行本合同而发生的争议，应协商、调解解决。协商、调解不成的，确定按以下第 2 种方式处理：

1．提交 广州 仲裁委员会仲裁；

2．依法向人民法院起诉。

第十五条：双方确定：本合同及相关附件中所涉及的有关名词和技术术语，其定义和解释如下：

无

第十六条：双方约定本合同其他相关事项为：

1. 本合同一方向另一方发出的全部通知和要求以及双方的文件往来等，必须用书面形式传递。

2.一方在本合同履行过程中向另一方发出或者提供的所有通知、文件、文书、资料等，均以本合同所列明的地址或邮箱送达；一方如果迁址或者变更电话、邮箱的，应当及时书面通知另一方，未履行通知义务的，另一方按原地址邮寄相关材料即视为已履行送达义务；当面交付上述材料的，在交付之时视为送达。只要相关材料送达至上述地址或邮箱的，无论受送达方是否实际拆阅或查阅的，均视为送达完成。

3. 所有乙方应当承担的违约金、赔偿金等责任，甲方均有权在应支付给乙方的款项中直接扣除，不足以扣除的，甲方有权继续要求乙方支付或赔偿。

第十七条：本合同一式 肆 份，具有同等法律效力。

第十八条：本合同经双方签字盖章后生效。

第十九条 本合同未尽事宜，可另行签订补充协议，如补充协议

甲方：工业和信息化部电子第五研究所（盖章）

法定代表人／委托代理人：（签名）

2022年 8月 10日

乙方：华南农业大学（盖章）

法定代表人／委托代理人：（签名）

年 月 日

合同验收书

<table>
<tr><td colspan="3">需方:工业和信息化部电子第五研究所</td><td colspan="3">供方:华南农业大学</td></tr>
<tr><td>合同名称</td><td colspan="5">印制电路板表面缺陷图像处理及模型构建</td></tr>
<tr><td></td><td colspan="5">交付情况</td></tr>
<tr><td rowspan="6">交付</td><td>序号</td><td>项目</td><td>规格说明</td><td>数量</td><td>是否符合</td></tr>
<tr><td>1</td><td>印制电路板缺陷分类及模型构建技术方案</td><td>电子文件</td><td>1</td><td>符合</td></tr>
<tr><td>2</td><td>印制电路板缺陷检测模型源代码</td><td>电子文件</td><td>1</td><td>符合</td></tr>
<tr><td>3</td><td>印制电路板缺陷检测模型封装与部署文档</td><td>电子文件</td><td>1</td><td>符合</td></tr>
<tr><td>备注</td><td colspan="4"></td></tr>
<tr><td colspan="5">接收人（签字）：钟书棋
交付地点：广州
交付人（签字）：
交付日期：2022.12.5</td></tr>
<tr><td>验收</td><td colspan="5">双方确认印制电路板表面缺陷图像处理及模型构建项目所交付的成果物符合合同技术要求，经测试验证，模型运行正常，一致同意通过合同验收
需方代表（签字）：钟书棋
验收地点：广州
供方代表（签字）：
验收日期：2022.12.5</td></tr>
</table>

横向科技合同受理号 HXKJHT20251757

合同编号：

技术服务合同

项目名称： 组织系统模型构建

委 托 方（甲 方）： 数孪模型科技（北京）有限责任公司

受 托 方（乙 方）： 华南农业大学

签订时间： 2025 年 8 月 10 日

签订地点： 数孪模型科技（北京）有限责任公司

有效期限： 2025 年 7 月 1 日——2025 年 12 月 31 日

中华人民共和国科学技术部印制

填 写 说 明

一、本合同为中华人民共和国科学技术部印制的技术服务合同示范文本，各技术合同认定登记机构可推介技术合同当事人参照使用。

二、本合同书适用于一方当事人（受托方）以技术知识为另一方（委托方）解决特定技术问题所订立的合同。

三、签约一方为多个当事人的，可按各自在合同关系中的作用等，在“委托方”、“受托方”项下（增页）分别排列为共同委托人或共同受托人。

四、本合同书未尽事项，可由当事人附页另行约定，并作为本合同的组成部分。

五、当事人使用本合同书时约定无需填写的条款，应在该条款处注明“无”等字样。

技术服务合同

委托方（甲方）：数孪模型科技（北京）有限责任公司

住　所　地：北京市朝阳区来广营乡中国铁建广场 B 座 908 号

法定代表人：程燕

项目联系人：程燕

联系方式：[illegible]

通讯地址：北京市朝阳区来广营乡中国铁建广场 B 座 908 号

电　　话：[illegible]　传真：

电子信箱：chengyan@emagecloud.com

受托方（乙方）：华南农业大学

住　所　地：广东省广州市天河区五山路 483 号

法定代表人：薛红卫

项目联系人：吕盛坪

联系方式：[illegible]

通讯地址：广东省广州市天河区五山路 483 号工程学院北 210B

电　　话：020-85280752　传真：

电子信箱：lvshengping@scau.edu.cn

本合同甲方委托乙方就　组织系统模型构建　项目进行的专项技术服务，并支付相应的技术服务报酬。双方经过平等协商，在真实、充分地表达各自意愿的基础上，根据《中华人民共和国民法典》的规定，达成如下协议，并由双方共同恪守。

第一条：甲方委托乙方进行技术服务的内容如下：

1．技术服务的目标：围绕复杂组织系统的建模与优化决策这一核心问题，构建"概念-逻辑-物理"三层递进式标准化建模体系，支持各类组织系统快速标准化建模、迭代优化以及动态决策分析。建立"抽象层次-建模领域-价值链覆盖"三维建模框架，支持纵向维度上的跨层级的智能映射和横向维度上跨领域的动态匹配，确保组织系统治理活动的全局一致性、灵活性和可扩展性。提出算法与规则双驱动的智能决策支持方法，为企业各层级决策者提供科学的决策支持理论方法。指导企业开发建模与决策支持系统，支持组织系统建模与优化决策。

2．技术服务的内容：1）研究组织系统分层分域建模框架，设计跨本体领域的要素一致性映射和关联规则，建立跨层域要素联动方法，形成统一元模型及其规范。2）基于要素特征与关联规则切分组织系统类型并定义模型类型，从而形成丰富的组织系统建模模板库（24 个以上），指导企业建立时序约束满足的模型演化引导范式，并构建基于逻辑规则的推理验证方法及多级冲突消解策略。3）选择业务领域具体场景，构建智能优化决策数学模型，设计相应算法，指导企业开发实现。

3．技术服务的方式：深入甲方企业进行两周的现场集中办公，平时以视频会议的形式进行沟通。

第二条：乙方应按下列要求完成技术服务工作：

1．技术服务地点：北京市朝阳区来广营乡中国铁建广场 B 座 908 号

2．技术服务期限：6 个月。

3．技术服务进度：

（1）乙方应在本合同生效后 10 日内向甲方提交研究开发计划。研究开

发计划应包括以下主要内容：项目实施途径、项目预期成果与周期、项目验收方式。

（2）乙方应按下列进度完成项目的设计及开发工作：

工作内容	完成时间	工作成果
统一元模型及其规范	合同签订后 10 日内	组织系统统一元模型总结文档 1 份
组织系统建模模板库	合同签订后 20 日内	模板库总结报告 1 份
选定具体业务场景，设计相应智能优化算法	合同签订后 30 日内	优化模型和算法流程设计文档 1 份

以上各阶段完成时间到期前 7 个工作日内乙方应提供本阶段所产生的技术成果。甲方仅对乙方提交的最终成果验收，在乙方按期提交最终成果后 10 个工作日内进行验收，并书面确认。

第三条：为保证乙方有效进行技术服务工作，甲方应当向乙方提供下列工作条件和协作事项：

1．提供技术资料：提供项目实施必要的技术文档

2．提供工作条件：合同签订后 5 个工作日内

3．其他：对乙方在研发过程中提出的问题，甲方应及时给予反馈。

4．本合同履行完毕后，上述技术资料按以下方式处理：乙方不得泄露甲方所提供给乙方所有资料。

第四条：甲方向乙方支付技术服务报酬及支付方式为：

1．技术服务费总额为：（大写）伍万 元整，计人民币（小写）¥50000 元。

2．技术服务费由甲方 分期 支付乙方。

具体支付方式和时间如下：

1）合同签订后 15 个工作日内甲方支付乙方 70%的合同金额；

2）项目通过验收后 15 个工作日内甲方支付乙方剩余 30%费用。

合同履约过程中所有价格变动的风险均由乙方承担，甲方不需再向乙方支付任何费用。在双方协商一致的前提下，因项目的内容、计划、验收标准或双方职责等发生变更而导致合同总价款的变更的，须另行签订补充协议。

在中国境内、外发生的与本合同执行有关的一切税费均由乙方负担。甲方按本条第 2 项约定，在每次支付费用前，乙方均应开具相应金额的 增值税专用发票。如因乙方延迟提供发票的，甲方的支付时间顺延，不因此承担逾期支付的违约责任。

乙方开户银行名称、地址和账号为：

开户银行：广州工行五山支行

地址： 广州市天河区五山路 483 号

账号：

【注意】：如乙方采用邮寄方式提供发票的，按如下信息邮寄，如因填写信息有误或非按如下内容填写，导致发票丢失，损失由乙方承担，并视乙方未提供发票：

收件人： 程燕

联系电话：

收件地址：北京市朝阳区来广营乡中国铁建广场 B 座 908 号

第五条：双方确定因履行本合同应遵守的保密义务如下：

1.保密内容（包括技术信息和经营信息）：涉及本合同的技术文件、资料、经营信息等商业秘密。

2．涉密人员范围：直接或间接涉及本合同项目的有关人员。

3．保密期限：本合同双方的保密义务不因本合同的解除或终止而免除。

4．泄密责任：除法律法规或政府主管部门、司法部门要求披露的以外，任

何一方违反本协议规定的，需承担全部责任且赔偿因此给甲方造成的全部损失，如该损失难以计量的，以甲方委托乙方开展本合同项下工作应支付全部款项的十倍计。

第六条：本合同的变更必须由双方协商一致，并以书面形式确定。

第七条：双方确定以下列标准和方式对乙方的技术服务工作成果进行验收：

1．技术服务工作成果的验收标准：按本技术协议进行逐项验收。

2．技术服务工作成果的验收方法：组织系统统一元模型总结文档 1 份、模板库总结报告 1 份；优化模型和算法设计报告 1 份。

3．验收的时间和地点：交付时间：2025 年 12 月 31 日前；北京市朝阳区来广营乡中国铁建广场 B 座 908 号。

第八条 乙方保证所提供的服务无侵害任何第三方的知识产权或其他合法权益。如有第三方指控乙方提供给甲方的服务侵犯了该方的知识产权或其他权利，乙方自行承担给甲方及第三方造成的一切损失（此损失包括但不限于甲方已经支付给第三方的或经法院生效判决、仲裁裁决确定的款项及甲方因此而产生的诉讼费、鉴定费、公证费、调查费、律师费等）。

第九条 双方确定，所产生的研究开发成果的知识产权归甲方所有。只有经甲方书面同意，方可将该研究开发成果向第三方转让或实施，对于实施、转让产生的收益由甲方享有。

甲方可对履行本合同所产生的研究开发成果进行后续改进，乙方不得将本合同产生交付的研究开发成果及由此后续改进产生的成果自行实施、转让给第三人。如有违反，需承担全部责任并且赔偿因此给甲方造成的损失，如该损失难以计量的，以甲方委托乙方开展本合同项下工作应支付全部款项的十倍计。

第十条 乙方应根据本项目成果的性质、特征，以及相关法律法规和甲方要求，对所提供的服务进行全过程的质量检查和检验，必要时接受甲方的监督和评审，以确保所提供的成果，符合甲方及相关法律法规的质量保证要求和保

障要求。

第十一条 双方确定，项目质保期为 1 年，自项目通过甲方验收之日起计算，质保期内，乙方有义务无偿为本项目的成果提供咨询解释工作，并根据甲方的请求，为甲方指定的人员提供技术指导和培训，或提供与使用该研究开发成果相关的技术服务。质保期到期后，根据甲方的请求，乙方可为甲方指定的人员提供技术指导和培训，或提供与使用该研究开发成果相关的技术服务，相关内容、费用双方另行协商，并签订服务合同，但该服务费用不得高于其收取的最低标准且不得高于当前市场普通水平。

第十二条：双方确定，在本合同有效期内，甲方指定程燕（[illegible]）为甲方项目联系人，乙方指定吕盛坪（18[illegible]）为乙方项目联系人。项目联系人承担以下责任：

1．组织参与人员开展项目交流、协商、实施和验收；

2．一方变更项目联系人，应当及时以书面形式通知另一方，未及时通知并影响本合同履行或造成损失的，应承担相应的责任。

第十三条：双方确定，出现下列情形，致使本合同的履行成为不必要或不可能的，可以解除本合同：

1．发生不可抗力；

2．合同依据国家、政府计划被取消。

第十四条：双方因履行本合同而发生的争议，应协商、调解解决。协商、调解不成的，确定按以下第 2 种方式处理：

1．提交 广州 仲裁委员会仲裁；

2．依法向人民法院起诉。

第十五条：双方确定：本合同及相关附件中所涉及的有关名词和技术术语，其定义和解释如下：无。

第十六条：双方约定本合同其他相关事项为：

1.本合同一方向另一方发出的全部通知和要求以及双方的文件往来等，必须用书面形式传递。

2.一方在本合同履行过程中向另一方发出或者提供的所有通知、文件、文书、资料等，均以本合同所列明的地址或邮箱送达；一方如果迁址或者变更电话、邮箱的，应当及时书面通知另一方，未履行通知义务的，另一方按原地址邮寄相关材料即视为已履行送达义务；当面交付上述材料的，在交付之时视为送达。只要相关材料送达至上述地址或邮箱的，无论受送达方是否实际拆阅或查阅的，均视为送达完成。

3. 所有乙方应当承担的违约金、赔偿金等责任，甲方均有权在应支付给乙方的款项中直接扣除，不足以扣除的，甲方有权继续要求乙方支付或赔偿。

第十七条：本合同一式 肆 份，具有同等法律效力。

第十八条：本合同经双方签字盖章后生效。

第十九条：本合同未尽事宜，可另行签订补充协议，如补充协议与本合同约定不一致的，以最新签署的有效的补充协议为准。

第二十条：本合同经双方签字盖章后生效。

甲方：数孪模型科技（北京）有限责任公司（盖章）

法定代表人 / 委托代理人：（签名）

2025 年 8 月 10 日

乙方：华南农业大学（盖章）

法定代表人 / 委托代理人：（签名）

年 月 日

二、科研项目---参与项目清单

编号：CARS-170405

国家糖料产业技术体系岗位任务书

(2021-2025 年)

岗位科学家姓名：张智刚

岗位名称：智能化管控

依托单位：华南农业大学

主管部门：广东省教育厅

农业农村部科技教育司

二〇二一年九月

填 写 说 明

1. 本任务书由农业农村部科技教育司、各体系首席科学家、产业技术研发中心依托单位联合签订。
2. 本任务书要求按照已给的格式，5 号宋体字填写，单倍行间距，段落间无间距，A 4 纸双面打印。
3. 本任务书封面不签字盖章，仅在签约各方页签字盖章。
4. 本任务书可视填报内容自行增加页码。
5. 本任务书由科技教育司统一编号，一式 5 份，农业农村部科教司 1 份，首席科学家 1 份，功能研究室 1 份、岗位科学家依托单位 1 份、岗位科学家 1 份。

一、岗位科学家基本情况表

<table>
<tr><td colspan="8">（一）总体概况</td></tr>
<tr><td>功能研究室</td><td colspan="2">机械化研究室</td><td colspan="2">岗位名称</td><td colspan="2">智能化管控</td></tr>
<tr><td>科学家姓名</td><td>张智刚</td><td>年 龄</td><td></td><td>学 历</td><td>博士</td><td>职 称</td><td>副教授</td></tr>
<tr><td colspan="2">岗位科学家的依托单位</td><td colspan="6">华南农业大学</td></tr>
<tr><td colspan="2">岗位专家的团队（人）</td><td>9</td><td colspan="2">合计经费（万元）</td><td colspan="3">275</td></tr>
<tr><td colspan="8">（二）团队组成情况</td></tr>
<tr><td colspan="8">1、姓名：吕盛坪，所在单位：华南农业大学工程学院</td></tr>
<tr><td>个人简介</td><td colspan="7">吕盛坪，男， 年 月，博士，副教授，研究方向：工/农业大数据分析，深度学习、知识图谱在工/农业中的应用。2005 年 6 月，毕业于湖南农业大学工程学院机电工程系，获学士学位；2008 年 3 月，毕业于北京交通大学机械与电子控制工程学院机械工程系，获工学硕士学位；2012 月 6 月，毕业于北京航空航天大学机械工程及自动化学院工业与制造系统工程系，获工学博士学位。现为华南农业大学工程学院车辆工程系副教授，华南农业大学青年骨干教师,华南农业大学工程学院优秀青年教师，美国农业生物工程师学会会员、中国机械工程学会工业大数据与智能系统委员（首届）。主要从事工业/农业大数据分析、智能制造、机器视觉在工业领域中应用等相关研究。主持国家自然基金项目 1 项、广东省自然科学基金项目 2 项、广东省教育厅育苗项目 1 项，企业横向项目多项。累计发表论文 40 余篇，其中 SCI 收录 10 篇，EI 收录 10 篇；参与编写教材 1 部；获软件著作权 8 件，其中第一申请人获批 7 件；申请公开专利 6 件，其中第一申请人授权 3 件。主持校级教学成果二等级 1 项。</td></tr>
<tr><td colspan="8">2、姓名：肖克辉，所在单位：华南农业大学数学与信息学院</td></tr>
<tr><td>个人简介</td><td colspan="7">肖克辉，男， 年 月出生，河南省商城县人，工学博士，博士后，高级实验师，副教授，硕士生导师，广东省农村科技特派员，主要从事农业信息化、农业物联网及农业大数据等相关技术领域的理论和应用研究。2002 年 6 月毕业于空军第一航空学院，获计算机及应用专业工学学士学位；2005 年 6 月毕业于中山大学，获计算机软件与理论专业工学硕士学位；2012 年 12 月毕业华南农业大学，获农业电气化与自动化专业工学博士学位；2014 年 7 月至 2015 年 7 月在美国华盛顿州立大学进行博士后研究工作，研究方向为精准农业及其自动化技术。现为广东省养殖物联网工程技术研究中心、广东省农业大数据工程技术研究中心骨干成员。近年来先后主持国家星火计划项目、广东省自然科学基金项目、广东省科技计划项目等省部级项目 5 项，参与省部级项目 10 余项，发表学术论文 40 余篇，申报发明专利 1 项，取得实用新型专利 1 项，取得软件著作权 5 项，获广东省农业技术推广奖二等奖 1 项。</td></tr>
<tr><td colspan="8">3、姓名：徐梅宣，所在单位：华南农业大学电子工程学院</td></tr>
<tr><td>个人简介</td><td colspan="7">徐梅宣，女， 年 月出生，工学博士，讲师。主要研究方向：机器视觉、多光谱探测、农情信息采集技术。2001 年 7 月毕业于重庆大学光电工程学院测控技术及仪器专业，获工学学士学位；2004 年 7 月毕业于重庆大学光学工程专业，获工学硕士学位；2007 年 7 月毕业于华南农业大学农业电气化及自动化专业，获工学博士学位。主持“十二五”农村领域国家科技计划（支撑计划）课题“水稻氮素诊断技术与在线监测设备”和广东省自然基金项目“基于立体视觉和光谱技术的水稻氮素营养检测方法研究”各 1 项。申请发明专利 1 项，发表论文 10 余篇。长期承担电子信息工程、通信工程、电子科学技术等本科专业的《信号与系统》、《数字信号处理》、《单片机接口及</td></tr>
</table>

	应用》、《DSP 技术及应用》等专业核心课的教学工作。作为第一主编编写《汽车生产中的 IT 技术》(ISBN：9787111479208)、作为第二主编编写教材《数字信号处理》(ISBN：978-7-5682-7195-0)。积极指导学生参与“大学生创新创业计划“等专业相关项目；积极指导学生参与”全国电子设计大赛“及”广东省电子设计大赛“等重要专业赛事，获得“广东省电子设计大赛一等奖”1 项。
4、姓名：廖娟，所在单位：华南农业大学工程学院	
个人简介	廖娟，女，　　年　月出生，工学博士，讲师。主要研究方向：作物航空植保作业质量评价与参数优选、作物植保喷施雾滴沉积与漂移特性研究、基于 LiDAR 的作物冠层监测方法研究、基于 LiDAR 作物冠层监测的智能施药系统研究。2009 年 6 月毕业于湖南科技学院，获得电子信息工程专业工学学士学位；2013 年 6 月毕业于华南农业大学工程学院，获得农业电气化与自动化专业工学硕士学位；2017 年 06 月毕业于华南农业大学工程学院，获得农业电气化与自动化专业博士学位，期间以联合培养方式留学澳大利亚昆士兰大学。2017 年 7 月至 2020 年 7 月，在华南农业大学工程学院从事博士后研究工作。主持国家自然科学基金（青年科学基金）“基于水稻冠层结构参数和光谱特性信息融合的褐飞虱危害程度判定新方法研究”1 项，主要参加广东省科学技术厅应用研究项目“农用无人直升机性能检测系统研发”、“基于农用无人直升机的新疆棉花脱叶剂喷施作业参数优化与示范”、“多旋翼农用无人机综合性能检测平台的开发”等，发表 SCI 论文 4 篇，EI 论文 3 篇，撰写的农业工程学部期刊论文“提高农业机械化水平促进农业可持续发展”荣获中国农业工程学会 40 周年优秀论文。
5、姓名：张闻宇，所在单位：华南农业大学工程学院	
个人简介	张闻宇，男，　　年　月生，博士，讲师。2014 年 9 月至 2018 年 6 月，攻读并获得华中农业大学与爱荷华州立大学联合培养农业工程专业博士学位，现为华南农业大学工程学院教师。主要从事农机自动导航、农机智能控制、人工智能农业应用和无人农场等关键技术研究。主持广东省区域联合基金-青年基金项目“基于深度强化学习的南方复杂边界环境田块农机导航控制方法”1 项，参加省部级以上科研项目 5 项，发表 SCI/EI 学术论文 12 篇，申请发明专利和软件著作权登记 9 项。
6、姓名：王辉，所在单位：潍柴雷沃重工股份有限公司	
个人简介	王辉，男，　　年　月，博士，工程师。潍柴雷沃智慧农业研究院副院长兼智能驾驶开发部部长，潍坊市鸢都产业领军人才。2015 年 6 月至 2019 年 12 月，攻读并获得华南农业大学农业电气化与自动化博士学位，2019 年 12 月至 2020 年 12 月，任雷沃重工股份有限公司液压电控技术中心工程师，2020 年 12 月至 2021 年 7 月，任潍柴雷沃重工股份有限公司精准农业模块主任工程师，2020 年 7 月至今，任潍柴雷沃重工股份有限公司智慧农业研究院副院长兼智能驾驶开发部部长。长期从事农机自动驾驶作业及精准农业方面研究，在传统农机装备智能化升级、农机自动驾驶和精准作业方面作了较多工作。参与省级以上项目 6 项，发表学术论文 3 篇，获授权专利 6 件，其中发明专利 4 件。获 2019 中国机械工业科学技术奖一等奖、2021 年工信部第四届 5G 应用“绽放杯”智慧交通赛道二等奖。
7、姓名：何留伟，所在单位：广东广垦农机服务有限公司	
个人简介	男，　　年出生，农艺师，广东广垦农机服务有限公司董事长、党总支书记。

8、姓名：苏俊波，所在单位：中国热带农业科学院南亚热带作物研究所	
个人简介	男，　　年　月生，中国热带农业科学院南亚热带作物研究所旱作种业与节水技术研究中心主任，博士，副研究员，所聘研究员。2006 年参加工作，在甘蔗育种方面，育成热甘 1 号、热甘 11713、热甘 11559、热甘 11713、热甘 1462 等甘蔗新品种 6 个，获植物品种权证书 5 件，登记甘蔗新品种 1 件。在甘蔗机械化工作方面，开展适宜机械化栽培和收获的甘蔗品种筛选和机械化配套栽培技术的研究与应用推广，总结出了一套适宜粤西蔗区推广的甘蔗机械化栽培技术。出版专著 2 部，发布行业标准 1 项，第一作者或通讯作者发表研究论文十余篇。2018 年获广东省甘蔗机械化工作先进个人称号，2021 年因科技特派员的出色工作获广东省科技厅通报表扬。
9、姓名：廖锡华，所在单位：广东广垦糖业集团有限公司	
个人简介	男，　　年出生，农艺师。农业工程系农机教研室主任。广东广垦糖业集团有限公司副部长。

二、体系重点任务与考核指标

（一）产业重大关键问题技术攻关

CARS-17-01A：基于机械化的甘蔗新品种选育及节本增效综合栽培模式集成与示范

1、任务名称：甘蔗生产机械智能作业关键技术研究与应用

2、研发背景：我国传统甘蔗生产机械劳动强度大、作业效率低、肥药利用率低、机收含杂率和损失率高，这导致甘蔗生产机械化水平低、蔗糖产业缺乏市场竞争力。本任务利用电子技术、计算机技术、自动化技术和机电液一体化技术改造传统甘蔗生产机械，提升其种植、田间管理和收获环节的作业效率和质量，同时降低农机驾驶员劳动强度，达到节本增效目的。

3、核心技术与实施内容：

（1）五年总体核心技术和实施内容：

拟研发的甘蔗生产机械智能作业关键技术包括：①面向规模化蔗田种植的智能作业路径规划技术、②基于北斗的甘蔗生产机械自动导航控制技术、③蔗田变量施肥控制技术、④蔗田变量施药控制技术等。通过对传统甘蔗生产机械进行技术改造，创制适配甘蔗生产机械的北斗导航控制装置和变量作业控制装置等技术成果。

（2）年度分解任务

2021 年实施内容：

以节本增效为目标，开展甘蔗生产机械智能作业需求调研，进一步明确甘蔗生产机械智能化作业的现实需求，制定甘蔗生产机械智能作业关键技术的研究方案。

2022 年实施内容：

开展蔗田智能作业路径规划技术研究，实现播种、田间管理和收获的作业路径自动规划。

2023 年实施内容：

开展甘蔗种植和收获机械北斗导航控制技术研究，研制适配甘蔗种植和收获机械的北斗导航控制装置，开展系统测试。

2024 年实施内容：

开展蔗田变量施肥技术研究，研制适配蔗田施肥机的变量作业控制装置，开展系统测试。

2025 年实施内容：

开展蔗田变量施药技术研究，研制适配蔗田施药机的变量作业控制装置，开展系统测试。

4、考核指标：

（1）五年总体考核指标：

突破蔗田作业路径智能规划、甘蔗生产机械北斗导航控制等关键技术 2 项；创制北斗导航控制装置 2 套、变量施肥控制装置 1 套、变量施药控制装置 1 套。

（2）年度分解考核指标：

2021 年考核指标：

撰写甘蔗生产机械智能作业需求调研报告 1 份。

2022 年考核指标：

突破甘蔗种植智能作业路径规划关键技术 1 项。

2023 年考核指标：

创制适配甘蔗种植和收获机械的北斗导航控制装置各 1 套。

2024 年考核指标：

创制适配蔗田施肥机的变量作业控制装置 1 套。

2025 年考核指标：

创制适配蔗田施药机的变量作业控制装置 1 套。

5、运行管理机制：

（1）任务分工：张智刚负责本岗位任务实施的组织与协调，并与各综合试验站对接。吕盛坪、肖克辉、徐梅宣、廖娟、张闻宇、王辉等负责制定智能作业关键技术研发方案，何留伟、苏俊波、廖锡华跟踪农机智能化作业实施情况。

（2）组织交流：参加由片区负责定期组织的有关岗位科学家和片区内综合试验站进行项目技术交流和工作经验交流；组织协调机械化研究室岗位专家紧密合作、资源共享，一起研讨技术路线；对技术难题协同攻关。

6、牵头和参加的机构和人员情况

技术总负责人：张智刚

岗位科学家	**岗位名称**	**任务分工**	
张智刚	智能化管控	组织实施，系统设计，体系内交流，总结	
参加的团队人员	**所在单位**	**任务分工**	
吕盛坪	华南农业大学	甘蔗种植智能作业路径规划技术	
肖克辉	华南农业大学	农机自动导航作业软件程序开发	
徐梅宣	华南农业大学	变量施肥（药）技术	
廖娟	华南农业大学	制定实施方案、跟踪机械化作业实施情况	
张闻宇	华南农业大学	联合收获机和转运车主从协调控制技术	
王辉	潍柴雷沃重工股份有限公司	甘蔗生产机械智能作业系统创制	
何留伟	广东广垦农机服务有限公司	负责农机智能化作业试验示范	
苏俊波	中国热带农业科学院南亚热带作物研究所	负责农机智能化作业试验示范	
廖锡华	广东广垦糖业集团有限公司	负责农机智能化作业试验示范	
参加的综合试验站	**参加站长**	**参加的团队人员**	**任务分工**
湛江综合试验站	刘建荣	揭　进	湛江农垦蔗区项目示范
崇左综合试验站	覃　勇	农永前	崇左蔗区项目示范
百色综合试验站	贺贵柏	黄文武	百色蔗区项目示范
来宾综合试验站	兰军群	黄家训	来宾蔗区项目示范
柳城综合试验站	卢文祥	卢李威	柳州蔗区项目示范
北海综合试验站	杨忠伟	陈家翔	北海蔗区项目示范
桂林综合试验站	李家文	钟　坤	桂林蔗区项目示范
金光综合试验站	李廷化	韦金凡	广西农垦蔗区项目示范
保山综合试验站	段兆祜	石红军	保山蔗区项目示范
临沧综合试验站	周　中	董有波	临沧蔗区项目示范

CARS-17-02A：基于甜菜全程机械化的新品种引（育）及节本增效综合栽培模式集成与示范

1、任务名称：甜菜收获机械智能对行收获关键技术研究与应用

2、研发背景：近年来，我国甜菜全程机械化生产方式在甜菜主产区已大面积推广应用。甜菜收获环节多以分段收获为主，农机驾驶员劳动强度大，收获作业质量受人为因素影响较大。为提高甜菜收获机械的作业效率和作业质量、降低驾驶员劳动强度，本任务提出研究甜菜收获机械智能对行收获关键技术和装备，为我国甜菜生产的智能化收获提供支撑。

3、核心技术与实施内容：

（1）五年总体核心技术和实施内容：

拟研究的关键技术包括：①甜菜生产机械作业路径智能规划技术、②基于北斗和 MEMS 惯导的甜菜收获机械位姿检测技术、③甜菜收获机械自动对行收获控制技术，研制甜菜收获机械自动对行收获控制装置，开展系统测试和应用示范。

（2）年度分解任务

2021 年实施内容：

开展甜菜收获机械智能化作业需求调研，制定甜菜收获机械对行收获技术路线。

2022 年实施内容：

开展甜菜作业机械智能路径规划技术研究，实现直播、田间管理和收获的作业路径自动规划。

2023 年实施内容：

研发甜菜收获机械位姿检测技术、甜菜收获机械对行收获技术，开发甜菜收获机械的对行收获控制装置。

2024 年实施内容：

系统调试、改进和完善，在规模化农场开展系统测试与应用示范。

2025 年实施内容：

在规模化农场开展应用示范。

4、考核指标：

（1）五年总体考核指标：

突破甜菜生产机械智能化作业路径规划、甜菜收获机械自动对行收获等关键技术 2 项，研制甜菜收获机械自动对行收获控制装置 2 套，开展系统测试和应用示范。

（2）年度分解考核指标：

2021 年考核指标：

撰写甜菜收获机械智能作业需求调研报告 1 份。

2022 年考核指标：

突破甜菜生产机械智能化作业路径规划技术。

2023 年考核指标：

突破甜菜收获机械位姿检测技术、甜菜收获机械对行收获技术，开发甜菜收获机械的对行收获控制装置 2 套。

2024 年考核指标：

开展系统测试和应用示范。

2025 年考核指标：

开展系统测试和应用示范。

5、运行管理机制：

（1）任务分工：张智刚负责本岗位任务实施的组织与协调，并与各综合试验站对接。吕盛坪、肖克辉、徐梅宣、廖娟、张闻宇、王辉等负责制定甜菜收获机械自动对行收获技术研发方案，跟踪系统运行实施情况。

（2）组织交流：参加甜菜片区内综合试验站组织的项目技术交流会和工作经验交流会；组织协调机械化研究室岗位专家紧密合作、资源共享，一起研讨技术路线；对技术难题协同攻关。

6、牵头和参加的机构和人员情况

技术总负责人：张智刚

岗位科学家	**岗位名称**	**任务分工**	
张智刚	智能化管控	组织实施，系统设计，体系内交流，总结	
参加的团队人员	**所在单位**	**任务分工**	
吕盛坪	华南农业大学	对行控制算法研究	
肖克辉	华南农业大学	对行控制软件程序开发	
徐梅宣	华南农业大学	农机作业位姿检测	
廖娟	华南农业大学	制定实施方案、跟踪技术研发实施情况	
张闻宇	华南农业大学	路径规划技术研究	
王辉	潍柴雷沃重工股份有限公司	自动对行控制装置开发	
参加的综合试验站	**参加站长**	**参加的团队人员**	**任务分工**
呼和浩特综合试验站	樊福义	李智	内蒙古甜菜产区项目示范
赤峰综合试验站	史树德	魏磊	内蒙古甜菜产区项目示范

（二）服务县域经济发展

CARS-17-03 A：云南省临沧市耿马县

1、产业分析：

耿马傣族佤族自治县 90%以上的土地分布在热带和亚热带，年均气温 19.2℃，光照充足，雨量充沛，雨热同季，无霜期长，非常适宜糖料甘蔗的生长。

耿马自治县是国家 51 个糖料蔗核心基地县（市）之一，也是国家糖料生产保护区的重要组成部分。目前蔗糖产业已成为耿马自治县关联度最大、涉及面最广、影响最深的重要支柱产业，成为全县经济发展的一张“绿色产业”王牌。该县围绕“一根甘蔗吃干榨尽”延伸产业链，形成了糖、酒、纸、饲、肥 5 大类 15 种产品的产业链格局，县内有涉糖企业 14 户，其中，制糖企业 4 户，日处理甘蔗达 2.55 万吨，每个榨季可处理甘蔗 300 万吨以上；有蔗渣制浆企业 1 户，年产蔗渣浆板 10 万吨；有糖蜜酒精生产线 1 条，年产酒精 5 万吨以上；在建的黄腐酸钾项目建成后，耿马将成为全国首家打通整个制糖产业链最后一环的地区。县内涉糖龙头企业主要有广西洋

浦南华集团下属 4 户制糖企业、1 户浆纸企业，均为规模企业，总产值达 26.8 亿元。在蔗梢利用环节，有省级农业龙头企业 1 户，总产值达 0.5 亿元。造纸和酒精生产的原料来源覆盖 3 个州市、7 个县（区）、13 个糖厂。

近年来，依托龙头企业的带动，着力开发了从甘蔗制糖到造纸、食用酒精、生物工程、蔗梢利用的循环经济产业链，带动就业 7000 人以上，带动农资、农机服务机构 128 个，并有效带动了交通运输、商贸物流等关联行业发展。

2、任务内容：

（1）五年总体任务内容：

耿马县甘蔗生产机械化与智能化调研与现场指导或培训。

（2）年度分解任务

2021 年任务内容：

2021 年度耿马县甘蔗生产机械化与智能化调研、现场指导或培训。

2022 年任务内容：

2022 年度耿马县甘蔗生产机械化与智能化调研、现场指导或培训。

2023 年任务内容：

2023 年度耿马县甘蔗生产机械化与智能化调研、现场指导或培训。

2024 年任务内容：

2024 年度耿马县甘蔗生产机械化与智能化调研、现场指导或培训。

2025 年任务内容：

2025 年度耿马县甘蔗生产机械化与智能化调研、现场指导或培训。

3、工作机制及任务分工：

张智刚为负责人，吕盛坪、肖克辉、徐梅宣、廖娟、张闻宇负责调研、现场指导或培训。

4、考核指标：

（1）五年总的考核指标：

耿马县“十四五”甘蔗生产机械化与智能化发展报告。

（2）年度分解考核指标：

2021 年考核指标：

2021 年度耿马县甘蔗生产机械化与智能化发展报告。

2022 年考核指标：

2022 年度耿马县甘蔗生产机械化与智能化发展报告。

2023 年考核指标：

2023 年度耿马县甘蔗生产机械化与智能化发展报告。

2024 年考核指标：

2024 年度耿马县甘蔗生产机械化与智能化发展报告。

2025 年考核指标：

2025 年度耿马县甘蔗生产机械化与智能化发展报告。

CARS-17-04 A：广西壮族自治区崇左市

1、产业分析：

崇左市地处北回归线以南，属亚热带季风气候区，气候温和，雨量充沛。年日照时数 1600 多时，1 月平均气温 13.8℃，7 月平均气温 28.1℃，年平均气温 20.8℃-22.4℃，年无霜期长达 340 多天，年降雨量 1200 毫米以上。全年光照充足，非常适宜糖料甘蔗的生长。

崇左主要产糖区是国家农业部划定的甘蔗“双高”优势产区之一，所辖 5 个县被国家发改委定为甘蔗生产基地县，糖料蔗面积常年维持在 400 万亩以上，甘蔗播种面积占崇左现有总耕地面积 610 万亩的 65.74%，从事与糖产业相关人口达 130 多万人，占总人口的 52.3%。2014 年以来崇左市积极开展糖料蔗基地建设，崇左市抢抓自治区实施 500 万“双高”基地建设的历史机遇，举全市之力推进全市 201 万亩“双高”基地建设。2014 年—2020 年，累计落实完成面积 206.51 万亩。为蔗糖产业可持续发展提供有力保障。

按照“十三五”规划目标要求，崇左市积极推进糖产品精深加工，推动企业高附加值新产品开发，打造制糖循环产业链，形成了制糖、造纸、酒精、酵母、酵母抽提物、味精、生物肥、生物质发电等多条蔗糖循环经济产业链。综合利用糖蜜生产酵母产能 8 万吨/年、酵母抽取物 2 万吨/年、酒精 5 万吨/年、味精 5 万吨/年，崇左市成为全国最大的酵母生产基地；综合利用蔗渣造纸产能 19 万吨/年、发电产能 90 兆瓦；综合利用滤泥生产生物肥产能 10 万吨/年。

2、任务内容：

（1）五年总体任务内容：

崇左市甘蔗生产机械化与智能化调研与现场指导或培训。

（2）年度分解任务

2021 年任务内容：

2021 年度崇左市甘蔗生产机械化与智能化调研、现场指导或培训。

2022 年任务内容：

2022 年度崇左市甘蔗生产机械化与智能化调研、现场指导或培训。

2023 年任务内容：

2023 年度崇左市甘蔗生产机械化与智能化调研、现场指导或培训。

2024 年任务内容：

2024 年度崇左市甘蔗生产机械化与智能化调研、现场指导或培训。

2025 年任务内容：

2025 年度崇左市甘蔗生产机械化与智能化调研、现场指导或培训。

3、工作机制及任务分工：

张智刚为负责人，吕盛坪、肖克辉、徐梅宣、廖娟、张闻宇负责调研、现场指导或培训。

4、考核指标：

（1）五年总的考核指标：

崇左市“十四五”甘蔗生产机械化与智能化发展报告。

（2）年度分解考核指标：

2021 年考核指标：

2021 年度崇左市甘蔗生产机械化与智能化发展报告。

2022 年考核指标：

2022 年度崇左市甘蔗生产机械化与智能化发展报告。

2023 年考核指标：

2023 年度崇左市甘蔗生产机械化与智能化发展报告。

2024 年考核指标：

2024 年度崇左市甘蔗生产机械化与智能化发展报告。

2025 年考核指标：

2025 年度崇左市甘蔗生产机械化与智能化发展报告。

（三）重大突发性事件应急和咨询服务

1.监测本产业生产和市场的异常变化，及时向农业农村部上报情况。
2.组织开展应急性技术指导和培训工作。
3.完成体系及研发中心交办的各项工作。
4.发生重大自然灾害或重大突发性事件，及时制订分区域的应急预案与技术指导方案，建立专家组，明确工作机制，并以体系的名义上报农业农村部科技教育司。

三、机械化研究室重点任务与考核指标

（一）CARS-17-07B：以降杂减损高效机收为目标的甘蔗全程机械化关键技术研究

1、任务名称：以降杂减损高效机收为目标的甘蔗生产机械智能化关键技术研究

2、研发背景：“十三五”期间（2015/2016～2019/2020），在中央及地方各类政策扶持下，甘蔗生产机械化总体呈现提升态势。但甘蔗机收率仅从 0.75%提高到 3.28%，5 个榨季只共提高了 2.53 个百分点。与国发 42 号文件提出的 2025 年“甘蔗收获机械化率达到 30%”的要求差距还很大。造成这种局面的原因有两个：一是机收蔗含杂（特别是泥土）较多，我国现有制糖工艺难以处理，造成糖厂不愿接收机收蔗入榨；二是机收造成的田间损失较大，造成蔗农采用机收的积极性不高。为此，开展以降杂减损高效机收为目标的甘蔗全程机械化关键技术研究，对推进我国甘蔗收获机械化的发展，解决甘蔗机械化的瓶颈问题，具有重要的意义。

3、核心技术与实施内容：

（1）五年总体核心技术和实施内容：

开展甘蔗生产机械智能化关键技术研究，具体包括：①中耕培土施肥机精准作业技术；②高地隙喷雾机自主作业技术；③甘蔗联合收获机与田间转运车主从协调控制技术等。针对甘蔗中耕培土施肥作业机械，创制甘蔗智能施肥作业系统；针对高地隙喷雾机，创制甘蔗自主施药作业系统；针对甘蔗收获机械和运输车，创制甘蔗智能收获作业系统。

（2）年度分解任务

2021 年实施内容：

开展甘蔗生产机械智能化作业需求调研，提出相关技术解决方案。

2022 年实施内容：

开展甘蔗联合收获机与田间转运车主从协调控制技术研究，创制甘蔗智能收获作业系统，实现双机主从导航和自动对位协同作业。

2023 年实施内容：

开展中耕培土施肥机精准作业技术研究，创制甘蔗智能施肥作业系统，实现自动导航控制和变量施肥作业的技术集成。

2024 年实施内容：

开展高地隙喷雾机自主作业技术研究，创制甘蔗自主施药作业系统，实现高地隙喷药机无人驾驶施药作业。

2025 年实施内容：

三大技术成果应用示范。

4、考核指标：

（1）五年总体考核指标：

突破高地隙喷药机自主作业技术、甘蔗联合收获机与田间转运车主从协调控制技术 2 项，创制甘蔗智能施肥作业系统 1 套、高地隙喷雾机自主施药作业系统 1 套、甘蔗智能收获作业系统 1 套。

（2）年度分解考核指标：

2021 年考核指标：

撰写甘蔗生产机械智能化作业需求调研报告。

2022 年考核指标：

突破甘蔗联合收获机与田间转运车主从协调控制技术 1 项，创制甘蔗智能收获作业系统 1 套。

2023 年考核指标：

创制甘蔗智能施肥作业系统 1 套。

2024 年考核指标：

突破高地隙喷药机自主作业技术 1 项，创制高地隙喷雾机自主施药作业系统 1 套。

2025 年考核指标：

创制甘蔗智能收获作业系统 1 套。

5、运行管理机制：

（1）任务分工：张智刚负责本岗位任务实施的组织与协调，并与各综合试验站对接。吕盛坪、肖克辉、徐梅宣、廖娟、张闻宇、王辉等负责制定甘蔗生产机械智能化关键技术研发方案，何留伟、苏俊波、廖锡华跟踪系统运行实施情况。

（2）组织交流：参加甘蔗片区内综合试验站组织的项目技术交流会和工作经验交流会；组织协调机械化研究室岗位专家紧密合作、资源共享，一起研讨技术路线；对技术难题协同攻关。

6、牵头和参加的机构、人员情况

岗位科学家	岗位名称	任务分工
张智刚	智能化管控	组织实施，系统设计，体系内交流，总结
参加的团队人员	**所在单位**	**任务分工**
吕盛坪	华南农业大学	蔗田作业路径规划技术
肖克辉	华南农业大学	蔗田作业机械自动导航控制技术

徐梅宣	华南农业大学	中耕培土施肥机精准作业技术	
廖娟	华南农业大学	制定实施方案、跟踪系统运行实施情况	
张闻宇	华南农业大学	甘蔗联合收获机与田间转运车主从协调控制技术	
王辉	潍柴雷沃重工股份有限公司	高地隙喷雾机自主施药作业系统	
何留伟	广东广垦农机服务有限公司	负责农机智能化作业试验示范	
苏俊波	中国热带农业科学院南亚热带作物研究所	负责农机智能化作业试验示范	
廖锡华	广东广垦糖业集团有限公司	负责农机智能化作业试验示范	
参加的综合试验站	**参加站长**	**参加的团队人员**	**任务分工**
湛江综合试验站	刘建荣	揭　进	湛江农垦蔗区项目示范
崇左综合试验站	覃　勇	农永前	崇左蔗区项目示范
百色综合试验站	贺贵柏	黄文武	百色蔗区项目示范
来宾综合试验站	兰军群	黄家训	来宾蔗区项目示范
柳城综合试验站	卢文祥	卢李威	柳州蔗区项目示范
北海综合试验站	杨忠伟	陈家翔	北海蔗区项目示范
桂林综合试验站	李家文	钟　坤	桂林蔗区项目示范
金光综合试验站	李廷化	韦金凡	广西农垦蔗区项目示范

（二）CARS-17-08B：基于北斗的甜菜智能化生产与管控关键技术研究

1、任务名称：基于北斗的甜菜生产机械自动导航关键技术研究

2、研发背景：近年来，我国甜菜全程机械化生产方式在甜菜主产区已大面积推广应用，农机北斗导航产品普及率也越来越高。甜菜种植的立地条件比较差，多为地势起伏较大的坡地和沙壤地，这对农机北斗导航系统的控制性能影响较大，常影响播种质量。甜菜全程机械化生产的重要配套动力是拖拉机，但以14.7-51.45kW中小型马力段拖拉机居多，因性价比原因，这个马力段的拖拉机北斗导航普及率不高。为此，本任务提出研究基于北斗的甜菜生产机械自动导航关键技术，提高农机北斗导航系统在坡地和沙地的适应性，同时降低系统成本，以满足我国甜菜生产机械高质量作业的需求。

3、核心技术与实施内容：

（1）五年总体核心技术和实施内容：

开展基于北斗的甜菜生产机械自动导航关键技术研究，具体包括：①坡地条件下的拖拉机位置姿态检测技术；②坡地条件下的拖拉机导航路径跟踪控制技术；③中小型轮式拖拉机低成本导航系统解决方案及实现技术等。在此基础上，创制适应坡地作业的拖拉机北斗导航系统和适应中小型拖拉机的低成本北斗导航系统。

（2）年度分解任务

2021 年实施内容：

初步研究坡地条件下的拖拉机导航路径跟踪控制技术，提出农机北斗导航系统可能的降成本解决方案。

2022 年实施内容：

开展坡地条件下的拖拉机位置姿态检测技术研究，提高拖拉机在地形起伏条件下的位姿检测

精度。

2023 年实施内容：

开展坡地条件下的拖拉机导航路径跟踪控制技术研究，提高拖拉机在坡地作业条件下的导航控制精度。

2024 年实施内容：

开展中小型拖拉机低成本北斗导航技术研究，提高农机北斗导航系统在中小型拖拉机上的普及率。

2025 年实施内容：

通过系统对比试验，验证系统的有效性。

4、考核指标：

（1）五年总体考核指标：

突破农机坡地位姿检测、农机坡地北斗导航路径跟踪控制等关键技术 2 项，创制适应坡地作业的拖拉机北斗导航控制系统 1 套、适应中小型拖拉机的低成本北斗导航控制系统 1 套。

（2）年度分解考核指标：

2021 年考核指标：

撰写甜菜生产机械北斗导航作业需求调研报告。

2022 年考核指标：

突破农机坡地位姿检测技术 1 项。

2023 年考核指标：

突破农机坡地北斗导航路径跟踪控制技术 1 项，创制适应坡地作业的拖拉机北斗导航控制系统 1 套。

2024 年考核指标：

创制适应中小型拖拉机的低成本北斗导航控制系统 1 套。

2025 年考核指标：

农机坡地导航性能指标优于改进前。

5、运行管理机制：

（1）任务分工：张智刚负责本岗位任务实施的组织与协调，并与各综合试验站对接。吕盛坪、肖克辉、徐梅宣、廖娟、张闻宇、王辉等负责制定甜菜生产机械自动导航关键技术研发方案，跟踪系统运行实施情况。

（2）组织交流：参加甜菜片区内综合试验站组织的项目技术交流会和工作经验交流会；组织协调机械化研究室岗位专家紧密合作、资源共享，一起研讨技术路线；对技术难题协同攻关。

6、牵头和参加的机构、人员情况

岗位科学家	岗位名称	任务分工
张智刚	智能化管控	组织实施，系统设计，体系内交流，总结
参加的团队人员	**所在单位**	**任务分工**
吕盛坪	华南农业大学	路径跟踪控制模型仿真研究
肖克辉	华南农业大学	中小型拖拉机低成本北斗导航系统设计

徐梅宣	华南农业大学	坡地条件下甜菜生产机械位姿检测技术	
廖娟	华南农业大学	制定实施方案、跟踪系统运行实施情况	
张闻宇	华南农业大学	坡地条件下甜菜生产机械路径跟踪控制技术	
王辉	潍柴雷沃重工股份有限公司	中小型拖拉机低成本北斗导航系统开发	
参加的综合试验站	**参加站长**	**参加的团队人员**	**任务分工**
呼和浩特综合试验站	樊福义	李智	内蒙古甜菜产区项目示范
赤峰综合试验站	史树德	魏磊	内蒙古甜菜产区项目示范

四、产业基础数据平台建设

CARS-17-20C：全国甘蔗/甜菜生产机械化与智能化数据库

（二）机械化研究室

数据库名称	负责人 （总负责人：刘庆庭　）	年度任务分解
甜菜生产智能化技术应用数据库	张智刚	各年度甜菜生产智能化技术应用数据库

五、其它研究任务

1、任务名称：甘蔗生产机械化发展战略研究（智能化管控部分）

2、研发背景：我国甘蔗生产机械化发展缓慢，特别是收获环节，机收率不到 5%。国务院国发 2018[42]提出到 2025 年甘蔗机收率达到 30%。由于我国甘蔗生产各主产区立地条件差异较大、经营模式和经营规模多样化，在甘蔗生产机械化推广中出现了各种各样的问题。研究我国甘蔗生产机械化的发展现状和趋势，在智能化方向提出可行发展建议，对指导我国甘蔗生产机械化发展具有重要现实意义。

3、核心技术与实施内容：

（1）五年总体核心技术和实施内容：

通过文献检索、专家咨询、现场调研等方式，研究我国甘蔗生产机械化的发展现状和趋势，分析甘蔗生产机械智能化的重点研发领域，提出甘蔗生产机械智能化的发展方向，形成研究报告。

（2）年度分解任务

2021 年实施内容：

制定研究方案。

2022 年实施内容：

调研我国甘蔗生产机械化的发展现状和趋势，分析存在的主要问题。

2023 年实施内容：

分析甘蔗生产机械智能化的重点研发领域。

2024 年实施内容：

分析甘蔗生产机械智能化的未来发展趋势。

2025 年实施内容：

形成我国 “十四五”甘蔗生产机械智能化调研报告。

4、考核指标：

（1）五年总体考核指标：

撰写形成我国 “十四五”甘蔗生产机械智能化调研报告。

（2）年度分解考核指标：

2021 年考核指标：

撰写形成 2021 年度我国甘蔗生产机械智能化调研报告。

2022 年考核指标：

撰写形成 2022 年度我国甘蔗生产机械智能化调研报告。

2023 年考核指标：

撰写形成 2023 年度我国甘蔗生产机械智能化调研报告。

2024 年考核指标：

撰写形成 2024 年度我国甘蔗生产机械智能化调研报告。

2025 年考核指标：

撰写形成 2025 年度我国甘蔗生产机械智能化调研报告，汇总形成我国 “十四五”甘蔗生产机械智能化调研报告。

六、经费预算表

2021 年-2025 年经费总预算表

科目名称	主 要 用 途	经费（万元）
1.材料和小型仪器设备购置费	在研究开发和试验示范过程中消耗的各种原材料、辅助材料等低值易耗品的采购和运输、装卸、整理等费用，以及单台(件)价值 5 万元以下(含 5 万元)的小型仪器设备购置费。	70
2．测试化验加工费	在研究开发和试验示范过程中对外支付(包括建设依托单位内部独立经济核算单位)的检验、测试、化验及加工等费用。综合试验站在 5 个示范县开展的试验示范工作，原则上按每个县 3 万元在试验站报账。特殊情况可采用任务委托方式，签订任务委托协议，明细预算支出。	50
3．燃料动力费	在研究开发和试验示范过程中相关大型仪器设备、专用科学装置等运行发生的可以单独计量的水、电、气、燃料消耗费用等。	5
4．差旅费	在研究开发和试验示范过程中开展科学实验(试验)、科学考察、业务调研、学术交流等所发生的差旅费等。差旅费的开支标准应当按照国家有关规定执行。	30
5．会议费	在研究开发和试验示范过程中为组织开展学术研讨、人员培训、咨询以及协调等活动而发生的会议费用。应当按照国家有关规定，严格控制会议规模、会议数量、会议开支标准和会期。	0
6.出版/文献/信息传播/知识产权事务费	在研究开发和试验示范过程中，需要支付的出版费、资料费、专用软件购买费、文献检索费、专业通信费、专利申请及其他知识产权事务等费用。	15
7．劳务费	在研究开发和试验示范过程中支付给没有工资性收入的相关人员(如在校研究生)和临时聘用人员等的劳务性费用。	82.5
8．管理费	在研究开发和试验示范过程中对使用依托单位现有仪器设备及房屋，日常水、电、气、暖消耗，以及其他有关管理费用的补助支出。管理费按照基本研发费预算分段超额累退比例法核定。	16.5
9．其他	除上述费用之外，在产业技术体系建设过程中发生的与产业技术体系建设和管理密切相关的其他支出。	6
合　计	275	

七、共同条款

签约各方共同遵守现代农业产业技术体系建设实施方案（试行）和专项资金管理试行办法及其它有关规定。

1. 现代农业产业技术体系建设经费要专账管理，专款专用。严格按照《现代农业产业技术体系建设专项资金管理试行办法》的有关规定和体系经费预算执行。若经费超支，由首席科学家自筹解决，不得影响体系任务执行。

2. 任务执行过程中，如需要修改原有任务和相关指标，须报主管部门审定同意。

3. 首席科学家因不可抗力不能履行任务职责时，应及时以产业技术研发中心正式文件形式报主管部门，并出具不能履行合同的证明材料。

4. 在履行任务职责过程中，由于人为因素导致任务无法完成，视情况追究有关人员责任。

5. 首席科学家依托单位应确保聘任人员和团队成员的稳定，不得随意调换。确需调换，须正式报主管部门同意。

6. 首席科学家要严格履行本任务书的各项指标，每年年底前，须提交体系年度任务执行情况总结报告、经费决算及下年度工作计划。

7. 首席科学家应细化任务书规定的各项指标，并与研究室主任、岗位科学家、综合试验站站长签订任务委托协议。

8. 体系任务书中的重点任务和数据库研发形成的知识产权及成果归国家所有，其管理及使用参照国家有关规定执行。形成的知识产权及成果统一标注“现代农业产业技术体系建设专项资金资助”（Supported by the earmarked fund for China Agriculture Research System）。

9. 在体系建设过程中，如有从国外引进的新品种或种质，必须交国家种质资源库统一登记。

10. 本任务书经各方签字、盖章后生效。在执行过程中如发生争议、纠纷时，由各方协商解决，或通过法律程序裁决。

八、任务书签约各方

产业技术体系研发中心依托单位：内蒙古自治区农牧业科学院 （公章） 依托单位法人代表（签字）： 2021 年 10 月 25 日
首席科学家（签字）： 2021 年 10 月 25 日
岗位科学家依托单位 ： （公章） 法人代表（签字）： 2021 年 10 月 25 日
岗位科学家（签字）： 2021 年 10 月 25 日

课题编号：2022YFD2001901　　**密　级：公开**

国家重点研发计划
课题任务书

课题名称：	主要　草　料全程智能化生产共性关　技术研究与应用
所属项目：	主要　草　料生产全程智能化作业装备创制与应用
所属专项：	工厂化农业关　技术与智能农机装备
项目牵头承担单位：	中国农业机械化科学研究　呼和浩特分　有公司
课题承担单位：	吉林大学
课题负责人：	袁洪方
执行期限：	2022 年 11 月 至 2027 年 10 月

中华人民共和国科学技术部制

2022 年 11 月 27 日

0003YF 2022YFD2001901 2022-11-27 21:46:38

填 写 说 明

一、任务书甲方即　目牵头承担单位，乙方即　　承担单位。

二、任务书　　“国家科技计划管理信息系统公共服务平台”，按照系统提示在线填写。

三、任务书中的单位名称，　按规范全称填写，并与单位公章一致。

四、任务书要求提供乙方与所有参加单位的合作协议，　对原件　行扫描后在线提交。

五、任务书中文字　用宋体小四号字填写。

六、凡不填写内容的栏目，　用“无”表示。

七、乙方完成任务书的在线填写，提交甲方审核确认后，用 A4 纸在线打印、装订、签章。一式八份报　目牵头承担单位签章，其中　　承担单位一份，

人一份，作为　目任务书　件六份。

八、如　目下仅　一个　　，　　任务书只　填报　　算　分。

九、涉密　　在“国家科技计划管理信息系统公共服务平台”下　任务书的电子版模板，按保密要求离线填写、报　。

十、《　目申报书》和《　目任务书》是本任务书填报的　要依据，任务书填报不得　低考核指标，不得自行对主要研究内容作大的　整。《　目申报书》、《　目任务书》和本任务书将共同作为　　程管理、综合绩效　价（　收）和监督　估的　要依据。

课题基本信息表

名称	主要　草　料全程智能化生产共性关　技术研究与应用			
编号	2022YFD2001901			
所属　目	主要　草　料生产全程智能化作业装备创制与应用			
所属专	工厂化农业关　技术与智能农机装备			
密级	■公开 □秘密 □机密	单位总数	3	
类型	□基础前沿■　大共性关　技术□应用示范研究□其他			
活动类型	□基础前沿■应用研究□　　发展			
研究 所属学科	自然科学相关工程与技术 农业工程			
成果应 用的主要国 民经济行业	制　业 专用　备制　业 农、林、牧、渔专用机械制 畜牧机械制			
的社会 经济目标	农林牧渔业发展 畜牧业			
经　　算	总　求 665.00 万元，其中中央　政专　　求 265.00 万元			
周期节点	始时	2022 年 11 月	结束时	2027 年 10 月
	实施周期	共 60 个月	计中期时　点	2025 年 07 月

承担 单位	单位名称	吉林大学	单位法定 代表人姓名	张希
	单位性	大专　校	组织机构代码	121000004232040648
	单位主管		属关系	地方
	单位所属地区	吉林省	地市（市、自治州、盟）	春市　朝　区
	信地址	春市前　大街 2699 号	政编码	130012
	单位开户名称	吉林大学		
	开户　行 （全称）	中国　行股份有　公司　春前　大街支行	汇入地点	吉林省　春市

	行　号	160402501175			行机构代码	104241017019
人	姓 名	袁洪方	性 别	■男 □女	出生日期	1984-09-20
	件类型	份	件号码			
	所在单位	吉林大学				
	最　学位	■博士□硕士□学士□其他				
	职 称	□正　级■副　级□中级□初级□其他			职 务	副主任
	电子　箱	yhf1984828@163.com		移动电		
联系人	姓 名	荣强	电子　箱	zrq@jlu.edu.cn		
	固定电	0431-85095253	移动电			
	件类型	份	件号码			
务人	姓 名	春明	电子　箱	gaocm@jlu.edu.cn		
	固定电	0431-85168008	移动电			
	件类型	份	件号码			

其他参与单位	序号	单位名称	单位性	组织机构代码
	1	华南农业大学	大专　校	124400004554165634
	2	山东巨明机械有　公司	国有企业	913703211644192l1L

参加人数		
	24 人。其中：	级职称 4 人，中级职称 6 人，初级职称 1 人，其他 13 人；
		博士学位 6 人，硕士学位 3 人，学士学位 11 人，其他 4 人。

简介（ 500 字以内）	
	对优　草　料生产对关　装备的　求，围绕种植、收获、储　等主要生产环节，探索不同环节作业　程整机与关　件的土壤激励　规律，研究茎叶在机械系统中的　移规律，突破　地自动仿形、低扰损平茬、减　耗仿生切碎、装备作业参数与作业　全程智能监控等共性关　技术，构建优化的农机农艺融合智能作业技术体系；研制切割　度和角度实时　整的割台自动仿形装置，　线　仿生圆盘式平茬装置，低耗仿生切割装置，开发具有作业　、品　以及机器扭矩、振动等作业状态参数的实时检测系统，创制我国　草　料生产核心技术，并　示范　行　，提升我国　草机械的自主研发能力与市场竞争力。 期指标：突破关　技术 4　及以上，研制关　装置 4　及以上，申　专利 7 件及以上，发表 SCI/EI　文 5 篇及以上，　件著作权 4 件及以上。

一、目标及考核指标、考核方式/方法

目标、 期成果与考核指标表

目标[1]	期成果			考核指标[2]				考核方式（方法）及 价手段[4]
	期成果名称		期成果类型	指标名称	立 时已有指标值/状态	中期指标值/状态[3]	完成时指标值/状态	
（ 500 字以内。） 对优 草 料生产对关装备的 求，围绕种植、收获、储 等主要生产环节， 探索不同环节作业 程整机与关 件的土壤激励 规律，基于草种、种茎、茎叶在机械系统中的 移规律，开发关 技术装备与 程 控制系统，实现 草料生产全程智能化，研制 线仿生圆盘式平茬装置、全程作业 在线测控系统等关 装置，创制我国 草 料生产核心技术，并 示范 行 ，提升我国 草机械的自主研发能力与市场竞争力。 期指标：突破关 技术 4 及以上，研制关 装置 4 及以上，申 专利 7 件及以上，发表 SCI/EI 文 5 篇及以上，申 件著作权 4 件及以上。	**主要成果**	1	主要草 料全程智能化生产共性关 技术及装备	□新理 □新原理 □新产品 ☑ 新技术 □新方法 ☑ 关 件 □数据库 □ 件 □应用解决方案 □实 装置/系统 □临床指南/规范 □工程工艺 □标准 ☑ 文 ☑ 发明专利 ☑ 其他实用新型专利、 件著作权	指标 1.1 数 指标	无	①关 技术2 及以上； ②关 装置1 及以上； ③申 发明专利 1 件及以上、实用新型专利 1 件及以上	①关 技术 4 及以上； ②关 装置 4 个及以上； ③申 发明专利 4 件及以上、实用新型专利 3 件及以上； ④ 件著作权 4 件及以上	①新技术、关 件：技术查新报告或申 专利； ②专利、 件著作权：符合 目研究内容，且取得受理 知书或授权 书
				指标 1.2 技术指标	切割茬口破损率≤ 7%	无	①留茬 度差小于 15mm； ②切割茬口破损率≤5%； ③监测准确率≥90%	有法定 的第三方检测机构依据相关标准、定大纲 行检测或 定
				指标 1.3 指标	无	SCI/EI 文 2 篇及以上	SCI/EI 文 5 篇及以上	符合 目研究内容，且第一标注 助 目为“国家 点研发计划 助”字样及本 目编号的录用 知或见刊 明

科技报告考核指标	序号	报告类型[5]	数	提交时	公开类别及时 [6]
	1	年度技术　展报告	1	2023.12	延期公开、3 年
	2	年度技术　展报告	1	2024.12	延期公开、3 年
	3	中期科技报告	1	2025.07	延期公开、3 年
	4	年度技术　展报告	1	2025.12	延期公开、3 年
	5	年度技术　展报告	1	2026.12	延期公开、3 年
	6	最终科技报告	1	2027.10	延期公开、3 年
其他目标与考核指标 培养研究生 6 人及以上，技术　干3 人及以上。					

备注：

1.“　　目标”，应从以下方　明确描　：（1）研发主要　对什么　　和　求；（2）将要解决哪些科学　、突破哪些核心/共性/关　技术；（3）　期成果；（4）成果将以何种方式应用在哪些　域/行业/大工程等，并拟在科技、经济、社会、环境或国　安全等方　发挥何种的作用和影响。（5）所列主要成果原则上不　　5　，如有其他　要成果放在“其他”成果中表　。

2.“考核指标”，指相应成果的数　指标、技术指标、　　指标、应用指标和产业化指标等，其中，数　指标可以为专利、产品等的数　，　文代表作应注　　，不以数　作为　价标准；技术指标可以为关　技术、产品的性能参数等；　　指标可以为产品的耐　动、　低温、无故　　行时　等；应用指标可以为成果应用的对　、范围和效果等；产业化指标可以为成果产业化的数　、经济效益等。同时，对各　考核指标　填写立　时已有的指标值/状态以及　　完成时要到　的指标值/状态。同时，考核指标也应包括支撑和服务其他　大科研、经济、社会发展、生态环境、科学普及　求等方　的直接和　接效益。如对国家　大工程、社会民生发展等提供了关　技术支撑，成果　让并带动了环境改善、实现了　售收入等。若某　成果属于开创性的成果，立　时已有指标值/状态可填写“无”,若某　成果在立　时已有指标值/状态　以界定，则可填写“/”。

3.“中期指标”，各专　根据管理特点，确定是否填写，　励　段目标明确的　目　　填写中期指标。

4.“考核方式方法”，应提出符合相关研究成果与指标的具体考核技术方法、测算方法等。

5.“科技报告类型”，包括　目综合绩效　价（　收）前撰写的全　描　研究　程和技术内容的最终科技报告、　目年度或中期检查时撰写的描　本年度研究　程和　展的年度技术　展报告以及在　目实施　程中撰写的包含科研活动细节及基础数据的专　科技报告（如实　报告、　　报告、　研报告、技术考察报告、　计报告、测　报告等）。其中，每个　目在综合绩效　价（　收）前应撰写一份最终科技报告；研究期　　2 年（含 2 年）的　目，应根据管理要求，每年撰写一份年度技术　展报告；每个　目可根据研究内容、期　和经　强度，撰写数　不等的专　科技报告。科技报告应按国家标准规定的格式撰写。

6.“公开类别及时　”，公开　目科技报告分为公开或延期公开，内容　要发表　文、申　专利、出版专著或涉及技术　窍的，可标注为“延期公开”。　要发表　文的，延期公开时　原则上在 2 年（含 2 年）以内；　要申　专利、出版专著的，延期公开时　原则上在 3 年（含 3 年）以内；涉及技术　窍的，延期公开时　原则上在 5 年（含 5 年）以内。涉密　目科技报告按照有关规定管理。

二、课题研究内容、研究方法及技术路线

（一）课题的主要研究内容

拟解决的关键科学问题、关键技术问题，针对这些问题拟开展的主要研究内容，限1000字以内。

1、拟解决的关键科学问题或关键技术问题

通过物理参数测定、光谱分析、力学分析等方法，揭示草种、饲草收获环节作业过程整机与关键部件的土壤激励谱、切割/切碎刀具-茎秆互作、茎叶在机械系统中的运移等规律。重点解决随地自动仿形、低扰损平茬、减阻降耗仿生切碎、全程作业质量在线测控等关键技术。

2、主要研究内容

（1）随地自动仿形技术及装置研究

基于深度学习的多模态信息融合算法，使用带有电位器的仿形地轮或采用多点超声波传感器与深度相机等获取地面地形数据，基于土壤-植物-机器系统的不同环节作业过程整机与关键部件的土壤激励谱规律，提出满足高速高效仿形工作的割台设计参数，搭建具有上下浮动、纵横向角度调整功能的高速响应闭环控制试验台，利用现代CAD 设计方法完成虚拟验证，研制随地自动仿形装置。

（2）低扰损平茬技术与装置研究

以棉蝗、天牛等昆虫动物口器生物结构为仿生原型，进行口器轮廓曲线提取，建立仿生切割部件模型，利用有限元软件，对具有低扰损平茬作用的切割部件进行参数优化，搭建仿生平茬试验台，进行效果验证，研究切割刀具-茎秆互作规律，并针对构树、柠条、杂交狼尾草等优质蛋白型饲料作物，研发具有仿生结构的低扰损高线速圆盘式平茬装置，提高切割茬口平整度，降低切割茬口破损率和平茬阻力。

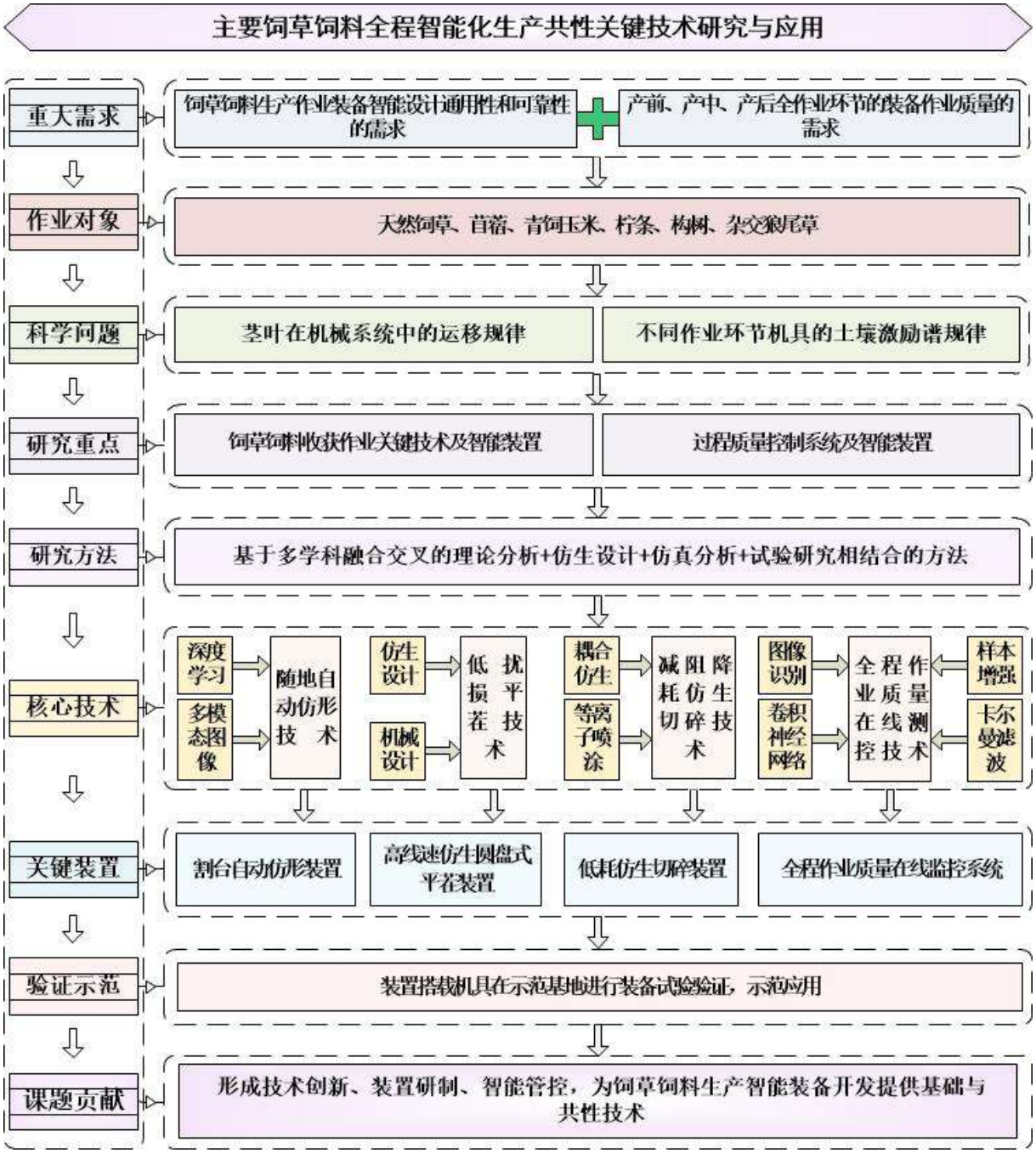
（3）减阻降耗仿生切碎技术与装置研究

以青饲玉米茎秆为研究对象，从切割特性和微观组织结构入手，研究其力学特性和切割特性。基于昆虫口器形状仿生原理，结合非光滑仿生表面减粘降阻机理，对茎秆切割与切碎部件进行耦合仿生设计，建立刃形和刃面双耦合仿生结构模型，进行仿真分析与参数优化。利用等离子喷涂、激光熔覆等技术对耦合仿生切碎部件进行表面强化处理，分析强化层的界面结合力、耐磨减阻等性能，搭建仿生切碎试验台并进行验证，研制低耗仿生切碎装置。

（4）全程作业质量在线测控技术及装置研究

基于多光谱视觉-IMU-GNSS 等多模态信息，通过深度学习、卡尔曼滤波等融合算法，采集工作通量、割茬高度、切段长度等作业参数，为作业参数与作业质量监控及闭环工作提供依据，并通过转轴扭矩、振动参数等实时检测实现对事故隐患预警；开发基于 CAN 总线与 USB3.0/IEEE 1394 总线组成的适应连接多路慢速传感器与高速相机的车载分级总线网络，满足即插即用可剪裁的灵活配置需求。按照非接触测量-分级总线传输-集散式信号处理的原则，研制作业参数与作业质量检测评价系统，提高装备的智能化程度。

主要研究内容依据以下技术路线展开：



（二）课题采取的研究方法

针对课题研究拟解决的问题，拟采用的方法、原理、机理、算法、模型等

限 1000 字以内。

针对饲草繁殖保育与营养保全机械化生产过程中智能作业装备适应性差、作业效率低等问题，课题基于多学科交叉，充分融合工程仿生、智能监控和机构优化设计等技术，开展饲草繁殖保育与营养保全机械化智能作业存在的关键共性技术研究，通过理论建模分析、运动学与动力学计算机模拟仿真、虚拟部件优化设计、参数优化台架试验、田间对比验证试验等研究方法，研究关键共性技术的机理与规律，突破关键共性技术瓶颈，研制出相对应的装置与系统。

1、机理研究方法，通过物理参数测定、光谱分析、力学分析等方法，探索不同环节作业过程机器与机具的土壤激励谱规律，揭示切割/切碎刀具-茎秆互作规律，研究茎叶在机械系统中的运移规律。

2、针对随地自动仿形技术，采用基于深度学习的多模态机器视觉信息融合方法，开展随地自动仿形技术研究。

3、针对低扰损平茬技术和减阻降耗仿生切碎技术，通过体视显微镜观察其口器形状，利用 Matlab 软件进行轮廓曲线提取，并进行切割过程力学特性分析，将曲线作为阵列单元，建立仿生切割部件模型，采用工程仿生与机械设计相结合的方法，开展低扰损平茬技术与减阻降耗仿生切碎技术研究。

4、针对全程作业质量在线测控技术，综合运用多光谱视觉-IMU-GNSS 等多模态信息通过深度学习、卷积神经网络、样本增强、多信息融合、卡尔曼滤波等理论与方法，开展工作通量、转轴扭矩、振动参数、割茬高度、切断长度、压缩密度、草捆长度等作业参数与作业质量监控研究。开发基于 CAN 总线与 USB3.0/IEEE 1394 总线组成的适应连接多路慢速传感器与高速相机的车载分级总线网络，满足即插即用可剪裁的灵活配置需求。按照非接触测量-分级总线传输-集散式信号处理的原则，设计作业参数与作业质量检测评价系统

5、针对关键装置，采用计算机仿真、台架试验、田间试验相结合的方式开展关键装置及系统研究。

三、主要创新点

围绕基础前沿、共性关键技术或应用示范等层 ，简述课 的主要创新点。具体内容应包括该 创新的基本形态及其前沿性、时效性等，并说明是否具备方法、理论和知识产权特征。每 创新点的描述 500 字以内。

1、创新点 1：针对 草 料生物性状与机械特性、多年生与周年多茬刈割、收获工艺与营养保全利用等差异化特征，阐释机械作业匹配作物属性相关科学问 。探索土壤-植物-机器系统的不同环节作业过程机器与机具的土壤激励谱与作业机器设计动态指标的的匹配规律，探究复杂环境 速 效 地自动仿形控制机理、多年生 草 料作物低扰损平茬切割损伤机理、昆虫口器轮廓曲线与 光滑表 耦合仿生减 耗切碎机理，突破作业装备研发理论基础薄弱瓶 ，拓展 草 料生产土壤-机器-植物互作理论，具有基础性创新。

2、创新点 2：低扰损仿生平茬切割与减 耗仿生切碎技术。基于对昆虫动物口器在咬合作物茎秆过程中的低耗减 机理分析，探索其结构形态和运动方式等对减 效果的影响因素，进而通过结构形态与材料表 的耦合仿生设计等方法，研制出 线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置，保证切割茬口平整度， 低切割作业能耗，提 草品质，为实现 草 料增产保质提供关键技术支撑，具有技术性创新。

3、创新点 3：全程作业质量在线测控技术。基于多光谱视觉-IMU-GNSS 等多模信息通过深度学习、卡尔曼滤波等融合算法，通过 接触测量-分级总线传输- 散式信号处理方式，开发作业参数与作业质量检测评价系统，满足即插即用可剪裁的灵活配置 求，具有一定通用性，为提 草 料作业装备的智能化程度提供关键技术支撑，具有技术性创新。

四、 预期经济社会效益

课题的科学、技术、产业预期指标及科学价值、社会、经济、生态效益。限 500 字以内。

1、项目的科学、技术、产业预期指标及科学价值

通过开展主要饲草饲料全程智能化生产共性关键技术研究，开发“传感模块-下位机-上位机”嵌入式控制系统，突破传统智能控制技术存在的通用性与可靠性不足的瓶颈，为智能装备开发提供全面可靠的设计依据、判别依据，提升我国饲草机械基础研究能力和智能化水平。

2、社会、经济、生态效益

针对优质饲草饲料生产对关键装备需求，围绕种植、收获、储运等主要生产环节，开发饲草繁殖保育与营养保全机械化智能作业技术与装备，实现饲草饲料生产全程智能化，缩短与国外饲草饲料生产装备的差距，可有效提高农业生产作业效率，降低作业能耗，提升饲草品质，减少饲草收获损失，提高天然草原生产能力，有利于提升我国饲草饲料全程作业装备技术水平，提高技术创新对农牧业发展的科技贡献率，保障国家粮食安全。因此，课题成果的实施可具有重要的社会、经济和生态效益。

五、课题年度计划

按每 6 个月制定形成课题的计划进度，应将课题的考核指标分解落实到年度计划中。

1、年度：2022 年 11 月—2023 年 04 月

任　　务：①制定课题总体实施方案，明确各参加单位任务分工；②开展子课题启动与研究工作，落实总体的进度安排；③开展饲草繁殖保育与营养保全机械化智能作业质量调研，探索不同环节作业过程机器与机具的土壤激励谱规律，研究草种、种茎、茎叶在机械系统中的运移规律；④撰写专利。

考核指标：申报发明专利 1 件及以上

成果形式：发明专利

2、年度：2023 年 05 月—2023 年 10 月

任　　务：①开展随地自动仿形技术、低扰损平茬技术、减阻降耗仿生切碎技术、全程作业质量在线测控技术研究；②进行割台自动仿形装置、高线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置、全程作业质量在线测控系统的第一轮设计与试制；③撰写专利。

考核指标：①完成割台自动仿形、高线速仿生圆盘式平茬，低耗仿生滚筒式切碎、全程作业质量在线测控系统等装置的设计；②申报实用新型专利 1 件及以上。

成果形式：实用新型专利

3、年度：2023 年 11 月—2024 年 04 月

任　　务：①进行割台自动仿形装置、高线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置、全程作业质量在线测控系统的台架试验与数据分析；②撰写论文。

考核指标：①完成割台自动仿形装置、高线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置、全程作业质量在线测控系统的第一轮试制；②撰写 SCI/EI 论文 1 篇及以上；③撰写年度科技报告 1 份。

成果形式：①装置样件；②论文；③年度科技报告。

4、年度：2024 年 05 月—2024 年 10 月

任　　务：①进行割台自动仿形装置、全程作业质量在线测控系统的第二轮优化改进、试制与试验；②撰写论文。

考核指标：①完成割台自动仿形装置、全程作业质量在线测控系统的第二轮试制；②撰写 SCI/EI 论文 1 篇及以上。

成果形式：①装置样件；②论文。

5、年度：2024 年 11 月—2025 年 04 月

任　　务：①进行高线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置的第二轮优化改进、试制与试验；②开展课题中期检查工作。

考核指标：①完成高线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置的第二轮试制；②撰写年度科技报告 1 份；③撰写中期进展报告 1 份。

成果形式：①装置样件；②年度科技报告；③中期进展报告。

6、年度：2025 年 05 月—2025 年 10 月

任　　务：①进行割台自动仿形装置、高线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置、全程作业质量在线测控系统的定型试制与试验；②撰写专利与论文。

考核指标：①完成割台自动仿形装置、高线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置、全程作业质量在线测控系统的定型试制；②申报发明专利和实用新型专利各 1 件及以上；③撰写 SCI/EI 论文 1 篇及以上。

成果形式：①定型装置；②发明专利；③实用新型专利；④论文。

7、年度：2025 年 11 月—2026 年 04 月

任　　务：①进行割台自动仿形装置、高线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置、全程作业质量在线测控系统的检验检测；②撰写专利与论文；③开展课题中期检查工作。

考核指标：①完成割台自动仿形装置、高线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置、全程作业质量在线测控系统的检验检测；②申报发明专利 1 件及以上；③撰写 SCI/EI 论文 1 篇及以上；④撰写年度科技报告 1 份。

成果形式：①检验报告；②发明专利；③论文；④年度科技报告。

8、年度：2026 年 05 月—2026 年 10 月

任　　务：①进行割台自动仿形装置、高线速仿生圆盘式平茬装置，低耗仿生滚筒式切碎装置、全程作业质量在线测控系统与机具融合；②撰写专利与软件著作权。

考核指标：①申报发明专利和实用新型专利各 1 件及以上；②撰写软件著作权 2 件及以上；③培养研究生 3 人及以上。

成果形式：①发明专利；②实用新型专利；③软件著作权。

9、年度：2026 年 11 月—2027 年 04 月

任　　务：①进行割台自动仿形装置、高线速仿生圆盘式平茬装置，低耗仿生滚筒式切

碎装置、全程作业质量在线测控系统的试验考核；②撰写论文与软件著作权。

考核指标：①撰写 SCI/EI 论文 1 篇及以上；②撰写软件著作权 2 件及以上；③培养技术骨干 3 人及以上；④撰写年度科技报告 1 份。

成果形式：①论文；②软件著作权；③年度科技报告。

10、年度：2027 年 05 月—2027 年 10 月

任　　务：①进行课题任务书各项指标完成度自查；②课题文件、资料准备，进行课题绩效评价与总结。

考核指标：①完成课题绩效评价；②培养研究生 3 人及以上；③撰写最终科技报告 1 份。

成果形式：①绩效评价意见；②最终科技报告。

六、课题组织实施机制及保障措施

1、课　的内部组织管理方式、协调机制等，　500 字以内。

（1）课　由吉林大学牵头实施，对课　的技术研究、资金筹措、课　进度、攻关质量实行全过程负责。课　参加单位华南农业大学和山东巨明机械有　公司严格履行“联合实施协议”的所有条款并承担相应的法律责任。吉林大学作为课　主持单位，确保国拨专　经费及时拨付给参与单位，保　课　的　利进行。

（2）各参与单位联合成立课　　导小组，由课　负责人任组长，课　组成员由各单位课　主管人员组成，　导小组切实做好课　运行管理协调工作，主要包括课　计划执行、经费　算控制、资金使用和组织管理、实施进度等内容考核，定期召开课　实施协调会，督促各参加单位实施工作，做到科学规划、精心设计、严密组织，搞好各环节的衔接，保证　目按期完成。

（3）完善和坚持严格的管理制度。实施定期检查和年报制度，对课　的执行情况进行动态跟踪管理，确保进度和质量；严格遵守有关国家课　的管理和财务等各　规定，加强对课　经费使用的监督和管理，设立科研专　资金账户，专款专用；严格使用程序，保证课　经费合法、合规和科学使用，保　课　　利开展。

2、课　实施的相关政策，已有的组织、技术基础，支撑保　条件，　500 字以内。

（1）政策保　：本课　依据《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》《“十四五”全国　草产业发展规划》（农牧发〔2022〕7 号）、《“十四五”全国农业机械化发展规划》（农机发〔2021〕2 号）、《农业农村部关于加快畜牧业机械化发展的意见》（农机发〔2019〕6 号）、《国务　办公厅关于改　完善中央财政科研经费管理的若干意见》（国办发〔2021〕32 号），立足“智能、　效、绿色”，按照“关键核心技术自主化，主导装备产品智能化，薄弱环节机械化”的发展思路，进行装置的研究开发。

（2）已有的组织、技术基础，支撑保　条件：课　由吉林大学牵头，联合华南农业大学、山东巨明机械有　公司等 2 家单位开展研究。课　团队汇　了我国　草　料生产机械化技术与装备创制的优势科研单位和生产企业，有较好的合作基础，具备圆满完成课　各　研究任务的基础和能力，保　课　　利实施。课　主持单位和参加单位已在本课　相关研究　域取得了一系列阶段性重要成果，建有工程仿生国家地方联合实　室、农业农村部　效播种收获装备重点实　室等相关研究　域的　水平实　和测试平台，为

完成课　提供了硬件保　。

3、对实现　目总目标的支撑作用，及与　目内其他课　的协同机制，　500 字以内。

（1）对实现　目总目标的支撑作用：针对　目总目标要求的“阐释机械作业匹配作物属性相关科学问　，突破作业装备研发理论基础薄弱瓶　，研制割台自动仿形、低耗仿生切碎、全程作业质量智能监测等关键装置”等内容。本课　从产前、产中、产后全作业环节的作业质量，综合运用工程仿生、信息技术、深度学习、智能监控和结构优化等技术，开展关键技术装备与过程质量控制系统研究，能够避免重复研究、减少研究成本、提　共性技术的通用性，为其他 4 个课　的智能装备开发提供全　可　的设计依据、判别依据，能够弥补我国　草　料装备智能化的共性技术短板。课　目标完全覆盖　目总目标中有关　草　料全程智能化生产共性关键技术部分的分目标，并对　目总目标的实现有着强有力的支撑作用。

（2）与　目内其他课　的协同机制：课　　要加强与　目内其他课　的协同、衔接。依托　目办公室的组织协调和各课　组间的沟通制度，有效实现各课　组之间协作和联合，实现信息与科技资源共享，协同推进　目实施。加强与国家重点研发计划工厂化农业关键技术与智能农机装备专　中其他　目合作，以汇聚资源、相互支撑，相互借鉴，充分吸收各方优秀成果和先进思想，加快课　研发进度。

七、知识产权对策、成果管理及合作权益分配

限 500 字以内。

1、课题实施过程中，按照国家科技成果相关规定，严格执行《科技成果登记办法》。

2、课题实施过程中，严禁弄虚作假、徇私舞弊、剽窃他人成果等科研不端行为。在不影响课题的专利申请或其他知识产权保护的前提下，明确要求依托本课题所取得的所有研究成果，包括但不限于论文、专著、样机、样品、报道、软件、数据库和奖项等，应标注“国家重点研发计划资助”字样及课题或项目编号，英文标注：“National Key R&D Program of China”。第一标注的成果作为验收或评估的确认依据。

3、课题实施过程中，根据课题任务分工，课题参与单位独立完成的科技成果及其相应知识产权归各方独自所有。

4、课题实施过程中涉及课题组内部合作、与课题组外单位合作的研究成果及其相应知识产权归合作各方共有，将由合作方就责、权、利达成共识，根据投资比例多少及智力投入程度依次进行分配，并签订合作协议。

5、各方对共有科技成果实施许可、转让专利技术、非专利技术而获得的经济收益由各方共享，收益共享方式应在行为实施前另行约定。

本课题不在各方之间建立任何商业上的代理、合作关系，如双方希望建立任何商业上的代理、合作关系的，应另行签订协议。

八、需要约定的其他内容

限 500 字以内。

课题数据汇交和安全管理按如下约定执行：

1、课题承担单位吉林大学与课题参与单位华南农业大学、山东巨明机械有限公司将按照科技计划项目科学数据汇交的有关要求，制定科技资源汇交方案，将科学数据汇交到有关方面认可的科学数据中心并出具汇交凭证。

2、课题承担单位吉林大学与课题参与单位华南农业大学、山东巨明机械有限公司将按照国家重点研发计划项目安全管理的有关要求，切实履行项目安全管理职责，加强人员培训教育，强化科研过程安全管理。

九、课题参加人员基本情况表

填表说明： 1.专业技术职称：A、正高级　B、副高级　C、中级　D、初级　E、其他；
2.投入本课题的全时工作时间（人月）是指在课题实施期间该人总共为课题工作的满月度工作量；累计是指课题组所有人员投入人月之和；
3.课题固定研究人员需填写人员明细；
4.是否有工资性收入：Y、是　N、否；
5.人员分类代码：B、课题负责人　C、项目/课题骨干　D、其他研究人员；
6.工作单位：填写单位全称，其中高校要具体填写到所在院系。

序号	姓名	性别	出生日期	证件类型	证件号码	专业技术职称	职务	最高学位	专业	投入本课题的全时工作时间（人月）	人员分类代码	在课题中分担的任务	是否有工资性收入	工作单位
1	袁洪方	男	- -	身份证		副高级	副主任	博士	农业机械化工程	40	课题负责人	课题总体方案制定与实施	是	吉林大学生物与农业工程学院
2	赵祚喜	男	- -	身份证		正高级	系主任	博士	农业机械化工程	30	课题骨干	作业质量在线测控技术研究	是	华南农业大学工程学院
3	崔守波	男	- -	身份证		副高级	总经理	其他	经济管理	30	课题骨干	关键装置试制	是	山东巨明机械有限公司
4	赵荣强	男	- -	身份证		中级	无	博士	农业机械化工程	30	课题骨干	仿生滚筒式切碎装置研究	是	吉林大学生物与农业工程学院
5	可欣荣	男	- -	身份证		中级	无	博士	农业机械化工程	18	其他研究人员	割台仿形控制系统研究	是	华南农业大学工程学院
6	吕盛坪	男	- -	身份证		副高级	无	博士	工业与制造系统工程	18	其他研究人员	自动仿形装置研究	是	华南农业大学工程学院

7	谢家兴	男	- -	身份证		中级	无	博士	农业电气化与自动化	18	其他研究人员	自动仿形装置研究	是	华南农业大学电子工程学院
8	高奎增	男	- -	身份证		中级	副总经理	其他	机械制造与工艺	30	其他研究人员	关键装置试验示范	是	山东巨明机械有限公司
9	张敏	男	- -	身份证		中级	技术开发部长	学士	机械设计制造与自动化	30	其他研究人员	关键装置试验示范	是	山东巨明机械有限公司
10	王克恒	男	- -	身份证		中级	技术总工	其他	农机机械	30	其他研究人员	关键装置试验示范	是	山东巨明机械有限公司
11	邵强	男	- -	身份证		初级	技术部长助理	其他	机械制造	30	其他研究人员	关键装置试验示范	是	山东巨明机械有限公司
12	曹庆秋	男	- -	身份证		其他	无	硕士	机械	30	其他研究人员	仿生滚筒式切碎装置研究	否	吉林大学生物与农业工程学院
13	薛钊	男	- -	身份证		其他	无	学士	农业机械化工程	30	其他研究人员	仿生滚筒式切碎装置研究	否	吉林大学生物与农业工程学院
14	郑宜强	男	- -	身份证		其他	无	学士	农业机械化工程	30	其他研究人员	仿生圆盘式平茬装置研究	否	吉林大学生物与农业工程学院
15	王云喆	男	- -	身份证		其他	无	学士	农业机械化工程	30	其他研究人员	仿生圆盘式平茬装置研究	否	吉林大学生物与农业工程学院
16	张岩	男	- -	身份证		其他	无	学士	农业机械化工程	30	其他研究人员	试验研究与数据分析	否	吉林大学生物与农业工程学院
17	袁凯	男	- -	身份证		其他	无	硕士	农业工程	30	其他研究人员	作业质量在线测控系统	否	华南农业大学工程学院

												研究		
18	罗阳帆	男	- -	身份证		其他	无	硕士	农业工程	30	其他研究人员	作业质量在线测控系统研究	否	华南农业大学工程学院
19	王乾	男	- -	身份证		其他	无	学士	农业工程	30	其他研究人员	计算机仿真分析与研究	否	华南农业大学工程学院
20	米亚龙	男	- -	身份证		其他	无	学士	机械	30	其他研究人员	计算机仿真分析与研究	否	华南农业大学工程学院
21	廖志辉	男	- -	身份证		其他	无	学士	机械	30	其他研究人员	试验数据分析与研究	否	华南农业大学工程学院
22	夏丹燕	女	- -	身份证		其他	无	学士	机械	30	其他研究人员	试验数据分析与研究	否	华南农业大学工程学院
23	赖仕俊	男	- -	身份证		其他	无	学士	农业工程	30	其他研究人员	计算机仿真分析与研究	否	华南农业大学工程学院
24	陆俊君	男	- -	身份证		其他	无	学士	农业工程与信息技术	30	其他研究人员	计算机仿真分析与研究	否	华南农业大学工程学院
固定研究人员合计										694	/	/	/	/
流动人员或临时聘用人员合计										0	/	/	/	/
累计										694	/	/	/	/

课题预算表

表B1　　课题编号：2022YFD2001901　　课题名称：主要饲草饲料全程智能化生产共性关键技术研究与应用　　金额单位：万元

序号	预算科目名称	金额
	（1）	（2）
1	一、中央财政专项资金	265.00
2	（一）直接费用	204.00
3	1.设备费	45.50
4	其中：购置设备费	
5	2.业务费	119.70
6	3.劳务费	38.80
7	（二）间接费用	61.00
8	二、其他来源资金	400.00
9	三、合计	665.00

注：1.间接费用无需编制预算说明；2.绩效支出在间接费用中无比例限制。承担单位在统筹安排间接费用时，要处理好合理分摊间接成本和对科研人员激励的关系，绩效支出安排与科研人员在课题工作中的实际贡献挂钩。

设备费——购置/试制设备预算明细表

表B2　课题编号：　2022YFD2001901　课题名称：　主要饲草饲料全程智能化生产共性关键技术研究与应用　金额单位：万元

填表说明：
1.设备分类：购置、试制；
2.购置设备类型：通用、专用；
3.试制设备不需填列本表（9）列、（10）列、（11）列、（12）列；
4.设备单价的单位为万元/台套，设备数量的单位为台套；
5.单价50万元以下的设备不用填写；
6.本表只填写中央财政资金购置（试制）的设备。

序号	设备名称	设备分类	功能和技术指标	单价	数量	金额	购置或试制单位	安置单位	购置设备类型	主要生产厂家及国别	规格型号	拟开放共享范围
	（1）	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
无记录												
单价50万元以上购置设备合计							/	/	/	/	/	/
单价50万元以上试制设备合计							/	/	/	/	/	/
累计							/	/	/	/	/	/

课题单位经费预算明细表

表B3 课题编号： 2022YFD2001901 课题名称： 主要饲草饲料全程智能化生产共性关键技术研究与应用 金额单位：万元

填表说明： 1.单位类型分课题承担单位、课题参与单位；
2.组织机构代码指企事业单位国家标准代码，单位若已三证合一请填写单位统一社会信用代码，无组织机构代码的单位填写“000000000”。

序号	单位名称	组织机构代码-统一社会信用代码		单位类型	任务分工	研究任务负责人	合计	中央财政专项资金		其他来源资金
								小计	其中：间接费用	
	（1）	（2）	（3）	（4）	（5）	（6）	（7）	（8）	（9）	（10）
1	吉林大学	统一社会信用代码	121000004232040648	课题承担单位	主要负责课题的总体实施，方案制定，低扰损平茬、减阻降耗仿生切碎和全程作业质量在线测控关键技术与装置研究，试验研究与分析，专利、论文等成果的完成，组织课题验收等工作。	袁洪方	170.00	170.00	39.00	
2	华南农业大学	统一社会信用代码	124400004554165634	课题参与单位	主要负责随地自动仿形、收获机械作业质量在线测控技术及装置研究，试验研究与分析，专利、论文等成果的完成，协助课题验收等工作。	赵祚喜	65.00	65.00	13.00	
3	山东巨明机械有限公司	统一社会信用代码	91370321164419211L	课题参与单位	主要负责关键部件的研制，试验示范，协助课题主持单位完成课题验收等相关工作。	崔守波	430.00	30.00	9.00	400.00
累计							665.00	265.00	61.00	400.00

预算说明

一、中央财政资金

预算的编制要坚持任务相关性、政策相符性和经济合理性，实事求是编制提出课题预算。填报时，直接费用应按设备费、业务费、劳务费三个类别填报，每个类别结合科研任务按支出用途进行说明。除 50 万元以上的设备外，其他费用只提供基本测算说明，不需要提供明细。

1.设备费（是指项目实施过程中购置或试制专用仪器设备，对现有仪器设备进行升级改造，以及租赁外单位仪器设备而发生的费用等。计算类仪器设备和软件工具可在设备费科目编列。填报时，50 万元以上的设备详细说明，50 万元以下的设备费用分类说明）

1.1 购置设备费： 0.00 万元。

1.2 试制设备费： 共计 37.00 万元，占专项经费比例 13.96%，具体测算依据如下：

（1）草种收获割台仿形控制试验台，用于割台上下浮动高度、割台横向倾角、割台纵向倾角等参数采集，进行割台仿形控制测试，计 8.00 万元。

（2）饲草仿生平茬切割试验台，用于测试仿生刀具切割构树、柠条、杂交狼尾草等饲草茎秆的切割力、切割时间、切割茬口整齐度和刀具运动状态、变形、磨损等参数，计 7.50 万元。

（3）青饲玉米切碎试验台，用于测试青饲玉米切段长度、切碎均匀性、切碎能耗、切碎阻力等性能参数，计 9.00 万元。

（4）苜蓿压缩在线测控试验台，用于测试苜蓿压缩过程中 XYZ 三个方向的压力参数、压缩时间以及压缩密度、草捆长度等作业参数，计 8.50 万元。

（5）饲草茎秆力学性能夹持装置，用于苜蓿、青贮玉米、构树、柠条、杂交狼尾草等单个饲草茎秆的拉伸、弯曲、剪切、压缩等力学性能试验，计 4.00 万元。

1.3 设备升级改造与租赁费：共计 8.50 万元，占专项经费比例 3.21%，具体测算依据如下：

（1）室内排种试验台升级改造，用于模拟实际作业条件测试振动频率、加速度等参数，进行牧草排种质量影响因素研究，计 3.70 万元。

（2）传动测试试验台升级改造，用于测试青饲收获、打捆收获等传动机构如带轮、齿轮、链轮的转速、扭矩、振动频率等参数，计 3.00 万元。

（3）用于租赁相关收获装备租赁，将研制的割台自动仿形、高线速仿生圆盘式平茬、低耗仿生切碎、全程作业质量在线测控等关键装置搭载在租赁的机具上进行田间试验，5 年预计共需试验 5 次，平均每次 4 天，平均租金 900 元/天，计 1.80 万元。

2.业务费（是指在项目实施过程中消耗的各种材料、低值易耗品等、发生的测试化验加工、燃料动力、出版文献、信息传播、知识产权事务、会议、差旅、国际合作与交流以及其他与项目实施直接相关的各项费用。编报时，对单笔大额支出、对外委托支出重点说明）

业务费共计 119.70 万元，占专项经费比例 45.17%，具体测算依据如下：

2.1 材料费：共计 74.33 万元，占专项经费总数的 28.05%，研究开发过程中消耗的各种原材料、辅助材料、低值易耗品等的采购及运输、装卸、整理等费用，具体测算依据如下：

（1）钢材等原材料费，14.28 万元。

本课题试制 5 种设备的自动仿形割台、3 种设备的高线速仿生圆盘式平茬装置、4 种设备的滚筒式低耗仿生切碎装置 12 套，每种 2 台，共计 24 台套，共计 12 吨，钢材占整机重量的 90%；从毛坯到工件，钢材损耗系数按 30%计，由于单件制作的损坏、损耗，风险备件系数按 30%，由于装置部件更改，其报废系数按 40%计，装置试制余量系数为 1.3 × 1.3 × 1.4 = 2.36；考虑试制装置的特殊性，单位钢材市场价格取不同区域价格，平均值为 0.56 万元/吨（吉林省、山东省 2022 年 8 月份市场报价）。试制钢材测算系数：0.9 × 2.36 = 2.124 钢材费计算方法如下：装置试制钢材费 = 所需钢材总重 × 试制钢材测算系数 × 钢材单价

钢材测算明细表

序号	名称	轮次	台数/轮次	单台重量/吨	测算系数	单价（万元）	总价（万元）
1	饲草种子采收自动仿形割台	2	1	0.8	2.124	0.56	1.90
2	青饲玉米收获自动仿形割台	2	1	0.8	2.124	0.56	1.90

3	构树收获自动仿形割台	2	1	0.7	2.124	0.56	1.67
4	柠条收获自动仿形割台	2	1	0.7	2.124	0.56	1.67
5	杂交狼尾草收获自动仿形割台	2	1	0.7	2.124	0.56	1.67
6	构树仿生平茬装置	2	1	0.3	2.124	0.56	0.71
7	柠条仿生平茬装置	2	1	0.3	2.124	0.56	0.71
8	杂交狼尾草仿生平茬装置	2	1	0.3	2.124	0.56	0.71
9	青饲玉米仿生切碎装置	2	1	0.5	2.124	0.56	1.19
10	构树仿生切碎装置	2	1	0.2	2.124	0.56	0.48
11	柠条仿生切碎装置	2	1	0.2	2.124	0.56	0.48
12	杂交狼尾草仿生切碎装置	2	1	0.5	2.124	0.56	1.19
合计							14.28

（2）液压元件材料费，31.20 万元。

本课题试制 5 种设备的自动仿形割台、3 种设备的高线速仿生圆盘式平茬装置、4 种设备的滚筒式低耗仿生切碎装置 12 套，每种 2 台，共计 24 台套，所需购买的液压缸、双向液压锁、液压变量泵、液压马达、液压集成阀、液压油箱等费用 31.20 万元，主要测算依据如下：

液压元件测算明细表

序号	名称	规格	单价（元）	数量	金额（万元）	用途
1	液压缸	活塞直径 30-200mm； 工作行程 0-800mm；	1500	24 个	3.60	用于割台、平茬装置、切碎装置的位置或角度调整

2	双向液压锁	公称压力 0-32MPa；流量 0-40L/min	500	48个	2.40	每个液压缸 2 个，用于液压缸工作位的锁定
3	液压变量泵	输出压力 0-32MPa；输出流量 0-40L/min	5000	12个	6.00	用于割台、 平茬装置、 切碎装置传动系统动力输入
4	液压马达	转速 0-3000r/min；扭矩 0-2000Nm	4000	12个	4.80	用于割台、平茬装置、切碎装置传动轴的转速控制
5	液压集成阀	公称压力 0-32MPa；流量 0-40L/min	8000	12套	9.60	用于控制割台、平茬装置、切碎装置液压系统液压油液压力、流量、方向
6	液压油箱	50-200L	4000	12套	4.80	用于割台、平茬装置、切碎装置液压系统供油、降温等
合计					31.20	/

（3）电子元件材料费，28.85 万元。

本课题试制的全程作业质量在线测控装置，包括：草籽收获作业质量在线测控、 捡拾打捆作业质量在线测控、饲草收获作业质量在线测控 3 种系统，试制 2 轮共 6 套，所需购买的电控单元通信模块、调制解调模块、信号模拟发生器、电控比例阀等费用 28.85 万元，主要测算依据如下：

电子元件测算明细表

序号	名称	单价（元）	数量	金额（万元）	用途
1	传感器模块	15000	6 套	9.00	专用传感器与模块：用于草籽收获、捡拾打捆、饲草收获过程中转
2	数据采集模块	9000	6 套	5.40	

3	调制解调模块	2000	6 套	1.20	速、压力、位移、角度、加速度、扭矩等信息测量与采集，信号调制、传输、处理，数据显示与参数调整等，用来监测工作通量、 割茬高度、切段长度、压缩密度、草捆长度等作业参数
4	信号模拟发生器	3000	6 套	1.80	
5	触控显示模块	6000	6 套	3.60	
6	电控单元通信模块	4000	6 套	2.40	
7	STM3240G 开发板	2000	6 个	1.20	专用耗材：用于草籽收获、捡拾打捆、 饲草收获过程中作业质量在线测控程序代码写入、命令输出
8	DSP 开发板	3000	6 个	1.80	
9	DSP 芯片	400	6 个	0.24	
10	JTAG 仿真器	1000	6 个	0.60	
11	USB 转 CAN 调试器	2000	6 个	1.20	
12	电阻本（含 200 种常用电阻）	1000	2 套	0.20	通用耗材：用于草籽收获、捡拾打捆、饲草收获过程中作业质量在线测控系统中 PLC 及设备保护
13	电容本（含 70 种常用电容）	1050	2 套	0.21	
合计				28.85	/

2.2 燃料动力费：共计 5.29 万元，占专项经费比例 2.00%，在项目实施过程中相关大型仪器设备、专用科学装置等运行发生的水、电、气、燃料消耗费用等，具体测算依据如下：

（1）开展田间试验所需柴油费 3.94 万元。按油价 8.2 元/升（中石油 2022 年 8 月份 - 35 号柴油报价）计算，研制的关键装置搭载在草籽收获机，捡拾打捆机，青饲玉米、构树、柠条等收获机进行田间试验，作业面积合计约为 600 亩，总需消耗柴油约 4800 L。

（2）室内试验台用电费 1.35 万元。按 1.00 元/kWh（吉林、山东两地均价）计算，试验台总功率为 30 kW，试验总时长约 450 h。

2.3 会议、差旅、国际合作与交流费：用于组织各类会议，研究人员差旅费、国际合作与交流费等，共计 29.48 万元，占专项经费比例 11.12%，具体测算依据如下：

（1）会议费 5.40 万元，组织召开课题启动、技术研讨、中期检查、课题绩效评价等会议 4 次，每次会议不超过 3 天，平均 15 人/次，会议标准参照

《中央和国家机关会议费管理办法》（财行[2016]214 号）四类会议执行，会议标准 300 元/(人·天)。

（2）差旅费 17.08 万元，项目执行期内安排用于课题调研、学术交流约 40 次，往返吉林、内蒙古、山东等常住地及开展课题研究需要的出差地和试验地开展装置试制、田间试验约 35 次，共计 90 人次。差旅费支出标准参照《中央和国家机关差旅费管理办法》（财行[2013]531 号）和关于印发《中央和国家机关工作人员赴地方差旅住宿费标准明细表》的通知（财行[2016]71 号）执行。差旅补助按照每人每天不超过 180 元；出差人员城市间交通费应当按该办法规定等级乘坐交通工具平均 1000 元/人次；出差人员应当按照所到城市住宿费标准限额内住宿，住宿费每人每天不超过 300 元。

（3）国际合作交流费 7.00 万元，参照《因公临时出国经费管理办法》（财行[2013]516 号）相关规定执行，主要参加德国汉诺威农机展览会、意大利博洛尼亚畜牧机械展等交流会 2 人次。

2.4 出版文献、信息传播、知识产权事务费：共计 10.60 万元，占专项经费比例 4.00%，具体测算依据如下：

（1）论文出版费：本课题计划发表 5 篇 SCI/EI 检索论文，每篇论文平均版面费按 4000 元计算，计 2.00 万元。

（2）知识产权事务费：本课题计划申请专利 7 项，每项申请费、代理费平均按 5000 元计算，计 3.50 万元。

（3）文献查新检索费：预计针对 4 种成果各进行 1 次文献查新检索，用于成果检验验收，每次 2500 元，计 1.00 万元。

（4）设计图纸资料费：主要包括割台自动仿形装置、高线速仿生圆盘式平茬装置、低耗仿生滚筒式切碎装置、全程作业质量在线测控系统（包括：草籽收获作业质量在线测控装置、捡拾打捆作业质量在线测控装置、饲草收获作业质量在线测控装置）等零部件设计导出 CAD 图纸打印专用纸张、彩色和黑白打印和装订等费用，计 2.60 万元。

（5）文献、资料购置费：用于购置国内外饲草机械智能控制关键技术及产品相关技术资料，专业期刊杂志和技术手册等，计 1.50 万元。

3.劳务费（是指在项目实施过程中支付给参与项目的研究生、博士后、访问学者以及项目聘用的研究人员、科研辅助人员、科研（财务）助理等的劳务性费用；支付给临时聘请的咨询专家的费用等。项目聘用人员由单位缴纳的社会保险补助、住房公积金等可纳入劳务费列支。）

劳务费共计 38.80 万元，占专项经费比例 14.64%，具体测算依据如下：

3.1 研究生、其他劳务人员等劳务性费用：共计 35.20 万元，占专项经费比例 13.28%，用于支付在项目实施过程中支付给参与项目的研究生、博士后、访问学者以及项目聘用的研究人员、科研辅助人员、科研（财务）助理等的劳务性费用。具体测算依据如下：

劳务标准博士生按每人 2250 元/月计算助研费用，投入时间 120 人月，硕士生按每人 800 元/月，投入时间 80 人月，计 33.40 万元；

聘用其他辅助技术人员劳务费，技术工人劳务费标准每人 3000 元/月，计 1.20 万元；

其他劳务人员，按临时雇工标准 200 元/天，计 0.60 万元。

3.2 专家咨询费：共计 3.60 万元，占专项经费比例 1.36%。用于支付给项目实施过程中临时聘请的咨询专家的费用。

开展课题启动、技术研讨、中期检查、课题绩效评价等会议及线上咨询，聘请技术专家、财务专家等进行技术方案、经费支出、课题成果评价等咨询活动，专家咨询费标准参照《中央财政科研项目专家咨询费管理办法》（财科教〔2017〕128 号）执行，按照 1500-2400 元/人天发放，约 20 人次，共计 3.60 万元。

（二） 间接经费： 61.00 万元。

间接经费按照财政部科技部《国家重点研发计划资金管理办法》（财教〔2021〕178 号）规定进行测算，主要用于承担单位为项目研究提供的房屋占用，日常水、电、气、暖等消耗，有关管理费用的补助支出，以及激励科研人员的绩效支出等，共计 61.00 万元。

二、其他来源资金

对其他来源资金主要用途、支出预算做简要说明。

1.设备费：共计 8.00 万元，占自筹经费比例 2.00%。

1.1 购置设备费：8.00 万元。

（1）高速运动摄像机 1 台，用于青饲玉米、构树、柠条等收获机切割茬口、切断长度等图像信息的采集，计 6.00 万元。

（2）图形工作站 1 台，用于割台、滚筒、切碎刀等计算机软件仿真分析，对工作参数进行优化设计，计 2.00 万元。

2. 业务费：共计 392.00 万元，占自筹经费比例 98.00%。

2.1 材料费：共计 167.00 万元。

购置 ARM11 开发测试板、FPGA 芯片、STM32CM4 处理器、光电编码器、AVX 钽电容、PESD5VOS1BAESD 保护二极管、贴片电感、高亮发光 LED、场效应管 MOS、数字示波器、万用表、拉线、手柄、仪表、按键、开关、继电器、限位器、光敏传感器、专用屏蔽信号线、隔离稳压电源等电子元件耗材，计 48.00 万元；购置液压油缸、液压油管、液压阀体、液压泵、及连接接头等液压系统耗材，计 35.00 万元；购置型材、护罩、齿轮箱、传动张紧机构等其他辅助部件，计 29.00 万元；购置液压油、齿轮油、油漆、防冻液、油漆等装置用油液，计 25.00 万元；购置轴承、密封件、螺栓、螺母、垫圈、挡圈、卡簧、皮带、链条等标准件，计 18.50 万元；购买试验用饲草茎秆、低值易耗品等，计 11.50 万元。

2.2 测试化验加工费：共计 180.00 万元。

用于高线速仿生圆盘式平茬装置、高速仿生切割部件、切割装置支架、切割装置传动机构，计 29.00 万元；

割台自动仿形装置、割台传动机构，计 14.00 万元；

低耗仿生滚筒式切碎装置、切抛部件、切抛传动机构，计 22.00 万元；

草籽收获作业质量在线测控系统、视觉识别系统，计 28.00 万元；

捡拾打捆作业质量在线测控系统、草捆长度检测系统、打捆故障检测系统，计 25.00 万元；

饲草收获装备作业质量在线测控系统、切断长度检测系统、留茬高度检测系统、茬口断面检测系统等测试费用，计 34.00 万元；

监控装置固定机构、其它辅助机构等加工费用，计 20.00 万元；

关键装置的第三方检测费用，计 8.00 万元。

2.3 其他支出：共计 45.00 万元。

主要用于青饲玉米、苜蓿、天然草原等作物的智能机械化作业装备在内蒙、吉林、山东等地田间试验土地、草场租赁及补偿，装置异地运输等。

其他来源资金承诺书

山东巨明机械有限公司（单位全称），为主要饲草饲料全程智能化生产共性关键技术研究与应用课题，提供 400 万元的资金，资金来源为 2.单位自筹资金（1. 地方财政资金 2. 单位自筹资金 3. 其他渠道获得资金）。

资金主要用于：设备费和业务费。

特此证明！

出资单位（公章）：山东巨明机械有限公司

2022 年 8 月 19 日

十一、相关附件

1. 乙方与参加单位有关协议（须加盖乙方与参加单位公章、法人签字签章；协议文件须扫描上传。如无参加单位，则不填）；

序号	单位名称	
1	课题承担单位	吉林大学
2	课题参与单位	华南农业大学
3		山东巨明机械有限公司

（1）吉林大学

2022 年度国家重点研发计划

“主要饲草饲料生产全程智能化作业装备创制与应用”项目

“主要饲草饲料全程智能化生产共性关键技术研究与应用”

课题联合实施协议

甲方（课题牵头单位）：吉林大学

乙方（课题参加单位）：吉林大学

甲方为课题牵头单位，乙方参与甲方牵头的国家重点研发计划“工厂化农业关键技术与智能农机装备”重点专项“主要饲草饲料生产全程智能化作业装备创制与应用”项目中的“主要饲草饲料全程智能化生产共性关键技术研究与应用”课题。为保证项目的顺利实施，经友好协商，达成如下合作协议：

一、共同条款

1.甲乙双方承诺遵守国家重点研发计划暂行管理办法、国家重点研发计划资金管理办法及其他相关法规制度的要求，严格执行国家重点研发计划项目申报书所承诺任务目标及考核内容。

二、任务分工

1.甲方为课题牵头单位，整体负责课题的组织实施，负责子课题研究进展与实施质量监督、总结上报及整体协调。

2.乙方作为“主要饲草饲料全程智能化生产作业质量测控关键技术研究与应用”子课题承担单位，负责子课题组织实施，对该子课题完成效果负责，按年度计划开展研究工作，保质保量地完成任务，按期完成考核指标，并按时提交相关材料，配合甲方进行课题验收。

三、经费分配与使用

1.根据课题总经费及乙方在课题中所承担的工作任务，乙方分配

总经费 170 万元，其中中央财政专项资金 170 万元，自筹经费 0 万元。

2.乙方保证专项经费的使用符合国家各项法规及国家重点研发计划资金管理办法，并按照与甲方约定的各个科目预算额度进行支出，保证资金使用规范，专款专用。

四、考核指标

课题完成时，乙方应达成如下目标：

1.突破低扰损平茬、减阻降耗仿生切碎、全程作业质量在线测控关键技术 3 项及以上；

2.研制高线速仿生圆盘式平茬装置 1 项，切割茬口破损率≤5%；

3.研制低耗仿生切碎装置 1 项；

4.研制全程作业质量在线测控系统 1 套，监测准确率≥90%；

5.申请发明专利 3 件及以上、实用新型专利 2 件及以上；

6.发表 SCI/EI 论文 3 篇及以上；

7.申请软件著作权 2 件及以上；

8.培养研究生 2 人及以上。

五、知识产权保密约定

1.项目执行过程中，甲、乙双方各自取得的研究成果和相关的知识产权归各单位自己所有，为甲乙双方共同研究形成的成果（专利、论文等）由双方另签协议约定知识产权共享；甲方有权因非商业目的（如：以政治性会议、报告、文件、统计资料等）使用乙方项目信息；乙方在课题实施过程中产生的专利、论文、著作、新技术、新工艺等登记申请和推广应用情况等，均须标注本项目资助编号。

2.乙方研究任务完成后，根据甲方要求提交技术成果，承诺对本单位提供的各种材料及其内容的真实性负责，承诺产生的所有科学数据无条件、按期递交到科技部指定的平台，在专项约定的条件下对专项各承担单位，乃至今后面向所有的科技工作者和公众开放共享。

六、违约责任

本协议签订后，双方共同遵守、不能全面、及时履行合同约定的即构成违约，守约方有权根据实际情况要求违约方继续履行合同或解除合同，同时违约方负责赔偿对方因此造成的全部损失。

七、其他

1.此协议为涉及的其他事项，由双方协商解决，并形成书面文件作为本协议的补充，补充协议与本协议具有同等法律效力。

2.本协议共叁页，一式肆份，甲乙双方各持贰份，自签字盖章之日起生效，课题结题验收后自行终止。

甲方（课题牵头单位）：吉林大学

法人（签字）：

课题负责人（签字）：袁洪亮

2022年11月16日

乙方（课题参加单位）：吉林大学

法人（签字）：

子课题负责人（签字）：袁洪亮

2022年11月16日

（2）华南农业大学

2022 年度国家重点研发计划

“主要饲草饲料生产全程智能化作业装备创制与应用”项目

“主要饲草饲料全程智能化生产共性关键技术研究与应用”

课题联合实施协议

甲方（课题牵头单位）：吉林大学

乙方（课题参加单位）：华南农业大学

甲方为课题牵头单位，乙方参与甲方牵头的国家重点研发计划“工厂化农业关键技术与智能农机装备”重点专项“主要饲草饲料生产全程智能化作业装备创制与应用”项目中的“主要饲草饲料全程智能化生产共性关键技术研究与应用”课题。为保证项目的顺利实施，经友好协商，达成如下合作协议：

一、共同条款

1.甲乙双方承诺遵守国家重点研发计划暂行管理办法、国家重点研发计划资金管理办法及其他相关法规制度的要求，严格执行国家重点研发计划项目申报书所承诺任务目标及考核内容。

二、任务分工

1.甲方为课题牵头单位，整体负责课题的组织实施，负责子课题研究进展与实施质量监督、总结上报及整体协调。

2.乙方作为“主要饲草饲料全程智能化生产作业参数测控关键技术研究与应用”子课题承担单位，负责子课题组织实施，对该子课题完成效果负责，按年度计划开展研究工作，保质保量地完成任务，按期完成考核指标，并按时提交相关材料，配合甲方进行课题验收。

三、经费分配与使用

1.根据课题总经费及乙方在课题中所承担的工作任务，乙方分配

总经费 65 万元，其中中央财政专项资金 65 万元，自筹经费 0 万元。

2.乙方保证专项经费的使用符合国家各项法规及国家重点研发计划资金管理办法，并按照与甲方约定的各个科目预算额度进行支出，保证资金使用规范，专款专用。

四、考核指标

课题完成时，乙方应达成如下目标：

1.突破随地自动仿形关键技术 1 项及以上；

2.研制割台自动仿形装置 1 项及以上，留茬高度误差小于 15mm；

3.申请发明专利 1 件及以上、实用新型专利 1 件及以上；

4.发表 SCI/EI 论文 2 篇及以上；

5.申请软件著作权 2 件及以上；

6.培养研究生 4 人及以上。

五、知识产权保密约定

1.项目执行过程中，甲、乙双方各自取得的研究成果和相关的知识产权归各单位自己所有，为甲乙双方共同研究形成的成果（专利、论文等）由双方另签协议约定知识产权共享；甲方有权因非商业目的（如：以政治性会议、报告、文件、统计资料等）使用乙方项目信息；乙方在课题实施过程中产生的专利、论文、著作、新技术、新工艺等登记申请和推广应用情况等，均须标注本项目资助编号。

2.乙方研究任务完成后，根据甲方要求提交技术成果，承诺对本单位提供的各种材料及其内容的真实性负责，承诺产生的所有科学数据无条件、按期递交到科技部指定的平台，在专项约定的条件下对专项各承担单位，乃至今后面向所有的科技工作者和公众开放共享。

六、违约责任

本协议签订后，双方共同遵守、不能全面、及时履行合同约定的即构成违约，守约方有权根据实际情况要求违约方继续履行合同或解

除合同，同时违约方负责赔偿对方因此造成的全部损失。

七、其他

1.此协议为涉及的其他事项，由双方协商解决，并形成书面文件作为本协议的补充，补充协议与本协议具有同等法律效力。

2.本协议共叁页，一式肆份，甲乙双方各持贰份，自签字盖章之日起生效，课题结题验收后自行终止。

甲方（课题牵头单位）：吉林大学

法人（签字）：

课题负责人（签字）：袁洪文

2022 年 11 月 17 日

乙方（课题参加单位）：华南农业大学

法人（签字）：

子课题负责人（签字）：

2022 年 11 月 16 日

（3）山东巨明机械有限公司

2022 年度国家重点研发计划

“主要饲草饲料生产全程智能化作业装备创制与应用”项目

“主要饲草饲料全程智能化生产共性关键技术研究与应用”

课题联合实施协议

甲方（课题牵头单位）：吉林大学

乙方（课题参加单位）：山东巨明机械有限公司

甲方为课题牵头单位，乙方参与甲方牵头的国家重点研发计划“工厂化农业关键技术与智能农机装备”重点专项“主要饲草饲料生产全程智能化作业装备创制与应用”项目中的“主要饲草饲料全程智能化生产共性关键技术研究与应用”课题。为保证项目的顺利实施，经友好协商，达成如下合作协议：

一、共同条款

1. 甲乙双方承诺遵守国家重点研发计划暂行管理办法、国家重点研发计划资金管理办法及其他相关法规制度的要求，严格执行国家重点研发计划项目申报书所承诺任务目标及考核内容。

二、任务分工

1. 甲方为课题牵头单位，整体负责课题的组织实施，负责子课题研究进展与实施质量监督、总结上报及整体协调。

2. 乙方作为“主要饲草饲料全程智能化生产共性关键装置试验示范”子课题承担单位，负责子课题组织实施，对该子课题完成效果负责，按年度计划开展研究工作，保质保量地完成任务，按期完成考核指标，并按时提交相关材料，配合甲方进行课题验收。

三、经费分配与使用

1. 根据课题总经费及乙方在课题中所承担的工作任务，乙方分配

总经费 430 万元，其中中央财政专项资金 30 万元，自筹经费 400 万元。

2. 乙方保证专项经费的使用符合国家各项法规及国家重点研发计划资金管理办法，并按照与甲方约定的各个科目预算额度进行支出，保证资金使用规范，专款专用。

四、考核指标

课题完成时，乙方应达成如下目标：

1. 培养技术骨干 3 人及以上。

五、知识产权机保密约定

1. 项目执行过程中，甲、乙双方各自取得的研究成果和相关的知识产权归各单位自己所有，为甲乙双方共同研究形成的成果（专利、论文等）由双方另签协议约定知识产权共享；甲方有权因非商业目的（如：以政治性会议、报告、文件、统计资料等）使用乙方项目信息；乙方在课题实施过程中产生的专利、论文、著作、新技术、新工艺等登记申请和推广应用情况等，均须标注本项目资助编号。

2. 乙方研究任务完成后，根据甲方要求提交技术成果，承诺对本单位提供的各种材料及其内容的真实性负责，承诺产生的所有科学数据无条件、按期递交到科技部指定的平台，在专项约定的条件下对专项各承担单位，乃至今后面向所有的科技工作者和公众开放共享。

六、违约责任

本协议签订后，双方共同遵守、不能全面、及时履行合同约定的即构成违约，守约方有权根据实际情况要求违约方继续履行合同或解除合同，同时违约方负责赔偿对方因此造成的全部损失。

七、其他

1. 此协议为涉及的其他事项，由双方协商解决，并形成书面文件作为本协议的补充，补充协议与本协议具有同等法律效力。

2. 本协议共叁页，一式肆份，甲乙双方各持贰份，自签字盖章之日起生效，课题结题验收后自行终止。

甲方（课题牵头单位）：吉林大学

法人（签字）：

课题负责人（签字）：袁洪久

2022年 11月 7日

乙方（课题参加单位）：山东巨明机械有限公司

法人（签字）：

子课题负责人（签字）：

2022年 11月16日

任务书签署

甲乙双方根据《国务院印发关于深化中央财政科技计划（专项、基金）管理改革方案的通知》（国发〔2014〕64号）、《国务院关于优化科研管理提升科研绩效若干措施的通知》（国发〔2018〕25号）、《国务院办公厅关于改革完善中央财政科研经费管理的若干意见》（国办发〔2021〕32号）、《科技部 财政部关于印发<国家重点研发计划管理暂行办法>的通知》（国科发资〔2017〕152号）、《财政部 科技部关于印发<国家重点研发计划资金管理办法>的通知》（财教〔2021〕178号）、《科学技术活动违规行为处理暂行规定》（科学技术部令第19号）、《科技部财政部关于印发<中央财政科技计划（专项、基金等）监督工作暂行规定>的通知》（国科发政〔2015〕471号）、《科技部 自然科学基金委关于进一步压实国家科技计划（专项、基金等）任务承担单位科研作风学风和科研诚信主体责任的通知》(国科发监〔2020〕203号)等有关文件规定，以及有关法律、政策和管理要求，依据项目立项通知，签署本任务书。

同时，本单位和项目负责人**郑重承诺**：对本项目所有成果产出（包括但不限于新产品、新技术、标准、论文、专利等）的真实性、与项目的关联性等负责，将按要求落实科研作风学风和科研诚信主体责任；项目经费全部用于与本项目研究工作相关的支出，不截留、挪用、侵占，不用于与科学研究无关的支出；严格按照政府采购和保密法律法规规定开展政府采购活动，规范信息公开工作；接受并积极配合相关部门的监督检查。如有违反，本单位和项目负责人以及相关成果产出者愿接受项目管理专业机构和相关部门做出的各项处理决定，包括但不限于终止项目执行、追回项目（课题）经费，取消一定期限国家科技计划项目申报资格，记入科研诚信严重失信行为数据库以及主要负责人接受相应党纪政纪处理等。

课题牵头承担单位（甲方）：

法定代表人签字（签章）：

（公章）

年　月　日

课题负责人签字（签章）：

2023年2月10日

子课题承担单位（乙方）：

法定代表人签字（签章）：

（公章）

年　月　日

子课题负责人签字（签章）：

2023年1月10日

项目批准号	31571561
申请代码	C130104
归口管理部门	
依托单位代码	51064208A0499-0932

国家自然科学基金委员会

资助项目计划书

资助类别：面上项目

亚类说明：

附注说明：常规面上项目

项目名称：基于仿生嗅觉和保鲜环境的荔枝货架多源信息反演机理研究

直接费用：61万元　　间接费用：10.904万元

项目资金：71.904万元　　执行年限：2016.01-2019.12

负 责 人：陆华忠

通讯地址：广州市五山路华南农业大学工程学院

邮政编码：510642　　电　　话：02085287438

电子邮件：huazlu@scau.edu.cn

依托单位：华南农业大学

联 系 人：全锋　　电　　话：020-85280070

填表日期：2015年09月02日

国家自然科学基金委员会制

Version：1.011.411

国家自然科学基金委员会资助项目计划书填报说明

一、项目负责人收到《关于国家自然科学基金资助项目批准及有关事项的通知》（以下简称《批准通知》）后，请认真阅读本填报说明，参照国家自然科学基金相关项目管理办法及《国家自然科学基金资助项目资金管理办法》（请查阅国家自然科学基金委员会官方网站首页“政策法规”-“管理办法”栏目），按《批准通知》的要求认真填写和提交《国家自然科学基金委员会资助项目计划书》（以下简称《计划书》）。

二、填写《计划书》时要求科学严谨、实事求是、表述清晰、准确。《计划书》经国家自然科学基金委员会相关项目管理部门审核批准后，将作为项目研究计划执行和检查、验收的依据。

三、《计划书》各部分填写要求如下：

（一）简表：由系统自动生成。

（二）摘要及关键词：各类获资助项目都必须填写中、英文摘要及关键词。

（三）项目组主要成员：计划书中列出姓名的项目组主要成员由系统自动生成，与申请书原成员保持一致，不可随意调整。如果批准通知中“项目评审意见及修改意见表”中“对研究方案的修改意见”栏目有调整项目组成员相关要求的，待项目开始执行后，按照项目成员变更程序另行办理。

（四）资金预算表：按批准资助的直接费用填报资金预算表和预算说明书，其中的劳务费、专家咨询费金额不应高于申请书中相应金额；间接费用及项目总经费由系统自动生成。国家重大科研仪器研制项目还应按照预算评审后批复的直接费用各科目金额填报资金预算表、预算说明书及相应的预算明细表。

（五）正文：

1. 面上项目、青年科学基金项目、地区科学基金项目：如果《批准通知》中没有修改要求的，只需选择“研究内容和研究目标按照申请书执行”即可；如果《批准通知》中“项目评审意见及修改意见表”中“对研究方案的修改意见”栏目明确要求调整研究期限和研究内容等的，须选择“根据研究方案修改意见更改”并填报相关修改内容。

2. 重点项目、重点国际（地区）合作研究项目、重大项目、国家重大科研仪器研制项目：须选择“根据研究方案修改意见更改”，根据《批准通知》的要求填写研究（研制）内容，不得自行降低、更改研究目标（或仪器研制的技术性能与主要技术指标以及验收技术指标）或缩减研究（研制）内容。此外，还要突出以下几点：

（1）研究的难点和在实施过程中可能遇到的问题（或仪器研制风险），拟采用的研究（研制）方案和技术路线；

（2）项目主要参与者分工，合作研究单位之间的关系与分工，重大项目还需说明课题之间的关联；

（3）详细的年度研究（研制）计划。

3. 国家杰出青年科学基金、优秀青年科学基金和海外及港澳学者合作研究基金项目：须选择“根据研究方案修改意见更改”，按下列提纲撰写：

（1）研究方向；

（2）结合国内外研究现状，说明研究工作的学术思想和科学意义（限两个页面）；

（3）研究内容、研究方案及预期目标（限两个页面）；

（4）年度研究计划；

（5）研究队伍的组成情况。

4. 对于其他类型项目，参照面上项目的方式进行选择和填写。

简表

申请者信息	姓　　名	陆华忠	性 别	男	出生年月	民 族	汉族
	学　　位	博士			职称	教授	
	电　　话	02085287438			电子邮件	huazlu@scau.edu.cn	
	传　　真	02085282693			个人网页		
	工作单位	华南农业大学					
	所在院系所	工程学院					
依托单位信息	名　　称	华南农业大学				代码	51064208A0499
	联 系 人	全锋			电子邮件	kycjhk@scau.edu.cn	
	电　　话	020-85280070			网站地址	http://web.scau.edu.cn/kjc/	
合作单位信息	单　位　名　称						代　码
项目基本信息	项　目　名　称	基于仿生嗅觉和保鲜环境的荔枝货架多源信息反演机理研究					
	资　助　类　别	面上项目			亚类说明		
	附　注　说　明	常规面上项目					
	申　请　代　码	C130104:农业信息学			F011305:智能信息处理		
	基　地　类　别						
	执　行　年　限	2016.01-2019.12					
	直　接　费　用	61万元			间接费用	10.904万元	
	项　目　资　金	71.904万元					

项目摘要

中文摘要(500字以内)：

荔枝易质变，具有“一日色变，二日香变，三日色香味俱变”的性质。加强荔枝货架信息监测，有利于实现对荔枝内部品质的全程监控，为荔枝的贮藏保鲜、分级、销售及产后深加工提供科学指导，从而推进荔枝采后处理的系统化、精致化，至今尚未圆满解决。针对传统的荔枝货架信息检测方法效率低且成本高、实际检测过程中荔枝果实相互遮挡等问题，提出了基于仿生嗅觉技术与保鲜环境的荔枝多源货架信息检测方法进行研究，结合智能化理论和技术构造具有较强分类能力和容错能力的荔枝货架信息反演模型。项目拟构建仿生嗅觉信息以及保鲜环境信息的快速获取平台，重点研究荔枝气体挥发物成分以及褐变程度随时间序列的发生、发展规律，揭示荔枝气体挥发物变化与褐变度之间的关系特性，结合其他保鲜环境信息传感器，寻找适合于描述荔货架信息的反演模型，提出一种基于仿生电子鼻和保鲜环境传感器的荔枝货架期智能监测的新方法，为荔枝采后的货架管理提供理论和实践上的依据。

关键词：果实挥发物；仿生嗅觉；环境信息；荔枝；在线监测

Abstract(limited to 4000 words)：

Litchi share the property of easily deteriorating, as the saying goes “One day after plucking changecolor, the second day fragrance and the third day flavor”. Strengthening the monitoring shelf lifeinformation of litchi is benefit for realizing whole process monitoring internal quality of litchi, which provides scientific guidance for storage, classification, sell and deep processing, and contributes the promoting of systematization and exquisite on litchi postharvest treatment. Accroding to the problems that conventional detection methods of litchi shelf informtion are low-efficiency and high-cost and litchi fruits shelter each other during detection, a inverse model of litchi shelf information is put forward which combines intelligent theory and structure technology with higher classification ability and fault-tolerance resilience, researching by bionic olfaction technology and detection of multi-source information of litchi in storage environment. This project is aimed at building a fast acquisition platform of bionic olfaction and storage environment information, finding out the suitable retrieval model describing litchi shelf life information, proposing a novel method of intelligently monitoring litchi shelf life based on bionic electronic nose and storage environment sensor, providing theoretical and practical basis on litchi shelf manager after harvest.

Keywords: Fruit volatile; Bionic olfaction; Environment information; Litchi; Real-time monitoring

项目组主要成员

编号	姓名	出生年月	性别	职称	学位	单位名称	电话	证件号码	项目分工	每年工作时间（月）
1	陆华忠	[illegible]	男	教授	博士	华南农业大学	02085287438	[illegible]	项目负责人	5
2	陈建业	[illegible]	男	教授	博士	华南农业大学	02085287438	[illegible]	方案评估、组织协调	5
3	吕恩利	[illegible]	男	教授	博士	华南农业大学	02085287438	[illegible]	保鲜环境信息的获取	7
4	吴慕春	[illegible]	女	副教授	硕士	华南农业大学	02085287438	[illegible]	保鲜参数获取平台的搭建	8
5	吕盛坪	[illegible]	男	讲师	博士	华南农业大学	02085287438	[illegible]	数据挖掘	8
6	徐赛	[illegible]	男	博士生	学士	华南农业大学	02085287438	[illegible]	仿生嗅觉信息获取	9
7	郭嘉明	[illegible]	男	博士生	硕士	华南农业大学	1[illegible]	[illegible]	特征分析与建模	9
8	李亚慧	[illegible]	女	硕士生	学士	华南农业大学	02085287438	[illegible]	荔枝褐变信息的获取	9
9	赵俊宏	[illegible]	男	硕士生	学士	华南农业大学	02085287438	[illegible]	算法实现及软件开发	9

总人数	高级	中级	初级	博士后	博士生	硕士生
9	4	1	0	0	2	2

国家自然科学基金项目资金预算表（定额补助）

项目名称： 基于仿生嗅觉和保鲜环境的荔枝货架多源信息反演机理研究

项目负责人：陆华忠

金额单位：万元

序号	科目名称	金额	备注
	(1)	(2)	(3)
1	一、 项目资金支出	71.9040	/
2	（一）直接费用	61.0000	没有5 万元以上大型仪器费用
3	1、 设备费	6.4800	用于传感器购置、软件试制、试验装置改造
4	(1)设备购置费	6.4800	用于购置保鲜环境信息传感器
5	(2)设备试制费	0.0000	无
6	(3)设备改造与租赁费	0.0000	无
7	2、 材料费	22.5700	原材料、试剂及试验平台构建所需耗材等费用
8	3、 测试化验加工费	2.8000	荔枝品质测定、平台加工
9	4、 燃料动力费	0.0000	气调保鲜运输车及专用科学装置的能耗
10	5、 差旅费	5.0500	省外调研、国内学术会议及市内交通相关费用
11	6、 会议费	2.5000	举办2次学术研讨会相关费用
12	7、 国际合作与交流费	9.0000	参加国际学术研讨会以及邀请国外专家费用
13	8、 出版/文献/信息传播/知识产权事务费	4.9000	查新、论文版面、专利申请、复印资料等费用
14	9、 劳务费	7.7000	用于参加项目研究生的补助发放
15	10、 专家咨询费	0.0000	无
16	11、 其他支出	0.0000	无
17	（二）间接费用	10.9040	用于实验室日常消费及绩效支出
18	其中：绩效支出	2.7259	发表论文、专利奖励及田间试验补贴等
19	二、 自筹资金	0.0000	无

预算说明书

（请对各项支出的主要用途和测算理由及合作研究外拨资金等内容进行详细说明，可根据需要另加附页。）

1. 仪器设备费用，共计 6.48 万元，其中：

（1）购置，6.48 万元

温度传感器，0.36 万元/只，6 只，共 2.16 万元

湿度传感器，0.31 万元/只，6 只，共 1.86 万元

氧气浓度传感器，0.28 万元/只，6 只，共 1.68 万元

二氧化碳浓度传感器，0.13 万元/只，6 只，共 0.78 万元

（2）设备试制费，0 万元

（3）设备改造与租赁费，0 万元

2. 材料费，22.57 万元

（1）原材料/试剂购置费，20.41 万元

荔枝果园鲜果购置费，2600kg，10 元/kg，共 2.6 万

塑料筐，400 个，30 元/个，共 1.2 万

包装箱，500 个，20 元/个，共 1 万

防雾膜，1400 个，1 元/个，共 1.4 万

气体购置费 8.16 万（每次试验用液氮 6 瓶，干冰 6 瓶，液氮 400 元/瓶，干冰 450 元/瓶，4 次/年，4 年共消耗气体费用 8.16 万）

常规化学分析试剂以及用于气相色谱仪的有机溶剂，共 3.55 万元

试验易耗材料（玻璃器皿、试管、保鲜袋、硅胶手套保鲜膜等）一批，共 2.5 万元

（2）其他材料费，2.16 万

机械标准件（工程塑胶、PVC 管、橡胶件等）一批，共 0.2 万元

数据通信模块，3000 元/个，2 个，共 0.6 万元

嵌入式开发板，3500 元/个，2 个，共 0.7 万元

电子元器件（电阻、电容、功率芯片、电源芯片、电缆等）一批，共 0.3 万元

PCB 电路板，200 元/块，6 块，共 0.12 万元

锂电池，600/块，2 块，共 0.24 万元

3. 测试/化验/加工费，2.8 万元

对不同保鲜环境下贮藏的荔枝品质测定分析，每年约检测 100 个样本，4 年共 400 个样本，50 元/样本，共 2 万元

荔枝货架多源信息获取平台的加工费用，约 0.8 万元

4. 差旅费，5.05 万元

省外业务调研，食宿每天 220 元/人，交通费 680 元/人次，2 天/次，8 人次/年，4 年计 2.88 万

参加国内学术会议，食宿每天 220 元/人，交通费 0.1 万元/人次，2 天/次，4 人次/年，4 年共 2 次计 0.97 万

市内交通，交通费 0.01 万元/人次，30 人次/年，4 年计 1.2 万

5. 会议费，2.5 万元

拟举办 2 次学术研讨会，会议规模为 25 人，会期 1 天，按每天 500 元/人/天的标准开支，1.25 万元/次，2 次预算支出 2.5 万

6. 国际合作与交流费，9 万元

参加美国农业与生物工程师学会（ASABE）学术年会，按 2 人计算，在国外停留 10 天，往返机票约 1.5 万元×2 人=3 万元；国外每人每天生活费按 1000 元计（住宿费 80 美元，伙食费 30 美元，公杂费等 20 美元），1000 元×2 人×10 天=2 万元，共计 5 万元

邀请国外专家来华，国外专家 2 人，来华合作交流 10 天，往返机票约 1.5 万元×2 人=3 万元；每天在华生活费按 750 元计（住宿费 220 元，伙食费 200 元，市内交通费 80 元），则： 2 人×10 天×500 元=1 万元，共计 4 万元

7. 出版/文献/信息传播/知识产权事务费，4.9 万元

专利申请费用：专利 2 个，5000 元/个×2 个=1 万元

核心期刊论文版面费：2000 元/篇×8 篇=1.6 万元

查新费：国内、外查新 2000 元/个，查新 2 次，2 次×2000 元/个=0.4 万元

项目技术总结材料复印装订、专题汇报宣传材料，共 1.5 万元

购买研究任务专用书籍、文献资料检索购买等约 0.4 万元

8、劳务费，7.7 万元

研究任务需要 4 名研究生，其中博士生 2 名，硕士生 2 名；每月补助标准博士生 600 元/人、硕士生 470 元/人，每人每年工作 9 个月，4 年劳务费支出为 7.7 万

9. 绩效支出，2.7259 万元

论文奖励 1000 元/篇，8 篇共 0.8 万元

专利奖励 3000 元/件，3 件共 0.9 万元

田间试验补贴及获奖补贴，共 1.0259 万元

项目负责人签字：

科研部门公章：

财务部门公章：

国家自然科学基金资助项目签批审核表

我接受国家自然科学基金的资助，将按照申请书、项目批准意见和计划书负责实施本项目（批准号：31571561），严格遵守国家自然科学基金委员会关于资助项目管理、财务等各项规定，切实保证研究工作时间，认真开展研究工作，按时报送有关材料，及时报告重大情况变动，对资助项目发表的论著和取得的研究成果按规定进行标注。 项目负责人（签章）： 2015年 9月 10日	我单位同意承担上述国家自然科学基金项目，将保证项目负责人及其研究队伍的稳定和研究项目实施所需的条件，严格遵守国家自然科学基金委员会有关资助项目管理、财务等各项规定，并督促实施。 依托单位（公章） 2015年 9月 21日

本栏目由基金委填写

科学处审查意见：

请按计划书内容执行

建议年度拨款计划（本栏目为自动生成，单位：万元）：

年度	总额	第一年	第二年	第三年	第四年	第五年
金额						

负责人（签章）：
2015年 10月 21日

科学部审查意见：

同意科学处审查意见

负责人（签章）：冯雪莲
2015年 10月 27日

本栏目主要用于重大项目等

相关局室审核意见：

负责人（签章）：
年 月 日

委领导审批意见：

委领导（签章）：
年 月 日

51064201-0932

国家自然科学基金

资助项目准予结题通知

陆华忠　同志：

您承担的国家自然科学基金项目：(基于仿生嗅觉和保鲜环境的荔枝货架多源信息反演机理研究)，批准号：(31571561)按有关规定已审核完毕，准予结题。

与本项目资助有关的后续成果，请您继续及时报送。

祝您在研究工作中取得更好的成绩！

国家自然科学基金委员会

生命科学部

2020年5月6日

附件1. 项目合同书

受理编号：c1530550100087

项目编号：2015A020209161

文件编号：粤科规财字[2015]150号

广东省省级科技计划项目

合同书

项目名称：基于混合群体智能的树状灌溉管网优化技术研究

计划类别：农村科技领域

项目起止时间：2016-01-01　至　2017-12-31

管理单位(甲方)：广东省科学技术厅

承担单位(乙方)：华南农业大学

乙方主管部门(丙方)：华南农业大学

通讯地址：广东省广州市天河区五山路483号

邮政编码：510642　　单位电话：020-38632819

项目负责人：吕石磊　　联系电话：020-85282269

项目联系人：吕石磊　　联系电话：[redacted]

广东省科学技术厅

二〇一四年制

一、项目实施内容

1.主要研究内容

（1）优化模型研究：

①以树状灌溉管网为研究对象，以管网总长度最小为目标，结合对管网总投资、灌溉水资源受限、地形等考虑，构建管网部署规划模型。

②以管网水资源利用率最大为目标，结合对管网总投资的考虑，以管道压力和流速为约束条件，构建管径优化模型。

（2）混合群体智能优化算法的设计：

①基于性能优势互补的角度，以PSO（Particle Swarm Optimization algorithm）算法和PGSA（Plant Growth Simulation Algorithm）算法为研究对象，进行混合策略研究。

②分析冬虫夏草生长机制，并将其用于指导模拟动物种群行为的PSO算法和模拟植物生长行为的PGSA算法间的混合过程。

（3）基于智能计算的树状灌溉管网优化机制研究：

搭建灌溉管网实验平台，将提出的混合优化算法用于求解优化模型，分别从仿真和实例的角度来验证优化机制的可行性。最后形成应用示范系统，并进行试点推广。

2.拟解决的关键问题及技术路线

关键问题：

（1）确定多个优化目标间的均衡策略

针对管网总投资与管网长度及管径参数等不同（矛盾）目标，考虑引入Pareto非支配解评价方法，避免多目标分别优化寻优无法获取 Pareto 最优解的缺点。

（2）确定PSO算法和PGSA算法间的混合策略

考虑将冬虫夏草生长过程划分为“类动物生长过程”（冬虫）和“类植物生长过程”（夏草），通过设置子囊孢子侵染率和子座存活率对两类物种的生长过程进行关联，并且将其用于PSO算法和PGSA算法的混合过程。

技术路线：

研究工作分两部分：管网优化模型研究和混合智能优化算法设计。在项目初期，两部分研究独立进行；在项目中期，将混合算法用于求解优化模型，并根据结果调整参数和模型；在项目后期，搭建实验平台，以实例验证优化机制的有效性。最后，实现通用、可靠的管网优化设计示范软件系统。项目组在节水灌溉技术、 部署规划、群体智能优化算法应用等方面有一定的技术优势， 已经完成了多个相关项目，这些技术的积累大幅度的降低了实现项目研发目标的技术风险。

3.创新点

（1）本项目首次将混合群体智能优化算法用于解决管网优化问题，并引入Pareto非支配解评价方法来均衡多目标优化方案。

（2）提出基于冬虫夏草生长机制的群体智能混合策略，利用两种不同物种的生长过程特征来设计混合群体智能优化算法。

二、项目考核指标

1.项目完成后提供的研究开发成果及形式（须明确产品、专利、版权、标准等成果的类型及数量）

成果形式		成果数量	成果形式		成果数量
发明专利	申请	1	引进人才(人)		
	授权		培养人才(人)		2
实用新型专利	申请		科技人才奖励(人)		
	授权		技术标准制定	牵头(个)	
外观设计专利	申请			参与(个)	
	授权		科技报告(篇)		
国外专利	PCT受理		软件著作权(项)		1
	授权		论文论著(篇)		4
获得国家级奖项(项)			其中：被收录论文数(篇)	SCI	1
获得省级奖项(项)				EI	3
新服务(项)				ISTP	
新产品（或新材料、新装备、新品种（系））			新工艺（或新方法、新模式、新技术）		

创新载体项目必填	技术服务数量（项）
	服务企业数量（家）
科技金融项目必填	开展培训宣讲活动场次(次)
	服务企业数量(家)
	帮助企业融资(万元)
	引进专业机构(家)
院士工作站项目必填	引进院士及其团队科技成果转化数量
	院士开展的战略咨询和技术指导次数
	院士年进站次数
	院士及院士团队年进站时间
软科学项目必填	决策咨询报告(篇)(至少1篇)
	研究总报告(篇)(至少1篇)
	研究中期报告(篇)
	研究分报告(篇)
	调研报告(篇)
	专著(篇)[须注明“广东省软科学研究计划项目(项目编号：)资助”]
	核心期刊论文(篇)[以第一作者发表，须注明“广东省软科学研究计划项目(项目编号：)资助”]
	培养人才(人)

	获国家级奖项(项)	
	获省级奖项(项)	
	其他（具体形式：用户填)	
其他成果及形式说明：		
（1）提供树状灌溉管网优化设计示范软件系统1套。 （2）项目成果将通过技术授权、方案设计等方式，服务相关企业，实现技术成果转化。		
2. 主要技术经济指标及社会效益		
累计新增销售收入（万元）		
累计新增利税（万元）		
其他主要技术经济指标及社会效益说明：		
项目负责人（签章）：[illegible]　2015年 10 月 30 日		

三、项目进度和阶段目标

开始日期	结束日期	主要工作内容
2016-01-01	2016-03-31	1. 开展项目的前期准备工作，查找资料，制定详细的研究方案。 2. 对树状灌溉管网的部署需求与网络特性进行分析，构建管网部署规划模型和管径优化模型。
2016-04-01	2016-06-30	1. 基于应用需求对所构建的树状灌溉管网优化模型进行完善。 2. 基于遗传算法、粒子群算法等成熟算法对构建的优化模型进行求解，根据计算结果对优化模型参数进行调整。 3. 分析与总结现有群体智能优化算法的优势与不足。
2016-07-01	2016-09-30	1. 基于冬虫夏草生长机制的内在规律，构建群体智能优化算法的混合策略。 2. 通过标准测试函数验证与调整所提出的混合群体智能优化算法的性能和参数。 3. 将混合群体智能优化算法应用于求解完善后的树状灌溉管网优化模型。
2016-10-01	2016-12-31	1. 分析与总结基于混合群体智能优化算法的计算结果。 2. 进一步调整算法参数及完善优化模型。 3. 撰写科研论文1篇。
2017-01-01	2017-03-31	1. 选择不同的实验环境和实验场地，采购实验材料。 2. 根据计算出的优化参数配置搭建树状灌溉管网实验平台。 3. 申报发明专利1件。
2017-04-01	2017-06-30	1. 分析与总结基于混合群体智能优化算法的实验结果。 2. 根据实验结果继续完善算法参数与优化模型。 3. 撰写科研论文2篇。
2017-07-01	2017-09-30	1. 基于计算机开展树状灌溉管网优化设计系统的代码编写与测试，形成独立的示范软件系统，并进行试点推广。 2. 撰写科研论文1篇。
2017-10-01	2017-12-31	项目总结及撰写研究报告。

四、承担、参与单位工作分工及经费分配情况

承担/参与单位名称（盖章）	工作分工	总经费分摊（万元）	省科技厅经费分配（万元）
华南农业大学	1. 组织开展项目的实施工作，制定详细的实施计划和人员分工方案。 2. 负责构建树状灌溉管网优化模型及设计混合群体智能优化算法等技术研究工作。 3. 组织撰写科研论文、申报专利和软件著作权等。 4. 组织项目验收。	15.00	15.00
	合计	15.00	15.00

五、项目总经费及省科技厅经费预算

1. 省科技厅经费下达总额： （大写）壹拾伍万圆整；（小写 ）15.00万元；
2. 省科技厅经费年度下达计划：（第一期）15.00万元；（余额）0.00万元
3. 总经费开支预算计划：
经费筹集情况：（单位：万元）
总投入经费：15.00

	省科技厅经费	自筹资金				合计
		自有资金	贷款	地方政府投入	其它	
已投入经费：						
新增经费：	15.00					15.00
政府部门、境外资金及其他资金投入情况说明：						

新增经费预算：（单位：万元）

	新增经费总额		省科技厅经费	
支出经费	经费额	用途说明	经费额	用途说明
基建费：	0	无		
1、直接费用：	14.25		14.25	
(1) 设备费：	1.50	包括购买农用灌溉水泵费用1.2万元（0.6万元/台×2台）和购买农用灌溉过滤器费用0.3万元（0.3万元/台×1台）	1.50	包括购买农用灌溉水泵费用1.2万元（0.6万元/台×2台）和购买农用灌溉过滤器费用0.3万元（0.3万元/台×1台）

(2)材料费：	3.00	购买实验所需的不同类型灌溉管道、管道接头、电子激光测距尺、角度测量仪、电磁阀、数据存储模块等费用	3.00	购买实验所需的不同类型灌溉管道、管道接头、电子激光测距尺、角度测量仪、电磁阀、数据存储模块等费用
(3)测试化验加工外协费：	0.50	委托校外加工厂制作灌溉管网实验平台支架费用	0.50	委托校外加工厂制作灌溉管网实验平台支架费用
(4)燃料动力费：	0	无	0	无
(5)差旅费：	3.50	到农田进行调研/实验费用，参加国内学术交流费用等	3.50	到农田进行调研/实验费用，参加国内学术交流费用等
(6)会议费：	0	无	0	无
(7)国际合作与交流费：	0	无	0	无
(8)出版/文献/信息传播/知识产权事务费：	4.00	专利申请费用，论文发表费用，申请软件著作权费用等	4.00	专利申请费用，论文发表费用，申请软件著作权费用等
(9)租赁费：	0	无	0	无
(10)人员费：	1.60	400元/人/月，2人2年共工作40个月，共计1.6万元	1.60	400元/人/月，2人2年共工作40个月，共计1.6万元
(11)专家咨询费：	0.15	500元/人次，3人次共计0.15万元	0.15	500元/人次，3人次共计0.15万元
(12)直接费用其他支出：	0	无	0	无
(13)科技金融服务体系其他费用：	0.00		0.00	
①信用评级补贴：				
②大赛场租：				
③特派员奖励与补贴：				
2、间接费用：	0.75		0.75	
学校管理费：	0.75	5%管理费	0.75	5%管理费
合计：	15.00		15.00	

六、人员信息

项目负责人情况

姓名	年龄	性别	职称	职务	学历	在项目中承担的任务	所在单位	签名
吕石磊		男	讲师	无	博士研究生	主持项目研究工作，具体负责混合群体智能优化算法设计与测试及灌溉管网实验	华南农业大学	吕石磊

主要研究开发人员

姓名	年龄	性别	职称	职务	学历	在项目中承担的任务	所在单位	签名
李震		男	副教授	学院副院长	博士	灌溉管网实验方案分析与设计	华南农业大学	李震
孙道宗		男	高级实验师	教研室主任	博士	灌溉管网优化模型分析与构建	华南农业大学	孙道宗
吕盛坪		男	讲师	无	博士	混合智能优化算法性能分析与测试	华南农业大学	吕盛坪
姜晟		男	实验师	无	硕士	组织搭建灌溉管网实验平台	华南农业大学	姜晟
徐培		男	未取得	无	本科	整理代码资料，实现（软件）示范系统	华南农业大学	徐培
翁江鹏		男	未取得	无	本科	协助进行灌溉管网实验，收集与分析实验数据	华南农业大学	翁江鹏

七、承担、参与单位合作协议（须与申报书中合作协议或意向书相一致）

/

八、合同条款

第一条 甲方与乙方根据《中华人民共和国合同法》及国家有关法规和规定，为顺利完成（2015）年**基于混合群体智能的树状灌溉管网优化技术研究**专项项目（项目编号：2015A020209161）经协商一致，特订立本合同，作为甲乙双方在项目实施管理过程中共同遵守的依据。

第二条 甲方的权利义务：

1. 按合同书规定进行经费核拨的有关工作协调。
2. 根据甲方需要，在不影响乙方工作的前提下，定期或不定期对乙方项目的实施情况和经费使用情况进行检查或抽查。
3. 根据《广东省科技计划项目信用管理办法(试行)》对乙方进行科技计划信用管理。

第三条 乙方的权利义务：

1. 确保落实自筹经费及有关保障条件。
2. 按合同书规定，对甲方核拨的经费实行专款专用，单独列账，并随时配合甲方进行监督检查。
3. 使用财政资金采购设备、原材料等，按照《广东省实施〈中华人民共和国招标投标法〉办法》有关规定，符合招标条件的须进行招标。
4. 项目实施完成或实施到一定程度，须按照《广东省省级科技计划项目结题管理的实施细则（试行）》提出验收或终止结题的申请，并按甲方要求做好项目结题工作。
5. 在每年1月向甲方如实提交上年度工作情况报告，报告内容包含上年度项目进展情况、经费决算和取得的效果等。
6. 按照国家和省有关规定，每年须提交年度科技报告；项目验收时，须提交验收科技报告。

第四条 在履行本合同的过程中，如出现广东省相关政策法规重大改变等不可抗力情况，甲方有权对所核拨经费的数量和时间进行相应调整。

第五条 在履行本合同过程中，需要对项目起止时间、项目经费使用（包括自筹经费、经费分配及经费支出预算等）、项目内容（包括研发内容、技术指标、经济指标及成果指标等）、项目名称、项目承担单位（包括承担单位更名、承担单位替换）、参与单位、项目负责人和成员等进行变更的，甲乙双方按照《广东省省级科技计划项目合同书管理的实施细则（试行）》有关规定执行。

第六条 在履行本合同的过程中，当事人一方发现可能导致项目整体或部分失败的情形时，应及时通知另一方，并采取适当措施减少损失，没有及时通知并采取适当措施，致使损失扩大的，应当就扩大的损失承担责任。

第七条 本项目技术成果的归属、转让和实施技术成果所产生的经济利益的分享，除双方另有约定外，按国家和广东省有关法规执行。

第八条 属技术保密的项目，甲乙双方应另行订立技术保密条款，作为本合同正式内容的一部分，与本合同具有同等效力。

第九条 根据项目具体情况，经双方另行协商订立的附加条款，作为本合同正式内容的一部分，与本合同具有同等效力。

第十条 本合同的争议应由双方本着协商一致的原则解决，如双方协商不成的，则应向甲方所在地法院提起诉讼。

第十一条 保密条款：

1. 本合同保密内容范围为：

/

2. 本合同保密期限为：

/

3. 乙方应与可能知悉保密内容的人员签订技术秘密保护协议。

4. 各方应建立技术秘密保护制度。

5. 属技术保密的项目必须经省负责技术保密部门审查后，确定可否发表或用于国际合作和交流。

第十二条 甲方可根据具体情况决定乙方是否需要单位担保，若需要保证单位，应订立担保条款，作为本合同正式内容一部分。当乙方不履行或不完全履行本合同，以及没有或没有完全承担违约责任时，乙方的保证单位承担连带保证责任。

第十三条 本合同一式六份，各份具有同等效力。甲方存三份，乙方存二份，丙方存一份，本合同自签字之日起生效，有效期至项目结题后一年内。各方均应负合同的法律责任，不应受机构、人事变动的影响。

说明：本合同书中，凡是当事人约定无需填写的内容，应在空白处划（/）。

九、本合同签约各方

管理单位（甲方）： 广东省科学技术厅 （盖章）

单位地址： 广州市连新路171号大院信息大楼

法定代表人（或授权代表）： 黄宁生 （签章）

联系人（经办人）姓名： 林振亮 （签章）

Email： linzl@gdstc.gov.cn

电话： 020-83163905

年 月 日

承担单位（乙方）： 华南农业大学 （盖章）

二级部门： 华南农业大学电子工程学院

单位地址： 广东省广州市天河区五山路483号

法定代表人（或法人代理）： 陈晓阳 （签章）

联系人（项目主管）姓名： 石睿 （签章）

Email： 77909213@qq.com

电话： 020-85283435

开户单位名称： 华南农业大学

开户银行及帐号： 广东广州工行五山支行 3602002609000310520

年 月 日

乙方主管部门（丙方）： 华南农业大学 （盖章）

单位地址： 广东省广州市天河区五山路483号

法定代表人（或法人代理）： 陈晓阳 （签章）

联系人（项目主管）姓名： 石睿 （签章）

Email： 77909213@qq.com

电话： 020-85283435

开户单位名称： 华南农业大学

开户银行及帐号： 广东广州工行五山支行 3602002609000310520

年 月 日

附件13. 广东省科技计划项目验收意见表

广东省科技计划项目验收结题专家组意见表

项目编号	2015A020209161	负责人	吕石磊
项目名称	基于混合群体智能的树状灌溉管网优化技术研究		

2018年3月8日，华南农业大学组织了该项目的材料验收工作。验收专家组审阅了项目实施工作总结报告等相关材料并进行了质询，经讨论形成验收意见如下：

一、提交的验收材料齐全，符合科技计划项目验收要求。

二、该项目通过构建规划模型和设计混合智能优化算法，研究树状灌溉管网工程规划最优方案，为灌溉管网系统规划设计提供科学的设计理论及合理的技术方法。

三、该项目发表与录用科研论文6篇，其中SCI论文1篇，EI论文3篇，会议论文2篇；获授权发明专利1项，公开或申请发明专利2项；获授权软件著作权3项；培养硕士2人，学士4人；研究成果获中国农业机械学会第五届青年科技奖1项。该项目已全部完成合同中规定的各项任务，达到了项目预期指标。

四、财政经费专款专用，符合科技计划项目经费使用管理要求，财务验收通过。

验收结论：☑通过（）不通过

验收等级：☑合格（）良好

验收专家组组长签字：黄XX

日期：2018.3.8

广东省科技计划项目验收结题专家组意见表

项目编号	2015A020209161	负责人	吕石磊
项目名称	基于混合群体智能的树状灌溉管网优化技术研究		

2018年3月8日，华南农业大学组织了该项目的材料验收工作。验收专家组审阅了项目实施工作总结报告等相关材料并进行了质询，经讨论形成验收意见如下：

一、提交的验收材料齐全，符合科技计划项目验收要求。

二、该项目实现了智能、可靠的树状灌溉管网优化机制，为农田节水灌溉的基础建设提供了理论依据，研究成果对促进智能农业发展具有一定的积极作用。

三、该项目发表与录用科研论文6篇，其中SCI论文1篇，EI论文3篇，会议论文2篇；获授权发明专利1项，公开或申请发明专利2项；获授权软件著作权3项；培养硕士2人，学士4人；研究成果获中国农业机械学会第五届青年科技奖1项。该项目已完成合同的各项任务，达到了项目预期指标。

四、财政经费专款专用，符合科技计划项目经费使用管理要求，财务验收通过。

验收结论：☑通过（）不通过

验收等级：☑合格（）良好

验收专家组签字：程建兴

日期：2018.3.8.

广东省科技计划项目验收结题专家组意见表

项目编号	2015A020209161	负责人	吕石磊
项目名称	基于混合群体智能的树状灌溉管网优化技术研究		

2018年3月8日，华南农业大学组织了该项目的材料验收工作。验收专家组审阅了项目实施工作总结报告等相关材料并进行了质询，经讨论形成验收意见如下：

一、提交的验收材料齐全，符合科技计划项目验收要求。

二、该项目实现了基于智能计算的树状灌溉管网优化机制，研究成果扩展了智能计算技术在农业工程领域的应用范畴。

三、该项目发表与录用科研论文6篇，其中SCI论文1篇，EI论文3篇，会议论文2篇；获授权发明专利1项，公开或申请发明专利2项；获授权软件著作权3项；培养硕士2人，学士4人；研究成果获中国农业机械学会第五届青年科技奖1项。该项目已完成合同的各项任务，达到了项目预期指标。

四、财政经费专款专用，符合科技计划项目经费使用管理要求，财务验收通过。

验收结论：(√)通过 ( ) 不通过

验收等级：(√)合格 ( ) 良好

验收专家组签字：周松斌

日期：2018.3.8

项目编号： 2018A030310216

资助类别： 广东省自然科学基金-博士启动
纵向协同

文件编号： 粤科规财字[2018]105 号

广东省自然科学基金

资助项目验收报告

项目名称： 基于混合教-学优化算法的多目标制造云服务组合优化方法研究

项目负责人： 金鸿 资助总经费： 10（万元）

计划完成时间： 2018-05-01 至 2021-04-30

实际完成时间： 2018-05-01 至 2021-04-30

依托单位： 华南农业大学工程学院

联 系 人： 吴双龙 联系电话：

填表日期： 2021 年 6 月 21 日

广东省自然科学基金管理委员会

一、 项目人员信息表

项目负责人：					
姓名	证件号码	职称	承担任务	所在单位	签名
金 鸿	[redacted]	讲师	算法总设计	华南农业大学	
主要研究人员：（ 须如实填写，以便检查核实）					
杨 洲	[redacted]	教 授	项目总规划指导	华南农业大学	
吕盛坪	[redacted]	副教授	约束函数优化分析	华南农业大学	
高锐涛	[redacted]	副教授	融合规则研究	华南农业大学	
汪 隽	[redacted]	讲 师	算法内在联系研究	华南农业大学	
汪 洋	[redacted]	未取得	算法编译与实验验证	华南农业大学	
欧治武	[redacted]	未取得	多目标优化方法分析	华南农业大学	

承担单位（盖章）：华南农业大学工程学院

参与单位 1（盖章）：

参与单位 2（盖章）：

二、 结题摘要

中文摘要（500 字以内）

制造云服务组合是成功实施云制造系统、实现制造资源优化配置的关键性问题之一。制造云服务组合是一类典型的组合优化问题，具有大规模、多极值、非线性、多目标、不确定等复杂性，是 NP-hard 问题。目前针对多目标制造云服务组合优化问题，智能优化算法所求得的最优解受算法初始值的影响，算法的效用与效率也有待于深入研究，且较少有研究针对简单加权和 Pareto 最优解两种多目标优化方法进行优劣分析。为此，本项目以多目标制造云服务组合优化问题为研究对象，引入对算法初始值不敏感的教-学优化算法，研究基于教-学优化算法与其他算法的融合规则，形成新的混合算法；并利用基于罚函数的约束函数处理方法，以提高混合算法在进行多目标制造云服务组合寻优时的效用与效率。然后，利用标准测试函数和制造云服务组合实例，来验证所提算法的有效性。最后，系统分析比较两类多目标优化方法，为其应用提供指导。本项目的研究成果将为多目标制造云服务组合优化问题提供理论、技术和方法的支持。

关键字：云制造；服务组合；教-学算法

三、 报告正文

（一）、总体进展情况

1．项目总体进展情况

对照项目任务书的目标和各项主要考核指标，阐明项目总体进展情况，经费的开支情况，项目重要产出和成果。若未完成目标任务的，说明未完成的原因，并附上相关证明材料。

本项目以多目标制造云服务组合优化问题为研究对象，采用教-学优化算法与其他智能优化算法相结合的手段，研究多目标协同优化的方法，探索教-学优化算法与其他智能优化算法的融合策略，形成新的混合优化算法，以提高智能算法在进行多目标制造云服务组合寻优时的效用与效率；为大规模的制造云服务组合优化问题提供理论、技术和方法的支持。

本项目完成了设定的目标。本项目提出了一种新的混合教-学优化算法来解决制造云服务组合问题，该算法结合了遗传算法的均匀变异、自适应的花授粉算法和教-学优化算法的优势，算法的全局搜索能力强、收敛速度快，该算法效用与效率得到了明显的提高。

项目总经费 10 万元，其中省科技厅下达经费 5 万元，本单位配套经费 5 万元。项目支出经费 5.95 万元，结余经费 4.05 万元。

项目原计划产出论文 2~3 篇，现已有 2 篇论文被录用，还有 1 篇论文在二审中。

2．项目重要调整情况

对项目主要研究内容和考核指标调整、项目牵头单位/参与单位变更、项目负责人变更、项目骨干变更、项目执行期变更等调整情况进行说明（如无调整此项可不写）。

增加了混合鲸鱼优化算法相关内容的研究。提出一种混合鲸鱼优化算法的制造云服务组合优化方法。利用遗传算法的多点交叉和单点变异策略，来提高所提算法在进化后期的全局搜索能力，以维持种群的多样性。

（二）、研究工作完成情况

1．项目的主要研究内容

（1）服务组合 QoS 评价方法。描述服务组合的过程，并阐述 QoS 属性的含义及其计算方法；探讨 QoS 属性值无量纲化处理方法；分析制造云服务组合四种基本结构及其转化方法，进而提出组合服务整体 QoS 的评价方法，给出制造云服务组合问题的数学模型。

（2）教-学优化算法与其他智能优化算法的融合策略。深入研究教-学优化算法与遗传算法、粒子群算法、花授粉算法的融合策略，构建出高性能的、不受初始参数值影响的混合智能优化算法，全面提高混合智能优化算法的全局和局部收敛能力。

（3）实验验证不同类型混合算法的效用与效率。教-学算法可与不同类型的智能优化算法相混合，形成多种混合智能优化算法。针对制造云服务组合求优问题，这就形成了许多种求优方法。利用实验验证，选择出效果较好的混合算法。

（4）混合鲸鱼优化算法的研究。提出一种混合鲸鱼优化算法的制造云服务组合优化方法。利用遗传算法的多点交叉和单点变异策略，来提高所提算法在进化后期的全局搜索能力，以维持种群的多样性。

2．项目的研究方法与技术路线

根据本课题的研究目标，采用如图 1 所示的技术路线来完成前述的研究工作。本课题采用分析与综合、理论探讨、和实验验证相结合的方法。具体方案如下：

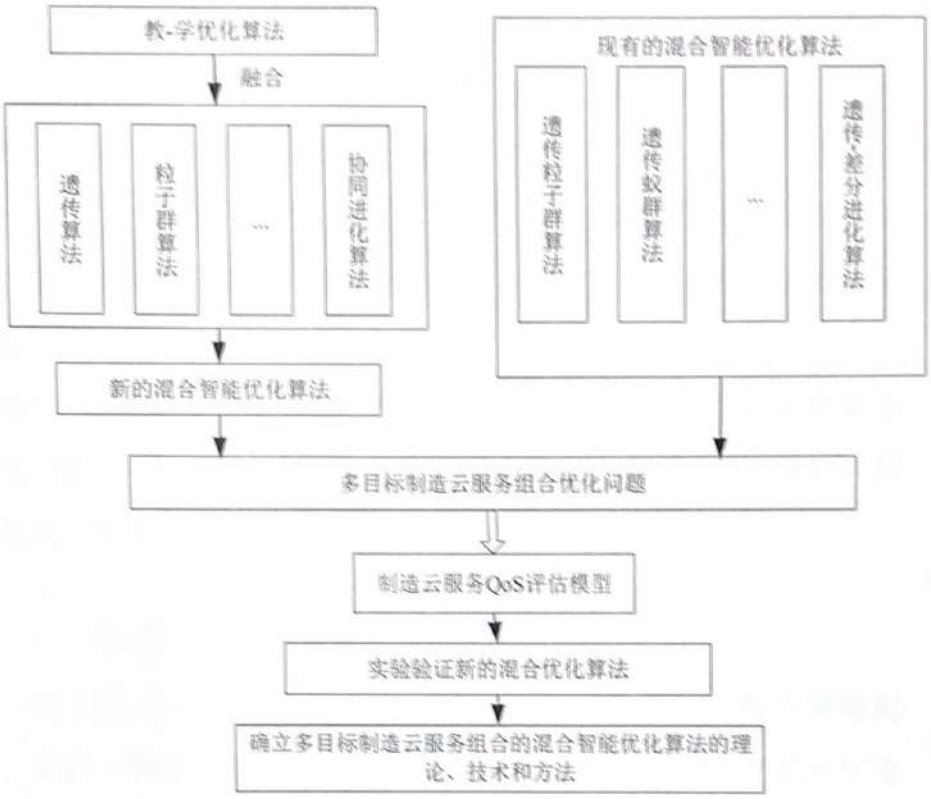


图1 本项目采取的技术路线

（1）服务组合 QoS 评价方法研究。考虑到 QoS 属性的数据类型众多，量纲相异，研究 QoS 属性值无量纲化（归一化）处理方法。为了从组合服务的整体 QoS 的角度来考察服务组合的效用 QoS 值，研究制造云服务组合四种基本结构及其转化方法，进而研究组合服务整体 QoS 的评价方法，提出制造云服务组合问题的数学模型。

（2）教-学优化算法与其他智能优化算法的融合策略研究。综合各种算法的差异性、互补性，扬长避短，研究教-学优化算法与其他智能优化算法（如遗传算法、粒子群算法、花授粉等）的融合策略，构建出高性能的，不受初始参数值影响的混合智能优化算法，全面提高混合智能优化算法的全局和局部收敛能力。

（3）实验验证不同类型混合算法的效用与效率。分别在两种多

目标优化方法的框架下，用大量的仿真实验，将本课题所提出的基于教-学优化算法的混合算法，同现有其他的混合算法相比较，以验证本课题所提出混合算法的效用与效率。

（4）混合鲸鱼优化算法的研究。利用遗传算法的多点交叉和单点变异策略来提高算法后期的全局搜索能力，维持种群的多样性，避免算法过早收敛。将服务组合优化问题形式化描述，利用 QoS 的评价指标，通过将混合鲸鱼算法与鲸鱼优化算法、改进后的花搜粉算法和遗传算法对比，验证混合鲸鱼算法在求解制造云服务组合优化问题的效用和效率。

3．项目解决的科学问题或关键技术问题

项目解决了教-学优化算法与其他智能优化算法的融合策略问题。提出一种新的混合教-学优化算法。该算法结合了遗传的均匀变异、自适应花授粉和教-学算法的优势。

所提的混合算法的表现较好，有两个原因：

一是，利用均匀变异来保持种群的多样性。种群中的个体始终有一定的概率来做变异操作，使得进化方向多变。

另一个是，种群被分为 2 个数目均等的子种群。利用教-学优化算法来进化 Top 子种群，教-学优化算法具有较强的局部搜索能力和较快的收敛速度。利用自适应的花授粉算法来进化剩余的个体，自适应的花授粉具有较强的全局搜索能力。

所提出的混合教-学优化算法集成了均匀突变，自适应的花授粉算法和教-学优化算法的优点，可以快速找到最佳解决方案。

4．项目的特色和创新突破点

项目的特色和创新之处如下：

（1）研究了遗传算法、粒子群算法和教-学算法在服务关联感知的制造云服务组合优化中的适用性，利用教-学优化算法来求解大规模制造云服务组合优化问题的方法，减少控制参数对优化求解结果的影响。

（2）提出了一种新型的混合教-学优化算法。并利用单峰函数、多峰函数和低维函数，来测试所提算法的收敛速度和寻优效果。同时，将其与遗传算法、粒子群算法、鲸鱼算法和花授粉算法等智能优化方法比较，验证所提算法的性能。

5．项目主要研究成果

（1）基于混合教-学算法的制造云服务组合优化方法

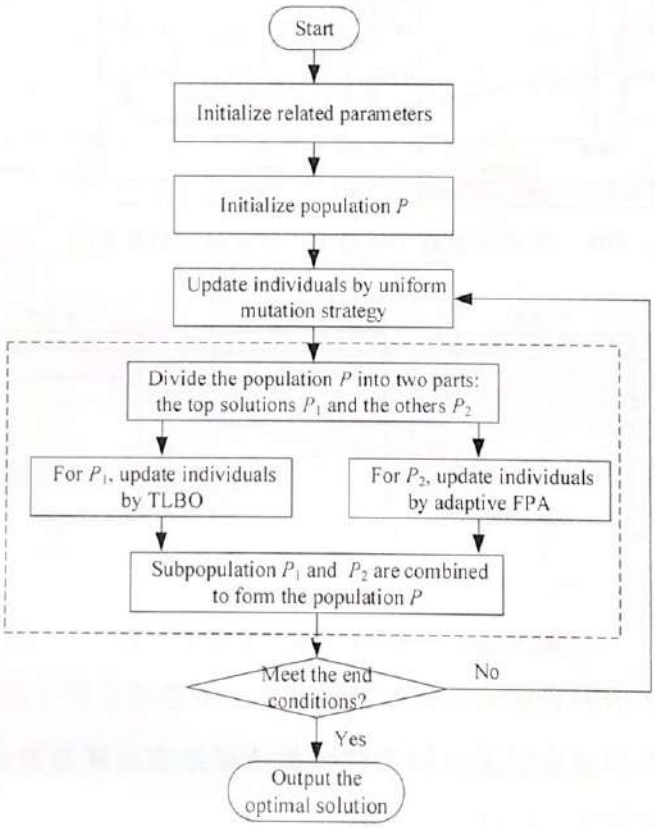


图 2 混合教-学算法流程图

所提混合教-学算法（HTLBO）的流程如图 2 所示。利用均匀变异来保持种群的多样性。同时，种群被分为 2 个数目均等的子种群，利用教-学优化算法来进化 Top 子种群，利用自适应的花授粉算法来进化剩余的个体。利用制造云服务组合实例，将所提混合教-学算法与相关算法的比较，来验证所提算法的有效性，结果如图 3 和图 4 所示。

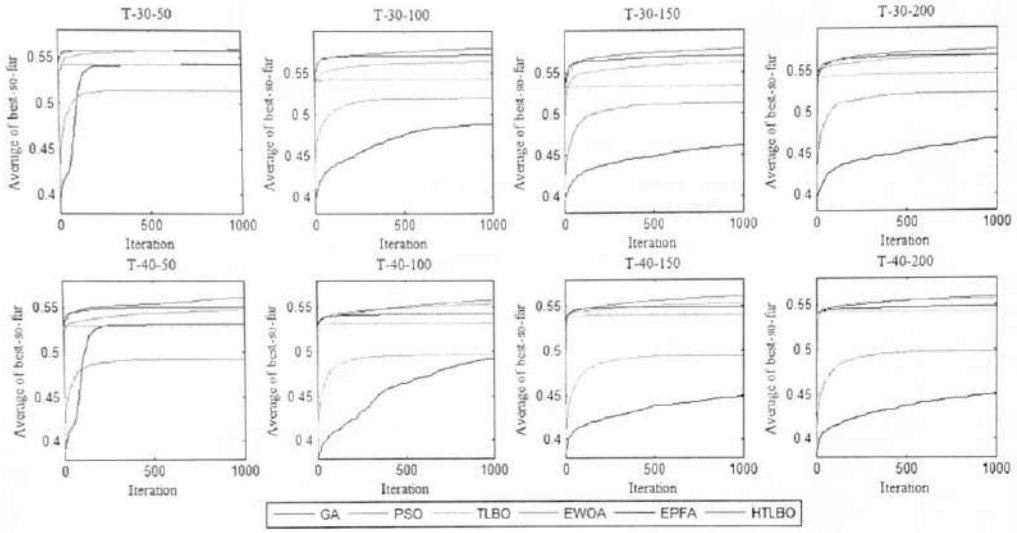


图 3 算法收敛曲线（子任务个数分别为 30、40）

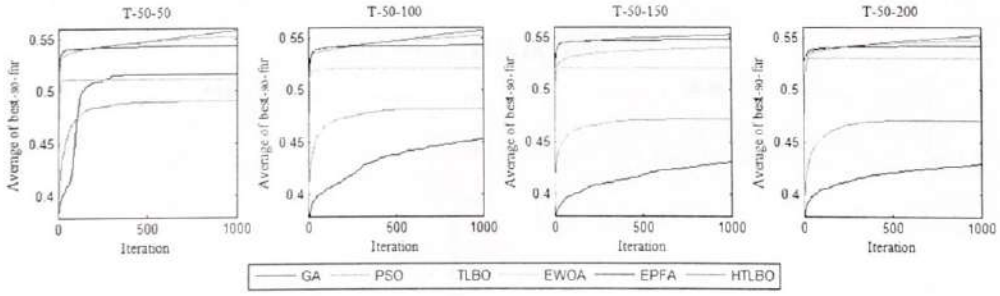


图 4 算法收敛曲线（子任务个数为 50）

（2）基于混合鲸鱼优化算法的制造云服务组合优化方法

为解决智能算法在求解制造云服务组合优化容易陷入局部最优解的问题，提出一种混合鲸鱼优化算法（GAWOA）的来求解制造云服务组合优化问题。所提混合鲸鱼优化算法流程如图 5 所示。利用遗

传算法的交叉和变异策略，来提高混合鲸鱼优化算法的全局搜索能力。最后通过一系列不同规模的制造云服务组合问题，将所提算法与常用的几种算法相比较，验证所提算法的性能，部分结果如图 6 所示。

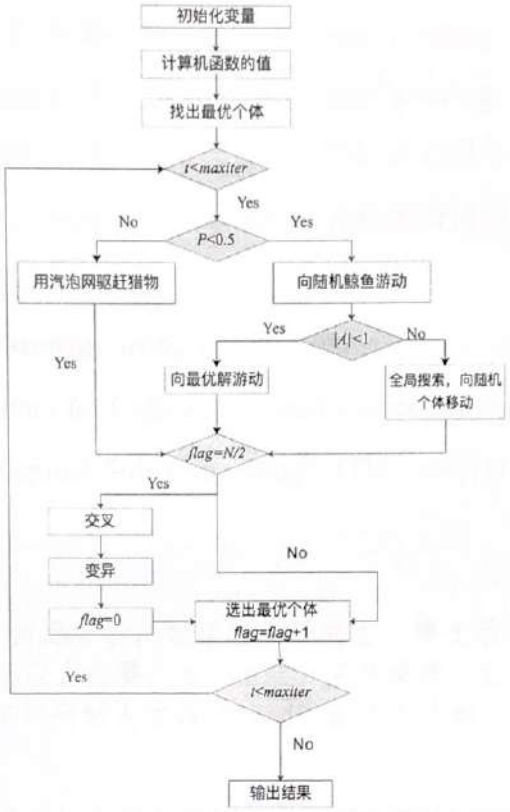


图 5 混合鲸鱼算法流程图

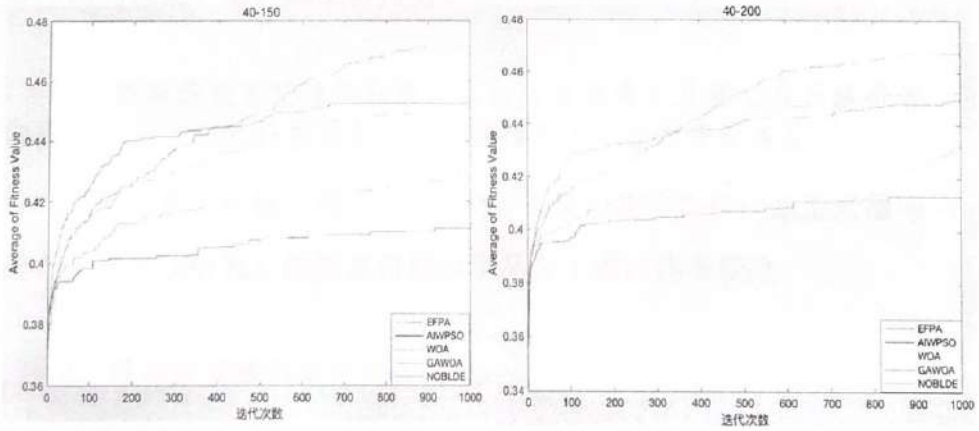


图 6 子任务个数为 40 时算法执的效用和效率

（三）、取得的成果及效益

1．项目科研产出分析，主要包括获奖、发表期刊论文、完成专著、撰写咨询（调研）报告、申请和授权专利、制定标准等，分析代表性科研成果质量、贡献和影响。

项目产出论文 3 篇。论文《A hybrid teaching–learning-based optimization algorithm for QoS-aware manufacturing cloud service composition》和《基于混合鲸鱼优化算法的制造云服务组合优化》已被第 4 届智能优化与调度学术会议录用，并被邀请做口头报告。获得了同行的一致好评。

论文《Eagle strategy using uniform mutation and modified whale optimization algorithm for QoS-aware cloud service compositio》已投稿至 SCI 源期刊《Applied Soft Computing》（Q1、中科院 2 区），目前已进入二审阶段。

2．人才培养情况分析，包括培养研究生、博士后数量，项目组成员所获省部级及以上主要人才（团队）荣誉情况。主要统计分析国家万人计划、科技部创新人才推进计划、省特支计划、珠江人才计划等。

系统地培养了参与人员在模型构建、算法设计与实现、系统开发等方面的技能。培养硕士研究生 2 名。

3．所获后续资助情况分析，包括项目负责人及核心成员获得省部级及以上基础和应用基础研究项目和重大项目的资助情况。

项目负责人申请了题为《面向服务质量预测与协作的制造云服务组合优化方法研究》的国家自然科学基金 1 项，尚未获批。

4．经济效益或社会效益分析。

尚未进行成果推广。

实际参加研究人数	高级职称	中级职称	初级职称	博士后	博士生	硕士生	其他人员
	3	2	0	0	0	2	0
计划执行情况	时间方面(B)		A、提前原计划 B、按原计划 C、滞后原计划				
	内容方面(A)		A、内容增加 B、内容减少 C、内容不变				
项目负责人：（签章）					2021年6月22日		

四、 研究成果目录

序号	成果类型	成果或论文名称	主要完成者	成果说明	标注状况
1	论文	A hybrid teaching–learning-based optimization algorithm for QoS-aware manufacturing cloud service composition	金鸿，江城，吕盛坪，何海平		录用
2	论文	基于混合鲸鱼优化算法的制造云服务组合优化	金鸿，江城，吕盛坪，何海平，朱紫纯		录用
3	论文	Eagle strategy using uniform mutation and modified whale optimization algorithm for QoS-aware cloud service composition	金鸿，吕盛坪，杨洲，刘莹	SCI 源期刊	二审中

五、 经费决算表

（金额单位：万元）

甲方经费下达总额：（大写） 五 万元 ； （小写） 5 万元			
项目编号： 2018A030310216 项目类型：广东省自然科学基金-博士启动级向协同			
项目名称： 基于混合教-学优化算法的多目标制造云服务组合优化方法研究 项目负责人：金鸿			
支出科目	预算经费	经费支出	备注（说明）
科研业务费	1.5	0	论文版面费
实验材料费	0.5	0	计算机配件升级
仪器设备费	1.0	0	购买图形工作站
实验室改装费	0	0	
协作费	0	0	
劳务费	0.75	0.7	研究生劳务费
管理费	0.25	0.25	按 5%计算
其他（具体说明）	1.0	0	绩效支出
合计	5.0	0.95	
甲方拨付经费结余	4.05		
与本项目相关的其他经费来源	预算经费		经费支出
其他计划资助经费	0		0
本单位配套经费	5.0		4.99
其他经费资助	0		0
其他经费来源合计	5.0		4.99
项目负责人： （签章） 2017 年 6 月 23 日	财务部门负责人： （公章） 年 月 日		所在学院负责人： （公章） 年 月 日

注：务必加盖公章，无公章无效。

附：经费使用说明表

填报说明：1.项目负责人需对经费使用情况作一般说明；
2.当预算经费与拨付（到位）经费不相符时，需要特别说明；
3.当经费预算金额与支出金额相差较大是，需要着重说明。

项目总经费 10 万元，其中省科技厅下达经费 5 万元，本单位配套经费 5 万元。预算经费与拨付经费相符。

本单位配套经费中，测试分析费预算 1.4 万元，实际支出 0.808 万元；差旅费预算 2.6 万元，实际支出 2.734 万元，主要用于参加国内学术会议、国内调研或工厂实践差旅费；研究生劳务费预算 1.0 万元，实际支出 0.94 万元；知识产权事务费预算 0 万元，实际支出 0.5 万元；合计支出 4.99 万元。

本项目总预算 10 万元，实际支出 5.95 万元，结余 4.05 万元。未使用完的经费中，论文录用，版面费待缴；设备仪器购置和计算机配件升级工作，有待进一步推进；绩效则需等项目结题后才能支出。其他各项经费，按照实际预算进行开支。

项目负责人：	财务部门负责人：	所在学院负责人：
（签章）	（公章）	（公章）
2021 年 6 月 13 日	年 月 日	年 月 日

六、 专家组意见表

专家信息表					
序号	姓名	年龄	职称	所在单位	签名
1	姚锡凡		教 授	华南理工大学	姚锡凡
2	刘建军		副教授	广东工业大学	刘建军
3	谭 伟		副教授	东莞理工学院	谭伟

专家组意见

受华南农业大学委托，专家组对华南农业大学工程学院金鸿承担的广东省自然科学基金项目“基于混合教-学优化算法的多目标制造云服务组合优化方法研究”（编号：2018A030310216）进行结题验收，专家组审阅了项目相关资料，经邮件、电话咨询和充分讨论后，形成如下验收意见：

1. 项目组提供的验收资料齐备，符合验收条件。

2. 本项目针对智能优化算法在求解制造云服务组合优化问题时，寻优结果容易受算法初始值影响的问题，提出了一种混合教-学优化算法；针对智能算法的容易陷入局部最优解的问题，探索了混合的鲸鱼优化算法；同时利用制造云服务组合实例，验证所提出算法的有效性。

3. 项目执行期间，项目组人员撰写论文 3 篇。科研经费专款专用，使用合理。

验收专家一致认为，项目组完成了合同书中的考核指标，经费使用合理，同意通过结题验收。

是否同意结题： 同意	专家组组长签名： 姚锡凡 2021 年 6 月 15 日

七、 项目负责人签字及审核意见表

项目负责人承诺：

我所承担的项目（编号：2018A030310216 名称：基于混合教-学优化算法的多目标制造云服务组合优化方法研究）验收报告内容填写实事求是，数据详实。在今后的研究工作中，如有与本项目相关的成果，将标注"广东省自然科学基金资助"，并报送广东省自然科学基金委员会。

负责人（签章）：

2021 年 6 月 23 日

所在学院审查意见：

同意，建议加快应用成果转化。

经办人（签章）：

单位公章：

年 月 日

学校意见：

负责人（签章）：

单位公章：

年 月 日

三、科研成果---第一作者发表论文清单

SCAULIB202625938

检索证明

根据委托人提供的论文材料，委托人工程学院 吕盛坪(学科类型:自然科学) 10 篇论文收录情况如下表。

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	A lightweight hierarchical aggregation task alignment network for industrial surface defect detection	EXPERT SYSTEMS WITH APPLICATIONS 出版年：2025 出版日期：MAR 5 卷期：263 页码：- 文献号：125727 文献类型：Article	1	T2 类	华南农业大学工程学院	SCI	IF2-year=7.5 IF5-year=7.8 （2024）	计算机科学 1区 Top 期刊：是 OA 期刊：否 标注：Mega-Journal （2025）
2	A dataset for deep learning based detection of printed circuit board surface defect	SCIENTIFIC DATA 出版年：2024 出版日期：JUL 22 卷期：11 1 页码：- 文献号：811 文献类型：Article	1	A 类	华南农业大学工程学院	SCI	IF2-year=6.9 IF5-year=8.7 （2024）	综合性期刊 2区 Top 期刊：否 OA 期刊：是 （2025）
3	An enhanced walrus optimization algorithm for flexible job shop scheduling with parallel batch processing operation	SCIENTIFIC REPORTS 出版年：2025 出版日期：FEB 17 卷期：15 1 页码：- 文献号：5699	1	B 类	华南农业大学工程学院	SCI	IF2-year=3.9 IF5-year=4.3 （2024）	综合性期刊 3区 Top 期刊：否 OA 期刊：是 标注：Mega-

		文献类型：Article						Journal （2025）
4	An FCM-GABPN Ensemble Approach for Material Feeding Prediction of Printed Circuit Board Template	APPLIED SCIENCES-BASEL 出版年：2019 出版日期：OCT 卷期：9 20 页码：- 文献号：4455 文献类型：Article	1	B 类	华南农业大学 工程学院	SCI	IF2-year=2.474 IF5-year=2.458 （2019）	工程技术 3区 Top 期刊：否 OA 期刊：是 （2019）
5	A Modified Bayesian Network Model to Predict Reorder Level of Printed Circuit Board	APPLIED SCIENCES-BASEL 出版年：2018 出版日期：JUN 卷期：8 6 页码：- 文献号：915 文献类型：Article	1	B 类	华南农业大学 工程学院	SCI	IF2-year=2.217 IF5-year=2.287 （2018）	工程技术 3区 Top 期刊：否 （2018）
6	A Review of Data Mining with Big Data towards Its Applications in the Electronics Industry	APPLIED SCIENCES-BASEL 出版年：2018 出版日期：APR 卷期：8 4 页码：- 文献号：582 文献类型：Review	1	B 类	华南农业大学 工程学院	SCI	IF2-year=2.217 IF5-year=2.287 （2018）	工程技术 3区 Top 期刊：否 （2018）

7	A cross-entropy-based approach for joint process plan selection and scheduling optimization	PROCEEDINGS OF THE INSTITUTION OF MECHANICAL ENGINEERS PART B-JOURNAL OF ENGINEERING MANUFACTURE 出版年：2016 出版日期：AUG 卷期：230 8 页码：1525-1536 文献类型：Article	1	B 类	华南农业大学工程学院	SCI	IF2-year=1.366 IF5-year=1.386 （2016）	工程技术 4区 Top 期刊：否 （2016）
8	深度学习在我国农业中的应用研究现状	计算机工程与应用 出版年：2019 出版日期：2019-08-23 15:08 卷期：55 20 页码：- 文献号： 文献类型：期刊论文	1	C 类	华南农业大学工程学院	北大核心	无	无
9	基于数据挖掘的印制电路样板投料优化	系统仿真学报 出版年：2018 出版日期：2018-07-08 卷期：30 07 页码：- 文献号： 文献类型：期刊论文	1	B 类	华南农业大学工程学院	北大核心	无	无

10	荔枝不同预冷方式降温特性研究	华南农业大学学报 出版年：2015 出版日期：2015-04-14 09:43 卷期：36 03 页码：- 文献号： 文献类型：期刊论文	1	B 类	华南农业大学 工程学院	北大核心	无	无

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明:如未盖章,报告无效

检索员：尹银怀

华南农业大学图书馆

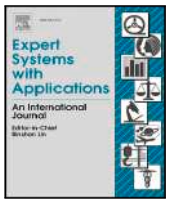
2026-03-11

Expert Systems With Applications 263 (2025) 125727

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



A lightweight hierarchical aggregation task alignment network for industrial surface defect detection

Shengping Lv [a,*], Tairan Liang [a], Kaibin Zhang [a], Shixin Jiang [b,c], Bin Ouyang [a], Quanzhou Li [b,c], Xiaoqing Li [a]

[a] School of Engineering, South China Agricultural University, Guangzhou, 510642, China
[b] CEPREI, No. 78, Zhucun Avenue West, Guangzhou, 511370, China
[c] Key Laboratory of Industrial Equipment Quality Big Data, MIIT, No. 78, Zhucun Avenue West, Guangzhou, 511370, China

ARTICLE INFO

Keywords:
Industrial surface defect
Real-time detector
Low-parameter backbone
Hierarchical multiscale feature enhancement path aggregation network
Friedman test

ABSTRACT

Industrial surface defect detection is crucial for maintaining product quality, but it faces challenges such as complex background interference, numerous small defects, and significant variations in defect characteristics. To address these challenges, this paper introduces a novel lightweight hierarchical aggregation task alignment network (LHATA-Net) designed to enhance detection accuracy, computational efficiency, and generalization. LHATA-Net includes four innovative features: (1) a fast-efficient layer aggregation network (F-ELAN) for efficient feature extraction; (2) a hierarchical multiscale feature enhancement path aggregation network (HMFE-PAN) to improve detection of small defects in complex backgrounds; (3) a lightweight task aligned head (LTA-Head) to optimize feature interaction between classification and localization; and (4) a slide loss function (Slideloss) that integrates slide weighting function with binary cross entropy with logits loss function to tackle sample imbalance. To better validate the detector, we compile a large-scale dataset, DsPCBSD+, which includes real images of surface defects on printed circuit boards from practical industrial production. Experimental results demonstrate that LHATA-Net, with only 3.5M parameters and 18.4G floating point operations per second, achieves an inference speed of 54.2 frames per second. It also achieves average precision of 79.6%, 70.0%, and 85.8% at an intersection over union threshold of 0.5 on two steel surface defect datasets and the DsPCBSD+ dataset, respectively. It ranks first, second, and third compared to state-of-the-art (SOTA) real-time detectors. The Friedman test confirms that LHATA-Net surpasses SOTA detectors in overall performance, highlighting its superiority in practical engineering applications. The code is available at https://github.com/Tarzan-Leung/LHATA-Net.

1. Introduction

Industrial surface defect detection (ISDD) is crucial across various fields, including steel (He et al., 2020; Lv et al., 2020), aluminum (Baidu, 2021; Guo et al., 2022), photovoltaic panel (PVP) (CCNUZFW, 2023; Li, Wang et al., 2023; Su, Zhou, Chen, 2023), and printed circuit boards (PCBs) (Huang et al., 2020; Lv et al., 2024; Tang et al., 2019) and so on. It plays a pivotal role in ensuring product quality. Throughout the manufacturing process, various defects can arise on industrial product surfaces due to factors such as material quality, technical faults, environmental contamination, equipment anomalies, and human error. These surface imperfections not only impact aesthetics but also have the potential to significantly impair industrial product performance, compromise safety, shorten lifespan, and result in wastage and substantial costs. However, these issues can be mitigated to some extent by accurately and efficiently identifying defects with high reliability during the production process.

ISDD in the past primarily relied on manual visual inspection. However, this approach suffers from several drawbacks, including dependence on experienced inspectors, subjectivity, high labor intensity, and inconsistency in efficiency (Moganti et al., 1996). With increasing demands for precise and efficient inspection of complex, intricate products, manual inspection is gradually becoming obsolete. Over the past decades, the field of computer vision (CV) has flourished, progressively replacing manual inspection across various industrial companies (Ling & Isa, 2023; Zhou, Yuan et al., 2023). Nonetheless, traditional CV-based methods require manually designed algorithms to extract defect features, which are cumbersome and exhibit limited generalization

* Corresponding author.
E-mail addresses: lvshengping@scau.edu.cn (S. Lv), liangtairan@stu.scau.edu.cn (T. Liang), zhangkaibing@stu.scau.edu.cn (K. Zhang), jiangshixin@ceprei.com (S. Jiang), ouyangbin2021@stu.scau.edu.cn (B. Ouyang), liquanzhou@ceprei.com (Q. Li), lixiaoqing@stu.scau.edu.cn (X. Li).

https://doi.org/10.1016/j.eswa.2024.125727
Received 18 September 2024; Received in revised form 29 October 2024; Accepted 5 November 2024
Available online 15 November 2024

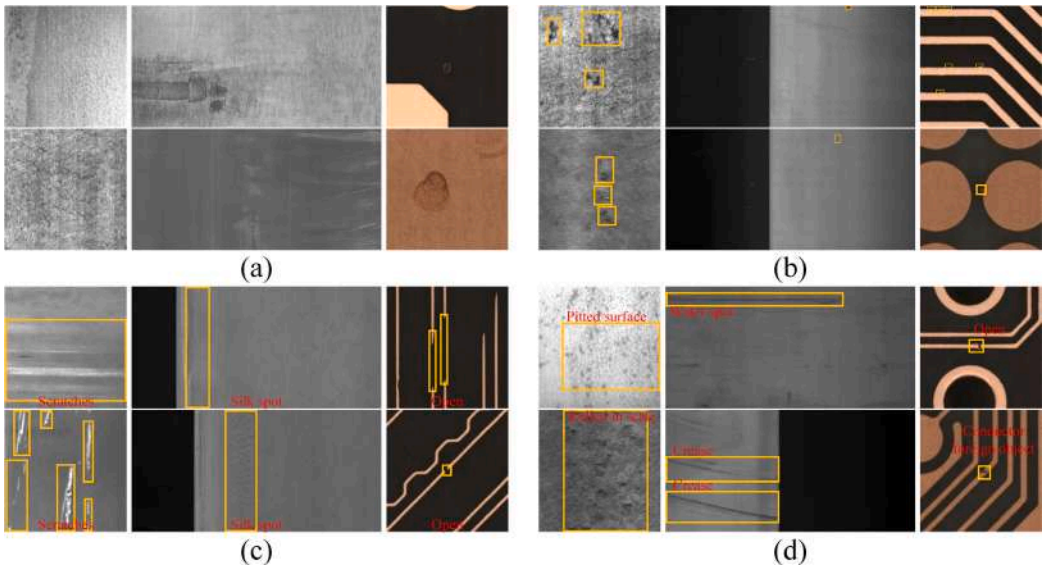


Fig. 1. Examples of surface defects on different industrial products.

performance. Recent advancements in deep learning (DL) technology have led to a growing emphasis on end-to-end CV-based detectors in ISDD research (Ameri et al., 2024; Gao et al., 2022), due to their high accuracy, fast detection speed, and superior generalization across various application scenarios.

DL-based detectors are categorized into two types: one-stage and two-stage detectors, both stemming from convolutional neural networks (CNNs). One-stage detectors regress bounding boxes and object probabilities directly without employing region proposals, which enhances their efficiency in terms of inference speed. In contrast, two-stage detectors use region proposals to improve localization accuracy and detection performance, albeit at the expense of slower inference speeds. The YOLO series (e.g.,YOLOv5 to YOLOv10) and the RCNN series (e.g., Faster R-CNN, Mask R-CNN, and Cascade RCNN) exemplify one-stage and two-stage object detection detectors, respectively. These detectors serve as foundational baselines in various object detection applications, including ISDD (Chen, Feng et al., 2023; Li, Kong et al., 2023; Lu, Wangqi et al., 2024; Lu, Yu et al., 2024), which poses specific challenges as follows:

(1) **Complex background interference**: In industrial environments, complex backgrounds are common and characterized by diverse colors, textures, and variations in lighting. Surface defects in such complex backgrounds may blend with the colors or textures, causing the boundaries of defective objects in images to appear blurred, as shown in Fig. 1(a).

(2) **Numerous small defects**: There is an increasing prevalence of small, intricate defects due to a growing trend towards greater intricacy and smaller sizes of products, as illustrated in Fig. 1(b). Consequently, this trend necessitates high-resolution detectors for effective identification of defects.

(3) **Significant variations in defects**: Industrial surface defects appear in various categories, with an imbalanced distribution of samples across these categories. Additionally, defects exhibit significant variations in shape, scale, and color resulting in higher intra-class differences and inter-class similarities, as illustrated in Fig. 1(c) and (d) respectively. Thus, developing a highly discriminative detector is crucial for accurate detection.

To meet practical requirements and address challenges in ISDD, several enhancement strategies have been proposed to improve fundamental detectors, focusing on the backbone, neck, and head networks of detectors. These enhancements include advanced techniques such as feature extraction (Li et al., 2024; Shao et al., 2024; Wang, Zhang et al., 2024; Zhang, Zhang et al., 2024; Zhang, Zhou et al., 2023; Zhao et al., 2023; Zhou, Yang et al., 2023), feature fusion (Li et al., 2024; Zhang, Hao et al., 2023; Zhang, Zhou et al., 2023; Zhao et al., 2023; Zheng et al., 2023) , feature regression (Zheng et al., 2024; Zhou, Yang et al., 2023), and attention mechanisms (Chen, Feng et al., 2023; Hou et al., 2023; Li et al., 2024; Lu, Yu et al., 2024; Xiao et al., 2024; Zhang, Li et al., 2024; Zhang, Zhang et al., 2024; Zhang, Zhou et al., 2023; Zhao, Chen et al., 2024; Zhao et al., 2023; Zhou, Lu et al., 2023; Zhu et al., 2023), among others. Continuously improving DL-based detectors has enhanced their performance in ISDD, including detection accuracy, computational efficiency, and generalization.

To tackle the challenges of ISDD, and enhance the performance of DL-based detector, this paper introduces a novel lightweight hierarchical aggregation task alignment network (LHATA-Net). The LHATA-Net comprises four key components tailored to enhance performance in challenging scenarios: First, a fast-efficient layer aggregation network (F-ELAN) is proposed as the core of the feature extraction module, which reduces parameters while maintaining high detection accuracy. Second, the hierarchical multiscale feature enhancement path aggregation network (HMFE-PAN) is introduced to address the challenge of detecting small, intricate defects amid complex backgrounds. Furthermore, a lightweight task aligned head (LTA-Head) is developed to improve feature interaction between classification and localization tasks and to mitigate variations in defects. Finally, the slide loss function (Slideloss) is developed by integrating the slide weighting function with the binary cross entropy with logits loss function (BCEWithLogitsloss) to address the issue of category imbalance. The corresponding results demonstrate the superior performance of the proposed LHATA-Net. The main contributions of this paper are as follows:

(1) To reduce parameters while maintaining detection accuracy, we integrate the generalized efficient layer aggregation network (GELAN) with FasterNet to develop the F-ELAN module for feature extraction.

(2) For improving detection of small and intricate defects amidst complex background interference, we propose HMFE-PAN, a novel feature fusion network. HMFE-PAN incorporates the context anchor attention (CAA) module and innovative feature fusion pathways to selectively process information across different scales, effectively reducing redundant features and enhancing defect detection accuracy.

(3) To address intra-class differences and inter-class similarities in defects, we introduce the LTA-Head, which combines a task decomposition module (TDM) with deformable ConvNets V2 (DCNv2) to enhance semantic linkage between classification and localization. Additionally, we use Slideloss to tackle sample imbalance across defect categories.

(4) To overcome the limitations of existing ISDD datasets, such as repetitive defect samples, artificial synthesis defects, and reliance on data augmentation, we construct the large-scale DsPCBSD+ dataset. This dataset comprises real images of PCBs surface defects from practical industrial production and is intended to improve the evaluation of various detectors.

(5) To compare the overall performance of various detectors in terms of detection accuracy, computational efficiency, and generalization, we introduce the Friedman test statistical method. This approach provides a thorough and valuable analysis for a comprehensive comparison of different detectors.

The structure of the remaining sections is as follows: Section 2 reviews the detectors and datasets used in ISDD, along with techniques involving the backbone, neck, and head networks of DL-based detectors. Section 3 provides a detailed description of LHATA-Net, covering F-ELAN, HMFE-PAN, LTA-Head, and the Slideloss. Section 4 presents the experimental validation of LHATA-Net and analyzes the results obtained. Finally, Section 5 concludes with a summary and outlines future research directions.

2. Related works

In ISDD, scholars have introduced numerous DL-based detectors and evaluated their efficacy across various datasets. These detectors primarily center on research concerning backbone, neck, and head networks. Consequently, this review will encompass two main aspects: the detectors and datasets, along with the methodologies involving backbone, neck, and head networks.

2.1. Detectors and datasets

Recent advancements in DL have led to a focus on end-to-end DL-based approaches for ISDD. Tian and Jia (2022) introduced DCC-CenterNet, which balances detection speed and accuracy, performing exceptionally well on the NEU-DET and GC10-DET datasets. Zhu et al. (2023) developed the LSwin Transformer, which outperforms other methods on the GC10-DET dataset. Li, Kong et al. (2023) presented EFD-YOLOv4, achieving 79.88% mean average precision (mAP) on NEU-DET dataset and 54.65% on GC10-DET dataset, demonstrating robust defect detection capabilities. Zhou, Lu et al. (2023) proposed a YOLOv5s-based detector that integrates CSPlayer and a global attention enhancement mechanism, surpassing YOLOv5s in precision, mAP, and speed on the GC10-DET dataset. Yu, Wang et al. (2024) introduced CAGLNet, which excels in accuracy and speed on the NEU-DET dataset, surpassing state-of-the-art (SOTA) methods. Du et al. (2024) developed AFF-Net, showing significant performance improvements over SOTA methods for diverse defect categories on the NEU-DET dataset. Lu, Yu et al. (2024) created SS-YOLO, a lightweight YOLOv7-based detector that enhances detection accuracy and efficiency on the NEU-DET dataset. Building on YOLOv8, Lu, Wangqi et al. (2024) developed WSS-YOLO, which outperforms SOTA methods in detection accuracy on both NEU-DET and GC10-DET datasets. Zhang, Li et al. (2024) presented an improved YOLOv5, achieving superior defect detection and localization on NEU-DET and GC10-DET datasets, with potential for further application on the APDDD dataset (Tianchi, 2018). These detectors primarily target steel surface defect detection.

Ding et al. (2019) developed TDD-Net to enhance tiny defect detection in industrial products, achieving competitive results on the augmented HRIPCB (Huang et al., 2020) dataset. Zeng et al. (2022) designed ABFPN for PCBs surface defect detection, showing superior performance and practicality on HRIPCB. Xia et al. (2023) introduced GCC-YOLO for PCBs surface defect detection, outperforming YOLOv5s in precision, recall, and mAP on the augmented HRIPCB, while also being more compact and faster. Xiao et al. (2024) created CDI-YOLO based on YOLOv7-tiny, which demonstrates better accuracy, speed, and efficiency on the augmented HRIPCB dataset. All these studies focus on PCBs surface defect detection.

Chen, Feng et al. (2023) developed a YOLOv5-based detector for industrial surface defects, showing improved mAP on the NEU-DET and PV-Multi-Defect datasets (CCNUZFW, 2023; Li, Wang et al., 2023). Su, Zhou, Wan et al. (2023) developed ICT-EDNet, surpassing SOTA methods on NEU-DET, DeepPCB (Tang et al., 2019), and Track Components dataset (Workspace, 2021). Zhou, Yang et al. (2023) devised ETD-Net, achieving competitive performance on NEU-DET, sewer defect detection dataset (SEDD), and Aliyun Tianchi fabric defect detection (AT-FDD) datasets (Tianchi, 2020). Liu et al. (2024) introduced a real-time anchor-free detector for ISDD, achieving SOTA performance on PVEL-AD (Su, Zhou, Chen, 2023), NEU-DET, and augmented HRIPCB datasets. Shao et al. (2024) created TD-Net, enhancing tiny defect detection with YOLOv5, and demonstrated competitive performance on NEU-DET, HRIPCB, and GC10-DET datasets. Li et al. (2024) presented IDP-Net, achieving high mAP on NEU-DET, augmented HRIPCB, and ASDD (Baidu, 2021) datasets while maintaining high inference speed. Zhang, Zhang et al. (2024) established DsP-YOLO, demonstrating superior accuracy and rapid inference on NEU-DET, HRIPCB, and GC10-DET datasets. The aforementioned detectors have been validated across multiple datasets encompassing two or three types of materials, such as steel, aluminum, PVP, and PCBs.

Recent advances in DL-based ISDD have seen significant progress, yet achieving high accuracy, efficiency, and robust generalization remains a challenge. Additionally, detectors often struggle to balance these aspects, and there is currently no unified analytical method to comprehensively evaluate their performance. To address these issues, we propose LHATA-Net, a novel detector with an innovative architecture incorporating a new backbone, neck, and head network design. Furthermore, we introduce the Friedman test statistical method to provide a comprehensive and insightful comparison of various detectors.

Research shows that NEU-DET (He et al., 2020), GC10-DET (Lv et al., 2020), ASDD (Baidu, 2021), APDDD (Tianchi, 2018), PV-Multi-Defect (CCNUZFW, 2023; Li, Wang et al., 2023), PVEL-AD (Su, Zhou, Chen, 2023), HRIPCB (Huang et al., 2020), and DeepPCB (Tang et al., 2019) datasets are widely used for detectors training and validation. NEU-DET and GC10-DET target steel, ASDD and APDDD focus on aluminum, PV-Multi-Defect and PVEL-AD are constructed for PVP, and HRIPCB and DeepPCB are defect datasets of PCBs. However, many ISDD datasets still exhibit certain shortcomings, such as repetitive defect samples, artificial synthesis defects, reliance on data augmentation, and extreme sample imbalance. For example, ASDD includes multiple views of the same defect, while APDDD contains duplicates images, leading to either repetitive defect samples or highly similar defect sample sets. HRIPCB and DeepPCB are synthetically generated, and PV-Multi-Defect and PVEL-AD suffer from significant imbalance of samples. Additionally, many studies (Ding et al., 2019; Li et al., 2024; Liu et al., 2024; Xia et al., 2023; Xiao et al., 2024) first augment and then partition datasets, resulting in high similarity between training and validation sets. The repetitive defect samples, artificial synthesis defects, and samples generated through data augmentation do not accurately reflect diverse real-world defects. Moreover, these datasets require users to partition training and validation sets, different partitions can lead to inconsistencies in detector outcomes, complicating the assessment of a detector's accuracy and generalization. The Track Components dataset (Workspace, 2021) and the AT-FDD dataset (Tianchi, 2020) have also been used in several studies for model validation. However, the Track Components dataset is dominated by scratch-type defects, while the AT-FDD dataset features a highly uneven distribution of surface defects. To address these issues, we present the large-scale DsPCBSD+ dataset (Lv et al., 2024), which comprises PCB surface defects collected from real images captured by automated optical inspection (AOI) systems, with standardized sample partitioning and validation. By utilizing datasets that closely align with practical engineering applications, this study aims to provide a more reliable and robust evaluation of LHATA-Net's accuracy and generalization capabilities.

2.2. Methodologies involving backbone, neck, and head networks

2.2.1. Feature extraction methods in backbone network

The backbone network, the main component of the detector, extracts features and provides input to the neck network. It is crucial for detector performance and contains most parameters. Researchers aim to enhance the backbone to improve accuracy and efficiency in ISDD.

The CNNs architecture is one of the widely used backbone structure for DL-based detectors. Li, Kong et al. (2023) enhanced YOLOv4 with a convolutional encoder–decoder module and an ECA-based feature alignment module for better feature representation and scale information. Chen, Feng et al. (2023) introduced the SPPFKCSPC module in YOLOv5 for improved multiscale feature fusion and replaced the C3 module to broaden the receptive field. Zhou, Lu et al. (2023) substituted the C3 module with CSPlayer in YOLOv5s and added global attention mechanism for improved small target detection. Shao et al. (2024) developed a defect downsampling module in TD-Net to mitigate defect information loss. Lu, Wangqi et al. (2024) added the C2f-DSC module to WSS-YOLO's backbone for better feature extraction. Xiao et al. (2024) integrated coordinate attention and DSConv in YOLOv7-tiny to boost feature extraction and detection speed. Zhang, Li et al. (2024) introduced the CRA block in YOLOv5 to enhance feature extraction and detail perception. Lu, Yu et al. (2024) replaced the CBS module in YOLOv7 with MobileNetv3 and added the D-SimSPPF module to reduce model size and parameters. Du et al. (2024) proposed the AFF-Block with diffusion-aggregation to adaptively focus on defect features, improving recognition across different morphologies.

The CNNs with Transformer (Vaswani et al., 2017) architecture is another commonly adopted backbone structure for DL-based detectors. Zhu et al. (2023) introduced a convolutional embedding and attention block merging module during downsampling to improve feature map channel connectivity, resolution reduction, and image detail preservation. They also developed a window shifting strategy to enhance defect modeling by increasing interactive computation chances. Su, Zhou, Wan et al. (2023) proposed the edge-interactive deep convolution (EIDC) and feature cyclic shift transformer (FCST) modules to capture local textures and global semantics in the ICT-EDNet encoder. Zhou, Yang et al. (2023) designed a modified lightweight vision transformer (M-LVT) for ETDNet to enhance global representation capabilities. Li et al. (2024) introduced the LLG-Net, which integrates self-attention and convolutional blocks in IDP-Net to boost detection under complex conditions.

The CNNs with Transformer architecture generally achieves higher detection accuracy, but the multi-head self-attention mechanism introduced by Transformer increases detector parameters and reduces detection speed. In contrast, the CNNs architecture has fewer parameters compared to CNNs with Transformer architecture. It can be combined with problem-specific characteristics to reduce detector depth and width further, thereby decreasing detector parameters while maintaining high detection accuracy. Therefore, we develop a novel F-ELAN module for feature extraction in the backbone network based on the CNNs architecture, aimed at streamlining parameters without compromising detection accuracy.

2.2.2. Feature fusion methods in neck network

The neck network acts as a feature fusion framework, typically integrating shallow and deep features to enhance positional accuracy while filtering noise. One mechanism to enhance the structure of feature fusion is to redesign the paths of feature forward propagation. Li, Kong et al. (2023) developed a hierarchical multi-scale module with residual connections in EFD-YOLOv4 to capture diverse scale features effectively. Xia et al. (2023) improved GCC-YOLO with high-resolution P2 layer features and BiFPN to boost small target detection. Su, Zhou, Wan et al. (2023) enhanced ICT-EDNet's decoder with feature similarity bridging (FSB) and feedback spatial information regulation (FSIR) to preserve local details and global perception. Zhou, Yang et al. (2023) proposed a channel-modulated feature pyramid network (CM-FPN) in ETDNet for multi-level feature fusion and information preservation. Shao et al. (2024) introduced the semantic information interaction module (SIIM) and scale information fusion module (SIFM) in TD-Net to integrate deep and shallow features and resolve multi-scale conflicts. Li et al. (2024) enhanced IDP-Net by incorporating the multiscale perceptual feature aggregation network (MPA-Net), which improves multi-scale target detection and connectivity. Additionally, IDP-Net incorporates adaptive cross-layer feature fusion (ACFF), which integrates features from adjacent layers, reducing semantic discrepancies and enhancing connectivity across multi-scale semantic information. Du et al. (2024) designed Foc-FPN to synthesize multi-scale defect features with a focusing strategy. Liu et al. (2024) introduced local feature enhancement module (LFEM) in neck to amplify small defect features.

Another mechanism to enhance the structure of feature fusion is replacing the original neck module with improved modules. Chen, Feng et al. (2023) integrated a coordinate attention mechanism into YOLOv5's neck to refine feature focus. Zhang, Zhang et al. (2024) added a lightweight and detail-sensitive PAN (DsPAN) module to YOLOv8 for better representation of small objects, and introduced the attention-based LCBHAM module to capture critical local details while filtering out less significant features. Lu, Wangqi et al. (2024) developed a slim-neck design for WSS-YOLO that maintains detection performance. Zhang, Li et al. (2024) applied multi-scale feature fusion (MSF) in YOLOv5 to merge shallow and deep features, enhancing detailed defect information. Lu, Yu et al. (2024) used the parameter-free attention mechanism SimAM in SS-YOLO's neck to improve detection of surface defects and reduce background interference.

Despite the notable effectiveness of these methods, they currently do not fully address challenges associated with cross-scale interactions, which results in a loss of global information and restricts the improvement of detection accuracy. To address this issue, we propose HMFE-PAN for feature fusion. HMFE-PAN effectively filters redundant features and strengthens semantic connections within the same scale, while also optimizing cross-scale feature fusion. This approach enhances the detection of small and intricate defects amidst complex background interference, all while reducing the detector's parameters.

2.2.3. Feature regression methods in head network

The head network receives feature representations from the neck network and is responsible for mapping these features to classification labels, object detection bounding box coordinates, and confidence scores. Customizing the head network can optimize the output layer to better suit specific tasks and data characteristics, thereby improving the detector's detection accuracy and efficiency in classification and localization tasks.

In EFD-YOLOv4, Li, Kong et al. (2023) used a decoupled head to address classification-regression conflicts. Xia et al. (2023) incorporated a ConvMixer module in GCC-YOLO's prediction head to enhance the receptive field and small target detection while maintaining efficiency. Su, Zhou, Wan et al. (2023) applied ZCIoU in ICT-EDNet for improved regression loss and convergence. Chen, Feng et al. (2023) upgraded YOLOv5's bounding box regression with EIOU to enhance target recognition and detection performance. Zhou, Yang et al. (2023) employed a task-oriented decoupled head with local feature representation (LFR) and global feature representation (GFR) modules in ETDNet for effective defect classification and localization. Liu et al. (2024) added a box refinement module (BRM) to the head for better defect shape refinement. Li et al. (2024) introduced a regional attention module (RAM) in IDP-Net to improve localization accuracy by focusing on critical regions and added a MEIoU loss function for better scale detection. Lu, Wangqi et al. (2024) utilized WIoU loss in WSS-YOLO to boost localization accuracy and generalization. Xiao et al. (2024) replaced CIoU with Inner-CIoU to accelerate bounding box regression. These studies collectively focus on decoupling tasks, attention mechanisms, and advanced loss functions to improve task interaction, reduce conflicts, and enhance accuracy.

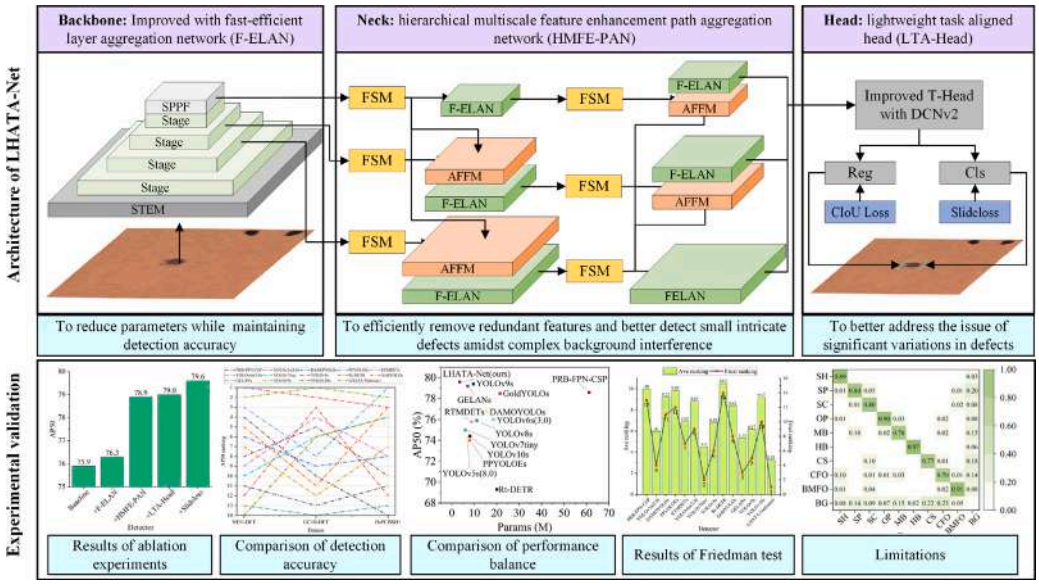


Fig. 2. Architecture of LHATA-Net along with its experimental validation.

Recent research indicates that replacing a coupled head for a decoupled one enhances regression accuracy in classification and localization tasks, though it does not completely address task disentanglement issues. Furthermore, refining the loss function can further boost both classification and localization accuracy. Therefore, this study introduces the decoupled LTA-Head, which enhances the interaction between classification and localization features, thereby reducing the impact of significant variations in defects. Additionally, we employs Slideloss to more effectively handle hard samples that are relatively sparse.

3. Methodology

3.1. General overview

The architecture of the proposed LHATA-Net along with its experimental validation are illustrated in Fig. 2. The LHATA-Net, using YOLOv8s as the baseline, consists of three parts: the backbone network, the neck network, and the head network. The backbone network features a multi-layer Stage pyramid structure, integrated with the STEM and SPPF modules from YOLOv8s. Each Stage primarily consists of the designed F-ELAN, which integrates GELAN, partial convolution (PConv), and pointwise convolution (PWConv) to extract features, thereby reducing parameters while ensuring detection accuracy. The neck network adopts the newly designed HMFE-PAN for feature fusion, with each path incorporating CAA-based feature selection module (FSM) or adaptive feature fusion module (AFFM). This approach effectively eliminates redundant features, enhances the detector's accuracy, and reduces the parameters. The head network introduces an innovative LTA-Head based on the decoupled head architecture. The LTA-Head combines TDM and DCNv2 to strengthen the semantic link between classification and localization, thereby better addressing the challenges posed by significant variations in defects. Additionally, the head network utilizes the Slideloss to address sample imbalances across different defect categories by incorporating the slide weighting function into the BCEWithLogitsLoss function. Building on this, ablation experiments will validate the effectiveness of the proposed improvements, while accuracy comparison, performance balance analysis, and Friedman test will confirm the superiority of LHATA-Net. The specific structures of these improvements will be detailed in Sections 3.2 through 3.5. The corresponding experiments and analyses will be discussed in Section 4.

3.2. Fast-efficient layer aggregation network (F-ELAN)

The structure of the Stage, with F-ELAN as the core, is shown in Fig. 3(a). The input feature $F_i \in \mathrm{R}^{C\times H\times W}$ of each Stage is passed through DownSample to obtain $F_d \in \mathrm{R}^{C\times H/2\times W/2}$ before being fed into F-ELAN. In Stage 1, this DownSample is achieved using a 3 × 3Conv with a stride of 2. For subsequent Stages (Stage 2 to Stage 4), the DownSample utilizes ADown, as shown in Fig. 4.

The proposed F-ELAN integrates the GELAN architecture (Wang, Yeh et al., 2024) with the PConv and PWConv structures from FasterNet (Chen, Kao et al., 2023). The GELAN architecture utilizes a design strategy known as "stack in computational block", which integrates CSPNet (Wang et al., 2020) and ELAN (Wang et al., 2022). Specifically, the F_D is split into two parts, $F_{l-1}^{(1)}$ and $F_{l-1}^{(2)}$, following 1 × 1Conv. The $F_{l-1}^{(1)}$ remains unchanged, while for the feature extraction of $F_{l-1}^{(2)}$, we employ one layer of F-RepCSP and two layers of F-RepCSP to obtain F_{l1} and F_{l2}. Subsequently, $F_{l-1}^{(2)}$, F_{l1} and F_{l2} are concatenated with $F_{l-1}^{(1)}$ to produce F_l. This concatenated feature is then transformed into the output feature F_o using a 1 × 1Conv.

F-RepCSP is constructed by incorporating FasterNetBlock into RepCSP (Wang, Bochkovskiy et al., 2023), as shown in Fig. 3(b). When feature F_i^{CSP} is input into F-RepCSP, it first passes through two parallel paths for channel transformation (1 × 1Conv) to generate $F_{l-1}^{CSP(1)}$ and $F_{l-1}^{CSP(2)}$. $F_{l-1}^{CSP(2)}$ undergoes feature extraction via RepConv and 3 × 3Conv, followed by residual connection to produce $F_{l-1}^{CSP(3)}$. Subsequently, $F_{l-1}^{CSP(3)}$ is concatenated with $F_{l-1}^{CSP(1)}$, undergoes channel transformation to obtain F_{l-1}^{CSP}, and then is fed into FasterNetBlock.

The FasterNetBlock, depicted in Fig. 3(c), divides input feature F_i^{FNB} along the channel into two parts in a 1:3 ratio to obtain $F_{l-1}^{FNB(1)} \in \mathrm{R}^{C/4\times H\times W}$ and $F_{l-1}^{FNB(2)} \in \mathrm{R}^{3C/4\times H\times W}$. The $F_{l-1}^{FNB(1)}$ undergoes 3 × 3Conv and is then concatenated with the $F_{l-1}^{FNB(2)}$ along the channel dimension to obtain $F_l^{FNB} \in \mathrm{R}^{C\times H\times W}$. The F_l^{FNB} is then processed by an MLP shown in Fig. 3(d), and its output features F_0^{MLB} are residual connected to the F_i^{FNB} to yield the F_0^{FNB}.

In conclusion, the F-ELAN design for each Stage employs the GELAN "stack in computational block" strategy, which organizes the network into smaller, stacked blocks. This structure allows for independent or hierarchical optimization of the blocks, facilitating the training of deeper networks while mitigating the gradient vanishing typically seen in very deep architectures. By enhancing gradient propagation, this approach allows the network to grow deeper without compromising

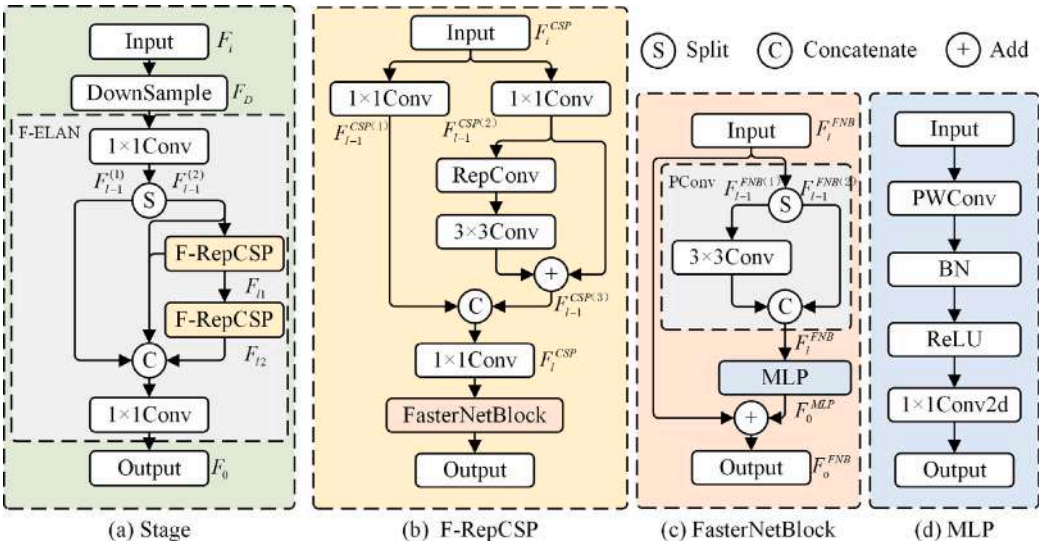


Fig. 3. Structure of each component in stage.

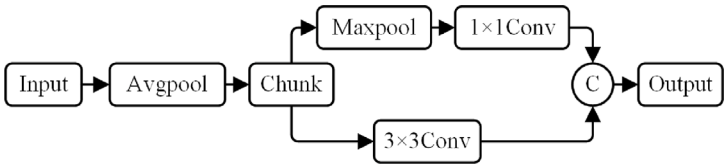


Fig. 4. Implementation process of adown.

detection accuracy. Additionally, the integration of PConv and PWConv within F-ELAN reduces kernel size and operates exclusively along the channel dimension, which decreases both the number of parameters and the computational burden. The synergy of these techniques allows the feature extraction module to minimize parameters while maintaining high detection accuracy.

3.3. Hierarchical multiscale feature enhancement path aggregation network (HMFE-PAN)

The multiscale feature fusion network combines features from different scales to capture comprehensive semantic information. Key networks include the feature pyramid network (FPN) (Lin et al., 2017), the path aggregation network (PAN) (Liu et al., 2018), and their variants, as shown in Fig. 5. FPN, with its top-down structure, merges high-level semantic features with low-level ones to improve shallow network semantics. PAN adds a bottom-up path to incorporate low-level location information into high-level features, boosting localization and classification. The hierarchical structure FPN (HS-FPN) (Chen et al., 2024) is an FPN variant designed to handle varying target sizes by filtering redundant features before cross-layer fusion, enhancing semantic integration across layers.

In industrial scenarios, surface defects often present numerous small features with significant variations in shape and scale, coupled with complex background interference. This can lead to the loss of small defects and object boundary details during feature fusion. HS-FPN handles small features and variations well but underutilizes shallow features. To improve this, we propose HMFE-PAN, as illustrated in Fig. 5(d). HMFE-PAN extends HS-FPN by integrating both top-down and bottom-up PAN structures. It features two "selection-fusion" stages: the first involves "selection-top-down feature fusion", while the second employs "selection-bottom-up feature fusion". FSM and AFFM are utilized for feature selection, and F-ELAN is applied for a lightweight representation of the output.

The structure of each component in HMFE-PAN is shown in Fig. 6. Each FSM employs the CAA module, initially introduced in PKINet (Cai et al., 2024), to extract multi-scale contextual information and enhance central feature. The CAA structure (Fig. 6(b)) processes input feature using F_i^{CAA} average pooling, followed by a 1 × 1Conv to obtain local feature F_o^{pool}. It then applies two inception-style 1 × 11 depthwise convolutions (DWConv) to capture long-range features. Finally, a 1 × 1Conv followed by Sigmoid is used to produce the weighted feature F_o^{CAA}. The two DWConv layers minimize extra parameters while enhancing feature extraction for elongated objects. This process is formulated as follows:

$$F_o^{pool} = Conv_{1\times1}\left(Avgpool(F_i^{CAA})\right) \tag{1}$$

$$F^{w} = DWConv_{1\times11}(F_o^{pool}) \tag{2}$$

$$F^{h} = DWConv_{11\times1}(F^{w}) \tag{3}$$

$$F_o^{CAA} = Sigmoid\left(Conv_{1\times1}(F^{h})\right) \tag{4}$$

In multi-scale features, high-level ones carry rich semantics, while low-level ones offer precise positioning. To align these across scales, this paper uses two AFFM types for cross-scale fusion (Fig. 6(a)). One type of AFFM uses DySample (Liu, Lu et al., 2023) for upsampling when fusing high-level features into low-level features, while another type uses ADown (Fig. 4) for downsampling when fusing low-level features into high-level features. Both sampling methods aim to perform rapid sampling and reduce feature loss during the fusion process.

The structure of DySample is illustrated in Fig. 6(c). The input feature $F_1^{Ds} \in \mathrm{R}^{C\times H\times W}$ undergoes a linear transformation to generate two features, F_2^{Ds} and F_3^{Ds}, of equal size $2\cdot 2^2\times H\times W$. The F_3^{Ds} is then activated by a Sigmoid function, scaled by 0.5, and multiplied by F_2^{Ds} to produce $F_4^{Ds} \in \mathrm{R}^{2\cdot 2^2\times H\times W}$. After F_4^{Ds} is processed through pixel shuffle (*PixShf*) (Shi et al., 2016), it is added to the original grid (*OrgGrid*) of the same size as $2\times 2H\times 2W$, resulting in a sampling set F_{ss}^{Ds}. Finally, the grid sample (*GridSmp*) function uses the positions from the sampling

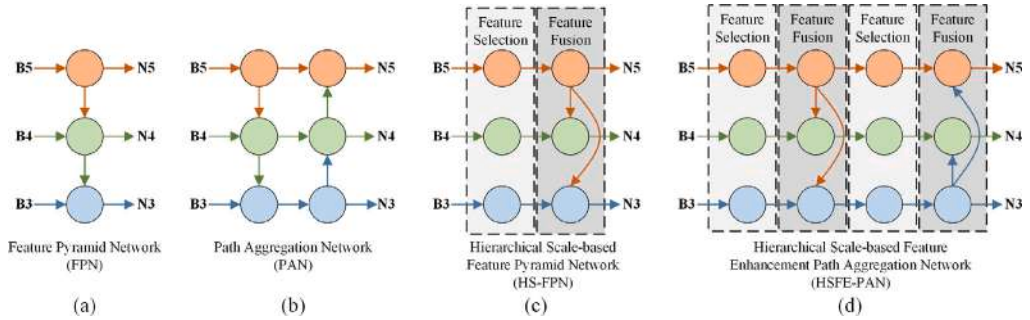


Fig. 5. Structures of FPN, PAN, HS-FPN, and HMFE-PAN.

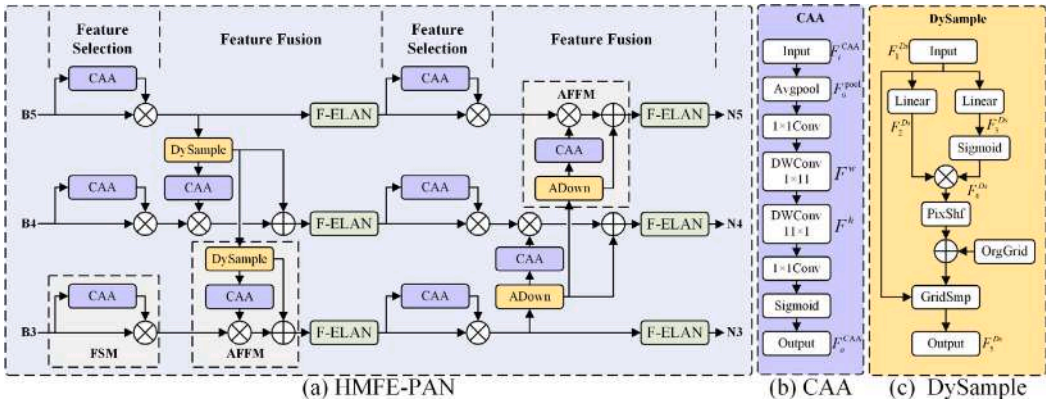


Fig. 6. Structure of each component in HMFE-PAN.

set to re-sample F_1^{Ds}, producing the output feature $F_5^{Ds} \in \mathrm{R}^{C\times 2H\times 2W}$. Its process is expressed as follows:

$$F_2^{Ds}, F_3^{Ds} = Linear(F_1^{Ds}) \in R^{2\cdot 2^2\times H\times W} \tag{5}$$

$$F_4^{Ds} = F_2^{Ds} + 0.5Sigmoid(F_3^{Ds}) \in R^{2\cdot 2^2\times H\times W} \tag{6}$$

$$F_{ss}^{Ds} = PixShf(F_4^{Ds}) + OrgGrid \tag{7}$$

$$F_5^{Ds} = GridSmp(F_1^{Ds}, F_{ss}^{Ds}) \tag{8}$$

Overall, HMFE-PAN enhances HS-FPN by adding a "selection-bottom-up feature fusion" to complement the "selection-top-down" in HS-FPN, linking high-resolution shallow features with semantically rich deep features to preserve spatial details for small object detection while retaining the essential semantics needed for classification and recognizing variations. Furthermore, HMFE-PAN incorporates the CAA during feature transmission for enhanced feature selection. This module applies two 1 × 11DWConvs in different directions in series, minimizing additional parameters while increasing the receptive field, thus facilitating the capture of long-range contextual information and supplementing local texture features extracted by multi-scale convolutional kernels. This approach improves feature extraction for elongated objects, reduces boundary feature loss, and mitigates background interference, resulting in enhanced detection of small and complex defects.

3.4. Lightweight task aligned head (LTA-Head)

Decoupled head is designed to optimize object localization and classification more effectively. However, industrial surface defect targets exhibit significant inter-class similarity and intra-class variation, potentially causing feature conflicts during classification and localization due to insufficient interaction. To address this issue, we propose the LTA-Head, which integrates TDM from the TOOD (Feng et al., 2021) and DCNv2 (Zhu et al., 2019) to strengthen semantic linkage between classification and localization, as illustrated in Fig. 7.

In the LTA-Head, a Convolutional Layer performs parameter-sharing convolution operations on features from different scales. Illustrated in the Convolutional Layer box of Fig. 7, this layer utilizes two concatenated 3 × 3Conv-Group Normal-SiLU (CGS) blocks to achieve parameter sharing across multiple scales. The outputs of these CGS blocks are merged along the channel dimension to enhance interaction between classification and localization tasks and reduce the number of parameters in the head. Subsequently, a TDM is utilized in conjunction with an Offset & Mask generator, followed by DCNv2, to perform the localization tasks. The TDM applies the task-aligned predictor (TAP) structure (Feng et al., 2021), with the distinction of incorporating group normal in the output to enhance task-specific performance through improved interaction features. The Offset & Mask Generator reduces input dimensionality via a 3 × 3Conv. It then directly generates a Mask for the first 18 channels and applies Sigmoid to produce Offsets based on the remaining channels. DCNv2 is used to adjust offsets in target localization and accommodate geometric variations in identified defects. Following this, the Conv_reg performs regression and outputs object boxes.

Simultaneously, another TDM combined with a feature dynamic selection module (FDSM) is used to deal with the classification task based on the output of Convolutional Layer. The FDSM dynamically selects features from the output of the Convolutional Layer to generate classification weights, which are then multiplied with the output feature of the TDM to enhance object classification. The FDSM process, as depicted by the FDSM box in Fig. 7, begins by reducing the dimensionality of input feature $F_{\mathrm{i}}^{\mathrm{FDSM}} \in \mathrm{R}^{C\times H\times W}$ using a 1 × 1Conv and a ReLU, generating preliminary selected features $F_m^{FDSM} \in R^{C/4\times H\times W}$. Subsequently, additional feature extraction and selection are performed using a 3 × 3Conv and a Sigmoid, which reduces the channel count to 1 and produces category prediction probabilities for each pixel. Based on Conv_cls, the category of the target is then determined. To resolve

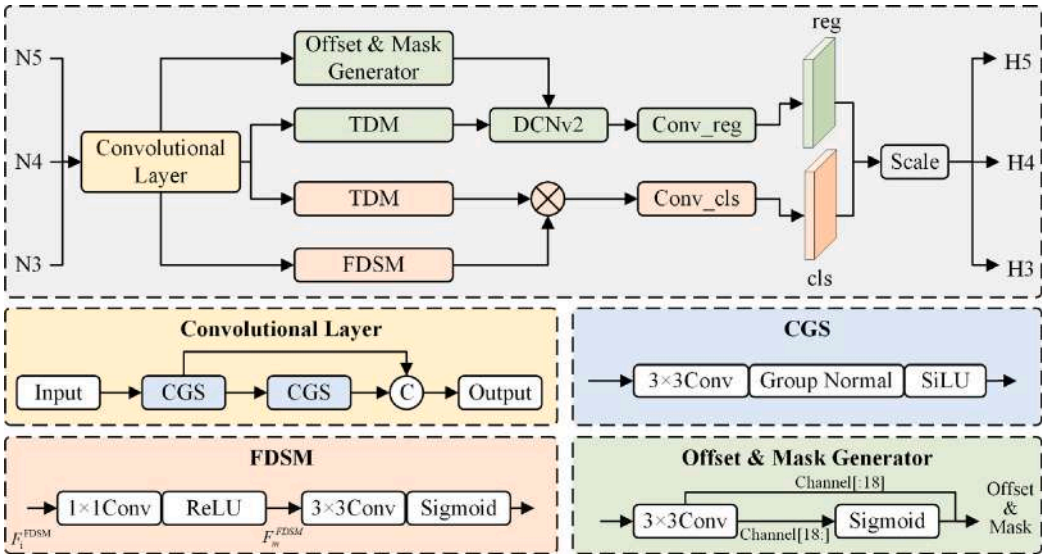


Fig. 7. Structure of each component in LTA-Head.

inconsistencies in object scales detected by each detection head, a Scale operator is applied to the outputs of Conv_reg and Conv_cls, adjusting feature scales to align with the input requirements of the LTA-Head.

The LTA-Head employs parameter-shared Convolutional Layer and two shared-feature TDMs, along with their respective specialized layers, to extract features for localization and classification tasks across different scales. In the localization branch, the DCNv2 utilizes deformable convolutions, incorporating the Offset & Mask of the Convolutional Layer's output features. This allows the network adjust the spatial distribution of sampling points, addressing geometric variations like object deformations. The classification branch's specialized layer employs FDSM to dynamically select features from the Convolutional Layer's output, generating classification weights. This enhances the classification feature representation in the TDM by filtering out less important features and concentrating on the most relevant and informative ones for each object instance. As a result, the network adapts more effectively to various object categories and variations.

3.5. Loss function

The loss function is used to quantify the discrepancy between the predicted values and the actual data. For the baseline YOLOv8s, the loss function is defined by the following expression.

$$L = \lambda_1 L_{loc} + \lambda_2 L_{cls} \tag{9}$$

where λ_1 and λ_2 represent the weight coefficients; L_{loc} and L_{cls} correspond to the bbox regression and classification losses, respectively.

The L_{loc} is computed through a combination of distribution focal loss (DFL) and CIoU loss functions. The L_{cls} is defined using BCEWithLogitsLoss as the follows:

$$L_{cls} = -\sum_{c=1}^{C} [y_c \log(x_c) + (1 - y_c) \log(1 - x_c)] \tag{10}$$

where c denotes a specific category to be detected, C is number of defects categories, x_c represents the predicted probability value of category c, y_c is the soft label where the label for the category c is determined by the IoU score, while the labels for all other categories are set to 0.

ISDD faces challenges due to sample imbalances across various defect categories. Training detectors on imbalanced samples can lead to biased feature learning, causing the detectors to prioritize categories with more available samples (Lv et al., 2024). To tackle this issue, we introduce slide weighting function (Yu, Huang et al., 2024) to adjust the IoU threshold μ for classifying various defect categories. Specifically, we computes the average IoU across all predicted bounding boxes as μ, categorizing values below μ as negative samples and those above it as positive samples. To effectively utilize defects with fewer samples during detector training, slide weighting function utilizes a weighting function that prioritizes these instances. This weighting function is shown in the following formula.

$$w(x) = \begin{cases} 1 & x \leqslant \mu - 0.1 \\ e^{1-\mu} & \mu - 0.1 < x < \mu \\ e^{1-x} & x \geqslant \mu \end{cases} \tag{11}$$

Building on this, Slideloss is developed for LHATA-Net by integrating the slide weighting function into BCEWithLogitsLoss, and can be formulated as follows:

$$L_{cls} = -\sum_{c=1}^{C} w(x_c) \cdot [y_c \cdot \log(x_c) + (1 - y_c) \cdot \log(1 - x_c)] \tag{12}$$

SlideLoss adaptively sets thresholds for positive and negative samples while increasing the relative weight of hard samples, effectively mitigating the imbalance between easy and hard samples and enhancing overall network performance. Additionally, the combination of LTA-Head and Slideloss addresses significant variations in defect sample distribution, shape, scale, and color.

Based on the approaches proposed above from Sections 3.1 to 3.5, the pseudocode of the proposed LHATA-Net is provided in Algorithm 1.

4. Experiments

4.1. Dataset description

This study evaluates the performance of LHATA-Net using three datasets: NEU-DET (He et al., 2020), GC10-DET (Lv et al., 2020), and DsPCBSD+ (Lv et al., 2024).

(1) NEU-DET: This dataset provides a publicly available resource for detecting surface defects on steel. It includes six defect types: Rolled-in scale (Rs), Pitted surface (Ps), Inclusion (In), Patches (Pa), Crazing (Cr), and Scratches (Sc), with a total of 1800 images at an original resolution of 200 × 200 pixels, and 300 images per defect type. For this study, the dataset is randomly divided into 1440 training images and 360 validation images, following an 80:20 ratio.

Algorithm 1 Pseudocode of the proposed LHATA-Net

Input: RGB images with surface defects

Output: The predicted bounding boxes and corresponding classification results of defects

Preprocess the input image (resize, normalize)
F-ELAN integrates a STEM, four Stages and a SPPF to extract multi-scale feature maps $B = \{B3, B4, B5\}$
HMFE-FPN leverages the CAA module to select and fuse $\{B3, B4, B5\}$, producing the output feature maps $N = \{N3, N4, N5\}$
for $i = 0$; $i < 3$; $i + 1$ **do**
    The T-Head, enhanced with DCNv2, processes $N(i+3)$ to generate the localization and classification branch feature maps, $Bl(i+3)$ and $Bc(i+3)$
    The Conv_reg and Conv_cls of the LTA-Head perform regression and compression on $Bl(i+3)$ and $Bc(i+3)$, respectively, to generate the feature map $H(i+3)$
    Obtain the predicted bounding box P_{bbox}, and predicted classification score P_{cls}
    Calculate $Lhead_{loc}$ (CIoU and DFL loss) based on P_{bbox}, and compute $Lhead_{cls}$ (Slideloss) according to Eq. (12) based on P_{cls}
    $L_{loc} = L_{loc} + L_{headloc}$
    $L_{cls} = L_{cls} + L_{headcls}$
end for
Calculate total loss L according to Eq. (9)
Apply NMS to the prediction boxes for refinement
Get the final bounding boxes and corresponding classification scores

(2) GC10-DET: This dataset, collected from real industrial scenarios, is a public repository of steel plate surface defects, encompassing 10 defect types: Welding line (Wl), Punching hole (Pu), Water spot (Ws), Oil spot (Os), Inclusion (In), Silk spot (Ss), Crescent gap (Cg), Rolled pit (Rp), Waist folding (Wf), and Crease (Cr), totaling 2294 images with an original resolution of 2048 × 100. In this study, the dataset is randomly divided into training and validation sets with an 80:20 ratio, comprising 1835 images for training and 459 images for validation.

(3) DsPCBSD+: This custom-compiled PCBs surface defect dataset comprises nine distinct defect categories: Short (SH), Spur (SP), Spurious copper (SC), Open (OP), Mouse bite (MB), Hole breakout (HB), Conductor scratch (CS), Conductor foreign object (CFO), and Base material foreign object (BMFO). The dataset contains 10,259 images (each 226 × 226 pixels) captured from actual PCB inner and outer layers post-etching via an AOI. These images are annotated with a total of 20,276 defect instances, exhibiting diversity in size, shape, color, and quantity, often blending with background colors or textures. The dataset is divided into 8,208 training images and 2,051 validation images, maintaining an 80:20 ratio. Table 1 presents the distribution of large, medium, and small defects across different categories and provides sample partitioning for each category. Five-fold cross-validation experiments were conducted using Co-DETR and YOLOv6-L6. Results indicated minimal impact of varying partitioning schemes on detection performance (Lv et al., 2024), confirming the representativeness of the samples, the rationality of the dataset distribution, and the high reliability of the dataset. This facilitates the training of robust and widely adaptable models. Notably, unlike other industrial surface defect datasets, a single manually identified defect in DsPCBSD+ may affect multiple PCBs components. These defects are further segmented into multiple distinct defects as shown in Fig. 8. Defects impacting multiple components are referred to as composite defects. Individual defects within a composite defect can present challenges such as dense distribution of many small defects, category occlusion or overlap, making it difficult for detectors to differentiate, accurately label, and bound each target. This can lead to missed or incorrect detections.

Table 1
Distribution of defect sizes and sample partitioning for each defect category.

Categories	Large	Medium	Small	All	Training set	Validation set
SH	0	205	710	915	746	169
SP	0	115	4469	4584	3655	929
SC	10	231	1352	1593	1308	285
OP	3	361	1406	1770	1432	338
MB	0	108	2421	2529	1983	546
HB	0	2848	35	2883	2275	608
CS	713	1043	734	2490	2042	448
CFO	110	582	1140	1832	1409	423
BMFO	68	304	1308	1680	1334	346
Total	904	5797	13 575	20 276	16 184	4092

Table 2
Hyperparameter settings of detectors in this study.

Parameters	Value	Note
Input images size	640 × 640	Input images size for training and validation
Batch	Train:64;Val:16	Number of images per batch
Device	Train:8;Val:1	Number of device
Epochs	300	Number of epochs to train
Optimizer	SGD	Optimizer to use
Weight decay	5e−4	Optimizer weight decay
Momentum	0.937	Optimizer momentum
Learning rate	1e−2 to 1e−4	Linear learning rate scheduler
Close mosaic	10	Close mosaic augmentation for final epochs
IoU	0.5	IoU threshold for NMS

4.2. Experimental environment and settings

The experimental environment for this study is based on the Ubuntu 20.04 64-bit operating system. The hardware configuration includes an Intel Xeon Gold 6242R CPU and a NVIDIA GeForce RTX 3090 GPU with 24 GB of VRAM. The experiments are performed using the PyTorch version 2.1.2 with CUDA version 12.2. All detectors in this study utilize the same hyperparameters. While fine-tuning hyperparameters is a challenging and under-explored area, the primary focus of this research is on developing a novel DL-based detector with a specifically designed architecture, rather than on hyperparameter optimization. Details of the hyperparameter settings of the involved DL-based detectors are provided in Table 2.

4.3. Evaluation metrics and evaluation method

Detectors in this study are analyzed from two perspectives: detection accuracy and computational efficiency. Accuracy is measured using COCO-format mAP, which includes metrics such as AP50, AP50:95, APs, APm, and APl. AP50 represents mAP at an IoU threshold of 0.5, while AP50:95 measures mAP across IoU thresholds from 0.5 to 0.95 in 0.05 intervals. APs, APm, and APl correspond to mAP for small (area $< 32^2$), medium ($32^2 <$ area $< 96^2$), and large (area $> 96^2$) objects, respectively. Note that APs, APm, and APl are calculated at IoU thresholds ranging from 0.5 to 0.95. In this study, AP50, AP50:95, APs, APm, and APl can be used not only to describe overall detection accuracy but also to detail the accuracy of specific category. AP50 is the primary metric for assessing detection accuracy. The computational efficiency of detectors is assessed by Params, floating point operations per second (FLOPs), and frames per second (FPS). Params indicate the size of the detector and are generally positively correlated with FLOPs. In resource-constrained scenarios, Params and FPS are particularly critical indicators of performance. Superior detection performance is reflected by a higher mAP, along with fewer Params, lower FLOPs, and higher FPS.

Detectors often struggle to balance detection accuracy, computational efficiency, and generalization across datasets. To thoroughly evaluate different detectors' overall performance, we employ the Friedman test (Derrac et al., 2011), a statistical method for multiple comparisons. This test assesses whether significant differences exist between

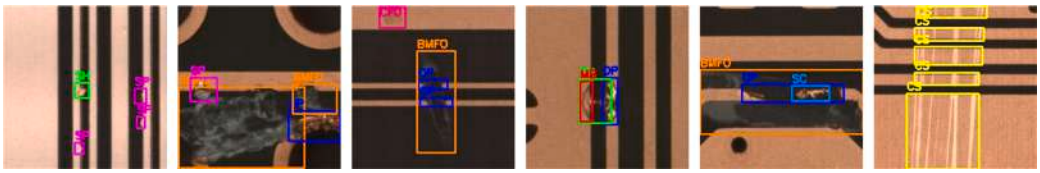


Fig. 8. Instances of defects consists of composite defect.

two or more algorithms. In this study, the Friedman test is applied to evaluate the performance of k detectors, considering accuracies across three datasets and computational efficiency metrics. In practice, different use cases may require emphasizing certain indicators over others. However, in the Friedman test, we treat accuracy, Params, FPS, and FLOPs as equally important to provide a more objective and comprehensive comparison of overall performance. The process is implemented as follows:

(1) **Data preparation**: Compute the AP50 for each detector across different datasets, as well as their Params, FLOPs, and FPS.

(2) **Rank detectors according to metrics**: Rank the detectors in descending order for AP50 and FPS, and in ascending order for Params and FLOPs. Record the ranking of the jth metric for the ith detector as $r_{ij}, 1 \le i \le k, 1 \le j \le 6$. Here, $r_{i1} \sim r_{i3}, 1 \le i \le k$ represent the rankings of the ith detector according to AP50 for NEU-DET, GC10-DET, and DsPCBSD+ datasets, respectively. Meanwhile, $r_{i4} \sim r_{i6}, 1 \le i \le k$ represent the rankings of the ith detector in terms of Params, FLOPs, and FPS respectively.

(3) **Average ranking computation**: Compute the average ranking for each detector using the following formula: $R_i = (\sum_{j=1}^{6} r_{ij})/6, i = 1, 2, \ldots .k$

(4) **Final ranking computation**: Determine the final rankings for the detectors based on their average rankings. Analysis of these results shows that a lower ranking corresponds to better performance.

(5) **Friedman test**

(1) At the significance level α, a one-sided test is conducted with the following hypotheses:

H_0: There is no significant difference in the average rankings of the detectors.

H_1: There are significant differences in the average rankings of the detectors.

(2) Calculate the Friedman statistic using the formula given below.

$$\chi^2 = 12n/[k(k+1)][\sum_{i=1}^{k}(R_i)^2 - k(k+1)^2/4] \tag{13}$$

where n represents the number of metrics (three AP50, FPS, Params and FLOPs), k represents the number of detectors, and R_i denotes the average ranking of the ith detector.

(3) Compare the computed statistic χ^2 with the critical value $\chi^2_{\alpha(k-1)}$ from the chi-square distribution table to determine if there is a significant difference. If $\chi^2 > \chi^2_{\alpha(k-1)}$, then reject H_0 conclude that the average rankings of the detectors involved in the comparison are significantly different; otherwise, accept H_0 and show that the average rankings of the detectors are not significantly different.

4.4. Ablation experiments

To comprehensively validate the effectiveness of improvements in LHATA-Net, we conduct ablation experiments on the NEU-DET dataset using YOLOv8s as the baseline detector (Baseline). We progressively integrate F-ELAN, HMFE-PAN, LTA-Head, and Slideloss into the Baseline and evaluate their performance. The results are detailed in Table 3.

Firstly, the results from the Baseline and +F-ELAN demonstrate that incorporating F-ELAN for feature extraction enhances AP50 by 0.4%, and reduces Params and FLOPs by 3.4M and 3.4G, respectively. This indicates that incorporating F-ELAN into the backbone not only benefits the development of a lightweight detector but also ensuring detection accuracy. The primary reason is that the integration of PConv and PWConv in F-ELAN reduces kernel size and operates along the channel dimension, which lowers both the number of parameters and computational load. Additionally, the "stack in computational block" strategy enhances gradient propagation, allowing the network to become deeper, thereby contributing to improved detection accuracy.

Secondly, comparing +F-ELAN with +HMFE-PAN reveals that AP50 and AP50:95 increase significantly by 2.6% and 2.28%. This improvement likely results from HMFE-PAN's enhancement of HS-FPN, which links high-resolution, shallow features with semantically richer, lower-resolution deep features. This design preserves essential spatial details required for detecting small objects while maintaining the semantic depth required for classification and variation recognition. Additionally, HMFE-PAN integrates CAA to optimize feature selection, improving the extraction of elongated object features, minimizing boundary feature loss, and reducing background interference—ultimately enhancing detection of small and complex defects. At the same time, Params and FLOPs decrease by 3.2M and 2.3G, respectively, due to the inclusion of F-ELAN in HMFE-PAN's Neck. Therefore, it can be concluded that HMFE-PAN not only enhances detection accuracy but also further reduces the detector's Params and FLOPs.

Thirdly, transitioning from +HMFE-PAN to +LTA-Head yields further improvements in accuracy metrics, with AP50 and AP50:95 increasing by 0.1% and 0.16%, respectively. This improvement likely results from integrating DCNv2 and FDSM into the LTA-Head. For localization, the LTA-Head utilizes DCNv2 to handle geometric variations, such as object deformations, while for classification, it employs FDSM to enhance feature representation, enabling better adaptation to diverse object categories and variations. This integration strengthens the interaction between classification and localization, improving the detector's overall performance. Additionally, Params and FLOPs are reduced by 1.0M and 4.3G, respectively, mainly due to the parameter-sharing mechanism in the LTA-Head's Convolutional Layer, which minimizes the need for excessive independent convolution operations for each task, thus reducing computational load and Params.

Lastly, using +Slideloss, LHATA-Net achieves 79.6% in AP50 and 42.66% in AP50:95, with 3.5M Params and 18.4G FLOPs. These results show an increase of 3.7% in AP50 and 1.88% in AP50:95, while reducing Params by 68.5% and FLOPs by 35.6% compared to the Baseline. Although LHATA-Net's FPS decreases from 210.3 to 54.2, it remains well within the current industrial standards for real-time defect detection in steel strips, which range from 18 to 83 FPS (Liu, Zhang et al., 2023). Overall, LHATA-Net demonstrates high detection accuracy and computational efficiency.

Generally, precision and recall are conflicting metrics. Thus, achieving higher precision while maintaining larger recall rates indicates better performance. This means that the closer the PR curve is to the upper right corner, the greater the robustness of the proposed LHATA-Net. Fig. 9 illustrates the PR curves of +F-ELAN, +HMFE-PAN, +LTA-Head, and +Slideloss. It can be seen that each enhancement extends the area under the PR curve (AUC-PR), indicating improved overall detection accuracy across different thresholds, thereby demonstrating its effectiveness in enhancing detection accuracy under conditions of significantly reduced Params and FLOPs.

In the HMFE-PAN, various attention mechanisms can be utilized for feature fusion. To identify the most suitable attention mechanism for

Table 3
Results obtained on NEU-DET datasets in ablation study.

Detector	AP50 (%)	AP50:95 (%)	Params (M)	FLOPs (G)	FPS
Baseline	75.9	40.78	11.1	28.4	210.3
+F-ELAN	76.3 (+0.4)	40.19 (−0.59)	7.7 (−3.4)	25.0 (−3.4)	167.0 (−43.3)
+HMFE-PAN	78.9 (+2.6)	42.47 (+2.28)	4.5 (−3.2)	22.7 (−2.3)	69.2 (−97.8)
+LTA-Head	79.0 (+0.1)	42.63 (+0.16)	3.5 (−1.0)	18.4 (−4.3)	53.9 (−15.3)
+Slideloss	79.6 (+0.6)	42.66 (+0.03)	3.5 (–)	18.4 (–)	54.2 (+0.3)
Total	+3.7	+1.88	−7.6	−10.0	−156.1

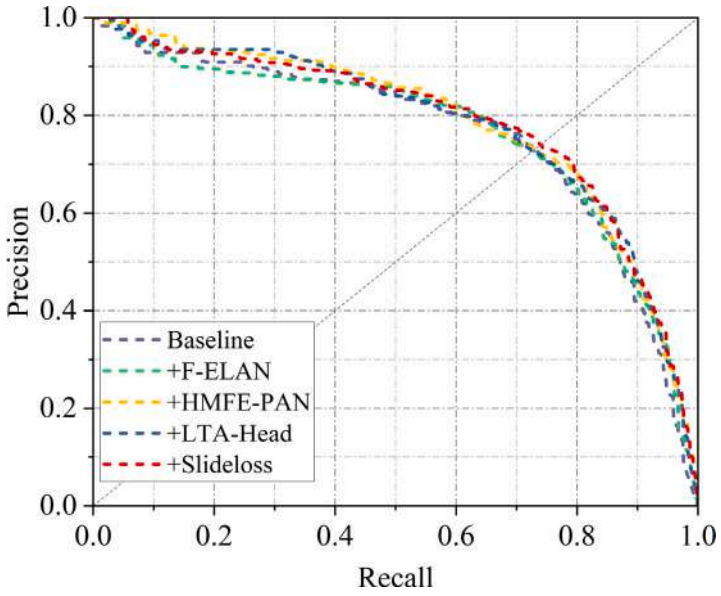


Fig. 9. PR curves of the enhancements in ablation study.

Table 4
Results obtained by different attention mechanisms in HMFE-PAN.

Metric	CA	BAM	ESE	CoordA	MLCA	ELA	CAA
AP50 (%)	78.6	76.6	78.0	78.4	76.8	76.6	**78.9**
AP50:90 (%)	39.4	38.6	**42.7**	42.3	39.2	38.5	42.5

Table 5
Error results obtained in ablation study.

Detector	Cls (%)	Loc (%)	Both (%)	Dupe (%)	Bkgd (%)	Miss (%)
Baseline	0.08	14.71	0.20	0.18	2.23	1.21
+F-ELAN	0.11	15.37	0.19	0.24	2.01	0.47
+HMFE-PAN	0.18	12.90	0.17	0.15	1.99	0.20
+LTA-Head	0.11	11.97	0.19	0.21	2.13	0.29
+Slideloss	0.06	12.17	0.15	0.26	2.09	0.08
Total	−0.02	−2.54	−0.05	+0.08	−0.14	−1.13

the HMFE-PAN, we compare channel attention (CA), bottleneck attention module (BAM) (Park et al., 2018), effective squeeze-and-excitation (ESE) (Lee & Park, 2020), coordinate attention (CoordA) (Hou et al., 2021), mixed local channel attention (MLCA) (Wan et al., 2023), efficient local attention (ELA) (Xu & Wan, 2024), and CAA (Cai et al., 2024). As shown in Table 4, CAA achieves the highest AP50 value at 78.9% and the second-highest AP50:90 value at 42.5%, due to its exceptional capability in capturing contextual information, expanding the receptive field, and enhancing feature interactions. Therefore, CAA is selected for FSM and AFFM in HMFE-PAN.

Additionally, we conduct an error analysis in the ablation study using the TIDE toolbox (Bolya et al., 2020). The results are detailed in Table 5. Overall, LHATA-Net shows reductions in classification error (Cls), localization error (Loc), combined classification and localization errors (Both), background error (Bkgd), and missing detection error (Miss), with only a 0.08% increase in duplicate prediction error (Dupe) compared to the Baseline. This indicates improved performance in defect classification and localization. Notably, Loc and Miss errors decrease by 2.54% and 1.13%, respectively, reflecting more precise localization and a significant reduction in missed detections. These improvements are crucial for defect detection in real production, as reducing missed detections and localization errors enhances product quality and minimizes the release of defective products.

Fig. 10 illustrates input images and output heat maps of each improvement (F-ELAN, HMFE-PAN, LTA-Head, Slideloss), visually demonstrating the effectiveness of the proposed enhancements. It is evident that the F-ELAN reduces the weights of specific defects through its lightweight design. HMFE-PAN effectively boosts weights in defect regions while suppressing background weights, thereby focusing more on defect areas. The subsequent integration of LTA-Head highlights the shape details of defects, which is advantageous for handling the diversity of defect sizes and shapes, facilitating more accurate classification and localization of various defects. Lastly, the Slideloss partly enhances defect weights despite background interference, enabling the detector to better differentiate between background elements and defects. HMFE-PAN, LTA-Head, and Slideloss also compensate for the feature weight reduction caused by F-ELAN. These visualizations further illustrate that the enhancements have successfully achieved their intended effects.

4.5. Comparative experiments

To validate the superiority of LHATA-Net, this paper compares it with 13 SOTA real-time detectors, including PRB-FPN-CSP (Chen et al., 2021), YOLOv5s, DAMOYOLOs (Xu et al., 2023), PPYOLOEs (Xu et al., 2022), RTMDETs (Lyu et al., 2022), YOLOv6s (Li, Li et al., 2023), YOLOv7tiny (Wang, Bochkovskiy et al., 2023), YOLOv8s, Rt-DETR (Zhao, Lv et al., 2024), GoldYOLOs (Wang, He et al., 2023),

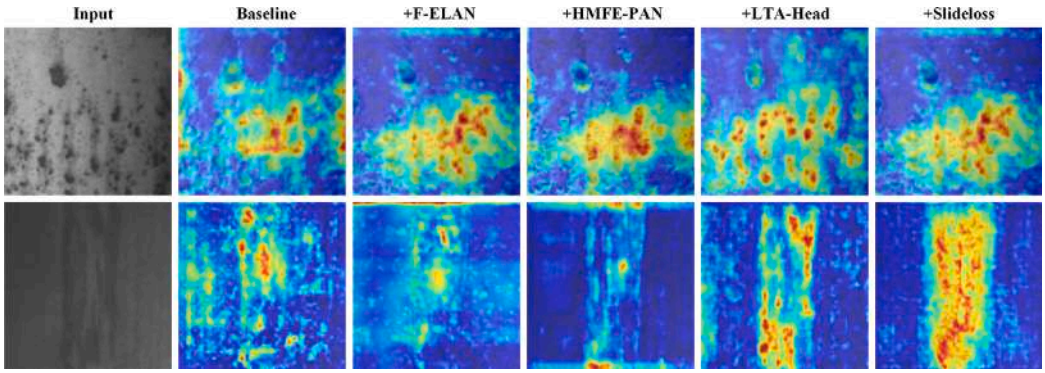


Fig. 10. Input images and output heat maps of each improvement in ablation study.

Table 6
Detection accuracy of different detectors on NEU-DET dataset.

Detector	AP50 (%)	AP50:95 (%)	AP50 of each defect category (%)					
			Rs	Ps	In	Pa	Cr	Cs
PRB-FPN-CSP	78.6	41.8	76.0	89.2	**77.7**	94.7	43.1	90.2
YOLOv5s(8.0)	74.2	38.8	67.8	86.8	70.1	89.9	37.0	93.2
DAMOYOLOs	76.8	39.8	70.7	81.7	74.5	**95.8**	44.8	92.7
PPYOLOEs	74.0	37.7	67.9	82.9	75.6	92.5	36.0	88.8
RTMDETs	75.8	43.0	72.6	83.8	73.9	92.8	37.8	93.7
YOLOv6s(3.0)	76.0	40.1	66.9	87.1	74.1	91.7	41.4	**94.5**
YOLOv7tiny	75.0	38.6	64.0	87.6	76.2	93.4	38.7	89.7
YOLOv8s	75.9	40.8	69.4	85.6	71.4	92.1	43.4	92.7
Rt-DETR	69.3	**44.2**	57.2	73.6	74.5	91.3	30.0	88.9
GoldYOLOs	78.5	40.0	73.4	**90.0**	74.5	92.0	47.8	92.8
GELANs	79.2	44.0	75.0	88.1	73.1	94.9	**50.8**	93.4
YOLOv9s	79.4	**44.2**	74.3	88.9	75.9	95.0	47.3	94.4
YOLOv10s	74.4	42.5	70.9	84.2	72.7	78.9	39.1	91.4
LHATA-Net(ours)	**79.6**	42.6	**78.8**	88.7	77.5	93.2	47.4	93.2

GELANs (Wang, Yeh et al., 2024), YOLOv9s (Wang, Yeh et al., 2024), and YOLOv10s (Wang, Chen et al., 2024) on the NEU-DET, GC10-DET and DsPCBSD+ datasets. All detectors use the same training parameters as LHATA-Net. Note that PPYOLOEs faced gradient vanishing issues on the GC10-DET dataset, leading to ineffective results.

4.5.1. Comparison of detection accuracy

The detection accuracy of different detectors on the NEU-DET and GC10-DET datasets is detailed in Tables 6 and 7. respectively. On the NEU-DET dataset, LHATA-Net achieves AP50 and AP50:90 scores of 79.6% and 42.6%, ranking highest in AP50 and fifth in AP50:90. On the GC10-DET dataset, LHATA-Net's AP50 and AP50:90 scores are 70.0% and 36.4%, respectively, securing second place out of 14 detectors. LHATA-Net also shows competitive AP50 scores for all defect categories, including challenging ones like Rs and Cr on the NEU-DET dataset, with scores of 78.8% and 47.4% respectively. This highlights LHATA-Net's effectiveness in detecting complex defects and its balanced performance across different categories in industrial surface detection.

Table 8 compares the performance of LHATA-Net on the DsPCBSD+ dataset with the 13 SOTA detectors. LHATA-Net achieves AP50, AP50:95, APs, APm, and APl scores of 85.8%, 52.2%, 42.8%, 57.2%, and 69.8%, respectively, ranking 3rd in AP50, 1st in AP50:95, and 2nd in APl. Table 9 details the AP50 scores for each category from various detectors on the DsPCBSD+ dataset, showing that LHATA-Net consistently delivers high AP50 values across all categories, maintaining robust detection performance. While certain detectors excel in specific categories with high AP50 scores, there are significant disparities between categories. For example, YOLOv7tiny, which achieves the same AP50 as LHATA-Net, shows similar AP50 values to LHATA-Net across the 9 categories. However, in the category with large size differences, CS, YOLOv7tiny's AP50 value is only 65.3%, significantly lower than LHATA-Net's result of 76.5%.

To offer a clearer comparison of detection accuracy across the three datasets, we perform visualization comparisons based on the quantitative results presented in Tables 6 to 9. Fig. 11 shows the AP50 rankings achieved by different detectors on the NEU-DET, GC10-DET, and DsPCBSD+ datasets. LHATA-Net ranks 1st, 2nd, and 3rd on these datasets, respectively. Among these detectors, only YOLOv9s has comparable AP50 rankings to our LHATA-Net on all three datasets (ranking 2nd, 1st, and 2nd, respectively). Other detectors either show lower mAP performance across the three datasets (e.g., PPYOLOEs and YOLOv10s,) or exhibit significant variability in their rankings. For instance, GELANs achieves the highest AP50 on the DsPCBSD+ dataset but ranks 8th on the GC10-DET dataset. GoldYOLOs also ranks 3rd on the DsPCBSD+ dataset but ranks 5th and 11th on the NEU-DET and GC10-DET datasets, respectively. YOLOv5s ranks 3rd on the GC10-DET dataset but ranks 12th on both NEU-DET and DsPCBSD+ datasets. These results underscore the high detection accuracy and robust generalization capabilities of LHATA-Net across diverse datasets. Fig. 12 shows the PR curves for various detectors on the DsPCBSD+ dataset. LHATA-Net distinguishes itself with the second-highest AUC-PR among all SOTA detectors, underscoring its competitive performance.

The DsPCBSD+ dataset was used as a case study to assess LHATA-Net's detection accuracy across various defect categories and sizes, with a focus on its improvements over the baseline detector, YOLOv8s. Table 10 presents the accuracy results of LHATA-Net and the Baseline across different defect categories on the DsPCBSD+ dataset. The corresponding results show that LHATA-Net demonstrates significant improvements over the Baseline. Specifically, AP50 demonstrates enhancements ranging from 0.9% to 3.5% in the SC, MB, CS, and BMFO categories. Notably, for defects in the CS and BMFO categories, which show considerable variation in size, shape, or color, the AP50 improvements are 3.5% and 3.2%, respectively. These gains are likely attributable to the LTA-Head, which strengthens the semantic connection between classification and localization within the network head, allowing for better handling of intra-class differences caused by such variations. Excluding the HB category, LHATA-Net achieves overall AP50:90 improvements ranging from 0.4% to 2.2% across the remaining eight categories.

The results in Table 10 also show that LHATA-Net achieves improvements in APs ranging from 0.1% to 2.8% across nine categories, excluding HB due to the very low proportion of small HB defects in the validation set (7 out of 608). Notably, in the OP and CFO categories, which are characterized by a high proportion of complex small defects (262 out of 338 and 263 out of 423, respectively), LHATA-Net improves APs by 2.1% and 2.8% compared to the Baseline. This improvement is likely due to the effectiveness of the proposed HMFE-PAN, which enhances the detection of small, intricate defects amidst complex background interference. Additionally, APl shows substantial improvements of 39.9%, 4.3%, and 5.4% in the SC, CS, and CFO

Table 7
Detection accuracy of different detectors on GC10-DET dataset.

Detector	AP50 (%)	AP50:95 (%)	AP50 of each defect category (%)									
			Wl	Pu	Ws	Os	In	Ss	Cg	Rp	Wf	Cr
PRB-FPN-CSP	65.5	32.4	69.2	93.9	81.2	56.3	43.6	59.4	95.4	29.9	91.9	34.6
YOLOv5s(8.0)	68.9	34.9	96.2	91.9	78.1	55.5	35.9	56.7	95.7	32.4	96.9	49.3
DAMOYOLOs	66.6	33.9	**97.0**	91.6	71.2	54.2	37.6	61.4	96.6	15.1	98.7	43.2
PPYOLOEs	–	–	–	–	–	–	–	–	–	–	–	–
RTMDETs	68.0	37.3	94.2	93.3	82.8	57.9	36.2	59.8	**96.8**	17.7	84.4	57.7
YOLOv6s(3.0)	64.6	33.7	73.7	94.5	82.3	56.8	33.1	61.0	**96.8**	10.0	96.0	42.3
YOLOv7tiny	67.4	35.0	81.4	92.8	**85.6**	58.1	33.2	**64.7**	96.5	15.2	94.8	51.1
YOLOv8s	67.8	33.3	83.2	92.6	83.6	54.7	43.1	57.7	96.0	15.5	92.7	**58.7**
Rt-DETR	67.4	36.1	95.9	91.5	81.2	56.8	47.1	47.8	**96.8**	25.2	76.4	55.7
GoldYOLOs	65.3	32.4	60.4	95.2	80.6	55.4	40.6	60.4	96.7	8.1	**99.3**	57.1
GELANs	67.2	35.1	91.1	92.6	82.4	57.7	**48.0**	61.0	95.2	13.3	93.3	37.6
YOLOv9s	**70.4**	35.5	84.9	96.7	84.2	59.3	47.9	61.6	96.1	**36.8**	92.3	44.4
YOLOv10s	62.1	32.8	73.7	93.6	75.8	49.6	36.2	49.3	95.7	16.8	92.2	38.6
LHATA-Net(ours)	70.0	**36.4**	89.1	**97.3**	84.9	**63.2**	37.7	61.9	96.4	21.8	98.9	49.0

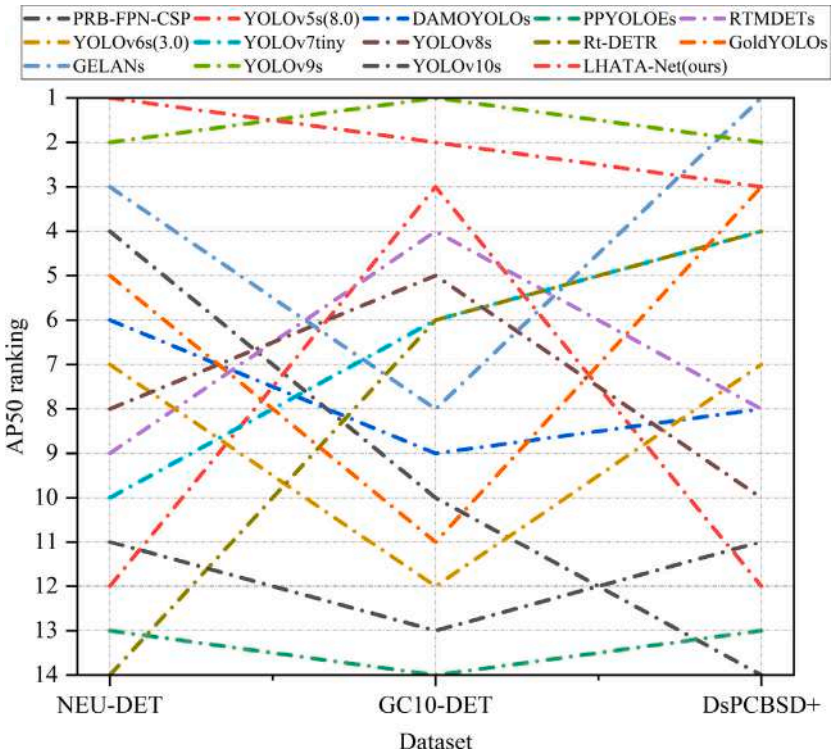


Fig. 11. AP50 rankings of various detectors across different datasets.

Table 8
Detection accuracy of different detectors on DsPCBSD+ dataset.

Detector	AP50 (%)	AP50:90 (%)	APs (%)	APm (%)	APl (%)
PRB-FPN-CSP	81.0	45.3	39.1	49.4	28.4
YOLOv5s(8.0)	84.3	48.3	42.9	53.1	41.5
DAMOYOLOs	84.8	48.5	40.8	54.6	50.1
PPYOLOEs	82.7	46.0	41.8	47.6	32.5
RTMDETs	84.8	48.6	40.1	55.9	64.3
YOLOv6s(3.0)	85.2	49.7	43.1	55.4	51.6
YOLOv7tiny	85.5	48.5	42.9	54.9	46.7
YOLOv8s	84.6	50.0	43.1	57.0	57.7
Rt-DETR	85.5	51.4	42.9	57.3	68.1
GoldYOLOs	85.8	50.0	42.3	55.3	49.7
GELANs	**86.5**	**52.2**	**45.4**	59.3	67.7
YOLOv9s	86.4	51.9	43.6	**59.4**	63.0
YOLOv10s	84.5	51.0	43.1	54.4	61.9
LHATA-Net(ours)	85.8	**52.2**	42.8	57.2	**69.8**

categories, respectively, while there is only a 1.4% decrease in the CFO category. The remaining five defect categories do not include large defects (areas $> 96^2$) in the validation set. In conclusion, LHATA-Net demonstrates superior accuracy and balanced performance across all categories compared to the Baseline, showcasing a strong capability in detecting small defects and those with significant variations in shape, size, and color.

Fig. 13 presents visual detection results from our LHATA-Net, the Baseline (YOLOv8s), and the top four YOLO detectors in AP50: YOLOv7tiny, GoldYOLOs, GELANs, and YOLOv9s, on the DsPCBSD+ dataset. The figure highlights that LHATA-Net effectively detects defects blend with the background or have blurred boundaries (e.g., 1st and 2nd rows). In contrast, other detectors show some inaccuracies; for instance, YOLOv8s and GoldYOLOs fail to identify part of BMFO that is similar to the background; YOLOv9s splits SC into SC and CFO, and YOLOv8s misclassifies SC as CFO (e.g., 2nd rows). LHATA-Net also excels in detecting small target defects like SP, OP, and MB (e.g., 3rd and 4th rows) with high accuracy, outperforming others that either miss detections or exhibit lower confidence. For defects with

Table 9
AP50 of each category obtained by different detectors on DsPCBSD+ dataset.

Detector	SH (%)	SP (%)	SC (%)	OP (%)	MB (%)	HB (%)	CS (%)	CFO (%)	BMFO (%)
PRB-FPN-CSP	84.7	81.3	81.9	88.1	80.3	97.2	61.4	66.9	87.2
YOLOv5s(8.0)	88.3	84.7	84.2	88.1	82.7	97.7	70.2	73.0	89.6
DAMOYOLOs	91.6	83.2	83.5	91.1	82.7	97.5	72.2	74.3	87.4
PPYOLOEs	87.7	82.6	81.3	88.5	80.8	97.3	66.9	70.8	88.5
RTMDETs	91.5	81.8	82.6	92.7	82.2	96.7	73.7	74.9	87.5
YOLOv6s(3.0)	91.4	84.0	83.8	91.4	83.7	97.9	73.9	72.0	89.0
YOLOv7tiny	89.6	83.8	**85.1**	91.8	85.0	97.7	65.3	73.6	88.9
YOLOv8s	91.0	84.4	83.0	90.4	81.2	**98.0**	73.3	73.2	86.8
Rt-DETR	**93.1**	85.5	84.5	91.2	83.9	97.5	73.2	73.1	87.9
GoldYOLOs	91.9	85.2	83.8	90.6	85.1	98.1	75.9	74.0	87.6
GELANs	90.1	**86.3**	81.9	**93.0**	**86.5**	**98.0**	78.0	**75.2**	89.6
YOLOv9s	91.8	85.4	84.1	90.7	83.9	97.9	**79.1**	74.1	**90.1**
YOLOv10s	91.1	84.6	83.1	88.2	83.3	97.7	71.9	72.2	88.5
LHATA-Net(ours)	90.9	84.4	83.9	90.3	83.8	**98.0**	76.8	73.0	90.0

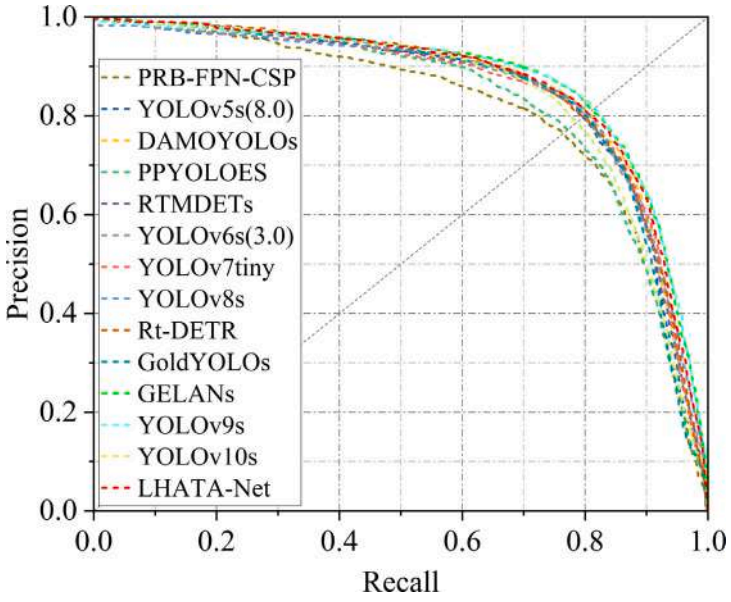


Fig. 12. PR curves of various detectors on DsPCBSD+ dataset.

Table 10
Results obtained by LHATA-Net and Baseline for each category on DsPCBSD+.

Detector	Category	AP50 (%)	AP50:90 (%)	APs (%)	APm (%)	APl (%)
Baseline	SH	91.0	57.6	58.5	55.6	–
	SP	84.4	37.3	36.5	63.7	–
	SC	83.0	50.4	47.2	74.0	55.1
	OP	90.4	53.4	51.4	59.5	–
	MB	81.2	39.1	40.1	33.7	–
	HB	98.0	82.7	48.2	83.0	–
	CS	73.3	43.7	27.3	43.0	60.3
	CFO	73.2	41.1	34.3	48.1	67.8
	BMFO	86.8	44.8	44.2	52.9	47.7
LHATA-Net	SH	90.9(−0.1)	58.0(+0.4)	58.6(+0.1)	57.7(+2.1)	–
	SP	84.4	38.2(+0.9)	37.5(+1.0)	60.5(−3.2)	–
	SC	83.9(+0.9)	51.0(+0.6)	47.4(+0.2)	70.4(−3.6)	95.0(+39.9)
	OP	90.3(−0.1)	54.2(+0.8)	53.5(+2.1)	58.1(−1.4)	–
	MB	83.8(+2.6)	39.8(+0.7)	40.4(+0.3)	30.6(−3.1)	–
	HB	98.0	82.5(−0.2)	37.4(−10.8)	82.9(−0.1)	–
	CS	76.8(+3.5)	47.5(3.8)	28.1(+0.8)	48.9(+5.9)	64.6(+4.3)
	CFO	73.0(−0.2)	42.5(1.4)	37.1(+2.8)	48.6(+0.5)	66.4(−1.4)
	BMFO	90.0(+3.2)	47.0(2.2)	44.9(+0.7)	57.3(+4.4)	53.1(+5.4)

varied sizes, shapes, and colors, such as CFO, LHATA-Net maintains high accuracy in both localization and classification. Some detectors, including GoldYOLOs, GELANs, and YOLOv9s, misclassify CFO as SC or mistake the background for defects (e.g., 5th and 6th rows). For composite defects where individual defects occlude or overlap (e.g., 7th and 8th rows), LHATA-Net detects them more completely, whereas other detectors like YOLOv7tiny, YOLOv8s, GoldYOLOs and GELANs show certain missed detections.

4.5.2. Comparison of performance balance

Conducting a balance analysis between detection accuracy and computational efficiency ensures that the detector can accurately identify

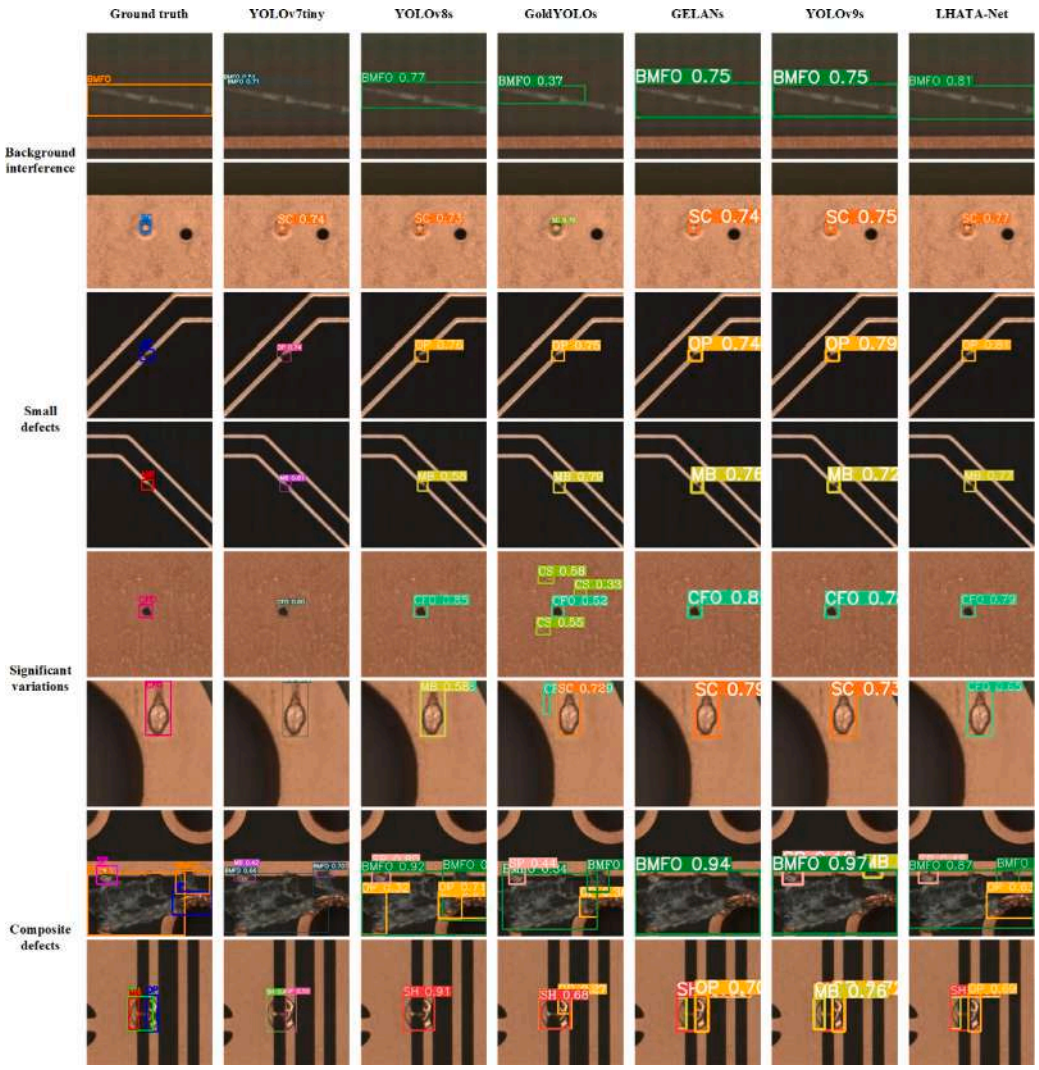


Fig. 13. Visualization of detection results of different detectors on DsPCBSD+ dataset.

Table 11
Computational efficiency of different detectors.

Detector	Backbone	Params (M)	FLOPs (G)	FPS	Detector	Backbone	Params (M)	FLOPs (G)	FPS
PRB-FPN-CSP	Darknet53	61.2	152.3	88.2	YOLOv8s	CSPDarknet	11.1	28.4	**210.3**
YOLOv5s(8.0)	CSPDarknet	7.0	15.8	144.3	Rt-DETR	Resnet18	20.1	60.0	35.0
DAMOYOLOs	MAE-Res	15.7	36.0	34.4	GoldYOLOs	EfficientRep	21.5	46.0	56.8
PPYOLOEs	CSRepResNet	7.9	17.3	43.2	GELANs	GELAN	7.1	26.2	43.8
RTMDETs	CSPDarknet	8.9	14.8	38.3	YOLOv9s	GELAN	9.6	38.7	32.2
YOLOv6s(3.0)	EfficientRep	18.5	45.2	59.3	YOLOv10s	CSPDarknet	8.0	24.5	47.7
YOLOv7tiny	E-ELAN	6.0	**13.1**	151.9	**LHATA-Net**	F-ELAN	**3.5**	18.4	54.2

defects while effectively utilizing computational resources. Table 11 presents the computational efficiency of various detectors. It shows that LHATA-Net has the fewest Params (3.5M), the 5th lowest FLOPs (18.4G), and the 7th fastest FPS (54.2) among the 14 detectors. Based on Tables 6 to 9 and Table 11, we further plot the AP50 distribution with respect to (w.r.t.) Params, FLOPs, and FPS for various detectors on the NEU-DET, GC10-DET, and DsPCBSD+ datasets, as shown in Fig. 14. Fig. 14 illustrates that LHATA-Net stands out at the top left in the AP50 w.r.t. Params and AP50 w.r.t. FLOPs plots, indicating superior accuracy with low Params and FLOPs.

From Table 11 and Fig. 14, it is evident that LHATA-Net not only achieves significantly better AP50 than DAMOYOLOs, YOLOv6s, Rt-DETR, and the latest YOLOv10s across the three datasets but also shows clear advantages in computational efficiency metrics such as Params, FLOPs, and FPS. Detectors with fewer FLOPs, including YOLOv5s, PPYOLOEs, RTMDETs, and YOLOv7tiny, show notably lower AP50 results across the datasets compared to LHATA-Net. Similarly, detectors with higher FPS, such as PRB-FPN-CSP, YOLOv5s, YOLOv6s, YOLOv7tiny, YOLOv8s, and GoldYOLOs, also achieve lower AP50 scores compared to LHATA-Net. The only exception is GoldYOLOs, which matches LHATA-Net's AP50 on the DsPCBSD+ dataset. Compared to YOLOv9s, which ranks higher than LHATA-Net on the GC10-DET and DsPCBSD+ datasets (1st and 2nd), LHATA-Net (2nd and 3rd) reduces Params and FLOPs by 63.5% and 52.5%, respectively, while increases FPS by 13.6%. When compared to GELANs, which ranks higher (2nd) than LHATA-Net on the DsPCBSD+ dataset, LHATA-Net reduces Params and FLOPs by 50.7% and 29.7%, respectively, and boosts FPS by 68.3%. Overall, LHATA-Net demonstrates highly competitive detection accuracy and computational efficiency validating its superiority.

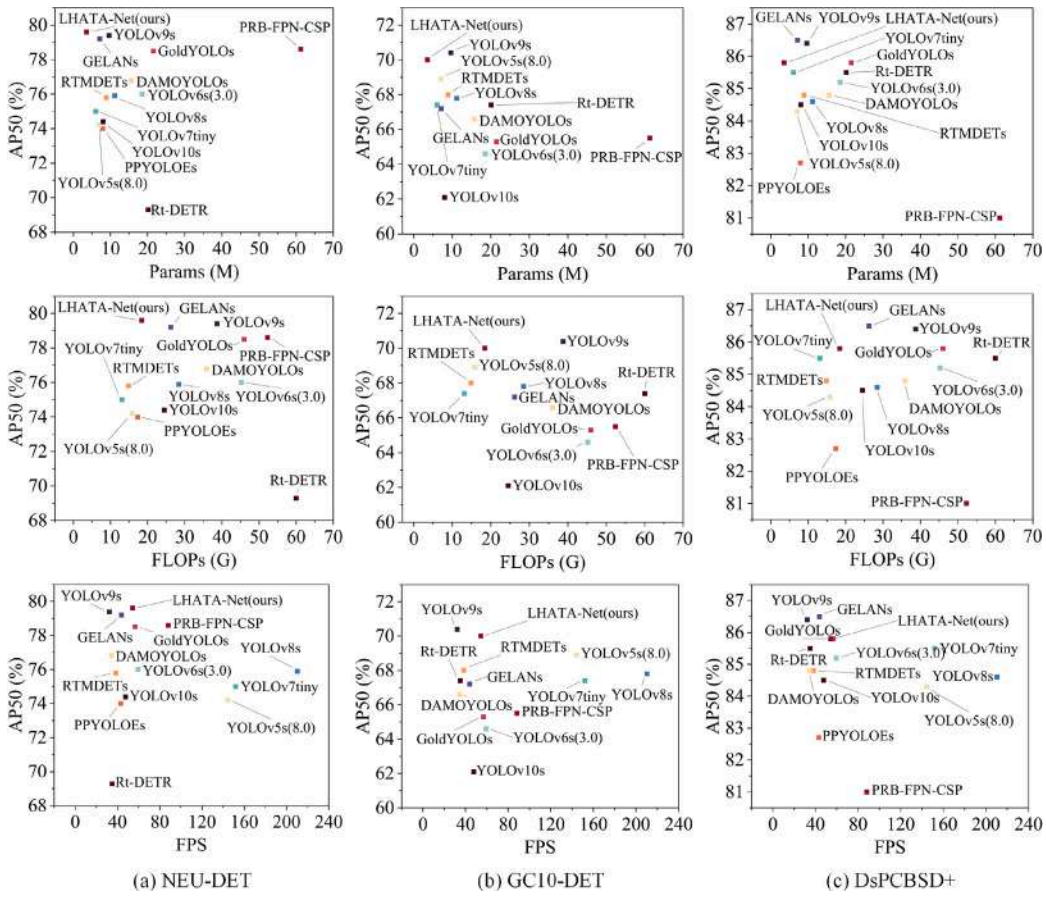


Fig. 14. Distribution of AP50 with respect to the Params, FLOPs and FPS of different detectors.

4.5.3. Friedman test

The Friedman test is used to compare the comprehensive performance of the 14 detectors by integrating the AP50 values on the NEU-DET, GC10-DET, and DsPCBSD+ datasets, along with the indicators Params, FLOPs, and FPS for each detector. It thoroughly evaluates the robustness of detection accuracy across different datasets and the computational efficiency of each detector. According to the steps of the Friedman test outlined in Section 4.3, the rankings of detectors according to various metrics are presented in Table 12. The average and final rankings of various detectors are depicted in Fig. 15. The results indicate that our LHATA-Net has the lowest final ranking value, signifying better comprehensive performance compared to other SOTA detectors.

The results of the Friedman statistic χ^2 are presented in Table 13. From Table 13, we observe that the degrees of freedom (df) is 13, and the computed value of χ^2 is 22.566. At a significance level α of 0.05, the critical value from the chi-square distribution table is 22.362. Since $\chi^2 = 22.566 > \chi^2_{\alpha(k-1)} = 22.362$, we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1). The Friedman test results indicate significant differences at $\alpha = 0.05$ in the average rankings obtained by the different detectors. It can be concluded that our LHATA-Net achieves a superior overall ranking compared to other detectors and shows a significant difference from them. It effectively addresses practical industrial requirements for rapid and high-accuracy surface defect detection in diverse scenarios under resource-constrained conditions.

Table 12
Rankings of detectors according to various metrics.

Detector	$AP50_N$	$AP50_G$	$AP50_D$	Params	FLOPs	FPS
PRB-FPN-CSP	4	10	14	14	14	4
YOLOv5s(8.0)	12	3	12	3	3	3
DAMOYOLOs	6	9	8.5	10	9	13
PPYOLOEs	13	14	13	5	4	10
RTMDETs	9	4	8.5	7	2	11
YOLOv6s(3.0)	7	12	7	11	11	5
YOLOv7tiny	10	6.5	5.5	2	**1**	2
YOLOv8s	8	5	10	9	8	**1**
Rt-DETR	14	6.5	5.5	12	13	12
GoldYOLOs	5	11	3.5	13	12	6
GELANs	3	8	**1**	4	7	9
YOLOv9s	2	**1**	2	8	10	14
YOLOv10s	11	13	11	6	6	8
LHATA-Net(ours)	**1**	2	3.5	**1**	5	7

Note: $AP50_N$, $AP50_G$, $AP50_D$ represent the ranking of AP50 for each detector on NEU-DET, GC10-DET, and DsPCBSD+ datasets, respectively.

Table 13
The results of Friedman statistic χ^2.

k	df	α	χ^2	$\chi^2_{\alpha(k-1)}$	H_0	H_1
14	13	0.05	22.566	22.362	×	✓

4.6. Limitations

Although the aforementioned analysis demonstrate that LHATA-Net achieves superior performance, LHATA-Net still has two main limitations. Firstly, as shown in Table 3, the FPS of LHATA-Net drops from 210.3 to 54.2 compared to the Baseline. This decrease is primarily due to the complex feature aggregation in F-ELAN and the integration of attention mechanisms in HMFE-NET and LTA-Head. To improve FPS, one potential approach is to apply module or channel pruning to the

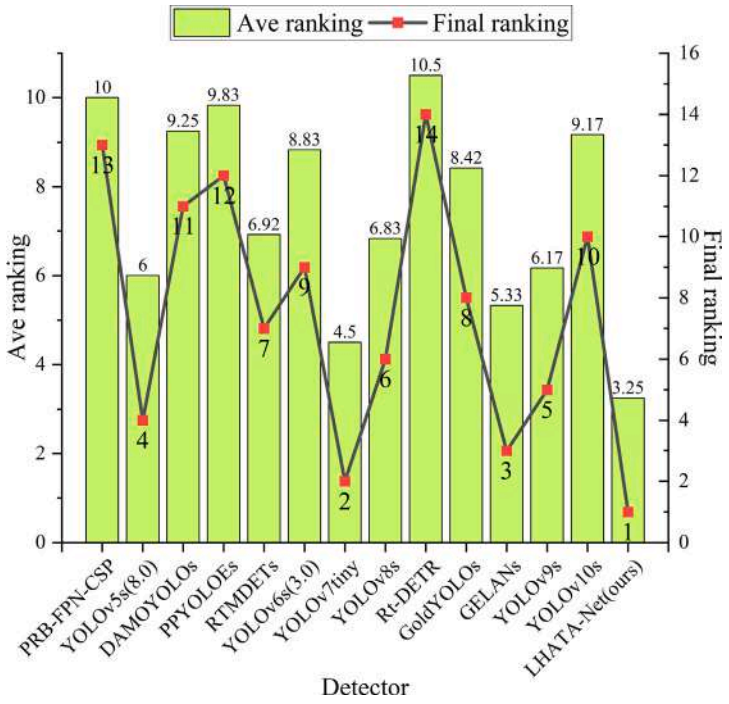


Fig. 15. Average and final rankings of various detectors.

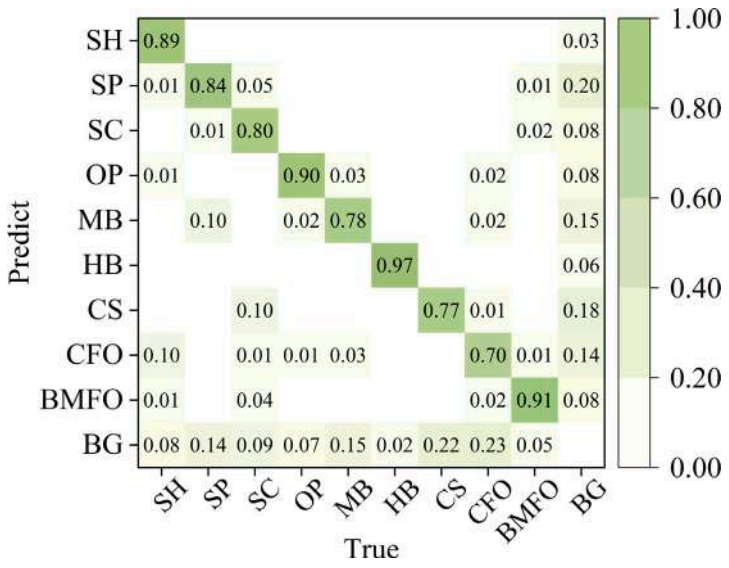


Fig. 16. Normalized confusion matrix of LHATA-Net including background (BG).

segments with complex feature aggregation. This involves selectively removing less critical paths or channels based on their contribution, thus reducing unnecessary computations. Another strategy is to replace attention mechanisms with efficient convolutional modules, which can further lower the detector's computational cost.

Second, despite its strong performance compared to other SOTA detectors, LHATA-Net still exhibits a high rate of missed detections for small defects, as shown in the last row of Fig. 16. Specifically, the miss rates for SP, MB, CS and CFO are 0.14, 0.15, 0.22 and 0.23, respectively. This issue is partly due to the loss of tiny defects during feature extraction, as evidenced in the 1st and 2nd columns of Fig. 17. Additionally, challenges such as dense distribution of defects, category occlusion or overlap contribute to missed detections of small defects in composite defects, as shown in the 3rd, 4th, and 5th columns of Fig. 17. These issues can be mitigated by employing techniques such as the occlusion-aware attention network and occlusion-aware repulsion loss.

5. Conclusions and future work

To address challenges like complex background interference, numerous small defects, and significant defect variations, this study introduces the LHATA-Net detector, which consists of four key components: F-ELAN, HMFE-PAN, LTA-Head, and Slideloss. The novel F-ELAN module reduces parameters while maintaining high detection accuracy. The proposed HMFE-PAN improves the detection of small defects amidst complex background interference while also reducing the parameters. The LTA-Head strengthens feature interaction between classification and localization, effectively addressing significant

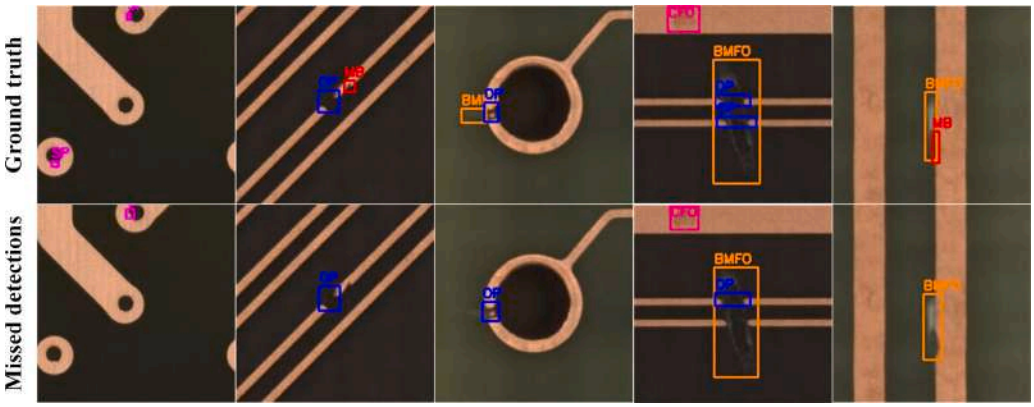


Fig. 17. Instances of missed detections by LHATA-Net.

variations in defects. Additionally, Slideloss tackles sample imbalance. Experiments are conducted on two public datasets, NEU-DET and GC10-DET, as well as a custom dataset from practical industrial production, DsPCBSD+. The results and conclusions are as follows:

(1) LHATA-Net achieves a Params and FLOPs of 3.5M and 18.4G, respectively, and achieves AP50 scores of 79.6%, 70.0%, and 85.8% on NEU-DET, GC10-DET, and DsPCBSD+ datasets, respectively. Compared to the Baseline, LHATA-Net shows reductions of 68.5% in Params and 35.6% in FLOPs, with corresponding improvements in AP50 of 3.7%, 2.2%, and 1.2% across the three datasets. Additionally, LHATA-Net operates at an inference speed of 54.2 FPS.

(2) Compared to other 13 SOTA real-time detectors, LHATA-Net has the lowest Params and ranks 1st, 2nd, and 3rd in AP50 across the three datasets, demonstrating high detection accuracy and robust generalization. The balance analysis further confirms that LHATA-Net delivers highly competitive detection accuracy and computational efficiency.

(3) The Friedman test results reveal that LHATA-Net achieves the lowest final ranking value compared to other SOTA detectors and shows a significant difference. This highlights LHATA-Net's superiority over other SOTA detectors in terms of its comprehensive performance when considering detection accuracy, generalization, and computational efficiency. It effectively meets practical industrial needs for rapid and high-accuracy surface defect detection in various scenarios under resource-constrained conditions.

Although the LHATA-Net performs well on the ISDD, further research is needed in several areas. These include developing higher-quality industrial surface defect datasets, designing detectors with better discriminative power, enhanced generalization capabilities, and faster detection speeds to meet the detection needs for various materials (such as fabrics, PCBs, steel, and aluminum) and types of defects; designing few-shot semi-supervised or even unsupervised detectors to reduce reliance on large-scale datasets; and improving validation in engineering applications.

CRediT authorship contribution statement

Shengping Lv: Investigation, Conceptualization, Data curation, Formal analysis, Methodology, Resources, Writing – original draft, Review & editing, Project administration, Supervision. **Tairan Liang:** Data curation, Methodology, Software, Validation, Visualization, Writing – original draft, Review & editing. **Kaibin Zhang:** Data curation, Validation, Formal analysis, Visualization. **Shixin Jiang:** Investigation, Conceptualization, Data curation, Review & editing, Platform supporter. **Bin Ouyang:** Data curation, Validation, Formal analysis, Visualization. **Quanzhou Li:** Investigation, Review & editing, Platform supporter. **Xiaoqing Li:** Data curation, Validation, Formal analysis, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 52275487), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2021A1515012395). The authors thank the anonymous reviewers for their valuable and constructive comments that greatly helped improve the quality and completeness of the paper.

Data availability

Data will be made available on request.

References

Ameri, R., Hsu, C.-C., & Band, S. S. (2024). A systematic review of deep learning approaches for surface defect detection in industrial applications. *Engineering Applications of Artificial Intelligence*, *130*, Article 107717. http://dx.doi.org/10.1016/j.engappai.2023.107717.

Baidu (2021). Aluminum defect dataset. https://aistudio.baidu.com/aistudio/projectdetail/3529511. (Accessed 26 April 2024).

Bolya, D., Foley, S., Hays, J., & Hoffman, J. (2020). TIDE: A general toolbox for identifying object detection errors. In *Computer vision - ECCV 2020* (pp. 558–573). Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-58580-8_33.

Cai, X., Lai, Q., Wang, Y., Wang, W., Sun, Z., & Yao, Y. (2024). Poly kernel inception network for remote sensing detection. arXiv preprint arXiv:2403.06258. https://arxiv.org/abs/2403.06258.

CCNUZFW (2023). PV-multi-defect dataset. https://github.com/CCNUZFW/PV-Multi-Defect. (Accessed 26 April 2024).

Chen, P.-Y., Chang, M.-C., Hsieh, J.-W., & Chen, Y.-S. (2021). Parallel residual bi-fusion feature pyramid network for accurate single-shot object detection. *IEEE Transactions on Image Processing*, *30*, 9099–9111. http://dx.doi.org/10.1109/TIP.2021.3118953.

Chen, Z., Feng, X., Liu, L., & Jia, Z. (2023). Surface defect detection of industrial components based on vision. *Scientific Reports*, *13*(1), 22136. http://dx.doi.org/10.1038/s41598-023-49359-9.

Chen, J., Kao, S.-h., He, H., Zhuo, W., Wen, S., Lee, C.-H., & Chan, S.-H. G. (2023). Run, don't walk: Chasing higher FLOPS for faster neural networks. In *2023 IEEE/CVF conference on computer vision and pattern recognition* CVPR, (pp. 12021–12031). http://dx.doi.org/10.1109/CVPR52729.2023.01157.

Chen, Y., Zhang, C., Chen, B., Huang, Y., Sun, Y., Wang, C., Fu, X., Dai, Y., Qin, F., Peng, Y., & Gao, Y. (2024). Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. *Computers in Biology and Medicine*, *170*, Article 107917. http://dx.doi.org/10.1016/j.compbiomed.2024.107917.

Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, *1*(1), 3–18. http://dx.doi.org/10.1016/j.swevo.2011.02.002.

Ding, R., Dai, L., Li, G., & Liu, H. (2019). TDD-net: A tiny defect detection network for printed circuit boards. *CAAI Transactions on Intelligence Technology*, *4*(2), 110–116, http://dx.doi.org/10.1049/trit.2019.0019.

Du, Y., Chen, H., Fu, Y., Zhu, J., & Zeng, H. (2024). AFF-Net: A strip steel surface defect detection network via adaptive focusing features. *IEEE Transactions on Instrumentation and Measurement*, *73*, 1–14. http://dx.doi.org/10.1109/TIM.2024.3398131.

Feng, C., Zhong, Y., Gao, Y., Scott, M. R., & Huang, W. (2021). TOOD: Task-aligned one-stage object detection. In *2021 IEEE/CVF international conference on computer vision* ICCV, (pp. 3490–3499). http://dx.doi.org/10.1109/ICCV48922.2021.00349.

Gao, Y., Li, X., Wang, X. V., Wang, L., & Gao, L. (2022). A review on recent advances in vision-based defect recognition towards industrial intelligence. *Journal of Manufacturing Systems*, *62*, 753–766. http://dx.doi.org/10.1016/j.jmsy.2021.05.008.

Guo, M., Xu, T., Liu, J., Liu, Z., Jiang, P., Mu, T., Zhang, S., Martin, R. R., Cheng, M., & Hu, S. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, *8*(3), 331–368. http://dx.doi.org/10.1007/s41095-022-0271-y.

He, Y., Song, K., Meng, Q., & Yan, Y. (2020). An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Transactions on Instrumentation and Measurement*, *69*(4), 1493–1504. http://dx.doi.org/10.1109/TIM.2019.2915404.

Hou, X., Liu, M., Zhang, S., Wei, P., & Chen, B. (2023). CANet: Contextual information and spatial attention based network for detecting small defects in manufacturing industry. *Pattern Recognition*, *140*, Article 109558. http://dx.doi.org/10.1016/j.patcog.2023.109558.

Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *2021 IEEE/CVF conference on computer vision and pattern recognition* CVPR, (pp. 13708–13717). http://dx.doi.org/10.1109/CVPR46437.2021.01350.

Huang, W., Peng, W., Zhang, M., & Liu, H. (2020). HRIPCB: A challenging dataset for PCB defects detection and classification. *The Journal of Engineering*, *2020*(13), 303–309. http://dx.doi.org/10.1049/joe.2019.1183.

Lee, Y., & Park, J. (2020). CenterMask: Real-time anchor-free instance segmentation. In *2020 IEEE/CVF conference on computer vision and pattern recognition* CVPR, (pp. 13903–13912). http://dx.doi.org/10.1109/CVPR42600.2020.01392.

Li, S., Kong, F., Wang, R., Luo, T., & Shi, Z. (2023). EFD-YOLOv4: A steel surface defect detection network with encoder–decoder residual block and feature alignment module. *Measurement*, *220*, Article 113359. http://dx.doi.org/10.1016/j.measurement.2023.113359.

Li, C., Li, L., Geng, Y., Jiang, H., Cheng, M., Zhang, B., Ke, Z., Xu, X., & Chu, X. (2023). YOLOv6 v3.0: A full-scale reloading. arXiv preprint arXiv:2301.05586. https://arxiv.org/abs/2301.05586.

Li, L., Wang, Z., & Zhang, T. (2023). GBH-YOLOv5: Ghost convolution with bottleneckcsp and tiny target prediction head incorporating YOLOv5 for PV panel defect detection. *Electronics*, *12*(3), http://dx.doi.org/10.3390/electronics12030561.

Li, G., Zhao, S., Li, M., Zhou, M., & Ying, Z. (2024). IDP-Net: Industrial defect perception network based on cross-layer semantic information guidance and context concentration enhancement. *Engineering Applications of Artificial Intelligence*, *130*, Article 107677. http://dx.doi.org/10.1016/j.engappai.2023.107677.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *2017 IEEE conference on computer vision and pattern recognition* CVPR, (pp. 936–944). http://dx.doi.org/10.1109/CVPR.2017.106.

Ling, Q., & Isa, N. A. M. (2023). Printed circuit board defect detection methods based on image processing, machine learning and deep learning: A survey. *IEEE Access*, *11*, 15921–15944. http://dx.doi.org/10.1109/ACCESS.2023.3245093.

Liu, Q., Liu, M., Jonathan, Q. M., & Shen, W. (2024). A real-time anchor-free defect detector with global and local feature enhancement for surface defect detection. *Expert Systems with Applications*, *246*, Article 123199. http://dx.doi.org/10.1016/j.eswa.2024.123199.

Liu, W., Lu, H., Fu, H., & Cao, Z. (2023). Learning to upsample by learning to sample. In *2023 IEEE/CVF international conference on computer vision* ICCV, (pp. 6004–6014). http://dx.doi.org/10.1109/ICCV51070.2023.00554.

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *2018 IEEE/CVF conference on computer vision and pattern recognition* CVPR, (pp. 8759–8768). http://dx.doi.org/10.1109/CVPR.2018.00913.

Liu, Y., Zhang, C., & Dong, X. (2023). A survey of real-time surface defect inspection methods based on deep learning. *Artificial Intelligence Review*, *56*(10), 12131–12170. http://dx.doi.org/10.1007/s10462-023-10475-7.

Lu, M., Wangqi, S., Zou, Y., Chen, Y., & Chen, Z. (2024). WSS-YOLO: An improved industrial defect detection network for steel surface defects. *Measurement*, *236*, Article 115060. http://dx.doi.org/10.1016/j.measurement.2024.115060.

Lu, J., Yu, M. M., & Liu, J. (2024). Lightweight strip steel defect detection algorithm based on improved YOLOv7. *Scientific Reports*, *14*(1), 13267. http://dx.doi.org/10.1038/s41598-024-64080-x.

Lv, X., Duan, F., Jiang, J., Fu, X., & Gan, L. (2020). Deep metallic surface defect detection: The new benchmark and detection network. *Sensors*, *20*(6), 1562. http://dx.doi.org/10.3390/s20061562.

Lv, S., Ouyang, B., Deng, Z., Liang, T., Jiang, S., Zhang, K., Chen, J., & Li, Z. (2024). A dataset for deep learning based detection of printed circuit board surface defect. *Scientific Data*, *11*(1), 811. http://dx.doi.org/10.1038/s41597-024-03656-8.

Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., & Chen, K. (2022). RTMDet: An empirical study of designing real-time object detectors. arXiv preprint arXiv:2212.07784. https://arxiv.org/abs/2212.07784.

Moganti, M., Ercal, F., Dagli, C. H., & Tsunekawa, S. (1996). Automatic PCB inspection algorithms: A survey. *Computer Vision and Image Understanding*, *63*(2), 287–313. http://dx.doi.org/10.1006/cviu.1996.0020.

Park, J., Woo, S., Lee, J.-Y., & Kweon, I. S. (2018). BAM: Bottleneck attention module. arXiv preprint arXiv:1807.06514. https://arxiv.org/abs/1807.06514.

Shao, R., Zhou, M., Li, M., Han, D., & Li, G. (2024). TD-Net: Tiny defect detection network for industrial products. *Complex & Intelligent Systems*, *10*(3), 3943–3954. http://dx.doi.org/10.1007/s40747-024-01362-x.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE conference on computer vision and pattern recognition* CVPR, (pp. 1874–1883). http://dx.doi.org/10.1109/CVPR.2016.207.

Su, B., Zhou, Z., & Chen, H. (2023). PVEL-AD: A large-scale open-world dataset for photovoltaic cell anomaly detection. *IEEE Transactions on Industrial Informatics*, *19*(1), 404–413. http://dx.doi.org/10.1109/TII.2022.3162846.

Su, Z., Zhou, M., Wan, H., Li, M., Zhang, Z., Han, D., Shao, R., & Li, G. (2023). Rethinking interactive networks and regression loss functions for industrial defect detection. *Journal of King Saud University - Computer and Information Sciences*, *35*(9), http://dx.doi.org/10.1016/j.jksuci.2023.101756.

Tang, S., He, F., Huang, X., & Yang, J. (2019). Online PCB defect detector on a new PCB defect dataset. arXiv preprint arXiv:1902.06197. https://arxiv.org/abs/1902.06197.

Tian, R., & Jia, M. (2022). DCC-CenterNet: A rapid detection method for steel surface defects. *Measurement*, *187*, Article 110211. http://dx.doi.org/10.1016/j.measurement.2021.110211.

Tianchi (2018). Aluminum profile surface detection database. Retrieved from https://tianchi.aliyun.com/dataset/148297 (Accessed 26 April 2024).

Tianchi (2020). Smart diagnosis of cloth flaw dataset. Retrieved from https://tianchi.aliyun.com/dataset/dataDetail?dataId=79336 (Accessed 26 April 2024).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Vol. 30*, In *Advances in neural information processing systems*. Curran Associates, Inc..

Wan, D., Lu, R., Shen, S., Xu, T., Lang, X., & Ren, Z. (2023). Mixed local channel attention for object detection. *Engineering Applications of Artificial Intelligence*, *123*, Article 106442. http://dx.doi.org/10.1016/j.engappai.2023.106442.

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *2023 IEEE/CVF conference on computer vision and pattern recognition* CVPR, (pp. 7464–7475). http://dx.doi.org/10.1109/CVPR52729.2023.00721.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458. https://arxiv.org/abs/2405.14458.

Wang, C., He, W., Nie, Y., Guo, J., Liu, C., Han, K., & Wang, Y. (2023). Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. arXiv preprint arXiv:2309.11331. https://arxiv.org/abs/2309.11331.

Wang, C.-Y., Liao, H.-Y. M., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., & Yeh, I.-H. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. In *2020 IEEE/CVF conference on computer vision and pattern recognition workshops* CVPRW, (pp. 1571–1580). http://dx.doi.org/10.1109/CVPRW50498.2020.00203.

Wang, C.-Y., Liao, H.-Y. M., & Yeh, I.-H. (2022). Designing network design strategies through gradient path analysis. arXiv preprint arXiv:2211.04800. https://arxiv.org/abs/2211.04800.

Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. arXiv preprint arXiv:2402.13616. https://arxiv.org/abs/2402.13616.

Wang, X., Zhang, Q., & Chen, C. (2024). Dual-branch information extraction and local attention anchor-free network for defect detection. *Scientific Reports*, *14*(1), 10886. http://dx.doi.org/10.1038/s41598-024-61324-8.

Workspace (2021). Defect detection 2 dataset. Retrieved from https://universe.roboflow.com/new-workspace-smiec/defect-detection-2 (Accessed 26 April 2024).

Xia, K., Lv, Z., Liu, K., Lu, Z., Zhou, C., Zhu, H., & Chen, X. (2023). Global contextual attention augmented YOLO with ConvMixer prediction heads for PCB surface defect detection. *Scientific Reports*, *13*(1), 9805. http://dx.doi.org/10.1038/s41598-023-36854-2.

Xiao, G., Hou, S., & Zhou, H. (2024). PCB defect detection algorithm based on CDI-YOLO. *Scientific Reports*, *14*(1), 7351. http://dx.doi.org/10.1038/s41598-024-57491-3.

Xu, X., Jiang, Y., Chen, W., Huang, Y., Zhang, Y., & Sun, X. (2023). DAMO-YOLO: A report on real-time object detection design. arXiv preprint arXiv:2211.15444. https://arxiv.org/abs/2211.15444.

Xu, W., & Wan, Y. (2024). ELA: Efficient local attention for deep convolutional neural networks. arXiv preprint arXiv:2403.01123. https://arxiv.org/abs/2403.01123.

Xu, S., Wang, X., Lv, W., Chang, Q., Cui, C., Deng, K., Wang, G., Dang, Q., Wei, S., Du, Y., & Lai, B. (2022). PP-YOLOE: An evolved version of YOLO. arXiv preprint arXiv:2203.16250. https://arxiv.org/abs/2203.16250.

Yu, Z., Huang, H., Chen, W., Su, Y., Liu, Y., & Wang, X. (2024). YOLO-FaceV2: A scale and occlusion aware face detector. *Pattern Recognition*, *155*, Article 110714. http://dx.doi.org/10.1016/j.patcog.2024.110714.

Yu, J., Wang, Y., Li, Q., Li, H., Ma, M., & Liu, P. (2024). Cascaded adaptive global localisation network for steel defect detection. *International Journal of Production Research*, *62*(13), 4884–4901. http://dx.doi.org/10.1080/00207543.2023.2281664.

Zeng, N., Wu, P., Wang, Z., Li, H., Liu, W., & Liu, X. (2022). A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–14. http://dx.doi.org/10.1109/TIM.2022.3153997.

Zhang, D., Hao, X., Wang, D., Qin, C., Zhao, B., Liang, L., & Liu, W. (2023). An efficient lightweight convolutional neural network for industrial surface defect detection. *Artificial Intelligence Review*, *56*(9), 10651–10677. http://dx.doi.org/10.1007/s10462-023-10438-y.

Zhang, H., Li, S., Miao, Q., Fang, R., Xue, S., Hu, Q., Hu, J., & Chan, S. (2024). Surface defect detection of hot rolled steel based on multi-scale feature fusion and attention mechanism residual block. *Scientific Reports*, *14*(1), 7671. http://dx.doi.org/10.1038/s41598-024-57990-3.

Zhang, Y., Zhang, H., Huang, Q., Han, Y., & Zhao, M. (2024). DsP-YOLO: An anchor-free network with DsPAN for small object detection of multiscale defects. *Expert Systems with Applications*, *241*, Article 122669. http://dx.doi.org/10.1016/j.eswa.2023.122669.

Zhang, Z., Zhou, M., Wan, H., Li, M., Li, G., & Han, D. (2023). IDD-Net: Industrial defect detection method based on deep learning. *Engineering Applications of Artificial Intelligence*, *123*, Article 106390. http://dx.doi.org/10.1016/j.engappai.2023.106390.

Zhao, B., Chen, Y., Jia, X., & Ma, T. (2024). Steel surface defect detection algorithm in complex background scenarios. *Measurement*, *237*, Article 115189. http://dx.doi.org/10.1016/j.measurement.2024.115189.

Zhao, S., Li, G., Zhou, M., & Li, M. (2023). ICA-Net: Industrial defect detection network based on convolutional attention guidance and aggregation of multiscale features. *Engineering Applications of Artificial Intelligence*, *126*, Article 107134. http://dx.doi.org/10.1016/j.engappai.2023.107134.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). DETRs beat YOLOs on real-time object detection. arXiv preprint arXiv:2304.08069. https://arxiv.org/abs/2304.08069.

Zheng, H., Chen, X., Cheng, H., Du, Y., & Jiang, Z. (2024). MD-YOLO: Surface defect detector for industrial complex environments. *Optics and Lasers in Engineering*, *178*, Article 108170. http://dx.doi.org/10.1016/j.optlaseng.2024.108170.

Zheng, Y., Lyu, W., Wang, C., Guo, Q., Zhou, D., & Xu, W. (2023). Efficient conflict-filtered network for defect detection. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–14. http://dx.doi.org/10.1109/TIM.2023.3293557.

Zhou, C., Lu, Z., Lv, Z., Meng, M., Tan, Y., Xia, K., Liu, K., & Zuo, H. (2023). Metal surface defect detection based on improved YOLOv5. *Scientific Reports*, *13*(1), 20803. http://dx.doi.org/10.1038/s41598-023-47716-2.

Zhou, H., Yang, R., Hu, R., Shu, C., Tang, X., & Li, X. (2023). Etdnet: Efficient transformer-based detection network for surface defect detection. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1–14. http://dx.doi.org/10.1109/TIM.2023.3307753.

Zhou, Y., Yuan, M., Zhang, J., Ding, G., & Qin, S. (2023). Review of vision-based defect detection research and its perspectives for printed circuit board. *Journal of Manufacturing Systems*, *70*, 557–578. http://dx.doi.org/10.1016/j.jmsy.2023.08.019.

Zhu, X., Hu, H., Lin, S., & Dai, J. (2019). Deformable ConvNets V2: More deformable, better results. In *2019 IEEE/CVF conference on computer vision and pattern recognition* CVPR, (pp. 9300–9308). http://dx.doi.org/10.1109/CVPR.2019.00953.

Zhu, W., Zhang, H., Zhang, C., Zhu, X., Guan, Z., & Jia, J. (2023). Surface defect detection and classification of steel using an efficient swin transformer. *Advanced Engineering Informatics*, *57*, Article 102061. http://dx.doi.org/10.1016/j.aei.2023.102061.

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A dataset for deep learning based detection of printed circuit board surface defect

Shengping Lv[1]✉, Bin Ouyang[1], Zhihua Deng[2], Tairan Liang[1], Shixin Jiang[3], Kaibin Zhang[1], Jianyu Chen[1] & Zhuohui Li[1]

Printed circuit board (PCB) may display diverse surface defects in manufacturing. These defects not only influence aesthetics but can also affect the performance of the PCB and potentially damage the entire board. Thus, achieving efficient and highly accurate detection of PCB surface defects is fundamental for quality control in fabrication. The rapidly advancing deep learning (DL) technology holds promising prospects for providing accurate and efficient detection methods for surface defects on PCB. To facilitate DL model training, it is imperative to compile a comprehensive dataset encompassing diverse surface defect types found on PCB at a significant scale. This work categorized PCB surface defects into 9 distinct categories based on factors such as their causes, locations, and morphologies and developed a dataset of PCB surface defect (DsPCBSD+). In DsPCBSD+, a total of 20,276 defects were annotated manually by bounding boxes on the 10,259 images. This openly accessible dataset is aimed accelerating and promoting further researches and advancements in the field of DL-based detection of PCB surface defect.

Background & Summary

The printed circuit board (PCB) plays a vital role in electronic devices, serving the dual purposes of mechanically supporting and establishing electrical connections among diverse electronic components[1]. PCB finds applications in virtually all types of electronic information devices, spanning from 3C (computer, communication, and consumer) devices, household appliances, automobile electronics and more[2]. The quality of PCB is crucial for the overall performance of electronic equipment. Consequently, PCB manufacturers are expected to provide products with high quality, high precision, and high reliability. Therefore, implementing rigorous quality control measures throughout the fabrication process and effectively detecting defects on PCB is of paramount importance. If recurring defects cannot be promptly and precisely detected, it is highly likely that many manufactured PCB will eventually be scrapped, resulting in both wastage and incurring significant costs[3].

The PCB fabrication process is complex, particularly for multilayer boards. Figure 1 illustrates the fabrication process of multilayer PCB. Due to factors such as technical faults, polluted work environment, device anomalies, and manual mishandling, various surface defects are inevitably introduced at different stages[4]. These defects encompass issues such as open, short, spur, spurious copper, foreign object, and more, affecting various element composition of PCB, including the copper wire, copper surface, hole, base material, and so forth. These surface defects not only affect the aesthetics but can also significantly impair the performance of PCB or even cause extensive damage to the entire board. Therefore, the PCB workshop explicitly mandates defect inspections for each inner/outer layer board after etching, as well as before shipping. Efficient and accurate detection of these various defects is crucial to ensure quality in the PCB fabrication.

Previously, surface defect detection of PCB was typically carried out through manual visual inspection. However, manual visual suffers from drawbacks such as heavily relying on experienced inspectors, significant subjectivity, high labor intensity, low consistency and efficiency[5]. As the demands for precise and efficient inspection of PCB increase, especially with the trend towards greater complexity, tiny, and intricacy, manual inspection is gradually being phased out. In the past decades, there has been a heightened emphasis on utilizing

[1]School of Engineering, South China Agricultural University, No. 483, Wushan Road, Guangzhou, 510642, China. [2]Guangzhou FastPrint Technology Co., Ltd, No. 33, Guangpuzhong Road, Guangzhou, China. [3]CEPREI, No. 78, Zhucun Avenue West, Guangzhou, 511370, China. ✉e-mail: lvshengping@scau.edu.cn

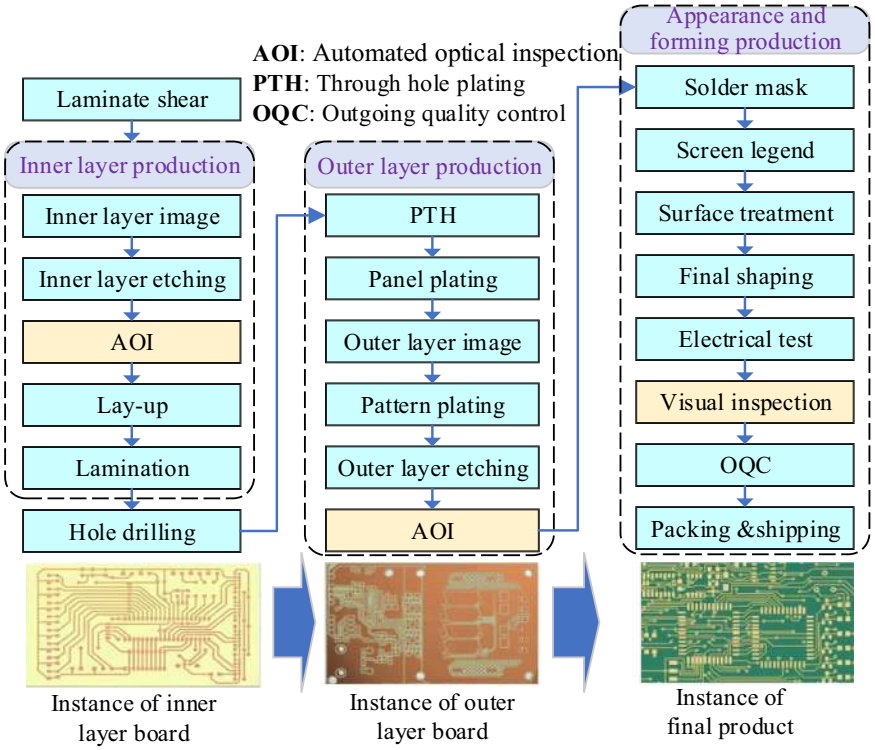


Fig. 1 Overview of fabrication process for multilayer PCB.

automated optical inspection (AOI) for defect detection in the PCB production, and empirical evidence suggests that it has significantly enhanced both the detection accuracy and efficiency[2,3].

The detection algorithm, which serves as the core component of AOI software, falls into three categories: traditional image processing approaches, machine learning approaches, and deep learning (DL)-based approaches. Traditional image processing approaches can still be classified into three types: referential methods, non-referential methods, and hybrid methods–which involve a combination of more than one of these methods[5]. The referential methods compare images or extracted features of the inspected PCB with predefined template or features to identify the locations of defects[6,7]. Referential methods have been widely used but suffer from several drawbacks[3], including heavy dependence on the quality of templates, the need for additional preprocessing, calibration, and post-processing, and time-consuming pixel-level comparison matching and so on. Non-referential methods recognize defects based on pre-designed rules or by assuming that features are simple geometric shapes, where defects manifest as unexpected irregular features. However, these methods might overlook significant flaws and distorted features[3].

Machine learning-based approaches such as decision trees[8] and support vector machines[9] have also been integrated into AOI to enhance detection accuracy. However, these approaches share the same issues as traditional image processing approaches[3]. Additionally, machine learning approaches require the configuration of various parameters and often necessitate additional preprocessing and post-processing for defect image handling. Based on the aforementioned achievements, the existing AOI detection accuracy, speed, and level of automation have greatly improved compared to manual visual inspection. However, practical experience has revealed that AOI equipment still has a high likelihood of falsely identifying or misjudging defects, necessitating the involvement of a significant number of specialized personnel for time-consuming visual rechecks.

With the rapid advancement of DL technology in recent years, there has been an increasing emphasis on end-to-end DL-based approaches in AOI detection algorithm research. The goal is to improve both efficiency and detection accuracy[10,11] while addressing the limitations associated with traditional image processing and machine learning-based approaches. To effectively support the training of DL models, there is a need to construct a dataset that comprehensively covers various types of PCB surface defects and has a significant scale.

In recent years, several publicly available PCB defect datasets have been created for training and evaluating DL models. Tang *et al.*[1] constructed a dataset called DeepPCB, which can be accessed at https://github.com/tangsanli5201/DeepPCB. This dataset consists of 1,500 pairs of defect images (template and tested images) covering 6 common types of surface defects, including open, short, mousebite, spur, copper, and pinhole. All images in the DeepPCB dataset were captured using a linear scan charge coupled device. Additionally, a number of artificial defects were manually introduced into each test image, resulting in approximately 3 to 12 defects per image. The MeiweiPCB dataset[12], available at https://github.com/youtang1993/MeiweiPCB), comprises 939 defect images randomly cropped from original images acquired using an industrial line scan camera. Defects in MeiweiPCB were not categorized, and labels for these defects were provided in two forms: pixel-wise mask annotation and bounding box annotation. Huang *et al.*[13] published a PKU-Market-PCB dataset, which contains

1,386 defect images categorized into 6 categories: missing hole, mouse bite, open, short, spur, and spurious copper. The dataset can be accessible at https://robotics.pkusz.edu.cn/resources/dataset/. Out of these images, 693 originated from 10 PCB, while the remaining 693 images were generated through rotation augmentation. The PKU-Market-PCB dataset has been extensively utilized to validate the performance of DL models[13–16], and has been augmented by some researches[15,16]. Ding *et al.*[15] expanded the dataset to 10,668 images using geometric and image transformations. Du *et al.*[16] constructed a dataset comprising 693 images of normally placed and 507 images randomly rotated based on PKU-Market-PCB.

Researchers have also developed some unpublicized datasets. Hu *et al.*[17] constructed a dataset comprising 1,500 defect images, covering 6 common types of defects: open, short, mouse bite, spur, pinhole, and solder ball. All the images in the dataset were cropped from the original images captured using a camera. Rotation and brightness adjustment were introduced to the dataset to augment the original images, resulting in a total of 12,000 defect images. Liao *et al.*[18] developed a dataset comprising 19,029 defect images. Each image encompasses one of the following surface defects: bumpy, clutter, scratch, line repair damage, hole loss, or over oil-filling. Among these, 2,008 images were randomly cropped from original images acquired using an industrial camera. The other 17,021 were generated using augmentation techniques, such as random rotation, cropping, translation, horizontal and vertical flipping, and luminance balancing. Adibhatla *et al.*[19] extracted 11,000 images from the AOI machine to compile a PCB dataset. The dataset comprises images with 11 types of defects; however, all the defects were labeled as a single defect type. Adibhatla *et al.*[20] built a dataset comprising 23,000 PCB surface defect images. All the defect images were collected from an automated visual inspection (AVI) machine. Each defective region in the image was labeled as DEFECT with bounding box. Pham *et al.*[21] created a dataset consisting of 22,909 surface defect images. The original PCB images were obtained through an AFVI system. Subsequently, defects were extracted and saved as cropped images, with the defects positioned at the center of each image. These defect images were classified as either true or false defects. Zhang *et al.*[11] provided a PCB-2 dataset consisting of 40,706 images and two classes of defects: real defect and pseudo defect. Li *et al.*[22] constructed a dataset containing 2,000 images and five categories of defects, including copper short, short, open, near open and near short. Each category has 400 defect images.

The aforementioned datasets support the training of DL models and accelerate the research on DL-based algorithms. However, the currently constructed datasets have several shortcomings.

(1) PKU-Market-PCB[13] and its extended versions[15,16], as well as DeepPCB[1] are primarily generated through artificial synthesis. The majority of defects in the datasets constructed by Ding *et al.*[15], Hu *et al.*[17] and Liao *et al.*[18] were generated by augmenting a small set of original defects. This synthesis or augmentation has led to very limited intra-class variability and has created significant disparities from real defects that occur in the PCB production process.

(2) The defects in certain datasets lack categorization[12,19–21] or they were roughly divided into two categories[11]. Meanwhile, some datasets encompass five[22] or six[1,13–18] defect categories but have a limited number of defect samples, with a maximum of 2,008 defects. Therefore, there is a need to enhance the coverage of surface defects on PCB and refine the categorization.

(3) The division of training and validation sets in some datasets[1,13–18] was conducted after augmenting the original defects. This results in the validation set containing a significant number of defect samples that are highly similar to those in the training set, making it challenging to validate the detection accuracy and generalization performance of DL models for real-world applications.

(4) The AOI/AVI system often crops multiple defect images using a sliding window, and these cropped images may display defect-free instances, duplicate defects, incomplete defects, and so on. However, datasets sourced from AOI/AVI[11,19,20,22] do not provide details on how to address this situation. Additionally, there is no guidance on handling defects that cannot be determined solely through visual inspection and highly unbalance among different categories of defects. Furthermore, these datasets have not been publicly shared.

To overcome these limitations, the goal of this study is to establish a publicly available dataset that focuses on the surface defects of inner and outer layer boards of PCB, covering multiple defect categories. The dataset, named DsPCBSD+ (Dataset of PCB surface defect), comprises images sourced from actual PCB produced at Guangzhou FastPrint Technology Co., Ltd. Experts in PCB have meticulously classified defects into 4 primary categories considering their causes: copper residue, copper deficiency, conductor scratch, and foreign object. These 4 categories are further subdivided into 9 categories considering factors such as locations and morphologies. The 9 defect categories comprise Short (SH), Spur (SP), Spurious copper (SC), Open (OP), Mouse bite (MB), Hole breakout (HB), Conductor scratch (CS), Conductor foreign object (CFO), and Base material foreign object (BMFO). In total, this dataset comprises 10,259 defect images and 20,276 manually annotated defect bounding boxes. This openly accessible dataset aims at accelerating and promoting further research and advancements in the field of intelligent detection of PCB defects.

Methods

The construction procedure for DsPCBSD+ is illustrated in Fig. 2 and is primarily divided into the following three steps: (1) Defect images collection; (2) Defect classification and data preprocessing; (3) Defect labeling and dataset partition.

Defect images collection. This study exclusively utilized images of actual PCB defects from the inner and outer layers of boards after etching, gathered from the AOI equipment AGLE'OL AOI-100 V8 in the workshop of Guangzhou FastPrint Technology Co., Ltd. These two equipment employ the multi-group of controllable LED spotlight system for illumination, coupled with a 16K high-resolution line scan system for image acquisition. The line scan image acquisition system comprises four cameras mounted on the top and an additional four on the bottom. These cameras are utilized to capture images of both sides of the PCB. When capturing PCB images,

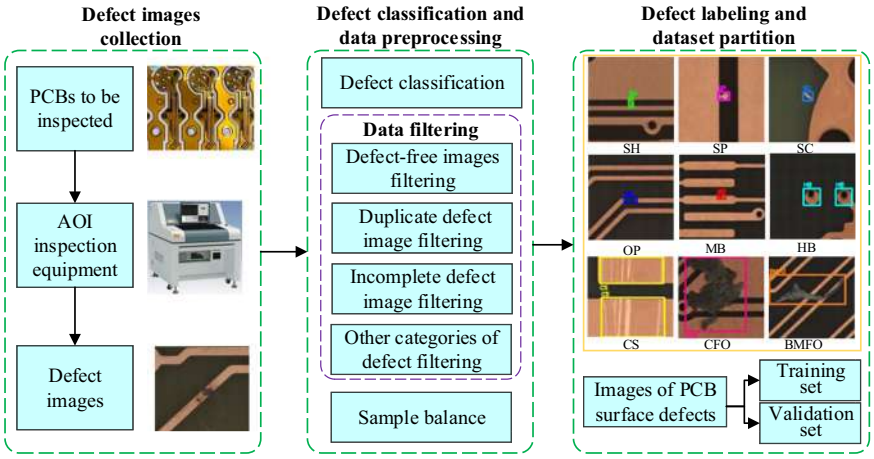


Fig. 2 Overview of the construction process for DsPCBSD+.

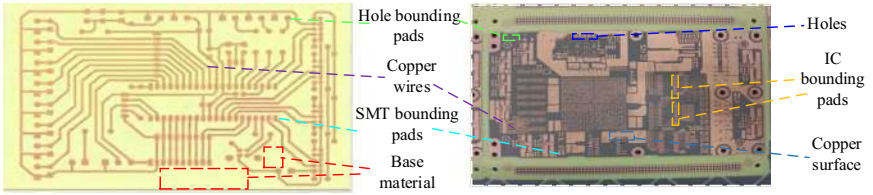


Fig. 3 Schematic diagram of element compositions on inner/outer layer.

the linear sensor scans line by line along the horizontal direction of the PCB under inspection. Image is captured and transmitted in digital form to the image processing unit. In the image processing unit, the captured images undergo preprocessing, which includes tasks such as noise removal, contrast enhancement, and brightness adjustment. Subsequently, key features are extracted from these images based on element feature learning and sub-pixel contour comparison approaches. The processor meticulously compares these extracted features with predefined reference images to identify any potential defects and their locations. All defect images are systematically archived in a management system, enabling comprehensive analysis, traceability, and continuous quality enhancement. A total of 32,259 images have been directly retrieved from the management system to compose the DsPCBSD+ dataset. Each image is formatted in JPG and has dimensions of 226 × 226 pixels.

Defect classification and data preprocessing. PCB surface defects are diverse and have traditionally been classified based on the understanding of personnel within PCB manufacturing facilities, rather than being categorized according to the requirements of DL-based detection. In this study, features of PCB surface defects are abstracted based on their causes, locations and morphologies, and defect types are redefined here.

Considering the causes, PCB surface defects are primarily influenced by factors such as copper residue, copper deficiency, conductor scratches, and foreign objects. Consequently, these defects have been classified into 4 distinct categories: Copper residue defect, Copper deficiency defect, Conductor scratch defect, and Foreign object defect. In terms of the locations where these surface defects manifest, they can occur across various elements and compositions on the PCB. Figure 3 provides a visual representation of the principal element compositions for inner and outer layer boards. Inner layer boards primarily consist of wires, hole bounding pads, surface mount technology (SMT) bounding pads, and the base material area. On the outer layer boards, the main element compositions encompass copper wires, copper surfaces, holes, and pads (including hole bounding pads, SMT bounding pads, and integrated circuit (IC) bounding pads). IC/SMT/hole bounding pads can be regarded as relatively small areas of copper surfaces. And copper wires, copper surfaces, IC/hole/SMT bonding pads can be collectively referred to as conductors. In summary, the locations on PCB where surface defects occur can be categorized into three main groups: conductors, holes, and the base material. Subsequently, these defects are further subdivided into 9 categories based on their locations and morphologies, as depicted in Fig. 4.

A Short on a PCB refers to an unintended connection between two or more distinct conductors resulting from the presence of residual copper. Short defects primarily occur between different wires, and they can also occur between copper surfaces and wires, or between copper surfaces. Spur defects are identified by irregular protrusions along the edges of conductors, often manifesting as sharp, pointed shapes. The Spur defect may lead to incorrect connections between conductors, potentially causing short or disrupting the normal operation of the circuit. Spurious copper refers to irregular or unwanted copper residue found on the base material, within

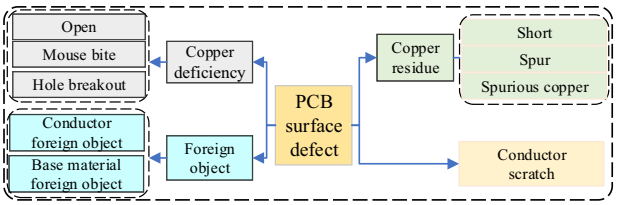


Fig. 4 PCB surface defect classification.

holes, or on copper surfaces. The presence of this unintended copper material can result in incorrect circuit connections, short circuits, and other potential circuit issues.

An Open defect on PCB occurs when the connection path within the conductor is interrupted, hindering the flow of electric current along that specific pathway. A Mouse bite on a PCB refers to small, localized depression or crack that appear at the edge of conductors. It may result in poor connections in the circuit or connections that are close to being interrupted. The Hole breakout typically refers to a situation where the center of a hole significantly deviates from the hole bounding pad, resulting in a noticeable lack of copper material along some edges of the hole. Hole breakout affects the functionality and performance of the holes, especially in multi-layer PCB where holes are used to connect different layers of circuits.

Defects on copper wires or copper surfaces stemming from scratches, appearing as linear or multiple linear defects, fall under the category of Conductor scratch. Such scratches on elements like wires and pads can result in suboptimal or interrupted electrical connections, impacting the overall functionality of the device. Concurrently, these scratches may undermine the mechanical strength between components, making the PCB more vulnerable to the effects of vibrations, impacts, or temperature fluctuations, ultimately diminishing the device's reliability.

The morphological characteristics of foreign objects appearing on conductors or the base material vary significantly. Impurities, bubbles, dirt, deposits, or other substances on the conductors are categorized as Conductor foreign object. Meanwhile, the unintentional presence of contamination such as oils, oxides, chemical etching, cleaning agents, solvents, fingerprints, stains and so forth on conductors is also classified as Conductor foreign object. When foreign materials, such as bubbles, chips, particles, or other substances, appear on the base material, they are classified as Base material foreign object. Similarly, contamination, such as grease, adhesives, oxidation, corrosion, and other forms of pollution, visually resembling foreign objects to a significant extent, is also categorized as the defect type of Base material foreign object. The background of holes is similar to base material, and the foreign objects within the holes generally exhibit morphological characteristics similar to those of Base material foreign object. Consequently, when foreign objects are present inside the holes, they are categorized as Base material foreign object. Foreign objects on a PCB have the potential to result in short circuits between components that should not be connected, obstruct connections between circuit elements, or interfere with circuit signals. These issues ultimately lead to reduced device performance.

During the dataset construction process, all the source images potentially containing abnormal data should be systematically eliminated, either manually or automatically. This includes defect-free images, duplicate defect images, incomplete defect images, and other categories of defect images in the original dataset.

Defect-free images represent cases where no defects are present or the defects are too subtle to be discerned by the naked eye. The defect-free images are manually screened and excluded. AOI often captures multiple photos of the same defect to enhance the accuracy of detecting and confirming potential defects. Consequently, this leads to the storage of duplicate defect images with partially overlapping regions. In such cases, the hash value matching method is employed to compare all images and one of images is retained. Subsequently, additional manual screening is conducted, and the image with the highest defect percentage is retained for dataset construction. The incomplete defect image signifies an image containing defects that require clear boundaries, but some boundaries related to the defects are missing in the image. For example, the image with Open object only includes one end of the conductor for this defect, while the image with Short object only contains one side of the conductor for this defect. Other categories of defect images involves images that do not include any of the aforementioned 9 specific defect categories, such as exceeding tolerances in conductor width/spacing, depressions in blind vias, or missing drilled holes. These defects are determined through template matching but cannot be identified solely through 2D visual inspection. Other categories of defect images are manually excluded from the source images.

After the aforementioned screening, the retained images exhibit a notable imbalance among the 9 types of defect objects. Specifically, there is a few of OP and SH defects, while a substantial number of CFO and BMFO defects are present. Training DL models on such imbalanced samples can introduce bias in feature learning, leading the model to prioritize better recognition of categories CFO and BMFO with more samples, resulting in suboptimal performance on OP and SH with fewer samples. Given that OP and SH defects can directly lead to the scrapping of PCB, whereas CFO and BMFO defects typically do not directly lead to PCB scrapping, it is crucial to ensure that the model avoids biasing towards learning features of CFO and BMFO at the expense of recognizing OP and SH defects. To address this issue, all images containing Open and Short defects are included in the DsPCBSD+ dataset. As for defects in other categories, an initial subset of images is randomly selected for defect labeling, and additional images are included based on the statistical results derived from labelled bounding boxes for various defect categories. This process endeavors to establish a more balanced distribution of defect types, thereby ensuring a more robust and equitable performance across all defect categories.

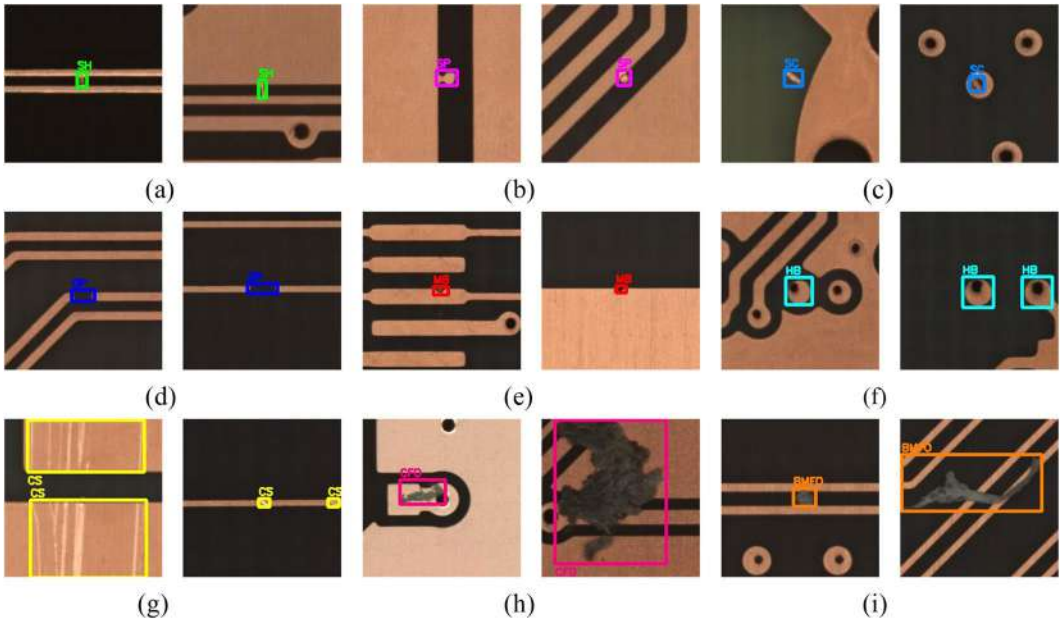


Fig. 5 Annotation instances of the 9 categories of defect. (**a**) SH; (**b**) SP; (**c**) SC; (**d**) OP; (**e**) MB; (**f**) HB; (**g**) CS; (**h**) CFO; (**i**) BMFO.

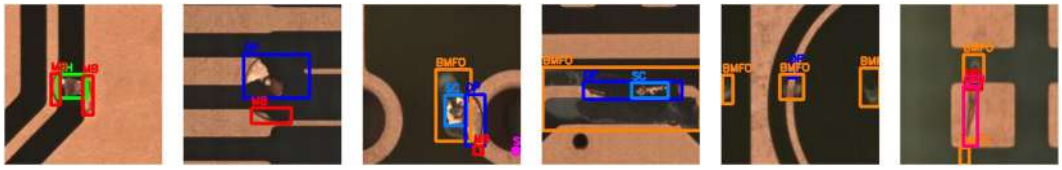


Fig. 6 Annotation instances of multi categories of defect on one image.

Defect labeling and dataset partition. Finally, the categorized defects were annotated and the labels were named using LabelImg software. An official open-source download link for the LabelImg software is provided here: https://github.com/heartexlabs/labelImg. Researchers have the flexibility to assign labels in YOLO, VOC, or CreateML format and annotate the images following the guidelines provided by the official documentation. The YOLO, VOC, or CreateML formats can be mutually converted. In this study, labels were initially formatted in the VOC style before annotation. For each defective image, a bounding box was meticulously drawn around the respective defect. Subsequently, the bounding box was labeled with abbreviations of defect categories, including SH (Short), SP (Spur), SC (Spurious copper), OP (Open), MB (Mouse bite), HB (Hole breakout), CS (Conductor scratch), CFO (Conductor foreign object), and BMFO (Base material foreign object). The annotation instances for the 9 categories of defects were presented in Fig. 5. In practical production, various defects of different shapes and sizes often occur on the surface of PCB. These defects may simultaneously affect multiple element compositions of the PCB, such as copper wires, copper surfaces, holes, and base materials, ultimately resulting in multiple categories of defect on one image. In such scenarios, each type of defect had to be individually annotated, and the annotation instances are depicted in Fig. 6.

After completing the annotation, we obtained a collection of images with labeled bounding boxes and VOC-structured XML files containing information about these bounding boxes and labels. Finally, a total of 20,276 defects were annotated by bounding boxes on the 10,259 images. The VOC datasets were then converted into YOLO and COCO datasets through script files and stored. According to the definition of COCO[23], objects with area of ground truth less than 32×32 pixels, between 32×32 pixels and 96×96 pixels and larger than 96×96 pixels are taken as small, middle and large-sized objects, respectively. The size distribution of objects is shown in Fig. 7. In addition, the distribution of three types defect labels across different categories in DsPCBSD+ has been compiled and given in Table 1. Notably, there is a significant variation in the size of defects within DsPCBSD+. It can also be seen that Short, Spur, Spurious copper, Open and Mouse bite exhibit relatively small differences in size, while others, like Conductor scratch, Conductor foreign object, and Base material foreign object show significant variations in size. By combining the defect instances given in Fig. 5, it is evident that the defect categories with significant size variations also exhibit substantial intra-class differences.

Based on the annotated dataset, images were randomly divided into training and validation sets at an 8:2 ratio. There are 8,208 images in the training set and 2,051 images in the validation set. The corresponding labels are 16,184 and 4,092 for the training and validation sets, respectively. Figure 8 provides the number of labels for the 9 categories of defects in both the training and validation sets.

Categories	Small	Medium	Large	All
Short (SH)	710	205	0	915
Spur (SP)	4469	115	0	4584
Spurious copper (SC)	1352	231	10	1593
Open (OP)	1406	361	3	1770
Mouse bite (MB)	2421	108	0	2529
Hole breakout (HB)	35	2848	0	2883
Conductor scratch (CS)	734	1043	713	2490
Conductor foreign object (CFO)	1140	582	110	1832
Base material foreign object (BMFO)	1308	304	68	1680
Total	13575	5797	904	20276

Table 1. Count of three size labels.

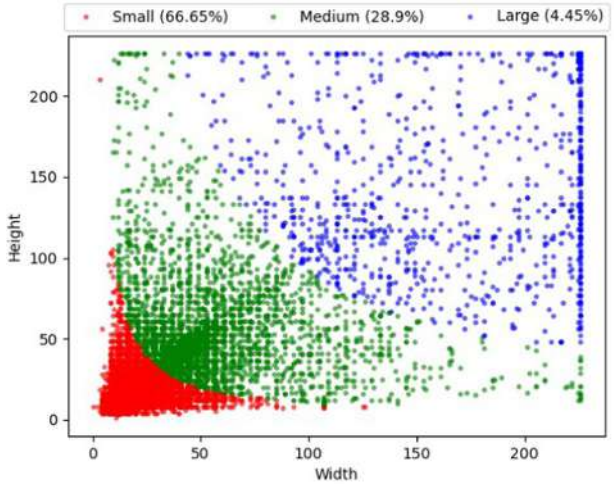


Fig. 7 Size distribution of defect bounding boxes in DsPCBSD+.

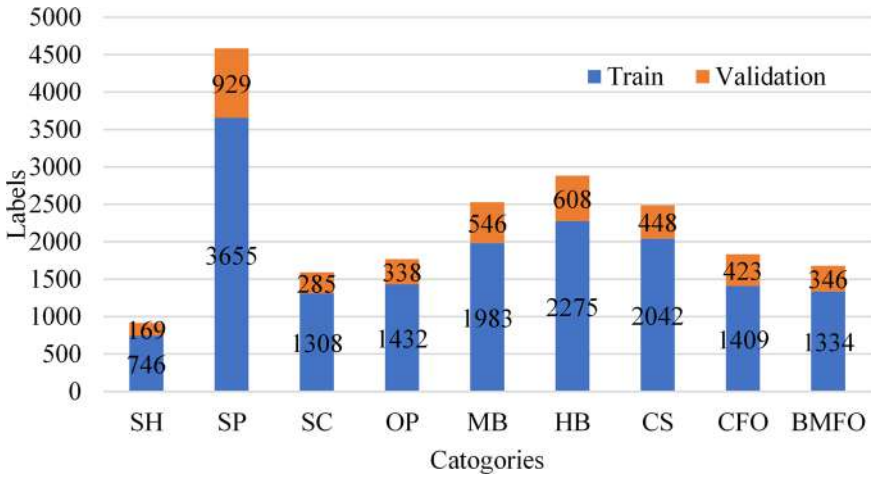


Fig. 8 Distribution of defect labels in the training and validation sets across different categories.

Data Records

The DsPCBSD+ are freely available in the Figshare repository[24]. Due to their attributes of simplicity, flexibility, and universality, YOLO and COCO dataset formats have garnered popularity in both academic and industry. Consequently, data annotations in this study are provided in YOLO and COCO formats, with images and annotation files stored separately in the Data_YOLO and Data_COCO folders.

There are two subfolders under Data_YOLO: images and labels, in which respectively stores image data and label data. Within these subfolders, there are two additional subfolders: train and val. These subfolders store

the defects data for training and validating the specified DL-based models respectively. The information contained in the label data mainly includes data type, number of labels and label coordinates. The files within the labels subfolder comprise details such as the file name, label category, center coordinates of the defective object's bounding box, as well as the width and height of the defective object's bounding box. These measurements are normalized, representing proportions relative to the image's width and height. The label file has the capability to encompass multiple defective objects, with each object delineated on an individual line.

The folder Data_COCO comprises three main subfolders: train2017 for the training set images, val2017 for the validation set images, and annotations for storing label files. In the annotations subfolder, there are two .json format label files: instances_train2017 and instances_val2017 which are used to store label information in the training and validation sets, respectively. These label files contain essential information such as images, annotations and categories. The images section includes image ID, file name, width, and height. Annotations part comprises annotation ID, associated image ID, defect object category ID, bounding box coordinates (bbox), defect object area, and more. The categories section provides details on category ID and category labels.

Technical Validation

To ensure the reliability of the DsPCBSD+ dataset for this study, a comprehensive manual examination was conducted on all images and their corresponding label annotations. This rigorous review process engaged five experts with substantial experience in the PCB manufacturing industry. These experts meticulously scrutinized each image in the dataset, carefully validating label information to identify potential omissions or inaccuracies. For situations prone to causing annotation discrepancies, such as similarities between different defect categories, a defect spanning multiple elements of PCB, or the overlap of multiple defect annotation boxes in the same region, a collective discussion involving five team members was conducted to determine the labeling categories and positions for these defects. In this collaborative discussion, a comprehensive consideration was given to factors such as the severity of each defect's impact on PCB performance, the proportion of defects in various locations (elements), and the visibility of defects.

The DsPCBSD+ dataset offers two dataset formats, YOLO and COCO, providing convenience for utilization with the currently prominent, top-ranked DL-based detection models. To evaluate the efficacy of the curated dataset, two state-of-the-art (SOTA) models, Co-DETR[25] and YOLOv6-L6[26], both ranked highly in object detection on COCO, were selected for training and validation on the DsPCBSD+. The verification link for the two models can be found at https://github.com/Sense-X/Co-DETR and https://github.com/meituan/YOLOv6. The dataset was trained on computer with Ubuntu 20.04 64-bit operating system, Intel Xeon Gold 6242R processor, GeForce RTX 3090 graphics processor. The hyperparameters for the two aforementioned detection models were initially set to the recommended default values. However, to better align with the characteristics of the dataset, certain hyperparameter values such as batch size, initial learning rate, and epochs were adjusted. The dimensions of the input images have been resized to 1333×800 and 1280×1280 respecitvely for the two mdoels in the training. These modifications were made in accordance with the respective recommendations from the original research papers of each model.

The training time for Co-DETR and YOLOv6-L6 on the test device configuration used in this study are approximately 69 minutes and 721 minutes, respectively. Figure 9 illustrates the precision-recall curves for Co-DETR and YOLOv6-L6, which are generated using their respective default tools. The benchmark results of the mean average precision (mAP) with different Intersection over Union (IoU) are summarized in Table 2, while Table 3 presents the average precision (AP), precision (P), recall (R), and F1 score for each defect category at IoU of 0.50 for the two models. Regarding the Co-DETR detection model, the mAP value for all defect categories at an IoU of 0.50 is approximately 0.848, as depicted by curve C50 in Fig. 9(a). On the other hand, for YOLOv6-L6, the overall mAP is approximately 0.851. Besides, Fig. 10 displays several instances of detection results generated by Co-DETR and YOLOv6-L6. These results validate the reliability and practicality of the DsPCBSD+ dataset, as demonstrated by the high detection performance and accuracy.

From Table 2, it can be observed that as the detected defect objects become smaller, both Co-DETR and YOLOv6-L6 show lower average detection precision (AP_S, AP_M, and AP_L) and recall (AR_S, AR_M, and AR_L). The challenges inherent in detecting small defects often involve constraints such as limited visual features, low resolution, and potential occlusion by surrounding elements. Given the increasing complexity of PCBs with a proliferation of tiny and intricate details, the corresponding defects have become finer and smaller. Therefore, models tailored for the detection of these small defects must adeptly navigate these challenges with sensitivity, capturing intricate details to ensure precise detection.

From Table 3, it is evident that both Co-DETR and YOLOv6-L6 demonstrate relatively high AP and R across various categories of defects. Figure 11 illustrates the confusion matrix for each defect category, including the background (BG), for both models. Based on the Table 3 and the confusion matrix in Fig. 11, it can be observed that the proportion of each type of defect being misclassified as other defect categories is relatively low, with only SC being misclassified as SP at a relatively higher rate (0.06 for Co-DETR and 0.07 for YOLOv6-L6). This could be because when SC is closely adjacent to the conductor, its features are similar to SP, making it prone to misclassification as SP. At the same time, it can be seen that both Co-DETR and YOLOv6-L6 exhibit relatively high rates of missing SP, CS, and CFO defects. From the perspective of defect features, the majority of SP defects are small in size, rendering them relatively inconspicuous. Conversely, CS defects exhibit significant internal size variations, making smaller scratches particularly prone to being overlooked, as evidenced by a CS defect that escaped detection by Co-DETR, as depicted in the second row and second column of Fig. 10. CFO defects display considerable disparities in color, size, and morphology, with some foreign objects being similar in color to the background. Consequently, CFO defects are susceptible to being erroneously identified as background, echoing the CFO defect that evaded detection by Co-DETR in the second row and three column of Fig. 10. The additional instances of misidentification or missed detection mentioned above are further depicted in Fig. 12.

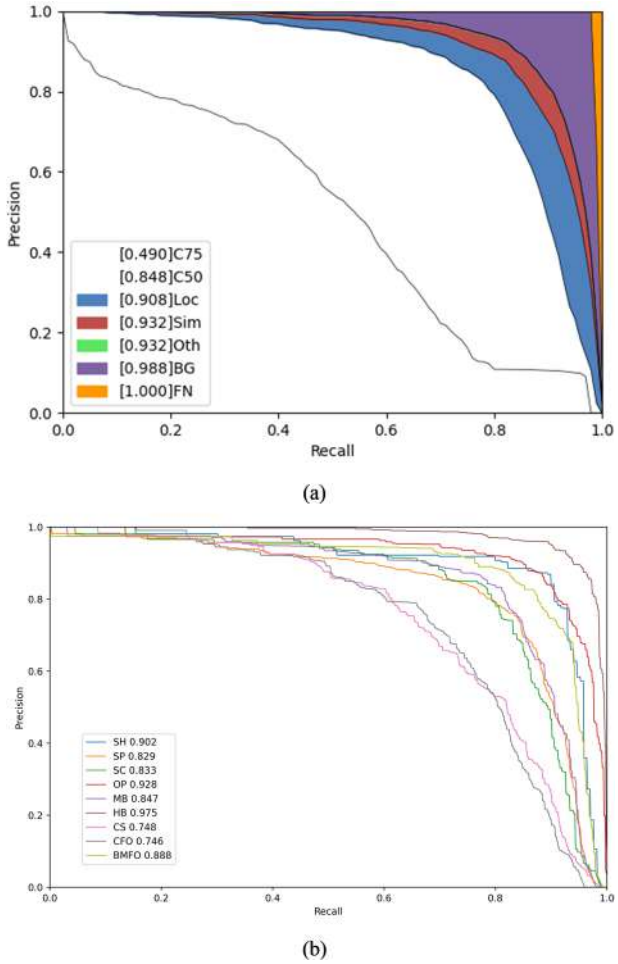


Fig. 9 The precision-recall curves and average AP of two models (IOU = 0.50). (**a**) The result obtained by Co-DETR. C75 denotes precision at an IoU threshold greater than or equal to 0.75, while C50 represents precision at an IoU threshold greater than or equal to 0.50; Loc signifies the accuracy of the model in predicting the detected object's position. Sim represents the model's error in predicting the shape or appearance similarity of the object. Oth encompasses errors in other aspects that do not fall under the specific types of Loc or Sim errors. BG indicates that the model erroneously predicted a background region as an object. FN represents objects that the model failed to detect. (**b**) The result obtained by YOLOv6-L6. The number following each class of defect represents the AP for that specific category.

Model	AP_{50}	AP_{75}	$AP_{50:95}$	AP_S	AP_M	AP_L	$AR_{50:95}$	AR_S	AR_M	AR_L
Co-DETR	0.848	0.490	0.492	0.425	0.554	0.671	0.668	0.600	0.743	0.846
YOLOv6-L6	0.851	0.525	0.514	0.405	0.597	0.681	0.654	0.590	0.748	0.812

Table 2. The mAP results of different models in DsPCBSD+. AP_{50}, AP_{75}, and $AP_{50:95}$ respectively represent the AP at IoU thresholds of 0.50, 0.75, and within the range from 0.50 to 0.95. AP_S, AP_M, and AP_L respectively denote the AP for small, medium, and large objects at IoU thresholds from 0.50 to 0.95. $AR_{50:95}$, AR_S, AR_M, and AR_L represent the Average Recall at IoU thresholds from 0.50 to 0.95, for total, small, medium, and large objects.

To better validate the robustness and reliability of the dataset, additional experiments were conducted using five-fold cross-validation. Based on the initial dataset partitioning, the training set underwent further random division into four equal segments. These segments were then merged with the original validation set, resulting in the creation of five distinct sets. Each set was cyclically utilized as a validation set once, while the remaining four sets served as training sets. Consequently, these sets not only represent the original dataset but also variations where each segment of the training set took turns as the validation set. Five-fold cross-validation experiments were carried out for each of the two models (represented respectively by Co-$DETR_5$ and YOLOv6-$L6_5$), and the resulting benchmark outcomes are detailed in Table 4. It can be observed that the performance metrics of Co-$DETR_5$ and YOLOv6-$L6_5$ exhibit minimal deviation from the results obtained with the original dataset partitioning as shown in Table 2. It can be inferred that the different folds of the constructed DsPCBSD+ contain

Model	Metric	SH	SP	SC	OP	MB	HB	CS	CFO	BMFO
Co-DETR	AP	0.885	0.842	0.825	0.898	0.836	0.973	0.727	0.741	0.900
	P	0.838	0.746	0.750	0.860	0.793	0.866	0.688	0.734	0.836
	R	0.824	0.824	0.736	0.819	0.783	0.980	0.761	0.659	0.854
	F1	0.831	0.783	0.743	0.839	0.788	0.920	0.723	0.695	0.845
YOLOv6-L6	AP	0.898	0.826	0.830	0.923	0.843	0.971	0.746	0.743	0.881
	P	0.649	0.676	0.627	0.677	0.644	0.726	0.442	0.547	0.687
	R	0.876	0.883	0.804	0.911	0.846	0.987	0.853	0.697	0.902
	F1	0.746	0.766	0.705	0.777	0.732	0.836	0.582	0.613	0.780

Table 3. The benchmark results of each defect category for the two models in DsPCBSD+.

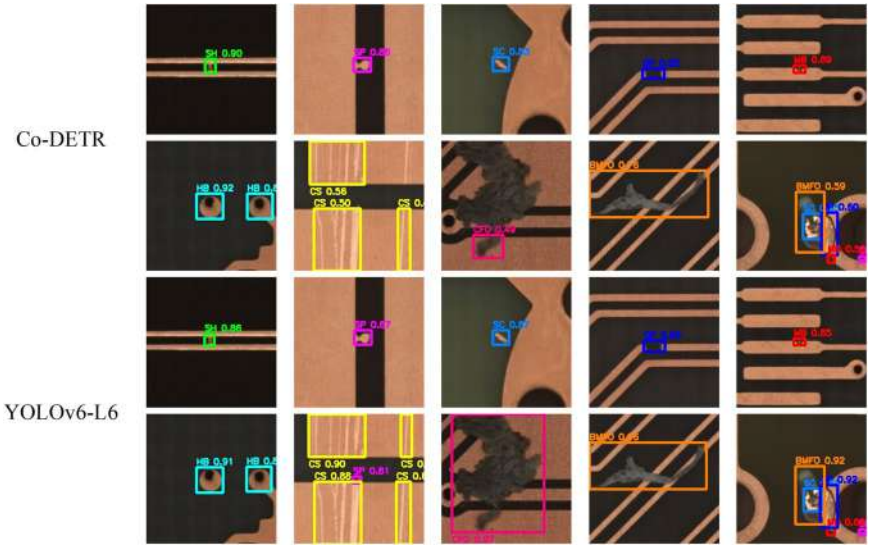


Fig. 10 Instances of detection result obtained by Co-DETR and YOLOv6-L6.

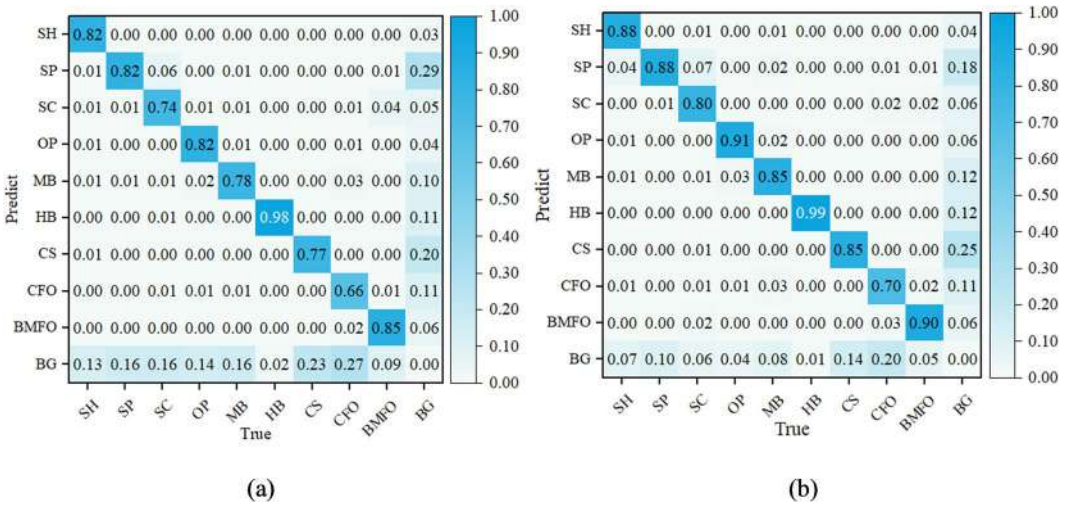


Fig. 11 The normalized confusion matrix of the two models. (**a**) Co-DETR. (**b**) YOLOv6-L6. In the normalized confusion matrix, rows represent the predicted categories, while columns represent the true categories. Diagonal cells indicate accurately predicted labels. Each number in a cell represents the proportion of the model predicting the true category as the corresponding rows category.

a sufficient number of representative samples, and the variations in the partitioning have minimal impact on the model results, demonstrating that the DsPCBSD+ effectively covers and represents the entire sample space.

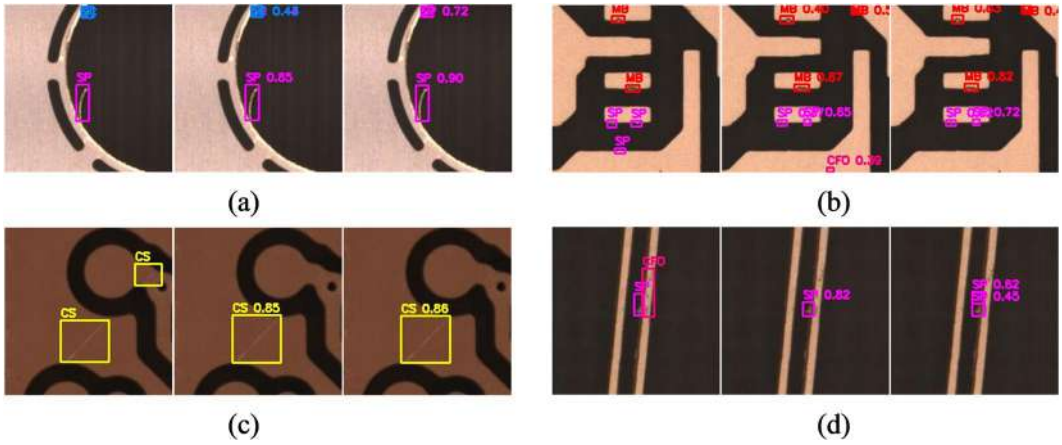


Fig. 12 The examples of misidentification or missed detection. (**a**) SC misidentified as SP; (**b**) Missed detection of SP; (**c**) Missed detection of CS; (**d**) Missed detection of CFO. The first column in each subfigure presents the annotated instance, while the second and third columns display the detection results obtained by Co-DETR and YOLOv6-L6, respectively.

Model	AP_{50}	AP_{75}	$AP_{50:95}$	AP_S	AP_M	AP_L	$AR_{50:95}$	AR_S	AR_M	AR_L
Co-$DETR_5$	0.840	0.483	0.484	0.420	0.551	0.547	0.648	0.612	0.746	0.858
YOLOv6-$L6_5$	0.837	0.512	0.502	0.405	0.591	0.614	0.651	0.596	0.748	0.823

Table 4. The five-fold cross-validation results of different models in DsPCBSD+.

Datasets	Defects categories
DeepPCB[1]	Open, Short, Mousebite, Spur, Copper, Pin-hole
MeiweiPCB[12]	Single type of defect
PKU-Market-PCB[13] and its extended version[14–16]	Missing hole, Mouse bite, Open circuit, Short, Spur, Spurious copper
Hu *et al.*[17]	Open circuit, Short course, Mouse bite, Spur, Pinhole, Solder ball
Liao *et al.*[18]	Bumpy or broken line, Clutter, Scratch, Line repair damage, Hole loss, Over oil-filling
Adibhatla *et al.*[19,20]	Single type of defect
Pham *et al.*[21]	
PCB-2[11]	Real defect, Pseudo defect
Li *et al.*[22]	Copper short, Short, Open, Near open, Near short
DsPCBSD+[24]	Short, Spur, Spurious copper, Open, Mouse bite, Hole breakout, Conductor scratch, Conductor foreign object, Base material foreign object

Table 5. Comparison of defect categories in DsPCBSD+ with those in existing PCB defect datasets.

Usage Notes

Table 5 compares the defect categories in DsPCBSD+ with those in existing PCB defect datasets. Typically, most existing PCB datasets overlook defects like hole breakout, foreign objects, and scratches. However, these defects are very common on the surface of PCBs in actual production. The images collected in this study also show that hole breakout, foreign objects, and scratches account for a large proportion. Therefore, by incorporating these three types of defects into its classification, DsPCBSD+ can better cater to the practical requirements of PCB product quality inspection. Additionally, existing datasets lack detailed explanations of classification standards. Typically, various defects are only introduced through carefully selected defect image instances that are easily distinguishable. However, descriptions detailing the defect formation location, cause, and morphological characteristics are not provided. The absence of specific classification standards during the dataset creation process poses a risk of misclassification, especially for defects sharing strong inter-class similarities, such as Spur and Spurious copper found at the edge of the conductor.

However, the classification scheme and the dataset have their own limitations. Firstly, all the defects are limited to 2D due to the AOI's cameras lacking 3D depth information. As a result, defects such as raised or recessed areas cannot be identified. Secondly, the images selected for DsPCBSD+ are from the inner and outer layers of boards after etching, without considering defect images after the solder mask. Thirdly, the defect images in DsPCBSD+ are all cropped images of local areas from the entire board. In practical applications, it is necessary to integrate these

Co-DETR		YOLOv6-L6	
Site-packages	**Versions**	**Site-packages**	**Versions**
Python	3.7.11	Python	3.8.18
Pytorch	1.11.0	Pytorch	1.13.1
Torchvision	0.12.0	Torchvision	0.14.1
Mmcv-full	1.5.0	Matplotlib	3.7.4
Mmdet	2.25.3	Numpy	1.23.5
Mmengine	0.10.2	Opencv-python	4.8.1.78
Numpy	1.21.6	Pillow	10.1.0
Openmim	0.3.9	Pycocotools	2.0.7
Opencv-python	4.9.0.80	Pyyaml	6.0.1
Pyyaml	6.0.1	Requests	2.31.0
Scipy	1.7.3	Scipy	1.10.1
Tqdm	4.65.2	Tqdm	4.66.1

Table 6. The site-packages and corresponding version for the two networks.

local images into the entire board and then mark the positions of defects on the entire board to facilitate localization by inspection personnel. These limitations should be considered when using DsPCBSD+ for practical applications.

Code availability

As mentioned earlier, the DsPCBSD+ dataset is accessible on the figshare data repository[24]. Additionally, the Python code for the hash value matching method, utilized to filter highly similar images, is provided alongside the dataset and named Hash.py. Researchers can perform label format conversion from VOC format to YOLO, and YOLO format to COCO format using the resources available at the following links: https://github.com/RapidAI/VOC2YOLO, https://github.com/RapidAI/YOLO2COCO. These links offer the necessary code for label format conversion, accompanied by a README file that serves as a helpful reference. The annotation tool LabelImg is available for download on the official website at https://github.com/heartexlabs/labelImg. For dataset verification using Co-DETR and YOLOv6-L6 codes, the following website links can be visited at: https://github.com/Sense-X/Co-DETR and https://github.com/meituan/YOLOv6. The site-packages and their corresponding versions used for the aforementioned two networks are provided in Table 6. The software packages can be obtained by accessing the links specified in the README files of the respective networks and can be easily installed using the Python package installer (pip).

Received: 12 January 2024; Accepted: 16 July 2024;
Published online: 22 July 2024

References

1. Tang, S., He, F., Huang, X. & Yang, J. Online PCB defect detector on a new PCB defect dataset. Preprint at https://doi.org/10.48550/arXiv.1902.06197 (2019).
2. Zhou, Y., Yuan, M., Zhang, J., Ding, G. & Qin, S. Review of vision-based defect detection research and its perspectives for printed circuit board. *Journal of Manufacturing Systems* **70**, 557–578, https://doi.org/10.1016/j.jmsy.2023.08.019 (2023).
3. Ling, Q. & Isa, N. A. M. Printed circuit board defect detection methods based on image processing, machine learning and deep learning: A survey. *IEEE Access* **11**, 15921–15944, https://doi.org/10.1109/ACCESS.2023.3245093 (2023).
4. Abu Ebayyeh, A. A. R. M. & Mousavi, A. A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry. *IEEE Access* **8**, 183192–183271, https://doi.org/10.1109/ACCESS.2020.3029127 (2020).
5. Moganti, M., Ercal, F., Dagli, C. & Tsunekawa, S. Automatic PCB inspection algorithms: A survey. *Computer Vision and Image Understanding* **63**, 287–313, https://doi.org/10.1006/cviu.1996.0020 (1996).
6. Wang, W.-C., Chen, S.-L., Chen, L.-B. & Chang, W.-J. A machine vision based automatic optical inspection system for measuring drilling quality of printed circuit boards. *IEEE Access* **5**, 10817–10833, https://doi.org/10.1109/ACCESS.2016.2631658 (2017).
7. Gaidhane, V. H., Hote, Y. V. & Singh, V. An efficient similarity measure approach for PCB surface defect detection. *Pattern Analysis and Applications* **21**, 277–289, https://doi.org/10.1007/s10044-017-0640-9 (2018).
8. Yuk, E. H., Park, S. H., Park, C.-S. & Baek, J.-G. Feature-learning-based printed circuit board inspection via speeded-up robust features and random forest. *Applied Sciences-Basel* **8**, 932, https://doi.org/10.3390/app8060932 (2018).
9. Ding, S., Liu, Z. & Li, C. AdaBoost learning for fabric defect detection based on HOG and SVM. In *2011 International conference on multimedia technology*, 2903–2906, https://doi.org/10.1109/ICMT.2011.6001937 (2011).
10. Kang, D., Lai, J. & Han, Y. Improving surface defect detection with context-guided asymmetric modulation networks and confidence-boosting loss. *Expert Systems with Applications* **225**, 120121, https://doi.org/10.1016/j.eswa.2023.120121 (2023).
11. Zhang, H., Jiang, L. & Li, C. CS-ResNet: Cost-sensitive residual convolutional neural network for PCB cosmetic defect detection. *Expert Systems with Applications* **185**, 115673, https://doi.org/10.1016/j.eswa.2021.115673 (2021).
12. Kang, D. MeiweiPCB surface defect dataset. Github https://github.com/youtang1993/MeiweiPCB (2021).
13. Huang, W. & Wei, P. A PCB dataset for defects detection and classification. Preprint at https://doi.org/10.48550/arXiv.1901.08204 (2019).
14. Zhang, Y. *et al*. A lightweight one-stage defect detection network for small object based on dual attention mechanism and PAFPN. *Frontiers in Physics* **9**, 708097, https://doi.org/10.3389/fphy.2021.708097 (2021).
15. Ding, R., Dai, L., Li, G. & Liu, H. TDD-net: A tiny defect detection network for printed circuit boards. *CAAI Transactions on Intelligence Technology* **4**, 110–116, https://doi.org/10.1049/trit.2019.0019 (2019).
16. Du, B. *et al*. YOLO-MBBi: PCB surface defect detection method based on enhanced YOLOv5. *Electronics* **12**, 2821, https://doi.org/10.3390/electronics12132821 (2023).

17. Hu, B. & Wang, J. Detection of PCB surface defects with improved faster-RCNN and feature pyramid network. *IEEE Access* **8**, 108335–108345, https://doi.org/10.1109/ACCESS.2020.3001349 (2020).
18. Liao, X. *et al*. YOLOv4-MN3 for PCB surface defect detection. *Applied Sciences* **11**, 11701, https://doi.org/10.3390/app112411701 (2021).
19. Adibhatla, V. A. *et al*. Defect detection in printed circuit boards using you-only-look-once convolutional neural networks. *Electronics* **9**, 1547, https://doi.org/10.3390/electronics9091547 (2020).
20. Adibhatla, V. A. *et al*. Applying deep learning to defect detection in printed circuit boards via a newest model of you-only-look-once. *Mathematical Biosciences and Engineering* **18**, 4411–4428, https://doi.org/10.3934/mbe.2021223 (2021).
21. Pham, T. T. A., Thoi, D. K. T., Choi, H. & Park, S. Defect detection in printed circuit boards using Semi-Supervised Learning. *Sensors* **23**, 3246, https://doi.org/10.3390/s23063246 (2023).
22. Li, Z., Gao, L., Gao, Y., Li, X. & Li, H. Zero-shot surface defect recognition with class knowledge graph. *Advanced Engineering Informatics* **54**, 101813, https://doi.org/10.1016/j.aei.2022.101813 (2022).
23. Lin, T.-Y. *et al*. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, 740–755, https://doi.org/10.1007/978-3-319-10602-1_48 (2014).
24. Lv, S. *et al*. A dataset for deep learning based detection of printed circuit board surface defect. *Figshare* https://doi.org/10.6084/m9.figshare.24970329 (2024).
25. Zong, Z., Song, G. & Liu, Y. Detrs with collaborative hybrid assignments training. In Proceedings of the IEEE/CVF352 international conference on computer vision, 6748–6758 (2023)
26. Li, C. *et al*. YOLOv6 v3.0: A full-scale reloading. Preprint at https://doi.org/10.48550/arXiv.2301.05586 (2023).

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (Grant No. 52275487) and Natural Science Foundation of Guangdong, China (Grant No. 2021A1515012395). The authors sincerely thank Guangzhou FastPrint Technology Co., Ltd. for their generous provision of the essential images depicting PCB surface defects for this study. Additionally, the authors extend their appreciation for the valuable guidance provided by the company in the process of defect classification and labeling. Special acknowledgments go to all individuals who contributed to the dataset construction and label annotation process.

Author contributions

Associate professor S.L. played a key role in overall planning, dataset organization, and manuscript writing. Z.D. was in charge of collecting dataset pictures. Defect classification and label review were carried out by S.L., Z.D., S.J., B.O. and T.L. Data filtering and dataset labeling were the responsibilities of B.O., T.L., K.Z., J.C. and Z.L. B.O. and T.L. specifically handled the Co-DETR and YOLOv6-L6 model training and results analysis. S. J. provided the training platform for model training. Manuscripts related to the dataset underwent collective review by all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-03656-8.

Correspondence and requests for materials should be addressed to S.L.

scientific reports

Check for updates

OPEN

An enhanced walrus optimization algorithm for flexible job shop scheduling with parallel batch processing operation

Shengping Lv[1], Jianwei Zhuang[1], Zhuohui Li[1], Hucheng Zhang[1], Hong Jin[1] & Shengxiang Lü[2]✉

The flexible job shop scheduling problem with parallel batch processing operation (FJSP_PBPO) in this study is motivated by real-world scenarios observed in electronic product testing workshops. This research aims to tackle the deficiency of effective methods, particularly global scheduling metaheuristics, for FJSP_PBPO. We establish an optimization model utilizing mixed-integer programming to minimize makespan and introduce an enhanced walrus optimization algorithm (WaOA) for efficiently solving the FJSP_PBPO. Key innovations of our approach include novel encoding, conversion, inverse conversion, and decoding schemes tailored to the constraints of FJSP_PBPO, a random optimal matching initialization (ROMI) strategy for generating diverse and high-quality initial solutions, as well as modifications to the original feeding, migration, and fleeing strategies of WaOA, along with the introduction of a novel gathering strategy. Our approach significantly improves solution quality and optimization efficiency for FJSP_PBPO, as demonstrated through comparative analysis with four enhanced WaOA variants, eleven state-of-the-art algorithms, and validation across 30 test instances and a real-world engineering case.

Keywords Flexible job shop scheduling, Parallel batch processing operations, Walrus optimization algorithm, Makespan

The flexible job shop scheduling problem (FJSP), first explored by Brucker and Schlie[1] and Brandimarte[2], evolved from the classic job shop scheduling problem (JSP). This problem poses a complex combinatorial optimization challenge involving multiple equalities and inequalities constraints. Due to its wide range of engineering applications and inherent complexity, FJSP has consistently attracted significant research attention[3,4].

The FJSP with parallel batch (p-batch) processing operation (FJSP_PBPO), explored in this study enables multiple jobs to be processed jointly on the same machine, thus posing a challenge to the traditional constraint of the FJSP, where each machine can handle only one job at a time. This problem is motivated by real-world scenarios observed in electronic product testing workshops. In electronic product testing, the workshop devises an overarching testing process plan for prototypes of the same product model. These prototypes are grouped into distinct categories, and each group undergoes sequential testing operations according to specified sub-routes within the plan. Figure 1 illustrates a performance testing process plan for a mobile phone, where 14 prototypes are divided into 7 groups. Each group of prototypes is treated as a single job. However, certain prototypes necessitate cross-group combination testing, leading to parallel batch processing operation (PBPO) on the same machine, as illustrated by (O_{14}, O_{24}) and (O_{33}, O_{44}) in Fig. 1 The introduction of PBPO to the FJSP further complicates the already NP-hard nature of the FJSP[4,5], making it significantly challenging to find viable and optimal solutions, which urgently needs resolution in electronic product testing workshops.

Recently, swarm-based metaheuristic algorithms have gained significant attention for addressing FJSP due to their efficiency in producing high-quality solutions[6–10]. Integrating FJSP_PBPO characteristics with advanced swarm-based metaheuristic mechanisms shows promise for improving optimization in this area. The walrus optimization algorithm (WaOA) is a relatively recent metaheuristic inspired by the behavior of walruses. It is known for its strong exploration capabilities and a balanced approach between exploration and exploitation. This balance allows the algorithm to avoid local optima and effectively explore the solution space, making it

[1]School of Engineering, South China Agricultural University, Guangzhou 510642, China. [2]School of Mathematics and Statistics, Hunan University of Finance and Economics, Changsha 410205, China. ✉email: lvshengxiang@hufe.edu.cn

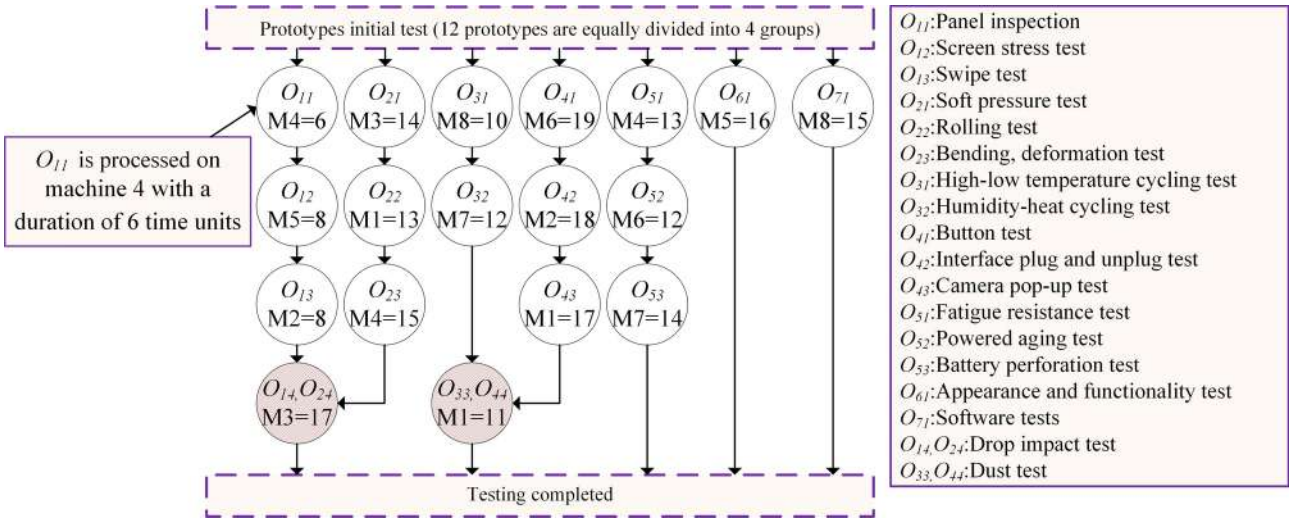


Fig. 1. Performance testing process plan of the mobile phone.

particularly well-suited for tackling global optimization problems. However, the algorithm uses a continuous encoding scheme, making it unsuitable for directly solving the FJSP_PBPO addressed in this study. Additionally, the No Free Lunch theorem asserts that no single algorithm is universally effective for all optimization problems, indicating that strong performance on one problem does not guarantee similar results on others[11]. In industrial applications, the selection, design, and fine-tuning of metaheuristic algorithms must consider the unique demands and characteristics of specific problems to optimize performance effectively[12]. Therefore, this study aims to leverage the potential of WaOA and improve it to better address the FJSP_PBPO and enhance its solution quality. The specific innovative contributions are as follows:

(1) An optimization model for FJSP_ PBPO is formulated using mixed-integer programming (MIP).
(2) An enhanced WaOA (eWaOA) is proposed specifically for FJSP_PBPO.
(3) New encoding, conversion, inverse conversion and decoding schemes tailored to FJSP_PBPO are developed.
(4) A random optimal matching initialization (ROMI) strategy is designed to generate diverse and high-quality initial solutions.
(5) Enhancements to feeding, migration, and fleeing strategies, coupled with the introduction of a novel gathering strategy, enhance the algorithm's effectiveness in both global exploration and local exploitation.

The subsequent sections are structured as follows: "Related works" section reviews literature on FJSP with p-batch processing and swarm-based metaheuristic algorithms for FJSP. "Description and modeling of FJSP_PBPO" section presents the problem description and model of FJSP_PBPO. "Walrus optimization algorithm" section details the mathematical modeling of WaOA. "The proposed improved WaOA for FJSP_PBPO" section outlines the design of the eWaOA, encompassing encoding, conversion and inverse conversion technique, decoding, population initialization, enhancements of WaOA's strategies, and newly designed gathering strategy. "Computational experiments and real-word case study" section presents the created benchmark instances, along with the conducted experiments, engineering case study, and results. "Conclusion and future research" section provides the conclusions and the future work.

Related works

FJSP and FJSP with p-batch processing

Since its inception in the beginning of the 90s[1,2], the FJSP has evolved significantly over the past three decades. During this time, researchers have incorporated additional resource constraints, including transport resources[13–16], molds[17], and dual human–machine resources[18,19]. Furthermore, new time-related constraints, such as setup times [20] and uncertain processing times[21,22], have been introduced. These advancements continuously broaden the applicability of FJSP, aligning the problem more closely with the optimization demands of real-world workshops[4]. In addition, the FJSP with p-batch processing has also been extensively studied[4,5], with a primary focus on wafer fabrication environments.

According to the standard three-field α|β|γ notation in scheduling, introduced by Graham et al.[23], the FJSP with p-batch processing can be represented as FJ|p-batch |γ, where γ denotes the objective to be optimized. The objectives in these studies mainly include minimizing makespan (C_{max}), total weighted tardiness (TWT), total tardiness (TT), total completion time (TC), total weighted completion time (TWC), maximum lateness (L_{max}), maximum tardiness ($T_{\max}$), and number of tardy jobs (NTJ).

Many studies have proposed heuristic methods based on disjunctive graph (DG), with the most widely explored being variations of the shifting bottleneck heuristic (SBH) initially introduced by Adams et al.[24]. Mason et al.[25,26] presented a modified SBH to address the FJ| p-batch |TWT. Experimental results show that their

modified SBH outperforms dispatching rules and surpasses an MIP heuristic in all but small instances. To reduce computational costs, Mönch and Drießel[27,28] adopted a two-layer hierarchical decomposition approach, using a modified SBH[27] and a SBH enhanced by a genetic algorithm (GA)[28] for sub-problems optimization. Mönch and Zimmermann[29] further applied the SBH to the same problem in a multi-product setting. Barua et al.[30] developed a SBH to optimize L_{max}, TT, TC of the problem within a stochastic and dynamic environment using discrete-event simulation, demonstrating superior performance compared to traditional dispatching methods. Upasani et al.[31] streamlined the problem by focusing on bottleneck machines in the DG while representing non-bottlenecks as delays, effectively balancing solution quality and computational effort. Sourirajan and Uzsoy[32] introduced a SBH that employs a rolling horizon approach to create manageable sub-problems. Upasani and Uzsoy[33] further integrated this rolling horizon strategy with the reduction approach proposed by Upasani et al.[31]. Pfund et al.[34] expanded the SBH to optimize TWT, C_{max}, and TC, employing desirability functions to evaluate criteria at both the sub-problems and machine criticality levels. Yugma et al.[35] proposed a constructive algorithm for a diffusion-area FJSP, incorporating iterative sampling and simulated annealing (SA), which demonstrated effectiveness on real-world instances. Knopp et al.[36] introduced a new DG and a greedy randomized adaptive search procedure (GRASP) metaheuristic combined with SA, yielding excellent results on benchmark and industrial instances. Ham and Cakici[37,38] developed an enhanced optimization model employing MIP and constraint programming (CP) for the FJ| p-batch $|C_{max}$, which was solved using IBM ILOG CPLEX. The computational results demonstrate that the proposed MIP model significantly reduces computational time compared to the original model, while the CP model outperforms all MIP models. Wu et al.[39] introduced an efficient algorithm based on dynamic programming and optimality properties for scheduling diffusion furnaces. The developed algorithm not only surpasses human decision-making but also enhances productivity compared to existing methods.

The FJ|p-batch $|\gamma$ has also been investigated across various other industrial environments. Boyer et al.[40] investigated this problem in seamless rolled ring production, where jobs are processed in batch furnaces, often in a first-in, first-out sequence, resulting in a PBPO structure. Zheng et al.[41] examined the JSP with p-batch processing, inspired by practical military production challenges, and proposed an auction-based approach combined with an improved DG for solution optimization. Xue et al.[42] developed a hybrid algorithm integrating variable neighborhood search (VNS) with a multi-population GA to address this problem, validated in a heavy industrial foundry and forging environment. Ji et al.[43] constructed a novel multi-commodity flow model for the FJ|p-batch $|C_{max}$, introducing an adaptive large neighborhood search (ALNS) algorithm with optimal repair and tabu-based components (ALNSIT) to effectively solve large-scale instances.

The above research provides valuable references for this study. However, existing optimization methods for FJSP with p-batch constraints are challenging to apply directly to the scheduling requirements of electronic product testing workshops. The main reasons are as follows: (1) Most of the research above, particularly studies on wafer fabrication scheduling, primarily focuses on batch processing machines (BPMs) scheduling and multi-level local optimization. In contrast, electronic product testing requires integrated scheduling of all jobs and machines throughout the workshop. (2) Except for the studies by Xue et al.[42] and Ji et al.[43], the process plans of all jobs are similar, and all jobs require processing through BPMs (e.g., acid bath wet sinks, heat treatment machines). In contrast, in this study, job process plans vary significantly, and only the jobs forming the PBPO need to undergo specified p-batch processing in the BPMs. (3) In existing studies, batch processing decisions are dynamically made based on each job's ready time and the capacity of the BPMs. In contrast, p-batching in this study pertains to specific operations from different jobs that must be processed jointly according to a predefined testing process plan.

Swarm-based metaheuristic algorithms for FJSP

Swarm-based metaheuristic algorithms are inspired by the collective behaviors observed in natural phenomena among mammals, birds, insects, and other organisms. Prominent examples include particle swarm optimization (PSO) (PSO)[44], ant colony optimization (ACO)[45], and artificial bee colony (ABC)[46], which are considered classical swarm-based metaheuristic algorithms. These algorithms have been extensively applied to the FJSP. For instance, Ding and Gu developed an enhanced PSO for addressing the FJSP[47]. Shi et al.[48] proposed a two-stage multi-objective PSO to tackle a dual-resource constrained FJSP. Zhang and Wong[49] addressed the FJSP in dynamic environments using a fully distributed multi-agent system integrated with ACO. Li et al.[50] introduced a reinforcement learning (RL) variant of the ABC for the FJSP with lot streaming.

In the past decade, swarm-based metaheuristic algorithms have seen rapid development, with new algorithms continually emerging. Notable examples include grey wolf optimization (GWO)[51], whale optimization algorithm (WOA)[52], satin bowerbird optimizer (SBO)[53], emperor penguin optimizer (EPO)[54], squirrel search algorithm (SSA) [55], harris hawks optimization (HHO)[56], red deer algorithm (RDA)[57], tuna swarm optimization (TSO)[58], remora optimization algorithm (ROA)[59], African vultures optimization algorithm (AVOA)[60], white shark optimizer (WSO) [61], WaOA[62], and walrus optimizer (WO)[12]. Some of these algorithms provide novel approaches for addressing (F)JSP.

Luo et al.[6] introduced an advanced multi-objective GWO (MOGWO) aiming at minimizing both makespan and total energy consumption for the multi-objective FJSP (MOFJSP). Lin et al.[63] introduced a learning-based GWO tailored for stochastic FJSP in semiconductor manufacturing. It employs an optimal computing budget allocation strategy to enhance computational efficiency and adaptively adjust parameters using RL.

Liu et al.[7] combined the WOA with Lévy flight and differential evolution (DE) strategies to tackle the JSP. The Lévy flight boosts global search and convergence during iterations, while DE enhances local search capabilities and maintains solution diversity to avoid local optima. Luan et al.[64] proposed an improved WOA (IWOA) for the FJSP, focusing on minimizing makespan. The IWOA features a conversion method to translate whale positions into scheduling solutions and employs a chaotic reverse learning strategy for effective initialization. Additionally,

it integrates a nonlinear convergence factor and adaptive weighting to balance exploration and exploitation, and incorporates a VNS for enhanced local exploitation.

Ye et al.[8] addressed the FJSP with sequence-dependent setup times and resource constraints by introducing a self-learning HHA (SLHHO) aimed at minimizing makespan. The SLHHO employs a two-vector encoding for machine and operation sequences, introduces a novel decoding method to handle resource constraints, and uses RL to intelligently optimize key parameters. Lv et al.[10] developed an enhanced HHO for both static and dynamic FJSP scenarios. This enhanced algorithm incorporates elitism, chaotic mechanisms, nonlinear energy updates, and Gaussian random walks to reduce premature convergence.

Fan et al.[65] introduced the genetic chaos Lévy nonlinear TSO (GCLNTSO) for the FJSP with random machine breakdowns, focusing on minimizing a combined index of maximum completion time and stability. He et al.[66] developed an improved AVOA for the dual-resource constrained FJSP (DRCFJSP). Enhancements to the AVOA include employing three types of rules for population initialization, establishing a memory bank to store optimal individuals across iterations for improved accuracy, and implementing a neighborhood search operation to further optimize makespan and total delay. Yang et al.[9] developed a hybrid ROA with VNS aimed at optimizing FJSP makespan. The algorithm incorporates a machine load balancing-based hybrid initialization method to enhance initial population quality and a host switching mechanism to improve exploration capabilities.

The advancements in FJSP research and engineering applications are notable, but the existing studies did not address PBPO constraints, which limits their applicability to FJSP_PBPO. Thus, new research is required to integrate the unique aspects of FJSP_PBPO with the selected swarm-based metaheuristic algorithms.

Description and modeling of FJSP_PBPO

The FJSP_PBPO extends the classic NP-hard problem FJSP[5]. It involves processing N jobs on M machines, with each job following a specific process plan composed of sequentially ordered operations. Each operation can be executed on a set of alternative machines with defined processing times. Additionally, some machines can process multiple operations from different jobs simultaneously, subject to PBPO constraints. The primary objective of the FJSP_PBPO is to determine the optimal processing order of each "task" (encompassing both operations and PBPO) on each machine to minimize the makespan (C_{max}), while also respecting precedence relationships among operations within the same job and among tasks on the same machine. Building on the complexities of FJSP, FJSP_PBPO further increases problem intricacy by incorporating PBPO constraints. Table 1 presents an example of an FJSP_PBPO scenario with four jobs and four machines, where the values under each machine indicate the processing time for each task. As with FJSP, FJSP_PBPO assumes that:

(1) All jobs can start processing at time 0, and all the jobs have the same priority.
(2) All machines are available at time 0.
(3) Each machine can handle only one task at a time.
(4) Each job is processed on only one machine at a time.
(5) Once a task begins on a machine, it must be completed without interruption.
(6) Each task can only start processing after its preceding tasks have been completed.
(7) All operations that form the PBPO must start and finish simultaneously.

The FJSP_PBPO is defined using specific notations. Below, we provide a concise overview of these notations and the corresponding problem formulations.

M: total number of machines;

Jobs	Tasks	Alternative machines			
		M1	M2	M3	M4
J_1	O_{11}	5	–	7	–
	O_{12}	5	–	–	–
	O_{13}	10	–	12	–
	O_{14}	14	13	–	14
J_2	O_{21}	8	–	7	10
	O_{23}	5	7	–	5
J_3	O_{31}	4	7	6	–
	O_{32}	8	12	–	–
	O_{34}	5	–	3	5
J_4	O_{41}	–	4	–	7
	O_{42}	6	–	7	–
	O_{44}	–	12	–	8
J_2, J_3	$\{O_{22}, O_{33}\}$	7	4	–	–
J_2, J_4	$\{O_{24}, O_{43}\}$	5	3	6	–

Table 1. Instance of FJSP_PBPO. J_i represents job i, O_{ij} represents the jth operation of J_i, $\{O_{22}, O_{33}\}$ and $\{O_{24}, O_{43}\}$ are two PBPOs.

N: total number of jobs;
m: machine index;
Tu: the task set for all the jobs;
u: task index,$u \in Tu$;
$JP[u]$: the immediate job predecessor task(s) of $u, u \in Tu$;
P_u^m: the processing time of task $u, u \in Tu$ on machine m;
S_u: the processing time of task $u, u \in Tu$;
C_u: the completion time of task $u, u \in Tu$;
P_o^m: the corresponding processing time of specific operation $o, o \in u$ on machine m;
S_o: the start time of specific operation $o, o \in u$;
C_o: the completion time of specific operation $o, o \in u$;
C^m: the completion time of the last task on machine m;
$C_{\max}$: makespan;
Q : a sufficiently large integer;
X_u^m: decision variable representing whether task u is processed on the machine m;

$$X_u^m = \begin{cases} 1 & \text{machine } m \text{ is assigned to task } u \\ 0 & \text{elsewise} \end{cases};$$

Y_{uv}: decision variable representing the order of two different tasks processed on the same machine;

$$Y_{uv} = \begin{cases} 1 & \text{if the task } u \text{ is processed before} \\ & \text{the task } v \text{ on the same machine} \\ 0 & \text{elsewise} \end{cases}.$$

Based on this, an optimization model is constructed using MIP, with the objective of minimizing the maximum completion time.

$$C_{max} = \min\{\max\{C^m\}\}, 1 \le m \le M. \tag{1}$$

Subject to:

$$C_u - S_u = P_u^m \times X_u^m, \forall u, m, \tag{2}$$

$$\sum_{m=1}^{M} X_u^m = 1, \forall u, \tag{3}$$

$$C_u \le C_{max}, \forall u, m, \tag{4}$$

$$C_u \le S_v + Q \times (1 - Y_{uv}), \forall u, v, m, \tag{5}$$

$$S_u \ge \max\{C_{u'}\}, \forall m, u' \in JP[u], \tag{6}$$

$$(S_u \ge 0) \cup (C_u \ge 0) \, \forall u, m, \tag{7}$$

$$(S_o = S_{o'}) \cup (C_o = C_{o'}), \forall m, o \in u, o' \in u. \tag{8}$$

Equation (1) represents the optimization objective function. Equation (2) indicates that tasks cannot be interrupted during processing. Equation (3) states that each task can only be processed on one machine. Equation (4) ensures that the completion time of any task does not exceed the maximum completion time C_{max} . Equation (5) ensures the precedence order between tasks on the same machine. Equation (6) guarantees the precedence order between tasks of the same job. Equation (7) asserts that the start and completion time of any task is non-negative. Equation (8) specify that the various operation of task u must start and finish simultaneously.

Walrus optimization algorithm

In WaOA, each walrus serves as a candidate solution in the optimization problem. Therefore, the position of each walrus within the search space determines the candidate values for the problem variables. The optimization process begins with a population of randomly generated walruses X, representing by D-dimensional random vectors, as defined by Eq. (9).

$$\begin{aligned} &X_p = lb + rand(ub - lb) \\ &X_p = [\ X_{p,1} \quad \ldots \quad X_{p,j} \quad \ldots \quad X_{p,D}\ ] \\ &1 \le p \le P, 1 \le j \le D, \end{aligned} \tag{9}$$

where, X_p is the pth initial walrus (candidate solution), lb and ub are the lower and upper boundaries of the problem, *rand* is a uniform random vector in the range 0 to 1, $X_{p,j}, 1 \le j \le D$ is the value of the jth decision variable of the initial walrus X_p, P is the number of walruses in the population, i.e., the population size, D is number of decision variables. Based on the suggested values for the decision variables, the objective function of the problem can be evaluated, and the resulting fitness function $F(X_p), 1 \le p \le P$ can be obtained.

Walruses are agents that perform the optimization process. Their positions are iteratively updated using feeding, migration, fleeing strategies until a termination condition is met. Each iteration follows a structured approach divided into three phases. In Phase 1, the WaOA utilizes feeding strategy to explore globally. In this phase, the best candidate solution so far is identified as the strongest walrus X_{str} according their fitness. Other walruses adjust their positions under the guidance of X_{str} according to the Eqs. (10) and (11).

$$X_{p,j}^{t+1} = X_{p,j}^{t} + rand_{p,j} \times (X_{str,j} - I_{p,j} \times X_{p,j}^{t}), 1 \le p \le P, 1 \le j \le D, \tag{10}$$

$$X_{p}^{t+1} = \begin{cases} X_{p}^{t+1}, & F(X_{p}^{t}) < F(X_{p}^{t+1}) \\ X_{p}^{t}, & else, \end{cases} \tag{11}$$

where $X_{p,j}^{t+1}$ is the new position for the *p*th walrus on the *j*th dimension, $X_{p,j}^{t}$ is the current position for the *p*th walrus on the *j*th dimension, $rand_{p,j}$ is a random number lies in the range (0,1), $X_{str,j}$ is the position for the strongest walrus X_{str} on the *j*th dimension, $I_{p,j}$ is an integer selected randomly between 1 or 2.

In Phase 2, each walrus migrates to a randomly selected walrus position in another area of the search space and the new position for each walrus can be generated according to Eqs. (12) and (13).

$$X_{p,j}^{t+1} = \begin{cases} X_{p,j}^{t} + rand_{p,j} \times (X_{k,j}^{t} - I_{i,j} \times X_{p,j}^{t}), & 1 \le p,k \le P, 1 \le j \le D, \ if\, F(X_{k}^{t}) \ge F(X_{p}^{t}) \\ X_{p,j}^{t} + rand_{p,j} \times (X_{p,j}^{t} - X_{k,j}^{t}), & 1 \le p,k \le P, 1 \le j \le D, \ if\, F(X_{k}^{t}) < F(X_{p}^{t}) \end{cases}, \tag{12}$$

$$X_{p}^{t+1} = \begin{cases} X_{p}^{t+1}, & F(X_{p}^{t}) < F(X_{p}^{t+1}) \\ X_{p}^{t}, & else, \end{cases} \tag{13}$$

where $X_{k}^{t}, 1 \le k \le P$ and $k \ne p$ is the position of the selected walrus to migrate the *p*th walrus towards it, $X_{k,j}^{t}, 1 \le k \le P, 1 \le j \le D$ is its *j*th dimension, and $F(X_{k}^{t})$ is its objective function value.

In Phase 3, the WaOA utilizes fleeing strategy to adjust the positions of each walrus within its neighborhood radius. This strategy is used to exploit the problem-solving space around candidate solutions. The new position can generate randomly in this neighborhood using Eqs. (14) and (15).

$$\begin{aligned} &X_{p,j}^{t+1} = X_{p,j}^{t} + \left(lb_{local,j}^{t} + \left(ub_{local,j}^{t} - rand \cdot lb_{local,j}^{t}\right)\right) \\ &local\ bound: \begin{cases} lb_{local,j}^{t} = lb_{j}/t \\ ub_{local,j}^{t} = ub_{j}/t \end{cases}, \end{aligned} \tag{14}$$

$$X_{p}^{t+1} = \begin{cases} X_{p}^{t+1}, & F(X_{p}^{t}) < F(X_{p}^{t+1}) \\ X_{p}^{t}, & else, \end{cases} \tag{15}$$

where lb_j and ub_j are the lower and upper bounds of the *j*th position, respectively, $lb_{local,j}^{t}$ and $ub_{local,j}^{t}$ are allowable local lower and upper bounds for the *j*th position, respectively. The pseudocode for WaOA is shown in Algorithm 1.

Input the information of the optimization problem
Set the population size of walruses (P) and the maximum iterations (T)
Initialize P search agents $X_p, 1 \le p \le P$ with D decision variables according Eq. (9), $X_p^0 = X_p, 1 \le p \le P$
Calculate the fitness of each search agent and set X_{str} as the best search agent
for t=1: T
  for p = 1: P
    Phase 1: Feeding strategy(exploration)
      Calculate $X_{p,j}^{t+1}, 1 \le j \le D$ using Eq. (10)
      Update X_p^{t+1} using Eq. (11)
    Phase 2: Migration strategy(exploration)
      Randomly choose an immigration destination X_k^t for the pth walrus $X_p^t, p \ne k$
      Calculate $X_{p,j}^{t+1}, 1 \le j \le D$ using Eq. (12)
      Update X_p^{t+1} using Eq. (13)
    Phase 3: Fleeing strategy (exploitation)
      Calculate $X_{p,j}^{t+1}, 1 \le j \le D$ using Eq. (14)
      Update X_p^{t+1} using Eq. (15)
  end for
  Update X_{str} if there is a better solution
  $t = t + 1$
end for
Output X_{str}

Algorithm 1. Walrus optimization algorithm (WaOA).

The proposed improved WaOA for FJSP_PBPO

Framework of the eWaOA

Due to the introduction of new constraints by FJSP_PBPO, existing encoding, conversion, and decoding methods for swarm-based metaheuristics used in FJSP are not directly applicable. Consequently, we first develop new encoding, conversion, inverse conversion and decoding schemes tailored to these constraints. Preliminary experiments have identified several shortcomings of the original WaOA when applied to FJSP_PBPO, such as premature convergence to local optima and inefficient updates. To address these issues, this study first create new initialization strategy and then enhance the WaOA's feeding, migrating, and fleeing strategies. Additionally, a gathering strategy is introduced to enhance both global and local optimization capabilities. The framework for the proposed eWaOA is illustrated in Fig. 2 and detailed below.

Step 1—Data input: Operation set, operation sequence for the N jobs, alternative machines with their associated processing times for each operation, and the PBPOs, with both operations and PBPOs collectively described as tasks.

Step 2—Parameter setting: Population size (P) of walruses, termination parameter (maximum iterations max$T$ or time limit T), control factor A, and matching parameter K for ROMI.

Step 3—Population initialization: The optimization process begins with P randomly generated walruses based on the ROMI strategy according to the parameter K. Each walrus is encoded as a real vector $X_p, 1 \le p \le P$ and an integer vector $X'_p, 1 \le p \le P$. On this basis, segmented conversion are developed to convert the X_p to X'_p, while the inverse conversion method is created to transform the X'_p into X_p. The X'_p are decoded using a designed semi-active decoding method to generate feasible scheduling scheme.

Step 4—Update the position of each walrus. Walruses serve as search agents in the optimization process, with their positions continuously updated through enhanced feeding, migration, fleeing strategies, or through the enhanced feeding and introduced gathering strategy. The decision between choosing the gathering strategy or the migration and fleeing strategies is controlled by A. For the feeding, migration, and fleeing strategies, real vectors X are updated directly during each iteration, with simultaneous conversion of the updated real vector X_{new} into corresponding integer vector X'_{new}. Conversely, gathering strategy involve direct updates of X' to generate X'_{new}, followed by inversely converting it to corresponding X_{new}. This ensures synchronization in updating the X and X' at each iteration.

Step 5—Updating the strongest walrus: Each X'_{new} are decoded into a feasible semi-active schedule for FJSP_PBPO, and the fitness values of the walruses is assigned the reciprocal of makespan corresponding to the schedule. And the walrus with the highest fitness so far is update as the strongest walrus X_{str}.

Step 6—Termination criterion: If the iterations reaches its preset max$T$ or the runtime reaches its preset T, the best solution is output, and the iteration stops; else, it proceeds to Step 4.

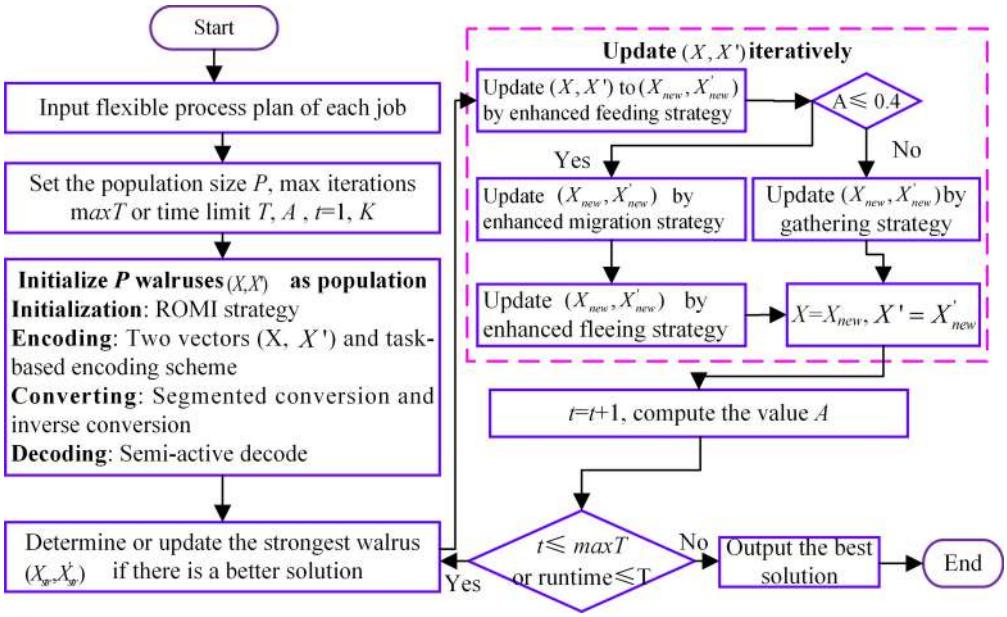


Fig. 2. Flowchart of the proposed eWaOA for the FJSP_PBPO.

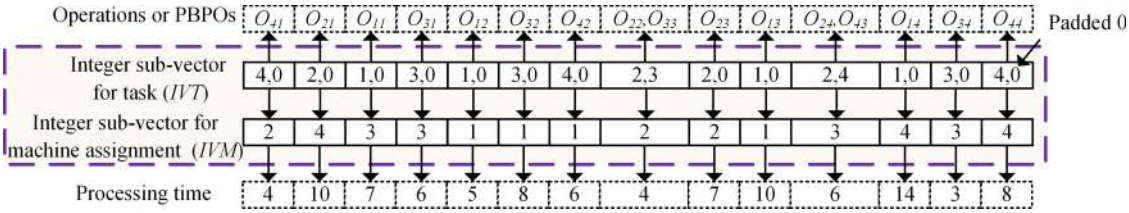


Fig. 3. An instance of two integer sub-vectors code for the FJSP_PBPO.

Representation of walrus and FJSP_PBPO

In our eWaOA, a vector $X_p = \{X_{p,1}, X_{p,2}, ...X_{p,D}\}$ is represented as a D-dimensional real vector, constrained by the specific requirements of the problem. FJSP_PBPO involves two sub-problems: task sequencing and machine assignment. Therefore, X_p should encompass information from both aspects. Let TN denote total number of all tasks in the FJSP_PBPO, then $D=2TN$. The first half part $X1_p = \{X_{p,1}, X_{p,2}, ...X_{p,TN}\}$ of X_p represent task sequencing, while the second half part $X2_p = \{X_{p,TN+1}, X_{p,TN+2}, ...X_{p,2TN}\}$ describes machine assignment for each task. Specifically, $X_{p,j}, 1 \leq j \leq TN$ denotes the value of the jth task decision variable in the vector X_p, $X_{p,j}, TN < j \leq 2TN$ represents the machine assignment decision variable for the (j-TN)th task of X_p. The $X1_p$ is defined as the real sub-vector for task (*RVT*) and $X2_p$ is defined as the real sub-vector for machine assignment (*RVM*) in this study. Additionally, the value of $X_{p,j}, 1 \leq j \leq 2TN$ is bound to be in the real range (-N, N), where N is the total number of jobs.

The WaOA is designed for continuous functions but is not directly applicable to discrete problems such as FJSP_PBPO. Additionally, the presence of PBPO implies that a single position in $X1_p$ may correspond to multiple operations across different jobs. Consequently, decoding X_p into a feasible schedule and evaluating the objective function value presents significant challenges. To address these issues, we further propose a task-based encoding method for FJSP_PBPO. This encoding scheme consists of an integer vector divided into two parts. The first part, the integer sub-vector for tasks (*IVT*), represents each position with job ID(s). To maintain consistency, the number of elements in each position is set to $s = \max\{|PBPO_k|\}, \forall k$, and the length of the vector is set to the number of tasks (*TN*), where $|PBPO_k|$ is the number of operations in the k-th PBPO. Positions with fewer than s elements are padded with zeros. If a position contains more than one job ID, it indicates that the position corresponds to a PBPO. For simplicity, padded zeros are omitted in the subsequent description. The second part, the integer sub-vector for machine assignment (*IVM*), has positions with potential values ranging from 1 to M. Each position in *IVM* corresponds to the processing machine for the task indicated by the same position in *IVT*. Figure 3 illustrates the integer vector code for FJSP_PBPO, showing the specific operations or PBPOs in *IVT* and their corresponding processing times. Thus, each walrus contains both continuous encoding vector $X_p = [RVT, RVM]$ and integer vector $X'_p = [IVT, IVM]$.

Conversion and inverse conversion scheme

The conversion process transforms a real vector to integer vector, allowing the eWaOA to solve FJSP_PBPO. The ranked order value (*ROV*) rule, originally designed for FJSP or JSP, uses order relationships and random keys to map a real vector into an integer operation sequence, which outlines the order of operations on all machines and forms a scheduling scheme[7,65]. However, the inclusion of PBPO in the FJSP_PBPO renders the traditional ROV method unsuitable. To address this, a novel segmented conversion algorithm is developed to convert the real vector $X_p = [RVT\ RVM]$ into integer vector $X'_p = [IVT, IVM]$. This algorithm introduces a task template (*TP*) segmented into three sections: the first section corresponds to tasks from jobs without PBPO; the second section consists of sequential PBPO tasks; and the third section includes tasks from jobs with PBPO, excluding the PBPOs themselves. For each segment, elements in the *RVT* are categorized and converted based on the methods outlined in Table 2 ensure that the constraints are maintained. Additionally, when converting the *RVM* to the *IVM*, the machine index for each task must be determined first, with the conversion formula provided in Eq. (16):

$$MI_j = \left\lfloor \frac{(RVM_j + N) \times s(j)}{2N} \right\rfloor + 1, TN + 1 \le j \le 2TN, \tag{16}$$

where N denotes the number of jobs, TN represents total number of all tasks,$s(j)$ indicates the total number of alternative machines for the task $IVT_{j-TN}, TN + 1 \le j \le 2TN$. The segmented conversion algorithm is described as follows:

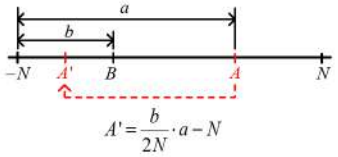
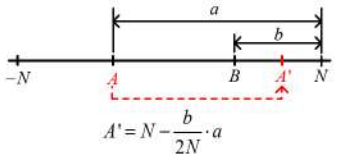
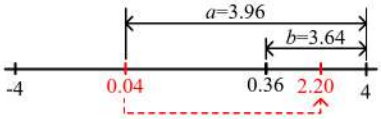
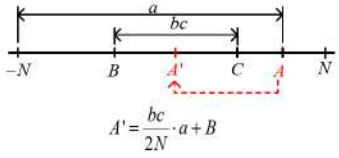
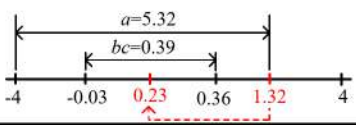
Type	Diagram and conversion formula	Example
Predecessor	 $A' = \frac{b}{2N} \cdot a - N$	B=0.36 represents value of PBPO {O24, O43} in the *RVT*, while A=0.52 denotes the value of O_{42} as a job predecessor task of this PBPO, then $A' = \frac{4.36}{2\times4} \times 4.52 - 4 = -1.54$. 
Successor	 $A' = N - \frac{b}{2N} \cdot a$	A=0.04 denotes the value of O_{44} as an immediate job successor task of PBPO {O24, O43}, and $A' = 4 - \frac{3.64}{2\times4} \times 3.96 = 2.20$ 
Middle	 $A' = \frac{bc}{2N} \cdot a + B$	B=-0.03 and C=0.36 represents the values of PBPO { O_{22}, O_{33}} and {O24, O43} in the *RVT* respectively. A=1.32 is the value of O_{23} as a middle task between the two PBPOs. Then the value of A' in *CRVT* can be calculated by $A' = \frac{0.39}{8} \times 5.32 - 0.03 = 0.23$ 

Table 2. Categorization and post-processing strategy of a *RVT*.
B and *C* represent the values of PBPOs. *A* is the value of either a job predecessor, successor of *B*, or the middle task between *A* and *B*. *N* denotes the number of jobs.

1:	Input a real vector X_p =[*RVT RVM*] for FJSP_ PBPO with *N* jobs and *TN* tasks, task template (*TP*)
2:	Let $TP_j, t \le j \le t'$ correspond to positions in the second section of *TP*, divide *RVT* into three sections following the *TP* structur
3:	Copy *RVT* to *CRVT*
4:	Sequentially update the **Predecessor** of $TP_j, t \le j \le t'$ in the second section of *CRVT* according to the strategy in Table 2
5:	Sequentially update the **Predecessors**, **Middles** and **Successors** of $TP_j, t \le j \le t'$ in the third section of *CRVT* according to the strategies outlined in Table 2
6:	The ROV of *CRVT* are obtained and stored in the ROV vector. Then, in ascending order of the ROV vector, job IDs are sequentially copied from the corresponding positions in the *TP* to construct an *IVT* without padded zeros
7:	Compute machine index $MI_j, TN+1 \le j \le 2TN$ using Eq.(16), and set the machine ID corresponding to MIS_j to $IVM_j, TN+1 \le j \le 2TN$
8:	Output the X'_p = [*IVT IVM*]

Algorithm 2. Segmented conversion.

Figure 4 illustrates an example of the segmented conversion process from *RVT* to *IVT*. First, the task template *TP* is created with three sections according to the job process plan: the first section, from TP_1 to TP_4, contains tasks for jobs without PBPO; the second section consists of sequentially arranged PBPO tasks TP_5 and TP_6; and the third section, from TP_7 to TP_{14}, includes tasks for jobs with PBPO but excludes the PBPOs themselves. The *RVT* is divided into three sections according to the *TP* structure, and its elements of *RVT* is copy to *CRVT*. Next, related values in the second and third sections of the *CRVT* are updated according to the conversion formulas given in Table 2. In this example, since O_{22} in $\{O_{22}, O_{33}\}$ is a predecessor of O_{24} in $\{O_{24}, O_{43}\}$, the value 3.28 in *RVT* for $\{O_{22}, O_{33}\}$ is initially converted to -0.03 in *CRVT* using the "Predecessor" conversion formula. Subsequently, the "Predecessor", "Middle", and "Successor" conversion formulas are applied sequentially to convert the predecessor tasks for $\{O_{22}, O_{33}\}$ and $\{O_{24}, O_{43}\}$, as well as the tasks O_{23} (middle of the two PBPOs) and the successor tasks. The original *RVT* values and their converted counterparts in the *CRVT* for the example in Table 2, are highlighted in red in Fig. 4 Finally, the ranked values of each element in the *CRVT* are obtained to generate the ROV vector. Following the ascending order of the ROV vector, job IDs are sequentially copied from corresponding positions in the *TP* to construct the IVT.

The inverse conversion is designed to transform X'_p to X_p while ensuring that the updates of X'_p remains consistent with the update of X_p. When converting the *IVT* to the *RVT*, a randomly generated *RVT* is introduced, and the *ROV* is determined based on the *TP* and *IVT*. Then, the *RVT* is reordered according to the *ROV*, and the corresponding values in *RVT* are updated based on the categorization strategies in Table 3, forming the new *RVT*. When converting *IVM* to *RVM*, the value corresponding to each task are first obtained using the following Eq. (17).

$$RVM_j = [\frac{2 \times N \times (MI_j - 1) - N \times s(j)}{s(j)}] + \frac{N}{s(j)}, TN + 1 \le j \le 2TN, \tag{17}$$

where MI_j is the machine index of the $j - TN$ th task, the mean of *N*, *TN* and $s(j)$ consistent with those in Eq. (16). This inverse conversion process is detailed in Algorithm 3.

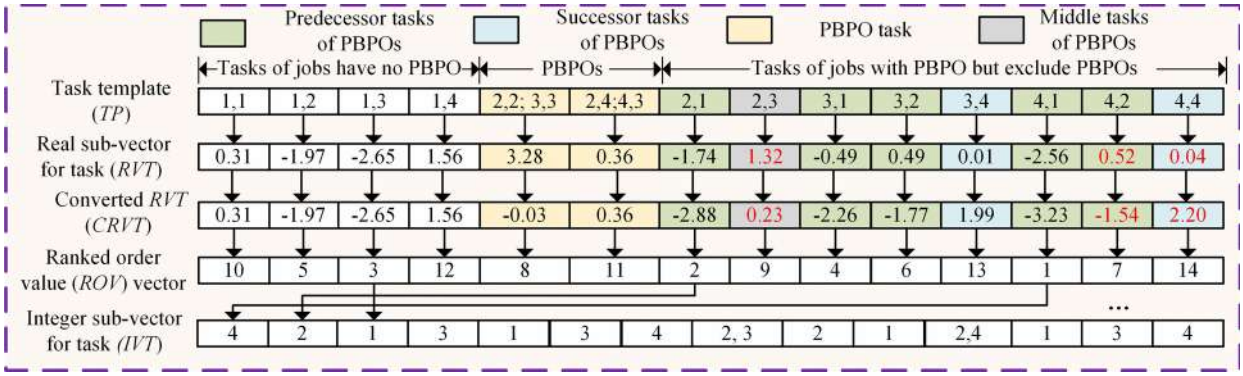


Fig. 4. Example of the segmented conversion from *RVT* to *IVT*.

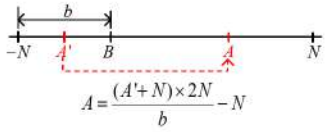
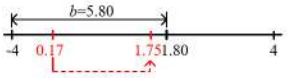
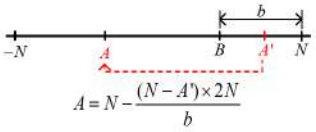
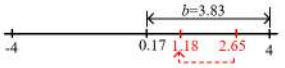
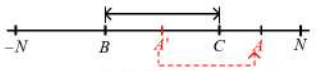
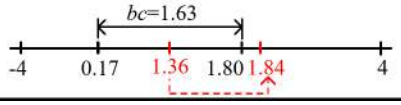
Type	Diagram and conversion formula	Example
Predecessor	 $A=\frac{(A'+N)\times 2N}{b}-N$	B=1.80 represents value of PBPO {O_{24}, O_{43}} in the RVT, while $A'=0.17$ denotes the value of PBPO {O_{22}, O_{33}} as a job predecessor task of {O_{24}, O_{43}}, then $A'=\frac{(0.17+4)\times 8}{5.80}-4=1.75$ 
Successor	 $A=N-\frac{(N-A')\times 2N}{b}$	$A'=2.65$ denotes the value of O_{34} as an job successor task of PBPO {O_{22}, O_{33}}, then $A=4-\frac{(4-2.65)\times 8}{4-0.17}=1.18$  $A=4-\frac{(4-2.65)\times 8}{4-0.17}=1.18$
Middle	 $A=\frac{(A'-B)\times 2N}{bc}-N$	B=0.17 and C=1.80 represents the values of PBPO {O_{22}, O_{33}} and {O_{24}, O_{43}} in the RVT respectively. $A'=1.36$ is the value of O_{23} as a middle task between the two PBPOs, then $A=\frac{(1.36-0.17)\times 8}{(1.80-0.17)}-4=1.84$ 

Table 3. Categorization and post-processing strategy of a *RVT* for inverse conversion.

RVT	-2.16	-2.11	-1.61	-0.27	-0.26	-0.18	-0.13	0.17	1.36	1.64	1.80	2.43	2.65	3.51
IVT	4,1	2,1	1,1	3,1	1,2	3,2	4,2	2,2; 3,3	2,3	1,3	2,4;4,3	1,4	3,4	4,4
TP	1,1	1,2	1,3	1,4	2,2; 3,3	2,4;4,3	2,1	2,3	3,1	3,2	3,4	4,1	4,2	4,4
ROV vector	3	5	10	12	8	11	2	9	4	6	13	1	7	14
Reordered *RVT*	-1.61	-0.26	1.64	2.43	0.17	1.80	-2.11	1.36	-0.27	-0.18	2.65	-2.16	-1.13	3.51
New*RVT*	-1.61	-0.26	1.64	2.43	1.75	1.80	-0.37	1.84	3.16	3.33	1.18	-1.46	-0.04	2.22

Fig. 5. Example of the inverse conversion from *IVT* to *RVT*.

1: Input the template TP, $X_p^{'}=[IVT,IVM]$ and its corresponding sequence of machine index MI
2: Obtain the corresponding ROV vector of IVT based on the TP
3: Re$orderedRVT$ ←Randomly generate a RVT vector, and reorder the RVT according to the ROV vector
4: Copy Re$ordered\ RVT$ to $NewRVT$
5: Let $TP_j, t \le j \le t'$ correspond to positions in the second section of TP
6: Sequentially update the **Predecessors**, **Middles** and **Successors** of $TP_j, t \le j \le t'$ in the third section of $NewRVT$ according to the strategies in Table 3
7: Sequentially update the **Predecessor** of $TP_j, t \le j \le t'$ in the second section of $NewRVT$ according to the strategy in Table 3
8: Compute $RVM_j, TN+1 \le j \le 2TN$ using Eq.(17)
9: Output $X_p=[NewRVT, RVM]$

Algorithm 3. Inverse conversion.

Figure 5 illustrates an example of the inverse conversion from *IVT* to the *RVT*. First, the positions of each *TP* element within the *IVT* are recorded to construct the *ROV* vector. For instance, the 12th and 7th elements of *TP*, with values (4, 1) and (2, 1) respectively, rank first and second in the *IVT*. Consequently, the 12th and 7th positions in the *ROV* vector are assigned the values 1 and 2, respectively. Next, a *Reordered RVT* is generated by sorting the *RVT* according to the *ROV* vector. For example, as shown in the figure, the first and second elements of the *ROV* vector are 3 and 5, respectively, so the third and fifth elements of the *RVT* are placed into the first and second positions of the *Reordered RVT*. Subsequently, the values in the *Reordered* RVT are converted according to the strategies outlined in Table 3 to generate *NewRVT*. The conversion process first applies to non-PBPO predecessor tasks, O_{23} (between the two PBPOs $\{O_{22}, O_{33}\}$, $\{O_{24}, O_{43}\}$), as well as the successor tasks for both PBPOs. Then, the conversion is performed for $\{O_{22}, O_{33}\}$, the predecessor tasks for $\{O_{24}, O_{43}\}$. The red-highlighted text in the *Reordered RVT* and *NewRVT* in Fig. 5 indicates the corresponding values before and after the inversion conversion, as shown in Table 3.

Semi-active decoding

Bierwirth and Mattfeld[67] introduced decoding methods that transform encoded permutations into semi-active, active, non-delay, and hybrid schedules. Among these, the semi-active schedule is particularly straightforward to implement, provides high decoding efficiency, and frequently produces high-quality solutions. Therefore, the semi-active decoding is adopt to decode the $X'_p = [IVT\ IVM]$ for a walrus. This decoding approach ensures that each task adheres to the precedence constraints both within the same job and on the same machine. However, for the FJSP_PBPO, it is crucial to thoroughly evaluate the completion times of all predecessors for each job in the PBPO. The constraints considered during the decoding process become more complex. The specific semi-active decoding designed for FJSP_PBPO is outlined in Algorithm 4.

1: Input integer sub-vector for tasks (*IVT*) and assignment machine (*IVM*)
2: Determine the task sequence (*TS*) and the processing time sequence (*PTS*)
3: Initialize $CMP = zeros(M)$, and scheduling plan $SPlan = \varnothing$, k=1// M is the number of machines
4: **for** $k = 1:|IVT|$
5: Retrieve the task u, machine m and processing time P_u^m from $TS(k)$, $IVM(k)$ and $PTS(k)$, respectively.
6: $S_u = \max\{C_{u'}, CMP[m]\}, u' \in JP[u], 1 \le m \le M, C_u = S_u + P_u^m$ // $JP[u]$ is the immediate job predecessor(s) of task u
7: $CMP[m] = C_u, SPlan = SPlan \cup \{u, m, S_u, C_u\}$
8: $k = k + 1$
9: **end for**
10: Return *SPlan* and the maximum C_u

Algorithm 4. Semi-active decoding.Random optimal matching-based initialization

To quickly obtain high-quality initial solutions, it is necessary to comprehensively balance the quality and diversity of the initial population during the initialization. For the optimization of FJSP_PBPO, the quality of the corresponding initial population is related to the matching of task and machine assignment. Based on this, a ROMI method is proposed to initialize the population. Specifically, for each walrus $X_p, 1 \le p \le P/2$ in the first half of the population, one *RVT* and *K RVM* are generated randomly accordingly to $RVT = -N + rand(2N)$, $RVM_k = -N + rand(2N), 1 \le k \le K$, respectively, where N is the number of jobs. Then, segmented conversion and semi-active decoding are employed to determine the makespan of the matched RVT and $RVM_k, 1 \le k \le K$, and the pair of *RVT* and $RVM_{k'}$ with the smallest makespan is selected to initialize $X_p = [RVT, RVM_{k'}], 1 \le p \le P/2$.

Conversely, for each walrus in the second half of the population $X_p, P/2 + 1 \le p \le P$, *K RVT* and one *RVM* are generated randomly accordingly to $RVT_k = -N + rand(2N), 1 \le k \le K, RVM = -N + rand(2N)$, respectively. Similarly, the pair of $RVT_{k'}$ and *RVM* with the smallest makespan is selected to construct $X_p = [RVT_{k'}, RVM]$. Since the *RVT* in the first half and the *RVM* in the second half of the population are generated randomly, the randomness and diversity of the initial population generation are ensured.

Enhanced feeding strategy

In the original feeding strategy of WaOA, each walrus moves toward the strongest individual X_{str} in the population. And a random number $rand_{p,j}$ within the interval (0,1) controls how each dimension of a walrus approaches the X_{str}, limiting the solution space. Preliminary experiments indicate that when solving the FJSP_PBPO, walruses often get stuck in local optima and experience a slower convergence speed. To enhance the global search capability and efficiency of WaOA, Lévy flight[68] is incorporated into the feeding strategy. Many animals, including walruses, perform fine-grained searches within a localized area for a period, followed by longer movements to explore other regions. Lévy flight, which alternates between short-distance searches and occasional long-range moves, effectively models this behavior and aligns well with the natural feeding patterns

of walruses. The feeding strategy, now integrated with Lévy flight for updating walrus positions, can be expressed by modifying the previous formula as follows:

$$X_{p,j}^{t+1} = X_{p,j}^{t} + sign[rand_{p,j} - 0.5] \times (X_{str,j} - I_{p,j} \times X_{p,j}^{t}) \oplus Levy(s), \quad 1 \le p \le P, 1 \le j \le D, \quad (18)$$

where *sign* [*rand* – 0.5] can take one of three values: – 1, 0, or 1. $\oplus$ means entry wise multiplication.

Lévy flight is a kind of non-Gaussian random process, and its step length obeys a Lévy distribution.

$$Levy(s) \sim |s|^{-1-\beta}, \quad 0 < \beta \le 2, \quad (19)$$

where *s* represents the step length of the Lévy flight, and β is an index parameter. The value of *s* can be calculated using Mantegna's algorithm as follows:

$$s \sim \frac{u}{|v|^{1/\beta}}, \quad (20)$$

where β is set to be 1.5, and both μ and v follow normal distributions.

$$\sigma = [\frac{\Gamma(1+\beta) \times sin(\pi\beta/2)}{\Gamma((1+\beta)/2) \times \beta \times 2^{(\beta-1)/2}}]^{1/\beta}, \quad (21)$$

where Γ denotes the standard Gamma function. According to Eqs. (18)–(21), Eq. (18) can be reformulated as:

$$X_{p,j}^{t+1} = X_{p,j}^{t} + sign[rand_{p,j} - 0.5] \times \frac{u}{|v|^{1/\beta}} \times (X_{str,j} - I_{p,j} \times X_{p,j}^{t}), \quad 1 \le p \le P, 1 \le j \le D. \quad (22)$$

Enhanced migration strategy

In the original migration strategy of the WaOA, each walrus randomly selects another walrus from the population as its migration destination. Throughout the iteration process, the updates of the walruses lack an adaptive adjustment mechanism, leading to slower convergence speeds or entrapment in local optima. To enhance global exploration in the early stages and strengthen local exploitation in the later stages of iteration, the migration strategy of WaOA is modified by introducing a self-adjusting factor *C* to replace *rand* in Eq. (12). The position update formula for walrus can then be rewritten as:

$$X_{p,j}^{t+1} = \begin{cases} X_{p,j}^{t} + C \times (X_{k,j}^{t} - I_{i,j} \times X_{p,j}^{t}), 1 \le p \le P, 1 \le j \le 2TN, \ if\ F(X_{k,j}^{t}) \ge F(X_{p,j}^{t}) \\ X_{p,j}^{t} + C \times (X_{p,j}^{t} - X_{k,j}^{t}), 1 \le p \le P, 1 \le j \le 2TN, \ if\ F(X_{k,j}^{t}) < F(X_{p,j}^{t}) \end{cases}, \quad (23)$$

$$C = \left[\frac{1}{2} + \cos\left(\frac{\pi}{2} \times \frac{t}{T}\right)\right] \times rand, \quad (24)$$

where *t* denotes the current iteration number and *T* represents the maximum number of iterations. In the early stages, when *t* is relatively small, the value of *C* is large, allowing individuals to explore with a greater step size during position updates. This facilitates rapid coverage of a broader search space and enhances global exploration. As the iterations progress, *t* gradually increases while *C* decreases. In the later stages, a smaller value of *C* promotes fine local exploitation within these improved regions, enhancing the algorithm's convergence efficiency.

Enhanced fleeing strategy

The original update expression for the fleeing strategy is shown in Eqs. (14) and (15). However, the local lower bound and upper bound in the Eq. (14) is controlled by the inverse proportional function $lb_{local,j}^{t} = lb_j/t$ and $ub_{local,j}^{t} = ub_j/t$ respectively. The inverse proportional function prioritizes global exploration at the beginning of the algorithm's iterations, with a larger radius to discover optimal regions within the search space. However, the neighborhood radius of the inverse proportional function decays rapidly, leading to a quick decline in global exploration capability, which makes it difficult for the fleeing strategy to play an exploitation role in the later stages of the algorithm. Therefore, this study replaces the original inverse proportional function with an arctangent function to control the local bound in fleeing. Then the fleeing strategy can be mathematically modelled by the Eq. (25).

$$X_{p,j}^{t+1} = X_{p,j}^{t} + \left(lb_{local,j}^{t} + \left(ub_{local,j}^{t} - rand \cdot lb_{local,j}^{t}\right)\right)$$
$$new\ local\ bound: \begin{cases} lb_{local,j}^{t} = lb_j \times \frac{\frac{\pi}{2} - \arctan(\frac{t-0.5 \cdot T}{0.03 \cdot T})}{\pi} \\ ub_{local,j}^{t} = ub_j \times \frac{\frac{\pi}{2} - \arctan(\frac{t-0.5 \cdot T}{0.03 \cdot T})}{\pi} \end{cases}, \quad (25)$$

where the mean of *t* and *T* consistent with those in Eq. (24).

Gathering strategy

Walruses enhance their foraging and movement efficiency by interacting and sharing location information with one another. To model this behavior, we propose a "gathering strategy" in which walruses form pairs and exchange information, thereby improving the herd's ability to identify areas with higher food availability. To assess this information-sharing process, walruses are paired through a random selection method. Based on these paired walruses, such as X_p^t and $X_{p'}^{\prime t}$, the position of each individual is updated according to the following Algorithm 5.

Input $X''_p = (X''_{p,1}, X''_{p,2}, \ldots, X''_{p,j}, \ldots, X''_{p,2TN})$, $X''_{p'} = (X''_{p',1}, X''_{p',2}, \ldots, X''_{p',j}, \ldots, X''_{p',2TN})$, $1 \le j \le 2TN$

Create a vector $X''^{t+1}_{p'}$ of length $2TN$ initialized with zeros

for $i = 1 : TN$

  if $R_i = U(0,1) > 0.5$

    $X''^{t+1}_{p,j} \leftarrow X''_{p,1}$, $X''^{t+1}_{p,j+TN} \leftarrow X''_{p,1+TN}$, find the first occurrence index j' of $X''_{p,1}$ in $X''_{p'}$, delete $X''_{p,1}, X''_{p,1+TN}, X''_{p',j'}, X''_{p',j'+TN}$

  else

    $X''^{t+1}_{p,j} \leftarrow X''_{p',1}$, $X''^{t+1}_{p,j+TN} \leftarrow X''_{p',1+TN}$, find the first occurrence index j' of $X''_{p'}$ in $X''_{p,1}$, delete $X''_{p',1}, X''_{p',1+TN}, X''_{p,j'}, X''_{p,j'+TN}$

end

Output $X''^{t+1}_p \leftarrow \arg\max\{F(X''^t_p), F(X''^t_{p'}), F(X''^{t+1}_p)\}$

Algorithm 5. Gathering strategy.

A control factor, denoted as *A*, is introduced to regulate the population's updating strategy. If *A* reaches or exceeds 0.4 after the feeding strategy, walruses will adopt migration, fleeing strategies to explore and exploit search area. Conversely, if *A* falls below 0.4, a gathering strategy will be employed. In this strategy, walruses search for new territories in pairs. Multiple pairs will form within the walrus population, thereby further enhancing search range. The value of *A* is controlled by the Eq. (26):

$$A = \frac{e - e^{[(t-1)/T]^2}}{e-1} \cdot \left|\sin(2\pi \times rand)\right|, \tag{26}$$

where the *rand* denotes a random number between 0 and 1.

Enhanced WaOA for FJSP_PBPO

Based on the above improvements, the pseudocode of the proposed eWaOA for FJSP_PBPO in this study is outlined in Algorithm 6. The eWaOA is initialized using the ROMI strategy, and the mathematical models for the enhanced feeding, migration, fleeing strategies are shown in Eqs. (18)–(25), in combination with Eqs. (11), (13), and (15). The gathering strategy is described in Algorithm 5.

Input flexible process plan of each job, the PBPOs, set population size P, max iterations max$T$, A=0, t=1
Initialize P walruses $(X_p^0, X_p'^0)$, $1 \le p \le P$, based on the ROMI approach. Let X_{str} is the best search agent
for t=1:max$T$
  Phase 1: Feeding strategy (exploration)
    Calculate $X_{p,j}^{t+1}, 1 \le p \le P, 1 \le j \le 2TN$ using Eq. (22)
    Convert $X_p^{t+1}, 1 \le p \le P$ to $X_p'^{t+1}, 1 \le p \le P$ by the Algorithm 2 in Section 5.3
    Decode $X_p'^{t+1}, 1 \le p \le P$ and evaluate each walruses by Algorithm 4 in Section 5.4
    Update $X_p^{t+1}, 1 \le p \le P$ using Eq. (11)
  if A$\leqslant$0.4 **then**
    Phase 2: Migration strategy(exploration)
      Randomly choose an immigration destination X_k^t for the pth walrus $X_p^t, 1 \le p \le P, p \ne k$
      Calculate $X_{p,j}^{t+1}, 1 \le p \le P, 1 \le j \le 2TN$ using Eq. (23)
      Convert $X_p^{t+1}, 1 \le p \le P$ to $X_p'^{t+1}, 1 \le p \le P$ by the Algorithm 2 in Section 5.3
      Decode $X_p'^{t+1}, 1 \le p \le P$ and evaluate each walruses by Algorithm 4 in Section 5.4
      Update $X_p^{t+1}, 1 \le p \le P$ using Eq. (13)
    Phase 3: Fleeing strategy (exploitation)
      Calculate $X_{p,j}^{t+1}, 1 \le p \le P, 1 \le j \le 2TN$ using (25)
      Convert $X_p^{t+1}, 1 \le p \le P$ to $X_p'^{t+1}, 1 \le p \le P$ by the Algorithm 2 in Section 5.3
      Decode $X_p'^{t+1}, 1 \le p \le P$ and evaluate each walruses by Algorithm 4 in Section 5.4
      Update $X_p^{t+1}, 1 \le p \le P$ using Eq. (15)
  end if
  if A>0.4 **then**
    Phase 2: Gathering strategy (enhancing phase)
      Generate $X_p'^{t+1}$, $X_{p'}'^{t+1}$ using Algorithm 5 based on the randomly selected $X_p'^t$ and $X_{p'}'^t$
      Convert $X_p'^{t+1}$ to X_p^{t+1}, $X_{p'}'^{t+1}$ to $X_{p'}^{t+1}$ by the Algorithm 3 given in Section 5.3
  end if
  Update X_{str} if there is a better solution
  t=t+1
  Compute A using Eq. (26)
end for
Output X_{str} and its objective function value makespan

Algorithm 6. eWaOA for FJSP_PBPO.

Computational complexity analysis

In the proposed eWaOA, the key components contributing to computational complexity include conversion and inverse conversion, population initialization, decoding, an enhanced feeding strategy, an enhanced migration strategy, an enhanced fleeing strategy, and a gathering strategy. Notably, population initialization, decoding, as well as the enhanced feeding, migration, fleeing, and gathering strategies, all involve either conversion or inverse conversion operations. Let P denote the population size, D represent the dimension of each individual, TN denote the total number of all tasks in the FJSP_PBPO (where $D=2TN$), K be the parameter in the ROMI strategy, and T be the maximum number of iterations.

Conversion involves task template partitioning, data duplication, updating *CRVT*-related values, constructing the *IVT*, and determining machine indices, each with a time complexity of O(D). The corresponding inverse conversion process includes obtaining the *ROV* vector, generating random vectors, updating *NewRVT* values, and computing the *RVM*, all with a time complexity of O(D). Therefore, the time complexity of both conversion and inverse conversion is $O(D)$. For decoding, the primary time consumption is spent iterating through the task

sequence. For each task, it is necessary to retrieve the task, machine, and processing time, as well as determine the task's start and completion time. Since the loop runs for a total of *TN* tasks, the time complexity is O(*TN*).

The computational complexity of population initialization using the ROMI strategy is O(*PKD*). For each individual in the first half of the population (a total of *P*/2 individuals), 1 task sequence random vectors (*RVTs*) and *K* machine assignment random vectors (*RVMs*) are generated. For each individual in the second half of the population (a total of *P*/2 individuals), *K RVTs* and 1 *RVM* are generated. The time complexity of generating each random vector is O(*TN*). Since the initialization process for each individual involves both conversion and decoding operations, their respective time complexities are O(*D*) and O(*TN*). Therefore, the total time complexity of generating random vectors is $O(P/2(1+K)\times(TN+TN+D)+P/2(1+K)\times(TN+TN+D)=O(PKD)$.

The enhanced feeding, migration, fleeing, and gathering strategies each have a computational complexity of O(PD) per iteration. This complexity arises because, for each dimension of every walrus, calculating the new position involves operations such as multiplication, addition, and random number generation, all with a constant time complexity of O(1). Furthermore, the conversion or inverse conversion process for each walrus has a complexity of O(*D*). Given a population size of *P*, updating the positions of all walruses leads to a total time complexity of O(*PD*). Hence, the overall complexity of these strategies is O(*PD*).

Therefore, the overall computational complexity of the eWaOA is expressed as O(*PDK* + *TPD*). Since the value of *K* is generally much smaller than the maximum number of iterations *T*, the total computational complexity can be simplified to O(*TPD*).

Computational experiments and real-word case study

We first develop 30 test instances based on existing benchmark FJSP instances. We then compare the performance of original WaOA with WaOA that incorporates the ROMI initialization strategy (WaOA-R) to assess the effectiveness of the ROMI approach. Subsequently, we design and conduct experiments with four enhanced WaOA variants and eleven state-of-the-art (SOTA) metaheuristic algorithms across these test instances, followed by a real-world engineering case study to evaluate the superiority of eWaOA. To ensure the stability and reliability of the results and minimize the effects of randomness, we run each test instances and engineering case ten times. The experiments are performed using MATLAB R2018a on a desktop computer equipped with an Intel Core i7-8700 processor, 16 GB of RAM, and Windows 10.

Instance generation

Due to the absence of benchmark for FJSP_PBPO, this study extends the MK01–MK15 benchmarks provided by Brandimarte[2] by introducing one or two randomly selected PBPOs to create new test instances, resulting in the EMK01-EMK15 benchmarks for FJSP_PBPO. Detailed information about these PBPOs is presented in Table 4.

All other data remain consistent with the original benchmarks. For example, in EMK02, the first PBPO consists of tasks O_{34} and O_{93}, which can be processed on machines 2, 5, and 6 with processing times of 4, 2, and 3 units, respectively. The second PBPO includes tasks O_{82} and $O_{10(3)}$, which are processed on machines 4 and 6 with processing times of 5 and 3 units, respectively. Test instances are identified with the suffixes "s" and "d", where "s" denotes instances considering only the first PBPO, and "d" denotes instances that include both PBPOs. Accordingly, the test instances are EMK01(s)-EMK15(s) and EMK01(d)-EMK15(d), totaling 30 instances. The extension to additional PBPOs follows the same principle as the scenario with two PBPOs, as these two PBPOs are generated randomly.

Parameter setting and notations

The performance of an algorithm is significantly influenced by its parameter configurations, which are selected based on extensive experimental validation and practical experience to ensure optimal results within a reasonable time. In this study, the parameters are configured as follows: the walrus population size is set to 200, maximum iterations is 250, different *T* are assigned based on the scale of each case when using time limit as the termination criterion, the parameter *K* in the ROMI is set to 7, as determined by the experiments discussed in "Effectiveness of ROMI" section.

To facilitate subsequent discussions and analyses, this paper standardizes the naming and descriptions for the algorithm variants as follows: WaOA-R denotes the original WaOA enhanced with ROMI strategy; WaOA-RF builds upon WaOA-R by incorporating the enhanced feeding strategy; WaOA-RFM further advances WaOA-RF by implementing the enhanced migration strategy; WaOA-RFMF, in turn, adds the improved fleeing strategy to WaOA-RFM. Finally, eWaOA integrates the gathering strategy into WaOA-RFMF. For performance evaluation, the following metrics are used for quantitative analysis in this section.

- $B(C_{max})$: the best makespan achieved across ten runs, assessing the optimal performance potential of algorithm.
- *Av*: the average makespan over ten runs, indicating the algorithm's overall performance.
- *Sd*: the standard deviation of C_{max} across ten runs, measuring performance of stability and consistency.
- *RPD* (%): the relative percentage difference between the current algorithm and the best-performing algorithm, calculated as $RPD = 100\% \times (C_{\max} - Min)/Min$, where *Min* is the smallest C_{max} value obtained by all algorithms on the same test instance. A lower *RPD* value signifies closer proximity to the optimal solution and better search capability.
- *SdMean*: the average *Sd* value for each algorithm across all test instances of varying sizes, reflecting the algorithm's stability and consistency across different problem scales.
- *RPDMean*: the average *RPD* value across all test instances of varying sizes, providing a comprehensive evaluation of the algorithm's search capabilities across diverse problem scales.

Instances	PBPOs	Alternative machines	Processing times
EMK01	O_{34},O_{75}	M2/M4	3/5
	O_{54},O_{95}	M2/M4	3/5
EMK02	O_{34},O_{93}	M2/M5/M6	4/2/3
	O_{82},$O_{10(3)}$	M4/M6	5/3
EMK03	O_{47},O_{54}	M2/M3/M4/M5	18/13/5/10
	O_{57},O_{82}	M4/M7/M8	13/2/18
EMK04	O_{34},O_{63}	M3/M4/M7	9/4/5
	O_{15},$O_{11(3)}$	M7/M8	5/9
EMK05	O_{32},O_{62}	M2/M3/M4	6/9/5
	O_{33},$O_{12(4)}$	M1/M2/M3	8/6/7
EMK06	O_{14},O_{37}	M2/M7	8/5
	O_{22},O_{94}	M1/M4/M6	6/6/2
EMK07	O_{12},O_{33}	M1/M2	5/1
	O_{15},O_{92}	M2/M3	4/8
EMK08	O_{74},O_{45}	M1/M10	10/19
	O_{38},$O_{17(5)}$	M3/M4	19/5
EMK09	O_{77},$O_{12(5)}$	M2/M6/M8	16/10/17
	$O_{12(8)}$,$O_{15(6)}$	M2/M3/M9	12/11/6
EMK10	O_{27},$O_{8(10)}$	M2/M6/M7	5/5/15
	O_{57},$O_{13(5)}$	M2/M4/M7	16/13/14
EMK11	O_{33},O_{54}	M4/M5	17/18
	O_{65},$O_{10(4)}$	M3/M4	28/22
EMK12	O_{13},O_{57}	M5/M10	22/15
	O_{15},O_{71}	M5/M7	18/24
EMK13	O_{84},$O_{10(8)}$	M1/M9	29/29
	$O_{15(6)}$,$O_{16(5)}$	M2/M10	21/18
EMK14	$O_{14(2)}$,$O_{15(7)}$	M4/M13	16/10
	$O_{20(1)}$,$O_{22(4)}$	M5/M9	25/28
EMK15	O_{31},O_{68}	M2/M15	25/27
	O_{37},O_{44}	M3/M4	24/28

Table 4. Description of the test instances.

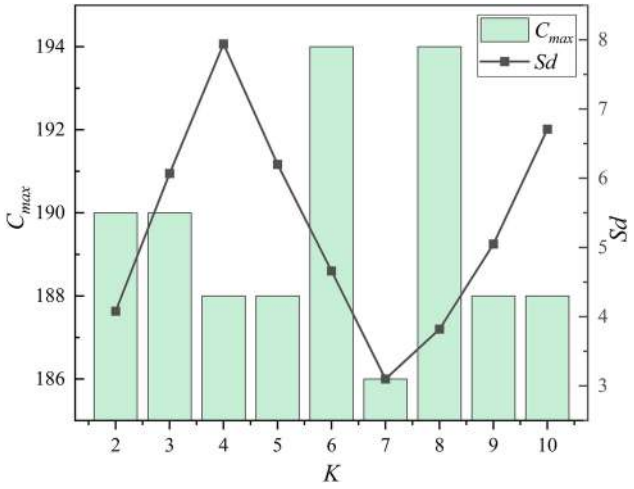


Fig. 6. Optimal selection of ROMI strategy parameter "*K*".

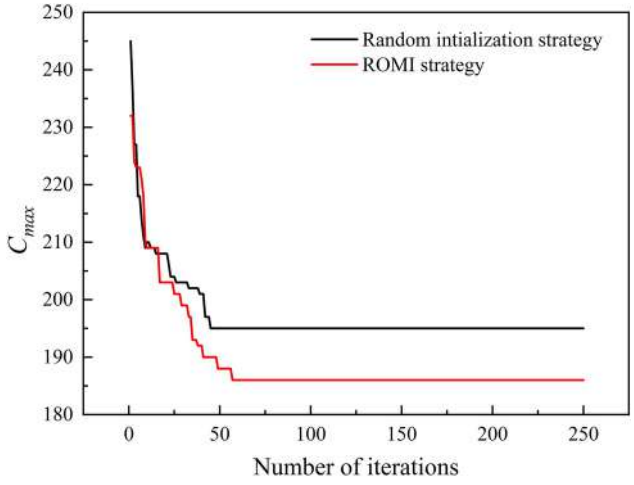


Fig. 7. Convergence for two initialization strategies in EMK05(s).

Instances	WaOA			WaOA-R			Instances	WaOA			WaOA-R		
	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*		$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*
EMK01(s)	47	49.1	2.4	**46**	**48.1**	**1.7**	EMK08(d)	550	569.4	**13.7**	**536**	**558**	14.7
EMK01(d)	44	47.7	3.1	**42**	**45.6**	3.1	EMK09(s)	428	447.5	12.3	**423**	**443.1**	**11.1**
EMK02(s)	40	45.2	2.8	**33**	**37.5**	**2.1**	EMK09(d)	423	458.7	**16.5**	**415**	**449.3**	18.2
EMK02(d)	39	42.8	2.6	**36**	**38.4**	**2.2**	EMK10(s)	368	384.6	15.8	**355**	**378.2**	**14.2**
EMK03(s)	238	258.6	**9.4**	**236**	**249.8**	11.2	EMK10(d)	**352**	395.8	16.9	357	**384.9**	**11.3**
EMK03(d)	260	274.3	**6.4**	**231**	**251**	10.0	EMK11(s)	**675**	691.2	20.4	**675**	**689.3**	**17.2**
EMK04(s)	82	95.3	3.9	**75**	**80**	**2.9**	EMK11(d)	676	689.4	**17.6**	**654**	**673.4**	19.8
EMK04(d)	82	92.7	4.2	**78**	**80.8**	**2.1**	EMK12(s)	564	598.3	23.7	**553**	**584.5**	**21.2**
EMK05(s)	195	203.8	5.1	**186**	**196**	**3.2**	EMK12(d)	**579**	**591.3**	24.8	589	599.1	**20.4**
EMK05(d)	202	224.6	8.5	**189**	**197**	**6.7**	EMK13(s)	569	625.2	40.7	**558**	**617.4**	**33.5**
EMK06(s)	124	148.3	7.5	**112**	**125**	**7.0**	EMK13(d)	**576**	**623.8**	**21.6**	589	633.3	24.3
EMK06(d)	122	134.5	11.7	**119**	**130.1**	**8.9**	EMK14(s)	758	792.3	32.8	**752**	**788.2**	**28.6**
EMK07(s)	173	187.4	8.5	**173**	**185.6**	**8.2**	EMK14(d)	**773**	810.2	36.2	778	**806.6**	**31.2**
EMK07(d)	194	208.7	10.4	**175**	**189.2**	**10.2**	EMK15(s)	567	587.5	**24.1**	**562**	**584.1**	25.2
EMK08(s)	566	584.1	**10.5**	**547**	**568**	14.8	EMK15(d)	532	581.1	27.2	**521**	**572.8**	**19.8**

Table 5. Comparison between WaOA-R and WaOA. Significant values are in bold.

Effectiveness of ROMI

To determine the optimal value for the "*K*" in the ROMI, experiments are conducted using the EMK05(s) benchmark. The "*K*" values range from 2 to 10, resulting in nine distinct experimental setups. The results, depicted in Fig. 6, show that when "*K*" is set to 7, the algorithm consistently achieves lower $B(C_{max})$ and *Sd* values across the ten runs. Therefore, 7 is selected as the optimal parameter for the ROMI strategy. Figure 7 illustrates a detailed comparison of convergence processes for WaOA-R and WaOA, with walruses initialized by ROMI converging faster and more efficiently to a better makespan than those initialized randomly in WaOA.

Table 5 presents the comparative experimental results of WaOA-R and the original WaOA across 30 test instances. The comparison reveals that incorporating the ROMI strategy improves $B(C_{max})$ in 25 instances, with 1 instance yielding identical results and only 4 instances performing worse. For *Av*, improvements are observed in 28 instances, with only 2 instances performing worse. Regarding the *Sd* metric, 22 instances show improvement, and 8 instances perform worse. These comparisons, illustrated in Table 5 and Fig. 7, suggest that WaOA-R with the ROMI strategy not only demonstrates superior search capability but also exhibits improved stability and consistency compared to WaOA, while further enhancing the algorithm's convergence efficiency.

Comparative experiments with enhanced WaOA variants

To validate the effectiveness and advantages of the enhanced feeding, migration, and fleeing strategies and the proposed the gathering strategy, we compare the metrics $B(C_{max})$, *Av* and *Sd* across ten runs for the algorithms WaOA-R, WaOA-RF, WaOA-RFM, WaOA-RFMF and eWaOA. The experimental results are detailed in Table 6.

Table 6 shows that WaOA-RF surpasses WaOA-R in terms of $B(C_{max})$ for 20 out of 30 test instances, with 8 instances achieving identical results and only 2 instances showing slightly lower performance. For the *Av* metric,

Instances	WaOA-R			WaOA-RF			WaOA-RFM			WaOA-RFMF			eWaOA		
	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*
EMK01(s)	46	48.1	1.7	45	47.6	2.7	43	46.3	2.2	**42**	44.3	1.9	**42**	**42.0**	**0.0**
EMK01(d)	42	45.6	3.1	42	46.3	2.8	39	43.6	1.8	40	42.8	1.5	**39**	**39.1**	**0.3**
EMK02(s)	33	37.5	2.1	35	39.2	2.5	34	37.9	2.3	34	37.1	2.0	**27**	**28.1**	**0.7**
EMK02(d)	36	38.4	2.2	36	41.3	3.3	35	38.5	2.7	32	35.8	2.6	**28**	**28.6**	**0.8**
EMK03(s)	236	249.8	11.2	230	248.1	8.5	227	241	7.1	226	234.4	7.4	**204**	**204.0**	**0.0**
EMK03(d)	231	251	10.0	232	252.9	9.7	228	245.4	8.3	217	229.2	7.5	**187**	**187.0**	**0.0**
EMK04(s)	75	80	2.9	74	82.1	3.2	73	79	2.5	73	76	**2.5**	**66**	**68.1**	**2.5**
EMK04(d)	78	80.8	2.1	74	78.3	1.6	74	77.7	2.9	73	75.8	**2.4**	**66**	**68.3**	2.5
EMK05(s)	186	196	5.4	186	192.2	4.7	186	193.7	4.1	184	187.7	3.3	**173**	**175.3**	**1.7**
EMK05(d)	189	197	6.7	188	195.7	6.8	188	193.5	5.7	180	188.4	4.1	**171**	**172.7**	**1.0**
EMK06(s)	112	125	7.0	112	123.9	8.5	111	121.9	6.8	107	119.8	6.0	**72**	**75.8**	**2.7**
EMK06(d)	119	130.1	8.9	117	125.4	8.5	114	122.1	10.6	110	119.6	8.1	**69**	**75.7**	**3.5**
EMK07(s)	173	185.6	8.2	170	184.8	11.8	175	183.4	9.5	166	179.1	8.2	**138**	**142.5**	**2.5**
EMK07(d)	175	189.2	10.2	173	185.5	8.4	169	183.3	6.0	171	178.9	5.2	**137**	**142.2**	**3.7**
EMK08(s)	547	568	14.8	545	565.6	16.5	545	563	14.5	533	554.8	13.6	**523**	**532.0**	**3.0**
EMK08(d)	536	558	14.7	536	557.5	18.9	534	557.3	17.2	523	545.2	16.7	**513**	**521.0**	**4.0**
EMK09(s)	423	443.1	11.1	409	428.7	21.6	387	413.8	21.0	380	402.2	21.3	**319**	**327.8**	**6.3**
EMK09(d)	415	449.3	18.2	410	441.2	17.0	407	436.1	15.3	391	413.5	14.9	**318**	**329.9**	**5.7**
EMK10(s)	355	378.2	14.2	351	371.3	15.6	347	368.5	19.5	334	370.8	20.3	**241**	**250.6**	**7.7**
EMK10(d)	357	384.9	11.3	347	375.1	12.7	338	366.6	18.4	325	359.7	20.7	**228**	**248.0**	**3.9**
EMK11(s)	675	689.3	17.2	661	681.8	16.4	656	673.3	14.1	639	664.3	13.7	**615**	**619.2**	**2.9**
EMK11(d)	654	673.4	19.8	654	673.2	15.2	646	671.1	15.9	646	666.7	14.4	**613**	**624.5**	**2.5**
EMK12(s)	553	584.5	21.2	542	576	14.7	535	572	10.5	540	554.6	10.0	**508**	**513.8**	**9.2**
EMK12(d)	589	599.1	20.4	571	587.4	22.8	567	579.1	19.7	532	561.5	19.9	**508**	**517.5**	**7.1**
EMK13(s)	558	617.4	33.5	558	603	27.4	554	598.1	20.6	552	578.8	19.9	**421**	**452.1**	**9.1**
EMK13(d)	589	633.3	24.3	572	602.9	17.5	544	595.5	17.1	568	592.6	16.7	**417**	**448.6**	**6.8**
EMK14(s)	752	788.2	28.6	739	787.2	27.3	745	784.6	26.4	**694**	730.6	20.3	**694**	**694.0**	**0.0**
EMK14(d)	778	806.6	31.2	757	783.4	21.8	745	784.6	21.3	**694**	734.9	22.7	**694**	**694.0**	**0.0**
EMK15(s)	562	584.1	25.2	549	562.5	23.6	520	549.7	22.0	539	567.2	21.7	**366**	**395.9**	**5.1**
EMK15(d)	521	572.8	19.8	521	571.7	18.3	519	547	17.5	549	570.8	21.4	**382**	**404.6**	**5.0**

Table 6. Results obtained by different WaOA variants. Significant values are in bold.

WaOA-RF demonstrates superior performance in 25 instances compared to WaOA-R, while 5 instances exhibit relatively lower *Av* values. Regarding the *Sd* metric, WaOA-RF outperforms WaOA-R in 17 instances, with 13 instances exhibiting comparatively higher *Sd* values. These results suggest that incorporating Lévy flight into WaOA's feeding strategy enhances both makespan and solution stability. The key benefit of Lévy flight is its combination of short- and long-distance moves, which enables more effective exploration of the search space and better balance between exploration and exploitation.

The comparative analysis between WaOA-RFM and WaOA-RF highlights a significant performance improvement due to the enhanced migration strategy. For the $B(C_{max})$ metric, 24 test instances show notable improvement, with 2 instances experiencing a minor decline and 4 remaining unchanged. Similarly, in the *Av* metric, 28 test instances demonstrate performance gains, while only 2 show slight declines. Regarding the *Sd* metric, WaOA-RFM outperforms WaOA-RF in 25 instances, with 5 instances exhibiting relatively higher *Sd* values. These findings underscore the enhanced migration strategy's effectiveness in achieving an optimal makespan, along with improved stability and consistency. This improvement is likely due to the self-adjusting factor in the migration strategy, which promotes global exploration in early iterations and strengthens local exploitation in later stages.

Comparing WaOA-RFMF with WaOA-RFM reveals that the enhanced fleeing strategy positively impacts WaOA. Specifically, for the $B(C_{max})$ metric, 21 test instances show performance gains, 6 instances experience slight declines, and 3 instances remain unchanged. In the case of the *Av* metric, 27 test instances show improvements, while only 3 instances perform worse than with the original fleeing strategy. For the *Sd* metric, 23 instances show improvements, while 7 instances perform relatively worse. These results conclusively demonstrate that the fleeing strategy with arctangent function-controlled local bounds significantly outperforms the original strategy, enhancing makespan, maintaining algorithmic consistency and stability, and reducing variability across runs.

Table 6 shows that eWaOA significantly improves the $B(C_{max})$ metric in 27 test instances compared to WaOA-RFMF, with performance in the remaining 3 instances being comparable. This result robustly demonstrates eWaOA's potential in global optimization. Additionally, eWaOA consistently improves the *Av* metric across all test instances. For the *Sd* metric, eWaOA achieves lower values in 28 instances, with 1 instance showing the

same value as WaOA-RFMF, and only 1 instance showing a minor 0.1 increase. These findings strongly indicate that the gathering strategy significantly enhances WaOA's ability to achieve a globally optimal makespan while markedly improving stability and consistency across multiple executions and diverse test scenarios.

The evolutionary trajectories of these algorithms on EMK09(s) are analyzed to assess whether the refined strategies accelerate WaOA's convergence, as illustrated in Fig. 8a. This figure demonstrates that each improvement, relative to the baseline (WaOA-R), enhances convergence speed and achieves a better makespan. As shown in Fig. 8b, Curve 1, representing the difference between WaOA-RF and WaOA-R, displays fluctuations in the early and middle iterations, suggesting that Lévy flight significantly enhances WaOA's global exploration capabilities. In the later stages, the differences in results continue to increase until reaching stability, indicating that Lévy flight also strengthens exploitation in the middle and later iterations, allowing the algorithm to escape local optima through occasional long-distance moves.

Curve 2, representing the difference between WaOA-RFM and WaOA-RF, oscillates above the baseline with a broader range of values than Curve 1 during the early and middle stages. This pattern indicates that the enhanced migration strategy strengthens WaOA-R's global exploration through the introduced adaptive parameter. In the middle and later stages, the differences continue to increase, reaching values significantly higher than those of Curve 1. This trend demonstrates that the adaptive parameter also facilitates more effective local search guidance.

Curve 3, which represents the difference between WaOA-RFMF and WaOA-R, shows pronounced fluctuations in the early and middle stages before transitioning into a phase of steady, stepwise improvement. Notably, Curve 3 consistently remains above Curve 2 throughout the process. This suggests that introducing the arctangent function to replace the inverse proportional function in the fleeing strategy does not diminish the global exploration capability provided by the enhanced feeding and migration strategies. Instead, it further enhances WaOA's local search capacity in the middle and later stages. This improvement is mainly attributed to the extended global search range produced by the combined effect of the three enhanced strategies in the early and middle stages. The arctangent function expands the local bound, enabling the algorithm to perform more detailed local searches within a larger search space, thereby enhancing solution optimization in the later stages.

The combined application of all three enhanced strategies substantially improves WaOA-R's exploration and exploitation capabilities, leading to a reduced makespan. However, results indicate that WaOA still risks

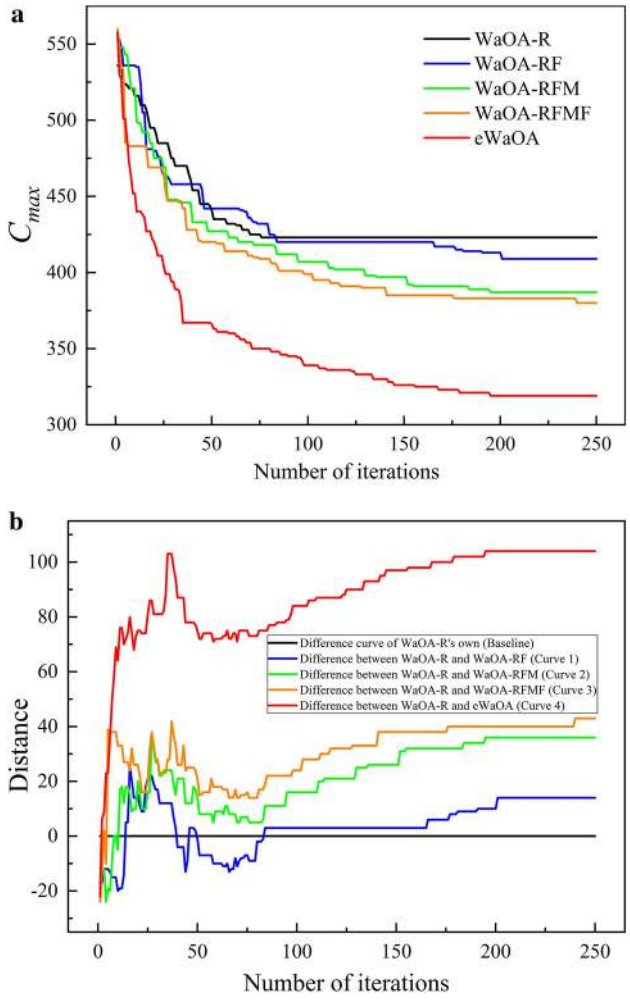


Fig. 8. The diversity curves for the five algorithms to solve EMK09(s).

becoming trapped in local optima, with local search improvements progressing slowly during the middle and later stages. The proposed gathering strategy effectively addresses this issue. As shown in Fig. 8b, Curve 4, representing the difference between eWaOA and WaOA-R, remains consistently above Curve 3 throughout the process. Alongside Fig. 8a, it is clear that the algorithm demonstrates strong global search capabilities, leading to a rapid reduction in makespan and producing significantly better results than WaOA-RFMF in the early and middle stages. In the mid-to-later stages (e.g., after 75 generations), the stepwise increases in Curve 4 occur more frequently than in Curve 3, stabilizing around 200 iterations. Overall, the gathering strategy significantly enhances both exploration and exploitation in WaOA. The primary advantage of the gathering strategy lies in its random pairing of agents for positional information exchange, which prevents excessive concentration in specific regions and reduces the risk of becoming trapped in local optima. As iterations progress into the middle and later stages, shared positional information among paired agents converges, allowing individuals to refine their positions within localized areas. This process enhances both convergence speed and solution accuracy.

Figure 9 depicts the variation in the value of A according to Eq. (26) over iterations when solving EMK09(s), showing that in the early stages, the gathering strategy is highly likely to be employed, thereby enhancing WaOA's exploration capability. Conversely, in the middle and later stages, the probability of using the gathering strategy decreases, shifting the focus toward improving exploitation. Thus, this strategy effectively balances exploration and exploitation throughout different phases. Figure 10 shows the Gantt chart for EMK09 (d), generated using the eWaOA algorithm. The chart illustrates that all operations comply with both the sequential order constraints of the process plan and the PBPO requirements. This demonstrates that the proposed methods for encoding, conversion, inverse conversion, and decoding effectively can handle the constraints of FJSP_PBPO.

Comparative experiments with other metaheuristic algorithms

Due to the novelty of FJSP_PBPO, there are no publicly available algorithms for direct comparison. Meanwhile, the eWaOA proposed in this study is a standalone algorithm rather than a hybrid one. Therefore, this study selected 11 SOTA standalone metaheuristic algorithms for evaluation. Each algorithm uniformly employs the encoding scheme, conversion scheme and semi-active decoding method. The main differences among the algorithms lie in their initialization and iterative processes. These algorithms can be categorized into four groups: evolutionary-based, swarm-based, physics-based, and human-based. Evolutionary-based algorithms mimic natural evolution using selection, crossover, and mutation to optimize solutions. The GA and DE[69], renowned for their robust global search capabilities, are the most prevalent evolutionary-based algorithms. They are widely applied in scheduling optimization and are chosen as comparison algorithms for this study. Swarm-based metaheuristic algorithms are developed by modeling the collective behaviors seen in natural phenomena. This study compares classic swarm-based metaheuristic algorithms, including PSO[44] and GWO[51], alongside newer algorithms such as HHO[56], artificial rabbits optimization (ARO)[70], and the latest WO[12]. Physics-based algorithms utilize principles from physics to address optimization challenges. In this paper, multi-verse optimization (MVO)[71] and optical microscope algorithm (OMA)[75] are chosen as comparison algorithms within the physics-based category. Human-based algorithms, inspired by human cognitive processes and behaviors, are represented here by the teaching learning based optimization (TLBO)[76] and poor and rich optimization (PRO)[72]. To eliminate variations from differences in initial candidate solutions, enhance the repeatability and stability of the experiments, and ensure fairness and consistency in evaluation, the initial population size is uniformly set to 200. Other parameters are configured according to the default settings of each algorithm, with the specific values presented in Table 7.

To comprehensively evaluate the performance of the algorithms across 30 extended test instances, we conducted two experiments: one with a fixed maximum number of iterations (Experiment 1) and the other with a time-limited termination criterion, where the algorithms terminate once the time limit is reached (Experiment 2). These experiments assess each algorithm's capability to find the global optimal solution and its

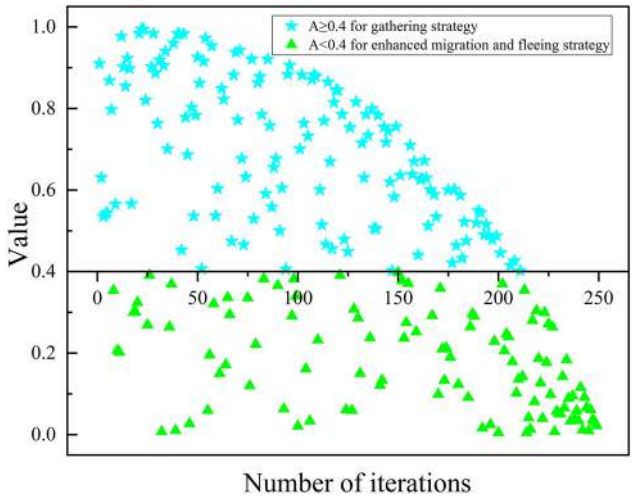


Fig. 9. The value of *A* over iterations of a run while solving EMK09(s).

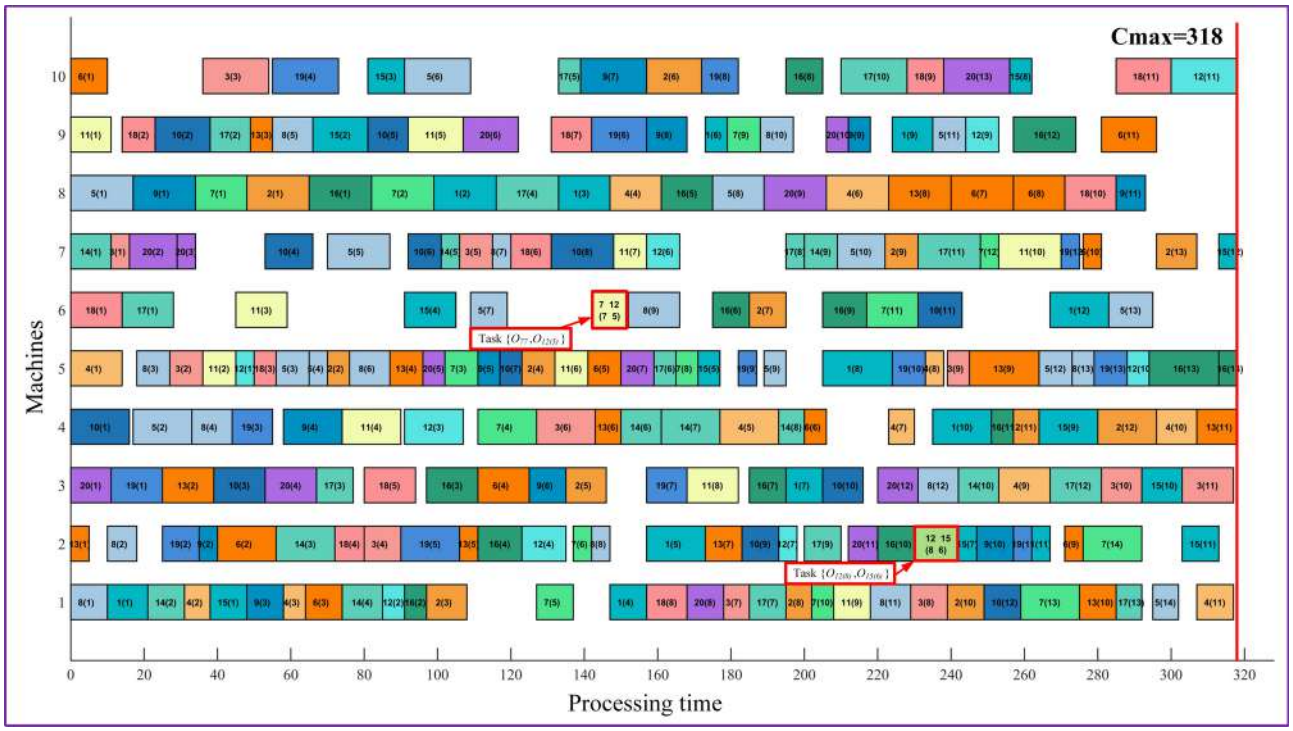


Fig. 10. Gantt chart of a scheduling scheme for EMK09 (d).

Category	Algorithm	Parameter (value)	Category	Algorithm	Parameter (value)
Evolutionary-based	GA	Crossover rate (0.8) Mutation rate (0.05)	Swarm-based	PSO	Cognitive coefficient (2) Social coefficient (2) Inertia weight (Linear reduction from 0.9 to 0.2)
	DE	Mutation factor (0.5) Crossover probability (0.2)		GWO	Convergence parameter (Linear reduction from 2 to 0)
Physics-based	OMA	–		HHO	Escape energy (Linear reduction from 2 to 0)
	MVO	WEP_Max (1) WEP_Min (0.2)		ARO	Energy factor ([0.1, 2]) Hiding parameter ([0.01, 0.5])
Human-based	TLBO	Teaching factor (1,2)		WO	Female rate ([0.4]) Danger signal (2)
	PRO	Mutation probability (0.06)			

Table 7. Parameter settings for the comparison algorithms.

trade-off between solution quality and search efficiency. We compare the metrics $B(C_{\max})$, *Av*, *Sd, RPD, SdMean, RPDMean* for all 12 algorithms.

The results of Experiment 1 are presented in Table 8. Notably, eWaOA achieves optimal values for the $B(C_{\max})$, *Av*, and *PRD* metrics across all 30 test instances. Specifically, eWaOA attains optimality in 26 instances for $B(C_{\max})$, while in the remaining 4 instances, it ties with other top algorithms. For the *Av* metric, eWaOA achieves optimality in all 30 instances. Furthermore, among the 12 algorithms, eWaOA records the lowest *Sd* value in 22 of the test instances. In terms of *SdMean* and *RPDMean* metrics across all instances, eWaOA demonstrates superior performance with values of 3.3 and 0, respectively. Table 9 presents the termination time settings (*Time*) and the results of Experiment 2. As shown, eWaOA also demonstrates strong competitiveness. Specifically, eWaOA achieves the minimum values for the $B(C_{\max})$, *Av*, and *Sd* metrics in 28, 30, and 15 instances, respectively. Notably, eWaOA also performs well in the *SdMean* and *PRDMean* metrics, with the lowest values of 3.9 and 0.1, respectively. These results indicate eWaOA's efficient optimization capability within a fixed termination time.

For each algorithm, we select the iteration data with the minimum C_{max} from 10 runs and then plot the C_{max} variation curves for different test instances, as shown in Fig. 11. In this figure, eWaOA attains the lowest C_{max} in all 30 instances. Moreover, the eWaOA achieves better C_{max} values with significantly fewer iterations (less than 50) for instances like EMK01(s), EMK01(d), EMK03(s), EMK04(s), EMK04(d), EMK05(s), EMK08(s), EMK08(d), EMK14(s), and EMK14(d). For instances EMK03(d), EMK05(d), EMK07(s), EMK11(d), EMK12(s), and EMK12(d), the convergence curve of eWaOA stabilizes within 50–100 iterations. For instances EMK02(s), EMK07(d), EMK09(s), EMK09(d), EMK13(d), EMK15(s), and EMK15(d), convergence stability is achieved

Instance	GA				DE				PSO				GWO				HHO				ARO			
	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*
EMK01(s)	48	54.2	2.7	14.3	46	46.6	0.7	9.5	**42**	47.9	3.8	**0.0**	45	47.7	2.5	7.1	43	50.3	3.5	2.4	43	45.6	0.9	2.4
EMK01(d)	51	52.2	1.2	30.8	42	43.7	1.1	7.7	42	44.1	2.3	5.1	40	44.8	3.2	2.6	46	48.3	1.5	17.9	**39**	41.5	1.2	**0.0**
EMK02(s)	43	46.4	2.5	59.3	38	40.1	1.0	40.7	34	37.4	1.7	25.9	37	39.9	1.8	37.0	39	44.9	2.5	44.4	36	37.8	1.7	33.3
EMK02(d)	44	46.6	2.0	57.1	37	39.2	1.3	32.1	35	38.1	2.2	21.4	35	39.3	2.8	25.0	46	48.0	2.6	64.3	34	37.1	1.6	21.4
EMK03(s)	284	295.4	8.9	39.2	277	286.6	5.4	35.8	219	225.2	7.7	6.9	216	231.7	10.6	5.9	252	279.8	12.1	23.5	228	241.1	6.0	11.8
EMK03(d)	270	296.6	11.6	44.4	267	274.8	6.1	42.8	213	220.1	5.3	11.8	218	235.7	10.5	16.6	266	287.3	10.4	42.2	223	235.7	8.3	19.3
EMK04(s)	72	74.1	1.5	9.1	79	81.1	**1.4**	19.7	75	78.4	3.5	10.6	74	77.3	2.8	12.1	79	85.1	5.2	19.7	73	77.9	2.9	10.6
EMK04(d)	83	90.7	5.3	25.8	81	82.5	**1.2**	22.7	73	79.3	3.8	10.6	72	77.2	5.1	9.1	76	82.9	3.6	15.2	72	74.5	2.1	9.1
EMK05(s)	209	219.4	6.6	20.8	195	200.8	3.5	12.7	181	187.7	4.0	4.6	186	193.5	4.9	7.5	189	200.2	7.6	9.2	186	188.8	1.9	7.5
EMK05(d)	207	214.7	4.2	21.1	199	203.7	2.9	16.4	185	189.3	2.4	7.0	180	190.5	5.6	5.3	196	204.4	5.5	14.6	185	189.3	2.4	8.2
EMK06(s)	142	151.3	5.6	97.2	138	140.1	**1.9**	91.7	107	116.0	7.6	43.1	102	114.4	7.4	41.7	138	145.8	7.6	91.7	107	116.5	5.7	48.6
EMK06(d)	142	154.6	9.1	105.8	140	143.3	**2.0**	102.9	108	113.0	3.8	58.8	105	118.8	6.5	52.2	136	146.4	7.0	97.1	109	117.1	4.3	58.0
EMK07(s)	198	215.2	8.6	43.5	194	199.3	3.7	40.6	165	175.2	7.7	19.6	159	167.0	7.0	15.2	179	194.9	8.7	29.7	164	171.1	4.6	18.8
EMK07(d)	202	216.8	10.9	47.4	199	210.7	4.9	45.3	168	179.4	9.2	18.2	165	173.8	5.0	20.4	209	215.8	10.0	52.6	171	178.2	5.6	24.8
EMK08(s)	599	624.8	18.8	14.5	583	596.3	7.5	11.5	535	542.5	5.8	1.9	538	556.1	15.7	2.9	572	596.1	11.8	9.4	548	560.7	7.2	4.8
EMK08(d)	599	609.6	9.2	16.8	582	595.6	8.8	13.5	523	542.1	10.7	1.9	541	553.8	9.5	5.5	583	595.5	12.5	13.6	549	561.8	7.4	7.0
EMK09(s)	495	513.8	14.9	55.2	484	495.9	6.7	51.7	380	413.0	23.5	18.2	387	423.1	21.2	21.3	454	484.6	18.5	42.3	411	428.0	13.4	28.8
EMK09(d)	484	522.9	17.0	51.3	492	503.2	5.4	53.8	403	423.4	11.4	23.0	391	417.6	12.5	22.2	459	483.7	14.7	43.4	418	441.5	16.4	30.6
EMK10(s)	420	437.7	9.3	74.3	407	413.0	**3.8**	68.9	341	363.4	19.2	36.4	347	353.6	11.8	44.0	402	422.3	17.9	66.8	348	370.6	9.8	44.4
EMK10(d)	421	436.3	13.1	84.6	402	416.0	8.0	76.3	327	348.8	20.6	40.8	345	361.9	13.0	51.3	412	431.1	14.0	80.7	338	361.6	13.0	48.2
EMK11(s)	722	772.1	24.6	17.4	727	735.1	5.9	18.2	646	663.1	12.5	5.0	651	673.2	17.4	5.9	695	717.0	19.7	13.0	665	675.9	7.4	8.1
EMK11(d)	704	752.0	19.6	14.8	696	718.3	8.7	13.5	657	664.1	5.0	7.2	643	671.6	15.8	4.9	680	702.6	12.4	10.9	644	655.8	6.2	5.1
EMK12(s)	634	670.7	18.0	24.8	603	610.8	**5.3**	18.7	531	566.7	16.1	4.5	534	551.0	12.3	5.1	562	595.4	29.5	10.6	538	561.5	12.4	5.9
EMK12(d)	668	688.5	16.4	31.5	599	609.3	**5.4**	17.9	541	572.2	16.6	6.5	535	558.1	16.0	5.3	603	635.4	29.3	18.7	540	558.1	16.0	6.3
EMK13(s)	679	735.3	31.7	61.3	676	692.3	9.4	60.6	528	565.8	16.3	24.0	531	561.0	18.8	26.1	605	649.8	29.7	43.7	563	586.6	17.7	33.7
EMK13(d)	716	733.3	13.9	71.7	677	686.5	6.5	62.4	556	582.5	19.6	29.5	542	561.1	14.3	30.0	605	683.3	40.7	45.1	582	594.5	8.6	39.6
EMK14(s)	932	972.9	24.1	34.3	804	822.6	8.6	15.9	721	761.4	26.8	3.9	707	733.0	11.4	1.9	752	816.6	29.1	8.4	**694**	724.1	17.6	**0.0**
EMK14(d)	947	985.5	29.1	36.5	765	792.4	12.4	10.2	707	764.8	29.0	1.9	707	729.3	18.1	1.9	772	837.2	48.7	11.2	712	728.5	9.2	2.6
EMK15(s)	618	654.6	21.4	68.9	616	632.4	9.4	68.3	527	552.6	17.9	39.6	514	543.8	20.1	40.4	575	614.9	23.6	57.1	533	561.2	17.3	45.6
EMK15(d)	633	672.8	32.7	65.7	626	644.9	10.7	63.9	515	531.3	14.1	34.3	521	542.8	16.8	36.4	590	629.0	21.7	54.5	528	560.3	14.5	38.2
SdMean	13.1				5.3				11.0				10.7				15.4				8.1			
RPDMean	44.6				38.2				18.8				18.8				35.1				20.8			

Instance	WO				MVO				OMA				TLBO				PRO				eWaOA			
	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*
EMK01(s)	48	51.1	2.0	14.3	47	49.4	2.4	11.9	49	51.1	1.1	16.7	45	46.8	1.7	7.1	48	52.1	1.8	14.3	**42**	**42.0**	**0.0**	**0.0**
EMK01(d)	43	47.3	3.3	10.3	41	46.1	3.2	5.1	46	48.1	1.7	17.9	**39**	40.6	1.3	**0.0**	48	53.0	2.4	23.1	**39**	**39.1**	**0.3**	**0.0**
EMK02(s)	36	43.6	3.2	33.3	36	40.1	2.0	33.3	46	47.4	1.2	70.4	33	38.0	2.6	22.2	39	43.3	2.7	44.4	**27**	**28.1**	**0.7**	**0.0**
EMK02(d)	39	43.8	3.5	39.3	34	40.3	3.6	21.4	43	45.1	1.0	53.6	35	39.2	3.0	25.0	39	42.3	2.1	39.3	**28**	**28.6**	**0.8**	**0.0**
EMK03(s)	238	263.2	26.3	16.7	222	248.3	19.5	8.8	268	279.4	6.0	31.4	216	249.3	21.2	5.9	281	291.9	5.4	37.7	**204**	**204.0**	**0.0**	**0.0**
EMK03(d)	249	285.9	25.2	33.2	219	240.7	14.9	17.1	266	280.3	7.4	42.2	204	237.0	20.5	9.1	286	298.3	9.1	52.9	**187**	**187.0**	**0.0**	**0.0**

Continued

Instance	WO				MVO				OMA				TLBO				PRO				eWaOA			
	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*	$B(C_{max})$	*Av*	*Sd*	*RPD*
EMK04(s)	73	80.6	5.1	10.6	79	82.2	3.3	19.7	81	83.6	1.9	22.7	68	74.5	3.7	3.0	85	89.7	2.9	28.8	**66**	**68.1**	2.5	**0.0**
EMK04(d)	73	79.8	5.1	10.6	77	80.9	2.8	16.7	80	83.7	1.7	21.2	73	76.9	2.0	10.6	84	88.3	3.3	27.3	**66**	**68.3**	2.5	**0.0**
EMK05(s)	190	200.2	7.6	9.8	188	198.0	7.8	8.7	203	207.2	3.1	17.3	186	195.7	6.3	7.5	213	223.8	4.6	23.1	**173**	**175.3**	**1.7**	**0.0**
EMK05(d)	180	192.7	5.6	5.3	191	197.6	6.1	11.7	202	204.0	1.5	18.1	175	189.1	6.6	2.3	213	221.1	5.3	24.6	**171**	**172.7**	**1.0**	**0.0**
EMK06(s)	119	138.9	18.0	65.3	115	127.2	7.4	59.7	142	147.3	3.4	97.2	113	127.4	9.0	56.9	153	160.7	5.5	112.5	**72**	**75.8**	2.7	**0.0**
EMK06(d)	111	139.5	14.5	60.9	114	127.6	9.3	65.2	141	147.9	3.7	104.3	107	126.1	12.1	55.1	146	156.1	6.5	111.6	**69**	**75.7**	3.5	**0.0**
EMK07(s)	176	201.2	13.4	27.5	172	185.8	7.3	24.6	191	200.4	4.8	38.4	164	179.5	11.4	18.8	204	214.8	7.5	47.8	**138**	**142.5**	**2.5**	**0.0**
EMK07(d)	182	201.8	12.8	32.8	179	189.6	11.5	30.7	206	209.3	4.6	50.4	156	171.1	9.2	13.9	201	214.1	8.4	46.7	**137**	**142.2**	**3.7**	**0.0**
EMK08(s)	548	582.2	30.2	4.8	553	573.9	17.5	5.7	582	590.3	4.5	11.3	563	581.6	10.9	7.6	604	613.4	8.2	15.5	**523**	**532.0**	**3.0**	**0.0**
EMK08(d)	526	586.6	45.3	2.5	552	576.5	21.4	7.6	573	584.8	6.7	11.7	535	574.4	17.9	4.3	599	634.6	15.6	16.8	**513**	**521.0**	**4.0**	**0.0**
EMK09(s)	423	483.9	32.3	32.6	413	439.2	14.0	29.5	493	503.2	**6.1**	54.5	439	462.1	17.2	37.6	481	523.3	20.9	50.8	**319**	**327.8**	6.3	**0.0**
EMK09(d)	485	495.3	28.9	51.6	418	441.1	15.6	30.6	500	505.9	**4.0**	56.3	396	443.3	39.2	23.8	503	538.8	17.9	57.2	**318**	**329.9**	5.7	**0.0**
EMK10(s)	357	414.3	41.4	48.1	345	375.5	20.2	43.2	417	429.3	7.7	73.0	383	421.3	15.4	58.9	436	453.0	11.9	80.9	**241**	**250.6**	7.7	**0.0**
EMK10(d)	378	417.7	41.4	65.8	344	372.3	15.0	50.9	434	443.0	7.1	90.4	368	410.0	24.3	61.4	430	454.4	15.3	88.6	**228**	**248.0**	**3.9**	**0.0**
EMK11(s)	677	702.1	24.3	10.1	666	686.7	14.9	8.3	691	708.2	8.4	12.4	682	695.9	10.4	10.9	714	738.5	16.4	16.1	**615**	**619.2**	**2.9**	**0.0**
EMK11(d)	683	717.7	33.2	11.4	670	682.3	10.6	9.3	714	726.5	6.9	16.5	695	718.6	15.4	13.4	723	757.7	14.5	17.9	**613**	**624.5**	**2.5**	**0.0**
EMK12(s)	552	605.0	36.2	8.7	573	591.7	15.1	12.8	572	581.2	5.6	12.6	529	549.7	13.6	4.1	639	655.4	9.8	25.8	**508**	**513.8**	9.2	**0.0**
EMK12(d)	564	589.0	34.0	11.0	540	584.0	28.7	6.3	572	596.4	11.1	12.6	533	558.1	17.6	4.9	617	654.7	18.6	21.5	**508**	**517.5**	7.1	**0.0**
EMK13(s)	605	665.3	49.9	43.7	564	612.2	33.7	34.0	687	716.9	13.3	63.2	541	635.0	50.3	28.5	711	745.6	23.1	68.9	**421**	**452.1**	**9.1**	**0.0**
EMK13(d)	552	668.0	74.3	32.4	566	612.6	30.8	35.7	718	727.1	**5.6**	72.2	493	607.3	54.2	18.2	680	725.6	30.3	63.1	**417**	**448.6**	6.8	**0.0**
EMK14(s)	727	756.1	22.4	4.8	727	782.4	41.5	4.8	734	757.1	14.1	5.8	**694**	716.2	10.2	**0.0**	882	915.0	25.5	27.1	**694**	**694.0**	**0.0**	**0.0**
EMK14(d)	733	779.0	38.2	5.6	759	808.7	22.9	9.4	748	767.6	15.8	7.8	**694**	705.1	11.8	**0.0**	873	926.2	35.8	25.8	**694**	**694.0**	**0.0**	**0.0**
EMK15(s)	541	628.1	46.0	47.8	517	565.8	29.0	41.3	621	643.2	10.7	69.7	630	650.8	13.2	72.1	648	673.8	16.1	77.0	**366**	**395.9**	**5.1**	**0.0**
EMK15(d)	548	611.8	58.6	43.5	535	573.9	32.1	40.1	640	657.1	7.8	67.5	548	620.7	31.2	43.5	673	688.9	11.3	76.2	**382**	**404.6**	**5.0**	**0.0**
sdMean	26.1				15.5				5.9				15.4				12.0				3.3			
RPDMean	26.5				23.5				41.3				20.9				45.4				0.0			

Table 8. Comparison results with the same maximum number of iterations. Significant values are in bold.

within 100–200 iterations. For instances EMK02(d), EMK06(s), EMK06(d), EMK10(d), EMK11(s), and EMK13(s), stability is achieved before 250 iterations.

To highlight the remarkable advantages of eWaOA over other algorithms, a paired *t*-test was conducted at a significance level of $\alpha = 0.05$ to explore the existence of statistically significant differences between eWaOA and various comparative algorithms. The data for this paired *t*-test were obtained from the outcomes of running eWaOA and each comparative algorithm 10 times per instance. Figure 12 shows the results of the paired *t*-test regarding the C_{max} of eWaOA and comparative algorithms for each test instance. In this figure, each group on the x-axis represents the paired t-test results of eWaOA against a specific comparative algorithm, and the log2FC shown is calculated based on the average values of C_{max}. Specifically, we first determine the fold change of the average C_{max} of eWaOA compared to that of each comparative algorithm and then take the logarithm to the base 2 of this fold change. Moreover, according to the p-values from the paired *t*-test, the scatter points in the graph are divided into two groups: the group with a "p-value ≤ 0.05" is represented by red scatter points, indicating a statistically significant difference, while the group with a "p-value > 0.05" is shown by blue scatter points. To avoid complete horizontal overlap of data points and improve the readability of the scatter plot, random perturbations have been applied to each data point during the plotting process. As can be seen from the figure, except for the EMK14(d) instance in the comparison between eWaOA and ARO, and the EMK07(d) instance in the comparison between eWaOA and TLBO, eWaOA shows highly significant differences from other algorithms in most instances, as demonstrated by the distribution of red and blue scatter points as well as the values of log2FC.

Based on the results in Tables 8 and 9, as well as Figd. 11 and 12, we conclude that the proposed eWaOA significantly outperforms the 11 SOTA algorithms in both optimization effectiveness and efficiency. Specifically, eWaOA demonstrates superior performance in terms of makespan, stability, consistency, and optimization efficiency-achieving better results within the specified time.

Engineering case study

This study aims to further validate the proposed eWaOA by applying it to a practical engineering scenario involving three distinct product categories tested at an electronic product performance lab. The products include mobile phones (MP), in-vehicle navigators (IVNs), and unmanned aerial vehicles (UAVs). The performance testing process plan for MP is illustrated in Fig. 1, while the testing process plan for IVNs and UAVs are detailed in Tables 10 and 11, respectively.

The results obtained by applying the 12 algorithms to the engineering case are presented in Table 12. All algorithms use an time-limited termination criterion, with the corresponding time limit set to 55(s). It is evident that PSO, GWO, ARO, TLBO, and eWaOA yield the smallest $B(C_{max})$, with eWaOA achieving the lowest *Av* and *Sd*. This further demonstrates that eWaOA not only minimizes $B(C_{max})$ but also shows superior stability and consistency. Figure 13 displays the Gantt chart of the optimal scheduling results from 10 runs of eWaOA, with PBPOs highlighted in red boxes. The chart demonstrates that all operations adhere to the sequential order constraints of the process plan and satisfy the PBPO requirements, reaffirming the feasibility and effectiveness of eWaOA in solving the FJSP_PBPO.

Conclusion and future research

To optimize the makespan for the FJSP_PBPO problem, this study develops an optimization model using MIP and introduces an enhanced swarm-based metaheuristic algorithm, eWaOA, which extends the WaOA framework. In eWaOA, new schemes for encoding, conversion, inverse conversion, and decoding tailored to the specific constraints of FJSP_PBPO are designed. Additionally, a ROMI strategy is designed to generate diverse and high-quality initial solutions. Enhancements are made to the feeding, migration, and fleeing strategies of WaOA, and a novel gathering strategy is introduced to improve both exploration and exploitation.

To evaluate these improvements, 30 test instance, extended from existing benchmark FJSP instances, are used. The ROMI initialization strategy shows superior search capability, stability, and consistency compared to WaOA, enhancing convergence efficiency. Comparisons are made with four enhanced WaOA variants and eleven SOTA metaheuristic algorithms on the 30 test instances, followed by a real-world engineering case study. Results from these comparisons confirm that the eWaOA effectively addresses the FJSP_PBPO, demonstrating superior optimization capability, stability, consistency, and efficiency.

The proposed eWaOA primarily addresses the FJSP with PBPO. However, electronic product performance testing introduces additional constraints, including multi-resource coupling and sequence-dependent setup times. Future research will focus on enhancing eWaOA to effectively handle these constraints, extending its applicability to more complex engineering scenarios.

Instance	Time (s)	GA				DE				PSO				GWO				HHO				ARO			
		$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD
EMK01(s)	60	49	53.6	3.1	16.7	43	44.3	0.8	2.4	43	47.9	2.4	2.4	44	46.5	1.8	4.8	46	48.8	2.0	9.5	**42**	44.7	1.9	**0.0**
EMK01(d)	60	48	51.8	2.9	23.1	40	41.5	**0.8**	2.6	41	43.9	2.2	5.1	41	44.4	2.6	5.1	41	46.4	3.1	5.1	**39**	39.7	0.9	**0.0**
EMK02(s)	60	40	44.3	2.4	42.9	36	38.1	0.9	28.6	35	36.3	1.1	25.0	37	40.0	2.1	32.1	38	41.6	2.1	35.7	34	36.6	1.9	21.4
EMK02(d)	60	43	44.9	1.8	53.6	36	38.1	1.2	28.6	34	35.9	1.6	21.4	36	39.2	2.2	28.6	41	44.5	2.5	46.4	33	35.5	1.4	17.9
EMK03(s)	120	272	285.4	8.7	33.3	264	271.6	4.1	29.4	206	215.0	6.8	1.0	214	229.5	8.7	4.9	256	273.4	7.9	25.5	219	223.8	5.2	7.4
EMK03(d)	120	262	279.3	10.6	40.1	265	271.2	3.9	41.7	203	221.2	10.6	8.6	204	228.9	15.3	9.1	249	270.7	13.0	33.2	208	218.3	6.8	11.2
EMK04(s)	120	82	89.1	3.3	22.4	75	77.6	1.6	11.9	73	75.3	2.2	9.0	73	78.2	3.3	9.0	76	83.0	5.5	13.4	72	73.7	**1.3**	7.5
EMK04(d)	120	85	90.5	4.2	28.8	75	78.2	**1.8**	13.6	73	76.2	3.2	10.6	73	76.2	2.6	10.6	75	83.0	4.9	13.6	70	72.8	1.4	6.1
EMK05(s)	60	204	214.9	7.6	17.2	193	199.2	2.4	10.9	180	185.9	3.1	3.4	187	191.1	2.8	7.5	193	202.6	7.7	10.9	179	183.8	3.4	2.9
EMK05(d)	60	204	215.1	5.9	18.6	194	199.1	2.6	12.8	181	186.1	4.2	5.2	179	191.1	5.2	4.1	195	204.8	7.5	13.4	180	184.9	3.6	4.7
EMK06(s)	150	132	143.9	5.9	78.4	130	134.3	2.9	75.7	100	109.5	4.7	35.1	103	109.7	4.3	39.2	131	142.6	8.2	77.0	100	105.6	**2.7**	35.1
EMK06(d)	150	134	151.2	9.9	71.8	134	137.4	**2.4**	71.8	101	109.3	5.4	29.5	103	111.5	7.1	32.1	127	139.7	7.8	62.8	95	107.7	4.6	21.8
EMK07(s)	100	197	212.7	6.7	36.8	183	190.0	3.8	27.1	155	163.8	5.5	7.6	163	167.2	3.7	13.2	181	193.4	7.4	25.7	159	165.7	4.9	10.4
EMK07(d)	100	204	220.0	10.3	39.7	190	194.0	**2.3**	30.1	161	167.8	4.8	10.3	158	171.6	7.7	8.2	181	198.3	10.9	24.0	159	165.6	6.4	8.9
EMK08(s)	200	596	614.8	19.0	12.5	580	586.5	5.2	9.4	**530**	546.4	8.8	**0.0**	533	550.8	12.1	0.6	545	577.5	15.4	2.8	533	535.4	3.3	0.6
EMK08(d)	200	580	613.4	18.3	13.1	562	579.8	9.1	9.6	523	541.8	11.5	1.9	524	547.4	20.4	2.1	557	576.9	15.2	8.6	523	528.1	5.0	1.9
EMK09(s)	200	478	511.4	22.1	48.4	459	476.4	7.7	42.5	368	390.8	15.3	14.3	400	415.1	12.8	24.2	440	475.2	20.4	36.6	387	401.0	13.1	20.2
EMK09(d)	200	480	509.4	21.4	48.1	464	476.5	**5.3**	43.2	381	397.0	10.6	17.6	397	416.4	9.7	22.5	446	467.8	15.9	37.7	358	395.5	15.2	10.5
EMK10(s)	300	403	427.8	16.5	60.6	394	403.6	6.2	57.0	318	332.6	10.2	26.7	333	350.9	13.3	32.7	391	404.4	11.3	55.8	326	339.3	10.0	29.9
EMK10(d)	300	408	436.5	15.9	62.5	390	405.3	7.5	55.4	316	333.2	10.4	25.9	324	346.8	14.9	29.1	382	408.7	12.7	52.2	319	334.2	9.6	27.1
EMK11(s)	150	721	755.4	23.2	16.9	702	710.0	5.5	13.8	641	655.5	8.6	3.9	642	657.2	10.7	4.1	678	706.5	19.5	9.9	636	649.6	7.9	3.1
EMK11(d)	150	717	757.5	25.8	16.2	692	705.6	6.8	12.2	645	658.4	10.2	4.5	644	656.9	8.7	4.4	675	702.2	17.1	9.4	641	647.2	5.7	3.9
EMK12(s)	300	610	659.8	27.0	20.1	577	583.0	**4.3**	13.6	524	558.8	16.3	3.1	524	550.2	16.0	3.1	572	599.2	16.4	12.6	531	544.6	10.0	4.5
EMK12(d)	300	627	672.1	20.4	23.4	576	588.1	**7.0**	13.4	524	552.0	17.3	3.1	534	550.4	12.0	5.1	564	597.8	29.7	11.0	524	550.4	12.0	3.1
EMK13(s)	300	698	738.9	17.8	53.4	624	654.0	15.8	37.1	511	542.4	24.5	12.3	529	552.3	15.7	16.3	608	644.7	28.6	33.6	512	543.4	13.0	12.5
EMK13(d)	300	696	732.1	20.8	56.8	646	660.5	7.0	45.5	501	534.5	16.5	12.8	535	560.2	14.9	20.5	586	652.9	32.0	32.0	528	543.8	14.4	18.9
EMK14(s)	450	895	963.3	38.0	29.0	750	771.9	12.3	8.1	**694**	738.2	38.4	**0.0**	**694**	730.5	16.8	**0.0**	745	789.0	32.0	7.3	**694**	723.1	16.0	**0.0**
EMK14(d)	450	929	963.8	29.7	33.9	747	771.7	10.5	7.6	714	757.5	27.9	2.9	**694**	724.8	18.5	**0.0**	720	802.2	36.4	3.7	**694**	700.6	11.1	**0.0**
EMK15(s)	450	644	664.0	14.2	63.9	608	620.4	8.8	54.7	473	498.9	12.9	20.4	498	525.0	15.2	26.7	571	612.3	27.0	45.3	486	507.4	15.7	23.7
EMK15(d)	450	625	656.4	25.9	56.6	583	623.4	16.9	46.1	494	515.8	17.8	23.8	511	537.5	14.1	28.1	576	628.1	26.4	44.4	494	511.2	14.7	23.8
SdMean		14.6				5.6				10.5				9.8				14.9				9.8			
RPDMean		38.0				28.5				11.6				14.3				26.6				14.3			

Instance	Time (s)	WO				MVO				OMA				TLBO				PRO				eWaOA			
		$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD
EMK01(s)	60	45	49.3	2.8	7.1	46	48.0	2.4	9.5	46	49.6	1.4	9.5	**42**	44.7	3.0	**0.0**	48	52.1	1.8	14.3	**42**	**42.0**	**0.0**	**0.0**
EMK01(d)	60	42	46.4	3.0	7.7	42	46.7	2.6	7.7	44	47.1	1.4	12.8	**39**	40.6	1.3	**0.0**	48	53.0	2.4	23.1	**39**	**39.4**	0.9	**0.0**
EMK02(s)	60	34	41.7	4.0	21.4	37	40.5	2.4	32.1	43	44.3	0.6	53.6	33	36.4	2.1	17.9	39	43.3	2.7	39.3	**28**	**29.0**	**0.4**	**0.0**
EMK02(d)	60	37	38.9	2.1	32.1	35	38.2	2.4	25.0	42	44.2	1.3	50.0	33	36.0	2.0	17.9	39	42.3	2.1	39.3	**28**	**29.3**	**1.1**	**0.0**
EMK03(s)	120	223	250.5	21.2	9.3	220	239.2	15.0	7.8	264	275.6	5.1	29.4	**204**	224.5	15.5	**0.0**	281	291.9	5.4	37.7	**204**	**204.0**	**0.0**	**0.0**
EMK03(d)	120	217	253.7	20.1	16.0	214	233.5	9.5	14.4	263	273.3	6.1	40.6	194	213.1	22.4	3.7	279	291.6	9.2	49.2	**187**	**188.2**	**3.6**	**0.0**

Continued

Instance	Time (s)	WO				MVO				OMA				TLBO				PRO				eWaOA			
		$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD	$B(C_{max})$	Av	Sd	RPD
EMK04(s)	120	73	76.6	3.9	9.0	74	80.8	4.9	10.4	80	81.4	1.4	19.4	**67**	72.2	2.4	**0.0**	85	89.8	2.9	26.9	**67**	**69.0**	2.8	**0.0**
EMK04(d)	120	73	81.3	5.9	10.6	73	81.2	3.8	10.6	79	80.4	**1.1**	19.7	**66**	70.6	2.9	**0.0**	83	87.3	3.3	25.8	67	**70.5**	3.1	1.5
EMK05(s)	60	183	195.1	10.0	5.2	193	199.3	6.2	10.9	195	198.6	2.6	12.1	179	186.5	5.4	2.9	206	216.9	4.5	18.4	**174**	**175.9**	**1.3**	**0.0**
EMK05(d)	60	185	198.0	7.9	7.6	188	194.3	5.7	9.3	190	196.8	3.5	10.5	176	180.7	2.8	2.3	207	214.4	5.1	20.3	**172**	**176.5**	**2.3**	**0.0**
EMK06(s)	150	110	124.5	11.2	48.6	112	123.9	5.9	51.4	137	145.4	3.5	85.1	94	109.3	13.0	27.0	146	154.4	5.1	97.3	**74**	**79.8**	4.1	**0.0**
EMK06(d)	150	114	128.8	14.8	46.2	109	122.8	8.4	39.7	145	147.1	2.7	85.9	93	101.8	7.1	19.2	144	154.7	6.1	84.6	**78**	**81.9**	4.0	**0.0**
EMK07(s)	100	174	183.4	8.8	20.8	167	187.5	10.5	16.0	188	194.5	**3.1**	30.6	148	158.0	7.3	2.8	202	212.8	7.5	40.3	**144**	**150.0**	3.9	**0.0**
EMK07(d)	100	166	192.1	25.3	13.7	159	178.5	10.5	8.9	185	196.6	4.6	26.7	148	155.5	5.2	1.4	205	214.4	6.8	40.4	**146**	**152.2**	4.2	**0.0**
EMK08(s)	200	533	555.1	20.1	0.6	544	572.7	16.9	2.6	559	580.5	9.6	5.5	533	556.2	13.5	0.6	604	613.4	8.2	14.0	533	**533.0**	**0.0**	0.6
EMK08(d)	200	540	560.9	25.2	5.3	528	569.1	27.2	2.9	562	581.1	9.4	9.6	524	540.8	12.6	2.1	585	617.6	14.1	14.0	**513**	**522.0**	**3.0**	**0.0**
EMK09(s)	200	399	466.3	49.8	23.9	419	436.7	16.3	30.1	482	493.6	**6.2**	49.7	365	406.8	29.3	13.4	481	522.2	20.9	49.4	**322**	**334.9**	8.5	**0.0**
EMK09(d)	200	393	435.9	31.9	21.3	417	435.4	10.3	28.7	476	496.2	8.3	46.9	370	407.1	35.8	14.2	518	530.2	12.4	59.9	**324**	**334.3**	6.9	**0.0**
EMK10(s)	300	341	400.2	46.2	35.9	342	358.3	11.8	36.3	416	423.2	**4.4**	65.7	311	386.0	28.1	23.9	432	449.0	11.9	72.1	**251**	**263.4**	8.0	**0.0**
EMK10(d)	300	337	393.9	41.7	34.3	339	363.5	13.2	35.1	418	425.7	**4.2**	66.5	299	355.8	37.8	19.1	428	453.2	11.3	70.5	**251**	**260.4**	5.8	**0.0**
EMK11(s)	150	651	689.5	33.4	5.5	656	688.2	21.5	6.3	695	700.7	**3.5**	12.6	633	659.0	18.6	2.6	736	756.9	13.1	19.3	**617**	**624.9**	4.2	**0.0**
EMK11(d)	150	643	680.0	29.1	4.2	664	685.3	12.3	7.6	689	703.4	6.5	11.7	622	651.8	20.3	0.8	730	752.3	15.5	18.3	**617**	**623.4**	**4.4**	**0.0**
EMK12(s)	300	524	553.1	16.1	3.1	550	593.3	32.5	8.3	540	549.0	7.2	6.3	513	532.3	10.4	1.0	609	640.3	17.8	19.9	**508**	**521.8**	7.5	**0.0**
EMK12(d)	300	540	573.2	27.7	6.3	562	581.8	19.9	10.6	566	582.7	9.6	11.4	524	532.2	10.3	3.1	630	669.4	21.4	24.0	**508**	**520.2**	8.5	**0.0**
EMK13(s)	300	592	661.4	51.7	30.1	558	589.3	22.6	22.6	661	677.9	12.5	45.3	507	553.9	43.0	11.4	666	717.0	26.3	46.4	**455**	**460.6**	**4.2**	**0.0**
EMK13(d)	300	558	623.0	53.8	25.7	558	599.4	28.6	25.7	668	692.9	11.5	50.5	467	547.6	48.8	5.2	685	716.3	18.5	54.3	**444**	**456.5**	**6.0**	**0.0**
EMK14(s)	450	707	733.0	22.7	1.9	752	789.1	21.0	8.4	707	734.6	15.3	1.9	707	718.5	8.2	1.9	868	915.6	23.7	25.1	**694**	**694.0**	**0.0**	**0.0**
EMK14(d)	450	**694**	710.3	16.9	**0.0**	772	803.4	23.9	11.2	712	735.7	14.3	2.6	**694**	703.6	12.6	**0.0**	850	896.0	36.3	22.5	**694**	**694.0**	**0.0**	**0.0**
EMK15(s)	450	550	610.9	44.2	39.9	521	554.9	22.0	32.6	612	636.5	11.7	55.7	462	586.1	51.2	17.6	648	676.2	16.1	64.9	**393**	**402.3**	**7.6**	**0.0**
EMK15(d)	450	541	605.6	39.4	35.6	528	561.7	24.1	32.3	631	645.0	**8.5**	58.1	471	547.9	54.6	18.0	650	682.5	19.9	62.9	**399**	**415.3**	9.8	**0.0**
SdMean		23.0				13.8				6.0				17.6				11.7				3.9			
RPDMean		17.6				18.5				32.9				7.7				39.8				0.1			

Table 9. Comparison results with identical time-limited termination. Significant values are in bold.

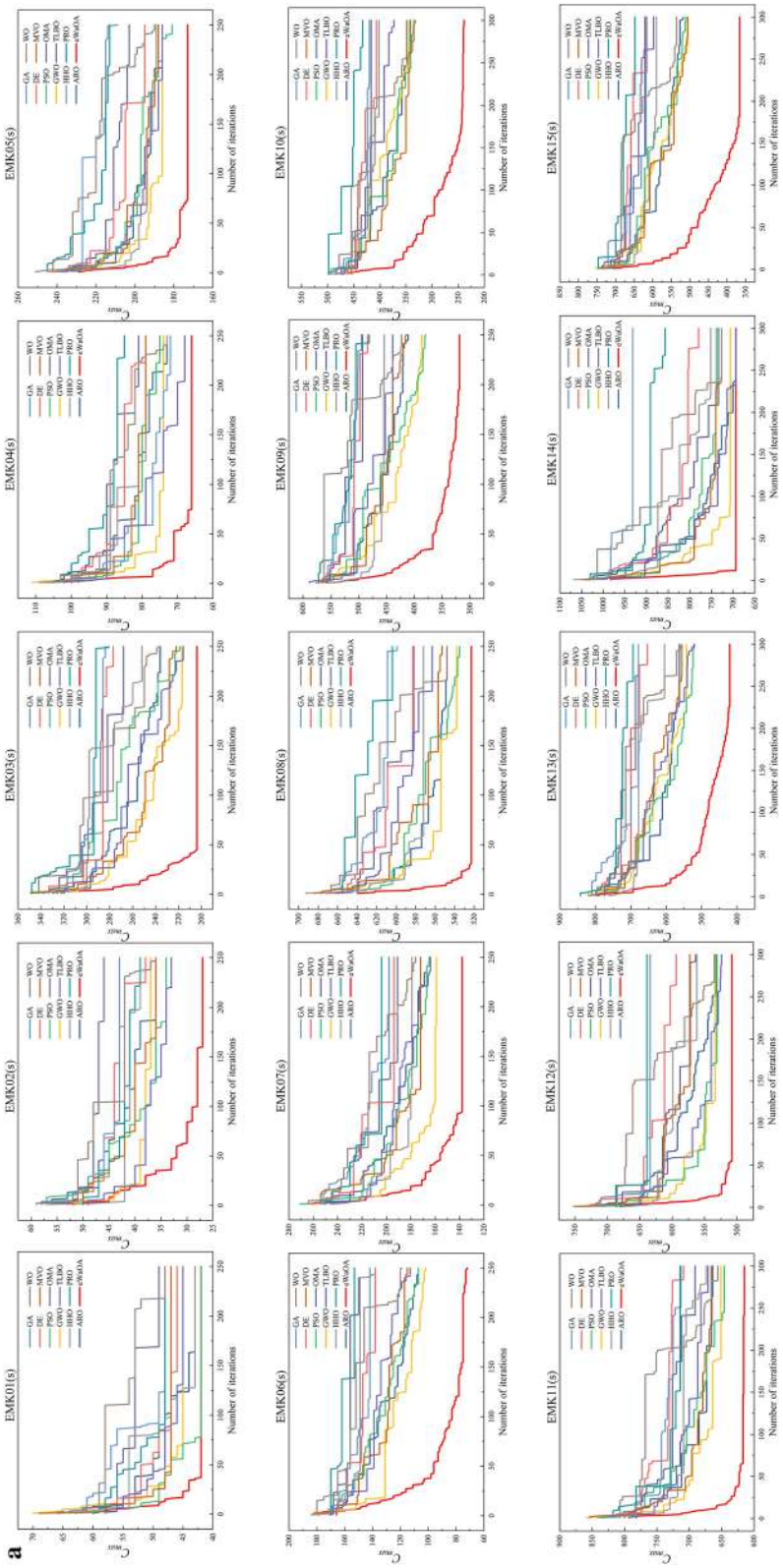


Fig. 11. Iterative results of various algorithms for solving FJSP_PBPO instances.

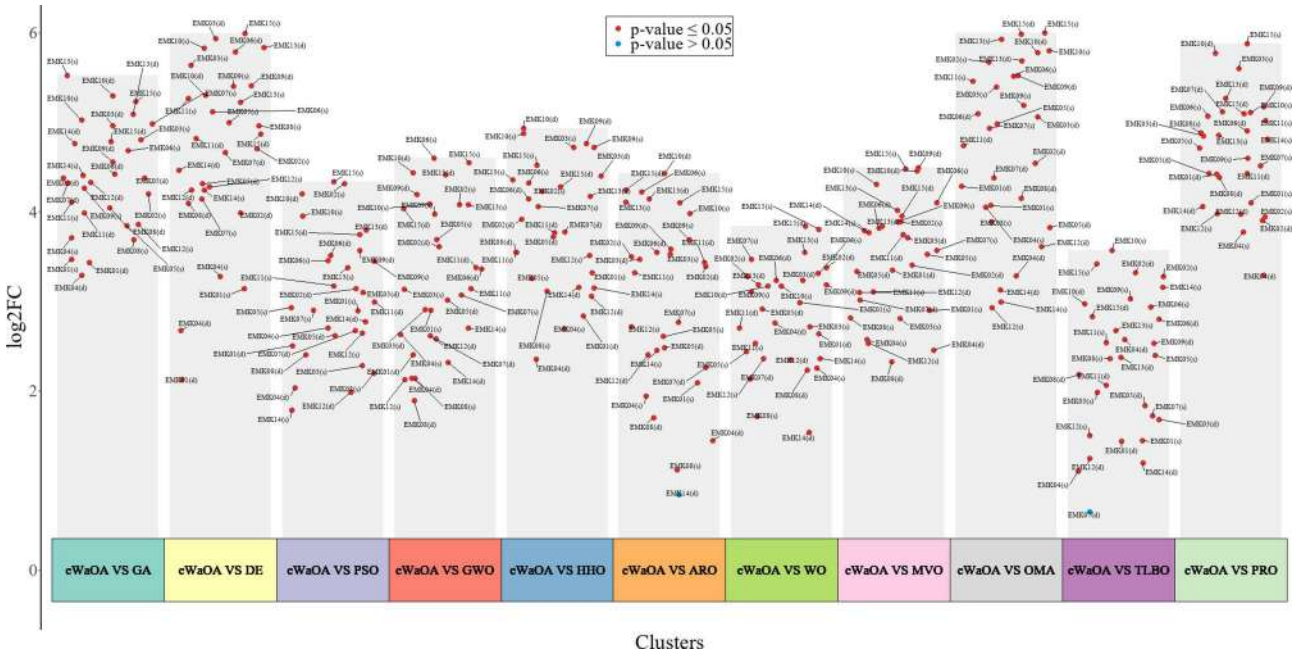


Fig. 12. The results of the paired *t* -test between eWaOA and the other 11 SOTA algorithms for the 30 test instances.

Tasks of IVN	Machine	Processing time	Tasks of IVN	Machine	Processing time
O_{11} (Electrical performance test)	M4	10	O_{42}(High temperature functional test)	M3	13
O_{12} (Button operation force test)	M7	12	O_{43} (Low temperature exposure test)	M6	17
O_{21} (Dimensional inspection)	M2	14	O_{44} (Low temperature functional test)	M8	17
O_{22} (Rapid thermal cycling test)	M3	19	O_{45} (Humidity and temperature cycling Test)	M7	19
O_{23} (Key operation durability test)	M4	13	{O_{37}, O_{46}}(Dust test)	M1	15
{O_{13}, O_{24}}(Drop impact test)	M3	18	O_{51} (Appearance function test)	M5/ M8	15/15
O_{31} (Static current test)	M1	12	O_{61} (Conductive emission test)	M7	17
O_{32} (Operating voltage range test)	M1	15	O_{62} (Radiated emission test)	M8	18
O_{33} (Alcohol screening)	M2	17	O_{63} (Radiated immunity test)	M3	15
O_{34} (Fuel injector adhesion point inspection)	M5	17	O_{64} (Electrostatic discharge test)	M3	16
O_{35} (Artificial sweat test)	M6	19	O_{65} (Temperature cycling test)	M6	16
O_{36} (Swabbing test)	M2	12	O_{71} (Lifetime testing)	M5/M8	16/16
O_{41}(High temperature exposure test)	M4	15			

Table 10. Process plan of the IVNs.

Tasks of IVN	Machine	Processing time	Tasks of UAVs	Machine	Processing time
O_{11} (High-low temperature charge–discharge test)	M1	17	O_{31} (Key/Button test)	M4	18
O_{12} (High-low temperature flight test)	M8	14	O_{32} (Transportation vibration Test)	M6	17
O_{13} (Swelling rainfall test)	M4	15	O_{33} (Handling Test)	M2	19
O_{21} (Corrosion resistance test)	M2	10	O_{34} (Circuit bending test)	M5/M8	12/12
O_{22} (Maximum load aging test)	M5	19	O_{41}(Battery insertion and removal test)	M2	14
O_{23} (Spraying aging test)	M3	16	O_{42}(Six-sided drop test)	M7	12
{O_{14}, O_{24}}(drop impact test)	M2	10	O_{43} (Dustproof test)	M1	17
			O_{44} (Immersion water test)	M6	15

Table 11. Process plan of the UAVs.

GA			DE			PSO			GWO		
$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*
135	144.7	5.59	129	129.5	0.81	**128**	130.6	4.13	**128**	129.2	0.75
HHO			**ARO**			**WO**			**MVO**		
$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*
129	133.0	5.50	**128**	128.8	0.60	129	131.2	2.79	129	132.3	2.65
OMA			**TLBO**			**PRO**			**eWaOA**		
$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*	$B(C_{max})$	*Av*	*Sd*
129	128.9	**0.30**	**128**	128.7	0.46	130	137.0	5.23	**128**	**128.1**	**0.30**

Table 12. Comparison of scheduling results for instance from testing workshop. Significant values are in bold.

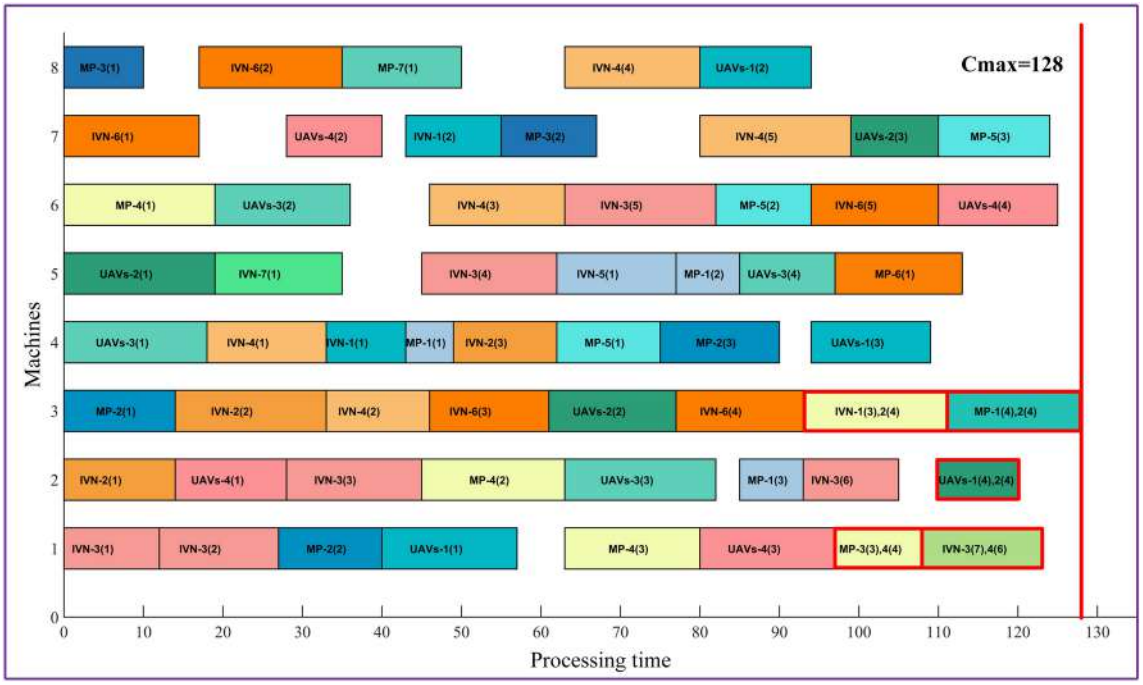


Fig. 13. Gantt chart of a scheduling scheme obtained by eWaOA for the practical example.

Data availability

The authors confirm that the data supporting the findings of this study are available within the paper.

Received: 12 December 2024; Accepted: 5 February 2025
Published online: 17 February 2025

References

1. Brucker, P. & Schlie, R. Job-shop scheduling with multi-purpose machines. *Computing* **45**, 369–375. https://doi.org/10.1007/BF02238804 (1990).
2. Brandimarte, P. Routing and scheduling in a flexible job shop by tabu search. *Ann. Oper. Res.* **41**, 157–183. https://doi.org/10.1007/BF02023073 (1993).
3. Xie, J., Gao, L., Peng, K. K., Li, X. Y. & Li, H. R. Review on flexible job shop scheduling. *IET Coll. Intell. Manuf.* **1**, 67–77. https://doi.org/10.1049/iet-cim.2018.0009 (2019).
4. Dauzère-Pérès, S., Ding, J. W., Shen, L. J. & Tamssaouet, K. The flexible job shop scheduling problem: A review. *Eur. J. Oper. Res.* **314**, 409–432. https://doi.org/10.1016/j.ejor.2023.05.017 (2024).
5. Fowler, J. W. & Mönch, L. A survey of scheduling with parallel batch (p-batch) processing. *Eur. J. Oper. Res.* **298**, 1–24. https://doi.org/10.1016/j.ejor.2021.06.012 (2022).
6. Luo, S., Zhang, L. X. & Fan, Y. S. Energy-efficient scheduling for multi-objective flexible job shops with variable processing speeds by grey wolf optimization. *J. Clean. Prod.* **234**, 1365–1384. https://doi.org/10.1016/j.jclepro.2019.06.151 (2019).
7. Liu, M., Yao, X. F. & Li, Y. X. Hybrid whale optimization algorithm enhanced with Lévy flight and differential evolution for job shop scheduling problems. *Appl. Soft Comput.* **87**, 105954. https://doi.org/10.1016/j.asoc.2019.105954 (2020).
8. Ye, S. & Bu, T. M. A self-learning Harris hawks optimization algorithm for flexible job shop scheduling with setup times and resource constraints. In *2021 IEEE Int. Conf. Syst., Man, Cybern. (SMC)* 2642–2649. https://doi.org/10.1109/SMC52423.2021.9659113 (2021).

9. Yang, Z., Liang, X., Li, Y. & Wang, H. A hybrid remora optimization algorithm with variable neighborhood search for the flexible job shop scheduling problem. In *2024 7th Int. Conf. Adv. Algorithms Control Eng. (ICAACE)* 942–950. https://doi.org/10.1109/ICAACE61206.2024.10548408 (2024).
10. Lv, Z. L., Zhao, Y. X., Kang, H. Y., Gao, Z. Y. & Qin, Y. H. An improved Harris Hawk optimization algorithm for flexible job shop scheduling problem. *Comput. Mater. Concr.* **78**, 2337–2360. https://doi.org/10.32604/cmc.2023.045826 (2024).
11. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**, 67–82. https://doi.org/10.1109/4235.585893 (1997).
12. Han, M. et al. Walrus optimizer: A novel nature-inspired metaheuristic algorithm. *Expert Syst. Appl.* **239**, 122413. https://doi.org/10.1016/j.eswa.2023.122413 (2024).
13. Fontes, D. B. M. M., Homayouni, S. M. & Gonçalves, J. F. A hybrid particle swarm optimization and simulated annealing algorithm for the job shop scheduling problem with transport resources. *Eur. J. Oper. Res.* **306**, 1140–1157. https://doi.org/10.1016/j.ejor.2022.09.006 (2023).
14. Zhang, F., Li, R. & Gong, W. Deep reinforcement learning-based memetic algorithm for energy-aware flexible job shop scheduling with multi-AGV. *Comput. Ind. Eng.* **189**, 109917. https://doi.org/10.1016/j.cie.2024.109917 (2024).
15. Chen, X., Li, J., Wang, Z., Li, J. & Gao, K. A genetic programming based cooperative evolutionary algorithm for flexible job shop with crane transportation and setup times. *Appl. Soft Comput.* **169**, 112614. https://doi.org/10.1016/j.asoc.2024.112614 (2025).
16. Li, J., Han, Y., Gao, K., Xiao, X. & Duan, P. Bi-population balancing multi-objective algorithm for fuzzy flexible job shop with energy and transportation. *IEEE Trans. Autom. Sci. Eng.* **21**, 4686–4702. https://doi.org/10.1109/TASE.2023.3300922 (2024).
17. Hu, C., Zheng, R., Lu, S., Liu, X. & Cheng, H. Integrated optimization of production scheduling and maintenance planning with dynamic job arrivals and mold constraints. *Comput. Ind. Eng.* **186**, 109708. https://doi.org/10.1016/j.cie.2023.109708 (2023).
18. Meng, L., Zhang, C., Zhang, B. & Ren, Y. Mathematical modeling and optimization of energy-conscious flexible job shop scheduling problem with worker flexibility. *IEEE Access* **7**, 68043–68059. https://doi.org/10.1109/ACCESS.2019.2916468 (2019).
19. Zhang, S. et al. Dual resource constrained flexible job shop scheduling based on improved quantum genetic algorithm. *Machines* **9**, 108. https://doi.org/10.3390/machines9060108 (2021).
20. Zhang, H., Xu, G., Pan, R. & Ge, H. A novel heuristic method for the energy-efficient flexible job-shop scheduling problem with sequence-dependent set-up and transportation time. *Eng. Optim.* **54**, 1646–1667. https://doi.org/10.1080/0305215X.2021.1949007 (2022).
21. Gao, K. Z. et al. An improved artificial bee colony algorithm for flexible job-shop scheduling problem with fuzzy processing time. *Expert Syst. Appl.* **65**, 52–67. https://doi.org/10.1016/j.eswa.2016.07.046 (2016).
22. Chen, N., Xie, N. & Wang, Y. An elite genetic algorithm for flexible job shop scheduling problem with extracted grey processing time. *Appl. Soft Comput.* **131**, 109783. https://doi.org/10.1016/j.asoc.2022.109783 (2022).
23. Graham, R. L., Lawler, E. L., Lenstra, J. K. & Kan, A. H. G. R. Optimization and approximation in deterministic sequencing and scheduling: A survey. *Ann. Discret. Math.* **5**, 287–326. https://doi.org/10.1016/S0167-5060(08)70356-X (1979).
24. Adams, J., Balas, E. & Zawack, D. The shifting bottleneck procedure for job shop scheduling. *Manag. Sci.* **34**, 391–401. https://doi.org/10.1287/mnsc.34.3.391 (1988).
25. Mason, S. J., Fowler, J. W. & Carlyle, W. M. A modified shifting bottleneck heuristic for minimizing total weighted tardiness in complex job shops. *J. Sched.* **5**, 247–262. https://doi.org/10.1002/jos.102 (2002).
26. Mason, S. J., Fowler, J. W., Carlyle, W. M. & Montgomery, D. C. Heuristics for minimizing total weighted tardiness in complex job shops. *Int. J. Prod. Res.* **43**, 1943–1963. https://doi.org/10.1080/00207540412331331399 (2005).
27. Mönch, L. & Drießel, R. A distributed shifting bottleneck heuristic for complex job shops. *Comput. Ind. Eng.* **49**, 363–380. https://doi.org/10.1016/j.cie.2005.06.004 (2005).
28. Mönch, L., Schabacker, R., Pabst, D. & Fowler, J. W. Genetic algorithm-based subproblem solution procedures for a modified shifting bottleneck heuristic for complex job shops. *Eur. J. Oper. Res.* **177**, 2100–2118. https://doi.org/10.1016/j.ejor.2005.12.020 (2007).
29. Mönch, L. & Zimmermann, J. A computational study of a shifting bottleneck heuristic for multi-product complex job shops. *Prod. Plann. Control* **22**, 25–40. https://doi.org/10.1080/09537287.2010.490015 (2011).
30. Barua, A., Raghavan, N., Upasani, A. & Uzsoy, R. Implementing global factory schedules in the face of stochastic disruptions. *Int. J. Prod. Res.* **43**, 793–818. https://doi.org/10.1080/00207540412331282024 (2005).
31. Upasani, A. A., Uzsoy, R. & Sourirajan, K. A problem reduction approach for scheduling semiconductor wafer fabrication facilities. *IEEE Trans. Semicond. Manuf.* **19**, 216–225. https://doi.org/10.1109/TSM.2006.873510 (2006).
32. Sourirajan, K. & Uzsoy, R. Hybrid decomposition heuristics for solving large-scale scheduling problems in semiconductor wafer fabrication. *J. Sched.* **10**, 41–65. https://doi.org/10.1007/s10951-006-0325-5 (2007).
33. Upasani, A. & Uzsoy, R. Integrating a decomposition procedure with problem reduction for factory scheduling with disruptions: A simulation study. *Int. J. Prod. Res.* **46**, 5883–5905. https://doi.org/10.1080/00207540601156215 (2008).
34. Pfund, M. E., Balasubramanian, H., Fowler, J. W., Mason, S. J. & Rose, O. A multi-criteria approach for scheduling semiconductor wafer fabrication facilities. *J. Sched.* **11**, 29–47. https://doi.org/10.1007/s10951-007-0049-1 (2008).
35. Yugma, C., Dauzère-Pérès, S., Artigues, C., Derreumaux, A. & Sibille, O. A batching and scheduling algorithm for the diffusion area in semiconductor manufacturing. *Int. J. Prod. Res.* **50**, 2118–2132. https://doi.org/10.1080/00207543.2011.575090 (2012).
36. Knopp, S., Dauzère-Pérès, S. & Yugma, C. A batch-oblivious approach for complex job-shop scheduling problems. *Eur. J. Oper. Res.* **263**, 50–61. https://doi.org/10.1016/j.ejor.2017.04.050 (2017).
37. Ham, A. M. & Cakici, E. Flexible job shop scheduling problem with parallel batch processing machines: MIP and CP approaches. *Comput. Ind. Eng.* **102**, 160–165. https://doi.org/10.1016/j.cie.2016.11.001 (2016).
38. Ham, A. Flexible job shop scheduling problem for parallel batch processing machine with compatible job families. *Appl. Math. Model.* **45**, 551–562. https://doi.org/10.1016/j.apm.2016.12.034 (2017).
39. Wu, K., Huang, E., Wang, M. & Zheng, M. Job scheduling of diffusion furnaces in semiconductor fabrication facilities. *Eur. J. Oper. Res.* **301**, 141–152. https://doi.org/10.1016/j.ejor.2021.09.044 (2022).
40. Boyer, V., Vallikavungal, J., Cantú Rodríguez, X. & Salazar-Aguilar, M. A. The generalized flexible job shop scheduling problem. *Comput. Ind. Eng.* **160**, 107542. https://doi.org/10.1016/j.cie.2021.107542 (2021).
41. Zeng, C. et al. Auction-based approach with improved disjunctive graph model for job shop scheduling problem with parallel batch processing. *Eng. Appl. Artif. Intell.* **110**, 104735. https://doi.org/10.1016/j.engappai.2022.104735 (2022).
42. Xue, L., Zhao, S., Mahmoudi, A. & Feylizadeh, M. R. Flexible job-shop scheduling problem with parallel batch machines based on an enhanced multi-population genetic algorithm. *Complex Intell. Syst.* **10**, 4083–4101. https://doi.org/10.1007/s40747-024-01374-7 (2024).
43. Ji, B. et al. Novel model and solution method for flexible job shop scheduling problem with batch processing machines. *Comput. Oper. Res.* **161**, 106442. https://doi.org/10.1016/j.cor.2023.106442 (2024).
44. Kennedy, J. & Eberhart, R. *Particle Swarm Optimization. Proc. ICNN'95* 942–1948. https://doi.org/10.1109/ICNN.1995.488968 (1995).
45. Dorigo, M., Maniezzo, V. & Colorni, A. Ant system: Optimization by a colony of cooperating agents. *IEEE Trans. Syst. Man Cybern. B* **26**, 29–41. https://doi.org/10.1109/3477.484436 (1996).
46. Karaboga, D. & Basturk, B. A powerful and efficient algorithm for numerical function optimization: Artificial bee colony (ABC) algorithm. *J. Glob. Optim.* **39**, 459–471. https://doi.org/10.1007/s10898-007-9149-x (2007).

47. Ding, H. J. & Gu, X. S. Improved particle swarm optimization algorithm based on novel encoding and decoding schemes for flexible job shop scheduling problem. *Comput. Oper. Res.* **121**, 104951. https://doi.org/10.1016/j.cor.2020.104951 (2020).
48. Shi, J. X., Chen, M. Z., Ma, Y. M. & Qiao, F. A new boredom-aware dual-resource constrained flexible job shop scheduling problem using a two-stage multi-objective particle swarm optimization algorithm. *Inform. Sci.* **643**, 119141. https://doi.org/10.1016/j.ins.2023.119141 (2023).
49. Zhang, S. C. & Wong, T. N. Flexible job-shop scheduling/rescheduling in dynamic environment: A hybrid MAS/ACO approach. *Int. J. Prod. Res.* **55**, 3173–3196. https://doi.org/10.1080/00207543.2016.1267414 (2017).
50. Li, Y. B. et al. A reinforcement learning-artificial bee colony algorithm for flexible job shop scheduling problem with lot streaming. *Appl. Soft Comput.* **146**, 110658. https://doi.org/10.1016/j.asoc.2023.110658 (2023).
51. Mirjalili, S., Mirjalili, S. M. & Lewis, A. Grey wolf optimizer. *Adv. Eng. Softw.* **69**, 46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007 (2014).
52. Mirjalili, S. & Lewis, A. The Whale optimization algorithm. *Adv. Eng. Softw.* **95**, 51–67. https://doi.org/10.1016/j.advengsoft.2016.01.008 (2016).
53. Samareh Moosavi, S. H. & Khatibi Bardsiri, V. Satin bowerbird optimizer: A new optimization algorithm to optimize ANFIS for software development effort estimation. *Eng. Appl. Artif. Intell.* **60**, 1–15. https://doi.org/10.1016/j.engappai.2017.01.006 (2017).
54. Dhiman, G. & Kumar, V. Emperor penguin optimizer: A bio-inspired algorithm for engineering problems. *Knowl. Based Syst.* **159**, 20–50. https://doi.org/10.1016/j.knosys.2018.06.001 (2018).
55. Jain, M., Singh, V. & Rani, A. A novel nature-inspired algorithm for optimization: Squirrel search algorithm. *Swarm Evol. Comput.* **44**, 148–175. https://doi.org/10.1016/j.swevo.2018.02.013 (2019).
56. Heidari, A. A. et al. Harris Hawks optimization: Algorithm and applications. *Future Gener. Comput. Syst.* **97**, 849–872. https://doi.org/10.1016/j.future.2019.02.028 (2019).
57. Fathollahi-Fard, A. M., Hajiaghaei-Keshteli, M. & Tavakkoli-Moghaddam, R. Red deer algorithm (RDA): A new nature-inspired meta-heuristic. *Soft Comput.* **24**, 14637–14665. https://doi.org/10.1007/s00500-020-04812-z (2020).
58. Xie, L. et al. Tuna swarm optimization: A novel swarm-based metaheuristic algorithm for global optimization. *Comput. Intell. Neurosci.* **2021**, 9210050. https://doi.org/10.1155/2021/9210050 (2021).
59. Jia, H. M., Peng, X. X. & Lang, C. B. Remora optimization algorithm. *Expert Syst. Appl.* **185**, 115665. https://doi.org/10.1016/j.eswa.2021.115665 (2021).
60. Abdollahzadeh, B., Gharehchopogh, F. S. & Mirjalili, S. African vultures optimization algorithm: A new nature-inspired metaheuristic algorithm for global optimization problems. *Comput. Ind. Eng.* **158**, 107408. https://doi.org/10.1016/j.cie.2021.107408 (2021).
61. Braik, M., Hammouri, A., Atwan, J., Al-Betar, M. A. & Awadallah, M. A. White shark optimizer: A novel bio-inspired meta-heuristic algorithm for global optimization problems. *Knowl. Based Syst.* **243**, 108457. https://doi.org/10.1016/j.knosys.2022.108457 (2022).
62. Trojovský, P. & Dehghani, M. A new bio-inspired metaheuristic algorithm for solving optimization problems based on walruses behavior. *Sci. Rep.* **13**, 8775. https://doi.org/10.1038/s41598-023-35863-5 (2023).
63. Lin, C., Cao, Z. & Zhou, M. Learning-based grey wolf optimizer for stochastic flexible job shop scheduling. *IEEE Trans. Autom. Sci. Eng.* **19**, 3659–3671. https://doi.org/10.1109/TASE.2021.3129439 (2022).
64. Luan, F. et al. Improved whale algorithm for solving the flexible job shop scheduling problem. *Mathematics* **7**, 384. https://doi.org/10.3390/math7050384 (2019).
65. Fan, C. S., Wang, W. T. & Tian, J. Flexible job shop scheduling with stochastic machine breakdowns by an improved tuna swarm optimization algorithm. *J. Manuf. Syst.* **74**, 180–197. https://doi.org/10.1016/j.jmsy.2024.03.002 (2024).
66. He, Z., Tang, B. & Luan, F. An improved African vulture optimization algorithm for dual-resource constrained multi-objective flexible job shop scheduling problems. *Sensors* **23**, 90. https://doi.org/10.3390/s23010090 (2022).
67. Bierwirth, C. & Mattfeld, D. C. Production scheduling and rescheduling with genetic algorithms. *Evol. Comput.* **7**, 1–17. https://doi.org/10.1162/evco.1999.7.1.1 (1999).
68. Viswanathan, G. M. et al. Lévy flights search patterns of biological organisms. *Physica A* **295**, 85–88. https://doi.org/10.1016/S0378-4371(01)00057-7 (2001).
69. Ponsich, A. & Coello, C. C. A hybrid differential evolution—tabu search algorithm for the solution of job-shop scheduling problems. *Appl. Soft Comput.* **13**, 462–474. https://doi.org/10.1016/j.asoc.2012.07.034 (2013).
70. Wang, L. Y., Cao, Q. J., Zhang, Z. X., Mirjalili, S. & Zhao, W. G. Artificial rabbits optimization: A new bio-inspired meta-heuristic algorithm for solving engineering optimization problems. *Eng. Appl. Artif. Intell.* **114**, 105082. https://doi.org/10.1016/j.engappai.2022.105082 (2022).
71. Mirjalili, S., Mirjalili, S. M. & Hatamlou, A. Multi-verse optimizer: A nature-inspired algorithm for global optimization. *Neural Comput. Appl.* **27**, 495–513. https://doi.org/10.1007/s00521-015-1870-7 (2016).
72. Samareh Moosavi, S. H. & Bardsiri, V. K. Poor and rich optimization algorithm: A new human-based and multi-population algorithm. *Eng. Appl. Artif. Intell.* **86**, 165–181. https://doi.org/10.1016/j.engappai.2019.08.025 (2019).
73. Phan, T. M., Duong, M. P., Doan, A. T., Duong, M. Q. & Nguyen, T. T. Optimal design and operation of wind turbines in radial distribution power grids for power loss minimization. *Appl. Sci.* **14**, 1462. https://doi.org/10.3390/app14041462 (2024).
74. Awad, A., Kamel, S., Hassan, M. H. & Zeinoddini-Meymand, H. Optimal allocation of flexible AC transmission system (FACTS) for wind turbines integrated power system. *Energy Sci. Eng.* **12**, 181–200. https://doi.org/10.1002/ese3.1628 (2024).
75. Cheng, M. Y. & Sholeh, M. N. Optical microscope algorithm: A new metaheuristic inspired by microscope magnification for solving engineering optimization problems. *Knowl. Based Syst.* **279**, 110939. https://doi.org/10.1016/j.knosys.2023.110939 (2023).
76. Rao, R. V., Savsani, V. J. & Vakharia, D. P. Teaching–learning-based optimization: A novel method for constrained mechanical design optimization problems. *Comput. Aided Des.* **43**, 303–315. https://doi.org/10.1016/j.cad.2010.12.015 (2011).

Acknowledgements

This work is supported by National Natural Science Foundation of China (Grant No. 52275487), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2021A1515012395), Natural Science Foundation of Changsha (No. kq2208001), and Hunan Provincial Department of Education (No. 21A0590). The authors thank the anonymous reviewers for their valuable and constructive comments that greatly helped improve the quality and completeness of the paper.

Author contributions

Shengping Lv: Conducted investigation, led conceptualization, managed data curation, performed formal analysis, established methodology, provided resources, was responsible for writing the original draft, carried out writing-review & editing, and oversaw project administration. Jianwei Zhuang: Handled data curation, executed formal analysis, developed methodology, worked on software, ensured validation, contributed to visualization, and participated in writing the original draft and writing-review & editing. Zhuohui Li: Managed data curation, carried out formal analysis, devised methodology, worked with software, achieved validation, and assisted with

visualization. Hucheng Zhang: Took care of data curation, conducted formal analysis, formulated methodology, used software, attained validation, and helped with visualization. Hong Jin: Performed investigation, contributed to conceptualization, carried out formal analysis, and engaged in review & editing. Shengxiang Lü: Participated in conceptualization, developed methodology, and was involved in writing-review & editing.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

applied sciences

MDPI

Article

An FCM–GABPN Ensemble Approach for Material Feeding Prediction of Printed Circuit Board Template

Shengping Lv, Rongheng Xian, Denghui Li, Binbin Zheng and Hong Jin *

College of Engineering, South China Agricultural University, Guangzhou 510642, China; lvshengping@scau.edu.cn (S.L.); xrh@stu.scau.edu.cn (R.X.); denghui.li@stu.scau.edu.cn (D.L.); zhengbinbin@stu.scau.edu.cn (B.Z.)

* Correspondence: hjin@scau.edu.cn; Tel.: +86-138-2603-0067

Received: 30 August 2019; Accepted: 17 October 2019; Published: 21 October 2019

Featured Application: The application of the work is to optimize the material feeding of a printed circuit board (PCB) template and therefore reduce the comprehensive cost caused by surplus and supplemental feeding.

Abstract: Accurate prediction of material feeding before production for a printed circuit board (PCB) template can reduce the comprehensive cost caused by surplus and supplemental feeding. In this study, a novel hybrid approach combining fuzzy c-means (FCM), feature selection algorithm, and genetic algorithm (GA) with back-propagation networks (BPN) was developed for the prediction of material feeding of a PCB template. In the proposed FCM–GABPN, input templates were firstly clustered by FCM, and seven feature selection mechanisms were utilized to select critical attributes related to scrap rate for each category of templates before they are fed into the GABPN. Then, templates belonging to different categories were trained with different GABPNs, in which the separately selected attributes were taken as their inputs and the initial parameter for BPNs were optimized by GA. After training, an ensemble predictor formed with all GABPNs can be taken to predict the scrap rate. Finally, another BPN was adopted to conduct nonlinear aggregation of the outputs from the component BPNs and determine the predicted feeding panel of the PCB template with a transformation. To validate the effectiveness and superiority of the proposed approach, the experiment and comparison with other approaches were conducted based on the actual records collected from a PCB template production company. The results indicated that the prediction accuracy of the proposed approach was better than those of the other methods. Besides, the proposed FCM–GABPN exhibited superiority to reduce the surplus and/or supplemental feeding in most of the case in simulation, as compared to other methods. Both contributed to the superiority of the proposed approach.

Keywords: printed circuit board (PCB); material feeding; fuzzy c-means; back-propagation networks; genetic algorithm

1. Introduction

Printed circuit board (PCB) is found in practically all electrical and electronic equipment, being the base of the electronics industry [1]. Due to the rapid development of computer, communication, consumer electronics, 5G, and automotive electronics, as well as the update of their products, the demand of PCB orders with specialized design features and manufacturing requirements, often referred to as a PCB template in the factory, has increased rapidly. The mode of production for a PCB factory with lots of template orders has changed from mass production to customer-oriented small-batch production, and therefore causes companies to face serious challenges. Accurate prediction of material feeding for each order is one of the critical problems.

After the feeding area and production panel (production unit) of each template order is accurately predicted, several goals (including the reduction of comprehensive cost caused overproduction or supplemental feeding, alleviation of environment pollution, improvement of on-time delivery, etc.) can be simultaneously achieved. However, it is difficult to determine the material feeding area of each PCB template order in advance of the production by manual feeding. Many factories undergo the violent fluctuation in both surplus and supplemental feeding by empirical manual feeding. Individualized surplus templates can be placed only in inventory or directly disposed, while supplemental feeding brings extra production costs and increases the probability of delivery tardiness compensation [2]. Furthermore, surplus products bring extra chemical and heavy metal pollution for production and disposal. This motivates us to explore the pattern of historical records that facilitate more reasonable and accurate prediction of material feeding for new template orders.

There are many applications of data mining (DM) or big data for quality improvement and optimization of PCB production [3]. Lee et al. [4] developed a data mining (DM)-based approach to predict the yield of a PCB, using the event sequence. Tsai [5] proposed a hybrid DM approach for soldering quality classification by using self-organizing map (SOM) and K-means, based on the statistical process control databases. DB-based PCB manufacturing process optimization has also attracted many researches, such as the parameter optimization of hot solder dip [6], stencil printing process [7,8], reflow soldering [9,10], fluid dispensing for microchip encapsulation [11], and wave soldering [12,13] for a component surface mount on a PCB. These models always combined artificial neural network (ANN), support vector machine (SVM), and multiple linear regression (MLR) for quality prediction with GA for parameters optimization simultaneously [7,10,11]. Meanwhile, many DM approaches like adaptive genetic algorithm (GA)-artificial neural network (ANN) [14], decision tree (DT) [15] have been employed for the defect diagnosis of PCB. The DM and/or big data were also widely adopted for smart production in different industries, not only for PCB manufacturing, and many reviews of these applications were reported in recent two years [3,16,17]. However, few of the aforementioned studies are on the prediction and optimization of material feeding in PCB template orders. In addition, the reviewed studies seldom considered the situation of diverse examples with different critical influence factors that require different prediction models to improve the prediction accuracy.

Considering the abovementioned requirement of a PCB template, Lv et al. [2] developed a hybrid model, multiple structural change (MSC)–ANN, to predict the feeding panel for each template, in which the template samples were pre-classified based on the required panel by the multiple structural change model. Then, the critical attributes for each category were selected based on neighborhood component approach; finally, the ANN prediction models were established for each category. The experimental results indicated that the attempt of the pre-classifying of inputs and establishing a prediction model for each category can indeed improve the prediction accuracy of material feeding for PCB template production. However, the MSC–ANN considered only one attribute to classify the sample, and the attributes were selected for each category by only one feature selection approach. Meanwhile, a template might be partitioned into multiple categories with different degrees, while the pre-partition based on MSC is a hard clarification. Therefore, it seems to be insufficient to predict the production-feeding panel of each template by using a prediction approach suited to a single category. Besides, MSC–ANN cannot handle a template order belonging to the border of two adjacent categories, because neither of the prediction models for the two categories is suitable for the template. Furthermore, the initialization parameter optimization of ANN benefits accuracy improvement [11,18,19] but was not considered.

For tackling the aforementioned difficult problems, a fuzzy c-means (FCM) classifier was adopted to handle the fuzzy classification and back-propagation network (BPN) ensemble with an aggregator BPN was employed to tackle the prediction by considering the membership degree of each template. The linear correlation (LC) [20], maximum information coefficient (MIC) [21], recursive feature elimination (RFE) [22], LR [23], lasso regression [24], ridge regression [25], and random forest regression (RFR) [26]

seven feature selection approaches were taken to select the critical attributes of each category divided by FCM. The GA was used to optimize the initialization parameters of BPN for each category. The reason for employing an FCM is that it accounts for the flexible classification (a template might be clustered into multiple categories with different membership degrees) and was widely used in many fields [27–29]. The reason for applying GA is that GA is easy to encode the problem and achieve good optimization results. It was also widely employed to optimize the structure (the number of layers and nodes in each hidden layer) and/or initial weight and bias [18,19,30] for the purpose of improving the prediction effectiveness of BPN. An aggregator BPN was adopted to conduct nonlinear aggregation because, theoretically, a BPN can approximate any nonlinear relationship [31].

In the proposed FCM–GABPN approach, input samples were first clustered with FCM, and seven feature selection methods were utilized to select critical attributes related to scrap rate for each category (a cluster is taken as a category) of PCB templates before they were fed into the BPN. Then, samples belonging to different categories were trained with different BPNs, in which the separately selected attributes were taken as their inputs and the initial parameters were optimized with GA. After training, an ensemble predictor formed with all GABPNs was taken to predict the scrap rate. Finally, another BPN was adopted to conduct nonlinear aggregation of the outputs from the component BPNs and determine the predicted feeding panel of the PCB template with a transformation. The proposed FCM–GABPN approach is illustrated in Figure 1.

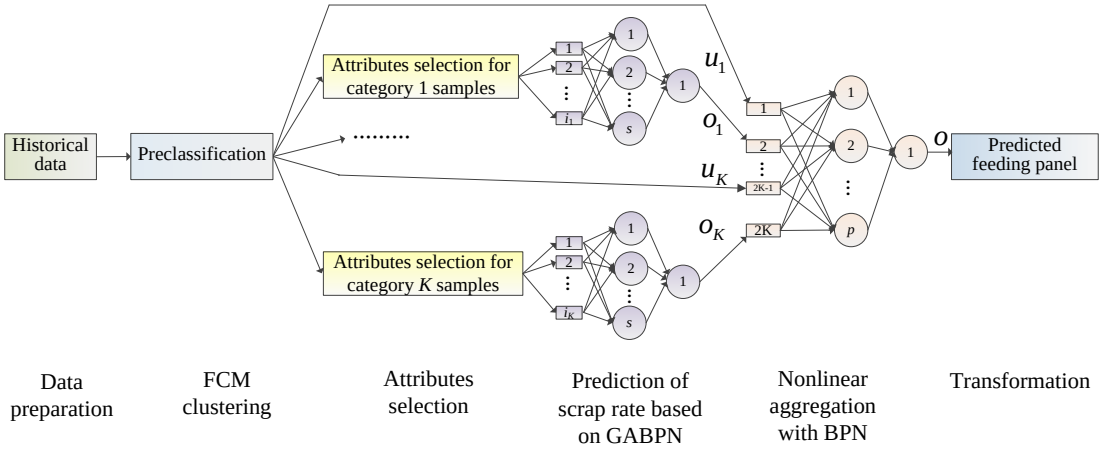


Figure 1. The architecture of the proposed fuzzy c-means–genetic algorithm with back-propagation networks (FCM–GABPN).

The remainder of the paper is organized as follows. In Section 2, variables specification and sample collection are described. The FCM, feature selection methods, GABPN, and the nonlinear aggregation BPN are introduced in Section 3, followed by experimental results and discussion in Section 4. Lastly, conclusions are given in Section 5.

2. Variables and Sample

The data used in this study were collected from Guangzhou FastPrint Technology Co., Ltd. A total of 56 variables inherited from an enterprise resource planning system combined with the derived variables were selected and specified in Table 1, in which variables 1 to 35 are the product/process attributes, while 36 to 56 are the statistic variables. The delivery unit in a panel, required quantity/panel/area, and delivery unit area, with No. 36, 38, 39, 47, and 46, respectively, can not only be taken as statistic items, but also attribute candidates for prediction model establishment. Set and unit are two types of delivery unit, whereas panel as a production unit will be partitioned into either set or unit according to the customer's requirement before delivery. If the number of final qualified set/unit (feeding set/unit minus the scrap set/unit) is larger than the demand number, it brings surplus sets/units; conversely, it causes supplemental feeding.

Table 1. Variables specification.

No.	Variable Name	Symbol	Description	Value Range
		Overall characteristics		
1	PCB thickness (mil)	*Pt*	Thickness of the ordered PCB	0.3–8
2	**Layer number**	*Ln*	Number of copper layer.	4–20
3	Rogers material	*Ro*	Whether substrate material is Rogers.	0/1
4	Plating frequency	*Plfr*	Number of plating operation.	0–4
5	**Number of operations**	*Noo*	Number of operations to produce the order.	16–71
6	Number of Prepreg	*NPP*	Number of Prepreg for lamination	1–50
7	Scrap units in a set	*Sus*	Allowed maximum scrap units in a set.	0–8
8	Photoelectric board	*Photb*	Whether the order is the specified board.	0/1
9	High frequency board	*Highfb*		
10	Test board	*Semictb*		
11	Negative film plating	*Nflp*	Whether the order takes negative film plating.	
12	Tinning copper	*Tinc*	Whether the order has tinning copper.	
13	IPCIII standard	*IPCIII*	Whether the order takes IPCIII or Huawei standard.	
14	Huawei standard	*Huawei*		
		Feature of internal/outer layer line		
15	**Minimum line width in internal layer (mil)**	*Mwil*	Minimum line width or space in core boards	3–100
16	**Minimum line space in internal layer(mil)**	*Mlsil*		1–137.66
17	**Minimum line width in outer layer(mil)**	*Mwol*	Minimum line width or space in outer layer	1–157.5
18	**Minimum line space in outer layer (mil)**	*Mlsol*		1.2–290
19	Average residual rate	*Arcr*	Average residual rate of copper layer	0.15%–94.75%
		Feature and operation information of hole		
20	Solder resist plug hole	*Srph*	Whether the order has the specified hole related operation.	0/1
21	Plug hole with resin	*Phwr*		
22	Second drilling	*Secd*		
23	Back drilling	*Bcdr*		
		Operation information of character/solder mask		
24	Character print	*Chaprt*	Whether the order has the specified character/solder mask related operation.	0/1
25	White oil solder mask	*White*		
26	Blue oil solder mask	*Blue*		
27	Black oil solder mask	*Black*		
		Surface finishing operation options		
28	**Hot-air solder leveling**	*Hasl*	Whether the order takes the specified surface finishing operation.	0/1
29	**Lead-free hot air solder leveling**	*Lfhasl*		
30	**Entek**	*Osp*		
31	**Cu/Ni/Au pattern plating**	*Cnapp*		
32	**Gold finger plating**	Gfig		
33	**Gold plating**	*Godp*		
34	**Soft Ni/Au plating**	*Snap*		
35	**Immersion Ag/Sn/Au**	*Iasa*		

Table 1. *Cont.*

No.	Variable Name	Symbol	Description	Value Range
		Statistic items		
36	Delivery unit in a panel	*Duap*	Number of delivery unit in a panel	1–262
37	Supplemental feeding frequency	*Supff*	Material feeding frequency minus 1	0–14
38	**Required quantity**	*Reqq*	Demand quantity of delivery unit minus delivery unit in inventory for the same order No.	1–3000
39	**Required panel**	*Reqp*	*Reqq*/*Duap* rounded up to the nearest integer	1–225
40	Feeding quantity	*Fedq*	Feeding number of delivery unit	2–6296
41	Least feeding panel	*Lfp*	*Reqq*/(1-*scrap rate*) rounded up to the nearest integer	1–245
42	Feeding panel	*Fedp*	Number of feeding panel	1–308
43	Scrap quantity	*Scraq*	Scrap number of delivery unit	0–712
44	Qualified quantity	*Qualq*	Qualified number of delivery unit	1–6226
45	Surplus quantity	*Surpq*	*Qualq*–*Fedq*	0–3226
46	Delivery unit area(m^2)	*Dunita*	Area of a delivery unit	0.001–0.393
47	**Required area(m^2)**	*Reqa*	*Reqq* × *Dunita*	0.001–25.74
48	Feeding area(m^2)	*Feda*	*Fedq* × *Dunita*	0.011–42.63
49	Scrap area(m^2)	*Scraa*	*Scraq* × *Dunita*	0–15.39
50	Qualified area(m^2)	*Quala*	*Qualq* × *Dunita*	0.009–37.49
51	Surplus area(m^2)	*Surpa*	*Surpq* × *Dunita*	0–25.45
52	Supplemental feeding rate	*Supfr*	*Supff* in a certain period/number of orders × 100%	18.83%
53	Scrap rate	*Scrar*	*Scraa*/*Feda* × 100%	0%–68.48%
54	Qualified rate	*Qualr*	*Quala*/*Feda* × 100%	31.52%–100%
55	Surplus rate	*Surpr*	*Surpa*/*Reqa* × 100%	0%–554.22%
56	Historical qualified rate	*Hquar*	The *Qualr* for the same order *No.* in the past 2 years	8.824%–100%

Note: New orders having no *Hquar* are replaced by the *Qualr* for orders having the same layer number and surface-finishing operation during the past 2 years.

On this basis, 30,117 samples of the orders were collected, multivariate boxplots [2] were conducted to detect the outliers, and, finally, 29,157 samples were left for this study. Performances of the proposed FCM–GABPN are compared to the other five approaches based on the same samples. Value range in the last column of Table 1 is the statistic result of the 29,157 samples, and variables 40 to 56 are the statistic results of the manual feeding adopted by FastPrint.

3. Methodology

The procedure of the proposed approach (FCM–GABPN) is shown in Figure 2, and various aspects of FCM–GABPN are discussed in the following subsections.

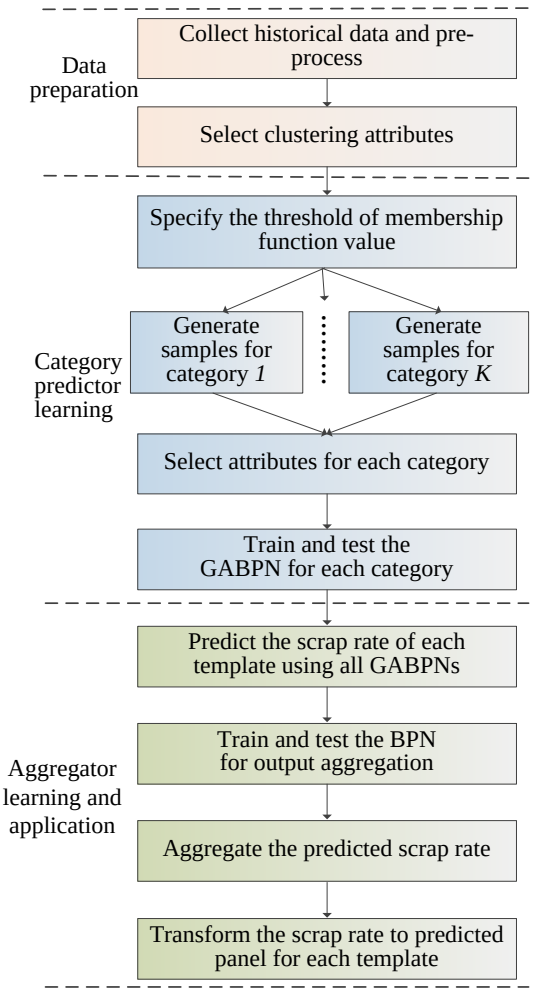


Figure 2. Procedure of the proposed FCM–GABPN.

3.1. Data Preparation and Template Classification with FCM

Data preparation is to collect the historical data of PCB templates for this study based on the variables given in Table 1. Then, 0–1 normalization was conducted for each variable for the purpose of reducing the influence of value-range difference. On this basis, the input attributes for FCM were selected based on the experience of experts from PCB workshops. The 17 attributes marked with boldface type in Table 1 were selected, in which the attributes of *Ln* and *Noo* represent the overall characteristics of the template; the *Mwil*, *Mlsil*, *Mwol*, and *Mlsol* are the design requirements of the hole and line; and the *Reqq*, *Reqp*, and *Reqa* are the production scale of each template order. Others are surface-finishing operation options.

Samples of templates were pre-classified into K categories with the selected 17 attributes by FCM before they were fed into the BPN. One recent example of FCM application is Tang et al. [27], in which FCM combining with adaptive neural network was applied to predict the lane changes by considering different simulation scenarios, and the results showed that the prediction performance and stability was considerably improved when compared with ANN, SVM, and MLR. Besides, Rezaee et al. [28] incorporated a dynamic FCM in ANN for the online prediction of companies in the stock exchange. According to experimental results, Rezaee et al.'s algorithm was efficient at clustering samples. In addition, Fathabadi [29] applied dynamic FCM clustering based ANN approach to reconfigure power-distribution networks. Experimental results indicated that Fathabadi's approach has some benefits, such as a short process time, a very simple structure, and higher accuracy compared to the others.

FCM performs clustering by minimizing $\sum_{c=1}^{C}\sum_{i=1}^{n}\mu_{i(c)}^{m}e_{i(c)}^{2}$, where C is the required number of clusters; n is the number of samples; $\mu_{i(c)}$ represents the membership of sample i belonging to cluster c; $e_{i(c)}$ measures the distance from samples i to the centroid of c; $m \in (1,\infty)$ is the hyper-parameter that controls how fuzzy the cluster will be. The procedure of applying FCM to cluster samples is as follows [31]:

(1) The cluster membership value, u_{ij} (the coefficient giving the degree of x_i being in the j^{th} cluster), are initialized randomly and establish an initial clustering result.

(2) (Iterations) obtain the centers of each cluster as $\overline{x}_{(c)} = \left\{\overline{x}_{(c)j}\right\}$, $\overline{x}_{(c)j} = \sum_{i=1}^{n} u_{i(c)}^{m}x_{ij} / \sum_{i=1}^{n} u_{i(c)}^{m}$, $1 \le j \le 17$, $u_{i(c)} = 1/\sum_{i=1}^{C}\left(e_{i(c)}/e_{i(c)}^{2}\right)^{2/(m-1)}$, $e_{i(c)} = \sqrt{\sum_{all\ j}\left(x_{ij}-\overline{x}_{(c)j}\right)^{2}}$, where x_{ij} is the j^{th} variable of the selected 17 attributes of sample i, and$\overline{x}_{(c)}$ is the centroid of cluster c.

(3) Re-measure the distance of each PCB template to the centroid of every cluster, and then recalculate the corresponding membership value.

(4) Stop if the number of iterations is larger than a set value. Otherwise, return to Step (2).

After clustering, samples of different categories (clusters) are then trained with different BPNs. First, a membership threshold value μ_L for selecting samples in network learning has to be determined. Only samples with $\mu_{i(c)} \ge \mu_L$ will be taken in training the BPN to obtain the weights and bias geared to the c^{th} category. As a result, a sample might be selected by multiple categories.

3.2. Attributes Selection for Each BPN Prediction Model

It is necessary to remove irrelevant and redundant attributes to reduce the complexity of analysis and the generated models, and also improve the efficiency of the whole modelling processes [2,32]. In this study, LC [20], MIC [21], RFE [22], LR [23], lasso regression [24], ridge regression [23], and RFR [24] seven feature selection approaches were employed to select critical attributes related to the scrap rate for each category of samples. The scarp rate can be taken as the dependent variable, and the independent variables are the attributes with No. 1–36, 38, 39, 46, and 47, given in Table 1. The score of independent variables obtained by each feature selection method were calculated and whose average score is greater than a certain threshold (e.g., 0.15) were taken as the input attribute of the prediction model.

The LC uses the linear correlation coefficient $lcc(x,y) = \mathrm{cov}(x,y)/\sqrt{\mathrm{var}(x)\mathrm{var}(y)}$ to measure the relationship between the (independent) variable x and variable y, where var is the variance of a variable and cov(x, y) denotes the covariance between x and y (namely scrap rate here) [20]. MIC is based on the idea that if a relationship exists between two variables, then a grid can be drawn on the scatterplot of the two variables that partition the data to encapsulate the relationship. To calculate the MIC of a set of two-variable data, all grids up to a maximal grid resolution are explored by computing for every pair of integers (x, y) the largest possible mutual information achievable by any x-by-y grid applied to the data. Then these mutual information (MI) values are normalized to ensure a fair comparison between grids of different dimensions and to obtain modified values between 0 and 1. Finally, the highest normalized MI achieved by any x-by-y grid as the value of MIC [21]. The main idea of RFE is to train an estimator based on the initial set of variables and weights are assigned to each one of them at first. Then, variables whose absolute weights are the smallest are pruned from the current set of variables. That procedure is recursively repeated on the pruned set until the desired number of variables to select is eventually reached [22].

The LR is to establish the regression equation of the dependent variable based on the independent variables, in which the importance of independent variables will be determined according to F-test. The smaller the value of F-test, the more important the variable is to the regression equation [21]. The lasso regression is a regularized LR by putting a L1 norm penalty on the regression coefficients.

Lasso regression will drive more coefficients of weak correlated independent variables to zero, and then facilitate the selection of variables with strong correlation [24]. The ridge regression is similar to lasso regression by putting a L2 norm penalty on the regression to penalize the weak correlated variables for the regression model establishment [25]. RFR is an ensemble of unpruned classification or regression trees, in which each branch of the trees will calculate the importance of each unused attribute in previous steps and then facilitate important-attribute selection simultaneously [26]. The above seven approaches were realized by the encapsulated functions in the machine learning library "sklearn" in this study.

3.3. GABPN-Based Scrap Rate Prediction for Each Category

The configuration of the BPN is established as follows:

(1) Input: the 0–1 normalized data of the selected attributes for each category.
(2) Architecture: Single hidden layer (number of nodes in the input layer + number of nodes in the output layer)/2 is one of the commonly used ways to determine the suitable number of neurons in the hidden layer. Therefore, the number of nodes in hidden layer is depended on the number of selected attributes in this study. In order to achieve better prediction accuracy (a large number of the hidden-layer nodes are theoretically conducive to improve the predicting accuracy) and to keep the consistency, the number of neurons in the hidden layer of each BPN was set to 12 for each category in the proposed approach, considering the number of selected attributes (up to 23 selected attributes for the samples that will be discussed in Section 4).
(3) Output: normalized scrap rate forecast of the template.
(4) Learning rule: Delta rule (the adjustment of weight and bias is proportional to the negative gradient of the error during the backward-propagation procedure).
(5) Propagation function: sigmoid activation function, $f(x_j) = 1/(1 + e^{-x_j})$.
(6) Learning rate: 0.05.
(7) Number of iterations: 25,000.

The performance of a BPN is sensitive to the initial condition. Therefore, the optimization of the initial weights and biases of BPN with GA was conducted. The design and configuration of GA is as follows:

(1) Encoding and decoding: The individual chromosome in the population was encoded as $[W_1, \Phi_1, W_2, \Phi_2]$ in which $W_1 = [w_{1,1}, w_{1,2}, .., w_{1,12}, w_{2,1}, w_{2,2}, .., w_{2,12}, w_{i,1}, w_{i,2}, .., .., w_{i,12}]$ (selected i attributes as input and the number of neurons in the hidden layer is 12) represents the weights between nodes in input layer and hidden layer; $W_1 = [w_{1,1}, w_{2,1}, ..., w_{12,1}]$ represents the weights between the nodes in hidden layer and output layer; $\Phi_1 = [\theta_1, \theta_2, ..., \theta_{12}]$ is the bias vector of nodes in the hidden layer; and Φ_2 is the bias of output node. The decoding is to assign corresponding weights and bias to each node based on the BPN structure, and then conduct the forward propagation to compute the output of each BPN.
(2) Population initialization: Each individual chromosome in the population was initialized randomly with its elements between −3 and 3, based on the encoding principle.
(3) Fitness evaluation: The sum of absolute error between reversely normalized scrap rate forecast $\hat{o}_k$ and actual scrap rate o_k was taken as the fitness $F = \sum|\hat{o}_k - o_k|$ for each individual. The smaller the fitness is, the more accurate prediction result it can obtain. Thereafter, the minimization objective function, which the problem seeks to optimize, is the same as the fitness function.
(4) Reproduction, crossover and mutation operation:

Reproduction: The roulette wheel selection was taken to select individuals for reproduction in which the fittest individuals have a greater chance of survival than weaker ones. The probability of each individual being selected is $p_i = (1/F_i)/\sum_{j=1}^{N}(1/F_j)$, where F_i is the fitness of the ith individual and N is the number of individuals.

Crossover: Two empty offspring chromosomes, O1 and O2, were initialized first, and two chromosomes, P1 and P2, were randomly selected from the reproduced population. The crossover location was randomly selected, and then the offspring O1 consisted of the genes of P1 before the crossover location and genes of P2 after the crossover location; while offspring O2 consisted of the genes of P2 before the crossover location and genes of P1 after the crossover location.

Mutation: One-point mutation was utilized as the mutation operator. The chromosome in the population was randomly selected, and one gene was chosen randomly from the selected chromosome. Then, a random r with the value in (0, 1) was generated to mutate the value. If $r > 0.5$, then $a_j = a_j + (a_j - a_{\max}) \times r$, otherwise $a_j = a_j + (a_{\min} - a_j) \times r$, where a_j is value of the j^{th} position in the chromosome selected for mutation, and $a_{\max}$ and $a_{\min}$ are the maximum and minimum of the j^{th} position of all chromosomes in current generation, respectively.

(5) Number of iterations: 100.

After the templates were clustered, a portion of the templates in each category were taken as "training samples" into the GABPN to determine the weights and bias values for the category. Three phases were involved at the training stage. First, the initial weight and bias were optimized according the GA. Second, the forward propagation is conducted, in which the inputs (selected attributes with bias) were multiplied with weights (weights of bias are 1), summated, and transferred to the hidden layer. The results of nodes in the hidden layer were further processed by sigmoid function and also transferred to the output layer with the same procedure. Finally, the output of GABPN was compared with the accurate scrap rate, and the accuracy of the GABPN, represented with mean squared error (MSE), was evaluated.

Subsequently, the backward pass which propagates derivatives (error between prediction and the actual value) from the output layer to hidden layers was conducted. The backward pass for a 3-layer BPN starts by computing the partial derivative for the output node (only one node here), and the error terms δ_j of nodes j in the hidden layers can be calculated according to $\delta_j = eW_j f'(x_j)$, in which e is error of the output node, W_j is the weight connecting node j to the output node, and $f'(x_j)$ is the derivative of the sigmoid activation function with the input x_j. On this basis, adjustments were made to the connection weights and bias to reduce the *MSE*. Network-learning stops when the iteration is greater than a given number in this study.

The trained GABPN was tested by the remaining portion of the templates in each category with the same performance indicator, *MSE.* Finally, the GABPN was used to predict the scrap rate of new templates that "completely" belonged to the clustered category. However, complete assignment of template to only a category is usually impossible. When a new template order is coming, the selected attributes associated with the new template are recorded, and the membership belonging to each category is calculated. Then, an ensemble predictor formed with all GABPNs can be taken to predict the scrap rate for the new template.

3.4. Nonlinear Aggregation with Another BPN and Transformation

For aggregating the predicted results from the component GABPNs into a single value representing the predicted scrap rate of the template, another BPN was employed in this study to conduct nonlinear aggregation, and the configuration is set as follows:

(1) Input: *2K* parameters consisted of the predicted results of each component GABPNs for the template and the membership values of the template belonging to each category.
(2) Architecture: Single hidden layer and the number of nodes in the hidden layer were set to the same as that in the input layer, *2K*.
(3) Output: normalized scrap rate forecast of the template.
(4) Learning rul*e*: Delta rule.
(5) Propagation function: sigmoid activation function.

(6) Learning rate: 0.05.
(7) Number of iterations: 10000.

The BPN also underwent training and testing. Then, the network output (i.e., the aggregation result) determined the normalized scrap-rate prediction of the template. Finally, the transformation of scrap rate to surplus rate and supplemental feeding rate were carried out. The reverse normalization was conducted for the output of the aggregation BPN and taking it as the predicted scrap rate (*Scrar_Pd*). Thereafter, the transformation for predicted feeding panel (*Fedp_Pd*) was conducted by $Fedq\_Pd = 100 \times \mathrm{Re}qq/(100 - Scrar\_Pd)$ and $Fedp\_Pd = \lceil Fedq\_Pd/Duap \rceil$, where *Reqq* is the required quantity, *Duap* is the delivery unit in a panel, and *Fedp_Pd* is the predicted panel.

3.5. Performance Indicators

In order to evaluate the effectiveness of the model, the *MSE*, mean absolute error (*MAE*), and mean absolute percentage error (*MAPE*) were adopted as the indicators to evaluate the performance of the approaches, in which the predicted data $\hat{o}_i$ are the predicted least feeding panel and the original data o_i are the least feeding panel. The *MSE, MAE*, and *MAPE* can be described as $MSE = \sum_{i=1}^{N} (\hat{o}_i - o_i)^2/N$, $MAE = \sum_{i=1}^{N} |\hat{o}_i - o_i|/N$, and $MAPE = \frac{1}{N}\sum_{i=1}^{N} \left|\frac{\hat{o}_i - o_i}{o_i}\right| \times 100$, respectively, where N is the number of samples.

The indicators surplus rate (*Surpr*) and supplemental feeding rate (*Supfr*) in the PCB template workshop were also considered. The predicted surplus rate (*Surpr_Pd*) and predicted supplemental feeding rate *Supfr_Pd* can be computed with Equations (10) and (11) in [2], respectively. The final performance is evaluated by the *MSE*, *MAE*, *MAPE*, *Supfr_Pd*, and *Surpr_Pd*.

4. Experimental Results and Discussions

The proposed FCM–GABPN was implemented by Python 3.6. The number of clusters was set to three while conducting FCM for the purpose of reducing the number of training, testing, and model maintenance in the workshop, but also to achieve good enough prediction accuracy based on some initial test. The hyper-parameter m that controls how fuzzy the cluster was commonly set to 2 [31], and it was adopted here. The maximum number of iterations of FCM was set to 800.

If FCM cluster samples fall into the category with the highest membership value, the templates will be cluster into C1, C2, and C3, with 20,773, 1354, and 7030 samples, respectively. The membership value giving the membership degree of each sample (samples were clustered into C1, C2, and C3 here for visualization) being in the three categories is illustrated in Figure 3. The membership degree of each sample will be taken as part of the input of the aggregator BPN to perform nonlinear aggregation, as shown in Figure 1.

The mean values of input attributes in the three categories are given in Figure 4. The mean values of *Reqq*, *Reqp*, and *Reqa* are comparatively different in the three categories; and they are the main attributes to distinguish and identify samples within each category, which is consistent with practice in which the workshop also regards order scale (*Reqq*, *Reqp*, and *Reqa*) as important variables to classify orders. Meanwhile, the mean values of *Mwil*, *Mlsil*, and *Mwol* in C2 is lower than the corresponding values in C1 and C3, but the *Ln* is higher, which indicates that, the higher *Ln* is, the denser the lines that coincide with the actual situation are.

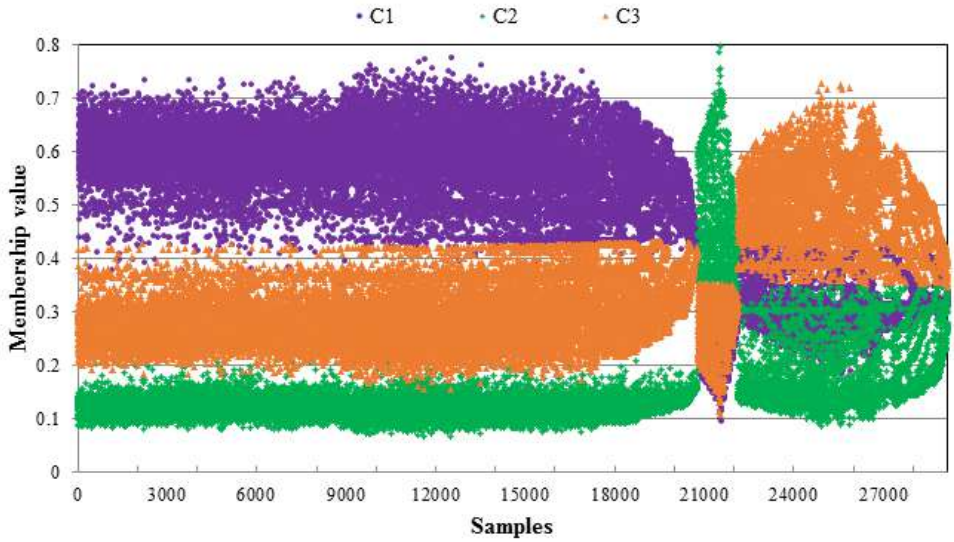
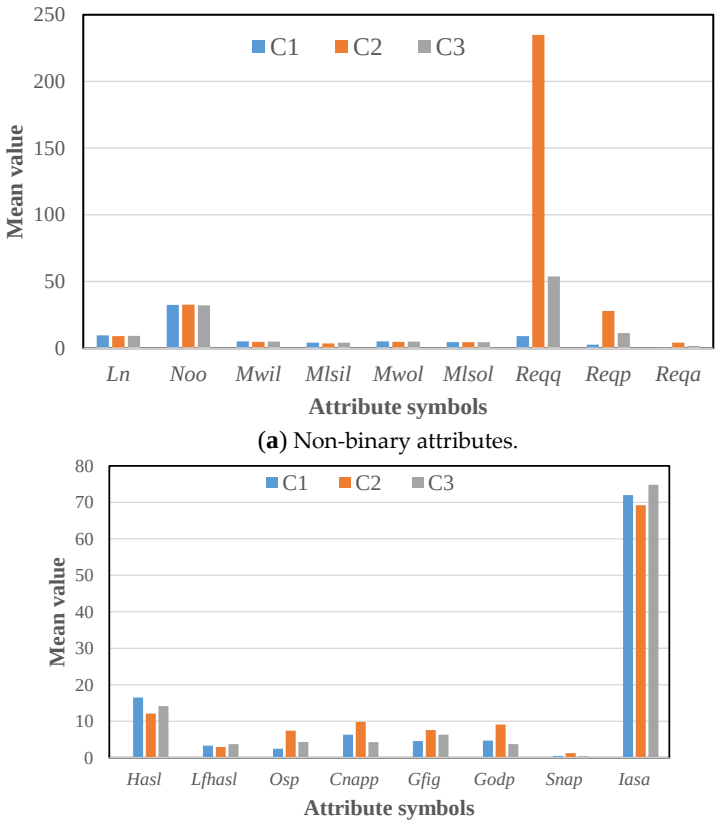


Figure 3. The membership value of each sample.



(a) Non-binary attributes.

(**b**) Binary attributes.

Figure 4. Comparison of mean value of attributes in each category.

The membership threshold μ_L should be specified for adopting samples in network learning. The numbers of samples within the three categories with different μ_L are given in Table 2. The 0.4 was selected as the threshold to generate training and testing samples, not only to make sure there were enough training and testing samples for each category, but also in case a template was clustered into multiple categories with different membership degrees. Then 2/3 and 1/3 of mutually exclusive samples were randomly selected as training and testing data for each category at each run. The unclassified

samples were not taken as the input to train each GABPN; however, they will be taken as the samples for final test. The numbers of training and testing samples for each category are given in Table 3.

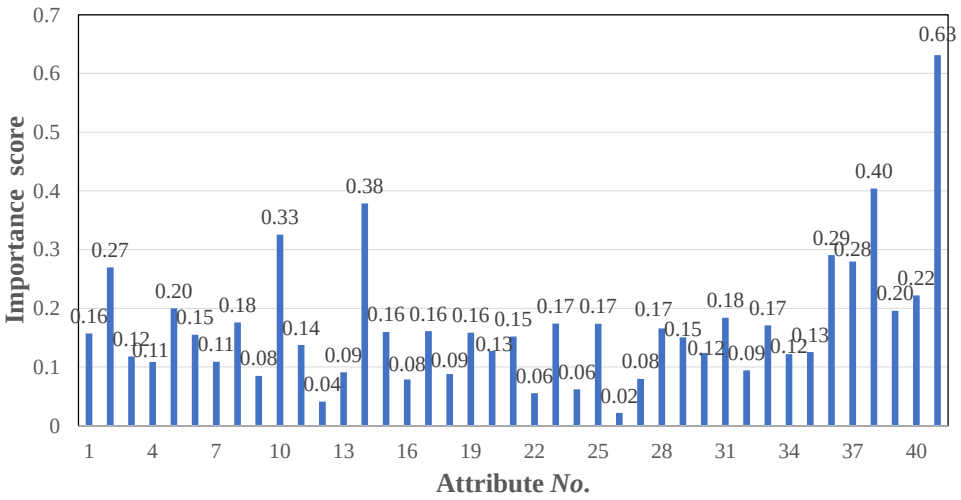
Table 2. Numbers of samples within the three categories with different μ_L .

μ_L	C1	C2	C3	Unclassified
0	29,157	29,157	29,157	0
0.1	29,156	27,608	29,157	0
0.2	28,922	3434	28,815	0
0.3	29,157	2193	15,529	0
0.4	21,230	973	7037	1355
0.5	17,717	393	3097	7951
0.6	8455	184	446	20,072
0.7	408	18	8	28,723
0.8	0	0	0	29,157
0.9	0	0	0	29,157
1	0	0	0	29,157

Table 3. Number of samples selected for training and testing.

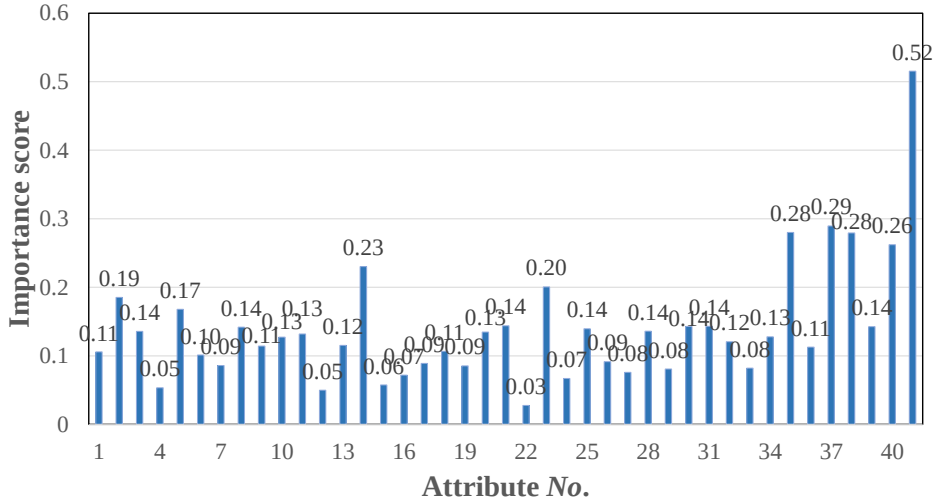
	Training Samples	Testing Samples
C1	14,153	8432
C2	649	1679
C3	4691	3701
All	19,448	9709

On the basis of the selected training samples and the 41 (variables No. 1–35, 36, 38, 39, 47, and 46 in Table 1) input attributes, the aforementioned seven feature selection mechanisms were employed to calculate the importance of each attribute on scrap rate. The importance (mean) score of each attribute for the three categories and all samples are given in Figure 5, and the corresponding No. is given in Table 4. The importance scores of attributes greater than 0.15 were chosen as the input of GABPN considering the number of selected attributes and confirmed by experts from the factory, and 23, 9, 20, and 16 attributes were selected for C1, C2, C3, and all data, respectively, that were marked with "▲" in Table 4. It can be seen that the critical attributes for different categories of samples are different, and one of the reasons is that the samples may have multiple complex distributions.

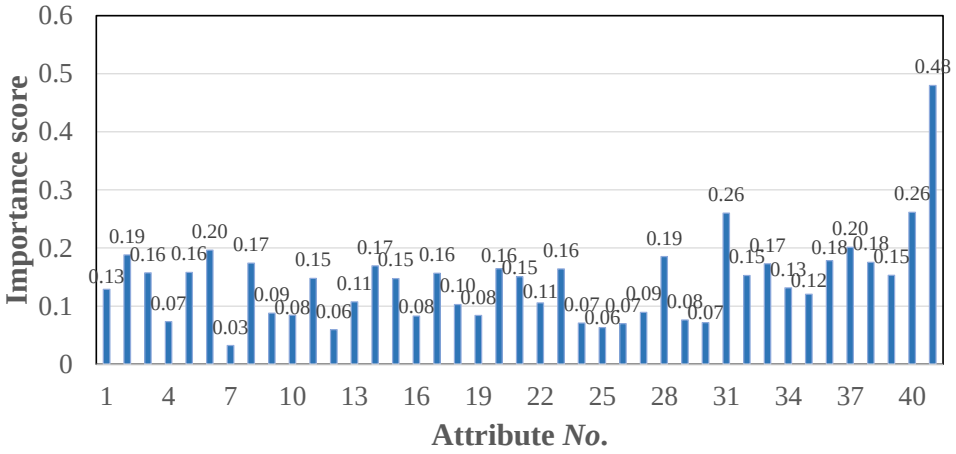


(**a**) Importance scores of attributes for samples in C1.

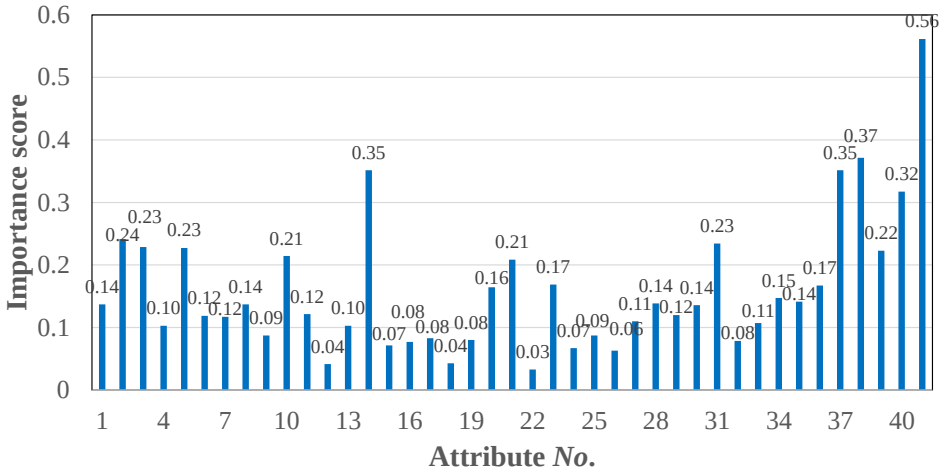
Figure 5. *Cont.*



(**b**) Importance scores of attributes for samples in C2.



(**c**) Importance scores of attributes for samples in C3.



(**d**) Importance scores of attributes for all samples.

Figure 5. Importance scores of attributes.

Table 4. Selected attributes for each category/all samples.

No.	Attributes	C1	C2	C3	All	No.	Attributes	C1	C2	C3	All
1	*Pt*	▲				22	*Secd*				
2	*Ln*	▲	▲	▲	▲	23	*Bcdr*	▲	▲	▲	▲
3	*Ro*			▲	▲	24	*Chaprt*				
4	*Plfr*					25	*White*	▲			
5	*Noo*	▲	▲	▲	▲	26	*Blue*				
6	*NPP*	▲		▲		27	*Black*				
7	*Sus*					28	*Hasl*	▲		▲	
8	*Photb*	▲		▲		29	*Lfhasl*	▲			
9	*Highfb*					30	*Osp*				
10	*Semictb*	▲			▲	31	*Cnapp*	▲		▲	▲
11	*Nflp*					32	*Gfig*			▲	
12	*Tinc*					33	*Godp*	▲		▲	
13	*IPCIII*					34	*Snap*				
14	*Huawei*	▲	▲	▲	▲	35	*Iasa*		▲		
15	*Mwil*	▲				36	*Duap*	▲		▲	▲
16	*Mlsil*					37	*Reqa*	▲	▲	▲	▲
17	*Mwol*	▲		▲		38	*Reqq*	▲	▲	▲	▲
18	*Mlsol*					39	*Reqp*	▲		▲	▲
19	*Arcr*	▲				40	*Hquar*	▲	▲	▲	▲
20	*Srph*			▲	▲	41	*Dunita*	▲	▲	▲	▲
21	*Phwr*	▲		▲	▲						

Each GABPN model was trained by the training samples and the selected attributes given in Table 4. All samples belonging to a category compete in the same way in training the GABPN geared to the category. Prediction models of GABPN were trained and tested for each category separately, while the aggregator BPN was trained with all the training samples and tested by all the testing samples.

The GA parameters of population size, crossover probability, mutational probability, and the number of iterations of the three GABPNs were set to 100, 0.8, 0.05, and 100, according to some initial test. The convergences of GA for the initial parameter optimization of the three BPNs are illustrated in Figure 6. On the basis of the optimized parameters, the three BPNs were trained in parallel, and the output of the three prediction models was set into the aggregator BPN, with the membership degree of each sample obtained by FCM given in Figure 3. The predicted feeding panel of each sample can be determined according to the transformation described in Section 3.4, based on the reversely normalized output of the aggregator BPN.

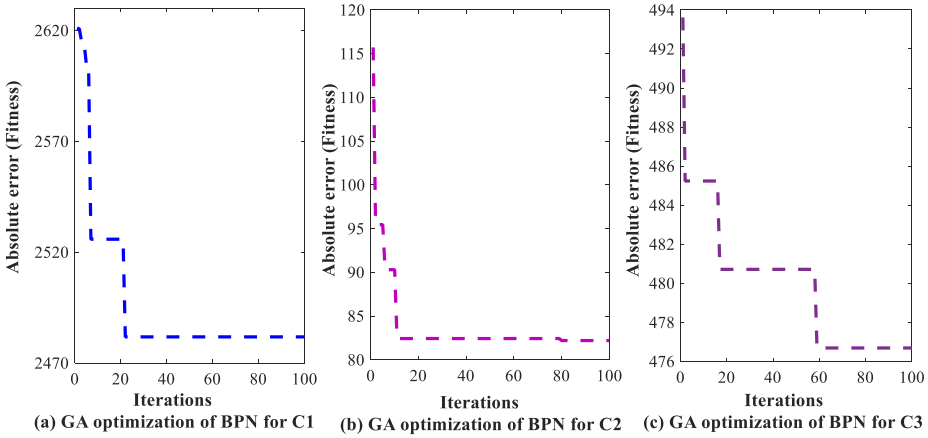


Figure 6. Convergences of GA for the initial parameter optimization of the three BPNs.

The regression of the predicted feeding panel versus the least feeding panel is given in Figure 7. Results indicated that the predicted feeding panel coincides well with the least feeding panel, and, therefore, the waste of surplus quantity and area can be reduced.

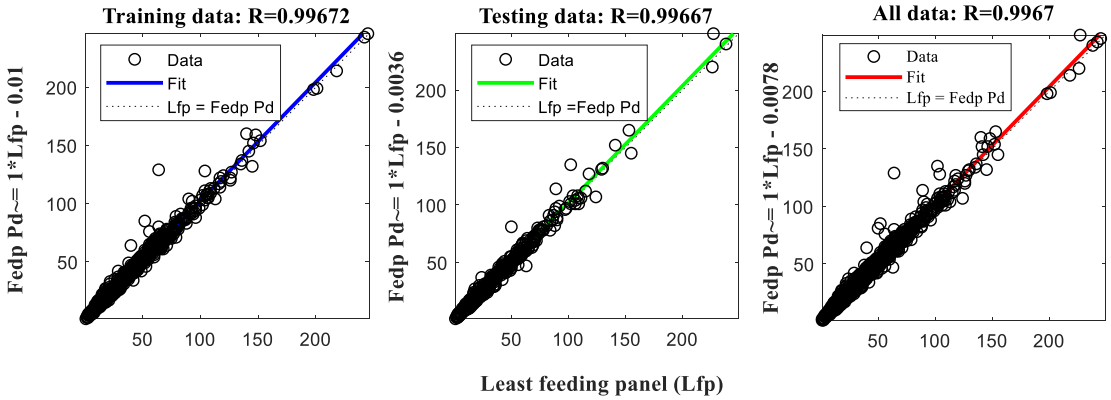


Figure 7. Regression of predicted feeding panel versus least feeding panel.

The FCM–BPBPN was compared to manual feeding, BPN, MSC–ANN, FCM–GABPN without aggregation (indicated with FCM–GABPN w/o aggregation), and FCM–BPN five approaches to quantify its performance. Manual feeding is to determine the feeding panel for each template based on worker in PCB factory. BPN is to establish a single BPN prediction model without pre-classification and takes the selected 16 attributes marked with "▲" in the column "All" of Table 4 as inputs. MSC–ANN [2] considered only required panel to classify the records and divide the samples into six groups. The FCM–GABPN w/o aggregation only applies the BPN to which the membership belonging is the highest and no BPN aggregation will be conducted. FCM–BPN has no GA to optimize the initial parameters of each BPN.

The testing samples were taken to evaluate the performance of the approaches, and the average *MSE*, *MAE*, *MAPE*, *Surpr_Pd*, and *Supfr Pd* of five runs for BPN, MSC–ANN, FCM–GABPN-w/o aggregation, FCM–BPN, and FCM–GABPN is given in Table 5. The improvement of different approaches comparing to the manual feeding (actual results of the factory) according to the performance indicators are also given, and the following discussions are made:

Table 5. Improvement of different approaches comparing to comparison basis—manual feeding.

Approaches	*MSE*	*MAE*	*MAPE*	*Surpr_Pd*(%)	*Supfr Pd*(%)
Manual feeding	22.862	1.467	29.161	28.49	18.53
BPN	2.143 (−90.63%)	0.759 (−48.26%)	17.962 (−38.40%)	16.85 (−40.86%)	13.02 (−29.74%)
MSC–ANN	1.272 (−94.44%)	0.396 (−73.01%)	5.542 (−81.00%)	12.25 (−57.00%)	12.78 (−31.03%)
FCM—GABPN-w/o aggregation	1.031 (−95.49%)	0.364 (−75.19%)	4.537 (−84.44%)	11.88 (−58.30%)	11.34 (−38.81%)
FCM–BPN	0.984 (−95.70%)	0.305 (−79.21%)	3.423 (−88.26%)	9.05 (−68.23%)	13.86 (−25.20%)
FCM–GABPN	0.935 (−95.91%)	0.249 (−83.03%)	3.041 (−89.57%)	8.50 (−70.16%)	12.78 (−31.03%)

(1) The prediction accuracy (measured with *MSE*, *MAE*, and *MAPE*) of the FCM–GABPN approach was significantly better than those of the other approaches, in most cases by achieving a 95.91%, 83.03%m and 89.57% reduction in *MSE*, *MAE*, and *MAPE*, respectively, over manual feeding. Meanwhile, the proposed FCM–GABPN exhibited superiority in the reduction of surplus and/or supplemental feeding in most of the case comparing to other methods by reducing 70.16% *Surpr_Pd* and 31.03% *Supfr Pd* over manual feeding.

(2) The advantages of FCM–GABPN over BPN without performing pre-classification were 5.28%, 34.77%, 51.17%, 29.30%, and 1.29% by reduction in *MSE*, *MAE*, *MAPE*, *Surpr_Pd*, and *Supfr Pd*,

respectively, and the superiority of MSC–ANN over BPN was 3.79%, 24.75%, 42.60%, 16.14 %, and 1.29%. The advantages of FCM–GABPN-w/o aggregation over BPN were 4.86%, 26.93%, 46.04%, 17.44%, and 9.07%, and the FCM–BPN over BPN were 5.28%, 30.95%, 49.86%, and 27.37% by reduction in *MSE*, *MAE*, *MAPE*, and *Surpr_Pd*, but with only a 4.54% increase in *Supfr.* Pre-classification and critical attribute selection for each category before prediction model establishment seem to have significant effect on the performance of material feeding prediction.

(3) The superiority FCM–GABPN-w/o aggregation, FCM–BPN, and FCM–GABPN over MSC–ANN according to the *MSE*, *MAE*, *MAPE*, and *Surpr_Pd* with only 5.83% inferiority in *Supfr Pd* for FCM–BPN approach and the same value for FCM-GABPN indicates that the pre-classification by clustering, which considers many attributes, surpassed the MSC classification, which considers only one attribute. In addition, the FCM–GABPN-w/o aggregation, FCM–BPN and FCM–GABPN only established three BPNs for the three categories of samples, while MSC–ANN pre-classified the samples into six categories and trained a prediction model for each category.

(4) FCM–BPN and FCM–GABPN achieved lower *MSE*, *MAE*, *MAPE*, and *Surpr_Pd* in comparison to FCM–GABPN-w/o aggregation, which indicates that applying the aggregator BPN to derive the representative value by considering the membership degree of each sample facilitates the prediction improvement for the four performance indicators. The 13.61% and 7.78% increase in *Surpr_Pd* for FCM–BPN and FCM–GABPN may be brought by the 9.07% and 11.86% reduction in *Supfr Pd*, respectively. In practice, the reduction of surplus feeding and supplemental feeding is conflicted because it is difficult to obtain the minimum value for both of them in the factory. However, the reduction of the surplus rate is a goal with the greatest cost impact in the factory because the individualized surplus template products can only be placed in inventory or directly destroyed, and the reduction of the surplus production will reduce the comprehensive cost caused by the waste of material, production, inventory, and disposal/ recycling.

(5) The FCM–GABPN surpassed FCM–BPN according to the five indicators that verify the effectiveness of the initialization optimization based on GA. The reason is that BPN is sensitive to the initial condition [30], especially for the samples in the three categories that were learned with different BPNs that may be influenced greatly by the combination of the BPN's initial parameters.

5. Conclusions

In order to enhance the accuracy of material feeding prediction of a PCB template, an ensemble predictor FCM–GABPN was proposed. In the proposed approach, the input templates were firstly clustered by FCM, and seven feature selection mechanisms were utilized to select critical attributes related to the scrap rate for each category of templates. Then, a GABPN was trained to predict the scrap rate for each category of templates, and the GABPNs for all the categories formed an ensemble predictor with a nonlinear aggregator BPN. Finally, the predicted feeding panel for each template was determined based on the predicted scrap rate with a transformation. The effectiveness and superiority were validated with many experiments based on the actual data. On the basis of the experimental results, conclusions and contributions are highlighted as follows:

(1) The accuracy of the proposed approach was better than those of the other approaches by achieving a 95.91%, 83.03%, and 89.57% reduction in *MSE*, *MAE*, and *MAPE*, respectively, over the comparison basis—manual feeding. Meanwhile, the FCM–GABPN's performance was superior to that of the other methods in the reduction of simulated surplus and/or supplemental feeding in most of the cases, by achieving a 70.16% reduction in *Surpr_Pd* and a 31.03% reduction in *Supfr Pd* over manual feeding.

(2) The material feeding prediction of PCB template problem considering category fuzziness of samples and the diverse samples with different influence factors is different from the existing production quality prediction and optimization problem, to the best of our knowledge. The novelty of the proposed FCM–GABPN is that we fuzzily clustered samples into different categories with FCM and specified a membership threshold to adopt samples for each category. Meanwhile,

component GABPN prediction model for each category was established with separately selected input attributes and GA optimized initial parameter. Furthermore, an aggregator BPN was employed to aggregate the predicted results of each GABPN by considering the membership values of each template.

Training an ensemble predictor with many sub-models that can extract shared attributes for similar templates automatically without explicit pre-classification needs to be studied, in which we do not have to divide samples, select critical attributes for each category, and build the prediction model separately. Meanwhile, the rapid development and evolution of PCB template should also be considered. The transfer and lifelong learning may be the mechanisms worthy of attempting, in order to handle the aforementioned problem.

Author Contributions: S.L. proposed the algorithm and wrote the paper; R.X. and D.L. implemented the algorithm; B.Z. conducted the experiments and analyzed the data; H.J. proposed the paper structure and wrote Section 4.

Funding: This research is funded by National Natural Science Foundation of China with grant number 51605169 and Natural Science Foundation of Guangdong, China with grant number 2018A030310216.

Acknowledgments: This paper is supported by the National Natural Science Foundation of China (Grant No. 51605169) and Natural Science Foundation of Guangdong, China (Grant No. 2018A030310216). The authors would like to express their appreciation to the agency. The authors wish to thank Guangzhou FastPrint Technology Co., Ltd. for providing data for the study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Marque, A.C.; Cabrera, J.M.; Malfatti, C.F. Printed circuit boards: A review on the perspective of sustainability. *J. Environ. Manag.* **2013**, *31*, 298–306. [CrossRef] [PubMed]
2. Lv, S.P.; Zheng, B.B.; Kim, H.; Yue, Q.S. Data mining for material feeding optimization of printed circuit board template production. *J. Electr. Comput. Eng.* **2018**, *2018*, 1852938. [CrossRef]
3. Lv, S.P.; Kim, H.; Zheng, B.B.; Jin, H. A review of data mining with big data towards its applications in the electronics industry. *Appl. Sci.* **2018**, *8*, 582. [CrossRef]
4. Lee, H.; Kim, C.O.; Ko, H.H.; Kim, M.Y. Yield prediction through the event sequence analysis of the die attach process. *IEEE Trans. Semicond. Manuf.* **2015**, *28*, 563–570. [CrossRef]
5. Tsai, T. Development of a soldering quality classifier system using a hybrid data mining approach. *Expert Syst. Appl.* **2012**, *39*, 5727–5738. [CrossRef]
6. Stoyanov, S.; Bailey, C.; Tourloukis, G. Similarity approach for reducing qualification tests of electronic components. *Microelectron. Reliab.* **2016**, *67*, 111–119. [CrossRef]
7. Khader, N.; Yoon, S.W.; Li, D.B. Stencil printing optimization using a hybrid of support vector regression and mixed-integer linear programming. *Procedia Manuf.* **2017**, *11*, 1809–1817. [CrossRef]
8. Tsai, T.; Liukkonen, M. Robust parameter design for the micro-BGA stencil printing process using a fuzzy logic-based Taguchi method. *Appl. Soft. Comput.* **2016**, *48*, 124–136. [CrossRef]
9. Kwak, D.; Kim, K. A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes. *Expert Syst. Appl.* **2012**, *39*, 2590–2596. [CrossRef]
10. Tsai, T. Thermal parameters optimization of a reflow soldering profile in printed circuit board assembly, A comparative study. *Appl. Soft. Comput.* **2012**, *12*, 2601–2613. [CrossRef]
11. Chan, K.Y.; Kwong, C.K.; Tsim, Y.C. Modelling and optimization of fluid dispensing for electronic packaging using neural fuzzy networks and genetic algorithms. *Eng. Appl. Artif. Intell.* **2010**, *23*, 18–26. [CrossRef]
12. Liukkonen, M.; Havia, E.; Leinonenb, H.; Hiltunena, Y. Quality-oriented optimization of wave soldering process by using self-organizing maps. *Appl. Soft. Comput.* **2011**, *11*, 214–220. [CrossRef]
13. Liukkonen, M.; Hiltunen, T.; Havia, E.; Leinonen, H.; Hiltunen, Y. Modeling of soldering quality by using artificial neural networks. *IEEE Trans. Electron. Packag. Manuf.* **2009**, *32*, 89–96. [CrossRef]
14. Srimani, P.K.; Prathiba, V. Adaptive data mining approach for PCB defect detection and classification. *Indian J. Sci. Technol.* **2016**, *9*, 1–9. [CrossRef]
15. Sim, H.; Choi, D.; Kim, C.C. A data mining approach to the causal analysis of product faults in multi-stage PCB manufacturing. *Int. J. Precis. Eng. Manuf.* **2014**, *15*, 1563–1573. [CrossRef]

16. Nagorny, K.; Lima-Monteiro, P.; Barata, J.; Colombo, A.W. Big data analysis in smart manufacturing: A Review. *Int. J. Commun. Netw. Syst. Sci.* **2017**, *10*, 31–58. [CrossRef]
17. Cheng, Y.; Chen, K.; Sun, H.M.; Zhang, Y.P.; Tao, F. Data and knowledge mining with big data towards smart production. *J. Ind. Inform. Integr.* **2018**, *9*, 1–13. [CrossRef]
18. Hashem, S.T.; Ebadati, E.O.M.; Kaur, H. A hybrid conceptual cost estimating model using ANN and GA for power plant projects. *Neural Comput. Appl.* **2017**, *2017*, 1–12. [CrossRef]
19. Tang, L.; Yuan, S.; Tang, Y.; Qiu, Z.P. Optimization of impulse water turbine based on GA-BP neural network arithmetic. *J. Mech. Sci. Technol.* **2019**, *33*, 241–253. [CrossRef]
20. Jiang, S.; Wang, L. Efficient feature selection based on correlation measure between continuous and discrete features. *Inf. Proc. Lett.* **2016**, *116*, 203–215. [CrossRef]
21. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, V.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting novel associations in large data sets. *Science* **2011**, *334*, 1518–1524. [CrossRef] [PubMed]
22. Gregorutti, B.; Michel, B.; Saint-Pierre, P. Correlation and variable importance in random forests. *Stat. Comput.* **2017**, *27*, 659–678. [CrossRef]
23. Hess, A.S.; Hess, J.R. Linear regression and correlation. *Transfusion* **2017**, *57*, 9–11. [CrossRef] [PubMed]
24. Zhang, Z.; Tian, Y.; Bai, L.; Xiahou, J.B.; Hancock, E. High-order covariate interacted lasso for feature selection. *Pattern Recognit. Lett.* **2017**, *87*, 139–146. [CrossRef]
25. Ohishi, M.; Yanagihara, H.; Fujikoshi, Y. A fast algorithm for optimizing ridge parameters in a generalized ridge regression by minimizing a model selection criterion. *J. Stat. Plan. Inference* **2019**. [CrossRef]
26. Ao, Y.; Li, H.Q.; Zhu, L.P.; Ali, S.; Yang, Z.G. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *J. Pet. Sci. Eng.* **2019**, *174*, 776–789. [CrossRef]
27. Tang, J.; Yu, S.W.; Liu, F.; Chen, X.Q. A hierarchical prediction model for lane-changes based on combination of fuzzy C-means and adaptive neural network. *Expert Syst. Appl.* **2019**, *130*, 265–275. [CrossRef]
28. Rezaee, M.J.; Jozmaleki, M.; Valipour, M. Integrating dynamic fuzzy C-means, data envelopment analysis and artificial neural network to online prediction performance of companies in stock exchange. *Phys. A Stat. Mech. Its Appl.* **2018**, *489*, 78–93. [CrossRef]
29. Fathabadi, H. Power distribution network reconfiguration for power loss minimization using novel dynamic fuzzy c-means (dFCM) clustering based ANN approach. *Int. J. Electr. Power* **2016**, *78*, 96–107. [CrossRef]
30. Jia, W.; Zhao, D.; Zheng, Y.; Hou, S.J. A novel optimized GA–Elman neural network algorithm. *Neural Comput. Appl.* **2019**, *31*, 449–459. [CrossRef]
31. Chen, T. Incorporating fuzzy c-means and a back-propagation network ensemble to job completion time prediction in a semiconductor fabrication factory. *Fuzzy Sets Syst.* **2007**, *158*, 2153–2168. [CrossRef]
32. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **2013**, *34*, 483–519. [CrossRef]

applied sciences

MDPI

Article

A Modified Bayesian Network Model to Predict Reorder Level of Printed Circuit Board

Shengping Lv [1], Hoyeol Kim [2], Hong Jin [1,*] and Binbin Zheng [1]

[1] College of Engineering, South China Agricultural University, Guangzhou 510642, China; lvshengping@scau.edu.cn (S.L.); zhengbinbin@stu.scau.edu.cn (B.Z.)
[2] Department of Industrial, Manufacturing and Systems Engineering, Texas Tech University, Lubbock, TX 79409, USA; hoyeol.kim@ttu.edu
* Correspondence: hjin@scau.edu.cn; Tel.: +86-187-1937-3880

Received: 11 May 2018; Accepted: 30 May 2018; Published: 2 June 2018

Featured Application: The research was motivated by the requirement of a printed circuit board (PCB) manufacturer and the application of the work is to identify the repeated PCB orders for batch production according to the predicted reorder level.

Abstract: Identifying the printed circuit board (PCB) orders with high reorder frequency for batch production can facilitate production capacity balance and reduce cost. In this paper, the repeated orders identification problem is transformed to a reorder level prediction problem. A prediction model based on a modified Bayesian network (BN) with Monte Carlo simulations is presented to identify related variables and evaluate their effects on the reorder level. From the historically accumulated data, different characteristic variables are extracted and specified for the model. Normalization and principal component analysis (PCA) are employed to reduce differences and the redundancy of the datasets, respectively. Entropy minimization based binning is presented to discretize model variables and, therefore, reduce input type and capture better prediction performance. Subsequently, conditional mutual information and link strength percentage are combined for the establishment of BN structure to avoid the defect of tree augmented naïve BN that easily misses strong links between nodes and generates redundant weak links. Monte Carlo simulation is conducted to weaken the influence of uncertainty factors. The model's performance is compared to three advanced approaches by using the data from a PCB manufacturer and results demonstrate that the proposed method has high prediction accuracy.

Keywords: printed circuit board; reorder level; principal component analysis; Bayesian network

1. Introduction

A printed circuit board (PCB) is found in practically all electrical and electronic equipment. It is the base of the electronics industry [1]. Due to increased competition and market volatility, demand for highly individualized products promotes a rapid growth of orders with a small batch of purchase and production. Some orders even with the relatively large volume have been placed separately and repeatedly at different times by customers. Dynamic fluctuation of market demands for PCB can easily bring great production imbalance, which is a waste of production capacity during the idle period with fewer orders from customers. However, during the busy period, it results in tardiness among many orders. Multi-batches of the same PCB product produced separately always require higher preparation and production cost with a higher scrap rate. Identifying orders with high reorder frequency and combining different batches of these orders during a reasonable period (e.g., an idle period) as batch

and inventory-oriented production can reduce production cost, benefit production capacity balance, and facilitate on time delivery.

Taking an example from a PCB manufacturer named Guangzhou FastPrint Technology Co., Ltd. (called FastPrint in this paper), few orders were manually selected each month for batch production during the idle period based on reorder frequency and cumulative delivery area in the past 30 months. The reorder frequency is the number of times a customer places the same type of orders to a manufacturer in a given period. The delivery area of each order corresponds to the amount (quantity) of PCB products the customer orders multiplied by the area of each piece of PCB. The cumulative delivery area is the accumulation of the delivery area of the same type of orders in a given period. If 80% of the manually selected orders for batch production can be purchased by customers within six months (i.e., the maximum storage period in the manufacturer's inventory can be ordered by most of customers), then the manufacturer can profit from better utilization of idle resources and the reduction of repeated production preparation and cost. However, the manual selection process is experience-dependent and time-consuming. Meanwhile, the accuracy needs to be improved because only the reorder frequency and the cumulative delivery area are taken into consideration.

The selection of orders for batch production is not based on the accurate reorder frequency within a certain period but always according to the range of predicted reorder frequency (e.g., reorder frequency ≥ 3) in practice. Moreover, it is difficult to accurately determine the reorder frequency within a certain period for each PCB order in advance. Therefore, we transform the repeated orders identification problem into a reorder level prediction problem in which the reorder frequency within six months was divided into four reorder levels (i.e., 1, 2, 3, and 4) corresponding to the reorder frequency (0, 1–2, 3–5, and >5, respectively). On this basis, orders with a highly predicted reorder level corresponding to the range of high reorder frequency placed within six months are taken as candidates for batch production.

The reorder level prediction is similar to the data mining based customer identification (also referred to as customer acquisition) problem as an important task of customer relationship management (CRM) [2]. The former can be conducted by analyzing characteristics of the orders and subdividing them into different groups in which the order groups with higher reorder levels (e.g., 3 and 4) can be taken as candidates for batch production. The latter, on the other hand, is to seek the profitable customer segments by analyzing their underlying characteristics and subdividing an entire customer base into smaller customer segments, which are comprised of customers who are relatively similar within each specific segment [2,3]. Identification of the most profit-generating customers and segmentation of customers are quite vital [3]. Previous studies reveal that recency, frequency, and monetary (RFM) analysis and frequent pattern mining can be successfully used or integrated to discover valuable patterns of customer purchase behavior [3–8]. Dursun and Caber [3] took the RFM analysis for profiling profitable hotel customers and related customers were divided into eight groups. Chen et al. [4] incorporated the RFM concept to define the RFM sequential pattern and developed a modified Apriori for generating all RFM sequential patterns from customers' purchasing data. Hu and Yeh [5] proposed RFM-pattern-tree to compress and store entire transactional database and developed a patterned growth-based algorithm to discover all the RFM-patterns in the RFM-pattern-tree. Coussement et al. [6] employed RFM analysis, logistic regression, and decision trees for the customers' segmentation and identification. Mohammadzadeh et al. [7] employed k-means clustering for identifying target patient customers and then conducted the prediction of customers churn behavior via the RFM model based on the decision tree classifier. Song et al. [8] employed RFM considering parameters with time series to cluster customers and identify target customers.

Other data mining approaches have also been developed and many special factors were considered to excavate the customer pattern purchase behavior. Liu [9] developed a fuzzy text mining approach to categorize textual data to analyze consumer behaviors for the accurate classification of customers. Sarti et al. [10] presented a consumer segmentation method using clustering based on consumers' purchase of sustainability and health-related products. Murray et al. [11] combined

clustering with time series analysis to create customer segments and segment-level forecasts and then applied the forecasts to individual customers. Caigny et al. [12] proposed a logit leaf model for customer churn prediction in which customer segments are identified using decision rules and then a model is created for every segment using logistic regression. Ngai et al. [2] provided a comprehensive review of CRM from four dimensions such as customer identification, customer attraction, customer retention, and customer development. Zerbino et al. [13] presented a review of Big Data-enabled CRM including research on customer evaluation and acquisition. However, the topic discussed in this paper has seldom been studied to the best of our knowledge and it was not involved in the two previously mentioned reviews either.

Customer identification and a reorder level prediction are similar in that both of them aim to develop classified treatment strategy according to historical transactions. Nevertheless, there are differences from the following three aspects. First, the customer identification problem is to develop more accurate sales and advertising strategies based on customer transaction history and, therefore, better for retaining target customers [10] while the final purpose of the problem discussed in this paper is to select orders for batch production with different misclassification risks based on accumulated manufacturing orders. Second, RFM of different purchase products (orders) should be considered for the customer pattern mining while, in this research study, we only consider the parameters of the same product ordered at different times. Third, RFM are the main parameters considered in the customer identification problem while the production scale including quantity, area, and lifecycle of orders should also be considered in this paper. However, the previously mentioned approaches cannot be employed directly for reorder level prediction. Therefore, more influential variables and misclassification loss should be considered and related approaches should be developed.

In this paper, a prediction model based on modified Bayesian network (BN) with Monte Carlo simulations is presented to predict a reorder level of PCB orders. More precisely, we apply BN to excavate the relationship between influential variables (factors) and the reorder level. The main reason for choosing BN is that it has the clearest common sense interpretation and can be viewed as causal models of the underlying domains. It also owns the powerful capability of dealing with uncertainty and causality inference and has been widely used in predicting and classifying problems [14–21]. Figure 1 illustrates the framework of the proposed approach in which all procedures will be discussed in detail except for decision making marked with the dashed boxes.

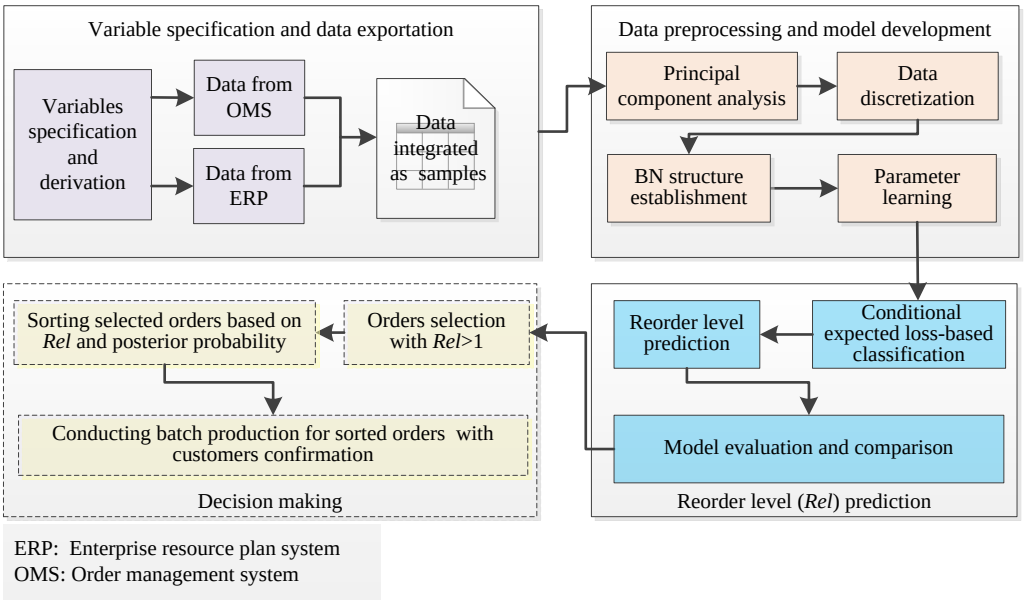


Figure 1. The framework of the reorder level prediction for batch production.

The remainder of this paper is organized as follows. Relevant variables specification and data preprocessing including principal component analysis (PCA)-based factors extraction and entropy minimization-based data discretization are introduced in Section 2. The combination of conditional mutual information (CMI) and link strength percentage (LSP) to avoid the defect of tree augmented naïve (TAN) BN and conditional expected loss for final classification are described in Section 3. The model evaluation and comparison are given in Section 4 in which Monte Carlo simulation is conducted to determine the confidence upper limits of reorder frequency and weaken the influence of uncertainty factors. Additionally, performance of the proposed approach is compared to TAN, AdaBoost, and artificial neural networks (ANN). Conclusions are drawn in Section 5.

2. Variables Specification and Data Preprocessing

2.1. Variables Specification

Reorder level related variables were inherited and derived from fields in enterprise resource plan (ERP) system and order management system (OMS) in which the same type of repeated orders placed at different date are labeled with the same production number but different order numbers generated by the manufacturer's coding rule. On this basis, statistics of delivery area, quantity, transaction money, and interval days of the past 30 months before a set date were derived and the related description is presented in Table 1. The set date is prepared for orders selection and batch production. The statistic excludes the accumulation before 30 months under the consideration of order's lifecycle based on expert experience. The reorder level is the classification objective and four levels are set based on the reorder frequency of a production number in the next six months after a set date.

Table 1. Variable specification.

Variables	Symbols	Description
Layer number	*Ln*	PCB is made of resin, substrate, and copper foil and the Ln is the number of copper foil layers.
Continued days	*Condays*	Interval days between the first order date and a set date.
Recency	*Rec*	A period between the last order date and a set date.
Maximum/minimum/mean of delivery interval days	*Delind_Max/Min/Mean*	Days between order date and required delivery date in the past 30 months.
Maximum/minimum/mean of interval days	*Interval_Max/Min/Mean*	Days between two adjacent order dates in the past 30 months.
Frequency in 30 months	*Fre3*	Reorder frequency within 30 months before a set date.
Frequency	*Fre*	Reorder frequency before a set date.
Maximum/minimum/mean/sum of delivery area (m^2)	*Area_Max/Min/Mean/Sum*	Delivery area of the past 30 months before a set date.
Maximum/minimum/mean/sum of money (CNY)	*Mon_Max/Min/Mean/Sum*	Transaction money of the past 30 months before a set date.
Maximum/minimum/mean/sum of delivery quantity	*Qaun_Max/Min/Mean/Sum*	Delivery quantity of the past 30 months before a set date.
Reorder level	*Rel*	1, 2, 3, and 4 levels corresponding to the reorder frequency 0, 1–2, 3–5, and >5 within six months, respectively.

Note: Statistic parameters of maximum/minimum/mean/sum were derived from the orders with the same production number accumulated in the past 30 months before a set date.

Data from three factories accumulated in ERP and OMS of Fastprint were collected and integrated. Then 33,542 training samples were selected randomly with the set date 31 March 2016 based on the

orders accumulated between 1 October 2013 and 31 March 2016. Meanwhile 14,484 test samples with another set date of 31 May 2016 were selected randomly based on the orders accumulated between 1 December 2013 and 31 May 2016 excluding the records with the same production number in training samples. The observed reorder levels for the records from training and test samples are transformed from the reorder frequency within six months after its set date (i.e., the frequency accumulated between 1 April 2016 and 30 September 2016 for training samples and between 1 June 2016 and 31 October 2016 for test samples). Each record was aggregated based on the orders placed during the past 30 months according to the production number. The records with reorder frequency 1 were not to be considered for batch production and have been deleted. Meanwhile few special orders with odd number layers and layer number greater than 20 have also been excluded because they are seldom taken for batch production in practice.

Sample size and the proportion of different reorder levels are presented in Table 2. It can be seen that sample proportions of different reorder levels are similar to those of the training and test samples, respectively. The statistic results show that only about 5.5% of the records with reorder level ≥ 3 in Table 2 in number were aggregated by a separately placed reorder. However, these small proportions of the records exerted significant influence on resource utilization and balance for the manufacturer in practice.

Table 2. Samples and their proportion of different reorder levels.

Samples	Number (Proportion of Different Reorder Levels %)				Total Number
	1	2	3	4	
All	38,678 (80.54)	6734 (14.02)	1777 (3.70)	837 (1.74)	48,026
Training	26,963 (80.39)	4735 (14.11)	1225 (3.65)	619 (1.85)	33,542
Test	11,715 (80.88)	1999 (13.80)	552 (3.80)	218 (1.52)	14,484

2.2. Principal Component Analysis

There are significant differences in values among variables given in Table 1. Some variables may be redundant or not have a significant influence on the reorder level prediction. Furthermore, continuous-valued variables with a large amount of input types easily generate too many conditional probability tables (CPTs) with sparse samples for each value, which negatively affects the establishment of a robust model. It is, therefore, necessary to perform preprocessing before building a model.

First, in order to eliminate the negative impact caused by the huge difference between each variable in terms of values, there is a need to normalize each variable ranging from 0 to 1. Second, the total data sample matrix 48,026 $\times$ 23 (i.e., the number of the samples multiplied by the number of the input variables for each records) would be considerably complicated and time consuming to model and test for such a high-dimension data samples [22]. It is, therefore, essential for reducing the dimension of the data samples and extracting the typical features from the original data samples. Third, in order to reduce input type and get better performance for variables, it is important to discretize variables for BN model development [14].

PCA is an effective statistical analysis method in multi-dimensional data compression and factors extraction. It can fuse relatively useful features and extract more sensitive factors through the evolution of the variance contribution rate and the cumulative variance contribution rate of each variable [22]. In this study, PCA was used for reducing variable redundancy for the proposed models. This could greatly reduce the modeling time and improve operational efficiency. The procedure of PCA is described below.

PCA was conducted by Algorithm 1 based on training samples according to the 21 input variables given in Table 1 except for the layer number and recency with some initial experiment. Seven factors were extracted with 87.87% of the cumulative variance contribution rate, which means the extracted factors can represent 87.87% of information of the original 21 input variables. The variance contribution

rate and cumulative variance contribution rate are shown in Figure 2. The factor loading matrix with each loading a_{ij} represents how many information factors f_j can explain the variable x_i, which is illustrated in Figure 3. The numbers represent the original variables in Figure 3 and the main variables that each factor explained can be found in Table 3. On this basis, factor values of the test samples were computed based on the weighted sum of the original variables.

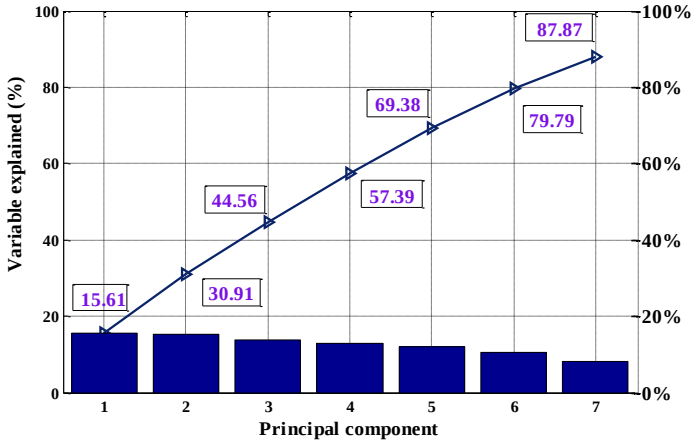


Figure 2. Variance explained based on principal component analysis.

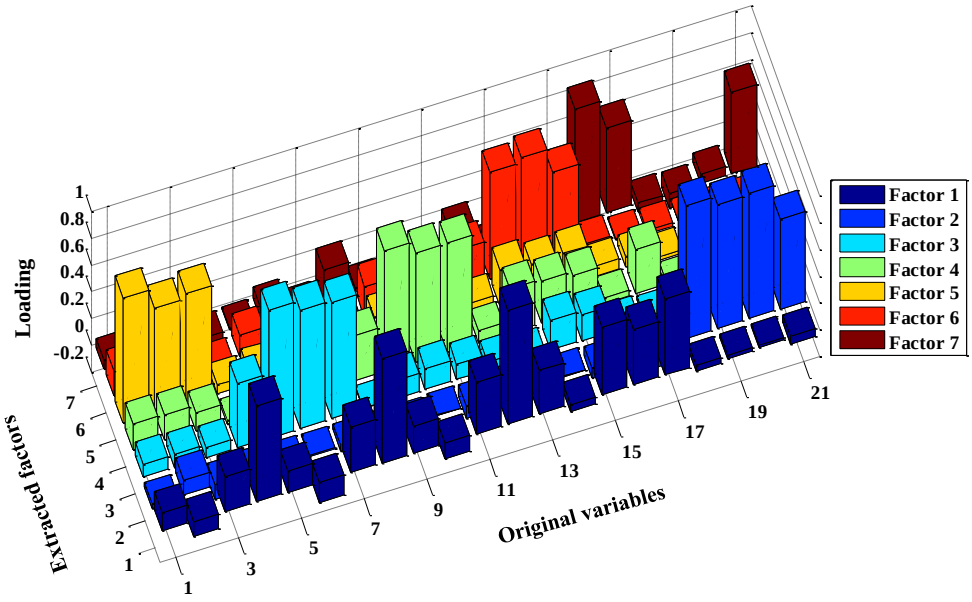


Figure 3. Component matrix.

Table 3. Number related variables and factor interpretation.

No.	Variables Name	Factors	Interpretation
1–3	*Delind_Mean/Min/Max*	1 (*F1*)	*Qaun_Sum/Mon_Sum/Area_Sum*
4–7	*Qaun_Sum/Mean/Min/Max*	2 (*F2*)	*Interval_Mean/Min/Max, Condays*
8–11	*Mon_Sum/Mean/Min/Max*	3 (*F3*)	*Qaun_Mean/Min/Max*
12–15	*Area_Sum/Mean/Min/Max*	4 (*F4*)	*Mon_Mean/Min/Max*
16–17	*Fre/Fre3*	5 (*F5*)	*Delind_Mean/Min/Max*
18–20	*Interval_Mean/Min/Max*	6 (*F6*)	*Area_Mean/Min/Max*
21	*Condays*	7 (*F7*)	*Fre/Fre3/Condays*

Algorithm 1. Factors extraction based on PCA.

(1) **Normalization**: For each column of the data sample x_i, min-max normalization was taken and $x'_i = (x_i - x_{imim})/(x_{i\max} - x_{imim})$ was computed, where x_i is the original data with x_{imim} and $x_{i\max}$ representing the minimum and maximum values in x_i, respectively.

(2) **Principal component analysis**: The calculation of the correlation coefficient matrix and its eigenvalues and eigenvectors was conducted. Subsequently, a variance explained matrix was constituted based on the eigenvectors. All the columns of this matrix were ranked according to the variance contribution rate in descending order. The cumulative variance contribution rate of principal components (factors) was calculated by $\alpha_m = \sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{n} \lambda_i$, where λ is the eigenvalue of each dimension with m representing the top m principal components and α_m is the cumulative variance contribution. In this study, the top m factors with a cumulative variance contribution rate of more than 85% were selected to replace the original n variables.

(3) **Computing scores of factors**. Scores of factors were computed according to the weight sum of original variables in which the weights were obtained based on least squares estimation.

2.3. Data Discretization

The entropy minimization based binning method employed in this paper has been widely applied in discretizing variables [23]. The core measures of entropy minimization based discretization include information entropy and gain [24,25]. Let k classes be $C_1, C_2, \ldots, C_k$ in samples set S and let $P(C_i, S)$ be the proportion of samples in S that has class C_i. The entropy of S is defined by the equation below.

$$Ent(S) = -\sum_{i=1}^{k} P(C_i, S) \log_2 P(C_i, S) \quad (1)$$

where *Ent(S)* measures the amount of information needed to specify the classes in S. The greater the *Ent(S)* value is, the more information it contains and the lesser purity it has. A binned interval with all values belonging to the same class has the highest purity [26,27].

Entropy of samples S partitioned by an arbitrary split point T of attribute X into two disjoint intervals is defined by the equation below.

$$Ent(X, T; S) == \sum_{j=1}^{2} \frac{|S_j|}{|S|} Ent(S_j) \quad (2)$$

where $|S_j|$ and $|S|$ are the sample size of subset S_j and S, respectively. The information gain for a variable X based on a given split point T can be defined by Equation (3).

$$Gains(X, T; S) = Ent(S) - Ent(X, T; S) \quad (3)$$

A partition induced by a split point T for a set S is accepted according to the minimum description length principle (MDLP) [24]. The binning algorithm for the discretization of each variable (i.e., *F*1–F7, layer number and recency) is described below.

The maximum number of the binned intervals was set to 10 and the discretization results of the variables obtained by Algorithm 2 are given in Table 4. Split points were used directly for the discretization of the test samples. Proportions of the different reorder levels for the training samples in the different binned intervals of the variables are illustrated in Figure 4.

It can be seen that proportions of the different reorder levels in the different binned intervals vary significantly, which indicates that the cumulative delivery quantity/area (*F1* and *F3*), delivery interval day (*F5*), and continued days (*F2*) are also important for classification of reorder level besides RFM (i.e. *Rec*, *F7*, and *F4*). Proportion of reorder level 2, 3, and 4 decreases with an increase in the value of discretized recency and the reorder level of the samples with the binned interval 10 for recency can

almost directly be classified as 1. *F7* performs the opposite tendency compared to recency in which the sample with the binned value 1 or 2 has high purity (probability) to be classified as 1. A sample with the binned recency as 1, *F7* as 6, *F2* or *F3* as 1, or *F7* as 9 has high probability to be classified as 4.

Table 4. Discretized results of variables.

Variables (Factors)	Spilt Points
Ln	4, 10
Rec	11, 36, 67, 114, 172, 195, 325, 440, 714
F1	−0.09, 0.002, 0.211, 1.221, 2.436
F2	−0.767, −0.634, −0.496, 0.026, 0.787, 1718
F3	−0.532, −0.311, −0.251, −0.173, −0.024, 0.534
F4	−0.802, −0.481, −0.09, 0.872
F5	−0.1097, 0.443
F6	−0.879, −0.586, −0.329, 0.009, 1.432
F7	−1.009, 0.628, 0.386, 0.023, 0.401,1.335, 1.865, 3.501

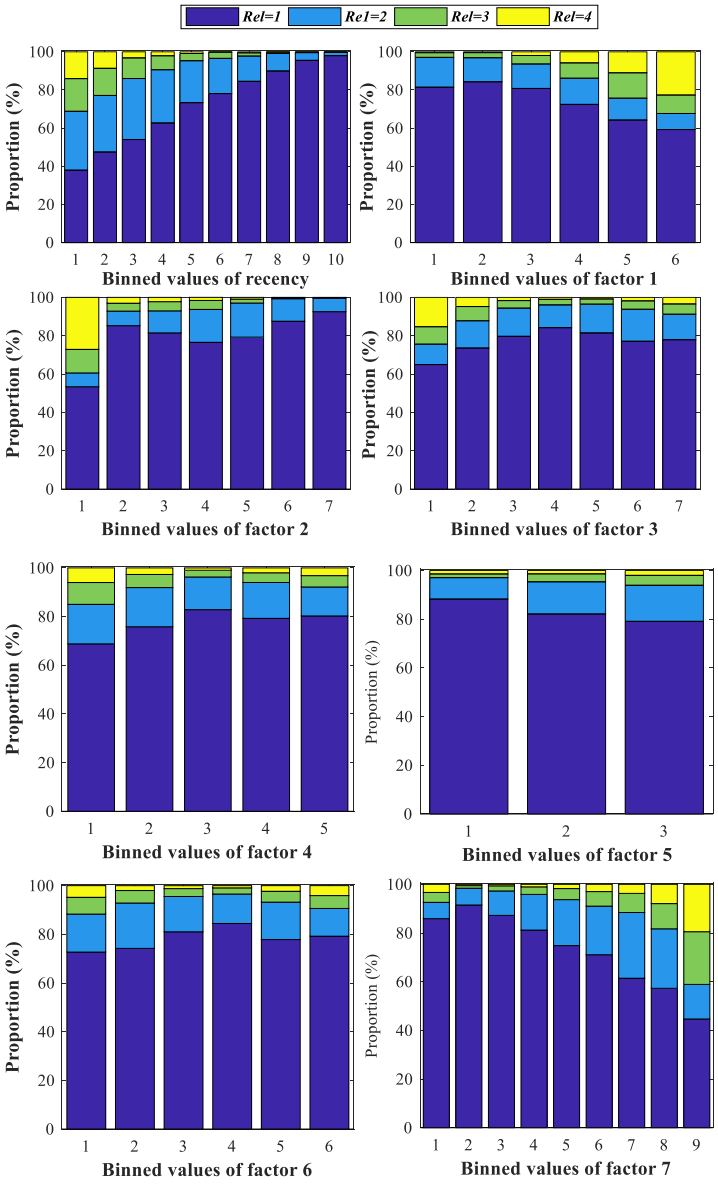


Figure 4. Proportions of different reorder level in different binned intervals.

Algorithm 2. Entropy-based data discretization.

Input:
Samples S, samples size N, variable x, classes of reorder level Re$l = \{1,2,3,4\}$, and maximum number of binned interval *MaxIntv*.
Output:
Split points for the variable $SP_x = \{sp_1, sp_2, \ldots, sp_{s-1}\}$, the number of final split point s, and binned values for variable $B_x = \{1,2,\ldots,s\}$
Iteration
Initialize empty set SP_x and B_x;
Rank sample S according to x in ascending order and the position $[1,\ldots,j,\ldots,N]$ are taken as its possible split points;
for k = 1 to *MaxIntv*;
Compute $E(S)$ according to Equation (1); *tempN* = N; *tempj* = *0*;
for j = 2 to N;
sp = 1; *tempj* = j
Get the value of x at the position j for (updated) sample S and suppose it is T_j;
$|S_j^1|$, $|S_j^2|$←the number of samples in the two intervals S_j^1 and S_j^2 separated by T_j;
$|S_{j1}^l|$, $|S_{j2}^l|$, $|S_{j3}^l|$, $|S_{j4}^l|$←the number of samples with Re$l = \{1,2,3,4\}$ in S_j^l, $l = 1,2$, respectively;
$P(\text{Rel}_k, S_j^l) = |S_{jk}^l|/|S_j^l|, 1 \le k \le 4, l = 1,2$;
Compute $E(S_j^l)$, $l = 1,2$, $Ent(x, T_j; S)$, and $Gains(x, T_j; S)$ according to Equations (1)–(3) respectively;
if $Gains(x, T_j; S)$ satisfies MDLP;
Append T_j to SP_x and sp to B_x; sp ++; update $S \leftarrow S_j^2$, $N = |S_j^2|$; break;
end;
end;
if $tempj = tempN$; break; end;
end;
return SP_x and B_x.

3. Modified Bayesian Network Model Development

3.1. Bayesian Network

Bayesian network (also known as belief network and causal network) is a probabilistic graphical model that represents a set of random variables and their conditional dependence by means of a directed acyclic graph (DAG) and CPTs [19,20]. Each node in DAG represents a variable of the ranges over a discrete set of domain and contacts with its parent's nodes [14] and directed arcs represent the condition or probability dependency between random variables [14,28]. BN has become a popular knowledge-based representational scheme in data mining [27–30]. This graphical structure, which expresses causal interactions and direct/indirect relations as probabilistic networks, has secured BN's popularity. Experts can easily understand such structures and (if necessary) modify them to improve the model [28].

The critical problem in establishing a BN is to determine the network structure S and corresponding set of parameters θ [28,29], which are always called structure learning and parameter learning, respectively. In order to reduce arcs between nodes (variables) with weak causal interactions and corresponding CPTs, CMI and LSP were combined with expert's experience to establish BN structure and, therefore, avoid the defects of TAN. The CMI was first introduced in TAN [31] by relaxing the conditional independence assumption of naïve Bayesian for the purpose of selecting particular dependences [15]. However, TAN links all the input variables (evidence node) to output variables (class node) and allows at most two parents nodes with one connection to the class node and one causal connection to another evidence node, which easily misses some strong links and sometimes generates redundant weak strength links [28] that are negative for the robustness and generalization of the BN model.

Suppose a set of discretized random variables is $X = \{x_1, x_2, \ldots, x_9\}$ corresponding to the variables such as a layer number, recency, and *F1–F7* and CMI between x_i and x_j can be computed below.

$$\mathrm{CMI}(x_i; x_j|\mathrm{Re}l) = \sum_{m,n,k} P(x_i^m, x_j^n, \mathrm{Re}l_k) \ln \frac{P(x_i^m, x_j^n|\mathrm{Re}l_k)}{P(x_i^m|\mathrm{Re}l_k)P(x_j^n|\mathrm{Re}l_k)}, i \neq j \quad (4)$$

where x_i^m, x_j^n, and$\mathrm{Re}l_k$ represent the mth, nth, and kth values of x_i, x_j, and Rel, respectively. $\mathrm{CMI}(x_i; x_j|\mathrm{Re}l)$ measures the information x_j provides on x_i when the value of reorder level Rel is known. The smaller the $\mathrm{CMI}(x_i; x_j|\mathrm{Re}l)$ value is, the weaker the connection between x_i and x_j is.

The LSP from parent node x to child node y is defined by the equation below.

$$\mathrm{LSP}(x \to y) = 100 \times \frac{Ent(y|Z) - Ent(y|x, Z)}{Ent(y|Z)} \quad (5)$$

where $Z = Pa_y/\{x\}$ denotes a set of all parents of y other than x, $Ent(y|Z) = \sum_z P(z) \sum_{x_2} P(y|z) \log_2(1/P(y|z))$, and $Ent(y|x, Z) = \sum_{x,z} P(x, z) \sum_{x_2} P(y|x, z) \log_2(1/P(y|x, z))$. LSP can be interpreted by how much the uncertainty (in percentage) in class variable is reduced by knowing the state of an input variable if the states of all other parent variables are known. LSP plays an important role to evaluate the quality of the BN structure and can facilitate experts to modify arrows based on link strength values [32]. The structure learning algorithm is depicted below (Algorithm 3).

Algorithm 3. Modified BN structure establishment.

(1) Compute the $\mathrm{CMI}(x_i; x_j|\mathrm{Re}l)$ between x_i and x_j according to Equation (4);
(2) Select input variables $x_{k1}, \ldots, x_{kt}$ with $\mathrm{CMI}(x_i; x_j|\mathrm{Re}l)$ being greater than a threshold, and manually link x_i to $x_{k1}, \ldots, x_{kt}$ with directed arcs if there are no arcs between the two nodes;
(3) Combining CMI by expert experience to determine the variables that links to Rel with directed arcs;
(4) Compute LSP according to Equation (5) for each link to evaluate the quality of BN structure and modified (deleted) arrows with small LSP, e.g., 10%.

The Bayesian estimation method was employed for parameter learning in this paper to estimate $\theta = \max p(\theta|X)$ based on the training samples. Initially, θ was treated as a random variable and prior knowledge of θ is expressed as a prior probability distribution $p(\theta)$. Furthermore, there is a likelihood that the function was utilized based on samples. Subsequently, the Bayesian formula was taken to determine the posterior probability distribution of θ. Dirichlet distribution was employed as the prior probability distribution of $p(\theta)$ [16].

CMI between x_i and x_j for the training samples based on Equation (4) is given in Table 5. The threshold was set as 10% and CMI equal to or greater than 10% was reserved to construct the link. It can be seen that (1) CMI is small between layer number, recency, and *F1–F7*, which can be taken as independent variables while constructing BN structure. (2) CMI is large between *F3*, *F4*, *F6*, and *F7*, which means that the cumulative delivery scale (*F1*) of repeated orders is not independent of mean/min/max statistic results of delivery quantity, area, transaction money (*F3*, *F6*, and *F4*), and frequency (*F7*). (3) Similarly, *F6* (mean/min/max delivery area) has great mutual information between *F3* (mean/min/max delivery quantity) and *F4* (mean/min/max transaction money).

On this basis, the structure of modified BN with entropy in each node and LSP for each link was constructed according to Algorithm 3, which was given in Figure 5. Entropy in each node reflects the purity of the node and they were computed based on $Ent(x) = -\sum_{x_i} P(x_i) \log_2 P(x_i)$ where x_i is the discretized value set of node x, which indicates how much uncertainty is in x if no evidence is given for any other nodes. LSP was computed based on Equation (5) and it can be seen that the LSP for links from *Ln*, *Rec*, *F1*, *F2*, *F5*, and *F7* to *Rel* are large, which means that these variables can help reduce

the high percentage of uncertainty for *Rel* when knowing the state of these nodes. Similarly, LSP of links from *F2*, *F3*, *F6*, and *F7* to *F1* are large, which indicates that *F2*, *F3*, *F6*, and *F7* have a close causal relationship with *F1*. However, weak LSP of links from *F7* to *F2* and *F4* to *F6* marked with dotted lines are less than 10% and, therefore, the directed arcs from *F7* to *F2* and *F4* to *F6* were deleted accordingly. Then parameter learning was conducted based on the training samples and the structure to determine the CPTs for each node.

Table 5. Conditional mutual information matrix.

Variables and Factors	*Ln*	*Rec*	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*
Ln	1	0.0018	0.002	0.0023	0.0703	0.0729	0.0307	0.0934	0.0023
Rec		1	0.01	0.019	0.005	0.041	0.0047	0.0108	0.0277
F1			1	0.0681	0.1004	0.141	0.0411	0.1098	0.1108
F2				1	0.0357	0.0125	0.0121	0.0226	0.1445
F3					1	0.099	0.041	0.2215	0.087
F4						1	0.0461	0.1121	0.0192
F5							1	0.0891	0.0144
F6								1	0.027
F7									1

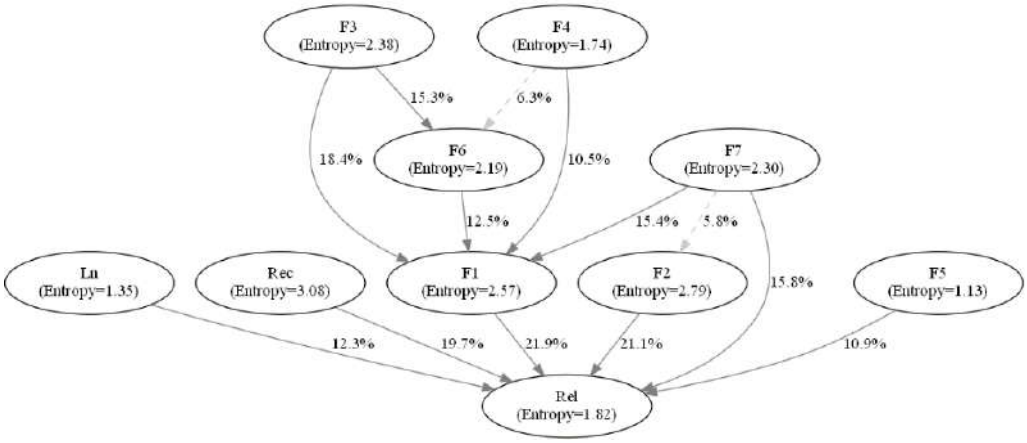


Figure 5. Bayesian network structure with entropy information and link strength percentage.

3.2. Conditional Expected Loss-Based Classification

Classification can be conducted based on learned BN structure and joint probability (the product of all the conditional probabilities of the network). Posterior probability of the reorder level can be calculated according to the Bayesian equation. Lastly, each sample can be predicted to the reorder level corresponding to the greatest posterior probability. However, a sample with a reorder level 1 misclassified as 2, 3, or 4 will bring economic risk if it is taken for batch production. At the same time, posterior probabilities of the biased reorder level may bring a different misclassification. Posterior probabilities of the training samples with observed reorder level 2, 3, or 4 based on the modified BN are given in Figure 6 according to initial experiments in which the posterior probabilities is generated by the formula below.

$$P(Pr\_\mathrm{Rel}_i|Ob\_\mathrm{Rel} = 2,3,4) = \frac{P(Pr\_\mathrm{Rel}_i)P(Ob\_\mathrm{Rel} = 2,3,4|Pr\_\mathrm{Rel}_i)}{\sum_{j=1}^{4} P(Pr\_\mathrm{Rel}_j)P(Ob\_\mathrm{Rel} = 2,3,4|Pr\_\mathrm{Rel}_j)} \quad (6)$$

where $P(Pr\_\mathrm{Rel}_i)$ is the probability of predicted reorder level i ($i = 1,2,3,4$), $P(Ob\_\mathrm{Rel} = 2,3,4)$ is the probability of observed reorder level with the value of 2, 3, or 4, $P(Ob\_\mathrm{Rel} = 2,3,4|Pr\_\mathrm{Rel}_i)$ is the posterior probabilities of observed reorder level with the value of 2, 3, or 4 on the condition of the predicted reorder level i and $P(Pr\_\mathrm{Rel}_i|Ob\_\mathrm{Rel} = 2,3,4)$ is the posterior probabilities of predicted

reorder level on the condition of observed reorder level as 2, 3, or 4. It can be seen that samples to be predicted as 3 and 4 are small and the corresponding posterior probabilities subjects to serious left skewed distribution with a mean value less than 0.25. In contrast, samples to be predicted as 1 or 2 are subject to right skewed distribution with a mean value greater than 0.5. This indicates that the posterior probability-based classification has high posterior probability to predict the reorder level of 1 with an observed value equal to or greater than 2 in many cases. Only a few instances with observed *Rel* = 3 or 4 have been predicted as 3 or 4.

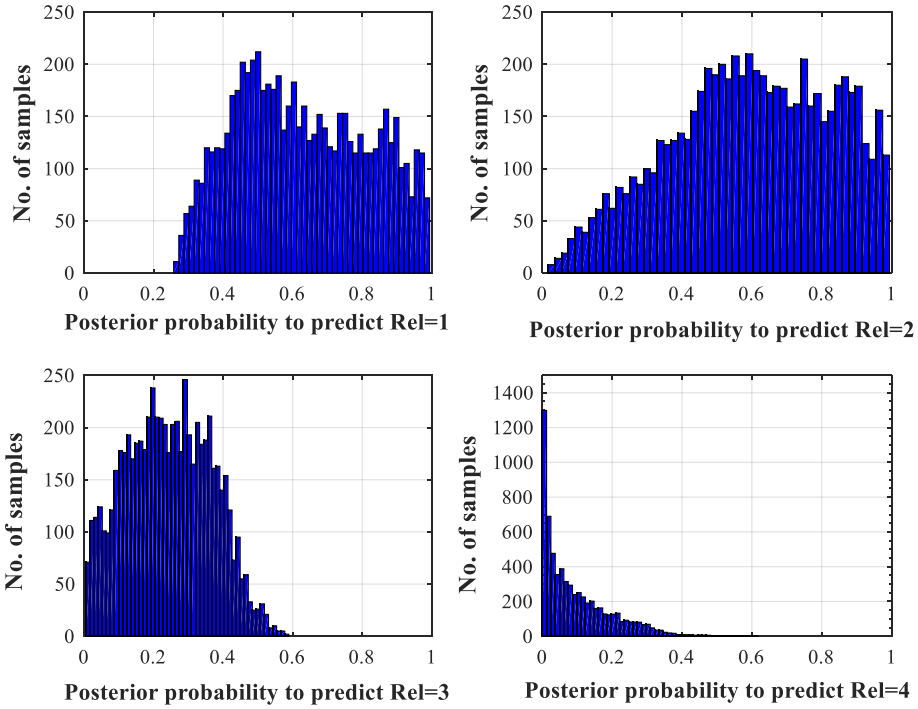


Figure 6. Posterior probabilities corresponding to different predicted reorder levels.

Figure 7 illustrates the posterior probabilities of 100 randomly selected samples obtained by Equation (6) with observed *Rel* = 2 and *Rel* = 4 in which the probability of 1, 2, 3, and 4 corresponds to a predicted reorder level 1, 2, 3, and 4, respectively. Posterior probability in Figure 7a illustrates that it is easy to predict the reorder level of 2 to 1. Figure 7b illustrates that many posterior probabilities corresponding to the predicted reorder level 4 have no significant possibility for classifying it as 4. Therefore, the conditional expected loss was introduced instead of a posterior probability for the final classification decision. Let α_i be the decision to classify sample X as α_i, $\lambda_{ij} = \lambda(\alpha_i, \omega_j)$represents the loss (risk) to classify X with observed value ω_j to α_i, and all the $\lambda_{ij} = \lambda(\alpha_i, \omega_j)$ $i, j = 1, 2, 3, 4$, consist of classification loss matrix. Conditional expected loss is defined to illustrate the expected risk for a decision to predict X as α_i.

$$R(\alpha_i|X) = E\left[\lambda(\alpha_i, \omega_j)\right] = \sum_{j=1}^{4} \lambda(\alpha_i, \omega_j) P(\omega_j|X), i = 1, 2, 3, 4. \quad (7)$$

The final decision can be conducted based on the minimization of the conditional expected loss.

$$R(\alpha_k|X) = \min_{i=1,2,3,4} R(\alpha_i|X) \quad (8)$$

The conditional expected loss-based classification can be described below (Algorithm 4).

Algorithm 4. Conditional expected loss-based classification.

(1) Compute the posterior probability $P\left(\omega_j|X\right) = P(\omega_j)P(X|\ \omega_j) / \sum_{j=1}^{4} P(\omega_j)P(X|\ \omega_j)$;
(2) Compute $R(\alpha_i|X)$ for the classification of α_i according to Equation (7);
(3) Classify reorder level of X to i with minimal $R(\alpha_i|X)$, $i = 1, 2, 3, 4$ by Equation (8).

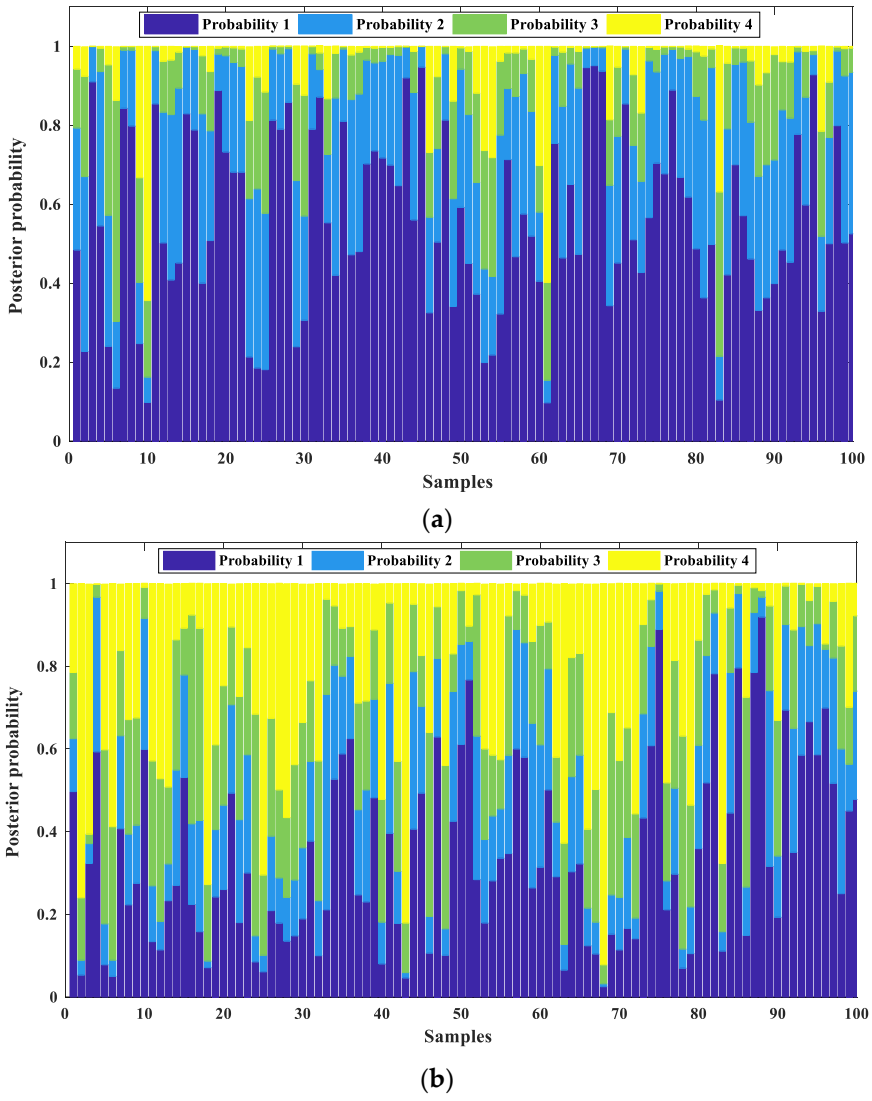


Figure 7. Posterior probabilities for 100 randomly selected samples. (**a**) Posterior probabilities for 100 randomly selected samples with observed Re*l* = 2. (**b**) Posterior probabilities for 100 randomly selected samples with observed Re*l* = 4.

Initial results also show that the probability to predict order with an observed reorder level of 1 to 2, 3, and 4 decreases with an increase in the value of binned recency and the risk to predict the larger reorder level to the smaller one will also decrease. Therefore, four loss matrices corresponding to the binned recency 1, 2–3, 4–5, and 6–10 were introduced for final classification based on Algorithm 4, in which the value in the upper half of the matrix decreases with an increase in the value of binned recency while the value in the lower half of the matrix increases with an increase in the value of binned recency. The values were set to 1 and 0 in the non-diagonal positions and diagonal position, respectively, when recency is 1. The other three matrices are shown in Table 6.

Table 6. Loss matrix for different binned recency values.

Predicted Value	Overserved Value											
	Recency = 2–3				Recency = 4–5				Recency = 6–10			
	1	2	3	4	1	2	3	4	1	2	3	4
1	0	1.2	1.25	1.3	0	1.15	1.2	1.25	0	1.1	1.15	1.2
2	1.3	0	0.8	0.9	1.35	0	0.75	0.85	1.4	0	0.7	0.8
3	1.35	0.75	0	0.9	1.4	0.8	0	0.85	1.45	0.85	0	0.8
4	1.45	0.85	0.75	0	1.5	0.9	0.8	0	1.55	0.95	0.85	0

4. Model Evaluation

4.1. Estimation of Reorder Frequency

In order to get the expected reorder frequency for a given order, we use the sum of the mean value of reorder frequency in each level weighted by conditional probability as the expected reorder frequency. The expected output of the model can be computed by the equation below.

$$E(\text{Re } Freq|Cluster = i) = \sum_{k=1}^{4} P(\text{Re}l = k|Cluster = i)M(\text{Re}l = k), i = 1, 2, \ldots. \quad (9)$$

where $M(\text{Re}l = k)$ is the mean value of reorder frequency within six months for the samples with $\text{Re}l = k$, which can be referred to Table 7. $P(\text{Re}l = k|Cluster = i)$ is the average conditional probability determined by the modified BN with $\text{Re}l = k$ given a specific cluster i and $Cluster = i$ represents the ith cluster of the samples determined by the clustering algorithm.

Table 7. Mean reorder frequency for different reorder levels.

Reorder Level	1	2	3	4
Mean value for training samples	0	1.28	3.67	10.03
Mean value for test samples	0	1.27	3.67	10.28
Mean value for all samples	0	1.28	3.67	10.09

The purpose of the clustering is to classify samples according to their similarity by considering the input features of discretized *F1–F7*, *Rec*, and *Ln*. The *k*-summary approach that can handle both categorical and numerical data was adopted for the clustering and the number of clusters was set to 7 based on an initial experiment. On this basis, the average conditional probability of different reorder levels given different clusters is presented in Table 8.

Table 8. Average conditional probability of different reorder levels given different clusters (%).

Reorder Level	Different Clusters						
	C1	C2	C3	C4	C5	C6	C7
1	89.67	91.88	65.87	58.87	89.60	42.60	92.49
2	9.42	7.32	24.84	31.50	7.03	23.25	6.67
3	0.78	0.64	7.06	7.99	2.18	17.79	0.65
4	0.13	0.16	2.23	1.64	1.18	16.36	0.19

4.2. Evaluation Indicators

The confusion matrix was taken to visualize the performance of different approaches in which each column of the matrix represents the instances in an actual class while each row represents the instances in a predicted class. All correct predictions are located in the diagonal of each table and errors can be visually inspected by values outside the diagonal. Related terminology and derivations are defined in Table 9 [33].

Table 9. Terminology and derivations of the confusion matrix.

Terminology	Description
True positive (TP)	Number of correctly predicted instances for each column
False negative (FN)	Number of incorrectly predicted instances for each column
False positive (FP)	Number of incorrectly predicted instances for each row
True negative (TN)	Number of correctly predicted instances for each row
Sensitivity or true positive rate (TPR)	TP/(TP+FN)
False negative rate (FNR)	1-TPR
Specificity or true negative rate (TNR)	TN/(TN+FP)
False positive rate (FPR)	1-TNR
Positive predictive value (PPV)	TP/(TP+FP)
False discovery rate (FDR)	1-PPV
Accuracy (ACC)	(TP+TN)/Total instances

In order to evaluate the performance of the proposed model, the following mean squared error (*MSE*), mean absolute error (*MAE*), and mean absolute percentage error (*MAPE*) evaluation indicators were used. *MSE* is the average of square sums between the predicted reorder level α_i and the observed value ω_i [14]. It defines the goodness of fit of the models and is given by the following equation.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\alpha_i - \omega_i)^2 \tag{10}$$

The *MAE* is the average of the sum of the absolute difference between observed values and the predicted reorder level, which can be expressed below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(\alpha_i - \omega_i)| \tag{11}$$

The *MAPE* is the average of the sum of the normalized absolute difference between observed values and estimated values. The formula is written below.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\omega_i - \alpha_i}{\omega_i}\right| \times 100 \tag{12}$$

4.3. Experimental Results

Reorder frequency of the samples six months after a set date can be taken as a random event for a manufacturer without prior knowledge and the models are expected to have errors in prediction. It is necessary for finding an upper limit and a lower limit to make as many as possible observed values lie in. Additionally, the small field data set may cause uncertain deviations between observed reorder levels and predicted ones. It is also impractical to fit the reorder frequency in a specific distribution absolutely. The counting nature of the reorder frequency makes it intuitive for using a Poisson distribution as the probability distribution function (PDF). Therefore, assuming the reorder frequency in each reorder level following a Poisson distribution is justifiable since it has a high repetition occurrence rate and more numbers under low reorder levels and a small probability at high reorder levels, which is shown in Table 2. The approximate upper limit for 95% confidence aligns with the Poisson cumulative distribution function (CDF) and is equal to or greater than 95%. The lower limits are considered to be zero because of the nonnegative counting property of Poisson distribution [14].

Monte Carlo simulation is used to weaken the influence of uncertainty factors in this research. It is an effective method for quantifying the variance resulting from the random nature of repetition events. For any sample (orders with the same production number) with a given cluster and the distribution of the reorder frequency, a random number can be generated through simulation to present the reorder frequency for this sample and the 95% confidence upper limit can be obtained. As a result, it is used to estimate reorder frequency and the 95% upper limit for each sample. Along with the increase in the

number of simulations, the prediction accuracy will increase gradually. Therefore, it can quantify the variance resulting from the randomness of repetition events. A procedure of Monte Carlo simulation is described below (Algorithm 5).

Algorithm 5. Monte Carlo simulation of reorder times.

(1) Given a specific order, determine the cluster of the sample;
(2) Determine the expected reorder frequency (within six months after a set date) as the parameter (lamda) of Poisson PDF and CDF for the sample according to Equation (9) based on Tables 7 and 8;
(3) Generate n (10,000 here) random number by Poisson PDF with lamda as its expectation and take the average result of the n random number as the simulated reorder frequency;
(4) Determine the least integer for Poisson CDF being greater than 0.95 as the 95% upper limit of reorder frequency for the order.

A total of 250 randomly selected samples with Monte Carlo simulation results obtained by Algorithm 5 are presented in Figure 8. The performance of the Bayesian network for training and test data can be seen from Figure 8a,b, respectively. When the reorder level is low, the estimated value is close to the observed one. In some extreme situations, it can cause a greater reorder frequency than what can be predicted. The presentation in figures is that the observed values are higher than the 95% upper limits. Figure 8 indicates that the difference between estimated values and actual values is small in most cases and almost all of them are less than 1.

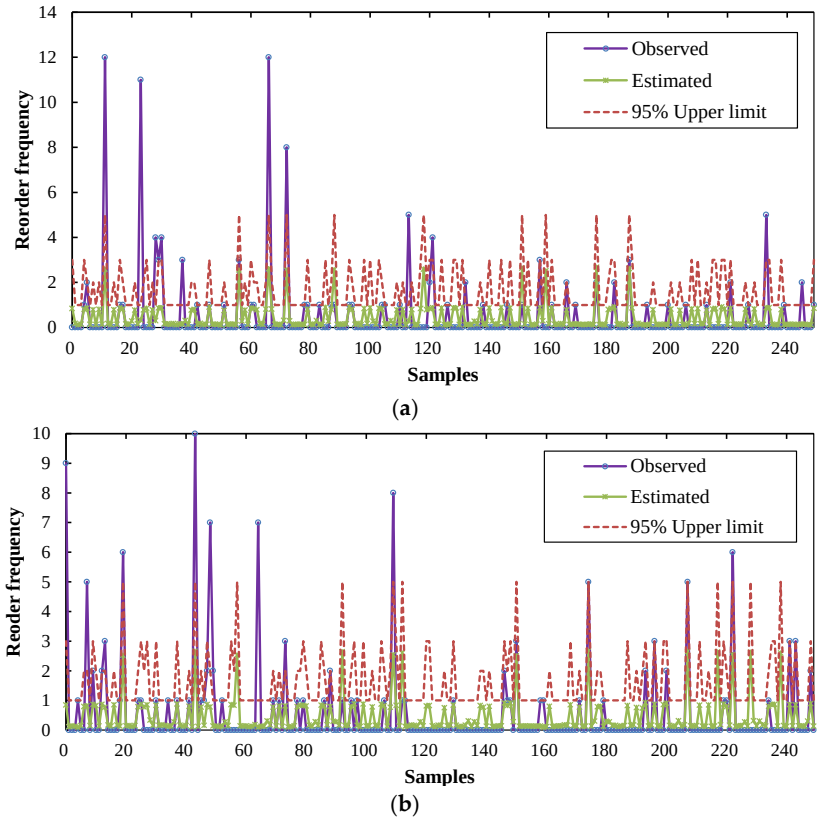


Figure 8. Observed and estimated reorder frequency with 95% upper limits by Monte Carlo simulations. (**a**) Training observed and estimated reorder frequency with 95% upper limits. (**b**) Test observed and estimated reorder frequency with 95% upper limits.

In order to verify the proposed ensemble approach in this research, the data preprocessing and modified BN prediction model were implemented and the performance was compared to other

classifiers including TAN, AdaBoost, and ANN. Among the four competing methods, TAN as a naïve Bayesian network has been widely utilized in classification [17,18]. AdaBoost is an ensemble method whose output is the weighted average of many weak classifiers and is the best-known and most widely applied boosting algorithm in both research and practice [34]. ANN has a strong learning ability and has also been widely used for prediction and classification [35–37]. TAN and ANN were implemented in the IBM SPSS Modeler and AdaBoost was developed by Matlab while the proposed modified BN was developed based on Matlab and package FullBNT.

Confusion matrices of different approaches are given in Figure 9. These confusion matrices show that TAN achieved the highest sensitivity for observed reorder levels 2, 3, and 4. However, the ability to identify a reorder level 1 was weak and the unbalanced and biased distribution of reorder level 1 can greatly increase the production risk if a large amount of orders were taken as batch production in advance without customers' confirmation. On the other hand, the modified BN achieved better results with sensitivity 98.5% and 98.1% for training and test samples, respectively. It can reduce the number of samples with a reorder level 1 to be incorrectly predicted as 2, 3, or 4, which reduces the risk of batch production. In addition, the modified BN can correctly identify a higher amount of orders with an observed reorder level 2 and 3 compared with AdaBoost and ANN both for training and test samples. The sensitivity of the modified BN for observed reorder level 4 deteriorated for the test sample. However, many samples have been predicted as 2 or 3, which can also be taken for batch production. Overall indicators of confusion matrices also show that the modified BN obtained the highest accuracy (81.9%) both for training and test samples.

The approaches were compared both for training and test samples according to indicators presented in Equations (10)–(12) and the comparison results can be referred to Table 10. It shows that the proposed modified BN obtained the lowest *MAE* and *MAPE* for training samples and the lowest *MSE* and *MAE* for test samples. The results in Figure 9 also illustrate that TAN obtained the maximum correctly classified instances with observed reorder levels of 2, 3, and 4 as well as the maximum incorrectly classified instances with an observed reorder level of 1 compared to the other classifiers. Therefore, the indicators show that TAN achieved the lowest *MSE* for training samples as well as almost the largest *MAE* and *MAPE* both for training and test samples. This may be caused by redundant links between the evidence node and the missing strong links such as the causal relationships between *F2*, *F3*, *F6*, and *F1*. Yet, it is worth noting that TAN deteriorated greatly on the test dataset according to *MAE* and *MAPE*, which indicates that the TAN considered in the current study lacked robustness and generalization ability. The ANN exhibited steady performance both for training and test samples but had no superiority according to the three indicators. The AdaBoost achieved slightly better performance for test samples for the indicator *MAPE* but the worst performance for the indicators *MSE* and *MAE*.

Table 10. Comparison of classifiers according to different indicators.

Classifiers	Training Samples			Test Samples		
	MSE	***MAE***	***MAPE***	***MSE***	***MAE***	***MAPE***
Modified BN	689.98	0.2306	10.6965	291.74	0.2239	11.0818
TAN	666.8964	0.2421	11.5641	302.2279	0.2509	11.9724
AdaBoost	738.0686	0.2407	11.2314	308.0735	0.2358	11.0472
ANN	692.7819	0.2349	11.3129	293.1946	0.2346	11.1991

The above results indicate that the modified BN combing CMI, LSP, and expert experience maintains the DAG requirement of BN and produces a more nuanced network that captures the main dependency relationships among evidence nodes (variables) while deleting some weak dependency relationships without allowing arbitrary graphical structures that would make it harder to interpret and extract relations to enhance the prediction model. At the same time, the conditional expected loss can benefit the final classification and it can exhibit better performance especially when compared to TAN. In addition, the modified BN has the clearest common sense interpretation.

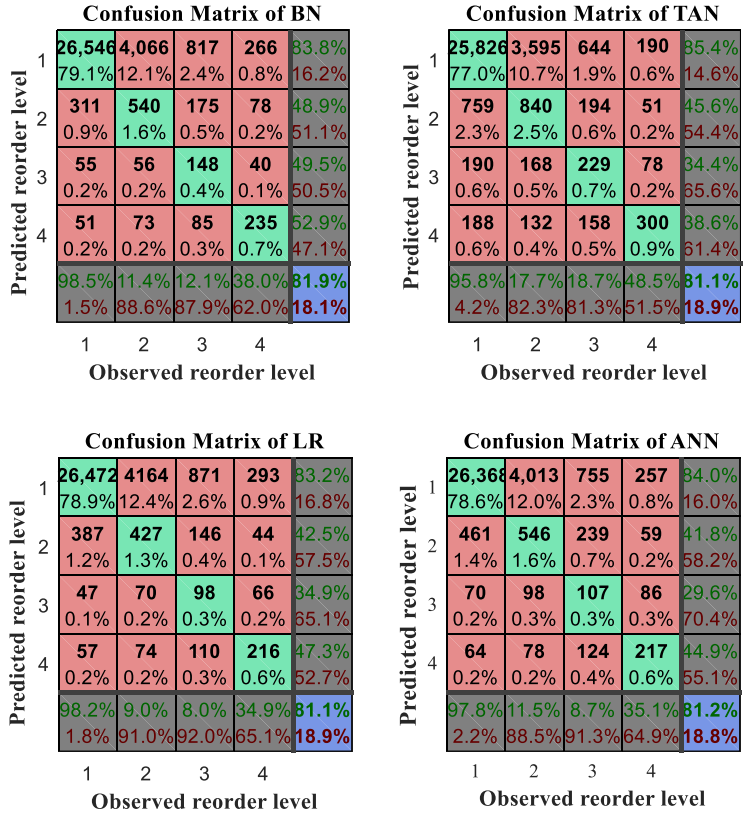


Figure 9. Confusion matrix of different approaches. (**a**) Confusion matrix of different methods for training samples. (**b**) Confusion matrix of different methods for test samples.

5. Conclusions

In this paper, the identification of repeated PCB orders for batch production was transformed into a reorder level prediction problem and a modified Bayesian network model with Monte Carlo simulations to study the relationship between different characteristic variables and reorder levels of PCB within six months was established. Reorder frequency was divided into four reorder levels and variables related to a reorder level were specified. Field data was exported and integrated from a PCB manufacturer with 33,542 training samples and 14,484 test samples. Normalization and PCA were employed to reduce differences and redundancy of the datasets, respectively. PCA results indicated that the causes of the reorder level are closely related to seven principal components and the other two variables, i.e., recency and layer number. Entropy minimization based binning method was employed to discretize model variables for the purpose of reducing input type and capturing better performance and results. The modified structure of BN was established by deleting redundant connections between nodes (with weak link strength) and corresponding conditional probability tables based on conditional mutual information and link strength percentage combining with expert experience. This can facilitate the manufacturer to comprehend causal interactions between variables. On this basis, the conditional expected loss was presented for final classification considering different misclassification risk.

Monte Carlo simulation was conducted to enable the determination with greater accuracy of a mean and confidence interval for reorder frequency estimations based on the predicted reorder level. The upper limits of reorder frequency are particularly useful for the PCB manufacturer as a basis of each reorder level. The performance of the proposed modified BN was visualized by confusion matrix, evaluated by three indicators, and compared to three advanced methods including TAN, AdaBoost, and ANN. It was found that the modified BN prediction model achieved steady and satisfactory results both for training and test samples with the clearest common sense interpretation. Therefore, the proposed model in this paper is an effective approach to capture the repetition pattern of PCB orders that have seldom been studied before. The established explicit relationship between the variables including extracted factors and the reorder level by the causal network can directly facilitate order selection for batch production that can be conducted according to the decision making step given in Figure 1.

The main contributions of this work are summarized below.

1. The tricky problem of identifying repeated orders for batch production was transformed into a reorder level prediction problem and then a reorder level prediction model based on modified causal Bayesian network was proposed. From the historically accumulated data in a PCB manufacturer, different characteristic variables were extracted and specified for the model.
2. PCA was employed for data compression and factors extraction. Yet, an entropy minimization based method was presented to discretize variable and extracted factors. They could facilitate data compression, input type reduction, and better classification performance.
3. In order to avoid the defect of TAN BN that easily misses strong links between nodes and generates redundant weak links, CMI and LSP were combined for the establishment of the BN structure.
4. By using Monte Carlo simulations, the confidence upper limits of reorder frequency within six months were determined and the influence of the random nature of reorder was reduced.

Further research will be made to design intelligent approaches that can predict and determine reasonable batch production area for each candidate order. Further attempts will also be made to apply this method to similar order-oriented production and develop other intelligent techniques for the repetition pattern excavation.

Author Contributions: S.L. implemented the algorithm and wrote the paper. H.K. edited the paper and improved the quality of the article. H.J. proposed the algorithm and the structure of the paper. B.Z. conducted the experiments and analyzed the data.

Acknowledgments: This paper is supported by the National Natural Science Foundation of China (Grant No. 51605169) and Natural Science Foundation of Guangdong, China (Grant No. 2014A030310345). The authors wish to thank Guangzhou FastPrint Technology Co., Ltd. for providing data for the study.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Marques, A.C.; Cabrera, C.J.; Malfatti, C.F. Printed circuit boards: A review on the perspective of sustainability. *J. Environ. Manag.* **2013**, *131*, 298–306. [CrossRef] [PubMed]
2. Ngai, E.W.T.; Xiu, L.; Chau, D.C.K. Application of data mining techniques in customer relationship management: A literature review and classification. *Exp. Syst. Appl.* **2009**, *36*, 2592–2602. [CrossRef]
3. Dursun, A.; Caber, M. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tour. Manag. Perspect.* **2016**, *18*, 153–160. [CrossRef]
4. Chen, Y.L.; Kuo, M.H.; Wu, S.Y.; Tang, K. Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. *Electron. Commer. Res. Appl.* **2009**, *8*, 241–251. [CrossRef]
5. Hu, Y.; Yeh, T. Discovering valuable frequent patterns based on RFM analysis without customer identification information. *Knowl. Syst.* **2014**, *61*, 76–88. [CrossRef]
6. Coussement, K.; Filip, A.M.V.B.; Bock, K.W. Dataac accuracy's impt on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *J. Bus. Res.* **2014**, *67*, 2751–2758. [CrossRef]
7. Mohammadzadeh, M.; Hoseini, Z.Z.; Derafshi, H. A data mining approach for modeling churn behavior via RFM model in specialized clinics case study: A public sector hospital in Tehran. *Procedia Comput. Sci.* **2017**, *120*, 23–30. [CrossRef]
8. Song, M.N.; Zhao, X.J.; Haihong, E.; Qu, Z.H. Statistics-based CRM approach via time series segmenting RFM on large scale data. *Knowl. Syst.* **2017**, *132*, 21–29. [CrossRef]
9. Liu, J. Using big data database to construct new GFuzzy text mining and decision algorithm for targeting and classifying customers. *Comput. Ind. Eng.* **2018**. [CrossRef]
10. Sarti, S.; Darnall, N.; Testa, F. Market segmentation of consumers based on their actual sustainability and health-related purchases. *J. Clean. Prod.* **2018**, *192*, 270–280. [CrossRef]
11. Murray, P.W.; Agard, B.; Barajas, M.A. Forecast of individual customer's demand from a large and noisy dataset. *Comput. Ind. Eng.* **2018**, *118*, 33–43. [CrossRef]
12. De Caigny, A.; Coussement, K.; Bock, K.W.D. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *Eur. J. Oper. Res.* **2018**, *269*, 760–772. [CrossRef]
13. Zerbino, P.; Aloini, D.; Dulmin, R.; Mininno, V. Big Data-enabled customer relationship management: A holistic approach. *Inform. Proc. Manag.* **2018**. [CrossRef]
14. Wang, G.; Xu, T.H.; Tang, T.; Yuan, T.M.; Wang, H.F. A Bayesian network model for prediction of weather-related failures in railway turnout systems. *Expert Syst. Appl.* **2017**, *69*, 247–256. [CrossRef]
15. Arias, J.; Gamez, J.A.; Puerta, J.M. Learning distributed discrete Bayesian network classifiers under MapReduce with Apache Spark. *Knowl. Based Syst.* **2017**, *11*, 16–26. [CrossRef]
16. Thorson, J.T.; Johnson, K.F.; Methot, R.D.; Taylor, I.G. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. *Fish. Res.* **2017**, *192*, 84–93. [CrossRef]
17. Mack, D.L.C.; Biswas, G.; Koutsoukos, X.D.; Mylaraswamy, D. Learning Bayesian network structures to augment aircraft diagnostic reference models. *IEEE Trans. Autom. Sci. Eng.* **2017**, *14*, 358–369. [CrossRef]
18. Alonso-Montesinos, J.; Martínez-Durbán, M.; Sagrado, J.; Águila, I.M.D.; Batlles, F.J. The application of Bayesian network classifiers to cloud classification in satellite images. *Renew. Energy* **2016**, *97*, 155–161. [CrossRef]
19. Hosseini, S.; Barker, K. Modeling infrastructure resilience using Bayesian networks: A case study of inland waterway ports. *Comput. Ind. Eng.* **2016**, *93*, 252–266. [CrossRef]
20. Hosseini, S.; Barker, K. A Bayesian network model for resilience-based supplier selection. *Int. J. Prod. Econ.* **2016**, *180*, 68–87. [CrossRef]

21. Li, B.C.; Yang, Y.L. Complexity of concept classes induced by discrete Markov networks and Bayesian networks. *Pattern Recognit.* **2018**. [CrossRef]
22. Liu, B.; Hu, J.; Yan, F.; Turkson, R.F.; Lin, F. A novel optimal support vector machine ensemble model for NOx emissions prediction of a diesel engine. *Measurement* **2016**, *92*, 183–192. [CrossRef]
23. Ramrez-Gallego, S.; Krawczyk, B.; Woniak, M.; Woniak, M.; Herrera, F. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing* **2017**, *239*, 39–57. [CrossRef]
24. Fayyad, U.; Irani, K. Multi-interval discretization of continuous-valued attributes for classification learning. *Int. J. Conf. Artif. Intel.* **1993**, 1022–1029.
25. Boonchuay, K.; Sinapiromsaran, K.; Lursinsap, C. Decision tree induction based on minority entropy for the class imbalance problem. *Pattern Anal. Appl.* **2017**, *20*, 769–782. [CrossRef]
26. Tahan, M.H.; Asadi, S. MEMOD: A novel multivariate evolutionary multi-objective discretization. *Soft Comput.* **2017**, *22*, 1–23. [CrossRef]
27. Zhao, Y.; Xiao, F.; Wang, S. An intelligent chiller fault detection and diagnosis methodology using bayesian belief network. *Energy Build.* **2013**, *57*, 278–288. [CrossRef]
28. Jun, H.B.; Kim, D. A Bayesian network-based approach for fault analysis. *Expert Syst. Appl.* **2017**, *81*, 332–348. [CrossRef]
29. Wang, S.C.; Gao, R.; Wang, L.M. Bayesian network classifiers based on Gaussian kernel density. *Expert Syst. Appl.* **2016**, *51*, 207–217. [CrossRef]
30. Gan, H.X.; Zhang, Y.; Song, Q. Bayesian belief network for positive unlabeled learning with uncertainty. *Pattern Recogn. Lett.* **2017**, *90*, 28–35. [CrossRef]
31. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian network classifiers. *Mach. Learn.* **1997**, *29*, 131–163. [CrossRef]
32. Imme, E.U. *Georgia Tech Research Report: Tutorial on How to Measure Link Strengths in Discrete Bayesian Networks*; Woodruff School of Mechanical Engineering, Georgia Institute of Technology: Atlanta, GA, USA, 2009.
33. Confusion Matrix. Available online: https://en.wikipedia.org/wiki/Confusion_matrix (accessed on 11 October 2017).
34. Liu, H.; Tian, H.Q.; Li, Y.F.; Zhang, L. Comparison of four Adaboost algorithm based artificial neural networks in wind speed predictions. *Energy Convers. Manag.* **2015**, *92*, 67–81. [CrossRef]
35. Cortes, C.; Gonzalvo, X.; Kuznetsov, V.; Mohri, M.; Yang, S. AdaNet: Adaptive structural learning of artificial neural networks. *arXiv*, 2016.
36. Sharma, A.; Sahoo, P.K.; Tripathi, R.K.; Meher, L.C. Artificial neural network-based prediction of performance and emission characteristics of CI engine using polanga as a biodiese. *Int. J. Ambient Energy* **2016**, *37*, 559–570. [CrossRef]
37. Saravanan, K.; Sasithra, S. Review on classification based on artificial neural networks. *Int J. Ambient Syst. Appl.* **2014**, *2*, 11–18.

applied sciences

MDPI

Review

A Review of Data Mining with Big Data towards Its Applications in the Electronics Industry

Shengping Lv [1,2], Hoyeol Kim [2], Binbin Zheng [1] and Hong Jin [1,*]

[1] College of Engineering, South China Agricultural University, Guangzhou 510642, China; lvshengping@scau.edu.cn (S.L.); zhengbinbin@stu.scau.edu.cn (B.Z.)

[2] Department of Industrial, Manufacturing and Systems Engineering, Texas Tech University, Lubbock, TX 79409, USA; hoyeol.kim@ttu.edu

* Correspondence: hjin@scau.edu.cn; Tel.: +86-187-1937-3880

Received: 11 March 2018; Accepted: 4 April 2018; Published: 8 April 2018

Featured Application: This review not only benefits researchers to develop strong research themes and identify gaps in the field but also helps practitioners for DM and Big Data application system development.

Abstract: Data mining (DM) with Big Data has been widely used in the lifecycle of electronic products that range from the design and production stages to the service stage. A comprehensive analysis of DM with Big Data and a review of its application in the stages of its lifecycle will not only benefit researchers to develop strong research themes and identify gaps in the field but also help practitioners for DM application system development. In this paper, a brief clarification of DM-related topics is presented first. A flowchart of DM and the main content of the flowchart steps are given in which commonly used data preparation and preprocessing approaches, DM functions and techniques, and performances indicators are summarized. Then, a comprehensive review covering 105 articles from 2007 to 2017 on DM or Big Data applications in the electronics industry is provided according to the flowchart from various points of view such as data handling, applications of DM, or Big Data at different lifecycle stages, and the software used in the applications. On this basis, a diagram of data content for different knowledge areas and a framework for DM and Big Data applications in the electronics industry are established. Finally, conclusions and future research directions are given.

Keywords: data mining; knowledge discovery in databases; big data; electronics industry; semiconductor; wafer; print circuit board; product lifecycle management

1. Introduction

Since the internet of things and advanced information technologies (for example, radio frequency identification (RFID) tags and smart sensors) are widely used in manufacturing enterprises for their daily production and management, the product lifecycle management (PLM) processes produce a huge amount of data [1]. Furthermore, the accumulation of historical data in enterprise resource planning (ERP), supply chain management (SCM), customer relationship management (CRM), and order management system (OMS), as well as the timely collected data by the widely used manufacturing execution system (MES) and distributed control system (DCS) contributed to the sharp increase of data over the decades. The era of industrial Big Data has come.

Leaders of manufacturing enterprises are becoming increasingly interested in benefiting their companies by effectively using Big Data [1]. Big data related technologies such as knowledge discovery in databases (KDD) and data mining (DM) have been widely employed to enhance the intelligence and efficiency of the design, production, and service processes in many manufacturing scenes such

as product design improvement, manufacturing process optimization, production management and optimization (PMO), production process monitoring and control, quality management, CRM, SCM, and so forth. Intel employs Big Data for predictive maintenance of equipment and greatly reduces the unnecessary equipment stop and idle time. A Taiwan Semiconductor Manufacturing Company adopts Big Data based advanced equipment control/advanced process control (AEC/APC) to improve production efficiency and wafer yield. Many reviews of these applications in the manufacturing industry have been reported and summarized in Table 1, from which we can see most of the achievements related to DM application in manufacturing before 2015 [2–6], and many researchers have started to adopt the concept of Big Data [7–11] in smart manufacturing since then. However, the aforementioned reviews provide no comprehensive analysis of DM with Big Data nor a summarization of them in the electronics industry from the view of their lifecycle, considering the special requirement of this manufacturing industry to the best of our knowledge.

Table 1. The reviews of data mining and big data application in the smart manufacturing industry.

Reference	Main Review Content	Year
Choudhary et al. [2]	Application of KDD and DM in manufacturing, the kinds of patterns to be mined, and data mining techniques (DMTs)	2009
Ngai et al. [3]	DM application in customer identification, attraction, retention, and development	2009
Gulser et al. [4]	DM application for product quality improvement tasks including quality description/predicting/classification and parameter optimization	2011
Liao et al. [5]	DMTs applications in CRM, product development, and fault pattern analysis	2012
Hamidey et al. [6]	Support vector machine (SVM) application in quality assessment in manufacturing	2015
Donovan et al. [7]	Application of Big Data in the area of design, process and planning, quality management, maintenance and diagnosis, scheduling, control, environment, and so forth.	2015
Li et al. [8]	Concept, characteristics, and potential application of Big Data in PLM	2015
Zhong et al. [9]	Big Data applications in finance, economics, healthcare, SCM, and the manufacturing sector. Current movements on the Big Data for SCM in service and manufacturing	2016
Nagorny et al. [10]	Big Data in smart manufacturing systems including related research roadmaps and projects in European, the infrastructures, Big Data analysis process, algorithm and tools, and so forth.	2017
Cheng et al. [11]	Development of DMTs, major functions of DMTs, applications of DMTs to production management in the Big Data era	2017

Electronics is one of the fastest evolving, most innovative, and most competitive industries. The research and development of new and improved products are of great importance, where companies often compete fiercely to bring the newest technology to the market first. The past five years, from 2012 to 2017, have been characterized by growth in emerging markets and introduction of new products, leading more people to buy consumer electronics. The global consumer electronics industry was valued at $283 billion in 2015 [12]. Grand view research predicted that the global consumer electronics market is expected to reach $838.85 billion by 2020 [13]. The newly developed products are featured by high precision, long and complex manufacturing/test processes with high purity environments, diverse and high-quality requirements from customers, and a large amount of data generated at different stages of their lifecycle from design and production to sale and service. Thus, the electronics industry is currently in the midst of a data-driven revolution [7] which has pushed

forward many data excavation related research over the past decades for the better utilization of these data that can facilitate quality or service improvement, production optimization, and so forth. [14]. A review of DM with Big Data application in the electronics industry not only benefits researchers to develop strong research themes and identify gaps in the field but also helps practitioners for DM application system development.

In the following sections, DM with Big Data and related techniques are given in Section 2 in which a brief introduction of the concepts of DM and Big Data is presented, and also the flowchart and the main content of the flowchart steps are summarized. In Section 3, the article selection condition and distribution of the selected articles in different years and different lifecycle stages are discussed. A comprehensive analysis of the reviewed literature from various points of view is provided subsequently, in Section 4, which summarizes data handling, discusses the DM with Big Data application in different stages of the product lifecycle, and surveys the software used in these applications. On this basis, the data content and a framework for DM application in the electronics industry are established in Section 5. Finally, the conclusions and future research directions are given in Section 6.

2. Data Mining with Big Data

2.1. Concepts of Data Mining and Big Data

There are many concepts such as DM, KDD, and Big Data that are closely related to each other. DM, as an interdisciplinary subject including database design, statistics, pattern recognition, machine learning, and data visualization [6], can be defined in many different ways. Romero and Ventura [15] specified DM as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Han et al. [16] defined DM as "the process of discovering interesting patterns and knowledge from large amounts of data".

Many researcher and practitioners treat DM as a synonym for KDD as IBM [17] deems KDD and DM the same as "an interdisciplinary area focusing on methodologies for extracting useful knowledge from data". However, others think that "KDD refers to the overall process of discovering knowledge from data while DM (in a narrow sense) refers to application of algorithms for extracting patterns from data without the additional steps of the KDD process" [16], in which the additional steps include data preparation, preprocessing, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining [16]. Here, we take DM as a synonym for KDD whereas DM in a narrow sense refers only to the step to generate a specific pattern using a particular algorithm within an acceptable computational efficiency limit [11,16].

There are various definitions of Big Data from 3 Vs to 4 Vs [18]. Volume, velocity, and variety are the well-known 3Vs and the fourth V can be value, variability, or virtual [8,18]. Wikipedia specifies that "Big Data is data sets that are so voluminous and complex that traditional data processing methods are inadequate to deal with them" [19]. Gartner gives a more detailed definition as follows: "Big Data is high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" [20]. Big Data analysis is strongly connected with classical data analysis and DM approaches to access and process these amounts of data very fast [2,10].

The flowchart of DM with Big Data is illustrated in Figure 1. The main content of each step includes data preparation, preprocessing, DM in a narrow sense, and evaluation. The interpretation of the results will be discussed in the following sections.

2.2. Data Preparation and Preprocessing

The data preparation includes problem clarification and collecting the targeted data. The problem clarification is to understand the industry domain including the relevant prior knowledge related to different applications and targeted goals [4]. The targeted data can be obtained by experimental

observations, historical accumulated records, online sensor measurement, real-time status of RFID tags, and simulation results. These data sets can be stored in different formats such as data warehouse, marts, database, files, and so on [4,16], and the data relevant to the mining tasks are retrieved and selected before data preprocessing.

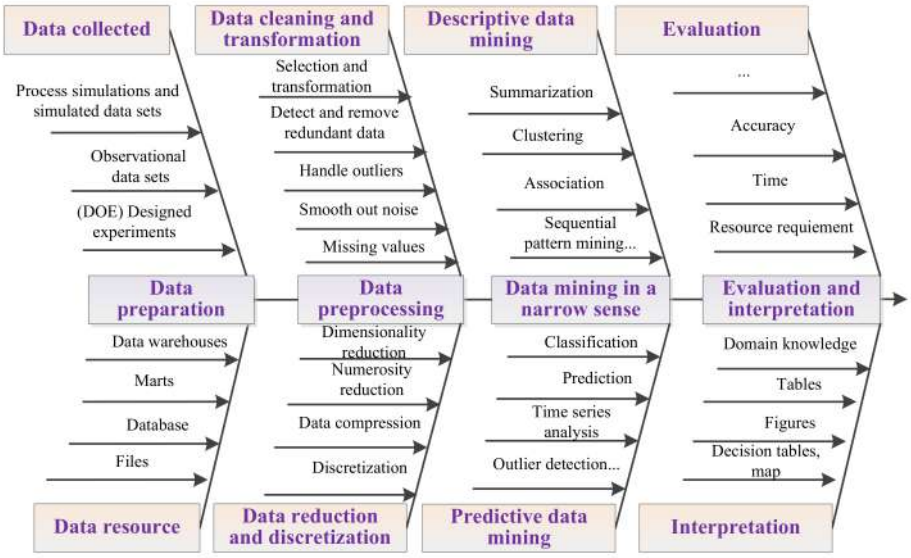


Figure 1. The data mining flowchart.

The preprocessing consists of data cleaning, transformation, reduction, and discretization. Data cleaning operation involves techniques for filling in missing values, smoothing out noise, handling outliers, detecting, and removing redundant data. Data transformation puts the data into appropriate forms for mining when necessary. Data reduction is performed to obtain a smaller representation of the original data without sacrificing its integrity. Dimensionality reduction, numerosity reduction, and data compression are the three ways for data reduction. Dimensionality reduction is a technique to detect and remove irrelevant, weakly relevant or redundant attributes [16]. Numerosity reduction replaces the original data volume by alternative and smaller forms of data representation. In data compression, transformations are applied so as to obtain a reduced or compressed representation of the original data, such as principal components analysis (PCA). Discretization reduces the number of levels of an attribute by collecting and replacing low-level concepts with high-level concepts [4].

2.3. Data Mining in a Narrow Sense

Data mining in a narrow sense, as the core of DM, is to derive the model and mining the patterns/knowledge in the data. The patterns to be mined determine the DM functions to be performed which can always be divided into descriptive and predictive DM. The descriptive function is to characterize properties of the data in a target data set that mainly includes the functions of summarization, clustering, and association/sequential pattern mining. While the predictive DM performs induction on the current data in order to make predictions that mainly consists of the functions of classification, prediction, outlier detection (anomaly detection), and time series analysis [4,11,16]. The corresponding data mining techniques (DMTs) to realize different functions can be categorized into statistical analysis-oriented (SA-oriented) and knowledge discovery-oriented (KD-oriented). SA-oriented techniques make assumptions about data distribution and relationships between variables based on prior knowledge in advance and verify or deny the assumptions. Common SA-oriented DMTs include the algorithms such as regression, k-nearest neighbor (k-NN), k-means, Bayesian classifier [21], and so on. On the contrary, KD-oriented DMTs search for the relationship

automatically under no clear assumptions [11]. The details of the DM functions and the related DMTs are summarized in Table 2 [4,11,16].

Table 2. The data mining functions and related techniques.

Type of Function	DM Functions	Description	Related DMTs
Descriptive DM	Summarization	Summarization of the general characteristics of a data set	Statistical measures and plots, online analytical processing, attribute-oriented induction, and so forth.
	Clustering	Grouping a set of data objects into multiple clusters so that objects within a cluster have high similarity	Centroid-based clustering, connectivity-based clustering, density-based clustering, and distribution-based clustering
	Association/ Sequential pattern mining	Mining frequent patterns to discover interesting associations and correlations (in a sequence for sequential pattern mining)	Apriori, AprioriAII, sampling, partitioning pattern growth, correlation rules, stream patterns, and so forth.
Predictive DM	Classification	A model or classifier is constructed to predict class (categorical) labels	DT, Bayesian, rule-based, SVM, ANN, CBR, k-NN, GA, RST, and Fuzzy Set
	Prediction	A model performing prediction function to forecast future values of continuous type data	Regression, ANN, SVM/SVR, DT, RST, and Fuzzy set
	Outlier detection	The process of finding data or objects that behave unexpectedly	Classification based, k-NN based, clustering based, and so forth.
	Time series analysis	Methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data	Regression, SVM, ANN, RST, and Fuzzy Set

DT: Decision tree; CBR: Case-based reasoning; GA: Genetic algorithm, RST: Rough set theory; SVM/SVR: Support vector machine/regression.

2.4. Performance Indicators

The knowledge extracted should be evaluated and interpreted correctly to obtain reliable results. The evaluation of the DM methods to reach a final decision requires a comparison of results obtained from various DM methods using several measures [4]. The performance indicators employed to evaluate classifiers based on a confusion matrix are illustrated in Figure 2. The indicators widely used for the measurement of prediction, clustering, and association of DM functions are summarized in Tables 3–5 respectively.

<table>
<tr><td></td><td></td><td colspan="2">True condition</td><td></td><td colspan="2"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence= (TP+FN)/Total population</td><td colspan="2">ACC=(ΣTP+ΣTN)/ Total population</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive (TP)</td><td>False positive (FP) Type I error</td><td>Positive predictive value (PPV) Precision =Σ TP/Σ(TP+FP)</td><td colspan="2">False discovery rate (FDR)= ΣFP/ (ΣTP+ΣFP)</td></tr>
<tr><td>Predicted condition negative</td><td>False negative (FN) Type II error</td><td>True negative (TN)</td><td>False omission rate (FOR)= ΣFN/(ΣFN+ΣTN)</td><td colspan="2">Negative predictive value (NPV) =Σ TN/ (ΣFN+ΣTN)</td></tr>
<tr><td colspan="2"></td><td>TPR, Recall, Sensitivity= ΣTP/(ΣTP+ΣFN)</td><td>FPR = ΣFP/(ΣFP+ΣTN)</td><td>Positive likelihood ratio (LR+) = TPR/FPR</td><td rowspan="2">Diagnostic odds ratio (DOR) = LR+/LR−</td><td rowspan="2">F1 score = 2/(1/TPR + 1/Precision)</td></tr>
<tr><td colspan="2"></td><td>False negative rate (FNR)= ΣFN/(ΣTP+ΣFN)</td><td>True negative rate (TNR), Specificity = ΣTN/(ΣFP+ΣTN)</td><td>Negative likelihood ratio (LR−) = FNR/TNR</td></tr>
</table>

Figure 2. The confusion matrix and performance indicators for classification [22].

Table 3. The performance indicators for the prediction function [23].

Indicators	Equation	Indicators	Equation
MAPE	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\hat{y}_i - y_i}{y_i}\right\| \times 100$	R^2	$R^2 = \sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2 / \sum_{i=1}^{N}(y_i - \overline{y})^2$
MSE	$MSE = \sum_{i=1}^{N}(\hat{y}_i - y_i)^2 / N$	ME	$ME = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert y_i - \hat{y}_i\rvert}{y_i}$
MAE	$MAE = \sum_{i=1}^{N}\lvert\hat{y}_i - y_i\rvert / N$	VARER	$VARER = \frac{1}{n-1}\sum_{k=1}^{n}(\frac{\lvert y_i - \hat{y}_i\rvert}{y_i} - ME)^2$
RMSE	$RMSE = \sqrt{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 / N}$	RE	$RE = \frac{E(y_i - \hat{y}_i)^2}{E(y_i - \overline{y})^2}$
RSE	$RSE = \sqrt{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 / N}$	IA	$IA = 1 - \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{N}(\lvert\hat{y}_i - \overline{y}\rvert + \lvert y_i - \overline{y}\rvert)^2$
RAE	$RAE = \sqrt{\sum_{i=1}^{N}\lvert\hat{y}_i - y_i\rvert / N}$	-	-

Note: y_i and $\hat{y}_i$ are the observed and predicted value of sample i respectively; $\overline{y}$ is the average result of samples. $\sum_{i=1}^{N}(y_i - \overline{y})^2$ is the total sum of squares, while $\sum_{i=1}^{N}(\hat{y}_i - \overline{y})^2$ is the explained sum of squares. E is the expectation value.

Table 4. The performance indicators for the clustering function [24].

Indicators	Equation	Description
DBI	$DBI = \frac{1}{n}\sum_{i=1}^{n}\max_{j \neq i}(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)})$	n is the number of clusters, c_x is the centroid of cluster x, σ_x is the average distance of all elements in cluster x to centroid c_x, and $d(c_i, c_j)$ is the distance between centroids c_i and c_j.
DI	$D = \frac{\min_{1 \leq i < j \leq n} d(i,j)}{\min_{1 \leq k \leq n} d\prime(k)}$	$d(i,j)$ represents the distance between clusters i and j; $d'(k)$ measures the intra-cluster distance of cluster k.
Purity	$Purity(\Omega, \mathbb{C}) = \frac{1}{N}\sum_k \max_j \lvert w_k \cap c_j\rvert$	$\Omega = \{w_1, w_2, \ldots, w_k\}$ is the set of clusters, $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ is the set of classes, I is mutual information, and H is entropy.
NMI	$NMI(\Omega, \mathbb{C}) = \frac{I(\Omega, \mathbb{C})}{\lvert H(\Omega), H(\mathbb{C})\rvert / 2}$	
RI	$RI = \frac{TP + TN}{TP + FP + FN + TN}$	The definitions of TP, TN, FP, FN, precision, and recall are the same as the specifications given in Figure 2; β is the penalty coefficient.
F measure	$F_\beta = \frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$	
QE	$QE = \frac{1}{N}\sum_{i=1}^{N}\left\|\left\|x_i - r_\beta\right\|\right\|$	N refers to the number of original data vectors, and r_β is the best matching unit of the data vector x_i; $u(x)$ gets the value of 1 if the best and the second best matching units of the input vector are non-adjacent, and 0 otherwise.
TE	$TE = \frac{1}{N}\sum_{i=1}^{N} u(x_i)$	

Table 5. The performance indicators for the association function.

Indicators	Equation	Description
Support	$\text{sup}(X) = \frac{\lvert\{t \in T; X \subseteq t\}\rvert}{\lvert T\rvert}$	X is an item set, $X \rightarrow Y$ is an association rule, and T is a set of transactions. Support of X (sup (X)) with respect to T is defined as the proportion of transactions t in the dataset which contains the item set X. $conf(X \rightarrow Y)$ is the proportion of the transactions that contains X which also contains Y.
Confidence	$conf(X \rightarrow Y) = \text{sup}(X \cup Y) / \text{sup}(X)$	
Lift	$lift(X \rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X) \times \text{sup}(Y)}$	
Conviction	$conv(X \rightarrow Y) = \frac{1 - \text{sup}(Y)}{1 - conf(X \rightarrow Y)}$	

Accuracy (ACC), precision, sensitivity or recall, specificity, and so forth, given in Figure 2, are the commonly employed indicators. Meanwhile, the receiver operating characteristic curve (ROC) created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings is always taken to illustrate the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The performance indicators for prediction mainly include the mean absolute percentage error (MAPE), the mean squared error (MSE), the mean absolute error (MAE), the root-mean-square error (RMSE), the root absolute error (RAE), the mean error (ME), the variance of errors (VARER), the relative error (RE), the goodness of fit (R^2), the index of agreement (IA), and so on. Typical objective functions to assess the quality of clustering include internal and external criteria. The internal criterion for the quality of a clustering can be evaluated by the Davies–Bouldin index (DBI), Dunn index (DI), and so on, while the most used external criteria includes purity, normalized mutual information (NMI), rand index (RI), *F* measure, and so on. Meanwhile, some indicators like the quantization error (QE) and the topographic error (TE) are for a special algorithm like self-organizing map (SOM). The support, confidence, lift, and conviction are pervasive performance indicators for association. The outlier detection can be taken as a binary classification, and the performance indicators for classification can be used to evaluate the results. Time series analysis can be used for clustering, classification, and anomaly detection, as well as forecasting, and therefore, the related performance can be verified by the corresponding indicators for clustering, classification, and prediction.

3. Article Selection and Distribution

The electronics industry is composed of organizations involved in the design, development, manufacture, assembly, and service of electronic equipment and components. These organizations offer a wide variety of products that range from government products, industrial products, consumer products, and electronic components as four primary segments. Each category serves a specific market, which allows it to focus on components and products geared toward their customers. The government market is primarily developed for aircraft and military products, as well as communication technology and medical devices. Industrial products include large-scale computers, radio and television broadcasting equipment, telecommunications equipment, and electronic office equipment, while consumer products are the well-known televisions, cell phones, DVD players, smartphones, radios, video game systems, personal computers, electronic ovens, and home intercommunication and alarm systems. The final segment the manufacturers produce and sell includes electron tubes, semiconductors, printed circuit boards (PCB), and passive components [25].

Based on the initial search from databases with keywords such as DM, Big Data, and electronics, we found that most of the articles were related to consumer products and components. Therefore, articles related to DM with Big Data applications in consumer electronics and components were selected here. On this basis, the article selection was conducted in which the period of interest for this literature survey ranges from 2007 to 2017. In October 2017, a search was made according to the following conditions:

(1) Database: Science Direct, IEEE Xplore Digital Library, Springer Link, Taylor & Francis Online, Wiley Online Library, SAGE Journal, Web of Science, and Google Scholar

(2) Stages: design, production, sale, service, and recycling

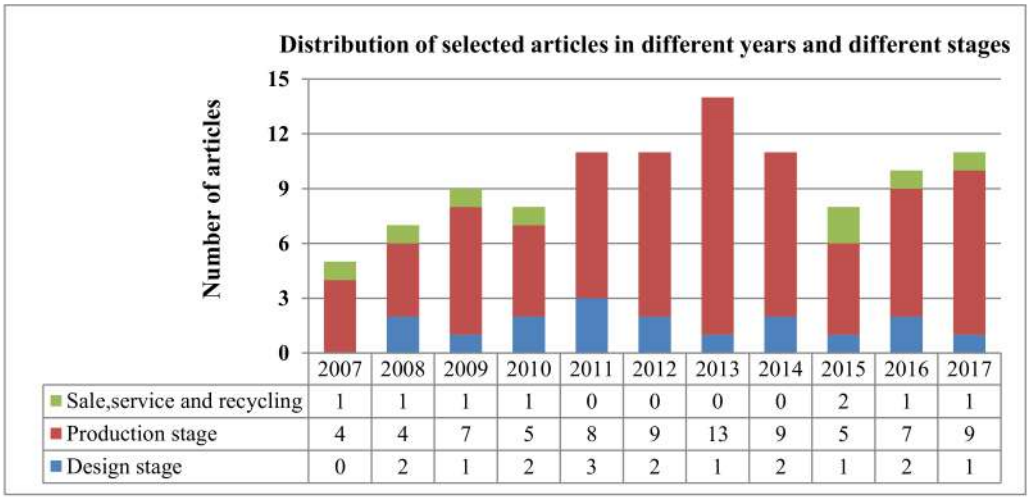
(3) Products: electronic products, integrated circuit, wafer, semiconductor, PCB, phone, and computer

(4) DM-related concepts: data mining, Big Data, and knowledge discovery

(5) DM functions: Prediction, classification, clustering, association, product/process characterization, time series analysis, outlier detection, and anomaly detection.

A total of 105 application studies within the scope of this review were found. The distribution of the selected articles in different years and different stages are illustrated in Figure 3. It can be seen

that 17% (17 articles) were related to the stage of product and manufacturing process design [26–42], and more than 75% (80 articles) applied DM and Big Data to production management and control in the stage of production [43–122], but less than 8% (8 articles) of applications focused on the stage of sale, service, and recycling [123–130]. The fluctuation in quantity of the selected articles in different years presents no obvious tendency, however, it indicates that the topic has attracted ongoing attention and research during the past decades, and the application areas have been extended and many new approaches have been developed.



	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Sale,service and recycling	1	1	1	1	0	0	0	0	2	1	1
Production stage	4	4	7	5	8	9	13	9	5	7	9
Design stage	0	2	1	2	3	2	1	2	1	2	1

Figure 3. The distribution of selected articles in different years and different stages.

4. Data Mining with Big Data Applications in the Electronics Industry

In the following, we examine and discuss the reviewed literature from various points of view based on the flowchart given in Figure 1. Data handling, or more specifically, data preparation and data preprocessing before performing the DM functions, are discussed first. Next, DM with Big Data applications in different stages of the electronics industry, including the knowledge area, DM functions, developed DMTs, and performance indicators, are summarized. In addition, findings of these applications in each knowledge area are given, and the summarization of these reviews is also presented. Finally, the software tools used in these applications are examined.

4.1. Data Handling

Data preparation is the initial step of DM to collect the necessary data recording the feature values directly from the experimental data and historical observations or indirectly from the simulation results [4], in which the experimental data are the records of full factorials or fractional factorials while historical observations can be obtained either through online measurements or from historical accumulated records. The data preparation from the reviewed literature is summarized in Table 6.

Through Table 6, we can see that the data for the verification of product design improvement and manufacturing process optimization were mainly based on experimental observation and historical records. The DM application in the production process monitoring and control for the tasks of fault detection and classification (FDC), run to run (R2R), statistical process control (SPC), and so on, worked mainly on the data obtained through online measurements while DM in production and quality management for the tasks such as scheduling, yield/cost/cycle time prediction, and so forth was conducted mainly based on historical records from ERP and MES along with some process simulation. The task of SCM and CRM is conducted mainly based on interactions and transaction records accumulated in the system of SCM, OMS, and CRM.

Table 6. The data preparation from the reviewed articles.

Data Preparation	Records Obtainment	Reference	Application
Experimental data	Full factorials	[26,29,33,35,36,39–41,84,107]	Product design improvement and manufacturing process optimization
	Fractional factorial	[27,114]	
	Orthogonal experiment	[31]	
Observational historical data	Historical records	[28,30,32,34,37,38]	Production management
		[43–46,48–52,54,56,59–64,66–77]	
	Accumulated records	[79,81,85,90,94,95]	Production process monitoring and control like FDC, R2R, SPC, and so forth.
	Online measured, traced or monitoring	[80,82–93,96–99,101]	
	Interactions and transaction records	[38,42,123–130]	CRM/SCM
	Accumulated records	[100,103–106,108–113,115–122]	Quality management
Simulation data	-	[47,53,55,58,78,121]	Process optimization, such as scheduling and cycle time prediction
Unspecified	-	[57,65,102]	-

Data preprocessing techniques used in the selected applications are summarized in Table 7, from which we can see that most of the cleaning techniques were used for the observational data sets. Some imputation techniques such as the missing values-patient rule induction method (m-PRLM) [30], k-NN [84,87], syndromes imputation [109] and so on were developed for filling in missing values. SVM [54], moving average smoothing [90], King-move neighborhood [93], Winter's exponentials smoothing [126,127], and so on were employed for noise smoothing. Meanwhile, the methods of box plot [79,88], PCA [97], clustering [122], and so forth were applied for outliers detection. However, the missing values, noise, outliers, and redundant data were omitted directly in most cases.

Table 7. The preprocessing techniques used in the reviewed literature.

Preprocess	Functions	Methods	Reference
Data cleaning	Filling in missing values	m-PRLM	[30]
		Delete	[37,52,81,88,91,96,117,123,124,130]
		Manually fill	[48]
		k-NN	[84,87]
		Omit/replace	[85,130]
		Missing syndromes	[109]
		SVM	[54]
	Smoothing out noise	Delete	[79,82,114,120]
		Moving average	[90]
		King-move neighborhood	[93]
		Winter's exponentials	[126,127]
	Handling outliers	Box plot	[79,88]
		Delete	[84,96]
		Online PCA	[97]
		Clustering	[122]
	Handling redundant data	Detecting and removing	[81,96]
Data transformation	-	Variance scaling	[35,63]
		Normalization	[37,39,46,48,51,55–58,60–62,64–78,80, 82,88,94,98,102,104,105,108,111,113]
		Text mining	[42,123]
		Fisher Z	[44,47]
		Box-Cox	[84]
		Numerical into binary	[85]
		Binary vector	[91,93,117]
		Spreadsheet format	[95]

Table 7. *Cont.*

Preprocess	Functions	Methods	Reference
Data reduction	Dimensionality reduction	ANOVA	[27,79]
		Multilayer perceptron	[35]
		Stepwise regression	[34,83,89,110,126,127]
		GA (GA+SVR)	[55,83]
		RST	[54]
		Regression-based	[44,45,86]
		SNBC	[53]
		Conditional mutual information	[63]
		Las Vegas filter	[60,78]
		By experts	[81,82]
		Pearson coefficient	[79]
		Mapper	[84]
		Cramer's V correlation coefficients	[85,87]
		Exclusive key parameter selection	[99]
		LASSO, Random forest, and PCA	[110,112]
		K-W test	[114]
		Eliminating variables	[120]
		Auxiliary variables derived	[124]
	Numerosity reduction	Aggregation	[34,87]
		Clustering	[38,90,101]
		Adjust imbalanced classes	[54]
		Sampling	[82]
		K-means and SOM clustering	[126]
	Compression	PCA	[64,65,83,92,94]
		Multi-dimensional scaling	[84]
	Discretization	Equal frequency discretization	[63,120]
		CHAID	[90]

LASSO: Least absolute shrinkage and selection operator; SNBC: Selective naive Bayesian classifier; CHAID: Chi-squared automatic interaction detection.

Data transformation is the process of converting data from one format or structure into another. The pervasive method is normalization for the selected articles but few were conducted based on variance scaling [35,63], text mining [42,123], Fisher Z-transformation [44,47], binary vector transformation [91,93,117], Box-Cox transformation [84], and numerical into binary [85].

Dimensionality reduction, as one of the important approaches to data reduction, is to remove the irrelevant and redundant variables to reduce the complexity of analysis and the generated models, and also to improve the efficiency of the whole modeling processes. The widely used approaches from the reviewed articles include regression [34,44,45,83,86,89,110,126,127], analysis of variance (ANOVA) [27,79], GA [55,83], Las Vegas filter [60,78], Pearson coefficient [79], Cramer's V correlation coefficients [85,87], and so on. Clustering [38,90,101,126], aggregation [34,87], and sampling [82] based approaches were applied to reduce the data numerosity. PCA or the modified PCA [64,65,83,92,94], and multi-dimensional scaling [84] were employed to compress the representation of the original data. Only a few of the researchers conducted discretization for continuous attributes at the stage of preprocessing.

4.2. Application of DM with Big Data in Different Stages

DM with Big Data has been applied in different stages including design, production, sale, service, and recycling for different scenes, such as product design improvement, manufacturing process optimization, PMO, production process monitoring and control, quality management, CRM, SCM, and so forth. The application of DM with Big Data for the procurement of electronics components at the production stage has not been studied in the reviewed articles. Meanwhile, few reviewed articles have devoted their research into product distribution and logistics that mainly includes order process, inventory management, and product transportation at the stage of sale and service, and thus, we take them into SCM as a whole. The order management as an extension of CRM will also be considered as

CRM. Quality improvement (QI), development time/cost estimation (DTCE), PMO, AEC/APC, CRM, and SCM considered in the review are the typical knowledge areas to enhance the intelligence and efficiency of lifecycle management and control in which the data-driven QI is closely related to product design improvement, manufacturing process optimization, and quality management. AEC/APC, as the core of production process monitoring and control in the electronics industry, is also used to enhance product quality or yield. The task of AEC/APC is always conducted online during the manufacturing process and has attracted a lot of research. The description of these knowledge areas and their tasks is summarized in Table 8.

In the following sections, from Section 4.2.1 to Section 4.2.3, the summarization will not be taken as a function alone because it is employed to characterize the product/process and then to facilitate the functions of prediction, classification, clustering, and so forth. The SA-oriented and/or KD-oriented categories of different DMTs in an article will also be included.

Table 8. The knowledge areas of DM application in the electronics industry.

Knowledge Area	Sub-Areas	Description	Applied Stage
QI [4]	Description of product/process	(1) Identifying attributes that affect quality significantly; (2) Comparing the end result of the whole process with the desired specifications, analyzing the root causes of low yield for adjusting the process parameters to ensure future quality [102], and we call it as post hoc (fault) diagnosis here.	Design and production stage
	Quality classification	For a given set of input parameters, predicting the class of the quality output.	
	Quality prediction	Predicting what the resulting quality (yield) characteristic will be for a given set of input parameters or process values.	
	Parameter optimization	Based on the learned features of the cases, yielding high-quality and finding optimal levels of process/product parameters that consistently yield target performance.	
DTCE	-	Predicting the development time and/or cost.	Design stage
PMO	Scheduling	Scheduling optimization or dispatch rules selection.	Production stage
	Production time prediction (PTP)	Predicting the production time (cycle time/lead time/due or complete date).	
	Resource optimization	Resource allocation optimization.	
AEC/APC [4,11,86]	Fault detection and classification	Fault detection (FD) is to monitor and analyze the variation in equipment, tool or process data and detect anomalies, and the fault classification is to determine its root cause.	Production stage
	R2R	Modifying recipe parameters or the selection of control parameters between runs to improve performance.	
	Virtual metrology (VM)	Prediction of post-process metrology variables using process and wafer state.	
	Equipment health monitoring (EHM)	Monitoring tool parameters to assess the tool health as a function of deviation from normal behavior.	
	Statistical process control	Using statistical methods to analyze processes or products to take appropriate actions to achieve a state of statistical control and continuously improve the process capability.	
CRM [3]	Customer identification, attraction, retention, and development	Analyzing and understanding customers' behaviors and characteristics.	Sale, service and, recycling stage
SCM	-	DM application for the management of the flow of goods and services	

4.2.1. Application of DM and Big Data for Design

The design stage includes the product design followed by process planning. Product design is to create a new product while process planning is to translate product design requirements to manufacturing process details that act as a bridge between product design and manufacturing. Capodieci [14] presented a review of the data analysis and machining learning for the design process yield optimization in electronic design and semiconductor manufacturing. Another 17 articles related to DM application in the stage of design have been retrieved and summarized in Table 9, and the following findings can be achieved:

(1) The quality improvement of product design [28], the prediction of development cost and time [34,37], and the product customization [38,42] are the main applications of DM in product design.

(2) The optimization of the manufacturing process parameter is the main task of process planning, such as the parameter optimization of stencil printing process (SPP) [26,27,36], reflow soldering [29,31,32], fluid dispensing for microchip encapsulation [33], wave soldering [35,41], and hot solder dip [39] for component surface mounts on PCBs. These models always combined ANN, SVR, and regression for the quality prediction with GA for parameters optimization [26,31–33].

(3) KD-oriented ANN is the widely used DMT. The pervasive function is prediction and has been widely employed for parameter optimization and determination of its effect [26,27,32–36] followed by clustering and association. Clustering was mainly employed to identify similarity products, process plans, and parameters and no supervision classification was conducted to support more efficient and reasonable manufacturing [39–41]. Association was mainly used to identify purchase behavior and therefore, develop marketing competitive products [38,42].

Table 9. DM with Big Data application in the design stage.

Function (Frequency)	DMTs	Categories	Knowledge Area/Task	Product/Process	Indicator	Ref.
Prediction (12)	SVR	KD-oriented	Quality prediction	SPP	-	[26]
	MRO, ANN + GA, Fuzzy logic + Regression	KD-oriented, SA-oriented		SPP	RMSE	[27]
	M5′, ANN	KD-oriented		Wafer	RMSE, RE	[28]
	ANN + GA	KD-oriented	Parameter optimization	SPP	RMSE	[29]
	m-PRLM	KD-oriented		Wafer etching	MSE	[30]
	ANN + GA	KD-oriented		SPP	RMSE	[31]
	ANN + GA	KD-oriented		SPP	MAPE	[32]
	FNN + GA	KD-oriented		Microchip encapsulation	ME, VARER	[33]
	MRA, ANN, CBR, MRA + ANN, ANN + CBR	KD-oriented, SA-oriented	Development time/cost estimation	Liquid-crystal display	MAER, RMSE	[34]
	ANN	KD-oriented	Quality prediction, Process description	SPP	IA	[35]
	ANN	KD-oriented	Quality prediction Parameter optimization	SPP	MSE	[36]
	ASVR, MLR	KD-oriented, SA-oriented	Development time/cost estimation	Electronic circuit	MSE	[37]
Classification (1)	Apriori, C5.0	KD-oriented	Product description	Digital camera	-	[38]
Clustering (3)	SOM	KD-oriented	Process description	Hot solder dip	QE	[39]
	K-means	SA-oriented	Parameter optimization	PCB	-	[40]
	SOM	KD-oriented	Process description	Wave soldering	QE, TE	[41]
Association (2)	Apriori	KD-oriented	Product description	Apple iPad	Support, confidence and lift	[42]
	Apriori, C5.0	KD-oriented		Digital camera		[38]

MRO: Multi-response optimization; FNN: Fuzzy neural network; MRA: Multiple regression analysis; ASVR: Adaptive SVR; MLR: Multiple linear regression.

4.2.2. Application of DM and Big Data for Production

The product in its final shape is obtained in the production phase. The knowledge areas of DM with Big Data application in the stage of production include PMO, AEC/APC, and quality improvement. The reviewed studies are summarized in Tables 10–12 for PMO, AEC/APC, and quality improvement, respectively. The following conclusions can be obtained for the application of DM with Big Data for PMO:

(1) The scheduling optimization, cycle time, complete time, and output time prediction for wafer fabs have attracted most of the research. The reason may be that wafer fab usually takes several months and is the top priority for improvement. Therefore, cycle time reduction is always an important task in controlling a wafer fab factory. To become an agile supplier, shortening the cycle time of every operation is critical [51].

(2) Hybrid approaches combining fuzzy logic/clustering with ANN have been developed for different applications because of the un-deterministic characteristic factors that require fuzzy expressions, such as the release time, average fab utilization, total queue length on the processing route, and cycle time. Since they cannot be determined accurately, a certain probability distribution is needed. The fuzzy based DM approaches facilitate more realistic pattern extraction.

(3) The tasks realized by ensemble approaches combining fuzzy c-means (FCM) or SOM-based clustering with ANN-based prediction were pervasive. The purpose of clustering is to classify objects according to its similarity considering various features and therefore, improve the accuracy of prediction. The results show that the hybrid approaches with clustering-based pre-classification or post-classification are some of the most accurate approaches used to estimate the cycle/lead time or the complete date and obtain an optimization scheduling plan [51].

Table 10. DM with Big Data application for production management and optimization.

Functions (Frequency)	DMTs	Categories	Knowledge Area/Task	Indicator	Ref.
Prediction (7)	ANN	KD-oriented	Allocation of resource	MSE, MAPE	[43]
	Regression	SA-oriented	Cycle time Prediction	MAE, MSE	[44]
	FNN	KD-oriented		RMSE, MAE, MAPE, RMSE	[45,46]
	FNN	KD-oriented	Rescheduling	RMSE	[47]
	GNR, ANN	KD-oriented, SA-oriented	Cycle time Prediction	MAPE	[48]
	ANN	KD-oriented	Assembly times perdition	MSE, MAE, RSE, RAE	[49]
Classification (6)	FACRs (Apriori + Fuzzy logic)	KD-oriented	Scheduling	-	[50]
	FNN + ANN + Apriori	KD-oriented	Cycle time prediction	-	[51]
	DT, ANN	KD-oriented		ACC	[52]
	SNBC	SA-oriented		ACC	[53]
	SVM, RST, DT	KD-oriented	Human management	ACC	[54]
	GA + SVM	KD-oriented	Scheduling	-	[55]
Prediction Clustering (23)	FCM + FNN	KD-oriented, SA-oriented	Scheduling	-	[56]
				MAE, MAPE, RMSE	[57]
	SOM + FNN	KD-oriented		-	[58]
				RMSE, MAPE	[59]
	SOM + ANN	KD-oriented		DBI	[60]
	FCM + ANN	KD-oriented, SA-oriented		RMSE	[61]
	SOM+ FNN	KD-oriented		RMSE	[62]

Table 10. *Cont.*

Functions (Frequency)	DMTs	Categories	Knowledge Area/Task	Indicator	Ref.
Prediction Clustering (23)	FNN + ANN + Apriori	KD-oriented	Cycle time Prediction	RMSE, MAE, MAPE	[51]
	FCM + ANN	KD-oriented, SA-oriented		MAE, MSE	[63]
	FCM + FNN			MAE, MAPE, RMSE	[64,65]
	FCM + RBFNN, FNN			RMSE, MAE, MAPE	[66]
	FCM + ANN	KD-oriented		RMSE	[67,68]
	SOM + FNN			RMSE	[69]
	FCM + FNN	KD-oriented, SA-oriented	Output time prediction	RMSE	[70]
	FNN	KD-oriented	Cycle time prediction	RMSE	[71]
			Due date prediction	RMSE	[72]
	SOM + FNN		Cycle time prediction	RMSE	[73]
	SOM + ANN	KD-oriented	Output time prediction	RMSE	[74]
	K-means + FNN	SA-oriented, KD-oriented		RMSE	[75]
	SOM + FNN	KD-oriented	Completion time prediction	RMSE	[76]
				RMSE	[77]
Clustering (1)	SOM	KD-oriented	Scheduling	-	[78]
Association (2)	FACRs			-	[50]
	FNN + ANN + Apriori	KD-oriented	Cycle time Prediction	Support, confidence	[51]

FACRs: Fuzzy association classification rules; GNR: Gauss-Newton regression; RBFNN: Radial basis function neural network.

Tens of thousands of monitoring and online detection measurement values, and hundreds of electrical test parameters timely measured at different positions on a wafer during the fab process facilitates the Big Data application for production control. The typical knowledge area of these applications is AEC/APC that is a collection of tasks including FDC, R2R control, SPC, and VM to reduce the process variation and meet the process target for yield (quality) enhancement.

The related literature is summarized in Table 11 from which we can see that the outlier detection was conducted online and the time series analysis was employed for anomaly detection while the prediction function was mainly used for VM and R2R. Classification and clustering have been widely used for FDC. Some preprocessing like regression [87–89] was conducted to identify the main effects on observation variables before classification or clustering model establishment.

Table 11. DM with Big Data application for advanced equipment control/advanced process control.

Functions (Frequency)	DMTs	Categories	Knowledge Area/Task	Indicator	Ref.
Prediction (5)	Regression	SA-oriented	R2R	R^2	[79]
	PLS + MLR		FDC, R2R, VM	ACC	[80]
	SVR	KD-oriented	VM	RMSE	[81]
	DT, ANN, SVR		VM	MAE, RMSE, R^2	[82]
	SLR, GA+SVM	SA-oriented KD-oriented	VM	MSE	[83]

Table 11. *Cont.*

Functions (Frequency)	DMTs	Categories	Knowledge Area/Task	Indicator	Ref.
Classification (12)	SVM	KD-oriented	FDC	ROC	[84]
	Logistic regression	SA-oriented	FDC	TPR	[85]
	A framework	SA-oriented, KD-oriented	FDC, R2R, SPC, EHM	ACC	[86]
	Forward stepwise regression, LASSO, Random forest	SA-oriented	FDC	R^2	[87]
	MLR		FDC	ACC	[88]
	Stepwise regression, CART	SA-oriented, KD-oriented	FDC	ACC	[89]
	K-means, CHAID			ACC	[90]
	PCA, SOM, CHAID			TPR, TNR	[91]
	Multi-sensor-based trace segmentation, PCA	SA-oriented	FDC	ACC	[92]
	Spatial statistics, ANN	KD-oriented, SA-oriented	FDC, SPC	-	[93]
	SOM, K-means, DT		SPC	DBI	[94]
	CART	KD-oriented	FDC	-	[95]
Clustering (4)	K-means, CHAID	SA-oriented KD-oriented	FDC	ACC	[90]
	PCA, SOM, CHAID		FDC	-	[91]
	SOM, K-means, DT		SPC	DBI	[94]
	SDC		SPC	FAR, FRR	[96]
Time series analysis (3)	CART	KD-oriented	FD	-	[95]
	osPCA, online PCA, ABOD, LB-ABOD, LOF	SA-oriented	FD	PPV, TPR	[97]
	EBIT, CUSUM	SA-oriented	FD	TP, FN	[98]
Outlier detection (3)	osPCA, onlinePCA, BOD, LB-ABOD, LOF	SA-oriented	FD	-	[97]
	EBIT, CUSUM		FD	TP, FN	[98]
	PSLA		FD	-	[99]

SLR: Stepwise linear regression; PLS: Partial least squares regression; SDC: Segmentation, detection, and cluster-extraction; osPCA: Online oversampling PCA; ABOD: Angle based outlier detection; LB-ABOD: Lower bound-ABOD; LOF: Local outlier factor; EBIT: Entropy-based information theoretic; CUSUM: Cumulative sum; PSLA: Process sensor log analysis.

The data-driven mechanism is one of the pervasive approaches to FDC [114] and the summarization in Table 11 also indicates that FDC (or FD only) is the most researched task of AEC/APC [84–93,95,97–99]. The wafer fab is a complex and lengthy process that involves hundreds of process steps, and early FD gives engineers more time to perform appropriately to avoid serious equipment abnormalities [84–86] while fault classification can be considered as the combination of fault identification and diagnosis in order to identify the main effects on observation variables, concentrate on the process variables related to diagnosing abnormalities, and then to determine the cause of the observed out-of-control status that can facilitate the process recovery by removing the cause of the fault to reduce yield loss [86,90].

R2R control consists of several levels including real-time control, single-process R2R control, inter-process R2R control, and factory-level R2R [79]. FDC stands for a representative technique of real-time control. Single-process R2R control focuses on an individual process module while the selected R2R related articles concentrate mainly on inter-process R2R that deals with the process control of two or more inter-related process modules [80,86] combined with other tasks like FDC. The factory-level R2R has only been considered by a few research papers [79] that are used to enhance the results of electronic tests in wafer acceptance tests and yield circuit probe tests.

The reviewed VM-related literature utilized MLR [80], ANN [81], SVM(R), and DT [82,83] based on the production equipment data and preceding metrology results to predict every wafer's metrology measurements, which fills a lack of physical measurement by prediction that enables the measurement of every wafer for every process step on all capable equipment available in the fab, thus, allowing significant improvement of process control and product quality, reduction of operational cost, and production cycle time [81,83].

In SPC, significant characteristics are monitored such as the failure percentage of wafer bin maps [93] and the soldering quality [94]. The process control chart, as a widely used approach to SPC [93,94], has been used to diagnose and identify the variability of the fab process. The statistical process system can help detect defects that might originate from the process steps to improve quality and eliminate the need for expensive post inspections [94,96]. With increasing the demand for high-quality products and reliable processes, multivariate statistical process control (MSPC) has been developed to ensure that equipment is "statistically controlled" by monitoring two or more related quality characteristics simultaneously [105].

The above review indicates that AEC/APC conducts monitoring of online measurements of specific process steps, and undertakes corrective action to ensure that the parameter being measured remains within the desired limits. However, the integration of FDC, R2R, SPC, and VM has been considered only in a few research papers [86], requiring further research from different aspects such as the consistency and integration of data, unified frameworks, high-efficiency algorithms and platforms, and so on.

The application of DM with Big Data for quality improvement of electronic products, especially for wafer fab at the production stage was summarized in Table 12. One of the research papers deals with predicting the performance (yield) of a manufacturing process or system in terms of critical functional characteristics. Months may pass before a chip is completed; hence, there is a great interest in mining production data to predict its performance prior to the final testing of the wafers [100–108]. In order to infer to the possible causes of faults and manufacturing process variations in semiconductor manufacturing after the whole fab process is completed, the clustering, classification, and association analyses are conducted based on different DMTs such as k-means, SOM, SVM, and decision tree to identify critical poor yield factors and determine the root cause of low yield. On this basis, the related process parameters can be adjusted to ensure future quality based on post hoc diagnosis [110–118,121]. Some studies combined with sequential pattern mining to identify the sequence association events between different operations during the manufacturing [119,120].

Table 12. DM with Big Data application for the quality improvement at the production stage.

Functions (Frequency)	DMTs	Categories	Knowledge Area/Task	Indicator	Ref.
Prediction (9)	FNN	KD-oriented	Yield prediction	MAE, RMSE, MAPE	[100]
	Regression, ANN, K-means clustering	SA-oriented KD-oriented		R^2	[101]
	PLS, CART			MAPE	[102]
	Generalized linear mixed models	SA-oriented		MSE	[103]
	FNN	KD-oriented	Quality prediction	MAE, RMSE, MAPE	[104]
	FCM + GA + DT			ME	[105]
	Fuzzy linear regression + BPN	SA-oriented, KD-oriented	Cost prediction	RMSE, MAE, MAPE	[106]
	Regression, ANN		Yield prediction	MAE, MAPE, MRSE, MAPE, R^2	[107]
	FNN	KD-oriented	Yield prediction	MAE, RMSE, MAPE	[108]

Table 12. *Cont.*

Functions (Frequency)	DMTs	Categories	Knowledge Area/Task	Indicator	Ref.
Classification (11)	SOM, K-means, DT	SA-oriented KD-oriented	Quality classification	DBI	[94]
	NB, SVM, ANN			ACC	[109]
	CHAID	SA-oriented		TPR TNR, FPR, ACC	[110]
	SVM	KD-oriented		TPR, FPR	[111]
	PCA, SVM, adaptive boosting, DT	SA-oriented, KD-oriented	diagnosis Post hoc diagnosis	FN, FP	[112]
	SVM	KD-oriented		TPR, FPR	[113]
	Statistical model	SA-oriented		-	[114]
	CART	KD-oriented		TPR, FPR	[115]
	SVM			-	[116]
	K-means, DT	SA-oriented, KD-oriented		-	[117]
	Spatial statistics + adaptive ANN, DT			-	[118]
Clustering (6)	SOM, K-means, DT	SA-oriented, KD-oriented	Quality classification	DBI	[94]
	SDC		Post hoc diagnosis	FAR, FRR	[96]
	Regression, ANN, K-means, clustering		Yield prediction	-	[101]
	FCM + GA + DT		Quality prediction	ME	[105]
	K-means, DT		Post hoc diagnosis	-	[117]
	Spatial statistics + adaptive ANN, DT			-	[118]
Association (Sequence Analysis) (3)	Association rule tree	SA-oriented, KD-oriented	Yield prediction	MAPE	[102]
	Bayesian network, PLS, Apriori		Post hoc diagnosis	ACC	[119]
	Decision correlation rules	KD-oriented		-	[120]
Time series analysis (1)	Co-clustering	SA-oriented	Quality prediction	RMSE	[121]
Outlier detection (1)	Hierarchical clustering, DT	SA-oriented, KD-oriented	Post hoc diagnosis	-	[122]

Moreover, more than hundred test items and millions of rows of data for wafers will be generated after testing, per day. According to the basic requirements of quality management, an essential work is to analyze these test items one by one according to different specifications and requirements. In accordance with the traditional mode of work, more than a hundred process capability indexes should be calculated step by step and the quality characteristics should be evaluated one by one with enormous and complicated operations. Meanwhile, it is difficult to determine the association between these indexes and present a comprehensive summary of the overall performance of the product. The application of Big Data for the quality management and analysis can easily generate a traditional single index process capability analysis report. More importantly, it can excavate many new results from the Big Data set [114].

4.2.3. Application of DM and Big Data for Sale, Service, and Recycling

The stage of sale, service, and recycling (SSR) is to store produced products in a warehouse and transport them to customers in logistics, and then the customers use the product while a manufacturer provides remote service. If it can no longer be used, it comes to the end of its life such as remanufacturing and disposal [8].

The summarization of DM application in the SSR stage is given in Table 13 and it can be seen that most of the applications related to CRM involve marketing and sales prediction [125–127], customer

service [129], and the SCM to achieve greater efficiencies and effectiveness in delivering customer value [130]. The detailed information indicates that one direction of the research is to mine the behavioral characteristics of customers on the product and maintenance, and therefore, identify customer's requirement for customer attraction and retention [123,129]. Another one is to predict the marketing demand and price for customer identification and development, and therefore, to facilitate the plan optimization of production, procurement, and resource [125–127]. Only one article is related to the recycling of electronic products considering the storage behavior of customers [124]. From Table 13, it can also be seen that the prediction of marketing requirement and determination of a more reasonable price are the main functions while the clustering and classification have been taken to classify products and customer's requirement and identify the purchase feature of different customers. Text mining was utilized to excavate the knowledge from interaction records in some cases [123].

Table 13. DM and Big Data application in the sale, service and recycling stage.

Functions (Frequency)	DMTs	Categories	Knowledge Area/Task	Product	Indicator	Ref.
Prediction (6)	Text mining + Regression	SA-oriented, KD-oriented	Purchase decisions prediction	iPhone, Mac, iPod, iPad and so forth. and components	-	[123]
	SVM	KD-oriented	Behavior prediction	Used hard disk	MAPE, MAE, MSE	[124]
	SVR + Bat	KD-oriented	Marketing and sale trends prediction	PCB	MAPE, RMSE	[125]
	K-means, SOM, FNN	SA-oriented, KD-oriented			MAPE, RMSE	[126]
	Fuzzy CBR	KD-oriented			MAPE, RMSE	[127]
	Weighted evolving FNN	KD-oriented			MAPE, MAE, RMSE	[128]
Classification (1)	Text mining + Regression	SA-oriented	Repair experience extraction	iPhone, Mac, iPod, iPad and so forth. and components	PPV, TPR	[123]
Clustering (2)	K-means	SA-oriented	Repaired products clustering	Camera, laptop, phone, printer, and so forth.	-	[129]
	K-means, SOM, FNN	SA-oriented, KD-oriented	Marketing and sale trends prediction	PCB	-	[126]
Time series analysis (1)	Nonlinear least square	SA-oriented	Demand prediction	Semiconductor	MAPE, R^2	[130]

4.2.4. Summarization of DM with Big Data Application in Different Stages

Figure 4 illustrates different functions used by the selected articles applied in different stages. It can be seen that the prediction, classification, and clustering functions are the top three functions employed for mining patterns at different stages. The six functions have been used in the production stage which indicates that there are diverse requirements of DM and Big Data application at this stage for different purposes, while the time series analysis and outlier detection function have seldom been used in the stage of design and SSR.

Figure 5 illustrates the distribution of different knowledge areas considering the tasks of QI for design/production, DTCE, PTP, FDC, VM, R2R, SPC, CRM, and SCM according to Tables 9–13. The frequency in Figure 5 indicates that the QI for design, scheduling optimization, production time prediction, FDC, post hoc diagnosis, production yield/quality prediction, and optimization of sale and service for CRM are pervasive knowledge areas and tasks.

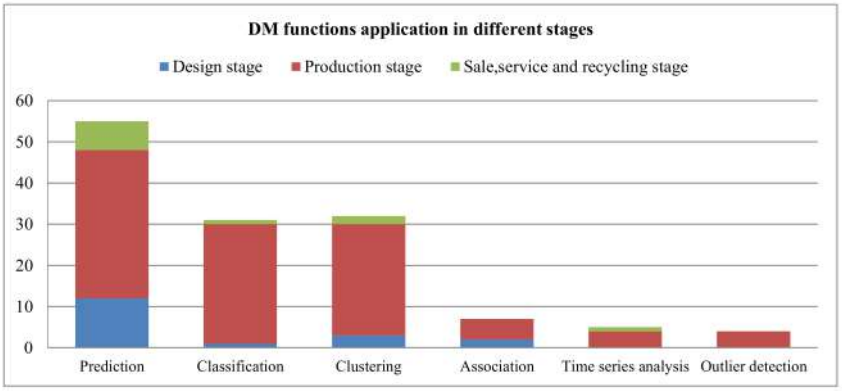


Figure 4. DM function applications in different stages.

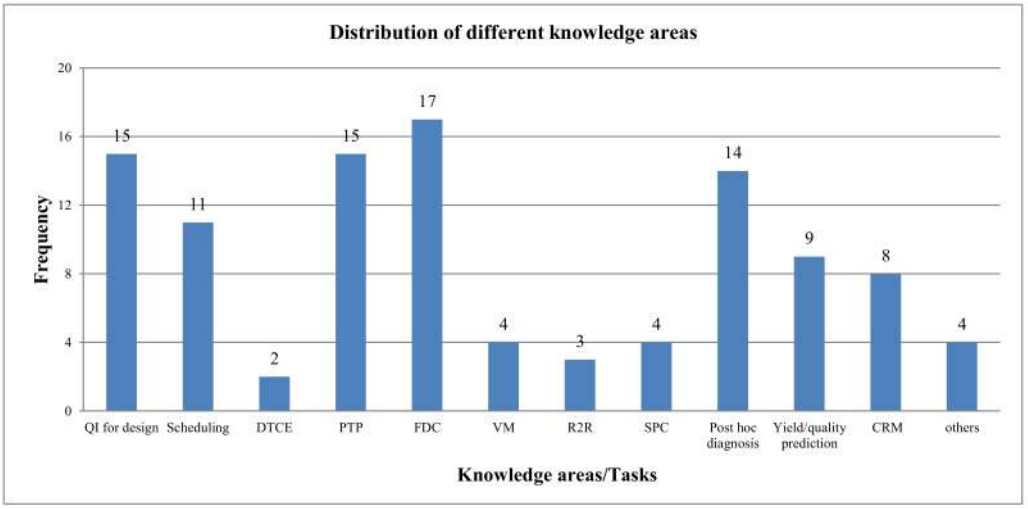
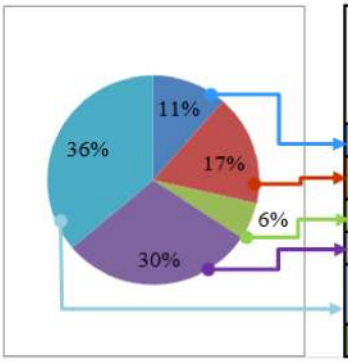


Figure 5. Distribution of different knowledge areas.

The statistic of different categories of DMTs adopted in the 105 articles for different knowledge areas are conducted and the results are illustrated in Figure 6 from which we can see that the pervasively used DMTs are hybrids or integrations of the SA-oriented and KD-oriented DMTs, especially for the knowledge areas of PMO, AEC/APC, and QI for production, followed by the combination of different KD-oriented DMTs or only one KD-oriented DMT. However, only one SA-oriented approach and the ensemble of SA-oriented DMTs have been widely adopted by researchers compared to other approaches.



Categories of DMTs	Knowledge areas						Sum
	QI for design	DTCE	PMO	AEC/ APC	QI for production	CRM& SCM	
Only one SA-oriented DMTs	1	0	2	4	3	2	12
Only one KD-oriented DMTs	7	0	3	3	4	1	18
Multiple SA-oriented DMTs	0	0	0	5	1	0	6
Multiple KD-oriented DMTs	6	0	16	2	4	3	31
Both SA-oriented and KD-oriented DMTs	1	2	11	8	14	2	38
Sum	15	2	32	22	26	8	105

Figure 6. Different categories of DMTs for different knowledge areas (DMTs: data mining techniques).

The commonly used 10 DMTs including ANN (back propagation neural network, fuzzy neural network, and so forth), fuzzy logic, DT (CART, CHAID, C5.0, and so forth), regression (MLR, MRA, stepwise regression, PLSR, logistic regression, and so forth), SVM (SVR, ASVR, and so forth), SOM, FCM, K-means, GA, and Apriori are given in Figure 7. Figure 8 presents different DMTs in different knowledge areas.

It can be seen that the top DMT used is ANN followed by fuzzy logic because many ANNs are combined with fuzzy logic to solve the scheduling optimization and production time prediction. ANN has been applied in eight areas of the above-mentioned knowledge areas except for SPC and R2R. Fuzzy logic has been used mainly for the production time prediction, yield/quality prediction, and the optimization of CRM/SCM. The DT has been widely employed for FDC and post hoc diagnosis. The regression has been pervasively used for feature selection and prediction of quality, yield, development cost, VM, and so on. The SOM, K-means, and FCM have been used for clustering, especially for the pre-classification of jobs while conducting scheduling optimization and production time prediction. GA has been used to find optimal levels of process/product parameters [26,27,31–36], which can also be used to optimize parameters of DMTs such as SVM [55,83] and fuzzy clustering [105].

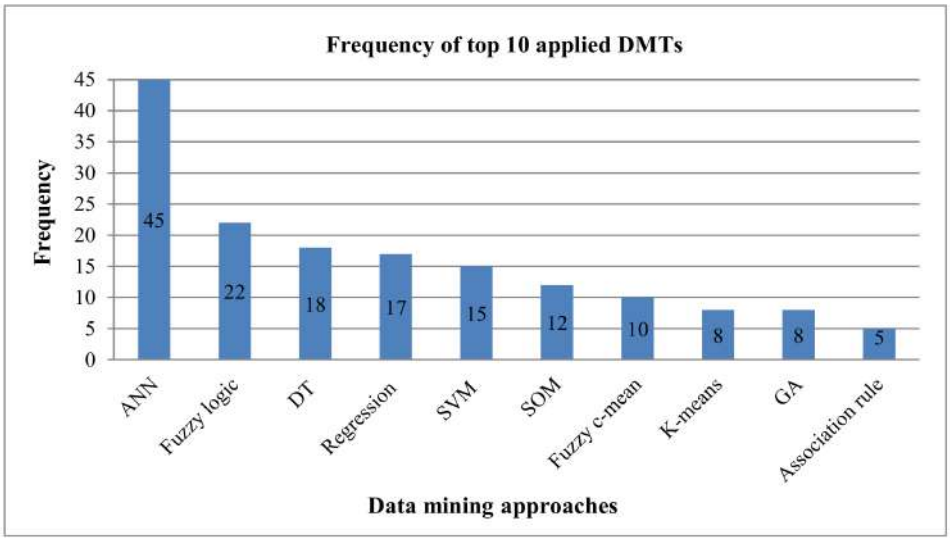
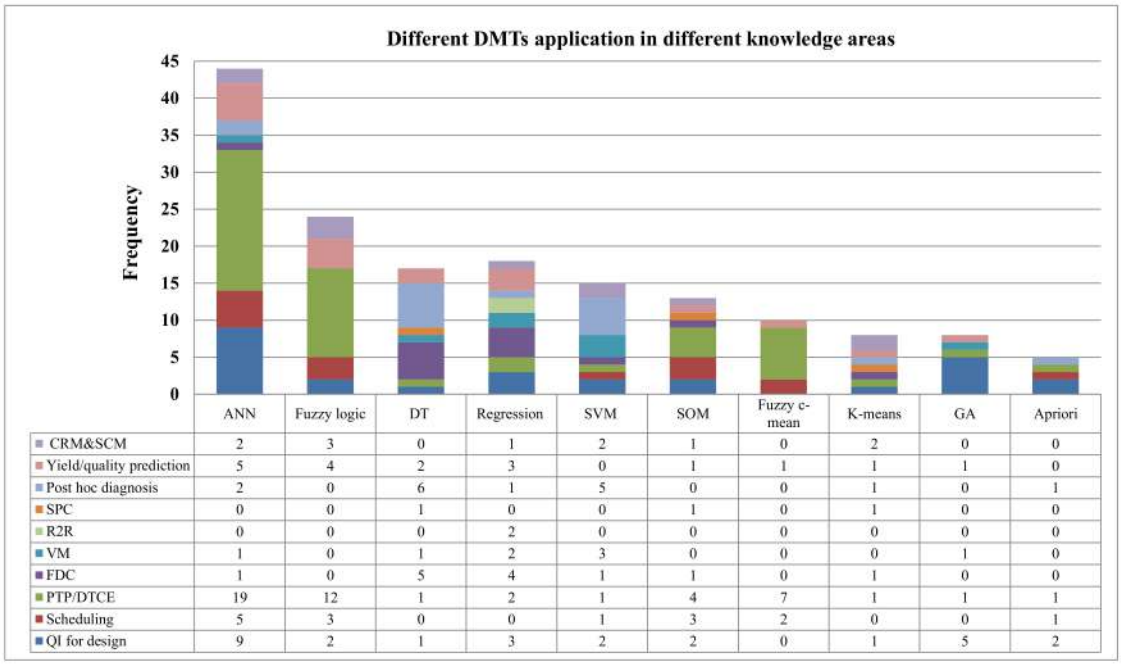


Figure 7. The frequency of the top 10 applied DMTs.



	ANN	Fuzzy logic	DT	Regression	SVM	SOM	Fuzzy c-mean	K-means	GA	Apriori
CRM&SCM	2	3	0	1	2	1	0	2	0	0
Yield/quality prediction	5	4	2	3	0	1	1	1	1	0
Post hoc diagnosis	2	0	6	1	5	0	0	1	0	1
SPC	0	0	1	0	0	1	0	1	0	0
R2R	0	0	0	2	0	0	0	0	0	0
VM	1	0	1	2	3	0	0	0	1	0
FDC	1	0	5	4	1	1	0	1	0	0
PTP/DTCE	19	12	1	2	1	4	7	1	1	1
Scheduling	5	3	0	0	1	3	2	0	0	1
QI for design	9	2	1	3	2	2	0	1	5	2

Figure 8. The different DMTs application in different knowledge areas.

4.3. Software Used for the Selected Articles

Many algorithm engines, tools, and platforms have been developed to implement functions and related DMTs. Predictive analytics today summarized the top 50 free DM software [131], including Orange, RapidMiner, Weka, KNIME, SpagoBI, Anaconda, Octave, and so forth. Some commercial software including Sisense, Oracle Data Mining, Microsoft SharePoint, IBM Cognos, Dundas BI, SAP Business Objects, Matlab, Statistic, SAS EM, SPSS Clementine (IBM SPSS Modeler after 2009), Tanagra, Qlik Sense, and so forth have also been widely used by researchers and practitioners.

The different tools shown in Table 14 have been used for various purposes in mining applications from the reviewed literature. The category of software in the reviewed articles can be categorized into spreadsheets, statistical software package, DM software package, general purpose software, special purpose tools, and high-level languages.

Some statistical software packages such as MiniTab, SAS, SPSS, and Statistics were preferred for implementing SA-oriented methods such as MRA and ANOVA. Spreadsheet-application excel was mainly used for data preparation and preprocessing. However, commercial software packages such as SPSS Modeler, SAS Enterprise Miner (SAS EM), were only used in a few of the applications.

The general purpose software Matlab and special purpose packages based on Matlab were used in various applications for the design and production of QI, PMO, and CRM. They were mostly utilized to realize ANN, fuzzy logic, SVM, and SOM supported by several open source toolboxes such as NeuroSolutions, Neural Network, NeuralPower, Fuzzy Logic, LibSVM, and SOM. The association, outlier detection, and time series analysis functions were mainly conducted by commercial software packages such as SAS EM [38], and RapidMiner [42].

Some high-level languages such as C/C++ [94,120] and Visual Basic [70–73,75–77] were used for SOM, fuzzy c-means, fuzzy logic, ANN, and the combination of these approaches for its flexibility for researcher to design or combine particular methodologies considering domain knowledge in handling and analyzing the data. Meanwhile, some platforms such as the online system [79], fab-wide FDC [80], VM system [83], online time series prediction system [88], and wafer bin of map clustering and classification systems [117] have been developed for different tasks of AEC/APC based on high-level languages. However, the commonly used platforms for developing DM or Big Data application system such as WEKA [28], RapidMiner [42], R software environment [122], and Python [84] have

been utilized by only a few of the researchers, indicating that the systematized applications of these results still require further development by practitioners.

Table 14. The software used for accomplishing DM and Big Data application in the electronics industry.

Type of Software	Name of Software	Reference	Usage
Spreadsheet application	Excel	[34,52,104,117]	Data preparation and data preprocessing
Statistical software package	MiniTab	[107,127]	Regression prediction
	SAS	[38,117,125]	SA-oriented methods such as MRA, ANOVA, and PCA
	SPSS	[34,126]	
	Statistics	[34,117,126]	
DM software package	SPSS Clementine	[38,52]	Preprocessing, prediction, classification, clustering, and association
	SAS EM	[38]	
General purpose software	Matlab	[26,33,35–37,45,47,55,70,78,82,104]	Prediction, classification, clustering, and optimization
	WEKA	[28]	Prediction
	Visual Mining Studio	[40]	Clustering
	RapidMiner	[42]	Association
	R software environment	[122]	Outlier detection
Special purpose tools	BrainMaker	[107]	ANN for prediction, classification, clustering, and so forth
	NeuralWorks Professional II/Plus	[127]	
	NeuroSolutions	[70,72,74,76,77]	
	Neural Network Toolbox	[45,57,58,64,67]	
	NeuralTools	[34]	
	Netlab Toolbox	[49]	
	NeuralPower	[27,31]	
	Fuzzy Logic Toolbox	[45,57,64]	Fuzzy logic
	LibSVM	[84,113,125]	SVM
	SOM toolbox	[39,41,60,78]	Clustering
	Lingo	[45,68,108]	Optimization
High-level language	C/C++	[94,120]	Various purposes such as SOM, fuzzy clustering, FBPN, and so forth
	Visual Basic	[59,62,70–73,75–77]	
	Python	[84]	Outlier detection

5. Diagram of Data Content for Different Knowledge Areas and DM Framework for the Electronics Industry

The product lifecycle processes carry a huge number of structured, semi-structured, and unstructured data. Big Data analytics and DM technology can be used to make a deep analysis of historical lifecycle data, to discover knowledge, and to optimize the process of PLM. A framework with four modules including data sensing and acquisition, data processing and storage, DM model development, and Big Data application in PLM was presented by Zhang et al. [1]. However, the summarization and classification of lifecycle related data and its utilization by different knowledge areas have not been discussed. Meanwhile, the special application scheme for electronics manufacturing has not been considered. Therefore, the establishment of a diagram of data content for different knowledge areas and DM with Big Data framework for the electronics industry can guide companies to accumulate related data and develop DM strategy from the view of lifecycle and overall business chain, which can also facilitate researchers and practitioners to select appropriate techniques and better utilization of data for knowledge discovery.

5.1. Diagram of Data Content for Different Knowledge Areas

From the view of electronics lifecycle, the main data for different knowledge areas can be divided into engineering data, enterprise resource and environment data, production plan and arrangement data, manufacturing result data, and transaction and interaction related data. Figure 9 illustrates the main content of each category and its application for different knowledge areas. The detailed description of each category is given as follows.

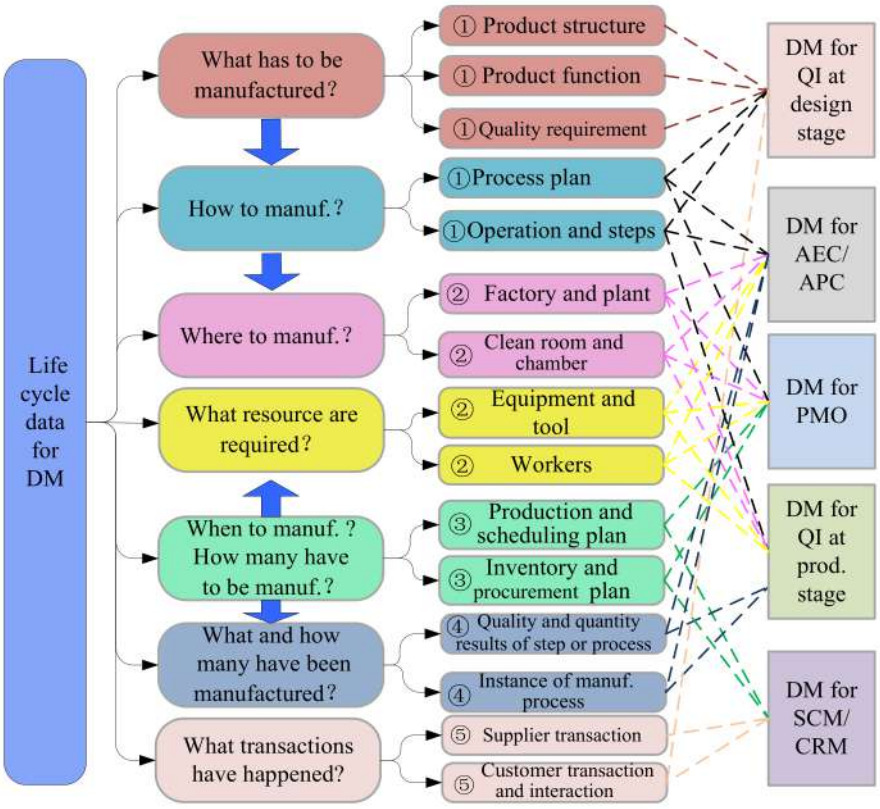


Figure 9. The data content for different knowledge areas.

① Engineering data: It includes product structure and function, manufacturing process plans, and quality requirements to define what is to be manufactured and how to manufacture. Relevant DM with Big Data applications have been conducted to improve product quality and customers' satisfaction or to optimize process parameters. This data can be stored in different systems such as PLM and computer-aided process planning system with structured (bill of materials), semi-structured (requirement reports), and unstructured (design model or drawing) styles.

② Enterprise resource and environment data: Resource data relates to the workplace, equipment, and tools that specify where and what resource are required to manufacture the product, which also includes data on process statuses, collected in real time by smart sensors and the traced data based on RFID placed on transportation robots. In common, these data are structurally stored in ERP, MES, and DCS that can be used for the optimization of process control such as AEC/APC and production management. Taking an example from wafer fab, the equipment status data such as chamber pressure, gases flow, and chuck temperature are collected in real time by sensors placed on tools, and valuable data that are generated from clean room environment monitoring [101].

③ Production plan and arrangement data: These data include the plan of the project, the hierarchy production plan, the inventory/material and procurement plan, and scheduling that defines when and how many products have to be manufactured. Different plans can be stored in ERP and MES with

a structured style, which has been widely used for the optimization of PMO tasks such as production time prediction and scheduling optimization at the production stage.

④ Manufacturing result records: Result records define the quality and quantity of products at a certain time and workplace. They are always accumulated in MES, quality management system, ERP, and storage management system with a structured style. RFID has been widely used for product lifecycle management in recent years, and the traced data generated automatically at different stages through RFID placed on materials, semi-products, and finished products can also be taken as the data of manufacturing result. Taking an example from the data involved in the wafer fab, it is generated at various steps including inline through metrology steps that measure test wafers and product wafers such as parameters of critical dimension, film thicknesses, film resistances, and so forth. It also includes electrical test and final yield data. DM-based post hoc diagnosis, yield prediction, and parameters adjustment are used to ensure the future quality has been conducted based on different steps of the result. They can also be combined with enterprise resource and environment data for AEC/APC.

⑤ Interaction and transaction data: Owing to the fast development of online trading and electronic commerce in the past decades, a large amount of records related to transactions and online interactions between upper stream supplier, middle collaborator, downstream customer have been accumulated. The structured transaction data, semi-structured or unstructured interactions have been widely used for the optimization of SCM and CRM such as marketing analyses and product design improvement based on the feedback from customers at the design stage, procurement and inventory optimization at the production stage, price and demand prediction, customer identification, attraction, retention, and development at the SSR stage. Text mining techniques have also been used to excavate the pattern from the interaction text and were combined with DM for the final knowledge discovery [123]. Meanwhile, RFID-based records can be used for product tracing in transaction, service, and recycling.

5.2. Data Mining with Big Data Frameworks for the Electronics Industry

On the basis of the aforementioned review, a framework of DM with Big Data applications in the electronics industry is presented in Figure 10 in which the stage of design and production corresponds to the beginning of lifecycle, and the sale and service can be taken as the middle of lifecycle, while recycling is at the end of lifecycle, respectively [1,8].

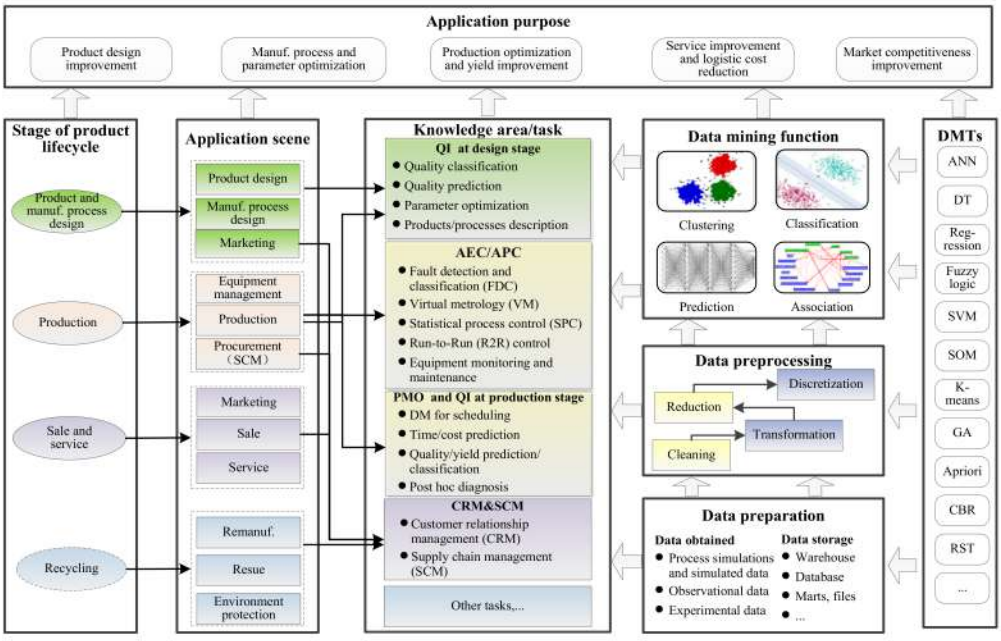


Figure 10. The framework of DM with Big Data towards IT applications in the electronics industry.

Each stage of the lifecycle corresponds to different application scenes. The DM application of product and manufacturing process design mainly includes the product design, manufacturing process design, and marketing with relevant knowledge areas, such as quality improvement, development cost and time prediction, product customization, manufacturing parameter optimization, and SCM. The equipment management, production management, and procurement are the main application areas of DM with Big Data in the production stage with typical knowledge areas such as AEC/APC, PMO, QI, and SCM. The application of the DM with Big Data for sale and service cannot only be support for quality improvement and customization design but also optimize logistics and facilitate customer service and maintenance. The recycling attracted less attention from DM with Big Data application in the electronics industry, which could be used in remanufacturing, reuse, and environment protection, considering the knowledge areas of product recovery, remaining life prediction, and reverse logistics optimization.

The details of knowledge areas of different stages have been summarized in Section 4.2. The quality improvement for design and production can be further divided into quality (yield) prediction, classification, description, and parameter optimization. Post hoc diagnosis can be taken as the quality description at the production stage with the purpose of process parameters adjustment to ensure future quality. The tasks of AEC/APC that consists of FDC, R2R control, SPC, VM, and so forth are also for quality enhancement, and therefore, the quality improvement at the production stage and AEC/APC are not a disjoint division here. The DM and Big Data application in PMO is a collection of scheduling optimizations, cost/time prediction, and so on.

SCM is used to optimize the logistics for material supply at the beginning stage of the lifecycle and it can also be used to achieve greater efficiencies and effectiveness in delivering customer value at the end of the lifecycle. The application of DM or Big Data tools in CRM is an emerging trend in the global economy. Analyzing and understanding customer behaviors and characteristics is the foundation of the development of a competitive CRM strategy so as to acquire and retain potential customers and maximize customer value [3]. The tasks of customer identification, attraction, retention, and development of CRM can be realized through Big Data-based marketing prediction, personalized service, predictive maintenance, remote online diagnosis and so on.

Data preparation such as data acquisition, accumulation, and storage for different knowledge areas and applications can be guided by the diagram of data content for different knowledge areas given in Section 5.1. The commonly used data preprocessing techniques including data cleaning, transformation, reduction, and discretization that can utilize the preprocessing approaches summarized in Sections 2.2 and 4.1, based on the requirement of application areas and the quality of data. DM, in a narrow sense, for each function, can be implemented based on some pervasive DMTs summarized in Section 4.2.4. The interpretation, evaluation, and implementation software can be conducted by combing experts' knowledge with performance indicators given in Section 2.4, which is not given in the framework because it has many selections in practice. The final purpose of the DM application has been proved by many researchers and practitioners. This framework provides an option for different types of companies and expects for further extension.

6. Conclusions and Future Research Directions

This paper presents a comprehensive review of DM with Big Data towards its applications in the electronics industry. We can see that the DM with Big Data has been applied to different scenes including product design improvement, manufacturing process optimization, PMO, production process monitoring and control, quality improvement, CRM, and so forth.

Customer-oriented product development and process plan optimization are the main applications for product design improvement and manufacturing process optimization in the stage of design. Prediction was the most frequently used DM function observed in the reviewed articles. ANN and regression were the widely used DMTs for the prediction.

The application of DM with Big Data for process monitoring and control, PMO, and quality improvement in the stage of production has attracted the interest of most research. On the one hand, sophisticated DM and Big Data related techniques such as FDC and R2R have been developed for the wafer production process monitoring and control to reduce defects and improve the quality/yield based on the data collected from manufacturing processes, equipment/tool/environment statuses, and process parameters. The functions of classification and clustering were widely used for FDC based on related DMTs such as DT, SVM and ANN, k-means and SOM, while the prediction function was widely presented for VM based on ANN, regression, and SVM. On the other hand, prediction, clustering, and the combination of the two are the most frequently employed functions for the optimizing scheduling plan and prediction of cycle time/due date based on ANN, FCM, SOM, and a hybrid of fuzzy logic and ANN. Additionally, post hoc diagnosis, quality prediction, and classification were conducted based on the functions of prediction, classification, clustering, and association for future production quality improvement.

Most of the DM applications are related to CRM at the stage of SSR for the purpose of acquiring and retaining potential customers and maximizing customer value based on the records of transaction and online feedback from customers. Prediction, classification, clustering, and time series analysis functions were conducted based on ANN, regression, and SVM for sale and service to mine the consumption habits and predict the marketing price.

The achievement of the reviewed articles facilitates theoretical study and practical application of DM with Big Data to the electronics industry. Nevertheless, the limitation and challenges still exist for future research.

(1) Data preparation and preprocessing. The data of the product lifecycle are characterized by multisource (for example, design, production, and service data), heterogeneity (for example, structured, semi-structured, and unstructured data), and "noise" (for example, incomplete, incorrect, redundant, and inconsistent data) [1]. These problems increase the difficulties of data preparation, preprocessing, and subsequent mining, and also generate misleading patterns. However, little effort has been devoted to handling these problems. Manufacturing organizations with well-established and integrated data collection systems would benefit from a larger application of DM and Big Data [4]. Unified management and storage of the multi-source and heterogeneous data are necessary, and this motivates enterprises to develop DM strategies with dedicated consideration to data accumulation, integration, and consistency. Multi-business requirements integration, concept standardization, unified model establishment, and data/system interface development should be conducted collaboratively to facilitate data utilization. The standardization of operations such as data entry, storage, and maintenance should also be conducted accordingly to ensure the data quality and reduce data redundancy.

(2) The knowledge area of DM application. DM has been widely used in the stage of design and production especially for wafer fab and PCB assembly, and the pervasive knowledge areas include QI, PMO, AEC/APC, and so forth. However, potential applications such as customization production, procurement, warehouse management and inventory balance, and equipment maintenance and repair require more relevant data accumulation and extended mining. The global logistics industry has a large ever-growing amount of Big Data and is flooded with real-time data ranging from smartphones, sensors, and digital machines [9]. However, the application of DM with Big Data in SCM and logistics for electronic products has attracted few special discussions. Meanwhile, little effort has been put on CRM and order management combining the features of electronics such as a large amount of consumers, fast replacement of new products, and fierce market competition.

The patterns and knowledge hidden in Big Data are multidimensional (for example, various departments and lifecycle stages) and scattered, which hinders the effective mining and utilization of the knowledge. Therefore, further studies can be conducted to mine consumer habits and market characteristics to support more reasonable decision for customization product development, market pricing, and maintenance based on the association, prediction, and time series analysis functions. The fast upgrading of electronic products resulting in a large number of e-waste and the use of DM and

Big Data to improve the efficiency and effectiveness of its energy saving, recycling, reverse logistics, and reduction of environmental risks are a worthwhile attempt. More importantly, the macro strategy for integrated mining and integration applications for the whole lifecycle should be considered and developed by enterprises.

(3) DM functions and DMTs. The prediction, classification, and clustering are the most frequently used DM functions while the other three functions (outlier detection, association and time series analysis) have been used only in a few situations. The extended investigation of outlier detection, sequential pattern mining, and time series analysis considering time information for online model development and updating could enable companies to respond promptly to dynamic and emerging situations. For DMTs, the parameter optimization of DMTs, such as ANN and SVM, requires continuous study. While FCM and fuzzy logic have been combined with ANN to handle uncertainty, they might be combined with other related mechanisms such as SVM and regression. Additionally, these approaches would handle Big Data with easy implementation and high performance, and more deliberate consideration for industrial applications is required.

(4) Algorithm performance. In general, it is difficult to obtain results with obviously competitive advantage in the existing single algorithm. Generally, a hybrid mining algorithm needs to be constructed based on the characteristics of the problem by integrating different functions and different DMTs so as to ensure the validity and advantage of the algorithm. How to set and optimize algorithm parameters, such as parameters of ANN and SVM, also remains to be further studied. Meanwhile, how to evaluate the advantages and disadvantages of the developed algorithm dynamically and ensure the robustness of the algorithm under certain data loss and redundancy needs to be further compared. How to evaluate the under-fitting and overfitting of algorithms and balance of the two has been paid less attention and requires further consideration.

(5) Software and implementation: Many researchers employed special purpose tools, such as NeuroSolutions, Neural Network Toolbox, LibSVM, Fuzzy Logic Toolbox, and SOM toolbox to implement the developed algorithms. Meanwhile, many approaches were developed by Matlab. A dedicated software package and Matlab integration of the basic engine allowed researchers to implement the proposed algorithm and verify the results more easily. The FDC was always conducted based on online analysis related platforms that were developed independently because of its high-efficiency requirements for data preprocessing and algorithm execution. However, application-oriented software platforms, such as Orange, IBM SPSS modeler, WEKA, and RapidMiner were employed only by few researchers in the reviewed articles. In order to strengthen the connection between enterprises and research, one of the important directions is to directly develop the application platform and then, to validate and optimize the results through practical feedback. In addition, DM technology should be combined with data management and visualization tools that can facilitate user understanding, operating, and utilizing data efficiently.

(6) Knowledge maintenance and updating. Most of the mining was conducted statically and the corresponding data handling was conducted based on batch data. These approaches were difficult to learn by themselves and the patterns obtained were often difficult to update dynamically based on newly accumulated data. Nowadays, data is generated continuously and typically sent in the data records simultaneously and in small sizes. This data needs to be processed sequentially and incrementally on a record-by-record basis or over sliding time windows and also used for a wide variety of analytics and mining. Online mining and learning will be an important challenge for further research.

Acknowledgments: This paper is supported by the National Natural Science Foundation of China (Grant No. 51605169) and Natural Science Foundation of Guangdong, China (Grant No. 2014A030310345). This study also supported by the State Scholarship Fund of China (Grant No. 201608440414).

Author Contributions: Shengping Lv wrote the paper. Hoyeol Kim edited the paper and improved the quality of the article. Binbin Zheng conducted literature retrieval and statistics. Hong Jin proposed the paper structure and wrote the Sections 5 and 6 of the paper.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Abbreviations

ABOD	Angle based outlier detection	MRA	Multiple regression analysis
ACC	Accuracy	MRO	Multi-response optimizations
AEC/APC	Advanced equipment control/advanced process control	MSE	Mean squared error
ANOVA	Analysis of variance;	NB	Naive Bayesian
ASVR	Adaptive support vector regression	NMI	Normalized mutual information
AIC	Akaike information criterion	NPV	Negative predictive value
CART	Classification and regression tree	OLS	Ordinary least square
CBR	Case-based reasoning	OMS	Order management system
CHAID	Chi-squared automatic interaction detection	onlinePCA	Online PCA
CRM	Customer relationship management	osPCA	Online oversampling PCA
CUSUM	Cumulative sum	PCA	Principle component analysis
DBI	Davies–Bouldin index	PCB	Printed circuit board
DCS	Distributed control system	PLM	Product lifecycle management
DI	Dunn index	PLS	Partial least square regression
DTCE	development time/cost estimation	PMO	Production management and optimization
DM	Data mining	PSLA	Process sensor log analysis
DMTs	Data mining techniques	PTP	Production time prediction
DOR	Diagnostic odds ratio	QE	Quantisation error
DT	Decision tree	QI	Quality improvement
EBIT	Entropy based information theoretic	RAE	Root absolute error
ERP	Enterprise resource planning	RBFNN	Radial basis function neural network
FACRs	Fuzzy association classification rules	RE	Relative error
FCM	Fuzzy c-means	RFID	Radio frequency identification
FD	Fault detection	RI	Rand index
FDC	Fault detection and classification	RMSE	Squared root of mean squared error
FDR	False discovery rate	ROC	Receiver operating characteristic curve
FN	False negative	RST	Rough set theory
FNN	Fuzzy neural network	R2R	Run to run
FOR	False omission rate	SCM	Supply chain management
FP	False positive	SDC	Segmentation, detection, and cluster-extraction
FPR	False positive rate	SLR	Stepwise linear regression
PPV	Positive predictive value	SNBC	Selective naive Bayesian classifier
GA	Genetic algorithm	SOM	Self-organizing map
GNR	Gauss-Newton regression	SPC	Statistical process control
IA	Index of agreement	SPP	Stencil printing process
KDD	Knowledge discovery in databases	SSR	Sale, service and recycling
LASSO	Least absolute shrinkage and selection operator	SVM	Support vector machine
LB-ABOD	Lower bound-ABOD	SVR	Support vector regression
LOF	Local outlier factor	TE	Topographic error
MAPE	Mean absolute percentage error	TN	True negative
ME	Mean error	TNR	True negative rate
MES	Manufacturing execution system	TP	True positive
MLR	Multiple linear regression	TPR	True positive rate
m-PRLM	Missing values-patient rule induction method	VARER	Variance of errors

References

1. Zhang, Y.; Ren, S.; Liu, Y.; Sakao, T.; Huisingh, D. A framework for Big Data driven product lifecycle management. *J. Cleaner Prod.* **2017**, *159*, 229–240. [CrossRef]
2. Choudhary, A.K.; Tiwari, M.K.; Harding, J.A. Data mining in manufacturing a review based on the kind of knowledge. *J. Intell. Manuf.* **2009**, *20*, 501–521. [CrossRef]
3. Ngai, E.W.T.; Xiu, L.; Chau, D.C.K. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Syst. Appl.* **2009**, *36*, 2592–2602. [CrossRef]
4. Koksal, G.; Batmaz, I.; Testik, M.C. A review of data mining applications for quality improvement in manufacturing industry. *Expert Syst. Appl.* **2011**, *38*, 13448–13467. [CrossRef]
5. Liao, S.H.; Chu, P.; Hsiao, P.Y. Data mining techniques and applications-A decade review from 2000 to 2011. *Expert Syst. Appl.* **2012**, *39*, 11303–11311. [CrossRef]
6. Rostami, H.; Dantan, J.Y.; Homri, L. Review of data mining applications for quality assessment in manufacturing industry: Support vector machines. *Int. J. Metrol. Qual. Eng.* **2015**, *6*, 1–18. [CrossRef]
7. Donovan, P.O.; Leahy, K.; Bruton, K.; O'Sullivan, D.T.J. Big data in manufacturing: A systematic mapping study. *J. Big Data* **2015**, *2*, 2–22.
8. Li, J.; Tao, F.; Cheng, Y.; Zhao, L.J. Big Data in product lifecycle management. *Int. J. Adv. Manuf. Technol.* **2015**, *81*, 667–684. [CrossRef]
9. Zhong, R.Y.; Newman, S.T.; Huang, G.Q.; Lan, S.L. Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Comp. Ind. Eng.* **2016**, *101*, 572–591. [CrossRef]
10. Nagorny, K.; Lima-Monteiro, P.; Barata, J.; Colombo, A.W. Big data analysis in smart manufacturing: A Review. *Int. J. Commun. Netw. Syst. Sci.* **2017**, *10*, 31–58. [CrossRef]
11. Cheng, Y.; Chen, K.; Sun, H.M.; Zhang, Y.P.; Tao, F. Data and knowledge mining with big data towards smart production. *J. Ind. Inform. Integr.* **2017**, *9*, 1–13. [CrossRef]
12. Global Consumer Electronics Manufacturing-Global Market Research Report. Available online: https://www.ibisworld.com/industry-trends/global-industry-reports/manufacturing/consumer-electronics-manufacturing.html (accessed on 10 October 2017).
13. Personal/Consumer Electronics Market Analysis by Product (Smartphones, Tablets, Desktops, Laptops/Notebooks, Digital Cameras, Hard Disk Drives, E-Readers) and Segment Forecasts to 2020. Available online: http://www.grandviewresearch.com/industry-analysis/personal-consumer-electronics-market (accessed on 12 October 2017).
14. Capodieci, L. Data analytics and machine learning for design process-yield optimization in electronic design automation and IC semiconductor manufacturing. In Proceedings of the China Semiconductor Technology International Conference (CSCTI), Shanghai, China, 12–13 March 2017; pp. 1–3.
15. Romero, C.; Ventura, S. Data mining in education. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2013**, *3*, 12–27. [CrossRef]
16. Han, J.W.; Kamber, M.; Pei, J. *Data Mining, Concepts and Techniques*, 3rd ed.; Morgan Kaufmann: Waltham, MA, USA, 2012; Chapter 1–3; pp. 6–12.
17. Knowledge Discovery and Data Mining. Available online: http://researcher.ibm.com/researcher/view_group.php?id=144 (accessed on 15 October 2017).
18. Philip, C.C.L.; Zhang, C. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Inform. Sci.* **2014**, *275*, 314–347. [CrossRef]
19. Big Data. Available online: https://en.wikipedia.org/wiki/Big_data (accessed on 15 October 2017).
20. Big Data. Available online: https://www.gartner.com/it-glossary/big-data (accessed on 15 October 2017).
21. Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.S.; et al. Top 10 algorithms in data mining. *Knowl. Inform. Syst.* **2008**, *14*, 1–37. [CrossRef]
22. Confusion Matrix. Available online: https://en.wikipedia.org/wiki/Confusion_matrix (accessed on 16 October 2017).
23. Wang, G.; Xu, T.; Tang, T.; Yuan, T.; Wang, H. A Bayesian network model for prediction of weather-related failures in railway turnout systems. *Expert Syst. Appl.* **2017**, *69*, 247–256. [CrossRef]
24. Evaluation of Clustering. Available online: https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html (accessed on 17 October 2017).

25. Electronics Manufacturing. Available online: http://www.vault.com/industries-professions/industries/electronics-manufacturing.aspx (accessed on 17 October 2017).
26. Khader, N.; Yoon, S.W.; Li, D.B. Stencil printing optimization using a hybrid of support vector regression and mixed-integer linear programming. *Procedia Manuf.* **2017**, *11*, 1809–1817. [CrossRef]
27. Tsai, T.; Liukkonen, M. Robust parameter design for the micro-BGA stencil printing process using a fuzzy logic-based Taguchi method. *Appl. Soft Comp.* **2016**, *48*, 124–136. [CrossRef]
28. Chien, C.; Hsu, C. Data Mining for optimizing IC feature designs to enhance overall wafer effectiveness. *IEEE Trans. Semicond. Manuf.* **2014**, *27*, 71–82. [CrossRef]
29. Sun, Z.L.; Guo, Y.; Pan, E.S.; Song, W. Reflow soldering process virtual test based on BPNN-GA and ANSYS. *Appl. Mech. Mater.* **2013**, *281*, 417–421. [CrossRef]
30. Kwak, D.; Kim, K. A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes. *Expert Syst. Appl.* **2012**, *39*, 2590–2596. [CrossRef]
31. Tsai, T. Thermal parameters optimization of a reflow soldering profile in printed circuit board assembly: A comparative study. *Appl. Soft Comp.* **2012**, *12*, 2601–2613. [CrossRef]
32. Pan, E.; Jin, Y.; Xu, H.; Liao, W.Z. Forecasting and parameters optimization of reflow soldering profile based on BPNN and GA. *Adv. Mater. Res.* **2011**, *139–141*, 990–995.
33. Chan, K.Y.; Kwong, C.K.; Tsim, Y.C. Modelling and optimization of fluid dispensing for electronic packaging using neural fuzzy networks and genetic algorithms. *Eng. Appl. Artif. Intell.* **2010**, *23*, 18–26. [CrossRef]
34. Chou, J.; Tai, Y.; Chang, L.J. Predicting the development cost of TFT-LCD manufacturing equipment with artificial intelligence models. *Int. J. Prod. Econ.* **2010**, *128*, 339–350. [CrossRef]
35. Liukkonen, M.; Hiltunen, T.; Havia, E.; Leinonen, H.; Hiltunen, Y. Modeling of soldering quality by using artificial neural networks. *IEEE Trans. Electron. Packag. Manuf.* **2009**, *32*, 89–96. [CrossRef]
36. Barajas, L.G.; Egerstedt, M.B.; Kamen, E.W.; Goldstein, A. Stencil printing process modeling and control using statistical neural networks. *IEEE Trans. Electron. Packag. Manuf.* **2008**, *31*, 9–18. [CrossRef]
37. Kwon, Y.; Omitaomu, O.A.; Wang, J.N. Data mining approaches for modeling complex electronic circuit design activities. *Comput. Ind. Eng.* **2008**, *54*, 229–241. [CrossRef]
38. Bae, J.K.; Kim, J. Product development with data mining techniques: A case on design of digital camera. *Expert Syst. Appl.* **2011**, *38*, 9274–9280. [CrossRef]
39. Stoyanov, S.; Bailey, C.; Tourloukis, G. Similarity approach for reducing qualification tests of electronic components. *Microelectron. Reliab.* **2016**, *67*, 111–119. [CrossRef]
40. Haneda, H.; Kodama, H.; Hirogaketi, T.; Aoyama, E.; Ogawa, K. Investigation of drilling conditions of printed circuit board based on data mining method from tool catalog data-base. *Adv. Mater. Res.* **2014**, *939*, 547–554. [CrossRef]
41. Liukkonen, M.; Havia, E.; Leinonenb, H.; Hiltunena, Y. Quality-oriented optimization of wave soldering process by using self-organizing maps. *Appl. Soft Comp.* **2011**, *11*, 214–220. [CrossRef]
42. Li, S.; Nahar, K.; Fung, B.C.M. Product customization of tablet computers based on the information of online reviews by customers. *J. Intell. Manuf.* **2015**, *26*, 97–110. [CrossRef]
43. Yu, C.; Kuo, C. Data mining approaches to optimize the allocation of production resources in semiconductor wafer fabrication. In Proceedings of the 2016 International Symposium on Semiconductor Manufacturing (ISSM), Tokyo, Japan, 12–13 December 2017; pp. 1–4.
44. Wang, J.; Zhang, J. A hybrid data driven approach for cycle-time forecasting in semiconductor wafer fabrication system. In Proceedings of the 20th world multi-conference on systemics, cybernetics and informatics (WMSCI), SeaWorld, Orlando, FL, USA, 5–8 July 2016; pp. 74–78.
45. Chen, T. An efficient and effective fuzzy collaborative intelligence approach for cycle time estimation in wafer fabrication. *Int. J. Intell. Syst.* **2015**, *30*, 620–650. [CrossRef]
46. Chen, T. An effective fuzzy collaborative forecasting approach for predicting the job cycle time in wafer fabrication. *Comp. Ind. Eng.* **2013**, *66*, 834–848. [CrossRef]
47. Zhang, J.; Qin, W.; Wu, L.H.; Zhai, W.B. Fuzzy neural network-based rescheduling decision mechanism for semiconductor manufacturing. *Comput. Ind.* **2014**, *65*, 1115–1125. [CrossRef]
48. Chien, C.; Hsu, C.; Hsiao, C.W. Manufacturing intelligence to forecast and reduce semiconductor cycle time. *J. Intell. Manuf.* **2012**, *23*, 2281–2294. [CrossRef]
49. Vainio, F.; Maier, M.; Knuutila, T.; Alhoniemi, E.; Johnsson, M.; Nevalainen, O.S. Estimating printed circuit board assembly times using neural networks. *Int. J. Prod. Res.* **2010**, *48*, 2201–2218. [CrossRef]

50. Zhang, L.; Liu, T.; Liu, M.; Wang, X.H. Scheduling semiconductor wafer fabrication using a new fuzzy association classification rules based on dynamic fuzzy partition. *Chin. J. Electron.* **2017**, *26*, 112–117. [CrossRef]
51. Chen, T. A job-classifying and data-mining approach for estimating job cycle time in a wafer fabrication factory. *Int. J. Adv. Manuf. Technol.* **2012**, *62*, 317–328. [CrossRef]
52. Tirkel, I. Cycle time prediction in wafer fabrication line by applying data mining methods. In Proceedings of the 2011 22nd Annual IEEE/SEMI Advanced Semiconductor Manufacturing Conference, Saratoga Springs, NY, USA, 16–18 May 2011; IEEE Press: New York, NY, USA, 2011; pp. 1–5.
53. Meidan, Y.; Lerner, B.; Rabinowitz, G.; Hassoun, M. Cycle-time key factor identification and prediction in semiconductor manufacturing using machine learning and data mining. *IEEE Trans. Semicond. Manuf.* **2011**, *24*, 237–248. [CrossRef]
54. Chen, L.; Chien, C. Manufacturing intelligence for class prediction and rule generation to support human capital decisions for high-tech industries. *Flex. Serv. Manuf. J.* **2011**, *23*, 263–289. [CrossRef]
55. Shiue, Y.R. Data-mining-based dynamic dispatching rule selection mechanism for shop floor control systems using a support vector machine approach. *Int. J. Prod. Res.* **2009**, *47*, 3669–3690. [CrossRef]
56. Chen, T. A fuzzy rule for job dispatching in a wafer fabrication factory—A simulation study. *Int. J. Adv. Manuf. Technol.* **2013**, *67*, 47–58. [CrossRef]
57. Wu, H.; Chen, T. A fuzzy-neural ensemble and geometric rule fusion approach for scheduling a wafer fabrication factory. *Math. Probl. Eng.* **2013**, *2013*, 956978. [CrossRef]
58. Chen, T. A fuzzy-neural DBD approach for job scheduling in a wafer fabrication factory. *Int. J. Innov. Comp. Inform. Control* **2012**, *8*, 4025–4044.
59. Chen, T. Intelligent scheduling approaches for a wafer fabrication factory. *J. Intel. Manuf.* **2012**, *23*, 897–911. [CrossRef]
60. Shiue, Y.; Guh, R.; Lee, K.C. Study of SOM-based intelligent multi-controller for real-time scheduling. *Appl. Soft Comp.* **2011**, *11*, 4569–4580. [CrossRef]
61. Chen, T. Dynamic fuzzy-neural fluctuation smoothing rule for jobs scheduling in a wafer fabrication factory. *Proc. Inst. Mech. Eng. I-J. Syst. Control Eng.* **2009**, *223*, 1081–1094. [CrossRef]
62. Chen, T.; Wang, Y. A nonlinear scheduling rule incorporating fuzzy-neural remaining cycle time estimator for scheduling a semiconductor manufacturing factory—A simulation study. *Int. J. Adv. Manuf. Technol.* **2009**, *45*, 110–121. [CrossRef]
63. Wang, J.; Zhang, J. Big data analytics for forecasting cycle time in semiconductor wafer fabrication system. *Int. J. Prod. Res.* **2016**, *54*, 7231–7244. [CrossRef]
64. Chen, T.; Wang, Y. An iterative procedure for optimizing the performance of the fuzzy-neural job cycle time estimation approach in a wafer fabrication factory. *Math. Probl. Eng.* **2013**, *2013*, 740478. [CrossRef]
65. Chen, T.; Romanowski, R. Precise and accurate job cycle time forecasting in a wafer fabrication factory with a fuzzy data mining approach. *Math. Probl. Eng.* **2013**, *2013*, 496826. [CrossRef]
66. Chen, T. Job cycle time estimation in a wafer fabrication factory with a bi-directional classifying fuzzy-neural approach. *Int. J. Adv. Manuf. Technol.* **2011**, *56*, 1007–1018. [CrossRef]
67. Chen, T.; Lin, Y. A collaborative fuzzy-neural approach for internal due date assignment in a wafer fabrication plant. *Int. J. Innov. Comp. Inform. Control* **2011**, *7*, 5193–5210.
68. Chen, T.; Wang, Y. Incorporating the FCM-BPN approach with nonlinear programming for internal due date assignment in a wafer fabrication plant. *Probl. Comp. Integr. Manuf.* **2010**, *26*, 83–91. [CrossRef]
69. Chen, T.; Wang, Y. A bi-criteria nonlinear fluctuation smoothing rule incorporating the SOM-FBPN remaining cycle time estimator for scheduling a wafer fab—A simulation study. *Int. J. Adv. Manuf. Technol.* **2010**, *49*, 709–721. [CrossRef]
70. Chen, T.; Lin, Y. A fuzzy back propagation network ensemble with example classification for lot output time prediction in a wafer fab. *Appl. Soft Comp.* **2009**, *9*, 658–666. [CrossRef]
71. Chen, T.; Wu, H.; Wang, Y.C. Fuzzy-neural approaches with example post-classification for estimating job cycle time in a wafer fab. *Appl. Soft Comp.* **2009**, *9*, 1225–1231. [CrossRef]
72. Chen, T.; Jeang, A.; Wang, Y.C. A hybrid neural network and selective allowance approach for internal due date assignment in a wafer fabrication plant. *Int. J. Adv. Manuf. Technol.* **2008**, *36*, 570–581. [CrossRef]

73. Chen, T.; Wang, Y. Lot cycle time prediction in a ramping-up semiconductor manufacturing factory with a SOM-FBPN-ensemble approach with multiple buckets and partial normalization. *Int. J. Adv. Manuf. Technol.* **2009**, *42*, 1206–1216. [CrossRef]
74. Chen, T. An intelligent mechanism for lot output time prediction and achievability evaluation in a wafer fab. *Comp. Ind. Eng.* **2008**, *54*, 77–94. [CrossRef]
75. Chen, T. An intelligent hybrid system for wafer lot output time prediction. *Adv. Eng. Inform.* **2007**, *21*, 55–65. [CrossRef]
76. Chen, T. A hybrid look-ahead SOM-FBPN and FIR system for wafer-lot-output time prediction and achievability evaluation. *Int. J. Adv. Manuf. Technol.* **2007**, *35*, 575–586. [CrossRef]
77. Chen, T. A SOM-FBPN-ensemble approach with error feedback to adjust classification for wafer-lot completion time prediction. *Int. J. Adv. Manuf. Technol.* **2008**, *37*, 782–792. [CrossRef]
78. Shiue, Y.; Guh, R.S.; Tseng, T.Y. Study on shop floor control system in semiconductor fabrication by self-organizing map-based intelligent multi-controller. *Comp. Ind. Eng.* **2012**, *62*, 1119–1129. [CrossRef]
79. Chien, C.; Chen, Y.; Hsu, C.Y. A novel approach to hedge and compensate the critical dimension variation of the developed-and-etched circuit patterns for yield enhancement in semiconductor manufacturing. *Comput. Oper. Res.* **2015**, *53*, 309–318. [CrossRef]
80. Tsuda, T.; Inoue, S.; Akihiro, K.; Shin-ichi, I.; Tomoya, T.; Naoaki, S.; Satoshi, Y. Advanced semiconductor manufacturing using big data. *IEEE Trans. Semicond. Manuf.* **2015**, *28*, 229–235. [CrossRef]
81. Lenz, B.; Barak, B. Data mining and support vector regression machine learning in semiconductor manufacturing to improve virtual metrology. In Proceedings of the 46th Hawaii International Conference on System Sciences(HICSS), Wailea, Maui, HI, USA, 7–10 January 2013; pp. 3447–3456.
82. Lenz, B.; Barak, B. Virtual metrology in semiconductor manufacturing by means of predictive machine learning models. In Proceedings of the 12th International Conference on Machine Learning and Applications, Miami, FL, USA, 4–7 December 2014; pp. 174–177.
83. Kang, P.; Lee, H.; Cho, S.; Kim, D.; Park, J.; Park, C.K. A virtual metrology system for semiconductor manufacturing. *Expert Syst. Appl.* **2009**, *36*, 12554–12561. [CrossRef]
84. Guo, W.; Banerjee, A.G. Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs. *J. Manuf. Syst.* **2017**, *43*, 225–234. [CrossRef]
85. Chien, C.; Liu, C.; Chuang, S.C. Analyzing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement. *Int. J. Prod. Res.* **2017**, *55*, 5095–5107. [CrossRef]
86. Moyne, J.; Iskandar, J. Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing. *Processes* **2017**, *5*, 39. [CrossRef]
87. Chien, C.; Chen, Y.; Wu, J.Z. Big data analytics for modeling WAT parameter variation induced by process tool in semiconductor manufacturing and empirical study. In Proceedings of the 2016 Winter Simulation Conference, Washington, DC, USA, 11–14 December 2016; pp. 2512–2522.
88. Hessinger, U.; Chan, W.K.; Schafman, B.T. Data Mining for significance in yield-defect correlation analysis. *IEEE Trans. Semicond. Manuf.* **2014**, *27*, 347–356. [CrossRef]
89. Chien, C.; Chuang, S.C. A framework for root cause detection of sub-batch processing system for semiconductor manufacturing big data analytics. *IEEE Trans. Semicond. Manuf.* **2014**, *27*, 475–488. [CrossRef]
90. Chien, C.F.; Diaz, A.C.; Lan, Y.B. A data mining approach for analyzing semiconductor MES and FDC data to enhance overall usage effectiveness (OUE). *Int. J. Comp. Intell. Syst.* **2014**, *72*, 52–65. [CrossRef]
91. Chien, C.; Hsu, C.; Chen, P.N. Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence. *Flex. Serv. Manuf. J.* **2013**, *25*, 367–388. [CrossRef]
92. Ko, J.M.; Hong, S.R.; Choi, J.Y.; Kim, C.O. Wafer-to-wafer process fault detection using data stream mining techniques. *Int. J. Precis. Eng. Manuf.* **2013**, *14*, 103–113. [CrossRef]
93. Chien, C.; Hsu, S.; Chen, Y.J. A system for online detection and classification of wafer bin map defect patterns for manufacturing intelligence. *Int. J. Prod. Res.* **2013**, *51*, 2324–2338. [CrossRef]
94. Tsai, T. Development of a soldering quality classifier system using a hybrid data mining approach. *Expert Syst. Appl.* **2012**, *39*, 5727–5738. [CrossRef]
95. Weiss, S.M.; Baseman, R.J.; Tipu, F.; Collins, C.N.; Davies, W.A.; Singh, R.; Hopkins, J.W. Rule-based data mining for yield improvement in semiconductor manufacturing. *App. Intell.* **2010**, *33*, 318–329. [CrossRef]

96. Ooi, M.P.; Sim, E.K.J.; Kuang, Y.C.; Demidenko, S.; Kleeman, L.; Chan, C.W.K. Getting more from the semiconductor test: Data mining with defect-cluster extraction. *IEEE Trans. Instrum. Meas.* **2011**, *60*, 3300–3317. [CrossRef]
97. Susto, G.A.; Terzi, M.; Beghi, A. Anomaly detection approaches for semiconductor manufacturing. *Procedia Manuf.* **2017**, *11*, 2018–2024. [CrossRef]
98. Li, Z.; Baseman, R.J.; Zhu, Y.; Tipu, F.A.; Slonim, N.; Shpigelman, L. A unified framework for outlier Detection in trace data analysis. *IEEE Trans. Semicond. Manuf.* **2014**, *27*, 95–103. [CrossRef]
99. Sohn, Y.; Lee, H.; Yang, Y.; Jun, C. A new method for wafer quality monitoring using semiconductor process big data. In Proceedings of the Society of Photo-Optical Instrumentation Engineers, San Jose, CA, USA, 28 March 2017; SPIE: Bellingham, WA, USA, 2017; p. 101450T.
100. Chen, T. A heterogeneous fuzzy collaborative intelligence approach for forecasting the product yield. *Appl. Soft Comp.* **2017**, *57*, 210–224. [CrossRef]
101. Butte, S.; Patil, S. Big data and predictive analytics methods for modeling and analysis of semiconductor manufacturing processes. In Proceedings of the IEEE Workshop on Microelectronics and Electron Devices (WMED), Boise, ID, USA, 15 April 2016; pp. 1–5.
102. Lee, H.; Kim, C.O.; Ko, H.H.; Kim, M.Y. Yield prediction through the event sequence analysis of the die attach process. *IEEE Trans. Semicond. Manuf.* **2015**, *28*, 563–570. [CrossRef]
103. Krueger, D.C.; Montgomery, D.C. Modeling and analyzing semiconductor yield with generalized linear mixed models. *Appl. Stoch. Models Bus. Ind.* **2014**, *30*, 691–707. [CrossRef]
104. Chen, T. Forecasting the yield of a semiconductor product with a collaborative intelligence approach. *Appl. Soft Comp.* **2013**, *13*, 1552–1560. [CrossRef]
105. Shukla, S.K.; Tiwari, M.K. GA guided cluster based fuzzy decision tree for reactive ion etching modeling: A data mining approach. *IEEE Trans. Semicond. Manuf.* **2012**, *25*, 45–56. [CrossRef]
106. Chen, T. Applying the hybrid fuzzy c-means-back propagation network approach to forecast the effective cost per die of a semiconductor product. *Comp. Ind. Eng.* **2011**, *61*, 752–759. [CrossRef]
107. Feng, C.J.; Gao, L.; Li, P.G.; Shao, X.Y. Selection and comparison of supervised predictive data mining models for electronics fabrication data. In Proceedings of the 2010 International Conference on Computing, Control and Industrial Engineering, Wuhan, China, 5–6 June 2010; pp. 3–7.
108. Chen, T.; Lin, Y. A fuzzy-neural system incorporating unequally important expert opinions for semiconductor yield forecasting. *Int. J. Uncertain. Fuzz. Knowl. Syst.* **2008**, *16*, 35–58. [CrossRef]
109. Guan, T.; Zhang, Z.B.; Dong, W.; Qiao, C.M.; Gu, X.L. Data-driven fault diagnosis with missing syndromes imputation for functional test through conditional specification. In Proceedings of the 22nd IEEE European Test Symposium (ETS), Limassol, Cyprus, 22–26 May 2017; pp. 1–6.
110. Lee, C.; Chen, B. Mutually-exclusive-and-collectively-exhaustive feature selection scheme. *Appl. Soft Comp.* **2017**. [CrossRef]
111. Chen, Y.; Fan, C.Y.; Chang, K.H. Manufacturing intelligence for reducing false alarm of defect classification by integrating similarity matching approach in CMOS image sensor manufacturing. *Comp. Ind. Eng.* **2016**, *99*, 465–473. [CrossRef]
112. Fan, S.S.; Lin, S.C.; Tsai, P.F. Wafer fault detection and key step identification for semiconductor manufacturing using principal component analysis, AdaBoost and decision tree. *J. Ind. Prod. Eng.* **2016**, *33*, 151–168. [CrossRef]
113. Liao, C.; Hsieh, T.J.; Huang, Y.S.; Chien, C.F. Similarity searching for defective wafer bin maps in semiconductor manufacturing. *IEEE Trans. Autom. Sci. Eng.* **2014**, *11*, 953–960. [CrossRef]
114. Chien, C.; Chang, K.H.; Wang, W.C. An empirical study of design-of-experiment data mining for yield-loss diagnosis for semiconductor manufacturing. *J. Intell. Manuf.* **2014**, *25*, 961–972. [CrossRef]
115. Chen, Y.; Lin, T.H.; Chang, K.H.; Chien, C.F. Feature extraction for defect classification and yield enhancement in color filter and micro-lens manufacturing: An empirical study. *J. Ind. Prod. Eng.* **2013**, *30*, 510–517. [CrossRef]
116. Hsieh, T.; Liao, C.; Huang, Y.S.; Chien, C.F. A new morphology-based approach for similarity searching on wafer bin maps in semiconductor manufacturing. In Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design, Wuhan, China, 23–25 May 2012; pp. 869–874.

117. Chien, C.; Wang, W.C.; Cheng, J.C. Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Syst. Appl.* **2007**, *33*, 192–198. [CrossRef]
118. Hsu, S.; Chien, C. Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing. *Int. J. Prod. Econ.* **2007**, *107*, 88–103. [CrossRef]
119. Sim, H.; Choi, D.; Kim, C.O. A data mining approach to the causal analysis of product faults in multi-stage PCB manufacturing. *Int. J. Precis. Eng. Manuf.* **2014**, *15*, 1563–1573. [CrossRef]
120. Casali, A.; Ernst, C. Discovering correlated parameters in semiconductor manufacturing processes: A data mining approach. *IEEE Trans. Semicond. Manuf.* **2012**, *25*, 118–127. [CrossRef]
121. Zhu, Y.; Xiong, J.J. Modern big data analytics for "old-fashioned" semiconductor industry applications. In Proceedings of the 2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), Austin, TX, USA, 2–6 November 2015; pp. 776–780.
122. Chen, H.H.; Hsu, R.; Yang, P.Y.; Shyr, J.J. Predicting system-level test and in-field customer failures using data mining. In Proceedings of the 2013 IEEE International Test Conference (ITC), Anaheim, CA, USA, 6–13 September 2013; pp. 1–10.
123. Mashhadi, A.R.; Esmaeilian, B.; Cade, W.; Wiens, K. Mining consumer experiences of repairing electronics: Product design insights and business lessons learned. *J. Clean. Prod.* **2016**, *137*, 716–727. [CrossRef]
124. Sabbaghi, M.; Esmaeilian, B.; Mashhadi, A.R.; Behdad, S.; Cade, W. An investigation of used electronics return flows: A data-driven approach to capture and predict consumers storage and utilization behavior. *Waste Manag.* **2015**, *36*, 305–315. [CrossRef] [PubMed]
125. Tavakkoli, A.; Rezaeenour, J.; Hadavandi, E. A novel forecasting model based on support vector regression and Bat meta-heuristic (Bat-SVR): Case study in printed circuit board industry. *Int. J. Inform. Technol. Des. Mak.* **2015**, *14*, 195–215. [CrossRef]
126. Chang, P.; Liu, C.H.; Fan, C.Y. Data clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry. *Knowl. Syst.* **2009**, *22*, 344–355. [CrossRef]
127. Chang, P.C.; Liu, C.H.; Lai, R.K. Fuzzy case-based reasoning model for sales forecasting in print circuit board industries. *Expert Syst. Appl.* **2008**, *34*, 2049–2058. [CrossRef]
128. Chang, P.; Wang, Y.W.; Liu, C.H. The development of a weighted evolving fuzzy neural network for PCB sales forecasting. *Expert Syst. Appl.* **2007**, *32*, 86–96. [CrossRef]
129. Sabbaghi, M.; Cade, W.; Behdad, S.; Bisantz, A. M. The current status of the consumer electronics repair industry in the U.S.: A survey-based study. *Resour. Conserv. Recyc.* **2017**, *116*, 137–151. [CrossRef]
130. Chien, C.; Chen, Y.J.; Peng, J.T. Manufacturing intelligence for semiconductor demand forecast based on technology diffusion and product life cycle. *Int. J. Prod. Econ.* **2010**, *128*, 496–509. [CrossRef]
131. Top-Free-Data-Mining-Software. Available online: https://www.predictiveanalyticstoday.com/top-free-data-mining-software/ (accessed on 10 November 2017).

Original Article

Institution of MECHANICAL ENGINEERS
Journal of ENGINEERING MANUFACTURE

Proc IMechE Part B:
J Engineering Manufacture
2016, Vol. 230(8) 1525–1536

DOI: 10.1177/0954405416640697
pib.sagepub.com
SAGE

A cross-entropy-based approach for joint process plan selection and scheduling optimization

Shengping Lv[1] and Wei Liu[2]

Abstract

The process plan selection and job shop scheduling are carried out separately and sequentially in many factories and the scheduling is always conducted after the process plan of each job has been determined. In fact, the activities for the determination of process plan and the scheduling plan are coupled with each other and actually complementary. Implementation of the two activities with an appropriate collaborative approach is essential to achieve greater performance and higher productivity for the manufacturing system. In this article, a novel cross-entropy-based approach for the joint process plan selection and scheduling optimization that can assist process planning and scheduling system to achieve optimal scheduling plan and determine the operations, machine for each operation and operation sequence for each job collaboratively was proposed. In order to facilitate the manipulation and improve the optimized performance of the approach, an efficient representation scheme and a generation method for samples were developed. Meanwhile, the updating mechanism for new introduced probability distribution parameters according to which the cross-entropy procedure generates samples was established. To verify the adaptability and performance of the proposed approach, experimental studies were conducted and comparisons were made between this approach and some previous methods. The experimental results indicate that the proposed approach is an alternative and acceptable method to solve the joint process plan selection and scheduling optimization problem.

Keywords

Cross-entropy, process planning, process plan selection, scheduling, optimization

Date received: 20 May 2015; accepted: 1 February 2016

Introduction

Manufacturing process planning (PP) and job shop scheduling are two pivotal planning steps which can greatly affect the performance of manufacturing systems. PP links product design and product manufacturing by specifying what resources are needed to produce a job and determining detailed instructions for transforming raw materials into the final product.[1] The selection of process plan as the inevitable step of flexible PP mainly includes the determination of operations, machine for each operation and operation sequence. With the selected process plans of jobs as inputs, the main task of scheduling is to assign operations of all the jobs on available machines with precedence relations defined in the process plans satisfied to optimize some predefined objectives.[1,2] Therefore, the selection of process plan as an important task of PP and scheduling are coupled with each other and actually complementary.

A PP system should interface with a scheduling system to generate more appropriate process and scheduling plans. In doing so, the efficiency of manufacturing system as a whole is expected to be improved.[3] In recent years, numerous efforts have been made to the research and application of integrated process planning and scheduling (IPPS) system, especially for the important and challenging problem of joint process plan selection and scheduling optimization (JPPSSO).[4–7] The joint implementation mechanisms indicate that the IPPS can

[1]College of Engineering, South China Agricultural University, Guangzhou, China
[2]Shenzhen Parameter Navigator Technology Co., Ltd, Shenzhen, China

Corresponding author:
Shengping Lv, College of Engineering, South China Agricultural University, No. 483 Wushan Road, Tianhe District, 510642 Guangzhou, Guangdong, China.
Email: lvshengping@scau.edu.cn

introduce significant improvement to the efficiency of the manufacturing facilities through elimination or reduction in scheduling conflicts, reduction in flow-time and work-in-process, improvement of production resources utilization and adaptation to irregular shop floor disturbances.[4] Therefore, it is ideal to integrate the PP and scheduling more tightly and implement the process plan selection and scheduling optimization jointly to increase the productivity and responsiveness of the manufacturing systems.[7]

The JPPSSO problem that is always directly referred as IPPS in many other papers is very different from the PP and the scheduling problem. Because the objectives are different, meanwhile, the constraints and the solution space of JPPSSO are more complicated than the PP problem and the scheduling problem.[6] The previous approaches for the scheduling cannot be utilized to solve the JPPSSO problem; therefore, several approaches have been developed to facilitate the optimization of the integrated problem. Some earlier work of IPPS had been summarized by Phanden et al.[8] The algorithm-based and agent-based methods are the two main optimization approaches for IPPS.

Due to their advantages in solving combinatorial optimization problems, meta-heuristic algorithms, mainly including the simulated annealing (SA),[9] genetic algorithm (GA),[10] object-coding GA,[11] particle swarm optimization (PSO),[12,13] imperialist competitive algorithm (ICA),[1] tabu search (TS)[14] and evolutionary algorithm (EA)-based approach,[2,5,6,15,16] have been utilized for the IPPS problem in the past few years. Meanwhile, the pattern search, GA, and SA have also been employed for the integrated problem considering energy consumption.[17] The hybrid approaches of graph-based ant colony (AC),[18] simulation-based GA,[19] combination of GA with TS,[7] game theory,[20] active leaning[21] and PSO[22] have also been established for the problem.

Agent-based approach is another important implementation method for the problem. Wong and colleagues[23–25] developed an online hybrid agent and an online multi-agent approach to integrate PP with scheduling/rescheduling. Leung et al.[26] combined an AC algorithm in the agent-based system to facilitate the integration of PP and scheduling. Shukla et al.[27] conceptualized a bidding-based multi-agent system for the IPPS problem. Li et al.[7] utilized job agent, machine agent and optimization agent to identify manufacturing operations, machine for each operation and scheduling plan simultaneously. Ueda et al.[28] introduced a multi-agent learning-based approach for the IPPS. Hsieh and colleagues[29,30] developed an integrated system considering the integration of PP and scheduling-based multi-agent architecture. Shen et al.[31] gave a comprehensive review for the agent-based approaches for PP, scheduling and the integration of the two.

These studies excluded the generation of flexible process plans while conducting PP and mainly focus on the JPPSSO optimization. In this article, a novel cross-entropy (CE)-based approach has been developed to tackle the JPPSSO problem. The CE-based method[32,33] as an effective approach for the combinatorial optimization problems has not yet been employed for the JPPSSO or IPPS problem to our best knowledge. The contributions of the proposed CE-based approach in this article are summarized below:

- The CE has been employed to optimize many combinatorial problems with one decision variable (vector). However, in this article, the systematic mechanisms including probability distribution design and updating, sample encoding/decoding and generation have been developed for the optimization of JPPSSO problem with two sub-problems (decision vectors) should be considered cooperatively.
- This approach supports intelligent decision making for the selection of process plan for each job and optimal scheduling plans from the view of manufacturing system collaboratively.
- Experimental studies have been conducted and comparisons have been made between CE and some previous methods. The experimental results indicate that the algorithm is effective and it demonstrates potential applicability in practice.

Proposed CE-based approach for JPPSSO

Problem description

The JPPSSO discussed in this article is stated as follows: given a set of N jobs which are to be processed on M machines with alternative operation sequences and alternative machines for operations, determine an operation sequence and corresponding machine sequence for each job and a schedule in which operations on the same machines are processed so that it satisfies the precedence requirements and the corresponding objectives can be achieved.

The scheduling in the JPPSSO is often assumed as the job shop scheduling problem (JSP). The optimization objective of the joint problem in this article is to minimize makespan. The following assumptions are made:[32]

1. Job preemption is not allowed and each machine can handle only one job at a time;
2. All jobs are simultaneously available at time zero;
3. Different operations of one job cannot be processed simultaneously;
4. After a job is processed on a machine, it is immediately transported to the next machine;
5. Setup times for the operations on machines are independent of operation sequence and are included in the processing time.

The mixed integer programming model of JPPSSO problem the same as IPPS has been given by the

previous work.[32] The JPPSSO that is more complicated than the JSP is a non-deterministic (NP)-hard problem, no polynomial time algorithm exists to find optimal solutions. Therefore, a CE-based meta-heuristic approach was developed to tackle the problem.

Flow chart of CE for JPPSSO

The CE algorithm which is known as efficient method for the combinatorial optimization problems was first proposed by Rubinstein.[33] The application of CE-based approach is verified by its capability and simplicity to extract a subset solutions satisfying the predefined quality criterion from the large space of all samples (feasible solutions) generated randomly. The probability distribution according to which the CE procedure generates samples assures both their quality and diversity necessary for the further optimization iteration.[34] The basic procedure of the proposed CE-based method for JPPSSO is described as follows:

Procedure of the proposed CE for JPPSSO

Step 1. Set the parameters (except the probability distribution parameter) of the algorithm.
Step 2. Select s process plans for each job. For the job has only $s'(s' < s)$ different process plans, then copying exist process plans should be conducted to ensure each job has the same s plans passed on to scheduling.
Step 3. Construct the representation for samples.
Step 4. Design and introduce probability distribution based on the representation of samples.
Step 5. Generate the samples randomly based on the probability distribution.
Step 6. Decode and calculate the performance function values of all the samples.
Step 7. Is the termination criteria satisfied?
If yes, go to Step 10. Else, go to Step 8.
Step 8. Update probability distribution parameters.
Step 9. Regenerate samples randomly based on the updated probability distribution, and go to Step 6.
Step 10. Output the optimal sample (solution).

The process plan selection, sample representation, probability distribution designing, probability distribution parameter updating, sample generating and decoding are described in the following sections.

Process plans selection

Taking all the process plans for the joint optimization problem will seriously influence the computational efficiency when each job has large number of alternative process plans; even more, the advantage gained by increasing the number of alternative process plans for a scheduling system diminishes rapidly.[32] The experimental results obtained by modified GA,[5] hybrid Algorithm (HA)[6] and improved GA[32] also indicates the effectiveness by determining a number of optimal or near optimal process plans for each job before the joint optimization of process plan and scheduling plan. Therefore, the proposed CE-based approach tries to select s (near) optimal process plans based on their production time before joint optimization.

AND/OR network is a widely used[1,4,32] and also been utilized in this article to represent the flexible process plans. Based on the AND/OR flexible process plan network, the selection procedure parses and generates many process plans according to the initial selection method described by Qiao and Lv.[32] These parsed process plans are sorted based on their production time in a non-decreasing order, and then the first s process plans are transferred into the joint optimization procedure. The initial selection method mainly includes three steps: parse the OR-link and AND-link path in AND/OR network to construct linear process plan with each operation choosing the shortest production time (including the processing time and transportation time) among its alternatives, and then adjust positions of nodes in AND-link paths randomly but satisfying the constraints, select operations to adjust its machine among its alternatives randomly and associate the related processing time and transportation time accordingly.[32] For the job with few alternative process plans, all the process plans can be passed on to scheduling system. For each selected process plan, the detailed information including the operations, the machine for each operation and operation sequence has been determined.

Sample representation

CE algorithm takes every possible solution of a combinatorial optimization problem as a sample. The representation of samples that can match well with the discussed problem can facilitate the designing and manipulation of the algorithm. In this article, a representation structure consisting of process plan section and scheduling plan section, two parts, has been constructed to represent the sample for the JPPSSO problem. The position of 1 to N in the process plan section of a sample represents the job from 1 to N, in which the ith position represents the selected alternative process plan (from the s inputted process plans) for job i. The scheduling plan section consists of the permutation of integer number corresponding to the operations in the chosen process plans for all the jobs and the length is set to $L = \sum_{i=1}^{N} u_i$, where u_i is the maximum operation number among all the alternative process plans of job i. The integer number opn between $[\sum_{i=1}^{n} u_i, \sum_{i=1}^{n+1} u_i]$ in the cell of scheduling plan section indicates the $opn - \sum_{i=1}^{n} u_i, (1 \le opn \le \sum_{i=1}^{n} u_i)$th operation in the selected process plan of job $n + 1$.

Figure 1 shows a feasible sample. In this example, the number of job N is equal to 3 and therefore the process plan section is made up of three elements. Assume the maximum operation number among the process plans for the three job is $u_1 = 4$, $u_2 = 4$, $u_3 = 3$,

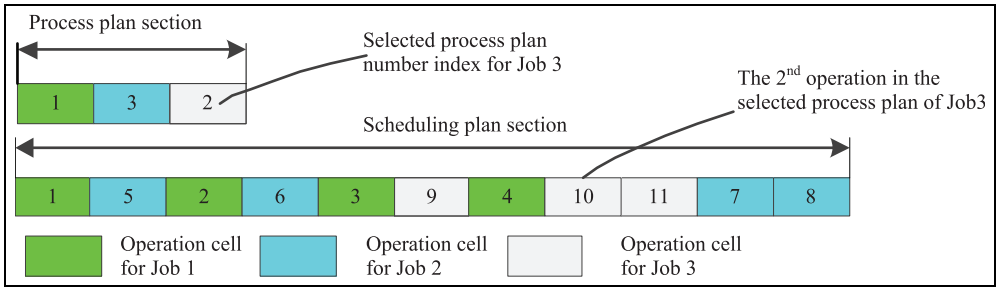


Figure 1. Instance of a sample.

respectively, then the scheduling plan section is made up of 11 elements. The process plan section in the sample indicates that Jobs 1, 2 and 3 select the first, third and second process plan from their s inputted process plans, respectively. The numbers 1–4, 5–8 and 9–11 in the scheduling plan section represent the first to fourth, first to fourth and the first to third operations in the selected process plan of jobs 1, 3 and 2, respectively.

Probability distribution designing and updating

The probability distribution according to which the CE procedure generates samples is closely related to the sample representation. Based on the sample structure, the process plan section and scheduling plan section in a sample can be represented by two random integer vectors $\mathbf{Y} = \{y_1, \ldots, y_i, \ldots y_N\}$ and $\mathbf{X} = \{x_1, \ldots, x_\nu, \ldots x_L\}$, respectively. The corresponding deterministic vectors are represented as $\mathbf{y}$ and $\mathbf{x}$, respectively. The elements in $\mathbf{Y}$ are independent discrete variables in $[1, s]$, and then a probability distribution parameters (matrix) $\mathbf{p}_\alpha = \{p_{\alpha,ij} | 1 \le i \le N, 1 \le j \le N\}$ is constructed to generate elements of $\mathbf{Y}$, in which $p_{\alpha,ij}$ represents the probability of job i, $1 \le i \le N$ selecting the jth, $1 \le j \le s$, process plan. The probability density function (PDF) with parameters $\mathbf{p}_\alpha$ is denoted by $f(\cdot; \mathbf{p}_\alpha)$; thus, the probability distribution of $\mathbf{Y}$ can be written as $\mathbf{Y} \sim f(Y; \mathbf{p}_\alpha)$. Similarly, the elements in $\mathbf{X}$ are discrete variables in $[1, L]$, and the probability distribution of $\mathbf{X}$ is described as $\mathbf{X} \sim f(\mathbf{X}; \mathbf{p}_\beta)$ by introducing the probability distribution parameters $\mathbf{p}_\beta = \{p_{\beta,\mu\nu} | 1 \le \mu \le L, 1 \le \nu \le L\}$, in which $p_{\beta,\mu\nu}$ is utilized to record the probability of the integer number ν at the position μ in the scheduling plan section of a sample.

The JPPSSO schemes represented by samples are determined by $\mathbf{Y}$ and $\mathbf{X}$ jointly, and then the optimization problem can be transformed into the following problem

$$S(\mathbf{x}^*, \mathbf{y}^*) = \gamma^* = \min_{y||x \in \chi} S(\mathbf{x}, \mathbf{y}) \tag{1}$$

where χ is the space of all the combination for the JPPSSO problem, S is the real-value performance function (its value is the makespan) in χ, $y||x$ represents a scheme consisting of $\mathbf{y}$ and $\mathbf{x}$ and $\gamma*$ is the minimum of S.

For certain $\mathbf{p}'_\alpha$, $\mathbf{p}'_\beta$, equation (1) is associated with the problem of estimating the number

$$\ell(\gamma) = P_{p'_\alpha, p'_\beta}(S(\mathbf{X}, \mathbf{Y}) \le \gamma) = E_{\mathbf{p}'_\alpha, \mathbf{p}'_\beta} I_{\{S(\mathbf{X}, \mathbf{Y}) \le \gamma\}} \tag{2}$$

where $P_{\mathbf{p}'_\alpha, \mathbf{p}'_\beta}$ is the probability measure under which the random vector $\mathbf{X}$, $\mathbf{Y}$ has PDF $f(\cdot; \mathbf{p}'_\alpha)$ and $f(\cdot; \mathbf{p}'_\beta)$, respectively, and $E_{\mathbf{p}'_\alpha, \mathbf{p}'_\beta}$ denotes the expectation operator, $\{I_{\{S(\mathbf{X}, \mathbf{Y}) \le \gamma\}}\}$ (0 or 1) is defined as a collection of indicator function on χ for various levels γ, $\gamma \in \mathbb{R}$.

For deterministic $\gamma = \gamma*$, parameter $\mathbf{p}^*_\alpha$, $\mathbf{p}^*_\beta$ can be estimated using the likelihood ratio estimator as

$$\hat{\mathbf{p}}_\alpha{}^* = \arg\max_{\mathbf{p}_\alpha} \frac{1}{Z} \sum_{z=0}^{Z} I_{\{S(\mathbf{X}_z, \mathbf{Y}_z) \le \gamma\}} \ln f(\mathbf{Y}_z; \mathbf{p}_\alpha) \tag{3}$$

$$\hat{\mathbf{p}}^*_\beta = \arg\max_{\mathbf{p}_\beta} \frac{1}{Z} \sum_{z=0}^{Z} I_{\{S(\mathbf{X}_z, \mathbf{Y}_z) \le \gamma\}} \ln f(\mathbf{X}_z; \mathbf{p}_\beta) \tag{4}$$

where Z is the sample size, $\mathbf{Y}_z$, $\mathbf{X}_z$, $z = 1, \ldots, Z$ is part of the sample in $\mathbf{Y}_z || \mathbf{X}_z$. With a specific $\mathbf{x}$ and $\mathbf{y}$, the PDF of $\mathbf{x}$ can be described as $f(\mathbf{x}; \mathbf{p}_\beta) = \prod_{\mu=1}^{L} \prod_{\nu=1}^{L} p_{\beta,\mu\nu}^{I_{\{x_\mu = \nu\}}}$, therefore the logarithm of $f(\mathbf{x}; \mathbf{p}_\beta)$ can be described as

$$\ln f(\mathbf{x}; \mathbf{p}_\beta) = \sum_{\mu=1}^{L} \sum_{\nu=1}^{L} I_{\{x_\mu = \nu\}} \ln p_{\beta,\mu\nu}$$

With the extra condition that the row of $\mathbf{p}_\beta$ needs to add up to 1, the maximization problem can be obtained using Lagrange multipliers $\omega_1, \ldots, \omega_L$

$$\max_{p_\beta} \min_{\omega_1, \ldots, \omega_L} \left[E_{\mathbf{p}_\beta, \mathbf{p}_\alpha} I_{\{S(\mathbf{x}, \mathbf{y}) \le \gamma\}} \ln f(\mathbf{x}; \mathbf{p}_\beta) + \sum_{\mu=1}^{L} \omega_\mu \left(\sum_{\nu=1}^{L} p_{\beta,\mu\nu} - 1 \right) \right]$$

On this basis, the corresponding estimator $\hat{p}_{\beta,\mu\nu}$, $\mu = 1, \ldots, L$, $\nu = 1, \ldots, L$, $\hat{p}_{\alpha,ij}$ of $p_{\alpha,ij}$, $i = 1, \ldots, N$, $j = 1, \ldots, s$ are

$$\hat{p}_{\beta,\mu\nu} = \frac{\sum_{z=1}^{Z} I_{\{S(\mathbf{X}_z, \mathbf{Y}_z) \le \gamma\}} I_{\{\mathbf{X}_{z\mu} = \nu\}}}{\sum_{z=1}^{Z} I_{\{S(\mathbf{X}_z, \mathbf{Y}_z) \le \gamma\}}} \tag{5}$$

$$\hat{p}_{\alpha,ij} = \frac{\sum_{z=1}^{Z} I_{\{S(\mathbf{X}_z,\mathbf{Y}_z)\le\gamma\}} I_{\{Y_{zi}=j\}}}{\sum_{z=1}^{Z} I_{\{S(\mathbf{X}_z,\mathbf{Y}_z)\le\gamma\}}} \quad (6)$$

where $\sum_{z=1}^{Z} I_{\{S(\mathbf{X}_z,\mathbf{Y}_z)\le\gamma\}}$ is the number of samples with makespan less than γ, and $\sum_{z=1}^{Z} I_{\{S(\mathbf{X}_z,\mathbf{Y}_z)\le\gamma\}} I_{\{\mathbf{X}_{z\mu}=\nu\}}$ is the number of samples with makespan less than γ taking into account those samples that have the number ν located at the position μ. $\sum_{z=1}^{Z} I_{\{S(\mathbf{X}_z,\mathbf{Y}_z)\le\gamma\}} I_{\{\mathbf{Y}_{zi}=j\}}$ is the number of samples with makespan less than γ considering those samples that have job i choosing the jth process plan from the s available alternatives.

In order to smooth out the parameter $\mathbf{p}_\alpha$ and $\mathbf{p}_\beta$ in each iteration, the smooth parameter η is introduced and the smoothed updating expressions are defined as follows

$$\hat{\mathbf{p}}_{\beta t} = \eta\hat{\mathbf{p}}_{\beta t} + (1-\eta)\hat{\mathbf{p}}_{\beta t-1} \quad (7)$$

$$\hat{\mathbf{p}}_{\alpha t} = \eta\hat{\mathbf{p}}_{\alpha 6} + (1-\eta)\hat{\mathbf{p}}_{\alpha t-1} \quad (8)$$

where $\hat{\mathbf{p}}_{\beta t}$ and $\hat{\mathbf{p}}_{\alpha t}$ obtained by equations (5) and (6) at the tth iteration are the estimators of $\mathbf{p}_{\beta t}$ and $\mathbf{p}_{\alpha t}$, respectively; while $\hat{\mathbf{p}}_{\beta t-1}$ and $\hat{\mathbf{p}}_{\alpha t-1}$ are the estimators of $\mathbf{p}_{\beta t-1}$ and $\mathbf{p}_{\alpha t-1}$ at the $(t-1)$th iteration, respectively.

Sample generating

The following two algorithms will introduce the procedure of generating random samples, which consist of $\mathbf{Y}$ and $\mathbf{X}$, based on the two probability distribution parameters $\mathbf{p}_\alpha$ and $\mathbf{p}_\beta$.The algorithm of generating $\mathbf{Y} = \{y_1, \ldots, y_i, \ldots y_N\}$ based on $\mathbf{p}_\alpha$ is stated as follows.

Procedure of generating $\mathbf{Y}$

Step 1. Set $i = 1$, normalize $\mathbf{p}_\alpha$ and make sure the sum of elements in each row equals to 1.
Step 2. Generate a double number r, $r \in [0, 1]$ randomly, and get the least t with $\sum_{j=1}^{t} p_{\alpha,ij} \geqslant r$, then $y_i = t$.
Step 3. If $i < N$, $i = i + 1$, return to Step 2; otherwise, set $\mathbf{Y} = \{y_1, \ldots, y_i, \ldots y_N\}$ and terminate.

The composition technique used to generate trajectory proposed by Rubinstein and Kroese,[33] which prove that the method can speed up substantially the generation while affecting very little the accuracy of the CE algorithm, is modified to generate $\mathbf{X}$ as follows:

Sample decoding

A sample can be scheduled only if the constraints have already been satisfied. In a sample, the operation precedence restriction is neglected at the generating stage and feasibility is ensured by the following decoding procedure. In the decoding process, a fixed alternative process plan for each job is given; therefore, the detailed information (operation, machine and the processing time) for each cell in the scheduling plan section based on the selected process plans has also been determined. If the number of operations (denoted by q_i) in the chosen process plans of job i, $i \in [1, N]$ is less than u_i, then the detailed information of $(q_i + 1)$th to u_ith operation in the selected process plan are filled with 0. The notations used to explain the procedure are described below.

Procedure of generating $\mathbf{X}$

Step 1. Set $\mu = 1$.
Step 2. Normalize the μth row of $\mathbf{p}_\beta$ and assure the sum of elements in the row equals to 1.
Step 3. Divide the columns with non-zero elements in the μth row into three groups, denoted $\Omega(\psi)$, $\psi = 1, 2, 3$ so that each $p_{\beta,\mu\nu}$ with $\nu \in \Omega(1)$ is $p_{\beta,\mu\nu} < 1/L$, each $p_{\beta,\mu\nu}$ with $\nu \in \Omega(2)$ is $1/L \leqslant p_{\beta,\mu\nu} \leqslant 2/L$, and each $p_{\beta,\mu\nu}$ with $\nu \in \Omega(3)$ is $1 \geqslant p_{\beta,\mu\nu} > 2/L$. Associate the ordered elements in $\Omega(1)$ with $\{1, \ldots, \tau_1\}$; $\Omega(2)$ with $\{\tau_1 + 1, \ldots, \tau_1 + \tau_2\}$, $\Omega(3)$ with $\{\tau_1 + \tau_2 + 1, \ldots, L - \mu\}$. Calculate $\delta_\psi = \sum_{\nu\in\Omega(\psi)} p_{\beta,\mu\nu}$, $\psi = 1, 2, 3$, and $\delta_1 + \delta_2 + \delta_3 = 1$, for $\nu \notin \Omega(\psi)$, $\psi = 1, 2, 3$, $p_{\beta,\mu\nu} = 0$.
Step 4. Reduce the difference between the probabilities for $\Omega(1)$ and $\Omega(2)$ by making them equal within each group, while the elements in $\Omega(3)$ remain untouched.
Step 5. Define a new discrete PDF ϕ on {1, 2, 3} with $(\psi) = \delta_\psi$, $\psi = 1, 2, 3$.
Step 6. Define two new discrete PDFs $h_1(\sigma)$, $h_2(\sigma)$, which are uniformly distributed on $\{1, \ldots, \tau_1\}$ and $\{\tau_1 + 1, \ldots, \tau_1 + \tau_2\}$, respectively. Meanwhile, define a new $L - \mu - \tau_1 - \tau_2$ point discrete PDF $h_3(\sigma)$, on es_{ik} that is associated with the original probabilities of the numbers in $\Omega(3)$.
Step 7. Generate a random x with the outcomes 1, 2, 3 from as_{ik}. Given $x = \psi$, then generate a random variable r, $r \in [1, L - \mu]$ from the corresponding PDF $h_x(\sigma)$. Let x_μ be the column number ν of $\mathbf{p}_\beta$ that is matched to r.
Step 8. If $\mu < L$, set $\mu = \mu + 1$, $p_{\beta,\mu' x_\mu} = 0$, $\mu \leqslant \mu' \leqslant L$, return to Step 2; otherwise, set $\mathbf{X} = \{x_1, \ldots, x_\mu, \ldots .x_L\}$, terminate.

o_{ik} kth operation of job i;
es_{ik} earliest starting time of o_{ik};
pt_{ik} processing time of o_{ik};
ec_{ik} earliest completion time of o_{ik};
as_{ik} allowed starting time of o_{ik}.

Experimental studies and discussions

The proposed CE algorithm procedure is coded in JAVA and implemented on a Dell Precision T7400. To illustrate the effectiveness and performance of the proposed approach, four sets of experiments are carried out. The smooth parameter η is found empirically that a value η between $0.3 \le \eta \le 0.9$ gives good results; the quantile parameter ρ is suggested to take a larger ρ say, $\rho = \ln L/L$; the sample size Z is suggested to take as $N = cL(5 \le c \le 10)$; termination parameter d is set to 5–10 in general.[32] In this article, the parameters set as $\rho = 0.1$, $Z = 500$, $d = 5$ for the instance in Experiments 1–3; the parameters in Experiment 4 are set as $\rho = 0.1$, $Z = 9 \times L$, $d = 8$ based on initial test.

Sample decoding

Step 1. Construct an operation set with the first operations in the chosen process plan for all the jobs as $A := \{o_{i1} | i \in [1, N]\}$ based on the sample and set $ac_{i1} = 0, es_{i1} = 0, i \in [1, N]$.
Step 2. Determine o_{ik} with its number in the left-most position of **X** among all the positions corresponding to the operations in A and obtain the machine m on which o_{ik} performed.
Step 3. Scheduling o_{ik}

Step 3.1. Compute the allowable starting time of o_{ik} as $as_{ik} = ec_{i(k-1)} + tp(m', m)$, where $ec_{i(k-1)}$ is the completion time of the pre-operation of o_{ik} from the same job, and m' is the machine on which $o_{(k-1)}$ manufactured.

Step 3.2. Check if there exists idle area $[i\_s, i\_e]$ for o_{ik} in turn, if there is $\max(as_{ik}, i\_s) + pt_{ik}) \leqslant i\_e$, then set $es_{ik} = \max(as_{ik}, i\_s)$, $ec_{ik} = es_{ik} + p_{ik}$; otherwise, $es_{ik} = \max(as_{ik}, t_m(o_{i_1k_1}))$ ($o_{i_1k_1}$ is the pre-operation of o_{ik} performed on the same machine m, and $t_m(o_{i_1k_1})$ is the available time of machine m equals to the completion time of $o_{i_1k_1}$). If $as_{ik} \geqslant t_m(o_{i_1k_1})$, set $ec_{ik} = as_{ik} + pt_{ik}$, $t_m = ec_{ik}$ (the available time of m) and add a new idle area $[i\_s, i\_e] = [t_m(o_{i_1k_1}), as_{ik}]$ for m; otherwise, set $ec_{ik} = t_m(o_{i_1k_1}) + p_{ik}$, $t_m = ec_{ik}$.
Step 4. Update $A := A/\{o_{ik}\}$. Check o_{ik}, if it has valid (with non-zero detailed information) succeed operation $o_{i(k+1)}$ in the selected process plans, set $A := A \cup \{o_{i(k+1)}\}$.
Step 5. Check A, if $A = \emptyset$, terminate; otherwise, return to Step 2.

The parameter η with a wide value range is important for the experimental result according to our experience. In order to determine a more reasonable value among its range $[0.3, 0.9]$, some comparative experiments based on the problems in experiment 3 will be conducted and analyzed, and the $\eta = 0.8$ is finally selected as the preferred value in this article for all the problems.

The best solution (with minimized makespan) obtained at each run was taken. For computational comparisons, each test-bed problem was repeated five times. The computational times are the mean times of all five independent runs recorded. In this article, The Gantt chart is utilized to illustrate the scheduling results. In the Gantt chart, the notations out of the parentheses in each block represent the operation of a job and the notations in the parentheses indicate the completion time of the operation. The selected process plan of each job can also been determined easily based on the notations in each block and the precedence relations of the blocks.

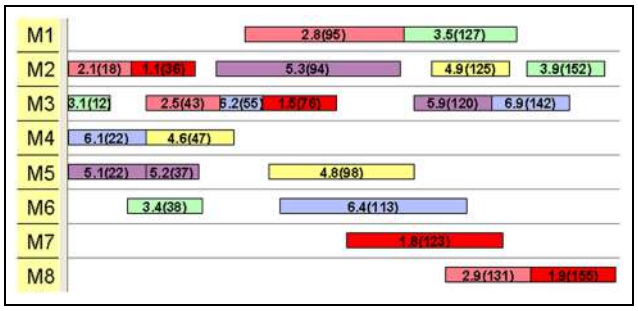


Figure 2. Gantt chart of problem 1 with CE (makespan = 155).

Experiment 1

In this Experiment, three problems are taken from the literature to compare the performance of the proposed CE with that of other approaches. The data of problem 1 is constructed with six jobs and eight machines designed by Shao et al.[4] Each job has nine operations in its process plan network. Problem 2 is constructed with 10 jobs and nine machines.[36] The data of problem 3 are constructed with six jobs and five machines.[5] The parameter s of CE is set to 4 for problems 1 and 2 based on some initial test. All the process plans of each job are passed on to the joint optimization for problem 3. The makespan values for these problems obtained by the proposed CE and methods in the literature are listed in Table 1. The values with bold type are used to emphasize the better computational results our proposed CE-based method achieved. Figures 2–4 illustrate the Gantt chart of the three problems in Experiment 1. The results indicate that the proposed CE-based method can obtain better result.

Experiment 2

The data of Experiment 2 which is constructed with six jobs and six machines by Saygin and Kilic,[37] job-related information and alternative machines for each operation of the jobs are given in Table 2. The number

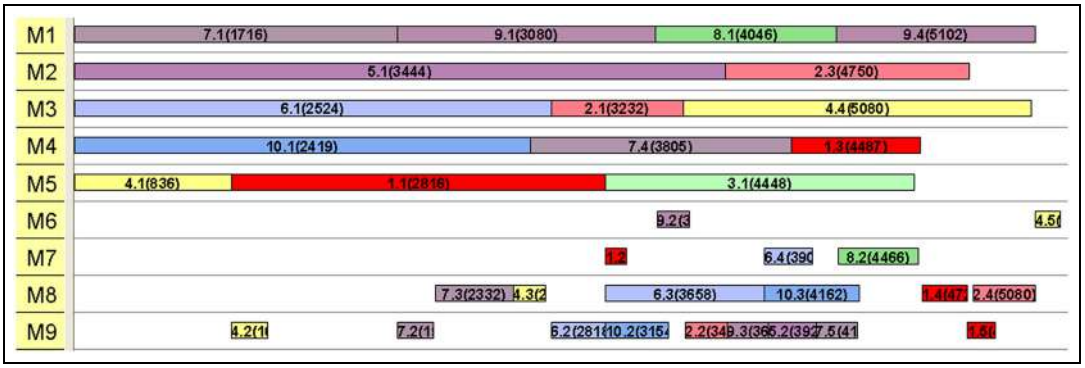


Figure 3. Gantt chart of problem 1 with CE (makespan = 5210).

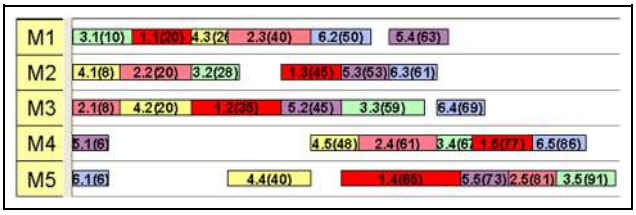


Figure 4. Gantt chart of problem 3 with CE (makespan = 91).

Table 1. Computational results of Experiment 1.

Pb	$n \times m$	Re	Approach	Makespan	CPU time (s)
1	6 × 8	4	MGA	162	_
		4	HIA	250	_
			CE	**155**	10.6
2	10 × 9	35	GA	6456	_
		23	oHAN	6574	_
		33	IGA	5268	_
			CE	**5210**	18.1
3	6 × 5	18	EA	92	_
		18	No integration	102	_
			CE	**91**	8.4

Pb: problem; Re: reference; *n* and: the total number of job and machine, respectively; –: data are not given in the Re; CE: cross-entropy; GA: genetic algorithm; EA: evolutionary algorithm; MGA: Modified Genetic Algorithm HIA: Hierarchical Approach oHAN: Online Hybrid Agent-based Negotiation IGA: Improved Genetic Algorithm.

s for Experiment 2 is set to 8 based on some initial experiments.

For the purpose comparison, an experiment for the separate implementation situation with which only one shortest process plan for each job has been passed on to CE, named as SPPCE here, has also been conducted. The shortest process plan in this experiment is the one with minimal production time from *s* process plans. Figures 5 and 6 illustrate the Gantt chart of Experiment 2 based on CE and SPPCE. The experimental results indicate that the joint implementation of process plan selection and scheduling optimization is superior to the situation with separate implementation. Meanwhile, the CE method can obtain different solutions with the same makespan (39) among five runs which indicates that the joint implementation of the problem has more alternatives and adjusting flexibility to react to the dynamic shop floor conditions to maintain the optimized result.

Experiment 3

The benchmark problem set used for Experiment 3 have been reported by Jain et al.[38] These problems involved 18 jobs, the jobs 1–3, 7–9 and 13–15 are with eight process plans, jobs 4, 5, 10, 11, 16 and 17 are with seven process plans and the other jobs are with six process plans. Shao et al.[4] constructed six problems with the 18 jobs. For the purpose of consistency, the job with process plans less than 8, the non-enough process plans are a copy of the last process plan for each job.

Table 2. Job-related information and alternative machines for each operation of Experiment 2.

Job	Process plan	Operation	Alternative machine (processing time)
1	1	1	1 (8) or 2 (12)
		2	3 (17) or 4 (11)
		3	3 (4) or 6 (6)
	2	1	3 (14) or 4 (18)
		2	5 (4)
		3	3 (6) or 4 (5)
2	1	1	4 (6)
		2	6 (4)
		3	1 (6) or 2 (10)
		4	5 (4) or 6 (6)
	2	1	4 (6)
		2	5 (3) or 6 (4)
		3	1 (9)
		4	6 (4)
3	1	1	3 (8) or 4 (10)
		2	5 (6)
		3	3 (4) or 4 (5)
	2	1	1 (4)
		2	4 (6)
		3	5 (6) or 6 (8)
		4	4 (6) or 3 (4)
4	1	1	1 (12) or 2 (18)
		2	6 (4)
		3	1 (4)
	2	1	3 (8) or 4 (12)
		2	2 (19)
		3	3 (6) or 4 (8)
		4	5 (4) or 6 (5)
5	1	1	1 (8)
		2	3 (10) or 4 (12)
		3	1 (6)
		4	5 (9) or 6 (9)
6	1	1	3 (7) or 4 (7)
		2	1 (9)
		3	2 (10)
		4	6 (6) or 5 (6)

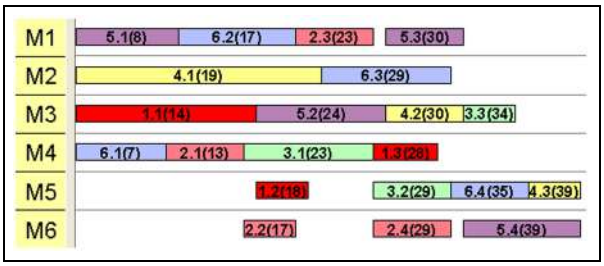


Figure 5. Gantt chart of Experiment 2 based on CE.

Figure 6. Gantt chart of Experiment 2 based on SPPCE.

Table 3. Computational result of Experiment 3.

Pb	Jobs	Makespan (CPU time/s)			
		CE	EA[a]	No integration[a]	ICA[a]
1	8	**499** (3.34)	520 (3.28)	615 (3.42)	499 (2.98)
2	10	607 (3.27)	621 (3.38)	831 (3.72)	586 (3.26)
3	12	697 (3.54)	724 (3.67)	934 (3.91))	679 (2.73)
4	14	**801** (3.73)	809 (3.69)	1004 (4.14)	803 (2.50)
5	16	**900** (4.01)	921 (3.73)	1189 (4.39)	900 (2.49)
6	18	**958** (4.58)	994 (4.09)	1249 (4.69)	976 (3.87)

Pb: problem; CE: cross-entropy; EA: evolutionary algorithm; ICA: imperialist competitive algorithm.
[a]Results obtained by the approaches are adopted from Lian et al.[1].

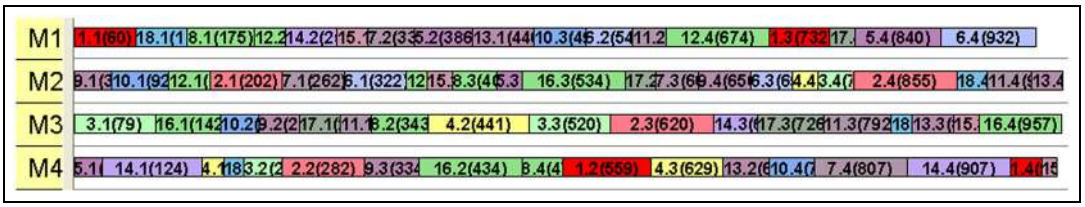


Figure 7. Gantt chart of problem 6 in Experiment 3 (makespan = 958).

Therefore, the number s for Experiments 3 is set to 8. The values set in bold type in Table 3 are utilized to emphasize the better computational results from our proposed approach.Table 3 shows the experimental results and Figure 7 illustrates the Gantt chart of the solution found for the problem 6. It can be seen that the CE outperforms the EA on all six problems and outperforms the ICA on two problems and two problems are with the same result. CE, EA and no integration methods perform no significant difference which is outperformed by ICA.

Figures 8 and 9 illustrate the convergence of matrix $\mathbf{p}_\alpha$ (with 12 jobs and each job with eight process plans and its dimension is 12×8) and $\mathbf{p}_\beta$ (dimension is 48×48), respectively, for the problem 3. It can be seen from Figures 8 and 9 that the two matrix parameters can converge to the optimal values simultaneously with only 53 iterations. Meanwhile, it can be observed obviously from Figure 8(d) that row vector elements of $\mathbf{p}_\alpha$ converge to near (0, 0, 0, 1, 0, 0, 0, 0), (0, 1, 0, 0, 0, 0, 0, 0), (0, 1, 0, 0, 0, 0, 0, 0), (1, 0, 0, 0, 0, 0, 0, 0), (1, 0, 0, 0, 0, 0, 0, 0), (0, 0, 0, 1, 0, 0, 0, 0), (0, 1, 0, 0, 0, 0, 0, 0), (0, 1, 0, 0, 0, 0, 0, 0), (1, 0, 0, 0, 0, 0, 0, 0), (1, 0, 0, 0, 0, 0, 0, 0), (1, 0, 0, 0, 0, 0, 0, 0) and (1, 0, 0, 0, 0, 0, 0, 0), respectively; it means that 4, 2, 2, 1, 1, 4, 2, 2, 1, 1, 1, 1 process plan has the most opportunities been selected for the 12 jobs in the problem by CE iteration, and it can be seen as the optimal selected process for the 12 jobs for this time computation. However, not all the elements in $\mathbf{p}_\beta$ converge to near 1 or 0 for the final iteration in Figure 9(d); it still can been observed that some elements rise apparently for each row and it means that the corresponding operation number has the most opportunities been selected to for the position.

In order to determine a more reasonable value among its range for the pivotal parameter η, comparative experiments based on the above six problems have been conducted with 0.1 as spacing step among its range [0.3, 0.9]. The experiments result are shown in Figure 10; it shows that as the smooth parameter η increases, mean makespans for major problems go downward and then upward. The valley point usually appears at around $\eta = 0.8$. The reason for the downward trend is clear that the decreased ration of previous matrix parameters reduces the probability of zero or one components in the matrix at the first few iterations, and then increases the probability to generate sample with better solutions. However, the following upward trend is mainly because that low ration of inherited matrix parameters may prematurely guide some components in the matrix to be zero or one which will cause the algorithm converges to a local optimum early. It can be seen that the balance of sample generation parameter from previous iteration and currently updated matrix is vital for CE in this problem, and the smooth parameter η is set to 0.8 in this article based on the above test.

Experiment 4

Experiment 4 is adopted from the benchmark reported by Kim.[39] In this experiment, 24 test-bed problems based on 18 jobs with various combinations of flexibility levels and 15 machines are constructed. The number

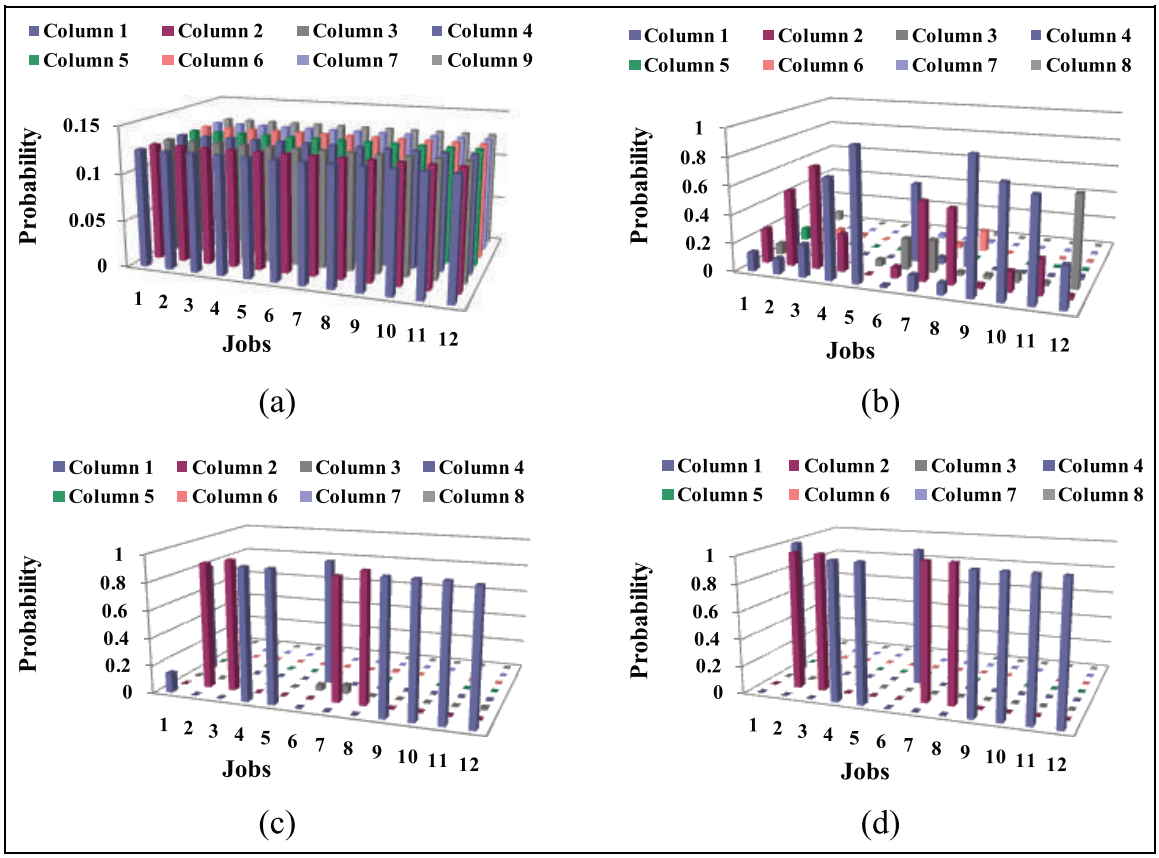


Figure 8. Convergence of matrix $\mathbf{p}_\alpha$ for the problem 3 in Experiment 3: (a) 0 iteration, (b) 10 iterations, (c) 30 iterations and (d) 52 iterations (final).

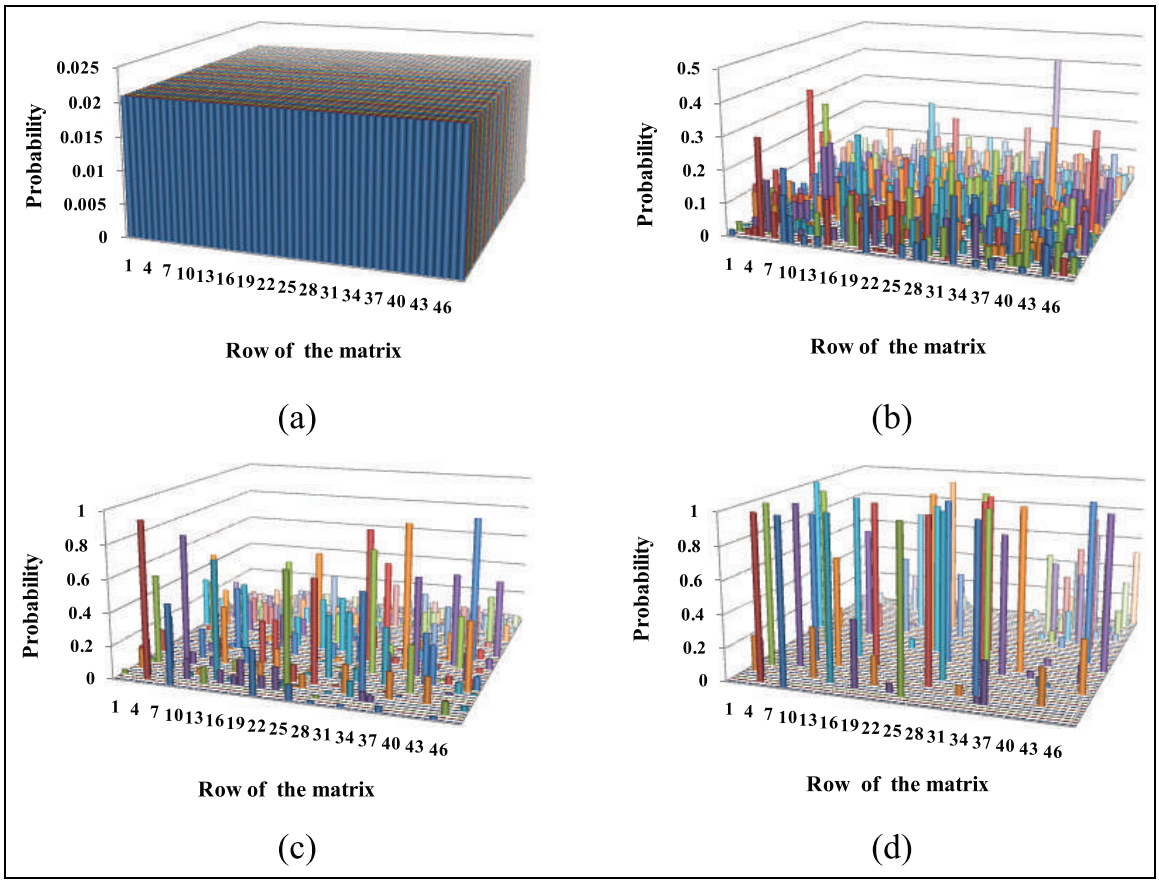
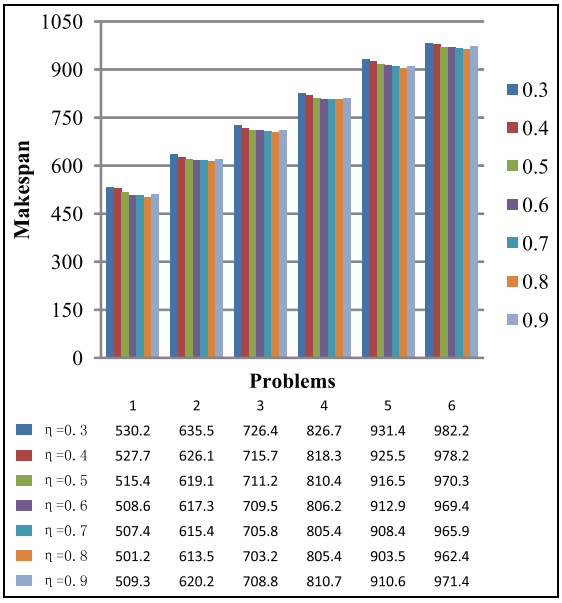


Figure 9. Convergence of matrix $\mathbf{p}_\beta$ for the problem 3 in Experiment 3: (a) 0 iteration, (b) 10 iterations, (c) 30 iterations and (d) 52 iterations (final).



	1	2	3	4	5	6
η=0.3	530.2	635.5	726.4	826.7	931.4	982.2
η=0.4	527.7	626.1	715.7	818.3	925.5	978.2
η=0.5	515.4	619.1	711.2	810.4	916.5	970.3
η=0.6	508.6	617.3	709.5	806.2	912.9	969.4
η=0.7	507.4	615.4	705.8	805.4	908.4	965.9
η=0.8	501.2	613.5	703.2	805.4	903.5	962.4
η=0.9	509.3	620.2	708.8	810.7	910.6	971.4

Figure 10. Mean makespans for the problems in Experiment 3 with different η.

Table 4. Computational results of Experiment 4.

No.	SEA[a]	HA[a]	ALGA[a]	ICA	CE
1	428	427	427	427	**427**
2	343	343	343	343	**343**
3	347	345	344	345	**344**
4	306	306	306	306	**306**
5	319	322	321	319	**315**
6	438	429	429	435	**429**
7	372	372	372	372	**372**
8	343	343	343	343	**343**
9	428	427	427	427	**427**
10	443	430	427	440	**427**
11	369	369	369	367	**365**
12	328	327	327	327	**322**
13	452	436	436	457	**433**
14	381	380	380	390	398
15	434	427	427	432	**427**
16	454	446	446	466	448
17	431	423	423	443	424
18	379	377	377	384	**375**
19	490	476	474	490	480
20	447	432	438	440	**430**
21	477	446	447	466	**442**
22	534	518	513	529	**512**
23	498	470	470	495	471
24	587	544	548	577	**528**

ICA: imperialist competitive algorithm; CE: cross-entropy.
[a]Results obtained by the approaches are adopted from Qiao and Lv.[32]
The values with bold type are used to emphasize the better computational results (or equal to the reported best results) our proposed CE based method achieved.

of operations of these problems varies from 79 to 300. The number s for Experiment 4 is set to 8. The best result for makespan obtained by the CE-based approach, ICA,[1] Symbiotic Evolutionary Algorithm (SEA),[2] HA[6] and Active Learning Genetic Algorithm (ALGA)[21] are listed in Table 4.

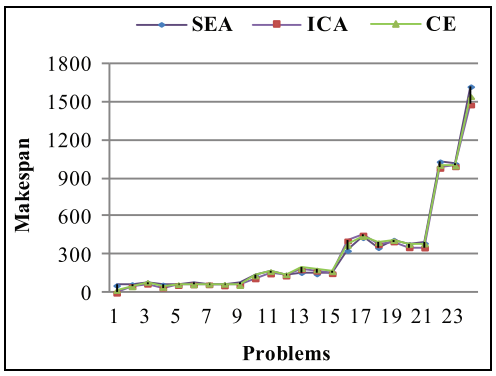


Figure 11. Comparison of CPU time for different algorithms.

It can be observed from Table 4 that the CE outperforms the ICA on 17 problems and 6 problems are with the same result. Compared to the best results obtained by SEA, HA, ALGA and ICA, the same or better solutions have been found by CE for 19 out of the 24 testbed problems. Figure 11 shows the comparison result of CPU time for SEA, ICA and CE, in which the data of ICA and SEA are adopted from Lian et al.[1] and Kim et al.[2] (HA and ALGA with no CPU time provided). It can be seen that no algorithm has significant advantage for the 24 problem from the CPU time view. Figure 12 illustrates the Gantt chart of the solution found for the problem 24.

Conclusion

In order to achieve greater performance and higher productivity of a manufacturing system, the novel CE method has been developed with new introduced probability distribution and updating approach, parallel sample encoding/decoding and generation mechanism for the JPPSSO problem with two sub-problems. The proposed CE-based approach can facilitate the joint implementation of process plan selection and scheduling optimization. The result of JPPSSO based on CE can give aid to the schedule planners to determine the optimal scheduling plans and assist the PP system to determine the final process plan (including the determination of operations, machine for each operation and operations sequence) for each job that will be passed on to the production systems from some alternative process plans simultaneously. Experimental studies have been conducted and comparisons have been made among CE and other developed methods to indicate the superiority and adaptability of the proposed approach. The experimental results show that the proposed CE method has advantage in the makespan objective for the test problems with reasonable CPU time, which provides an alternative and acceptable method in the research of JPPSSO.

The outstanding performance of the CE in makespan is mainly due to the following two factors. First of all, the reasonably introduced two probability

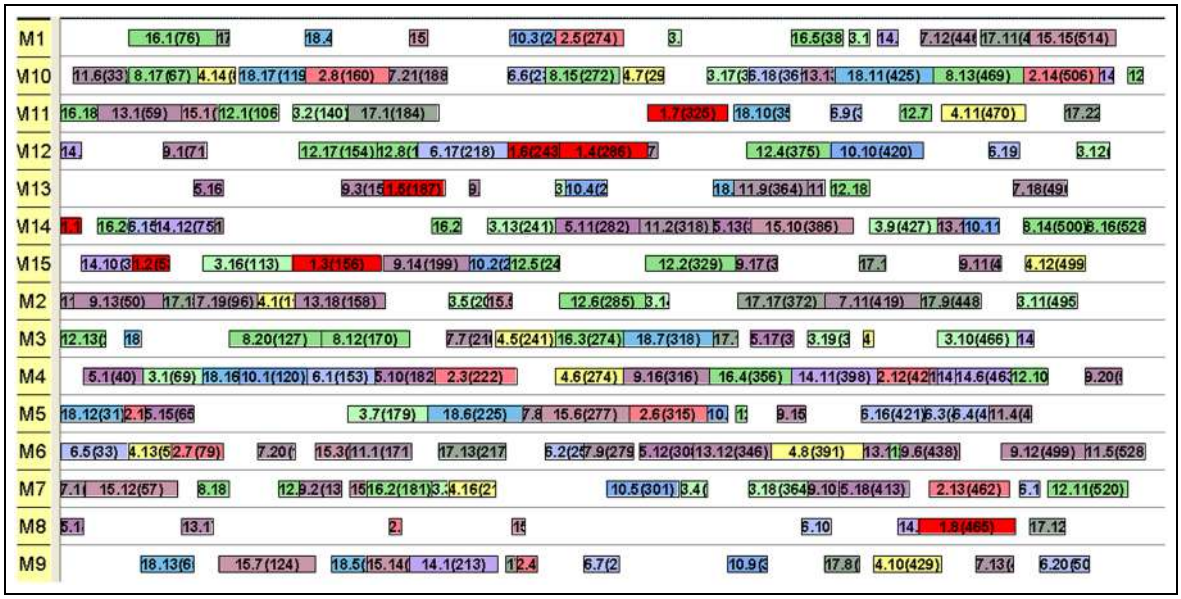


Figure 12. Gantt chart of problem 24 in Experiment 4 (makespan = 528).

matrices and the related sample generation mechanism enable whole solution space can be covered. Then the smoothed updating mechanism can reduce the probability that some of component in the two matrices to be zero or one at the first few interactions, and then help the iteration procedure avoid premature and increase the probability to generate sample with better solutions. However, the CE method has no significant advantage in CPU time comparing to other algorithms. The reason is that the generation of operation-based scheduling plan section in a sample according to the new introduced probability matrix consumes much CPU time.

The further research direction can be conducted from the following aspects:

- To design more efficient method to express and generate samples especially for larger scale problems and conduct experiment with different parameter levels to determine their reasonable values.
- The proposed algorithm should be studied continually for its practical application in manufacturing system by considering uncertainties, such as new job arrival, machine breakdown and so on.
- The mean flow time, resource utilization and so on multi-objectives should also been investigated.
- Furthermore, CE can be extended to cope with the JPPSSO requirements of other scenarios, for instance, assembly operations, integration of electronic product testing process and scheduling.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was financially funded by the Natural Science Foundation of Guangdong, China (grant number: 2014A030310345).

References

1. Lian KL, Zhang CY, Gao L, et al. Integrated process planning and scheduling using an imperialist competitive algorithm. *Int J Prod Res* 2011; 50(15): 4326–4343.
2. Kim YK, Park K and Ko J. A symbiotic evolutionary algorithm for the integration of process planning and job shop scheduling. *Comput Oper Res* 2003; 30(8): 1151–1171.
3. Kumar M and Rajotia S. Integration of scheduling with computer aided process planning. *J Mater Process Tech* 2003; 138(1–3): 297–300.
4. Shao XY, Li XY, Gao L, et al. Integration of process planning and scheduling—a modified genetic algorithm-based approach. *Comput Oper Res* 2009; 36(6): 2082–2096.
5. Li XY, Shao XY and Zhang CY. Mathematical modeling and evolutionary algorithm-based approach for integrated process planning and scheduling. *Comput Oper Res* 2010; 37(4): 656–667.
6. Li XY, Shao XY, Gao L, et al. An effective hybrid algorithm for integrated process planning and scheduling. *Int J Prod Econ* 2010; 126(2): 289–298.
7. Li XY, Zhang CY, Gao L, et al. An agent-based approach for integrated process planning and scheduling. *Expert Syst Appl* 2010; 37(2): 1256–1264.
8. Phanden RK, Jain A and Verma R. Integration of process planning and scheduling: a state-of-the-art review. *Int J Comp Integ M* 2011; 24(6): 517–534.
9. Li WD and McMahon CA. A simulated annealing-based optimization approach for integrated process planning and scheduling. *Int J Comp Integ M* 2007; 20(1): 80–95.

10. Cai N, Wang L and Feng HY. GA-based adaptive setup planning toward process planning and scheduling integration. *Int J Prod Res* 2009; 47(10): 2745–2766.
11. Zhang LP and Wong TN. An object-coding genetic algorithm for integrated process planning and scheduling. *Eur J Oper Res* 2015; 244: 434–444.
12. Guo YW, Li WD, Mileham AR, et al. Applications of particle swarm optimisation in integrated process planning and scheduling. *Robot Cim: Int Manuf* 2009; 25(2): 280–288.
13. Guo YW, Li WD, Mileham AR, et al. Optimisation of integrated process planning and scheduling using a particle swarm optimisation approach. *Int J Prod Res* 2009; 47(14): 3775–3796.
14. Baykasoğlu A and Özbakr L. A grammatical optimization approach for integrated process planning and scheduling. *J Intell Manuf* 2009; 20(2): 211–221.
15. Moon C and Seo Y. Evolutionary algorithm for advanced process planning and scheduling in a multi-plant. *Comput Ind Eng* 2005; 48(2): 311–325.
16. Moon C, Lee YH, Jeong CS, et al. Integrated process planning and scheduling in a supply chain. *Comput Ind Eng* 2008; 54(4): 1048–1061.
17. Wang S, Lu X, Li XX, et al. A systematic approach of process planning and scheduling optimization for sustainable machining. *J Clean Prod* 2015; 2015(87): 914–929.
18. Wang JF, Fan XL, Zhang CW, et al. Graph-based ant colony optimization approach for integrated process planning and scheduling. *Chinese J Chem Eng* 2014; 22: 748–753.
19. Zhang R, Ong SK and Nee AYC. A simulation-based genetic algorithm approach for remanufacturing process planning and scheduling. *Appl Soft Comput* 2015; 37: 521–532.
20. Li XY, Gao L and Li WD. Application of game theory based hybrid algorithm for multi-objective integrated process planning and scheduling. *Expert Syst Appl* 2011; 39(1): 288–297.
21. Li XY, Gao L and Shao XY. An active learning genetic algorithm for integrated process planning and scheduling. *Expert Syst Appl* 2012; 39(8): 6683–6691.
22. Yu MR, Zhang YJ, Chen K, et al. Integration of process planning and scheduling using a hybrid GA/PSO algorithm. *Int J Adv Manuf Tech*. Epub ahead of print 10 December 2014; DOI: 10.1007/s00170-014-6669-7.
23. Wong TN, Leung CW, Mak KL, et al. An agent-based negotiation approach to integrate process planning and scheduling. *Int J Prod Res* 2006; 44(7): 1331–1351.
24. Wong TN, Leung CW, Mak KL, et al. Integrated process planning and scheduling/rescheduling—an agent-based approach. *Int J Prod Res* 2006; 44(18–19): 3627–3655.
25. Wong TN, Leung CW, Mak KL, et al. Dynamic shopfloor scheduling in multi-agent manufacturing system. *Expert Syst Appl* 2006; 31(3): 486–494.
26. Leung CW, Wong TN, Mak KL, et al. Integrated process planning and scheduling by an agent-based ant colony optimization. *Comput Ind Eng* 2010; 59(1): 166–180.
27. Shukla SK, Tiwari MK and Son YJ. Bidding-based multi-agent system for integrated process planning and scheduling: a data-mining and hybrid tabu-SA algorithm-oriented approach. *Int J Adv Manuf Tech* 2008; 38(1–2): 163–175.
28. Ueda K, Fujii N and Inoue R. An emergent synthesis approach to simultaneous process planning and scheduling. *CIRP Ann: Manuf Techn* 2007; 56(1): 463–466.
29. Hsieh FS and Chiang CY. Collaborative composition of processes in holonic manufacturing systems. *Comput Ind* 2011; 62(1): 51–64.
30. Hsieh FS and Lin JB. A dynamic scheme for scheduling complex tasks in manufacturing systems based on collaboration of agents. *Appl Intell* 2014; 41(2): 366–382.
31. Shen W, Wang L and Hao Q. Agent-based distributed manufacturing process planning and scheduling: a state-of-the-art survey. *IEEE T Syst Man Cy C* 2006; 36(4): 563–577.
32. Qiao LH and Lv SP. An improved genetic algorithm for integrated process planning and scheduling. *Int J Adv Manuf Tech* 2012; 58(5–8): 727–740.
33. Rubinstein RY. The cross-entropy method for combinatorial and continuous optimization. *Methodol Comput Appl* 1999; 1(2): 127–190.
34. Rubinstein RY and Kroese DP. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. New York: Springer, 2004.
35. Jain AK and Elmaraghy HA. Production scheduling/rescheduling in flexible manufacturing. *Int J Prod Res* 1997; 35: 281–309.
36. Saygin C and Kilic SE. Integrating flexible process plans with scheduling in flexible manufacturing systems. *Int J Adv Manuf Tech* 1999; 15(4): 268–280.
37. Jain A, Jain PK and Singh IP. An integrated scheme for process planning and scheduling in FMS. *Int J Adv Manuf Tech* 2006; 30(11–12): 1111–1118.
38. Kim YK. A set of data for the integration of process planning and job shop scheduling, 2003, http://syslab.jnu.ac.kr/index.php?mid=board_PDAS73&document_srl=528(data-pp&s-C&OR(2003))

ISSN 1002-8331
CODEN JGYYAT

Vol.55 No.20 15 Oct. 2019

计算机工程与应用

Computer Engineering and Applications

中国计算机学会会刊
北大中文核心期刊
中国科技核心期刊
中国科学引文数据库（CSCD）来源期刊
RCCSE中国核心学术期刊
《中国学术期刊文摘》源期刊
英国SA/INSPEC收录期刊
俄罗斯《文摘杂志》收录期刊
美国《剑桥科学文摘》收录期刊
美国《乌利希期刊指南》收录期刊
波兰《哥白尼索引》收录期刊
《日本科学技术振兴机构数据库》收录期刊

第55卷 第20期 2019年10月15日

20
2019年

华北计算技术研究所

计算机工程与应用

Jisuanji Gongcheng yu Yingyong

第55卷　第20期　2019年10月15日

目　　次

·工程与应用·

计算机工程与应用

Jisuanji Gongcheng yu Yingyong

(半月刊,1964年创刊)

第55卷　第20期(总第939期)　2019年10月15日

Computer Engineering and Applications

(Semimonthly)

Started in 1964

Vol.55　No.20(Sum No.939)　15 Oct. 2019

主管单位　中国电子科技集团公司	China Electronics Technology Group Corporation
主办单位　华北计算技术研究所	Sponsored by North China Institute of Computing Technology
编委会主任　刘学林	Director of Editorial Board:LIU Xuelin
主　　编　谭继红	Editor-in-Chief:TAN Jihong
总 编 辑　陶小雪	Chief Editor:TAO Xiaoxue
执行总编辑　丁宇萍	Executive Chief Editor:DING Yuping
编辑出版　北京《计算机工程与应用》期刊有限公司 通信地址:北京619信箱26分箱　邮编:100083 电话:(010)89055542　http://www.ceaj.org 电子信箱:ceaj@vip.163.com	Edited and Published by Journal of *Computer Engineering and Applications* Beijing Co., Ltd. Mail Address:No.26,P.O.Box 619,Beijing 100083,P.R.China Tel:(8610)89055542　http://www.ceaj.org E-mail:ceaj@vip.163.com
国内总发行　中国邮政集团公司北京市报刊发行局	Distributed by Beijing Bureau for Distribution of Newspapers and Journals
订 阅 处　全国各地邮局(82-605)	Subscribed by All Local Post Offices in China
国外总发行　中国国际图书贸易集团有限公司	Foreign Distributed by China International Book Trading Corporation
印　　刷　廊坊市祥丰印刷有限公司	Printed by Langfang Xiangfeng Printing Limited Company

深度学习在我国农业中的应用研究现状

吕盛坪,李灯辉,冼荣亨
华南农业大学 工程学院,广州 510642

摘 要:深度学习(Deep Learning,DL)已广泛应用于智能农业的病虫害检测、植物和水果识别、农作物及杂草检测与分类等研究中。对2014年至2019年国内发表的65篇有关DL在农业中应用研究成果进行综述。简要介绍DL的基本概念及其发展历史,给出了所选论文检索方法及其分布;对所选论文从研究对象与目的、数据来源、类间差异、预处理、数据扩增、模型框架以及性能对比等角度进行了综述;对DL的优缺点进行了分析,并指明了其在智能农业研究中的发展趋势。

关键词:深度学习;智能农业;检测;识别;分类;预测
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1907-0089

吕盛坪,李灯辉,冼荣亨.深度学习在我国农业中的应用研究现状.计算机工程与应用,2019,55(20):24-33.
LV Shengping, LI Denghui, XIAN Rongheng. Research status of deep learning in agriculture of China. Computer Engineering and Applications, 2019, 55(20):24-33.

Research Status of Deep Learning in Agriculture of China

LV Shengping, LI Denghui, XIAN Rongheng
College of Engineering, South China Agricultural University, Guangzhou 510642, China

Abstract: Deep Learning(DL) has been widely used in intelligent agriculture for plant disease detection, plant and fruit recognition, crop and weed detection and classification and so on. 65 articles from 2014 to 2019 on the application of DL in agriculture of China are presented. At first, the basic concept and development history of DL are briefly introduced, and the article retrieval and distribution of the reviewed articles are given. Subsequently, the articles are reviewed from various points of view such as research object, data source, inter-class differences, preprocessing, data-augmentation, framework and performance comparison. Eventually, the advantages and disadvantages of DL are analyzed, and its development trends in agriculture are demonstrated.

Key words: deep learning; intelligent agriculture; detection; recognition; classification; prediction

1 引言

为了更好地监测和分析各种作物和动物生长状态,新的信息和通信技术被大量使用,比如基于摄像机的图像采集和基于传感器的环境监控等[1]。如何快速识别处理这些图像和结构化的监测数据以支持智能决策是智能农业领域的重要研究方向。传统处理技术包括机器学习(*K*-means聚类、支持向量机、人工神经网络等),线性极化,小波滤波。近年来,深度学习(Deep Learning,DL)被大量采用,特别是在病虫害检测、植物和水果识别、农作物及杂草检测与分类等智能农业领域[2-3]。

DL是机器学习研究中的一个分支,其通过组合低层特征形成更抽象的高层表示属性类别或特征,以发现数据的分布式特征[4]。至今,DL已经广泛应用于图像识别[5]、物体分类与检测[6]、人脸识别[7-8]和语音识别[9]等。

相对于传统机器学习,DL能更好地提取农业领域所采集图像和结构化数据的各种特征,并与农业机械有效结合,更好地支持农业智能机械装备的开发。因此,近年来,DL受到农业领域的高度重视,相应研究成果不断涌现。Kamilaris等[10]对国外近年来DL在农业领域中的应用进行了全面综述。本文对近年来国内农业领域

基金项目:国家自然科学基金(No.51605169)。
作者简介:吕盛坪(1982—),通讯作者,男,博士,副教授,主要从事农业/工业数据挖掘、机器学习、深度学习研究,E-mail:lvshengping@scau.edu.cn。
收稿日期:2019-07-08 **修回日期:**2019-08-22 **文章编号:**1002-8331(2019)20-0024-10
CNKI**网络出版:**2019-08-23, http://kns.cnki.net/kcms/detail/11.2127.TP.20190823.1141.016.html

DL的应用现状进行综述，一方面，为农业研究者提供可用的DL方法参考；另一方面，以便于研究者快速精确地检索与所研究问题相关的文献。本综述框架如图1所示。

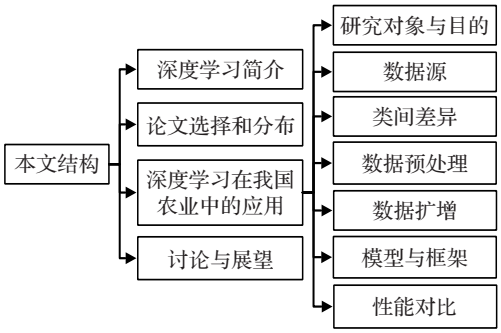


图1　本文框架

2　深度学习简介

DL最早由Hinton及其团队于2006年提出，Hinton等[11]通过逐层初始化人工神经网络解决了大规模多层神经网络在训练速度上的难题，奠定了DL的基础。2012年，Hinton及其学生通过Rectified Linear Neurons（RLN）和Dropout正则化改进了卷积神经网络，并在ImageNet分类竞赛上，将错误率降低至16%[12]。在接下来的几年中，研究者对其DL神经网络进行了不断改进，并将ImageNet分类错误率降低至零点几个百分点。2012年的突破标志着基于DL的人工智能繁荣的开始。2015年，LeCun，Bengio及Hinton在Nature上共同发表了《Deep learning》的综述，对DL进行了定义：DL是一种更复杂的表示学习，具有多个级别的表示，它通过组合简单但非线性的模块获得，每个模块将一个级别的表示（从原始输入开始）转换为更高、稍微抽象的级别的表示；有了足够多的这种变换的组合，就可以学习非常复杂的模式；对于分类任务，较高的表示层会放大输入中对识别重要的特征，并抑制无关变化[4]。2019年3月27日，ACM（Association for Computing Machinery）将2018年的图灵奖授予给了Hinton、LeCun和Bengio，以奖励这三位科学家在DL基本概念的发明、实验中惊人结果的发现及其在工程应用中的重要突破等方面做出的重要贡献。

DL的强大优势是特征学习，即从原始数据中自动提取特征，由较低层次特征的组合形成更高层次的特征[4]。不同的DL由各种不同的组件（例如卷积、池化层、完全连接层、门、内存单元、激活函数、编码/解码器等）构成，具体取决于所使用的网络类型。当前主要网络类型有多层感知器（Multi-Layer Perceptron，MLP）[13]、卷积神经网络（Convolutional Neural Network，CNN）[14-15]、深度置信网络（Deep Belief Network，DBN）[16]、递归神经网络（Recursive Neural Network，RNN）[17]等，其中CNN是农业中最常用的一种网络模型。

MLP是一种前馈人工神经网络模型，其将输入的多个数据集映射到单一的输出的数据集上。CNN是一类包含卷积计算且具有深度结构的前馈神经网络，具有表征学习能力，能够按其阶层结构对输入信息进行平移不变分类。DBN为概率生成模型，通过联合概率分布推断出数据样本分布，其中生成模型通过训练网络结构中的神经元间的权重使得整个神经网络依据最大概率生成训练数据，形成高层抽象特征，提升模型分类性能。RNN是具有树状阶层结构且网络节点按其连接顺序对输入信息进行递归的人工神经网络，其可以引入门控机制以学习长距离依赖，具有灵活的拓扑结构且权重共享，适用于包含结构关系的机器学习任务，在自然语言处理领域有重要应用[4]。

随着DL的快速发展，各种网络架构被提出。常用的网络架构有Lenet[18]、AlexNet[19]、CaffeNet[20]、VGGNet[21]、GoogleNet[22]、ResNet（Residual Neural Network）[23-24]、Network in network[25-26]、ResNeXt[27]等，研究人员在这些网络架构基础上又相继提出了RCNN[28]、SPPNet[29]、SSD[30]、Fast R-CNN[31]、Faster R-CNN[32]、YOLO[33]等架构。这些架构的核心思想主要体现在两个方面：一是它们的神经元间的连接是非全连接的；另一个是同一层中某些神经元之间的连接的权重是共享的。这种非全连接和权值共享的网络结构使它们更类似于生物神经网络，由此可以降低网络模型的复杂度，减少权值的数量。上述架构均被一些数据集预先训练过网络参数，能为某些特定问题提供较好的分类、检测和识别效果。用于训练DL的常见数据集是ImageNet[34]、PASCAL VOC、Labelme、COCO、SUN、PlantVillage等。

同时，各种DL框架[35]被开发出来，例如加州大学伯克利分校视觉与学习中心维护的Caffe[36]，谷歌的TensorFlow[37]，微软研究院的CNTK2.0[38]，Facebook、Twitter、Google等维护的Torch[39]，蒙特利大学的Theano[40]等。这些框架为DL模型的开发、训练、测试、微调提供了统一平台；且每一框架各自具有统一的代码风格、模板化的结构，能减少DL开发大量重复代码的编写[41]。

3　论文选择和分布

农业是指利用动植物的生长发育规律，通过人工培育来获得产品的领域，研究对象主要是有生命的动植物及其场地与环境等。相应DL主要集中于影响植物生长发育的土壤水分及营养、温湿度、病虫害和影响动物健康生长的饲料营养、病害等领域。通过国内数据库对深度学习、卷积神经网络、农业、农田、动物、土地覆盖、土壤水分、温度、病虫害等关键词进行搜索，发现大部分相关研究成果发表于2015年之后，且集中于种植和养殖业。因此，本文选取了2015年到2019年3月间在这些领域中应用DL的相关文章。2019年3月，开展相应检

索,相应条件设置如下:

(1)数据库:中国知网、万方数据库;(2):深度学习;(3)植物类别:小麦、水稻、玉米、棉花、甜菜、黄瓜、烟叶、油茶、菊花等农作物,苹果、柑橘、番茄等水果;(4)动物类别:猪、牛;(5)其他类别:土地覆盖、土壤水分、温度。

通过人工进一步筛选最终确定65篇发表于核心期刊上的研究成果,所选论文在不同年份、不同类别研究对象的分布情况如图2所示。从图2中论文分布看,近2年DL在农业中的应用研究快速增加,其中2018年相关研究成果为40篇,占比超过61%。从研究对象看,80%(52篇)研究对象为植物,17%(11篇)涉及土壤水分、温度等资源环境分析,不到3%(2篇)研究对象涉及动物;说明现阶段DL在我国农业中的应用主要侧重于植物分类识别等研究。

	2015年	2016年	2017年	2018年	2019年
植物	2	1	13	30	6
动物	0	0	0	2	0
其他	1	1	0	7	2

图2　所选论文分布

4　深度学习在我国农业中的应用

下面将从研究对象与目的、数据源、数据差异、预处理、数据扩增、模型与框架以及性能对比等角度对所选论文进行综述。

4.1　研究对象与目的

广义农业包括种植业、林业、畜牧业、渔业、副业。表1给出了DL应用研究对象及其应用目的。所综述论文其研究对象和应用目的分布如图3所示。由表1及图3可知DL在农业中应用主要集中于种植业和畜牧业。其中DL在种植业中主要集中于研究对象的分类、检测、识别;耕作场地和耕作环境的预测;而畜牧业主要集中于动物对象的识别。

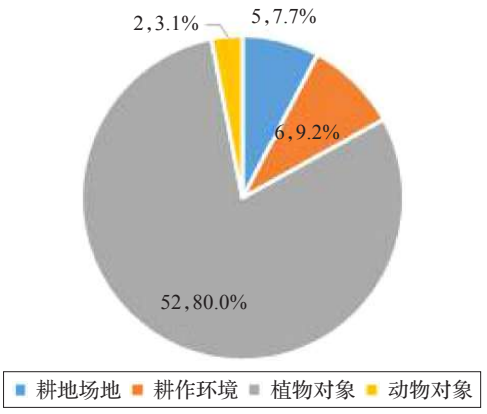
4.2　数据源

DL较其他算法能提高精准度,但其前提是有足够大的可用数据集来描述问题。所综述论文中数据类型及获取方式分类如表2所示。由表2中数据类型可知,农业中用于构建DL模型的常用数据类型有图像和结构化数值数据,其中以图像为主。

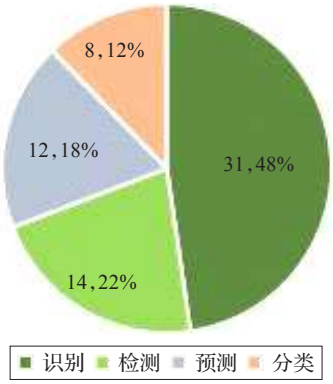
从表2中数据获取方式看,图像数据集的获取可分为自行采集和公开数据集,自行采集图像数据常通过无人机遥感、地面相机拍照或录像、搭载相机的无人机航

表1　所综述论文研究对象与目的

类别	对象	目的	参考文献
植物	植物病害	检测	[42-61]
	植物	识别	[62-74]
	水果(番茄/苹果/柑橘/荔枝/皇帝柑)	检测	[75-80]
	杂草	分类	[81-84]
	作物遮挡	检测	[85-86]
	龙眼叶片叶绿素含量	预测	[87-88]
	茶叶/青梅分级	分类	[89-90]
	苹果病变	识别	[91-92]
	林地图像	预测	[93]
土地	耕地面积	预测	[94-95]
	大棚/玉米倒伏区	检测	[96-97]
	农用地基准地价	预测	[98]
土壤	土壤水分	预测	[99-100]
	土壤孔隙	检测	[101]
温室	温室环境	预测	[102-103]
农田	农田障碍物	识别	[104]
动物	肉牛形体部位	识别	[105]
	猪只饮水行为	识别	[106]



(a)研究对象分布



(b)应用目的分布

图3　所选论文研究对象和应用目的分布

拍、高光谱成像仪、近红外光谱仪等方式获得。公开数据集一般来源于现有知名公开的标准库,如MIT[104]、Oxford-17 flower[62]、Oxford-102 flower[62]、PlantVillage[43]、Flavia[64]、ICL[69]、ImageNet[72]和Kaggle[74]等。结构化数值数据主要通过传感器在线监测获取。

从表2中样本规模看,研究针对具体应用场景自行

表2 数据源

类型	获取方式	样本规模	应用	参考文献
图像数据	公开数据库	3 000~28 000张	植物叶片/病害识别	[43,62,64,69,72,74,83,104]
	网上收集	400~2 472张	植物病害检测	[46,50,53,104]
	无人机遥感/搭载相机航拍	5~5 149张	土地覆盖分类、植物识别	[47,67-68,84,93-94,96-97]
	支架承载相机拍照/录像	550~40 000张	水果病变、动物生长、植物病害识别,植物分类	[42,45,49-50,52,55-59,61,63,65-66,70,73,75-82,85-86,89-92,105-106]
	高光谱成像仪	300张	龙眼叶片叶绿素含量预测	[60,87-88]
	近红外光谱仪	289~600张	烟叶分类、土壤含水率预测	[71,100]
	螺旋CT机	4 956张	土壤断层孔隙分割	[101]
数值数据	传感器监测	4 858~31 076条	温室温度预测、农田障碍物识别、植物病害检测	[44,48,51,54,99,102-103]
	调查/GPS定位	11 496条	农用地基准地价评估	[98]
	统计年鉴数据	14年	耕地面积预测	[95]

表3 类间差异分析

类型	特征部位	差异	参考文献
不同品种	生存条件差异	时间、位置、物种、温度等条件	[46,54,93,95-96,98-99,102-104]
	茎叶	植株茎叶外观形状	[42-43,45,47-53,55-59,64-65,67-69,71,81-89,94,97]
	花	花朵外观形状及颜色	[62-64,70,72]
	果	果实外观形状及颜色	[44,60-61,74,77,79-80,90-92]
同一品种	芽	芽的朝向	[66]
	不同部位	动物的头、躯干、尾等部位/植物的茎、叶、花、果等部位外观形状及颜色	[75-76,78,105-106]

采集的图像一般规模较小。比如研究产量估算[93]、大田稻穗分割[65]、森林虫害监测[47]、杂草识别[84]使用的图像只有几幅至几十幅图;因为通过无人机遥感或搭载相机航拍的地面范围比较大,像素比较高,这些图像经过预处理后也能得到几百或者上千张用于训练和测试的图像。而知名公开数据库的样本规模在3 000~28 000张之间。

4.3 类间差异

一般,检测、识别和分类等的准确率与各类间的差异程度呈正相关关系。就考虑类间差异来看,DL在农业中的应用可以分为生物和非生物,其中生物的类间差异主要是指生物不同种类、不同个体之间的外观特征差异;非生物的类间差异是地理位置特征差异和对生物的特征影响上的差异。类间差异的存在,是DL识别各类特征的基础。

所综述论文中有关类间差异分析如表3所示,其中花卉的分类存在着种间相似和种内差异的现象[62,72];果体病理图像几何特征差异比较明显[61,91];同类疾病,在致病环境相差不大时,病果图像往往表现出共性,也就会呈现出非常相似的特征[91]。刺儿、灰菜与早熟禾的外形较为相似,莎草与玉米的外形较为相似,这种类间差异比较小的植物会导致DL识别准确率下降[82]。一些植物的特殊视图(如番茄的花、果、茎、叶之间有明显差异)提供了不同的茎、叶、花和果实的分类标准,能够提高DL的分类准确率[75]。

4.4 数据预处理

所综述论文97%(63篇)涉及数据预处理。预处理环节相应预处理方法如表4所示,数据预处理过程包括数据清洗、数据转换和降维处理。其中数据清洗技术主要是用于保证数据特定特征的完整性;数据转换是为了满足深度学习模型的要求,将数据从一种格式或结构转换为另一种格式或结构的过程;降维是去除不相关和冗余的变量,降低分析和生成模型的复杂性,提高建模效率[107]。最常见的预处理方法是调整图像大小,包括图像分割、缩放和归一化(48篇)。根据DL模型的要求,图像像素大小为600×600、256×256、128×128、95×95和48×48是最常见的尺寸。

4.5 数据扩增

DL模型一般是由多层非线性学习器组成,模型较为复杂;要分析的数据是从复杂的自然环境中获得。为了使DL模型具有较好的泛化性能,需要尽可能多地增加训练样本规模,数据补充和数据转换等数据扩增技术被提出。本文所综述应用研究中有37%(24篇)的文献采用了数据扩增技术。

由表5可知,应用最多的数据扩增技术有图像随机旋转、剪裁、平移、水平和垂直翻转等方法,以向模型提供不同环境的数据,从而改善模型学习过程,提高模型泛化性能。特别是对那些只采集了少量数据的研究;比如在黄瓜叶部病害识别中采用随机旋转、水平翻转图像[52],在植物叶片图像识别中采用随机水平、垂直翻转

表4 数据预处理

类型	目的	方法	参考文献
数据清除	去噪	中值滤波/去除毛刺/孔洞等噪声	[42,45,52-53,75,87]
	图像补边/拼接处理	正射校正/影像拼接/填充目标区孔洞	[50,65,67,94,96,100]
	处理异常值	对异常数据进行纠正/剔除	[98]
	数据统计	提取有价值信息	[48]
	特征补充	增加时间特征	[103]
数据转换	图像分割	分割成若干个像素一致的图像	[47,49-50,55,60,65,68,72-73,76,84,92-93,96-97,101,105-106]
	归一化	数据归一化处理,尺寸统一	[43-44,49,51-52,54,58,60-61,63,69,74,80,82-83,85,89,95,98-99,104]
	图像缩放	缩放转换,图像大小统一尺寸	[42,45,47,50,62,66,75,77,79,81,94]
	灰度转换	把原始的彩色图像转换为灰度图像	[53,76,84,87-88,93,104]
	空间转换	将RGB图像转变HSI彩色空间的图像	[45,86]
	格式转换	转换成tfrecord数据文件	[70]
降维	降低维度	高斯滤波	[42,51,71,85,87]

表5 数据扩增技术

功能	方法	参考文献	样本划分
数据补充	分别模拟不同角度和背光场景下对同一种病变果体的成像	[65,76,91]	3:1:1
	从互联网上下载对应类别图片用于扩充图片集	[43]	4:1
	再次采集6幅特征光谱图像和3幅主成分图像	[87]	4:1
	采集其他地区的玉米田间杂草图像	[82]	5:1
	再次采集健康草莓叶片的图像	[55]	不同的比例
数据转换	随机水平或垂直翻转、随机旋转角度、随机缩放原图等操作	[43,50,52,58,63,69,74-75,77,81,83-84,97]	3:1:1或4:1
	仿射变换、透视变换、颜色抖动、对比度增强、叠加噪声等操作引入轻微的扰动而实现数据扩充	[45,60-61,79-80]	4:1

注:3:1:1是训练集、验证集、测试集的比例,4:1或5:1为训练集与测试集的比例。

及缩放图像[69]等,将扩增的图像和实际采集的图像共同构成数据集,然后在真实图像上进行测试。因此,运用数据扩增技术使他们的模型能够更一般化和更好地应对现实中的复杂场景。

4.6 模型与框架

DL在农业中的应用研究一般包括模型优化、框架选择和模型训练与测试。所采用网络结构模型如表6所示。其中52%(34篇)的研究成果直接从头开始训练针对特定研究对象的CNN,例如:多特征融合的CNN[62,87,93,104]等,以提高模型对特定研究对象的检测、分类、识别等准确率。31%(20篇)的研究成果是基于经过大规模数据集预训练的经典网络结构模型,比如AlexNet、VGGNet、ResNet、Faster R-CNN、GoogLenet、LeNet等。还有7篇论文使用改进的DBN。

所选论文所采用框架角度如表7所示。其中大部分研究集中于Caffe(18篇,占比28%)、Tensorflow(12篇,占比18%)和Keras(2篇)。Caffe被广泛使用的一个可能原因是它包含了各种卷积神经网络模型和数据集,用户可以轻松地调用这些数据集。

其中模型训练和测试主要包括样本的划分、训练策略的制定、初始参数设置与调优等。常见的划分方式如表5最后一列所示。一些论文中采用10折交叉验证的策略,即每次选择9个子集作为训练数据,1个子集作为测试数据,这种训练和验证策略能够充分利用数据集中的所有数据[46,51-52]。

初始化参数一般包括学习率、权重、动量等,学习率一般设置在0.001~0.01之间。为了进一步优化模型,Dropout正则化、梯度下降等调优技术常被采用,比如为了避免求解器陷入局部极小值而显著降低模型性能,较通用的做法是开始设定一个较高的学习率,随着训练的进行而自适应地降低。

4.7 性能对比

为了评价DL效果,准确率(Accuracy,ACC)、召回率(Recall,R)、平均正确率(mean Average Precision,mAP)、交除并(Intersection over Union,IoU)、均方误差(Root Mean Square Error,RMSE)、平均绝对误差(Mean Absolute Error,MAE)、$F1$值等评价指标被采用,具体如表8所示。

大部分研究显示基于DL所获得的结果优于与之比较的其他实现机制。DL技术在植物病虫害检测、植物识别和分类等领域中的应用均表现出非常好的性能,一般识别准确率大于95%、识别速度快、鲁棒性强、泛化性

表6　深度学习网络模型选择

模型	网络结构	网络特点	参考文献
自行构建的卷积神经网络	多特征融合的CNN	提取的融合特征维度低于传统的人工设计特征	[62,87,93,104]
	端到端的CNN	直接作用于原始图像数据，通过逐层特征学习，进而利用多层网络获取特征信息	[63]
	7层结构的CNN	共享权值和逐渐下降的学习速率	[89]
	时变冲量学习的CNN	参数训练过程实现网络自我优化，自动提取果园物联网传感器采集的果体图像病变特征	[81,92,99]
	二进制哈希码的CNN	可有效地将高维杂草特征进行压缩，以便于实际田间杂草特征的存储和后续计算	[82]
	MobileNet科优先的CNN	轻量CNN，能降低CNN的权重大小	[64]
	深度卷积神经网络	优选一种8层网络用于番茄主要器官特征提取与表达	[47,49,52-53,55-56,59,61,65,72,75,84,94,96,100,103]
	编码器-解码器为基础的	能够自动从环境信息中学习到主要的非线性组合特征	[54,83]
	基于RGB和HSI关系阈值法优化的CNN	基于区域的分割技术，获取前景目标与背景在像素灰度值特征上的差异，构造一个区分不同区域的分水岭	[86]
	YOLO的CNN	通过单个CNN遍历整个图像，回归目标的类别和位置，实现了直接端到端的目标检测	[80]
	Inception Net的CNN	对得到的不同尺度特征图进行分类和位置回归	[70]
	全卷积神经网络	通过卷积和池化运算输出不同尺度的孔隙特征图，将孔隙的深层特征和浅层特征相融合	[101]
	自学习特征的CNN	对图像块采用线性稀疏自动编码器进行自动学习，获取局部特征的权值矩阵	[73]
已构建的卷积神经网络	AlexNet	将训练好的模型继续进行迁移训练，保留预训练模型所有卷积层的参数，只替换最后一层全连接层	[43,45,50,57-58,69,74]
	VGGNet	优化全连接层层数，用6标签SoftMax分类器替换原有VGG-16网络中的分类器优化模型结构和参数	[42,68,97]
	ResNet	对块图像的特征进行抽象与学习，以自动获取更加深层抽象更具代表性的图像块深层特征	[68,77,79]
	Fast R-CNN	5个卷积层的网络即可具有较高的特征提取和分类性能，增加或降低卷积层数都会使网络性能下降	[46,76,78,105]
	GoogLeNet	利用多尺度卷积核提取不同尺度穗瘟病斑分布式特征并进行级联融合	[60,106]
	LeNet	将方形矩阵卷积核改为适用于一维近红外光谱的向量卷积核，简化网络结构	[71]
深度置信网络	多个限制玻尔兹曼机（RBM）堆叠而成	引入神经胶质改进深度信念网络，并将分解信号结合光照和二氧化碳进行多因子的特征提取	[44,48,51,67,85,98,102]

表7　深度学习框架

框架	主要功能	参考文献
Caffe	应用在视频、图像处理方面	[43,50,53,57-59,62,65,74,77-80,86,93,97,100,106]
Tensorflow	应用于各类机器学习算法的编程实现	[45-46,51,55,63,69-70,99,101,103,105]
Keras	应用于将创意迅速转换为结果的编程实现	[66,99]
微软DL框架CNTK2.0	主要应用于作为语音识别的应用上	[75-76]
Theano	在Python中用于定义、优化、求值数学表达式	[102]
Chainer	允许用简单直观的方式编写出复杂的架构	[96]

能好。从识别准确率和识别速度方面看，例如在植物叶片病害识别中[43]，测试一张图片的时间仅20.79 ms，且其对图像空间位置变化的适应性较好，在扩增图片集上的测试准确率高达99.56%；可能原因是所获取的图形中植物叶子形状、生病叶子具有较明显特征，相对易于识别。在运动中肉牛形体部位识别[105]、龙眼叶绿素含量[87]、作物产量估计[93]、番茄主要器官分类识别[75]、花卉种类识别[72]和林业图像分类[73]等领域的应用中准确率和平均精度相对较低，一般准确率和平均精度均低于85%。这可能是由于使用的数据中包含有动态模糊的图像、叶片采摘后叶绿素有少量变化等造成。从鲁棒性和泛化性能方面看，例如在基于自学习特征的林业图像分类中[73]，底层局部特征是通过自动学习得到的，泛化性更好；在水稻虫害识别中[56]，设计的10层的CNN模

表8　性能指标

指标	定义	说明	参考
准确率	$Acc=(TP+TN)/(TP+TN+FN+FP)$，$TP$ 和 TN 分别为将正类分类为正类和负类，FN 和 FP 分别为将负类划分为负类和正类	识别、分类或预测的正确程度	[42-45,48-49,52-53,56,58-63,65,67-70,72-74,79-82,84-86,89,91-92,96-97,103-104,106]
召回率	$R=TP/(TP+FN)$	将正类预测为正类与所有正类的比率	[91,97]
平均正确率	$mAP=1/C(\sum_{i=1}^{N}Acc(k)\Delta R(k))$，$C$ 为类别数，N 为引用阈值的数量，k 为阈值，$Acc(k)$ 为准确率，$R(k)$ 为召回率	预测目标位置及类别的准确度	[42,44,46-47,50-51,54,57,63,71,75,77,83,90,94,101,105]
交除并	$IoU=AO/AU$，AO 和 AU 分别为目标识别与目标标注的交集与并集	图像中识别相应目标的准确度	[55]
均方根误差	$RMSE=\sqrt{\sum_{i=1}^{n}\frac{e_i^2}{n}}$，$n$ 为预测总次数，e_i 为第 i 个样本预测值和观测值的偏差	预测值和观测值之间残差的样本标准偏差	[88,99,102]
平均绝对误差	$MAE=\frac{\sum_{i=1}^{n}\lvert e_i\rvert}{n}$，$e_i$ 为第 i 个样本预测的绝对误差，n 为样本个数	预测值和观测值之间绝对误差的平均值	[99]
$F1$ 值	$F1=2\times\frac{Acc\times R}{Acc+R}$，$ACC$ 和 R 分别是准确率和召回率	准确率和召回率的调和平均数	[65,77-78,97]

型，可有效地提取图像的特征，对水稻二化螟害虫识别具有很好的抗干扰性和鲁棒性。

5　讨论与展望

通过上述综述可进一步总结DL在农业中的应用具有如下几个方面的优势：首先，它能提高分类/检测/识别等准确率，例如，在植物叶片病害识别[43]中仅经过3次训练迭代，就能达到90%以上的识别准确率；4.7节中所介绍的性能对比也显示较常规算法其能得到更高的准确率。其次，DL具有很好的泛化性和通用性。例如，在水稻虫害识别[56]、果蔬果体病理图像识别[91]中，可有效地提取图像的特征，对目标识别具有很好的抗干扰性和鲁棒性。此外，虽然它较传统方法（如支持向量机、随机森林等）训练时间更长，但它的识别时间非常短。例如，在黄瓜叶部病害识别[52]中，基于CNN训练时间为56 h，但识别只需2.7 s。最后，可以通过运用图像旋转和剪裁进行扩增数据集来训练模型，以节省在复杂环境中收集图像信息的工作量，例如，在冬枣病害识别[61]、番茄主要器官分类识别[75]中，通过旋转、颜色和亮度变化、尺寸缩放等，对数据集进行扩增，DL仍能学习到较好的稳定的分类特征，避免了传统特征提取方法的不足。

分析发现DL在农业中的应用还存在如下几个方面的不足：首先，DL需要大量数据集用于模型的训练、验证和测试，这就需要搭建相机或传感器设备采集不同环境下的数据信息。例如，在棉花病害识别[42]、大蒜鳞芽朝向识别[66]、花卉种类识别[72]中，都需要采集大量的图像。其次，基于DL的大部分农业问题为有监督学习，相应样本数据需要标签标识，一般需要较为专业的人员参与并对目标类别进行人工标记。例如，在玉米田间杂草快速识别[81]、草莓叶部白粉病病害识别[55]中，均需要对所采集图像进行耗时的人工标记。最后，虽然DL可以很好地学习训练数据集中的特征，但是不能在数据集的表达能力之外进行一般化。例如在菊花花型和品种识别[63]中，把菊花的识别作为一个封闭的系统，需要进一步研究该模型是否能迁移到其他花型和品种的识别。

整体上，DL在农业中的应用场景和研究对象仍有待进一步扩展：

当前研究成果主要集中于植物在形态学、病态学、生长环境信息学等方面的检测、分类及预测。而DL在动物的分类、识别和检测中的研究成果相对较少。一个原因可能是动物的动态运动特征使得其应用场景更加复杂，一般需要结合兽医或动科专家参与分析对应动物的生理和行为特征；同时也增大了图像获取、预处理以及快速精准识别的难度，一般需要采用视频分析手段，这给DL的适时性和鲁棒性提出了更高的要求。从媒体报道看，基于DL的猪、牛脸部识别和行为特征分析是当前应用研究的一个热点。随着人们对动物健康状况及肉制品质量安全的重视，DL技术也将为动物生长环境的监控及改善提供便利。

另外，近年来，智慧农业正在我国兴起，其在推动农业生产领域的智能化、经营领域的差异化以及服务领域的全方位信息化过程中产生了大量的图片和数据，如何

融合并综合利用这些数据还面临着较大挑战，DL在这方面的应用仍有待深入研究。

再者，有待进一步将DL研究成果融入农机装备和装置，以真正落地相应理论成果。例如基于DL定位水果的位置并识别水果的成熟度，以支持智能采摘和分类；应用DL技术对土壤含水率、大气温湿度、CO_2含量、土壤酸碱度、肥料营养等作物生长的环境信息进行挖掘分析，通过云端服务实时提供给农场管理者以辅助其精准决策。

就理论方法而言，在如下三个方面仍有待深入研究。一个是专家经验和DL算法有待进一步融合，比如将手工制作的特征与使用各种技术自动提取的特征结合在一起，以提高整体性能。另一个是未来还可能利用时间维度进行更高的特征分类或预测，以适应模型的终身学习；例如可以根据先前连续观察到的植物或动物的生长情况，动态预测它们的产量、评估它们的需水量或避免疾病的发生等。最后，算法的执行速度有待提高，以满足实时性要求，比如视频识别、应用于除草机和水果采摘装置的图像识别算法对适时性都具有非常高的要求。

参考文献：

[1] Kamilaris A，Gao F，Prenafeta-boldú F X，et al.Agri-IoT：a semantic framework for internet of things-enabled smart farming applications[C]//3rd World Forum on Internet of Things（WF-IoT） IEEE，Reston，VA，USA，2016：442-447.

[2] Saxena L，Armstrong L.A survey of image processing techniques for agriculture[C]//Proceedings of Asian Federation for Information Technology in Agriculture，Australian Society of Information and Communication Technologies in Agriculture，Perth，Australia，2014.

[3] Singh A，Ganapathysubramanian B，Singh A K，et al. Machine learning for high-throughput stress phenotyping in plants[J].Trends in Plant Science，2016，21(2)：110-124.

[4] Lecun Y，Bengio Y，Hinton G.Deep learning[J].Nature，2015，521(7553)：436-444.

[5] 袁嘉杰，张灵，陈云华.基于注意力卷积模块的深度神经网络图像识别[J].计算机工程与应用，2019，55(8)：9-16.

[6] 刘栋，李素，曹志冬.深度学习及其在图像物体分类与检测中的应用综述[J].计算机科学，2016，43(12)：13-23.

[7] 阮凯，邱卫根.多信息融合的深度学习人脸表情识别算法研究[J].计算机工程与应用，2019，55(5)：192-196.

[8] 王小玉，韩昌林，胡鑫豪.加权特征融合的密集连接网络人脸识别算法[J].计算机科学与探索，2019，13(7)：1195-1205.

[9] 侯一民，周慧琼，王政一.深度学习在语音识别中的研究进展综述[J].计算机应用研究，2017，34(8)：2241-2246.

[10] Kamilaris A，Prenafeta-Boldú F X.Deep learning in agriculture：a survey[J].Computers and Electronics in Agriculture，2018，147：70-90.

[11] Hinton G E，Osindero S，Teh Y W.A fast learning algorithm for deep belief nets[J].Neural Computation，2006，18(7)：1527-1554.

[12] Krizhevsky A，Sutskever I，Hinton G E.ImageNet classification with deep convolutional neural networks[C]// Annual Conference on Neural Information Processing Systems.Cambridge，USA：MIT Press，2012：1097-1105.

[13] Rossi F，Conan-Guez B.Functional multi-layer perceptron：a non-linear tool for functional data analysis[J]. Neural Networks，2005，18(1)：45-60.

[14] Tu Z，Wei X，Qin Q，et al.Multi-stream CNN：learning representations based on human-related regions for action recognition[J].Pattern Recognition，2018，79：32-43.

[15] Paliy I.Face detection using Haar-like features cascade and convolutional neural network[C]//Proceedings of International Conference on Modern Problems of Radio Engineering，Telecommunications and Computer Science，2008：375-377.

[16] Mohamed A R，Dahl G E，Hinton G E.Acoustic modeling using deep belief networks[J].IEEE Transactions on Audio，Speech，and Language Processing，2012，20(1)：14-22.

[17] Hush D，Abdallah C，Horne B.The recursive neural network and its applications in control theory[J].Computers & Electrical Engineering，1993，19(4)：333-341.

[18] Lecun Y L，Bottou L，Bengio Y，et al.Gradient-based learning applied to document recognition[J].Proceedings of the IEEE，1998，86(11)：2278-2324.

[19] Krizhevsky A，Sutskever I，Hinton G E.ImageNet classification with deep convolutional neural networks[J].Communications of the ACM，2017，60(6)：84-90.

[20] Hadjis S，Abuzaid F，Zhang C，et al.Caffe con Troll：shallow ideas to speed up deep learning[C]//Proc Fourth Workshop on Data Analytics in the Cloud，2015：1-4.

[21] Simonyan K，Zisserman A.Very deep convolutional networks for large-scale image recognition[J].Computer Science，2014.

[22] Szegedy C，Liu W，Jia Y，et al.Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition，Boston，MA，USA，2015：1-9.

[23] He K，Zhang X，Ren S，et al.Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition，2016：770-778.

[24] Sun W，Yao B，Chen B，et al.Noncontact surface roughness estimation using 2D complex wavelet enhanced ResNet for intelligent evaluation of milled metal surface quality[J].Applied Sciences，2018，8(3)：381.

[25] Pang Y，Sun M，Jiang X，et al.Convolution in convolu-

tion for network in network[J].IEEE Transactions on Neural Networks & Learning Systems,2018,29(5):1587-1597.
[26] Chen Y,Yang X,Zhong B,et al.Network in network based weakly supervised learning for visual tracking[J].Journal of Visual Communication and Image Representation,2016,37:3-13.
[27] Xie S,Girshick R,Dollár P,et al.Aggregated residual transformations for deep neural networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition,2017:5987-5995.
[28] Girshick R,Donahue J,Darrell T,et al.Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2014:580-587.
[29] He K,Zhang X,Ren S,et al.Spatial pyramid pooling in deep convolutional networks for visual recognition[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(9):1904-1916.
[30] Liu W,Anguelov D,Erhan D,et al.SSD:single shot multibox detector[C]//European Conference on Computer Vision,2016:21-37.
[31] Girshick R.Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision,2015:1440-1448.
[32] Ren S,He K,Girshick R,et al.Faster R-CNN:towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems,2015:91-99.
[33] Redmon J,Divvala S,Girshick R,et al.You only look once:unified,real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:779-788.
[34] Deng J,Dong W,Socher R,et al.ImageNet:a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition,2009:248-255.
[35] Bahrampour S,Ramakrishnan N,Schott L,et al.Comparative study of deep learning software frameworks[J].Computer Science,2015.
[36] Jia Y,Shelhamer E,Donahue J,et al.Caffe:convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd International Conference on Multimedia,Orlando,FL,USA,2014:675-678.
[37] Abadi M.TensorFlow:learning functions at scale[C]//Proceedings of ICFP,2016.
[38] Seide F.Keynote:the computer science behind the Microsoft cognitive toolkit:an open source large-scale deep learning tool kit for Windows and Linux[C]//IEEE/ACM International Symposium on Code Generation and Optimization,2017.
[39] Sankaran A,Aralikatte R,Mani S,et al.DARVIZ:deep abstract representation,visualization,and verification of deep learning models[C]//2017 IEEE/ACM 39th International Conference on Software Engineering:New Ideas and Emerging Technologies Results Track(ICSE-NIER),2017:47-50.
[40] Boufenar C,Batouche M.Investigation on deep learning for off-line handwritten Arabic character recognition using Theano research platform[C]//2017 Intelligent Systems and Computer Vision(ISCV),2017:1-6.
[41] 付文博,孙涛,梁藉,等.深度学习原理及应用综述[J].计算机科学,2018,45(S1):11-15.
[42] 张建华,孔繁涛,吴建寨,等.基于改进VGG卷积神经网络的棉花病害识别模型[J].中国农业大学学报,2018,23(11):161-171.
[43] 孙俊,谭文军,毛罕平,等.基于改进卷积神经网络的多种植物叶片病害识别[J].农业工程学报,2017,33(19):209-215.
[44] 张善文,张传雷,丁军.基于改进深度置信网络的大棚冬枣病虫害预测模型[J].农业工程学报,2017,33(19):202-208.
[45] 龙满生,欧阳春娟,刘欢,等.基于卷积神经网络与迁移学习的油茶病害图像识别[J].农业工程学报,2018,34(18):194-201.
[46] 魏杨,毕秀丽,肖斌.基于区域卷积神经网络的农业害虫检测方法[J].计算机科学,2018,45(S2):226-229.
[47] 孙钰,周焱,袁明帅,等.基于深度学习的森林虫害无人机实时监测方法[J].农业工程学报,2018,34(21):74-81.
[48] 王秀美,牟少敏,邹宗峰,等.基于深度学习的小麦蚜虫预测预警[J].江苏农业科学,2018,46(5):183-187.
[49] 秦丰,刘东霞,孙炳达,等.基于深度学习和支持向量机的4种苜蓿叶部病害图像识别[J].中国农业大学学报,2017,22(7):123-133.
[50] 杨国国,鲍一丹,刘子毅.基于图像显著性分析与卷积神经网络的茶园害虫定位与识别[J].农业工程学报,2017,33(6):156-162.
[51] 王献锋,张传雷,张善文,等.基于自适应判别深度置信网络的棉花病虫害预测[J].农业工程学报,2018,34(14):157-164.
[52] 张善文,谢泽奇,张晴晴.卷积神经网络在黄瓜叶部病害识别中的应用[J].江苏农业学报,2018,34(1):56-61.
[53] 蒋丰千,李旸,余大为,等.基于Caffe的生姜病害识别系统研究与设计[J].中国农机化学报,2019,40(1):126-131.
[54] 张善文,黄文准,张传雷.基于环境信息和深度自编码网络的农作物病害预测模型[J].江苏农业学报,2018,34(2):288-292.
[55] 杨晋丹,杨涛,苗腾,等.基于卷积神经网络的草莓叶部白粉病病害识别[J].江苏农业学报,2018,34(3):527-532.
[56] 梁万杰,曹宏鑫.基于卷积神经网络的水稻虫害识别[J].江苏农业科学,2017,45(20):241-243.
[57] 刘婷婷,王婷,胡林.基于卷积神经网络的水稻纹枯病图

像识别[J].中国水稻科学,2019,33(1):90-94.
[58] 孙云云,江朝晖,董伟,等.基于卷积神经网络和小样本的茶树病害图像识别[J].江苏农业学报,2019,35(1):48-55.
[59] 尹晔,尚媛园,邵珠宏,等.基于迁移学习的甜菜褐斑病识别方法[J].计算机工程与设计,2018,39(9):2748-2752.
[60] 黄双萍,孙超,齐龙,等.基于深度卷积神经网络的水稻穗瘟病检测方法[J].农业工程学报,2017,33(20):169-176.
[61] 张善文,黄文准,尤著宏.基于物联网和深度卷积神经网络的冬枣病害识别方法[J].浙江农业学报,2017,29(11):1868-1874.
[62] 林思思,叶东毅,陈昭炯.多特征融合的花卉图像DL分类算法[J].小型微型计算机系统,2018,39(7):1446-1450.
[63] 袁培森,黎薇,任守纲,等.基于卷积神经网络的菊花花型和品种识别[J].农业工程学报,2018,34(5):152-158.
[64] 曹香滢,孙卫民,朱悠翔,等.基于科优先策略的植物图像识别[J].计算机应用,2018,38(11):3241-3245.
[65] 段凌凤,熊雄,刘谦,等.基于深度全卷积神经网络的大田稻穗分割[J].农业工程学报,2018,34(12):202-209.
[66] 方春,孙福振,任崇广.基于深度学习的大蒜鳞芽朝向识别研究[J].计算机应用研究,2019,36(3):1-2.
[67] 陆永帅,李元祥,彭希帅.深度置信网络模型的机载多光谱数据罂粟识别[J].遥感信息,2017,32(4):98-103.
[68] 尼加提·卡斯木,师庆东,刘素红,等.基于卷积网络的沙漠腹地绿洲植物群落自动分类方法[J].农业机械学报,2019,50(1):217-225.
[69] 郑一力,张露.基于迁移学习的卷积神经网络植物叶片图像识别方法[J].农业机械学报,2018,49(S1):354-359.
[70] 吴佳,许立兵,孙立新,等.基于多尺寸特征图卷积方法的玉米雄穗检测[J].科学技术与工程,2018,18(27):48-52.
[71] 鲁梦瑶,杨凯,宋鹏飞,等.基于卷积神经网络的烟叶近红外光谱分类建模方法研究[J].光谱学与光谱分析,2018,38(12):3724-3728.
[72] 沈萍,赵备.基于深度学习模型的花卉种类识别[J].科技通报,2017,33(3):115-119.
[73] 李英杰,张广群,汪杭军.基于自学习特征的林业业务图像分类方法[J].林业科学,2018,54(5):78-86.
[74] 孙俊,何小飞,谭文军,等.空洞卷积结合全局池化的卷积神经网络识别作物幼苗与杂草[J].农业工程学报,2018,34(11):159-165.
[75] 周云成,许童羽,郑伟,等.基于深度卷积神经网络的番茄主要器官分类识别方法[J].农业工程学报,2017,33(15):219-226.
[76] 周云成,许童羽,邓寒冰,等.基于双卷积链Fast R-CNN的番茄关键器官识别方法[J].沈阳农业大学学报,2018,49(1):65-74.
[77] 彭红星,黄博,邵园园,等.自然环境下多类水果采摘目标识别的通用改进SSD模型[J].农业工程学报,2018,34(16):155-162.
[78] 熊俊涛,刘振,汤林越,等.自然环境下绿色柑橘视觉检测技术研究[J].农业机械学报,2018,49(4):45-52.
[79] 王丹丹,何东健.基于R-FCN深度卷积神经网络的机器人疏果前苹果目标的识别[J].农业工程学报,2019,35(3):156-163.
[80] 赵德安,吴任迪,刘晓洋,等.基于YOLO深度卷积神经网络的复杂背景下机器人采摘苹果定位[J].农业工程学报,2019,35(3):164-173.
[81] 王璨,武新慧,李志伟.基于卷积神经网络提取多尺度分层特征识别玉米杂草[J].农业工程学报,2018,34(5):144-151.
[82] 姜红花,王鹏飞,张昭,等.基于卷积网络和哈希码的玉米田间杂草快速识别方法[J].农业机械学报,2018,49(11):1-13.
[83] 刘庆飞,张宏立,王艳玲.基于深度可分离卷积的实时农业图像逐像素分类研究[J].中国农业科学,2018,51(19):3673-3682.
[84] 王术波,韩宇,陈建,等.基于深度学习的无人机遥感生态灌区杂草分类[J].排灌机械工程学报,2018,36(11):1137-1141.
[85] 邬美银,陈黎.基于深度学习的监控视频树叶遮挡检测[J].武汉科技大学学报,2016,39(1):69-74.
[86] 张加楠,张雪芬,简萌,等.先验阈值优化卷积神经网络的作物覆盖度提取算法[J].信号处理,2017,33(9):1230-1238.
[87] 岳学军,凌康杰,洪添胜,等.基于高光谱图像的龙眼叶片叶绿素含量分布模型[J].农业机械学报,2018,49(8):18-25.
[88] 甘海明,岳学军,洪添胜,等.基于深度学习的龙眼叶片叶绿素含量预测的高光谱反演模型[J].华南农业大学学报,2018,39(3):102-110.
[89] 高震宇,王安,刘勇,等.基于卷积神经网络的鲜茶叶智能分选系统研究[J].农业机械学报,2017,48(7):53-58.
[90] 李帷韬,曹仲达,朱程辉,等.基于深度集成学习的青梅品级智能反馈认知方法[J].农业工程学报,2017,33(23):276-283.
[91] 谭文学,赵春江,吴华瑞,等.基于弹性动量深度学习神经网络的果体病理图像识别[J].农业机械学报,2015,46(1):20-25.
[92] 王细萍,黄婷,谭文学,等.基于卷积网络的苹果病变图像识别方法[J].计算机工程,2015,41(12):293-298.
[93] 许等平,任怡,闫哲,等.基于CNN的无人机遥感影像质量评价[J].林业工程学报,2018,3(5):121-127.
[94] 鲁恒,付萧,贺一楠,等.基于迁移学习的无人机影像耕地信息提取方法[J].农业机械学报,2015,46(12):274-279.
[95] 王洪,刘伟铭.深度信任支持向量回归的耕地面积预测方法[J].郑州大学学报(理学版),2016,48(1):121-126.
[96] 孙钰,韩京冶,陈志泊,等.基于深度学习的大棚及地膜农田无人机航拍监测方法[J].农业机械学报,2018,49(2):133-140.
[97] 郑二功,田迎芳,陈涛.基于深度学习的无人机影像玉米倒伏区域提取[J].河南农业科学,2018,47(8):155-160.

(下转第51页)

[2] MacQueen J.Some methods for classification and analysis of multivariate observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.Berkeley，CA：University of California Press，1967：281-297.
[3] Rodriguez A，Laio A.Clustering by fast search and find of density peaks[J].Science，2014，344(6191)：1492-1496.
[4] Ertöz L，Steinbach M，Kumar V.A new shared nearest neighbor clustering algorithm and its applications[C]// Workshop on Clustering High Dimensional Data & Its Applications at SIAM International Conference on Data Mining，2002：105-115.
[5] Du M，Ding S，Jia H.Study on density peaks clustering based on k-nearest neighbors and principal component analysis[J].Knowledge-Based Systems，2016，99：135-145.
[6] 鲍舒婷，孙丽萍，郑孝遥，等.基于共享近邻相似度的密度峰聚类算法[J].计算机应用，2018，38(6)：1601-1607.
[7] Guo Z，Huang T，Cai Z，et al.A new local density for density peak clustering[C]//Advances in Knowledge Discovery and Data Mining，2018：426-438.
[8] 朱庆峰，葛洪伟.K 近邻相似度优化的密度峰聚类[J].计算机工程与应用，2019，55(2)：148-153.
[9] 邱保志，辛杭.一种基于共享近邻亲和度的聚类算法[J].计算机工程与应用，2018，54(18)：184-187.
[10] Jarvis R A，Patrick E A.Clustering using a similarity measure based on shared near neighbors[J].IEEE Transactions on Computers，1973，22(11)：1025-1034.
[11] Chang H，Yeung D Y.Robust path-based spectral clustering[J].Pattern Recognition，2008，41(1)：191-203.
[12] Cover T，Hart P.Nearest neighbor pattern classification[J].IEEE Transactions on Information Theory，1967，13(1)：21-27.
[13] Korn F，Muthukrishnan S.Influence sets based on reverse nearest neighbor queries[J].ACM SIGMOD Record，2000，29(2)：201-212.
[14] Carpaneto G，Toth P.Algorithm 548：solution of the assignment problem[J].ACM Transactions on Mathematical Software，1980，6(1)：104-111.
[15] Strehl A，Ghosh J.Cluster ensembles—a knowledge reuse framework for combining multiple partitions[J].Journal of Machine Learning Research，2002，3：583-617.
[16] UCI machine learning repository[EB/OL].[2018-10-25].http://archive.ics.uci.edu/ml/datasets.html.
[17] Singh D，Febbo P G，Ross K，et al.Gene expression correlates of clinical prostate cancer behavior[J].Cancer Cell，2002，1(2)：203-209.
[18] Nene S A，Nayar S K，Murase H，et al.Columbia object image library(coil-20)：CUCS-005-96[R].1996.
[19] Deng L.The mnist database of handwritten digit images for machine learning research[J].IEEE Signal Processing Magazine，2012，29(6)：141-142.
[20] Hull J J.A database for handwritten text recognition research[J].IEEE Transactions on Pattern Analysis and Machine Intelligence，1994，16(5)：550-554.
[21] Samaria F S，Harter A C.Parameterisation of a stochastic model for human face identification[C]//Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision，Sarasota，1994：138-142.

（上接第33页）

[98] 王华，罗平，赵志刚，等.基于深度置信网络的农用地基准地价评估模型[J].农业工程学报，2018，34(21)：263-271.
[99] 谭建灿，毛克彪，左志远，等.基于卷积神经网络和AMSR2微波遥感的土壤水分反演研究[J].高技术通讯，2018，28(5)：399-408.
[100] 王璨，武新慧，李恋卿，等.卷积神经网络用于近红外光谱预测土壤含水率[J].光谱学与光谱分析，2018，38(1)：36-41.
[101] 韩巧玲，赵玥，赵燕东，等.基于全卷积网络的土壤断层扫描图像中孔隙分割[J].农业工程学报，2019，35(2)：128-133.
[102] 周翔宇，程勇，王军.基于改进深度信念网络的农业温室温度预测方法[J].计算机应用，2019(4)：1053-1058.
[103] 孙耀杰，蔡昱，张馨，等.基于WDNN的温室多特征数据融合方法研究[J].农业机械学报，2019，50(2)：273-280.
[104] 薛金林，闫嘉，范博文.多类农田障碍物卷积神经网络分类识别方法[J].农业机械学报，2018，49(S1)：35-41.
[105] 邓寒冰，许童羽，周云成，等.基于DRGB的运动中肉牛形体部位识别[J].农业工程学报，2018，34(5)：166-175.
[106] 杨秋妹，肖德琴，张根兴.基于机器视觉的猪只饮水行为识别[J].农业机械学报，2018(6)：1-8.
[107] Lv S，Kim H，Zheng B，et al.A review of data mining with big data towards its applications in the electronics industry[J].Applied Sciences，2018，8(4)：582-616.

中国仿真学会会刊
A journal of China Simulation Federation

系统仿真学报

Journal of System Simulation

7
2018 Vol. 30

中文核心期刊要目总览
中国科学引文数据库(CSCD)核心刊
中国科技期刊统计源核心刊
RCCSE中国核心学术期刊
全球文献检索数据库SCOPUS
英国科学文摘SA/INSPEC
美国Ulrich's Periodicals Directory

短文

动态与信息

期刊基本参数：CN11-3092/V*1988*m*A4*393*zh*P*50.00*47*2018-07

基于数据挖掘的印制电路样板投料优化

吕盛坪[1]，乐强生[1]，刘涛[2]
(1. 华南农业大学 工程学院，广州 510642；2. 安徽建筑大学 机械与电气工程学院，合肥 230601)

摘要：为了更准确确定 PCB (Printed Circuit Board)样板投料，基于车间历史数据开展挖掘分析。*梳理报废率关联参数，利用假设检验优选报废率预测建模参数。*构建多元线性回归、卡方自动相互作用检测器、人工神经网络和支持向量机预测模型；*定义余数入库率和补投率及两者加权和评价指标，*开展投料仿真，对比优选多元线性回归预测机制。*引入调节系数，结合企业成本模型进行优选；开发实施投料控件并进行验证；*结果证明余数入库率和补投率较实际值均有明显降低，可减少样板生产物料投入、库存浪费、补投拖期等成本，为 PCB 样板投料优化提供新的参考手段。

关键词：PCB；报废率；数据挖掘；预测；投料优化
中图分类号：TN41; TP391.7 文献标识码：A 文章编号：1004-731X (2018) 07-2656-10
DOI: 10.16182/j.issn1004731x.joss.201807028

Optimization of Material Release for Printed Circuit Board Template Based on Data Mining

Lü Shengping[1], *Yue Qiangsheng*[1], *Liu Tao*[2]
(1. School of Engineering, South China Agricultural University, Guangzhou 510642, China;
2. School of Mechanical and Electrical Engineering, Anhui Jianzhu University, Hefei 230601, China)

Abstract: Data mining were employed for the optimization of material release of PCB (Printed Circuit Board) template. *PCB scrap ratio related parameters were specified and prediction model variables were chosen according to hypothesis test*. Multiple linear regression (MLR), Chi-squared automatic interaction detector, artificial neural network and support vector machine approaches for the prediction of scrap ratio were employed. *Evaluation indictors called as superfluous ratio, supplement release ratio and weighted sum of the two were presented*; the material release simulation was conducted and then the four approaches were compared and MLR was taken as the preferred one. *Adjust coefficient was introduced and optimized according to factory's cost model.* Finally, material release tool were developed and verified. Comparison results shown that superfluous and supplement release ratio has significant reduction which indicates that the approach can systematically reduce the cost of material release, waste of inventory, tardiness and so on.
Keywords: PCB; scrap ratio; data mining; prediction; optimization of material release

引言

PCB (Printed Circuit Board)是电气和电子设备的基体[1]。客户个性化的需求使得 PCB 样板订单大量增加，相应的生产模式也从传统的大规模批量生产转化为以大规模个性化生产为主要特征的长尾生产[2]。更精准地确定各订单投料是影响车间上下游综合成本的关键。合理确定每个订单的投料面积将消减车间物料、生产、库存、回收处理和拖期等成本。但各订单具有个性化的功能特征，所需要

收稿日期：2017-04-20 修回日期：2017-09-11；
基金项目：国家自然科学基金(51605169)，广东省自然基金(2014A030310345)；
作者简介：吕盛坪(1982-)，男，湖南新邵，博士，副教授，研究方向为生产计划、工艺规划与调度优化、工业数据挖掘等。

经历的工艺流程也不同, 确切的投料面积在生产前难以提前确定。车间投料人员通常基于不同层板的历史报废率均值计算确定各订单的投料面积, 调研中发现人工投料易导致车间超投和补投均较高。

剩余个性化 PCB 样板只能置于库存或直接销毁, 这导致物料、生产、存储和销毁成本的浪费。通过频繁的补投可以降低面积剩余, 但补投将带来车间在制品种增加和调度的不稳定, 影响车间生产周期和订单准交率。

显然, 同时减少超投和补投是消减车间成本的全局优化策略。但车间目前难以平衡这对 Pareto 目标, 主要以控制超投为主。寻求一种更优化的投料策略, 同时降低这两个目标具有重要的科学和工程实际意义。在此提出采用数据挖掘的手段对 PCB 样板进行报废率预测和投料优化。

数据挖掘是采用一系列的技术以抽取数据中蕴藏的有价值信息, 现已成功应用于工业、农业、商业等诸多领域[3]。近年来, 随着大数据和云制造的兴起, 相关学者针对大数据的制造范式[2]、大数据资源的云服务组合[4]、基于大数据的智慧工厂框架构建[4]等宏观角度开展了较深入研究。但与 PCB 的精益制造结合的数据挖掘研究相对较少。本研究从企业实际需求出发, 基于数据挖掘手段对其 PCB 样板投料进行优化, 主要内容包括 3 个方面 :1) 梳理 PCB 报废率相关参数, 利用假设检验和 MLR (Multiple linear regression)参数检验机制优选 PCB 报废率预测建模参数 ; 2) 利用 MLR、CHAID (Chi-squared automatic interaction detector)、ANN (Artificial neural network) 和 SVM (Support vector machine)等预测机制开展报废率预测 ; 结合 PCB 产品特点, 定义相应评价指标, 对比优选预测机制 ; 3) 引入调节系数, 结合企业成本模型进行优选 ; 开发实施投料控件并进行实施验证。

1 样本和方法

1.1 整体流程和抽样

本文整体流程如图 1 所示。其中数据来源于广州某公司的 ERP (Enterprise Resource Planning) 系统。抽取 51 192 条有效记录, 其中下单时间在 2014-01-01~2015-12-31 之间的 40 526 条记录用作训练样本集(S1) ; 第一组测试样本(S2)为下单时间在 2016-01-1~2016-03-31 之间的 3 726 条记录, 第二组测试样本(S3)为订单下单时间在 2016-04-01~2016-07-31 的 6 940 条记录, 各组样本之间互不重叠。

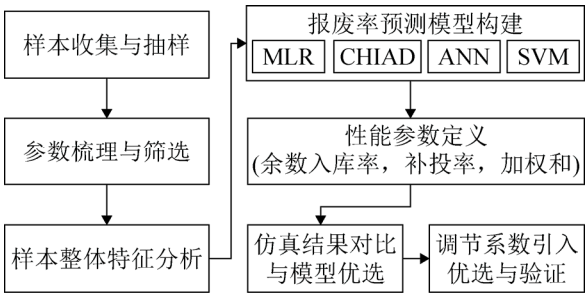


图 1 整体流程
Fig.1 Main flowchart

1.2 参数梳理与筛选

综合 ERP 中现有参数和数据预分析, 利用继承、派生、转换等方式, 共梳理与报废率具有较强关系的参数 53 个。为减少分析复杂度并提高建模效率, 以寻找对报废率具有积极贡献的重要参数[6], 在此对离散和连续型并分别采用 *F*-检验和 *t*-检验进行筛选, 筛除连续型变异系数过小或离散型某一取值比例过大的参数。然后基于 MLR 的参数 *t*-检验筛除对报废率影响不显著参数, 最终保留 26 个参数(具体在 3.1 节讨论)。相应参数集合如表 1 所示。结合现有导出数据, 确定各数据范围, 并在“描述(取值范围)”列给出。

1.3 预测方法

统计分析和机器学习是常用测机制[3]。MLR 作为常用的统计分析方法通过构建一个因变量与多个自变量的线性函数进行预测, 其中模型未知参数基于样本估计获得[3]。

表 1 参数规范

Tab. 1 Parameters specification

序号	参数名称	符号	描述(取值范围)	*p*-值
1	板厚	*Pt*	PCB 板厚(0.3~6.1)	<0.001[t]
2	层数	*Ln*	PCB 铜箔层数(2~10)	<0.001[f]
3	板镀次数	*Plfr*	板镀工艺频次(0~4)	<0.001[f]
4	工序数	*Noo*	完成 PCB 板所需经历工序数(4~55)	<0.001[f]
5	半固化片数	*NPP*	层压时所采用的半固化片总数(0~20)	<0.001[f]
6	是否光电板	*Photb*	光电板为 1，反之为 0.	<0.001[f]
7	是否高频板	*Highfb*	高频板为 1，反之为 0.	<0.001[f]
8	是否 IPCIII 标准	*IPCIII*	采用 IPCIII 验收标准时为 1，反之为 0	<0.001[f]
9	内层最小线宽	*Mwil*	各层芯板中线宽最小值(0.9~137.8)	<0.001[t]
10	内层最小线间距	*Msil*	各层芯板中线间距最小值(1.0~99)	<0.001[t]
11	外层最小线宽	*Mwol*	上下层铜箔上线宽最小值(3-140)	<0.001[t]
12	外层最小线间距	*Msol*	上下层铜箔上线间距最小值(3~140)	<0.001[t]
13	是否有减薄铜	*Crd*	采用减薄铜工序时为 1，反之为 0	<0.001[f]
14	残铜率均值	*Arcr*	内层铜箔各层残铜率均值(0.15%~95%)	<0.001[f]
15	是否内层干膜	*Dfil*	采用了内层干膜工艺为 1，反之为 0	<0.001[f]
16	是否有负片电镀	*Nflp*	有负片电镀为 1，反之为 0	<0.001[f]
17	总孔数	*Noh*	总钻孔孔数(0~116 406)	<0.001[f]
18	沉铜通孔厚径比	*Hddr*	电镀前孔径尺寸比(0.816~12)	<0.001[f]
19	是否需要二钻	*Secd*	需要二钻为1，反之为0	<0.001[f]
20	是否有树脂塞孔	*Phwr*	有树脂塞孔为 1，反之为 0	<0.001[f]
21	是否有背钻	*Bcdr*	有背钻为 1，反之为 0	/
22	是否埋盲孔板	*Bbuv*	是埋盲孔板为 1，反之为 0	/
23	有铅喷锡	*Haslwl*	采用有铅喷锡表面处理为 1，反之为 0	<0.001[f]
24	无铅喷锡	*Haslol*	采用无铅喷锡表面处理为 1，反之为 0	<0.001[f]
25	护铜剂	*Osp*	采用有机保焊膜表面处理为 1，反之为 0	<0.001[f]
26	图镀铜镍金	*Cnapp*	采用图镀铜镍金表面处理为 1，反之为 0	<0.001[f]
27	镀金手指	*Gfp*	采用镀金手指表面处理为 1，反之为 0	<0.001[f]
28	电镀硬金	*Godp*	采用电镀硬金表面处理为 1，反之为 0	<0.001[f]
29	电镀软金	*Snap*	采用电镀软金表面处理为 1，反之为 0.	/
30	化学镍钯金	*Imnpa*	采用化学镍钯金表面处理为 1，反之为 0	/
31	沉金/银/锡	*Iasa*	采用沉金、沉银、沉锡中的一种为 1，反之为 0	0.001[f]
32	是否有碳油	*Cboil*	有碳油为 1，反之为 0	0.285
33	是否采用字符打印	*Chaprt*	采用字符打印为 1，反之为 0	<0.001
34	是否采用阻焊塞孔	*Srph*	采用阻焊塞孔为 1，反之为 0	<0.001
35	是否非绿油墨颜色		采用非绿油墨颜色为 1，反之为 0	<0.001[f]
36	Panel 包含交货单元数量	*Duap*	多个交货单元布局于一个 Panel 进行生产, 该厂只生产单拼板(1~1 008)	<0.001[t]
37	补投次数	*Suprt*	同一订单号投料次数-1 (1~9)	
38	要求生产数量	*Reqq*	要求交货数量-引用库存数量(1~42 000)	<0.001[t]
39	综合投入数量	*Relq*	生产中各订单所有投入交货单元数量(6~45 360)	
40	报废数量	*Scraq*	各订单报废数量(0~1 040)	
41	入库数量	*Qualq*	各订单合格入库数量(5~44 688)	
42	余数入库数量	*Surfq*	入库数量-要求生产数量(0~4 480)	
43	交货单元面积	*Punita*	各订单交货单元面积(0.001~0.267)	<0.001[t]
44	要求生产面积	*Reqa*	*Reqq*×*Punita* (0.054~41.553)	<0.001[t]
45	综合投入面积	*Rela*	*Relq*×*Punita* (0.175~76.532)	

续表

序号	参数名称	符号	描述(取值范围)	*p*-值
46	报废面积	*Scraa*	*Scraq*×*Punita* (0.000~20.976)	
47	入库面积	*Quala*	*Qualq*×*Punita* (0.14~75.842)	
48	余数入库面积	*Surfa*	*Surfq*×*Punita* (0~47.092)	
49	补投率	*Suprr*	一定周期内所有补投次数/总订单数×100%	
50	报废率	*Scrar*	*Scraa/Rela*×100% (0%~39.394%)	
51	综合合格率	*Qualr*	*Quala/Rela*×100% (60.606%~91.66%)	
52	余数入库率	*Surfr*	*Surfa/Reqa*×100% (0%~91.67%)	
53	历史良率	*Hquar*	2 年内同一生产型号综合合格率的均值(2.903%~100%)	<0.001[t]

注：板厚、线宽/间距、铜厚、面积的单位分别为 mm、mil、OZ 和 m^2。[f]和 [t]符号分别表示 *F*-检验和 *t*-检验小于 0.05，“/”表示连续型参数变异系数过小或离散型参数中某一取值比例过大；*p*-值为空表示该参数只做统计用。

机器学习主要目标是设计模型使得计算机能基于数据确定相应行为。决策树、ANN 和 SVM 是常见的机器学习算法[7]。决策树采用递归划分训练样本以评估指定变量对因变量的影响, 从而将样本分隔为具有相似特征的不同分组并预测结果[3]。本文所采用的 CHAID 决策树通过反复地将样本划分为两个或两个以上的子节点, 用于分类或连续型因变量的预测，任何预测变量(连续预测转化为序数预测)将被合并，直到对目标变量没有显著差异，划分过程中的合并与否将基于 Bonferroni-检测计算调整 *p*-值确定[8]。

ANN 基于大量的神经单元模拟大脑解决问题的方式进行训练并挖掘相应规律[9]。本研究将采用具有一层隐含层的 BP(Back Propagation)人工神经网络进行预测。BP 人工神经网络隐含层和输出层中每一个节点都具有一个输入的加权求和函数 ;同时定义了一个激活函数用以确定一个节点在一个或多个输入节点条件下的输出。所有节点集可视为一个超平面, 而隐含层神经元实质上是将非线性的样本转换为线性样本[9]，算法参数设置为默认值。

SVM 是一种用于分类和回归的监督式机器学习模型。类似于 MLR，SVM 回归预测的目的是通过确定一个因变量与多个自变量之间的超平面 ;目标是减少实际值和预测值之间的偏差平方和；但 SVM 的超平面参数估计遵循的是 ε-不敏感损失函数参数估计最小化原则[10]。实践证明 SVM 具有较好的预测精度和预测效果[11]。本研究通过 SPSS Modeler 14.1 实现。

1.4 性能度量

常用的均方误差在此无法完全反应预测效果。为了更好地评价预测性能，结合 PCB 生产特点，引入余数入库率和补投率相应预测值以及两者加权和作为预测性能评价指标。相应评价指标计算流程如图 2 所示，具体参数如表 2 所示。

投料数量预测定义为：

$$Relq\_\mathrm{P}d=\frac{\mathrm{Re}qq\_\mathrm{P}d\times(1+adj\_cof)}{(100-Scrar\_\mathrm{P}d)/100} \tag{1}$$

式中：*adj_cof* 为平衡余数入库率和补投率并优化综合成本引入的调节系数，将设置不同水平，并通过企业成本模型确定优选值。

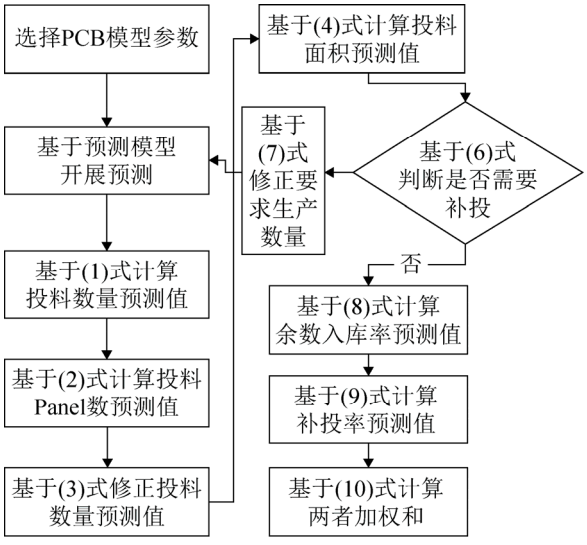


图 2 性能指标计算流程

Fig.2 Flowchart for the calculation of indicators

表 2 评价指标相关参数
Tab. 2 Evaluation indicators related parameters

参数	符号
报废率预测值	*Scrar_Pd*
投料数量预测值	*Relq_Pd*
投料 Panel 数预测值	*Relp_Pd*
投料面积预测值	*Rela_Pd*
要求生产数量预测值	*Reqq_Pd*
要求生产面积预测值	*Reqa_Pd*
报废面积预测值	*Scraa_Pd*
余数入库率预测值	*Surfr_Pd*
订单补投次数预测值	*Suprt_Pd*
补投率预测值	*Suprr_Pd*

相应投料 Panel 数预测值通过下式计算：

$$Relp\_\mathrm{P}d=\left\lceil\frac{Relq\_\mathrm{P}d}{Duap}\right\rceil \tag{2}$$

式中：*Duap* 为表 1 中所定义的一个 Panel 中包含交货单元数量。

在此基础上，投料数量将更新为：

$$\mathrm{Re}lq\_\mathrm{P}d=\mathrm{Re}lp\_\mathrm{P}d\times Duap \tag{3}$$

投料面积预测值将直接基于修正后的投料数量计算：

$$\mathrm{Re}la\_\mathrm{P}d=\mathrm{Re}lq\_\mathrm{P}d\times Punita \tag{4}$$

相应报废面积预测值通过下式获得：

$$Scraa\_\mathrm{P}d=\mathrm{Re}la\_\mathrm{P}d\times Scrar \tag{5}$$

定义某一次预测投料中订单补投次数：

$$Suprt\_\mathrm{P}d=\begin{cases}1, Rela\_\mathrm{P}d-\mathrm{Re}qa\_\mathrm{P}d-Scraa\_\mathrm{P}d<0\\ 0,\text{反之}\end{cases} \tag{6}$$

更新要求生产数量预测值：

$$\mathrm{Re}qq\_\mathrm{P}d=\left\lceil\mathrm{Re}lq\_\mathrm{P}d-Scraa\_\mathrm{P}d/Duap\right\rceil \tag{7}$$

相应的余数入库率预测值定义为：

$$Surfr\_\mathrm{P}d=\frac{\sum Rela\_\mathrm{P}d-\mathrm{Re}qa-\sum Scraa\_\mathrm{P}d}{\mathrm{Re}qa}\times 100\% \tag{8}$$

$\sum Rela\_\mathrm{P}d$ 为所有投料面积预测值之和，$\sum Scraa\_\mathrm{P}d$ 为所有报废面积预测值之和。

n 个订单的补投率预测值为：

$$Suprr\_\mathrm{P}d=(\sum_{i=1}^{n}\sum_{j}Surfr\_\mathrm{P}d_i)/n\times 100\% \tag{9}$$

$\sum_{j}Surfr\_\mathrm{P}d_i$ 为订单 i 的补投次数。

在此基础上，余数入库率和补投率的预测值的加权和定义为：

$$WSum=w_1\times Surfr\_\mathrm{P}d+w_2\times Suprr\_\mathrm{P}d,\quad w_1+w_2=1, 0\le w_1\le 1, 0\le w_2\le 1 \tag{10}$$

加权和是对余数入库率和补投率这对 Pareto 目标的综合；当对比不同预测模型时，如果无法保证某一模型在这两个目标上均优于其它模型，难以评定算法的优劣。而在确定的权值下，加权和越大相应预测机制所导致的超投和补投成本越大；反之则越小，对应预测模型越好。

2 结果与讨论

2.1 *t*-检验和参数优选

基于 *F*-检验和 *t*-检验优选的 35 个参数，进一步利用 MLR *t*-检验筛除对报废率影响不显著参数，作为各预测模型构建参数输入。表 3 中给出 *t*-检验结果，基于显著性判断筛选：层数、板镀次数、工序数、半固化片数、是否光电板、是否高频板、是否 IPCIII 标准、是否有减薄铜、残铜率均值、是否内层干膜、是否有负片电镀、孔数、沉铜通孔厚径比、是否有树脂塞孔、有铅喷锡、无铅喷锡、护铜剂、图镀铜镍金、镀金手指、电镀硬金、沉金/银/锡、是否采用阻焊塞孔、是否非绿油墨颜色、交货单元面积、要求生产面积、历史良率 26 个报废率预测建模参数。

2.2 预测性能对比

基于上述预测机制和优选参数，开展虚拟仿真投料，测试报废率预测模型的训练和预测效果。针对训练样本，表 4 给出了不同预测机制均方误差；可以看出 ANN 的均方误差最小，随后依次为 MLR、CHAID 和 SVM。

进一步对比 3 组样本的余数入库率(*Surfr*)和补投率(*Suprr*)预测值，具体结果如表 5 所示。其中 S2、S3 的 *Suprr* 较 S1 分别增加了 13.17%、30.84%，

而相应的 *Surfr* 分别降低了 28.38%、52.07%。其主要原因来自于车间管理改革，2016 年 1 月和 4 月开始(分别对应于样本组 S2、S3)车间管理层强制车间投料人员一次投料不能超过要求生产数量的 9%和 6%。调整后的结果显示目前车间的投料策略和管理技术难以同时优化余数入库率和补投率。

表 3　MLR 参数检验
Tab. 3　MLR parameter testing

序号	变量	*B*	*SE*	*Beta*	*t*	Sig.
1	板厚	0.075	0.079	0.017	0.949	0.343
2	层数	0.632	0.024	0.421	26.582	<0.001
3	板镀次数	1.012	0.12	0.108	8.405	<0.001
4	工序数	0.156	0.01	0.564	15.184	<0.001
5	半固化片数	–0.042	0.017	–0.028	–2.455	0.014
6	是否光电板	0.367	0.142	0.013	2.583	<0.001
7	是否高频板	0.550	0.13	0.013	4.226	<0.001
8	是否 IPCIII 标准	0.976	0.087	0.046	11.220	<0.001
9	内层最小线宽	–0.005	0.004	–0.010	–1.188	0.235
10	内层最小线间距	–0.012	0.01	–0.016	–1.238	0.216
11	外层最小线宽	–0.005	0.005	–0.010	–1.155	0.248
12	外层最小线间距	0.008	0.01	0.011	0.815	0.415
13	是否有减薄铜	0.587	0.089	0.021	6.564	<0.001
14	残铜率均值	0.021	0.001	0.211	14.891	<0.001
15	是否内层干膜、	0.378	0.083	0.013	4.565	<0.001
16	是否有负片电镀	1.127	0.141	0.115	8.020	<0.001
17	总孔数	7.49E-005	0.0	0.153	26.154	<0.001
18	沉铜通孔厚径比	0.180	0.02	0.124	9.155	<0.001
19	是否需要二钻	–0.058	0.062	–0.003	–0.936	0.35
20	是否有树脂塞孔	–1.104	0.209	–0.025	–5.275	<0.001
21	有铅喷锡	2.631	0.268	0.172	9.809	<0.001
22	无铅喷锡	2.697	0.271	0.136	9.943	<0.001
23	护铜剂	2.652	0.282	0.076	9.397	<0.001
24	图镀铜镍金	3.915	0.274	0.159	14.268	<0.001
25	镀金手指	0.417	0.122	0.010	3.411	<0.001
26	电镀硬金	0.748	0.155	0.027	4.813	<0.001
27	沉金/银/锡	2.664	0.266	0.246	10.005	<0.05
28	是否采用字符打印	0.125	0.135	0.003	0.927	0..354
29	是否采用阻焊塞孔	–0.389	0.062	–0.045	–6.265	<0.001
30	是否非绿油墨颜色	0.139	0.065	0.007	2.134	0.033
31	Panel 中包含交货单元数量	–0.004	0.003	–0.011	–1.729	0.084
32	要求生产数量	4.08E-005	0.0	0.003	0.542	0.588
33	交货单元面积	37.202	0.903	0.233	41.213	<0.001
34	要求生产面积	–0.121	0.006	–0.119	–19.540	<0.001
35	历史良率	–0.099	0.003	–1.227	–33.801	<0.001

注：*B* 为各参数系数估计值；*SE* 为标准偏差；*Beta* 为标准化系数；*t* 为 *t*-检验值；*Sig*.为显著性值的临界水平。

表 4 均方误差对比
Tab. 4 Comparison of mean squared error (*MSE*)

	MLR	CHAID	ANN	SVM
SE	3.08	3.18	3.05	3.22

表 5 不同预测算法余数入库率和补投率对比
Tab. 5 Comparison of *Surfr* and *Suprr* for the four prediction approaches

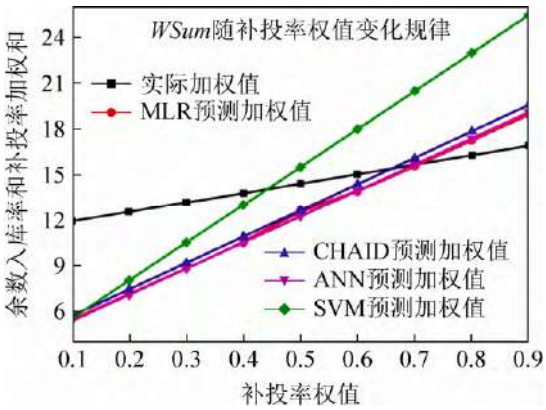
样本	实际值	MLR	CHAID	ANN	SVM
S1	17.54%	20.64%	21.33%	20.83%	27.97%
	11.31%	3.77%	3.98%	3.66%	3.05%
S2	19.85%	21.84%	21.38%	21.65%	26.36%
	8.1%	3.89%	4.01%	3.81%	3.24%
S3	22.95%	21.38%	22.68%	21.63%	26.78%
	5.42%	3.87%	3.91%	3.85%	3.14%

注：各组样本上下行分别为补投和余数入库率。

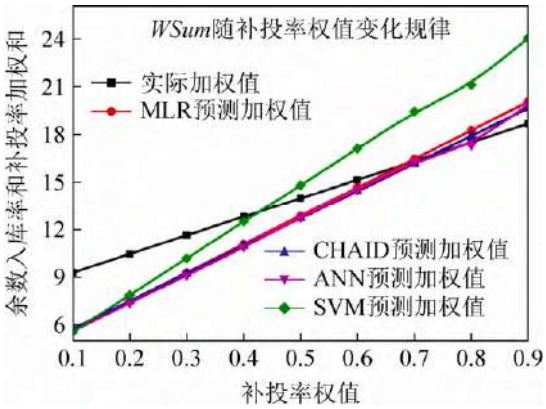
对比结果同时显示，S2、S3 预测余数入库率明显低于实际余数入库率，SVM 获得最低的余数入库率；对于 S1，其中 MLR、CHAID、ANN 以及 SVM 所得余数入库率预测值较实际值分别下降了 66.67%，64.81%，67.63%，73.03%；对于 S2，分别下降了 51.98%，50.49%，52.96%，60%；对于 S3，分别下降了 28.59%，27.85%，28.97%，42.07%。4 种预测机制均可减少余数入库率，从而降低物料、生产、库存和销毁处理成本。

补投率对比结果显示, 对于 S1, MLR、CHAID、ANN、SVM 所得预测补投率较实际值分别上升了 17.67%，21.61%，18.76%和 59.46%；对于 S2，4 种预测模型所得预测补投率较实际值分别上升了 10.02%，7.71%和 9.07%，32.76%；对于 S3，MLR，CHAID 以及 ANN 所得预测补投率较实际值分别下降了 6.84%，1.17%和 5.75%，而 SVM 所得值增加了 16.69%。

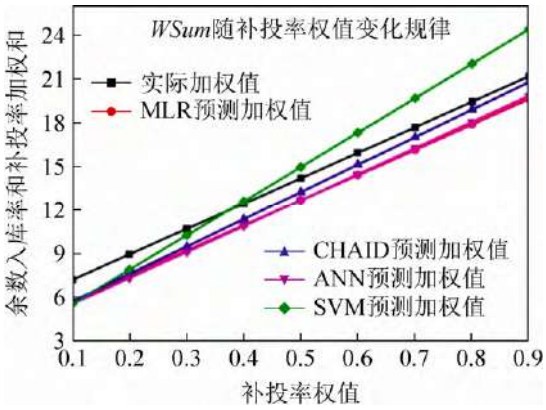
上述对比仍无法直接确定预测机制优劣。进一步基于(10)式加权和进行对比。因为 w_1 和 w_2 无法明确, 所以设置 *Suprr_Pd* 权值 0.1~0.9 九个水平(间隔为 0.1，且满足 $w_1+w_2=1$)，对于样本 S1、S2 和 S3，分别计算 *WSum*，并拟合 *WSum* 随 *Suprr_Pd* 权值变化趋势，具体结果如图 3 所示。



(a) S1 的 *WSum* 随 *Suprr_Pd* 变化趋势



(b) S2 的 *WSum* 随 *Suprr_Pd* 变化趋势



(c) S3 的 *WSum* 随 *Suprr_Pd* 变化趋势

图 3 加权和随补投率预测值权值变化趋势
Fig.3 Wsum changes along with the weight of *Suprr_Pd*

从图 3(a)~(c)拟合曲线可以看出，不同权值条件下，SVM 所得 *WSum* 最大，CHAID 次之，MLR 和 ANN 最小且拟合曲线几乎重合；对于 3 组样本，在补投率权值小于 0.4 时, 4 种预测机制所得 *WSum*

均小于 *WSum* 实际值；对于 S1、S2，在 *Suprr_Pd* 权重小于 0.7 时，4 种预测机制所得 *WSum* 均小于 *WSum* 实际值；对于 S3，除 SVM 外，其它预测机制所得 *WSum* 均小于 *WSum* 实际值。按照企业经验 *WSum* 中 w_2 一般小于 0.7, 整体上可以选择 MLR 和 ANN 作为预测模型。因 ANN 对外是一个黑箱且参数配置较为困难，而 MLR 能通过权值系数直观反映各参数对报废率的影响，在此优选 MLR。

2.3 调节系数优选

基于报废率多元回归预测模型, 设置调节系数 0%，0.5%，1%，1.5%，2.0%，2.5%，3.0%七个水平，开展投料仿真，计算仿真投料后相应余数入库率和补投率。同时结合车间实际成本模型计算对应的余数入库成本、拖期成本和补投成本，基于上述 3 种成本优选综合损失最小所对应调节系数。表 6 给出了计算相应成本的参数描述 ;表 7 是实施车间针对第三组样本(S3)计算成本所涉及的核心参数及其结果, 对比综合损失发现当预设余数入库率为 1%时，综合成本最低。

2.4 实施与验证

为了进一步验证上述方法，结合企业应用需求，基于 MLR 所优选参数和预测机制，设置其调节系数 1%，开发实施投料优化控件，该控件与车间 ERP 和制造执行系统具有数据集成双向接口，根据新订单相应参数，可手工选择订单或后台自动触发预测引擎并向 ERP 和制造执行系统反馈预测报废率、下料 Panel 数、综合投入数量等关键结果。

基于该控件开展实施，截取实施以来 2 372 个订单作为验证样本(S4)，实施后的余数入库率和补投率分别为 4.89%和 16.61%，较 S2、S3 的余数入库率分别下降 39.62%、9.78%，补投率分别降低 16.32%、38.17%。验证结果显示本研究所确定的预测机制能更精准的自动确定各订单的投料面积和数量，从而进一步消减车间物料、生产、库存、回收处理、补投和拖期等成本。同时可减少车间投料经验影响和投料人员。

表 6 成本计算相应参数说明
Tab. 6 Specification of parameters related to cost calculation

No.	参数	符号	说明
1	需要补投次数	*TSuprt_Pred*	该样本预测补投累计次数
2	需要补投面积	*RSupra_Pred*	按照预测各订单需要补投面积和
3	余数入库面积	*Surfa_Pred*	同表 1 种定义
4	引用库存面积比例	*Ivuger_Pred*	引用库存中的余数入库面积和预测新产生的余数入库面积之比
5	单位面积平均售价	*Apua_Pred*	平均售价 2000/ m^2 左右
6	**余数入库成本**	***Surfpc_Pred***	***Surfa_Pred*×(1- Ivuger_Pred)×Apua_Pred***
7	拖期>3 天的订单数	*Anom3d_Pred*	约有 1/4 的补投订单其拖期大于 3 天(一般 3 天即以上需要赔偿)
8	拖期订单平均面积	*Aato_Pred*	平均值约为 *5.5* m^2
9	单个订单拖期订单折款	*Disctpo_Pred*	拖期订单平均折款损失 2 200 元左右
10	**拖期成本**	***Tdc_Pred***	***Anom3d_Pred*×Disctpo_Pred***
12	补投面积	*Supra_Pred*	基于需要补投面积，考虑 Panel 圆整后实际需要重新投料面积
13	**补投成本**	***Prodinca_Pred***	***Prodinc_Pred*×Apua_Pred***
14	**综合损失**	***Compl_Pred***	***Surfpc_Pred+ Tdc_Pred+ Prodinca_Pred***

注：其中统计结果基于车间 2016 年 1-8 月累计历史数据。

表 7 成本统计计算表

Tab. 7 Statistic table for the computation of cost

Adjst_coef	*Surfr*	*Suprr*	*TSuprt*	*RSupra*	*Surfa*	*Ivuger*	*Apua*	***Surfpc***
0%	3.89%	21.84%	1 516	565.389 4	1 194.289 9	48.43%	2 000	**1 231 790.6**
0.5%	4.18%	19.97%	1 386	535.215 7	1 283.324 4	48.43%	2 000	**1 323 620.8**
1%	4.56%	17.74%	1 231	503.405 6	1 399.990 2	48.43%	2 000	**1 443 949.9**
1.5%	4.93%	16.24%	1 127	481.217 2	1 513.585 9	48.43%	2 000	**1 561 112.5**
2%	5.32%	14.89%	1 033	458.340 1	1 633.321 9	48.43%	2 000	**1 684 608.2**
2.5%	5.63%	14.19%	985	441.202 5	1 728.496 7	48.43%	2 000	**1 782 771.5**
3%	5.92%	13.39%	929	425.996 3	1 817.531 2	48.43%	2 000	**1 874 601.6**

Adjst_coef	*Anom3d*	*Aato*	*Prodinc*	***Tdc***	*Supra*	***Prodinca***	***Compl***
0%	379	5.5	2 200	**833 800**	565.389 4	**1 130 778.8**	**3 196 369.4**
0.5%	346.5	5.5	2 200	**762 300**	535.215 7	**1 070 431.4**	**3 156 352.2**
1%	307.75	5.5	2 200	**677 050**	503.405 6	**1 006 811.3**	**3 127 811.2**
1.5%	281.75	5.5	2 200	**619 850**	481.217 3	**962 434.56**	**3 143 397.1**
2%	258.25	5.5	2 200	**568 150**	458.340 1	**916 680.25**	**3 169 438.5**
2.5%	246.25	5.5	2 200	**541 750**	441.202 6	**882 405.11**	**3 206 926.6**
3%	232.25	5.5	2 200	**510 950**	425.996 3	**851 992.65**	**3 237 544.3**

注：表头_Pred 均被省略；要求生产面积(*Reqa*)为 30 701.54 m^2。

3 结论

本文提出了基于数据挖掘优化 PCB 样板投料的新机制。首先梳理和规范定义了与 PCB 报废样本直接相关的核心参数, 并基于假设检验优选了与报废率预测建模相关的 26 个参数 :主要包括层数、板镀次数、工序数、半固化片数、是否光电板、是否高频板、是否 IPCIII 标准、是否有减薄铜、残铜率均值、是否内层干膜、是否有负片电镀、孔数、沉铜通孔厚径比、是否有树脂塞孔、有铅喷锡、无铅喷锡、护铜剂、图镀铜镍金、镀金手指、电镀硬金、沉金/银/锡、是否采用阻焊塞孔、是否非绿油墨颜色、交货单元面积、要求生产面积、历史良率等，为类似挖掘分析和参数优选提供了重要参考。

构建了 MLR、CHAID、ANN 和 SVM 四种预测模型 ;引入了余数入库率、补投率及两者加权和评价指标，开展虚拟投料仿真，对比优选 MLR 预测机制。引入调节系数，结合企业实际成本模型进行优选，设置 0%~3.5% 7 个不同调节水平，基于 MLR 预测模型和优选调节系数，通过仿真投料发现 1%时能最小化车间投料综合损失 ;在此基础上，基于 MLR 和 1%调节系数开发投料优化控件，以实施车间实际数据进行验证, 结果显示余数入库率较两组测试样本分别下降 39.62%，9.78%，补投率降低 16.32%，38.17%，具有明显的经济效益。

本研究主要基于一个车间数据, 鉴于该车间生产特点，其数据具有一定的局限性，比如极少的埋盲孔板(包括高密度板)、10 层以上板和多拼板 ;同时 PCB 产品和工艺参数非常之多，其它未列举参数缺乏一定量的高质量的数据积累, 所筛选的参数需要结合不同车间甚至同行业其它企业数据进行适当增删，以提高预测模型参数的通用性。

参考文献:

[1] Marques A C, Cabrera C J, Malfatti C F. Printed circuit boards: A review on the perspective of sustainability [J]. Journal of Environmental Management (S0301-4797), 2013, 131: 298-306.

[2] 姚锡凡, 周佳军, 张存吉, 等. 主动制造—大数据驱动的新兴制造范式[J]. 计算机集成制造系统, 2017, 23(1): 172-185.
Yao X F, Zhou J J, Zhang C J, et al. Proactive manufacturing-a big-data driving emerging manufacturing paradigm [J]. Computer Integrated Manufacturing Systems, 2017, 23(1): 172-185.

[3] Han J W, Kamber M, Pei J. Data mining: concepts and techniques [M]. 3rd Edition. San Francisco: Morgan

Kaufmann, 2011.

[4] 张影, 翟丽丽, 王京. 大数据背景下的云联盟数据资源服务组合模型[J]. 计算机集成制造系统, 2016, 22(12): 2920-2929.
Zhang Y, Zhai L L, Wang J. Cloud computing federation data resource sever composition in big-data background [J]. Computer Integrated Manufacturing Systems, 2016, 22(12): 2920-2929.

[5] 吕佑龙, 张洁. 基于大数据的智慧工厂技术框架[J]. 计算机集成制造系统, 2016, 22(11): 2691-2697.
Lyu Y L, Zhang J. Big-data-based technical framework of smart factor [J]. Computer Integrated Manufacturing Systems, 2016, 22(11): 2691-2697.

[6] Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A. A review of feature selection methods on synthetic data [J]. Knowledge and Information Systems (S0219-1377), 2013, 34(3): 483-519.

[7] Philip C C L, Zhang C. Data-intensive applications, challenges, techniques and technologies: a survey on big data [J]. Information Science (S0020-0255), 2014, 275: 314-347.

[8] Ture M, Tokatli F, Kurt I. Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients [J]. Expert Systems with Applications (S0957-4174), 2009, 36 (2, Part 1): 2017-2026.

[9] Wikipedia. Artificial neural network [EB/OL]. (2016-10) [2017-09-05]. https://en.wikipedia.org/wiki/Artificial_neural_network.

[10] Smola A J, Schölkopf B. A tutorial on support vector regression [J]. Statistics and Computing (S0960-3174), 2004, 14(3): 199-222.

[11] 任艳, 周小敏, 关威, 等. 支持向量回归机在颜色测温中的仿真应用[J]. 系统仿真学报, 2016, 28(11): 2736-2741.
Ren Y, Zhou X M, Guan W, et al. Simulation approach to temperature measuring using image color based on support vector regression [J]. Journal of System Simulation, 2016, 28(11): 2736-2741.

(上接第 2655 页)

[9] 张继华, 邓研, 郭凤仪, 等. 永磁操作机构真空断路器的智能控制器的设计[J]. 辽宁工程技术大学学报(自然科学版), 2011, 30(12): 751-756.
Zhang Jihua, Deng Yan, Guo Fengyi, et al. Design of controller for permanent magnetic actuator vacuum circuit breaker[J]. Journal of Liaoning Technical University (Natural Science), 2011, 30(12): 751-756.

[10] 吕锦柏, 王毅, 谢将剑, 等. 基于线圈电流的永磁真空断路器控制方法[J]. 高电压技术, 2013, 39(4): 860-868.
Lv Jinbo, Wang Yi, Xie Jiangjian, et al. Control Method for Permanent Magnetic Vacuum Circuit Breaker Based on Coil Current[J]. High Voltage Engineering, 2013, 39(4): 860-868.

[11] 刘博, 张建峡. 电容器恒流充电方法的分析与研究[J]. 工业控制计算机, 2013, 26(1): 116-118.
Liu Bo, Zhang Jianxia. Analysis and Research of Constant Current Charging Method for Capacitor[J]. Industrial Control Computer, 2013, 26(1): 116-118.

[12] 阳峰. 基于双管 Buck-Boost 变换器的电容器充电电源研究[D]. 武汉: 华中科技大学, 2012.
Yang Feng. A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Engineering[D]. Wuhan: Huazhong University of Science and Technology, 2012.

[13] 赵红强, 徐建源, 秦祖荫. 单线圈永磁操动机构中驱动器的研制[J]. 华通技术, 2005 (1): 11-15.
Zhao Hongqiang, Xu Jianyuan, Qin Zuyin. The Permanent Magnetic Machine Driver of Vacuum Circuit Breaker[J]. Huatong Technology, 2005 (1): 11-15.

[14] 危立辉. 储能电容开环 PWM 控制恒流充电装置[P]. 中国, 201010555730.X.

[15] 汪先兵, 费树岷, 徐清杨, 等. BP 神经网络 PID 控制的永磁真空开关储能电容恒流充电特性分析[J]. 电工技术学报, 2015, 30(10): 212-218.
Wang Xianbing, Fei Shumin, Xu Qingyang, et al. Constant Current Charging Characteristic Analysis of Storage Capacitor Based on BP Neural Network PID Control for Permanent Magnet Vacuum Switch[J]. Transactions of China Electrotechnical Society, 2015, 30(10): 212-218.

華南農業大學学报

JOURNAL OF SOUTH CHINA AGRICULTURAL UNIVERSITY

ISSN 1001-411X
CODEN HNDXBH
中国·广州
GUANGZHOU CHINA

全国中文核心期刊

2015
5月 第36卷 第3期
May. Vol.36 No.3

华南农业大学学报

第 36 卷　第 3 期　2015 年 5 月

目　　次

期刊基本参数：CN44-1110/S * 1959 * b * A4 * 124 * zh + en * P * ￥15.00 * 1600 * 21 * 2015-05　　**封面设计**：林　敏

网络出版时间：2015-04-14 09:43
网络出版地址：http://www.cnki.net/kcms/detail/44.1110.s.20150414.0943.017.html
华南农业大学学报 2015,36(3):114-119 http://xuebao.scau.edu.cn
Journal of South China Agricultural University doi:10.7671/j.issn.1001-411X.2015.03.020

吕盛坪，吕恩利，陆华忠，等.荔枝不同预冷方式降温特性研究[J].华南农业大学学报,2015,36(3):114-119.

荔枝不同预冷方式降温特性研究

吕盛坪，吕恩利，陆华忠，杨松夏，方思贞
（南方农业机械与装备关键技术教育部重点实验室/华南农业大学 工程学院，广东 广州 510642）

摘要：【目的】研究荔枝不同预冷方式的降温特性.【方法】建立差压预冷试验箱，以“淮枝”为材料，采用冰水（L1）、冷库（L2）、差压（L3）以及高湿差压（L4）进行预冷试验.【结果和结论】L1、L2、L3、L4 分别需耗时 35、55、64 和 345 min 将平均果温降至目标温度（5 ℃）. L1 不同位置降温无显著差异. L2 分别用 195、258 和 228 min 将左右侧和上层果温降至 5 ℃，345 min 后中下层和中间位置果温仍分别高达 5.37、6.16 和 7.37 ℃；左右与中间处降温差异显著. L3 分别用 39、52、42 min 将左右侧和上层果温降至 5 ℃，55 min 后中下层和中间位置果温仍分别高达 6.03、5.67 和 9.03 ℃，上层与中下层果温差异显著；L4 分别用 39、41 min 将左侧和上层果温降至 5 ℃，64 min 后中下层和中右位置果温仍分别高达 5.86、8.83、7.87 和 6.63 ℃，左侧和中间处降温差异显著. L1 预冷效率高、果温均匀性好，是荔枝较适合的预冷方式.

关键词：荔枝；预冷方式；降温特温
中图分类号：S379.1 **文献标志码**：A **文章编号**：1001-411X(2015)03-0114-06

Cooling characteristics of different precooling methods for litchi

LÜ Shengping, LÜ Enli, LU Huazhong, YANG Songxia, FANG Sizhen
(Key Laboratory of Key Technology on Agricultural Machine and Equipment/College of Engineering, South China Agricultural University, Guangzhou 510642, China)

Abstract: 【Objective】To study the cooling characteristics of different precooling methods for litchi. 【Method】A pressure-difference precooling test chamber was established. Four precooling methods, including ice water (L1), cold storage (L2), pressure-difference (L3) and forced-air pressure-difference with high humidity (L4), were adopted for “Huaizhi” litchi. 【Result and conclusion】The results showed that L1, L2, L3 and L4 spent 35, 55, 64 and 345 min respectively to precool the litchi down to the target temperature (5 ℃). The cooling procedure of litchi fruit at different positions for L1 performed no significant difference. L2 took 195, 258 and 228 min to precool litchi at left and right positions and top layer respectively. However, the fruit temperature at middle and bottom layer and middle position were still up to 5.37, 6.16 and 7.37 ℃ respectively after 345 min precooling; and the cooling procedure of litchi fruit showed significant differences between the left, right position and its middle position. L3 took 39, 52, 42 min to precool litchi at the left and right positions and top layer respectively. However, the fruit temperature at middle and bottom layers and middle position were still up to 6.03, 5.67 and 9.03 ℃ respectively after 55 min precooling. The cooling procedure of litchi fruit showed significant difference among the top layer, left position and its middle and bottom layers. L4 spent 39 and 41 min to precool litchi at the left position and top layer respectively. However, the fruit temperature at the middle

收稿日期：2014-03-18 **优先出版时间**：
优先出版网址：
作者简介：吕盛坪（1982—），男，讲师，博士，E-mail：lvshengping@scau.edu.cn
基金项目：国家自然科学基金（31101363）；国家科技支撑计划项目子课题（2013BAD19B01-1-3）；广东省自然科学基金（S2012010010388）；广东省科技计划项目（2012B020313007）；广东省高等学校学科与专业建设专项资金项目（2013LYM_0001）

and bottom layers, middle and right positions were still respectively up to 5.86, 8.83, 7.87 and 6.63 ℃ after 64 min precooling. The cooling procedure of litchi fruit showed a significant difference between the left position and its middle position. L1 has high cooling efficiency and good fruit temperature uniformity, which is suitable for cooling litchi.

Key words: litchi; precooling methods; cooling characteristics

荔枝是我国南方亚热带名优水果，采后急需进行预冷，一般要求在采收后 6 h 内完成包装、预冷、入冷库贮藏[1]. 预冷对降低荔枝采后呼吸强度和生理代谢频率，抑制酶和乙烯释放，减少生理病害，降低腐烂和贮运能耗具有重要意义.

荔枝常用的预冷方式有冰(冷)水预冷、冷库预冷、差压(加湿差压)预冷等[2-3]. 王倩等[4]设计开发了基于机械制冷冷风机组为冷源和以冰为冷源的荔枝产地复合预冷装置. 段洁利等[5]研究了荔枝差压预冷温变特性. 杨洲等[6]对荔枝差压预冷环境气流场进行了研究. 宋晓燕等[7]研究了上海青叶子表面温度在真空预冷过程中的温度变化规律. 宋小勇等[8]对非洲菊真空预冷过程中舌状花瓣、管状花瓣和茎秆 3 个部位的降温速度和均匀性进行了研究. 对果蔬差压预冷过程数学模型和降温特性也有较多研究[9-11]，但针对荔枝不同预冷方式降温特性的研究较少. 本文研究冰水、冷库、低湿差压(简称差压)和高湿差压预冷荔枝果肉的降温规律和温度均匀性，为荔枝预冷方式的选择提供参考.

1　材料与方法

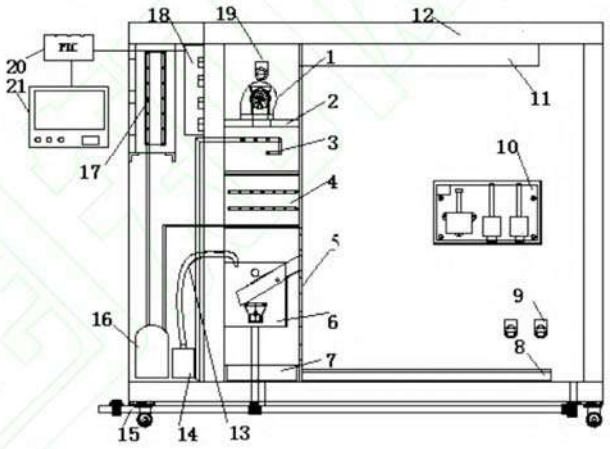
1.1　材料及预处理

试验用荔枝品种为“淮枝”，于 2013 年 7 月 23 日清晨采自广州市从化果园，果实成熟，着色充分. 采后立即运回实验室，剪去果枝、去除伤病果. 为保证不同预冷方式荔枝后续储藏品质，调制 φ 为 0.11% 的施保克进行消毒处理. 因冰水浸泡荔枝会清洗消毒液，并重新带入病毒，所以冰水预冷完成后才进行消毒处理.

1.2　主要仪器设备

冷库预冷采用低温冷库；差压和高湿差压预冷需差压装置来实现冷风强迫对流. 冰水预冷采用尺寸为 530 mm × 320 mm × 400 mm 的储水水箱进行.

冷库预冷采用华南农业大学南方农业机械与装备关键技术教育部重点实验室自主开发的试验厢作为平台，结构如图 1 所示. 该试验平台尺寸为 2 380 mm × 1 280 mm × 1 400 mm，贮藏区尺寸为 1 180 mm × 940 mm × 1 340 mm. 试验平台采用 2 匹制冷机组(四菱制冷设备有限公司)进行制冷，利用冷风机(KINGBO ZNF295-G 24V 直流风机)实现气流循环，超声波雾化振子(JAS-20-B 型，中山市红星电子厂)进行加湿. 可编程控制器(SIMENS S7-300 型 PLC)根据设置的初始参数值和传感器采集的厢内温湿度，对制冷机组、加湿装置、风机等进行控制，智能调控贮藏室内保鲜环境.

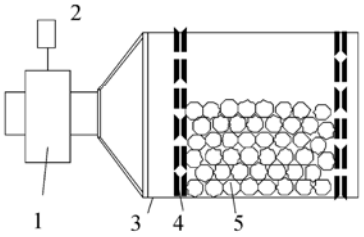


1：风机；2：风机安装板；3：蒸发器；4：汽化盘管；5：开孔隔板；6：加湿器；7：积水槽；8：气流导轨；9：排气阀；10：传感器盒；11：回风道；12：压差式厢体；13：进水管；14：补水箱；15：排水管；16：制冷压缩机；17：冷凝器；18：继电器盒；19：进气阀；20：可编程控制器；21：记录仪.

图 1　试验平台结构示意图

Fig. 1　A schematic diagram of the experimental platform

差压和高湿差压预冷采用自主建立的如图 2 所示差压箱实现差压送风. 预冷差压箱采用 8 mm 厚的有机玻璃板制成. 试验区尺寸为 422 mm × 294 mm × 354 mm. 根据前期针对番茄[12]和龙眼[13]研究所确定的开孔率，结合初步试验，选择开孔率为 13.9%. 根据开孔率在两侧开孔板上均匀设置 25 个直径为 15 mm 的圆孔. 利用 DPT10-35B 型圆型管道风机(佛山南海南洋电机电器有限公司)吸力在箱体内外产生压差，迫使冷空气从箱内快速通过. 试验时，差压箱置于图 1 所示平台中，利用压差抽取试验平台中冷风预冷荔枝. 出口风速通过调速器(湘潭充畅电子电器厂生产的 3000W 可控无极调节王)实现，风速由 AZ8901 风速仪(台湾衡欣科技股份有限公司)测定，误差 ±2%.



1:风机; 2: 调速器; 3 :差压试验箱体; 4: 开孔隔板; 5:荔枝.

图2 预冷差压试验箱

Fig. 2 The forced-air precooling experimental box

试验时,冷库和差压预冷采用同一冷库平台(编为1号试验台),通过控制器开启制冷机组,关闭加湿装置;高湿差压预冷在2号试验台中进行,同时开启加湿和制冷功能. 冰水预冷果温和水温采用 Anymetre PT3002 型探针式温度计测量,测量误差 ±(1~5) ℃,测量范围 50~300 ℃. 其他预冷方式果温采用 WRNT-02 型 K 型热电偶测定,测量误差 ±(1~5) ℃,测量范围 0~500 ℃.

1.3 处理和测定方法

取 60 kg 荔枝均匀分装在 12 个塑料筐中,随机分成 4 组,每组 3 筐. 从第 1 组中随机选取 9 颗荔枝测定并记录预冷前初始温度,然后将 3 筐荔枝垂直堆垛放置于预冷水箱中,快速加入冰水覆盖筐中荔枝. 每隔 5 min,从上中下每个筐的左中右分别随机选择 1 颗荔枝,快速测定荔枝果温. 同时测定冰水温度;如果冰温超过 5 ℃,在预冷水箱中快速加入冰块. 当荔枝平均温度降低至近 5 ℃的目标温度结束.

选第 2 组 3 筐荔枝作为冷库预冷材料,将该组 3 筐荔枝垂直堆垛快速置于 1 号试验台中,保证筐的长度方向平行于 1 号试验台长度方向,筐的最左侧靠近图 1 所示开孔隔板右侧 30 mm. 然后从 3 个筐中沿长度方向左中右位置各选 1 颗荔枝,分别插入 1 个 K 型热电偶.

第 3、4 组各 3 筐荔枝分别用于差压和高湿差压预冷. 将差压预冷 3 筐荔枝分批倒入图 2 所示差压箱,每倒入 1 筐作为 1 层(共包括上中下 3 层),并从每层的左中右位置分别选择 1 颗荔枝,各插入 1 个 K 型热电偶,完成后将差压箱置于 1 号试验台(差压箱长度方向与冷库预冷组塑料筐长度平行,最右侧离试验台开孔隔板 30 mm). 并将冷库和差压预冷热电偶数据线一并连接到数字记录仪上,利用电脑保存数据. 调整变频开关使差压箱出口风速为 4 $m \cdot s^{-1}$,相应差压箱横截面上风速约 1 $m \cdot s^{-1}$(等于出口处所测风速乘以截面比,截面比为风机出口与差压箱横截面面积比值,约为 0.27). 开启 1 号试验台电源和差压箱风机电源,设置制冷温度为 0 ℃后进行差压和冷库预冷试验. 差压预冷热电偶均温降到 5 ℃时,打开平台 1 取出差压箱关闭库门继续进行冷库预冷. 高湿差压预冷在 2 号试验台进行,试验时同时开启制冷和加湿(制冷温度 0 ℃,湿度 85%~95%),待热电偶所测温度平均值降低到 5 ℃关闭 2 号试验台,其他操作与差压预冷过程类似.

1.4 数据处理

试验数据处理软件为 Excel 和 SPSS(16.0).

2 结果与分析

不同预冷方式荔枝温变从上中下 3 层和左中右 3 个不同位置处果肉降温过程进行分析;温度均匀性通过不同层和位置处温度的差异性和温度标准差反应. 其中每一层果温为同一层左中右 3 颗荔枝同一次测定所得温度的均值,左中右不同位置的温度为同一位置上中下不同层 3 颗荔枝同一次测定所得结果的均值.

2.1 冰水预冷荔枝降温过程和温度均匀性分析

图 3、4 分别给出了冰水预冷过程不同层(包含整体均温降温)和左中右不同位置处果肉平均温变过程. 可以看出,平均果温从 27.3 ℃降至 5.06 ℃只需 35 min,降温迅速;且左中右不同位置处果温降温曲线非常接近. 同时可以看出,预冷过程中,上层荔枝果温较中层低,下层果温最高. 可能是预冷过程冰浮于水上,上层荔枝与冰接触多,降温快;越到下层,荔枝接触冰的机率越小,降温越慢. 但同一时期,上中下、左中右不同位置处荔枝果温均无显著性差异. 分析不同层处左中右不同位置温度标准差发现:越到下面荔枝果温越均匀. 可能原因是中下层水温较一致,被冷水包围的荔枝温度一致性好;而上层荔枝与浮于水面冰块或碎冰接触不均匀,预冷过程温差较大.

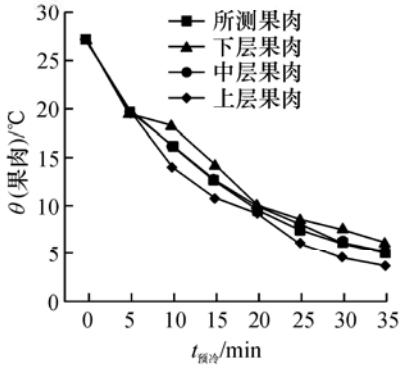


图3 冰水预冷不同层处果肉温度变化过程

Fig. 3 The fruit flesh temperature changes of different layers for ice precooling

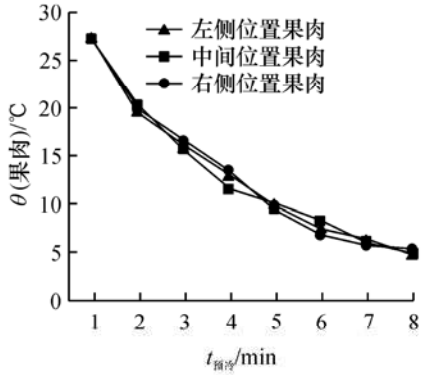


图 4　冰水预冷不同位置处平均温度变化过程

Fig. 4　The fruit flesh temperature changes of different positions for ice precooling

2.2　冷库预冷荔枝降温过程和温度均匀性分析

冷库预冷耗时 345 min 才将荔枝均温从 24.74 ℃（消毒处理后荔枝表面携带水分蒸发降温导致荔枝初始温度稍低于冰水预冷荔枝初温）降到 5.02 ℃. 图 5、6 给出了冷库预冷上中下不同层和左中右不同位置处果肉均温变化过程.

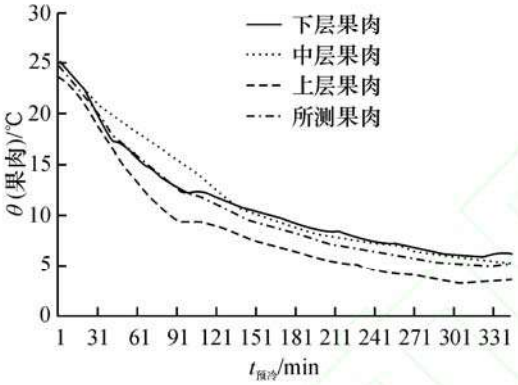


图 5　冷库预冷不同层处果肉温度变化过程

Fig. 5　The fruit flesh temperature changes of different layers for room precooling

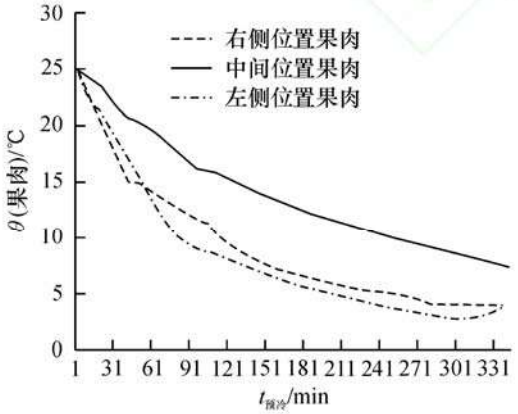


图 6　冷库预冷不同位置处荔枝果肉温度变化过程

Fig. 6　The fruit flesh temperature changes of different positions for room precooling

对比发现，上层荔枝降温较中下层快，约 228 min 果温降至近 5 ℃，而中下层果温 345 min 后仍分别高达 5.37、6.16 ℃. 原因可能是荔枝垂直堆垛，上层荔枝与冷空气接触机会多、降温快；中下层荔枝与冷空气直接接触机会少，降温慢. 同时，左侧靠近冷气出口位置荔枝降温最快，右侧次之，两者分别耗时 195 和 258 min 将左右侧果温降至近 5 ℃；中间荔枝接触冷空气困难，降温最慢，345 min 后果温仍高达 7.37 ℃.

每隔 50 min 选取 1 个时间点，对不同层和位置处荔枝果温均匀性分析发现：同一时期，上中下层荔枝果温无显著性差异；左侧荔枝均温最低，左右两侧荔枝均温无显著性差异，但左右位置与中间位置处荔枝果温差异显著. 可能原因是左右两侧荔枝与冷空气接触较充分，温度变化较一致；但中间位置处荔枝较难与冷空气接触，降温慢，温度较高. 冷库预冷靠近冷风口处易发生冻害. 所以，最好将荔枝置于远离冷风出口处，并尽量置于温度均匀的预冷区域.

2.3　差压和高湿差压预冷荔枝降温过程和温度均匀性分析

差压预冷过程将荔枝均温从 22.1 ℃降到 5.08 ℃约 55 min；高湿差压预冷速度较差压预冷降温速度慢，将荔枝果肉均温从 22.7 ℃降到 5.01 ℃需 64 min. 图 7、8（图 9、10）给出了差压（高湿差压）预冷不同层和不同位置处果肉温度变化过程.

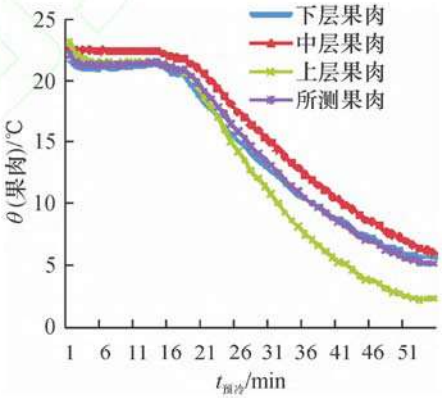


图 7　差压预冷不同层处荔枝果肉温度变化过程

Fig. 7　The fruit flesh temperature changes of different layers for forced-air precooling

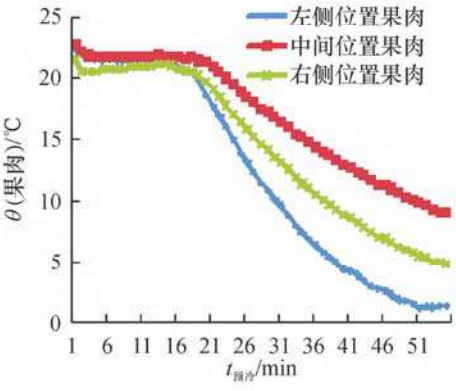


图 8　差压预冷不同位置处荔枝果肉平均温度变化过程

Fig. 8　The fruit flesh temperature changes of different positions for forced-air precooling

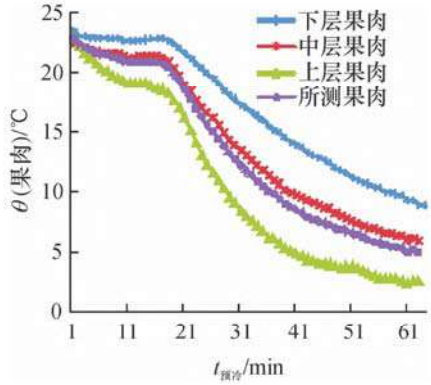


图9 高湿差压预冷不同层处果肉温度变化过程

Fig. 9 The fruit flesh temperature changes of different layers for forced-air precooling with high humidity

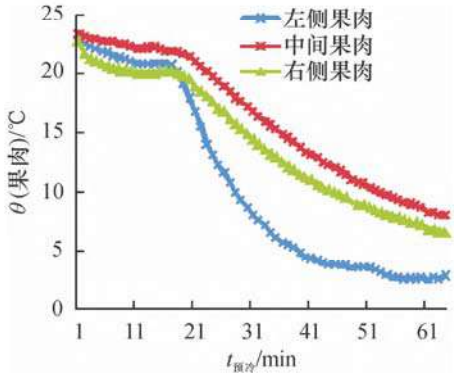


图10 高湿差压预冷不同位置处荔枝果肉平均温度变化过程

Fig. 10 The fruit flesh temperature changes of different positions for forced-air precooling with high humidity

由图7~10可以看出,2种预冷方式降温过程非常相似.15 min前,2种预冷方式荔枝不同层和不同位置处果温变化不大.不同层比,2种预冷方式均表现为上层降温速度最快(差压预冷和高湿差压预冷将上层果温降至近5 ℃分别耗时42和41 min;预冷结束时2种预冷方式中下层果温仍分别高达6.03、5.67和5.86、8.83 ℃).不同位置比,中间位置荔枝均温降速最慢(预冷结束时2种预冷方式中间层果温仍分别高达9.03、7.87 ℃),右侧次之,左侧最快(差压和高湿度差压均使用39 min将左侧果温降至近5 ℃).2种预冷方式上层降温最快的可能原因是荔枝并未填满差压箱试验区形成空穴,上层空穴通风阻力小,冷风快速流过,加速了上层荔枝降温.越到下层和中间,越难接触冷空气,降温越慢;左侧因靠近冷气出口,降温较快.

每隔10 min取1个预冷时间点,对差压和高湿差压预冷不同层和不同位置处荔枝果温均匀性进行分析发现:在温度开始稳定下降后(15 min后),差压和高湿差压预冷上层荔枝果温明显低于中下层荔枝果温.前20 min,左中右不同位置处温度无显著性差异;20 min后,左侧与右侧、中间与右侧位置处荔枝果温各无显著性差异,但左侧温度最低,且与中间位置处果温差异显著(其中差压预冷最大平均温差达8.8 ℃,高湿差压预冷最大平均温差达9.1 ℃).为实现荔枝完全预冷,2种预冷方式上层和靠近冷风口处荔枝往往易受冻害.

2.4 不同预冷方式降温过程与均匀性对比分析

图11给出了不同预冷方式荔枝平均温度变化过程.总体看,冰水、差压、高湿差压预冷和冷库预冷果温降速依次减缓;并一致表现出温度越低,降温速率越慢.冰水预冷主要通过热传导降温,水的热流密度大,所以降温迅速.而其他预冷主要通过空气(自然和强迫)对流降温,空气的热对流系数远小于水的热对流系数(200~1 000 W/m^2·℃),所以降温相对慢.同时,由于空气差压强迫对流时热对流系数(20~100W/m^2·℃)大于自然对流的(5~25 W/m^2·℃)的换热系数,所以冷库预冷果肉降温较差压和高湿差压预冷慢.且高湿环境影响荔枝热交换,高湿差压降温速度较差压预冷慢.

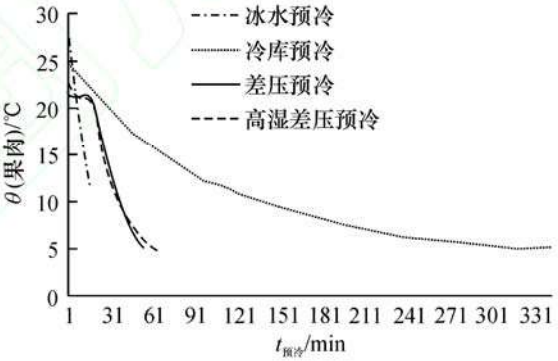


图11 不同预冷方式果肉平均温度变化过程

Fig. 11 The fruit fresh temperature changes of different precooling methods

从各预冷方式降温过程中果温标准差(图12)可以看出冰水预冷不同位置和层处总体温度最均匀,冷库预冷次之,差压预冷较高湿差压预冷均匀.同时可以看出,冷库、差压和高湿差压预冷方式温度标准差均表现出先增加、后下降的趋势.可能原因是开始降温时,荔枝初始温度较高,堆垛筐不同位置和层处荔枝接触冷源机会不同,上层和左侧荔枝热对流降温迅速,不同位置荔枝逐渐形成较大温度梯度,并不断增大;当预冷一段时间果温降到一定程度后,温度低处荔枝降温速度减缓,且较高温度差荔枝逐层接触亦发生热传导降温,不同层和位置处荔枝温度梯度逐渐缩小.因差压和高湿差压机理基本一致,温度不均匀特性相似.

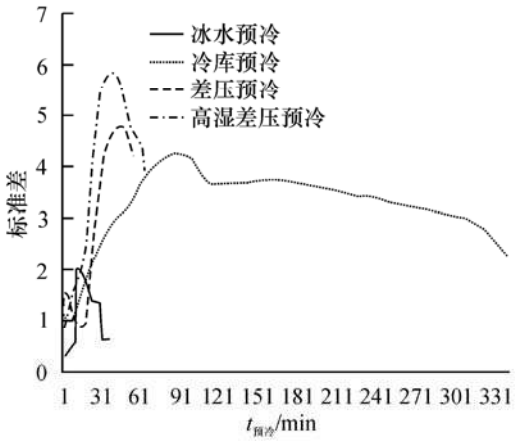


图 12　不同预冷方式果肉温度标准差

Fig. 12　Standard deviations of the rfuit flesh temperature for different precooling methods

3　结论

建立了差压预冷试验装置，采用冰水、冷库、差压以及高湿差压方式对荔枝预冷的降温规律和温度均匀性进行了对比分析，结果发现：

1）冰水预冷降温最快、差压次之、高湿差压更慢、冷库预冷最慢；且温度越低，降温速率越慢.

2）同一预冷时期，冰水预冷不同位置处、不同层处荔枝果温均无显著性差异. 冷库预冷左右与中间位置处荔枝果温差异显著，左侧靠近冷风口温度最低，左侧荔枝易受冻害. 差压和高湿差压预冷上层荔枝降温过程明显快于中下层降温过程；左侧荔枝降温较中右位置快，且与中间位置处荔枝果温差异显著；上层和左侧靠近冷风口的荔枝易受冻害.

3）温度标准差反应的温度均匀性显示冰水预冷温度最均匀，冷库预冷次之，差压预冷较高湿差压预冷均匀.

从预冷效率、均匀性和防冷冻害角度看，冰水预冷是较合适的预冷方式.

参考文献：

[1]　佚名. 荔枝预冷的目的与作用[EB/OL]. (2013-08-08)[2014-03-17]. http://lzlytx.scau.edu.cn/html/yanfazhongxin/zongheshiyanzhan/zhangzhouzongh/2012/0314/106.html.

[2]　阮文琉，刘宝林，宋晓燕. 荔枝的冷却方式选择[J]. 食品工业科技，2012，11：352-353.

[3]　LIN H T，CHEN S J，XI Y F. Commercial postharvest handling and storage technology of litchi fruit[J]. Trans CSAE，2003，5(19)：126-134.

[4]　王倩，戴绍碧，徐娓，等. 荔枝产地预冷装置的开发研究与实验[J]. 农机化研究，2012，7：100-104.

[5]　段洁利，杨洲，马征，等. 荔枝果实通风预冷试验研究[J]. 食品科学，2007，28 (7)：504-507.

[6]　杨洲，陈朝海，段洁利，等. 荔枝压差预冷包装箱内气流场模拟与试验[J]. 农业机械学报，2012，10 (42)：215-217.

[7]　宋晓燕，刘宝林. 真空冷却中的上海青表面温度变化规律[J]. 农业工程学报，2012，28 (1)：266-269.

[8]　宋小勇，李云飞，邓云，等. 鲜切花真空预冷过程温度的红外热成像检测[J]. 农业机械学报，2009，40 (11)：129-132.

[9]　DEHGANNYA J. NGADI M. VIGNEAULT C. Mathematic al modeling of airflow and heat transfer during forced convection cooling of produce considering various package vent areas[J]. Food Contr，2011，22:1393-1399.

[10]　ZOU Q A，OPARA L U，MCKIBBIN R. A CFD modeling system for air flow and heat transfer in ventilated packaging for fresh foods：II：Computational solution，software development，and model testing[J]. J Food Eng，2006，77 (4)：1048-1058.

[11]　DEFRAEYE T，VERBOVEN P，NICOLAI B. CFD modeling of flow and scalar exchange of spherical food products：Turbulence and boundary-layer modeling[J]. J Food Eng，2013，114(2)：495-504.

[12]　吕恩利，陆华忠，杨洲，等 番茄差压预冷过程中的通风阻力特性[J]. 农业工程学报，2010，26 (7)：341-345.

[13]　杨洲，赵春娥，汪刘一，等. 龙眼果实差压预冷过程中的阻力特性[J]. 农业机械学报，2007，38 (1)：104-107.

【责任编辑　霍　欢】

三、科研成果---通信作者发表论文清单

SCAULIB202625943

检索证明

根据委托人提供的论文材料，委托人工程学院 吕盛坪(学科类型:自然科学) 10 篇论文收录情况如下表。

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	A genetic algorithm enhanced with neighborhood structure for general flexible job shop scheduling with parallel batch processing machine	EXPERT SYSTEMS WITH APPLICATIONS 出版年：2025 出版日期：MAR 15 卷期：265 页码：- 文献号：125888 文献类型：Article	通讯作者	T2 类	华南农业大学工程学院	SCI	IF2-year=7.5 IF5-year=7.8 （2024）	计算机科学 1区 Top 期刊：是 OA 期刊：否 标注：Mega-Journal （2025）
2	Detection of breakage and impurity ratios for raw sugarcane based on estimation model and MDSC-DeepLabv3+	FRONTIERS IN PLANT SCIENCE 出版年：2023 出版日期：NOV 8 卷期：14 页码：- 文献号：1283230 文献类型：Article	通讯作者	A 类	华南农业大学工程学院	SCI	IF2-year=4.1 IF5-year=5.3 （2023）	生物学 2区 Top 期刊：是 OA 期刊：是 （2023）
3	YOLO-DSD: A YOLO-Based Detector Optimized for Better Balance between Accuracy, Deployability and Inference Time in Optical Remote	APPLIED SCIENCES-BASEL 出版年：2022 出版日期：AUG 卷期：12 15 页码：-	通讯作者	B 类	华南农业大学工程学院	SCI	IF2-year=2.7 IF5-year=2.9 （2022）	综合性期刊 4区 Top 期刊：否 OA 期刊：是

	Sensing Object Detection	文献号：7622 文献类型：Article						（2022）
4	A hybrid teaching-learning-based optimization algorithm for QoS-aware manufacturing cloud service composition	COMPUTING 出版年：2022 出版日期：NOV 卷期：104 11 页码：2489-2509 文献类型：Article	通讯作者	B 类	华南农业大学工程学院	SCI	IF2-year=3.7 IF5-year=3.2 （2022）	计算机科学 3区 Top 期刊：否 OA 期刊：否 （2022）
5	YOLOv4-MN3 for PCB Surface Defect Detection	APPLIED SCIENCES-BASEL 出版年：2021 出版日期：DEC 卷期：11 24 页码：- 文献号：11701 文献类型：Article	通讯作者	B 类	华南农业大学工程学院	SCI	IF2-year=2.838 IF5-year=2.921 （2021）	工程技术 4区 Top 期刊：否 OA 期刊：是 （2021）
6	PCB 表面缺陷数据集与基于 YOLOv5s-P6SE 的检测	计算机工程与科学 出版年：2025 出版日期：2025-02-15 卷期：47 02 页码：- 文献号： 文献类型：期刊论文	通讯作者	C 类	华南农业大学工程学院	北大核心	无	是

7	考虑强制同机并行作业的广义作业车间调度优化	计算机应用研究 出版年：2024 出版日期：2024-03-18 13:45 卷期：41 08 页码：- 文献号： 文献类型：期刊论文	通讯作者	C 类	华南农业大学工程学院	北大核心	无	无
8	基于 GENI-SD 的定制化印制电路板工序重要性评估	计算机应用研究 出版年：2023 出版日期：2023-01-12 15:46 卷期：40 05 页码：- 文献号： 文献类型：期刊论文	通讯作者	C 类	华南农业大学工程学院	北大核心	无	无
9	基于时间加权改进的 LDTW 算法	计算机应用研究 出版年：2022 出版日期：2021-12-14 16:35 卷期：39 04 页码：- 文献号： 文献类型：期刊论文	通讯作者	C 类	华南农业大学工程学院	北大核心	无	无

10	基于自组织映射-反向传播网络的PCB样板投料预测	计算机应用与软件 出版年：2020 出版日期：2020-08-12 卷期：37 08 页码：- 文献号： 文献类型：期刊论文	通讯作者	C类	华南农业大学工程学院	北大核心	无	无

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明:如未盖章,报告无效

检索员：尹银怀

华南农业大学图书馆

2026-03-12

Expert Systems With Applications 265 (2025) 125888

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A genetic algorithm enhanced with neighborhood structure for general flexible job shop scheduling with parallel batch processing machine

Hucheng Zhang, Shengping Lv [*,1], Dequan Xin, Hong Jin

School of Engineering, South China Agricultural University, 510642, China

ARTICLE INFO

Keywords:
Parallel batch processing
General flexible job shop scheduling
Genetic algorithm
Neighborhood structure

ABSTRACT

In this study, a general flexible job shop scheduling problem with parallel batch processing machine (GFJSP_PBPM) is presented. GFJSP_PBPM allows multiple jobs, whether mandatory or flexible, to be simultaneously processed on the same machine, challenging the conventional constraints of the flexible job shop scheduling problem where a single machine can only undertake one job at any given time. The motivation for this problem arises from real-world scenarios encountered in electronic product performance testing and mold manufacturing workshops. Firstly, the problem of GFJSP_PBPM is defined, and an optimization model is established with the objective of minimizing the makespan using mixed-integer programming. Subsequently, a genetic algorithm enhanced with neighborhood search (GANS) is developed to efficiently tackle the problem at different scales. To evaluate its performance, benchmark instances are created for testing and comparative analysis. Through testing on these instances and a real-world engineering case, the feasibility and superiority of the proposed GANS are demonstrated.

1. Introduction

The general flexible job shop scheduling problem (GFJSP) with parallel batch processing machine (GFJSP_PBPM), studied in this work, is an extension of the classical flexible job shop scheduling problem (FJSP) (Brucker & Schlie,1990; Brandimarte, 1993). GFJSP_PBPM allows the same machine to concurrently process multiple jobs, thereby challenging the traditional constraints of the FJSP, where a machine can only handle one job at any given time. This problem is inspired by real situations observed in electronic product performance testing and mold manufacturing workshops.

In the process of electronic product performance testing, the workshop designs an overall testing process plan for prototypes of the same product model. These prototypes are categorized into different groups. Each group of prototypes undergoes testing sequentially according to the specified sub-routes in the overall process plan. Fig. 1 depicts a performance testing process plan of an in-vehicle navigator, where 14 prototypes are divided into 7 groups. Each group of prototypes or testing units is regarded as a single job, with the quantity of prototypes (units) in each group considered as its weight. However, certain prototypes require cross-group combination testing, thus establishing mandatory parallel batch processing operation (PBPO) on the same machine, exemplified by ($O_{1,3}$,$O_{2,4}$) and ($O_{3,7}$,$O_{4,6}$) in Fig. 1. Furthermore, aiming to minimize processing time, cost, and energy consumption, specific machines within the workshop permit prototypes from different models or even different types of products to undergo parallel testing on the same machine, provided they meet constraints such as load capacity and parameter ranges. This leads to the flexible PBPO on the same machine.

Similarly, certain processes in mold production require mandatory PBPO. For instance, mold cavity structures are frequently designed as assemblies of multiple inserts. During the cavity feature machining process, it is common practice to concurrently machine the precisely assembled inserts to meet accuracy requirements, thereby leading to the mandatory PBPO. Other features on each insert are typically processed separately according to flexible process plans. Meanwhile, for certain machine such as heating furnace, multiple jobs can be processed in the furnace simultaneously. However, there is no pre-imposed constraint on the combination of jobs for this batch processing. This flexibility leads to flexible PBPO. Lin et al. (2022) developed a job-constraint genetic algorithm (GA) tailored for job shop scheduling problem (JSP), focusing particularly on the parallel processing of mold manufacturing and assembly. Liang et al. (2018) introduced an improved hybrid immune

* Corresponding author.
E-mail address: lvshengping@scau.edu.cn (S. Lv).
[1] 0000–0001-6480–4110.

https://doi.org/10.1016/j.eswa.2024.125888
Received 14 June 2024; Received in revised form 4 October 2024; Accepted 20 November 2024
Available online 26 November 2024

algorithm for FJSP, and validated its effectiveness through a practical case in a mold workshop. Zhong et al. (2020) proposed a hierarchical production control framework for mold manufacturing. However, it is worth noting that these studies did not address the practical constraint of parallel batch processing.

Solving the GFJSP_PBPM is a crucial problem urgently needing resolution in electronic product performance testing and mold manufacturing workshops. GFJSP_PBPM introduces additional complexities by incorporating parallel batch processing machines (PBPM) with capacity limitations, further complicating the already established NP-hard nature of FJSP (Sotskov, 1991), thus escalating the challenge of its resolution. To address this problem, the present study conducts corresponding research on the modeling and optimization method of GFJSP_PBPM. The specific innovative contributions are as follows:

(1) The GFJSP_PBPM is defined with the objective of minimizing the maximum completion time. The optimization model for GFJSP_PBPM is established based on mixed-integer programming (MIP).
(2) A genetic algorithm (GA) enhanced with neighborhood search (GANS) is proposed for GFJSP_PBPM.
(3) A hybrid initialization strategy is developed to generate diverse and high-quality feasible solutions. This strategy is supported by a customized task-based encoding and active schedule-based decoding tailored to address constraints originating from GFJSP_PBPM.
(4) A novel Split-One by One Order Crossover (Split-OOOX) and a hybrid mutation are designed to ensure the feasibility and diversity of new solutions during the GA iteration, thereby guaranteeing the GA's global search capability.
(5) To address constraints related to both FJSP and PBPO, better optimize problems of different scales, and leverage the advantages of N5 and N7 neighborhood structures, a hybrid N5/7 is proposed. Subsequently, this hybrid N5/7 is integrated into GA to bolster its local search capabilities.

The subsequent sections are structured as follows: Section 2 reviews related works on FJSP and FJSP with parallel batch processing machine, GA enhanced with neighborhood structure for FJSP. Section 3 presents the problem description and model of GFJSP_PBPM. Section 4 delineates the design of the GANS, encompassing population hybrid initialization, encoding, decoding, genetic operations, and the hybrid N5/7 neighborhood structure. Section 5 presents the constructed benchmark instances, corresponding experiments, case study, and results, followed by an analysis and discussion of the experimental outcomes. Finally, Section 6 concludes with a summary of key findings and suggestions for future research.

2. Literature review

2.1. FJSP and FJSP with parallel batch processing machine

The FJSP has evolved from the classical JSP, incorporating flexible features such as machine substitutability, operation substitutability, and sequencing flexibility (Ozguven et al., 2010; Xie et al., 2019; Dauzère-Pérès et al., 2023). The FJSP is prevalent in discrete manufacturing industries, including aerospace (Zhou et al., 2019), automotive (Huang et al., 2023), and electronics (Fan & Su, 2022). It presents a combinatorial optimization challenge constrained by multiple equations and inequalities. Since its emergence in the early 1990 s (Brucker and Schlie, 1990; Brandimarte, 1993), the FJSP has attracted ongoing scholarly research. Recent studies have introduced additional constraints for FJSP, such as transport resources (Fontes et al., 2023; Zhang et al., 2024), molds (Hu et al., 2023), human–machine resources (Meng et al., 2019; Zhang et al., 2021), preparation time (Zhang et al., 2022) and uncertain processing times (Gao et al., 2016; Chen et al., 2022). These advancements have significantly enhanced the FJSP's applicability to real-world workshop optimization needs (Dauzère-Pérès et al., 2023).

Building on this foundation, corresponding scholars further explored the FJSP with PBPM (denoted as FJSP_PBPM) (Fowler & Mönch, 2022). Ham and Cakici (2016) constructed an optimization model using both MIP and constraint programming (CP) for the FJSP_PBPM in semiconductor manufacturing processes. Subsequently, they utilized IBM ILOG CPLEX to solve it, and the results demonstrated the superior performance of the CP model. Similarly, Ham et al. (2017) modeled the FJSP_PBPM, taking into account both machine capacity constraints and the impact of job priority using MIP. Once again, they employed IBM ILOG CPLEX for the solution. However, these studies only optimized 2

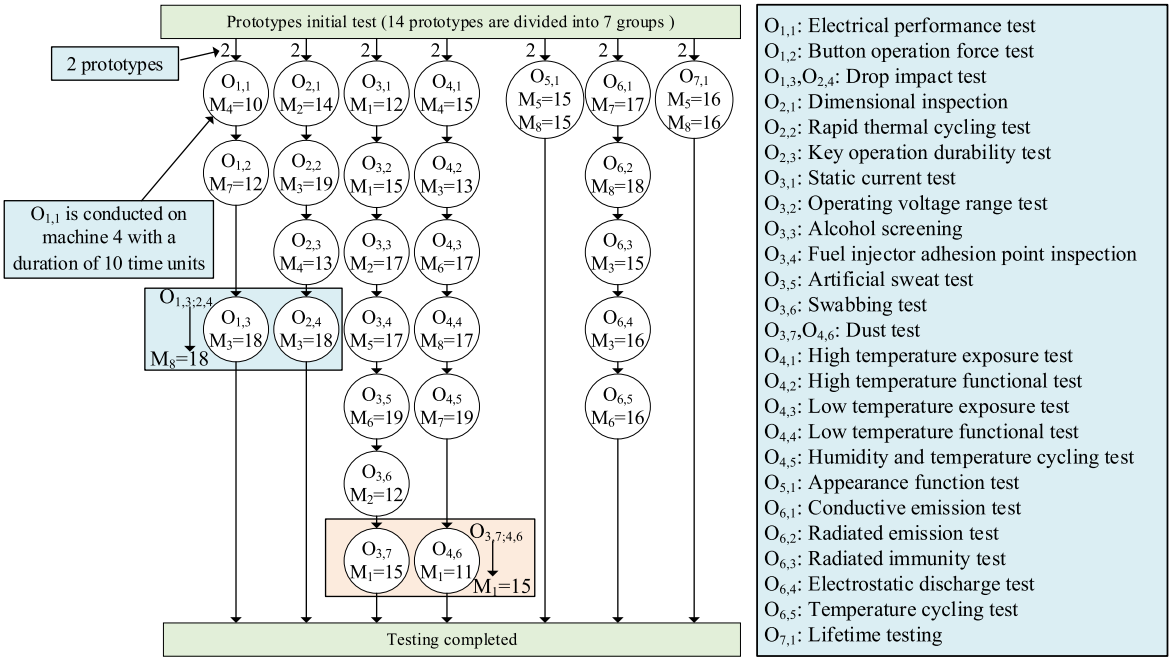


Fig. 1. Performance testing process plan of the in-vehicle navigator.

~ 4 consecutive operations involving PBPM, such as burn-in operation or diffusion process, rendering it a local scheduling for a few manufacturing stages of the jobs. Inspired by similar issues, Knopp et al. (2017) investigated the FJSP_PBPM with the constraints of reentrant flows, sequence-dependent setup times, and release dates. To address this problem, they introduced novel disjunctive graph for feasible solution representation and employed a greedy randomized adaptive search procedure (GRASP) for its solution.

Boyer et al. (2021) studied the FJSP_PBPM in seamless rolled ring manufacturing, referring to it as GFJSP. They devised an optimization model tailored for this manufacturing process using both MIP and CP, which encompassed considerations such as machine capacity, time lags, holding times, and sequence-dependent setup times. Additionally, the researchers proposed a GRASP for solution exploration. Within this framework, PBPM includes heating jobs in a furnace, usually following a first-in, first-out order, which leads to a PBPO. Moreover, operations such as cutting, pressing, or rolling serve as fixed preceding operations, and pressing or rolling also act as fixed succeeding operations of the PBPO. In contrast, the operations examined in our study require the completion of predecessor operations for all relevant jobs before commencing the PBPOs.

The existing studies of FJSP_PBPM have directly inspired our research. However, current studies mainly focus on the new constraint of PBPM with flexible PBPO and only optimize a few consecutive steps involving PBPM. In contrast, our study requires the comprehensive consideration of both mandatory and flexible PBPO. Furthermore, our research aims to globally optimize the scheduling for all operations of the involved jobs.

2.2. GA enhanced with neighborhood structure for FJSP

Due to its notable features such as robust parallelism, adaptability, and autonomous learning, the GA has attracted significant interest from scholars worldwide for addressing the FJSP (Li and Gao, 2016; Huang and Yang, 2019; Chen et al., 2020; Park et al., 2021). The GA exhibits robust global search capability, enabling rapid identification of numerous solutions within the search space for straightforward problems. However, its effectiveness is constrained when tackling more intricate challenges (Sun et al., 2023). The "No Free Lunch" theorem (Wolpert & Macready, 2005) indicates that optimization mechanisms designed with domain-specific neighborhood structures tailored to the FJSP yield advantages in both exploration and exploitation, enhancing the efficiency of obtaining high-quality solutions. As a result, various applications of neighborhood structures have been incorporated into GA for addressing the FJSP, aiming to improve both solution quality and efficiency (Gao et al., 2008; Türkyılmaz & Bulkan, 2014; Palacios et al., 2015; Zhang et al., 2019; Liu et al., 2021; Sun et al., 2023; Xie et al., 2023a).

In the research of GA enhanced with neighborhood structures for FJSP, Gao et al. (2008) utilized dual vectors to represent solutions and incorporates sophisticated crossover and mutation operators tailored to the unique chromosome structure and characteristics inherent in the FJSP. To bolster the search capability, individual solutions within the GA undergo refinement through a variable neighborhood descent (VND). A comprehensive computational study, involving 181 benchmark problems, was conducted to assess the efficacy of the proposed approach.

Türkyılmaz and Bulkan (2014) developed a novel approach for FJSP by combining a GA with a variable neighborhood search (VNS) strategy. The GA incorporates advanced crossover and mutation operators to tailor the chromosome structure to the specific characteristics of the problem. The VNS employs local search techniques, which involve assigning operations to alternative machines and altering the order of selected operations on the assigned machine, to enhance result quality while ensuring feasibility. The effectiveness of the proposed method is verified through numerical experiments on diverse representative problems, and its performance is compared with results obtained from adapted constructive heuristic algorithms.

Palacios et al. (2015) presented an efficient GA integrated with Tabu search (TS) and heuristic seeding to minimize the makespan of FJSP with uncertain processing times. The proposed approach involves introducing a heuristic method to generate a diverse and high-quality set of initial solutions. Each generated chromosome subsequently undergoing TS. The TS relies on a neighborhood structure, proposed and analyzed in this paper, with proven properties such as feasibility and connectivity. Additionally, a filtering mechanism was incorporated to reduce the neighborhood size and accelerate the evaluation of new chromosomes. The experimental outcomes unequivocally demonstrate that the hybrid algorithm not only capitalizes on the synergy among its components but also stands out as a formidable competitor against SOTA methods in addressing both crisp and fuzzy instances. Remarkably, the hybrid algorithm establishes novel best-known solutions for a significant number of test instances.

Zhang et al. (2019) proposed a hybrid algorithm combines GA with VNS to enhance search ability and balance intensification and diversification. The hybrid algorithm utilizes systematic neighborhood search structures and an improved external library to save optimal or near optimal solutions. Computational results and comparisons demonstrate the efficiency and effectiveness of the proposed hybrid algorithm.

Liu et al. (2021) established a novel algorithm called the variable neighborhood descent hybrid genetic algorithm (VND-hGA) to solve the FJSP. The proposed VND-hGA integrates a barebones particle swarm optimization-based mutation operator, a hybrid heuristic initialization strategy, and a VND based on an improved multilevel neighborhood structure into the standard GA framework. The VND-hGA was tested on benchmark cases, with the results showing superior solution accuracy and convergence performance compared to existing approaches.

Sun et al. (2023) offered a hybrid GA with VNS (HGA-VNS) for solving the FJSP. The HGA-VNS represents each solution with a chromosome consisting of two parts: the code of the machine number and the code of the operation number. It incorporates combined crossover and mutation operators that take into account machine workload balance. Additionally, it employs a local search approach for critical path operations to minimize the number of invalid transformation. The HGA-VNS was tested on extended instances based on well-known benchmarks and the results demonstrate its superior performance compared to other algorithms.

Xie et al. (2023a) introduced a hybrid genetic tabu search algorithm (HGTSA) tailored for the distributed FJSP. HGTSA strategically combines the global search capability of the GA with the local search proficiency of Tabu Search (TS). Two genetic operators are specifically designed based on the critical factory, enabling effective population discretization. Additionally, a novel neighborhood structure N8 is integrated into TS, expanding the search space for neighborhood solutions. The experimental results indicate the superiority of HGTSA in terms of solution quality and computational efficiency.

This study attempts to integrate GA and neighborhood structures to solve the GFJSP_PBPM. However, the GA and neighborhood structure proposed in the aforementioned studies were designed specifically for the FJSP. Following the introduction of the additional constraints of PBPO, the representation of individuals, decoding, population initialization, crossover, mutation, and neighborhood solution generation, among others, all require a redesign.

3. Problem description

GJSP_MPBPO is an extension of the well-known NP-hard problem FJSP. It involves a set of machines M and a set of jobs J, where each job has a predetermined process plan. These process plans comprise not only a series of ordered operations, each of which can be processed on a set of alternative machines with specific processing times, but also involve constraints of mandatory and flexible PBPOs on the same machine, as long as the batch size does not exceed the machine's capacity. The objective of the GFJSP_PBPM is to determine the processing order of

operations on each machine with the aim of optimizing the makespan.

The GFJSP_PBPM inherits the complexities of the FJSP and introduces additional intricacy by incorporating constraints of PBPO. Fig. 2 depicts the process plans of five jobs, in which $O_{2,2}$ (operation 2 of job 2), $O_{3,3}$, $O_{4,5}$ and $O_{5,5}$ form a flexible PBPO (FPBPO). The $O_{2,4}$ and $O_{3,4}$ form a mandatory PBPO (MPBPO). For clarity, we collectively define the operation of a job, FPBPO, and MPBPO as task. Similar to FJSP, the GFJSP_PBPM assume that all jobs can be processed at time zero, with each machine able to handle only one task (an operation, a MPBPO, or a FPBPO) at a time, and each job being processed on only one machine at a time. Once a task begins on a machine, it cannot be interrupted until completion. Additionally, each task can only be processed after its preceding tasks have been completed.

Based on the above definition and assumption, MIP is employed to formulate an optimization model with the aim of minimizing the maximum completion time (C_{max}). For clarity, the notation used in this model is summarized as follows:

J	Jobs(j)
O_j	Operations of job j (o)
M	Machines (m)
$MPBPO_k$,$FPBPO_k$	The k^{th} MPBPO, and FPBPO respectively
N_{MPBPO},N_{FPBPO}	The number of MPBPO, and FPBPO respectively
O_l^m	Task (an operation or a PBPO) processing at the positon (consecutive time interval) l on the machine m
$O_{j,o}$	The operation o of job j
$Pt_{j,o}^m$	The processing time of $O_{j,o}$ on the machine m
$MP_{m,l}^{pt}$	The processing time at the positon l on the machine m
Cap^m	Capacity of machine m
$w_{j,o}^m$	Weight of $O_{j,o}$ on the machine m (weight or number of units for job j)
Q	An infinite real number
$[S_{j,o}^{m,l}, E_{j,o}^{m,l}]$	Start time and completion time of $O_{j,o}$ at the positon l on the machine m
$[MP_{m,l}^{st}, MP_{m,l}^{ed}]$	Start time and completion time of the l^{th} positon on the machine m
$X_{j,o}^{m,l}$	1 if $O_{j,o}$ occupies positon l on the machine m; 0 otherwise
$Y_{j,o}^k$	1 if $O_{j,o}$ is the operation in $MPBPO_k$; 0 otherwise
$Z_{j,o}^k$	1 if $O_{j,o}$ is the operation in $FPBPO_k$; 0 otherwise

The proposed MIP model for GFJSP_PBPM is based on the one established by Ham (2017) for FJSP_PBPM with only FPBPO. It is formulated as follows:

$$C_{\max} = \min\{\max\{E_{j,o}^{m,l}\}\}, \forall j, o, m, l \tag{1}$$

Subject to:

$$\sum_{m,l} X_{j,o}^{m,l} = 1 \tag{2}$$

$$\sum_{j,o} w_{j,o}^m \times X_{j,o}^{m,l} \times Y_{j,o}^k \le Cap^m \forall j, o, l, k, m \tag{3}$$

$$\sum_{j,o} w_{j,o}^m \times X_{j,o}^{m,l} \times Z_{j,o}^k \le Cap^m \forall j, o, l, k, m \tag{4}$$

$$MP_{m,l}^{pt} = \max\{Pt_{j,o}^m \times X_{j,o}^{m,l}\} \forall j, o, m, l \tag{5}$$

$$S_{j,o}^{m,l} \ge 0 \forall j, o, m, l \tag{6}$$

$$MP_{m,l}^{st} \ge E_{j,o-1}^{m',l'} \times X_{j,o-1}^{m',l'} + Q(X_{j,o}^{m,l} - 1) \ \forall j, o, m, l, m', l' \tag{7}$$

$$MP_{m,l}^{ed} = MP_{m,l}^{st} + MP_{m,l}^{pt} \forall j, o, m, l \tag{8}$$

$$S_{j,o+1}^{m,l} \ge MP_{m',l'}^{ed} \times X_{j,o}^{m',l'} + Q(X_{j,o}^{m',l'} - 1) \ \forall j, o, m, l, m', l' \tag{9}$$

$$MP_{m,l+1}^{st} \ge MP_{m,l}^{ed} \forall m, l \tag{10}$$

$$C_{\max} \ge MP_{m,l}^{ed} + Q(X_{j,o}^{m,l} - 1) \ \forall j, o, m, l \tag{11}$$

$$\sum_{O_{j,o} \in MPBPO_k} X_{j,o}^{m,l} \times Y_{j,o}^k = |MPBPO_k| \forall m, l : 1 \le k \le N_{MPBPO} \tag{12}$$

$$1 \le \sum_{O_{j,o} \in FPBPO_k} X_{j,o}^{m,l} \times Z_{j,o}^k \le |FPBPO_k| \forall m, l : 1 \le k \le N_{FPBPO} \tag{13}$$

Eq. (1) represents the objective function for optimization, where Eq. (2) indicates that any operation on a machine occurs only once. Eqs. (3) and (4) ensure that the total weight of a mandatory PBPO and flexible PBPO on a given machine does not exceed the machine's capacity, respectively. Eq. (4) ensures that the total weight of operations in a PBPO on a machine does not exceed the machine's capacity. Eq. (5) define the processing time of a task on a machine. Eq. (6) ensures that the start time of all operations is non-negative. Eq. (7) guarantees that the available time of an operation is greater than or equal to the completion time of its immediate job predecessor. Eq. (8) ensures that once a machine starts processing, it cannot be interrupted. Eq. (9) guarantees the precedence order between operations of the same job, implying that the start time of any operation must not be earlier than the completion time of its immediate job predecessor. Eq. (10) ensures precedence order constraints between tasks on the same machine. Eq.

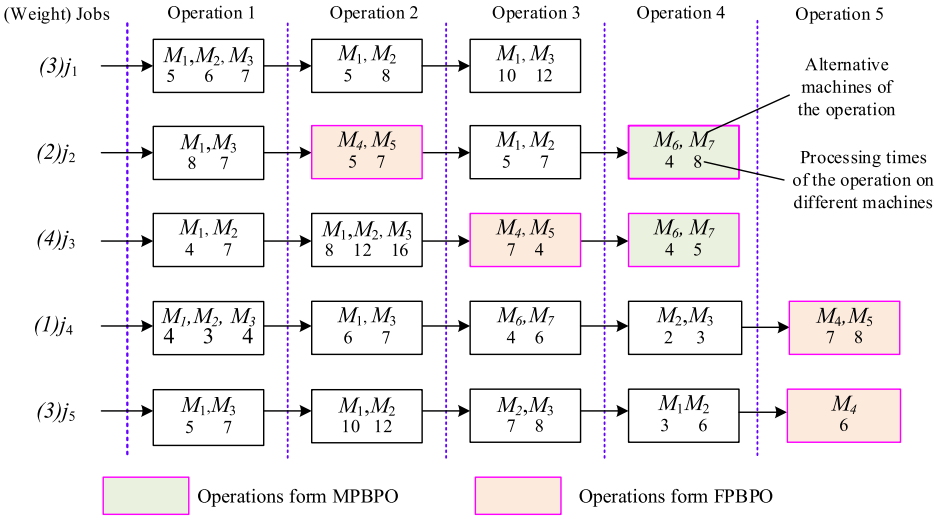


Fig. 2. Process plans of five jobs.

(11) ensures that the completion time of any machine does not exceed the makespan. Eq. (12) require that operations belonging to the same MPBPO must occur simultaneously on the same machine. Eq. (13) denotes that operations belonging to the same FPBPO can occur simultaneously on the same machine.

4. Proposed GANS for GFJSP_PBPM

4.1. Framework of GANS

To ensure the applicability of the proposed GANS to GFJSP_PBPM and achieve optimal results, a task-based encoding and active schedule-based decoding method, tailored for constraints originating from FJSP and PBPO, are designed to respectively represent solutions and resolve chromosomes into scheduling scheme. On this basis, a new hybrid initialization strategy is developed to initialize diverse and high-quality feasible solutions. Additionally, novel Split-OOOX and hybrid mutation operator are designed to maintain feasibility and promote diversity throughout the GA iteration process, thereby enhancing the algorithm's global search capability. Furthermore, a neighborhood structure, hybrid N5/7, is specifically crafted and seamlessly integrated into the GA to strengthen local search capability. The framework of GANS is presented in Fig. 3 and is described in detail below.

Step 1: The flexible process plan (*PPlan*) of each job, including job and operation set, operation sequence, alternative machines with their associated processing times for each operation, and the PBPOs, are input.
Step 2: Parameters such as the population size, crossover/mutation rate P_m/P_C, max iterations, elapsed time, preset time, scale factor for selecting chromosome for neighborhood search, are setting according to some initial experiments.
Step 3: An initial population (Pop) is generated using the hybrid initialization strategy. In this strategy, each individual chromosome in the Pop is represented according to the customized task-based encoding method. Job sequences represented by job numbers are created utilizing a classification random strategy, with distinct conditions designed for non PBPO task, MPBPOs and FPBPOs. A hybrid random and greedy strategy is applied to allocate machines and corresponding processing time for each task in the sequence.
Step 4: Each chromosome is decoded through the designed active schedule-based decoding method to generate feasible scheduling scheme, and the objective value C_{max} is get for evaluation.
Step 5: If the iterations reaches the max iterations or the elapsed time exceeds its preset time, the best solution is output, and the process stops; else, it continues to Step 6.
Step 6: Roulette wheel selection is applied to generate Pop1 based on Pop. Subsequently, intentionally designed Split-OOOX and hybrid mutation are employed as the crossover and mutation operators, respectively. These mechanisms ensure the feasibility and diversity of new newly generated offspring, contributing to the creation of Pop2.
Step 7: Chromosomes from Pop2 are selected according to the scale factor, and the hybrid N5/7 neighborhood structure is applied to these selected individuals during local search, ultimately resulting in the generation of the Pop3.
Step 8: Elite retention is employed to generate a new Pop from {Pop1, Pop2, Pop3} for next iteration. The procedure is repeated until the iterations reaches max iterations.

4.2. Encoding and decoding

4.2.1. Task-based encoding

In the field of GA-based approaches for FJSP, the operation-based encoding method is a commonly employed for encoding. In this method, each individual chromosome in the population is represented by a permutation of job numbers. These numbers serve as job IDs, and their frequencies indicate the number of operations within the process plan. Additionally, each operation in the encoded sequence is specifically assigned to a particular machine and processing time (Li and Gao, 2016; Chen et al., 2020), thus generating the machine and time sequences. These encodings conventionally stipulate that each encoding position in a job ID sequence corresponds to an operation for a job. However, the PBPO involved in GFJSP_PBPM introduce coupled effects across multiple jobs during encoding, leading to inconsistent elements in each gene and flexible variations in the length of the encoded sequence.

To address the aforementioned issues and drawing inspiration from operation-based encoding, this study designs a new task-based encoding method for GFJSP_PBPM. The encoding consists of two segments of integer coding sequences, as illustrated in Fig. 4. The first segment is the job sequence (*JPlan*), each gene is represented by the job IDs. To ensure consistency in each gene and the length of each chromosome, the number of elements in each gene is set to $s = \max\{|MPBPO_k|, |FPBPO_k|\}, \forall k$; and the length is set to $l = \sum_{j\in J}|O_j| - \sum_{k=1}^{N_{MPBPO}}|MPBPO_k| + N_{MPBPO}$. For elements less than s, 0 is used to pad the gene. If the number of jobs in a

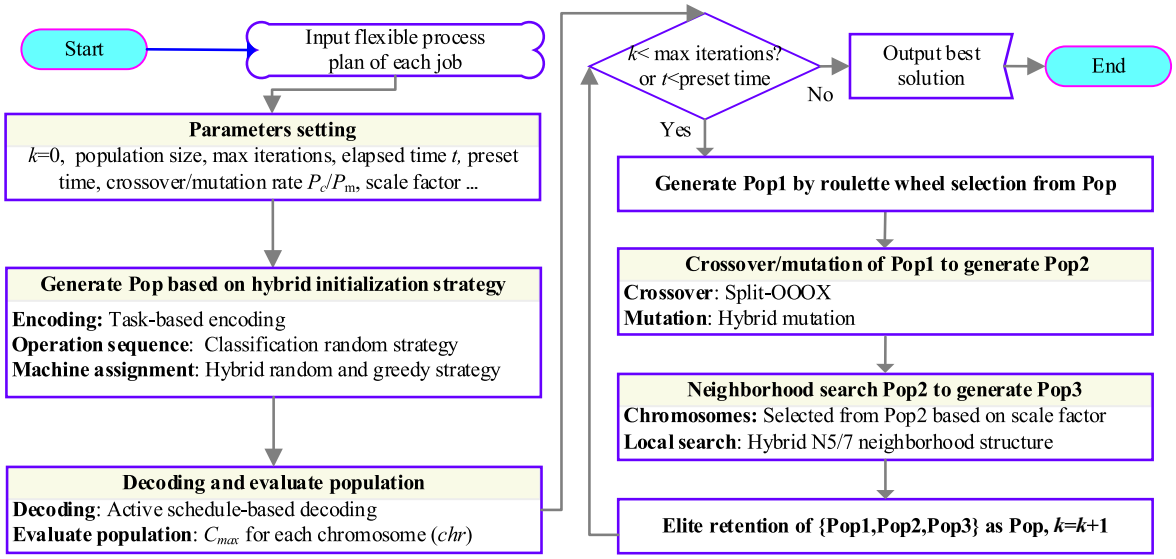


Fig. 3. Framework of GANS for GFJSP_PBPM.

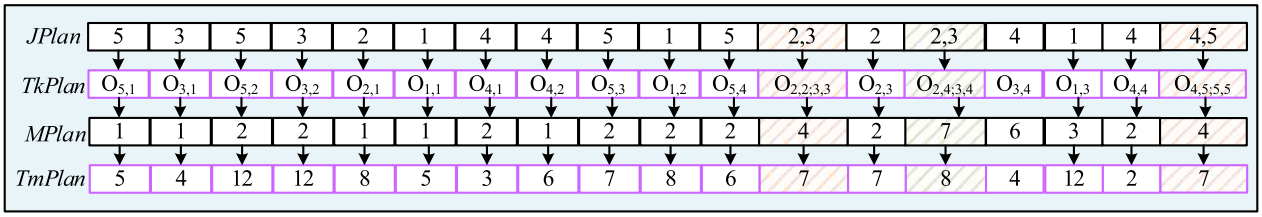


Fig. 4. Encoding instance of schedule plan for jobs in Fig. 2 (omitting 0 s in each gene).

gene is exceeds 1, it indicates that the gene corresponds to a PBPO. To simplify the description, the padded 0 s and the all-zero genes are omitted in the figure and the following description. When mapping *JPlan* to the task sequence (*TkPlan*), the frequency of occurrence for each job indicates the corresponding operation. For PBPO task, it is imperative to encompass all the associated operations. For example, job 2 and 3 in the shaded box of *JPlan* represent a FPBPO of $O_{2,2}$ and $O_{3,3}$, gene with number 2 and 3 represent a MPBPO of $O_{2,4}$ and $O_{3,4}$, and gene coded with 4 and 5 represent another FPBPO of $O_{4,5}$ and $O_{5,5}$. The second segment is the machine sequence (*MPlan*), where each encoding position has a range of potential values from 1 to *M*, and each gene in *MPlan* corresponds to the machine for task in *TkPlan*. On the basis of *TkPlan* and *MPlan*, the time sequence (*TmPlan*) representing the required processing time for each task on the designated machine can be obtained as well. Therefore, the complete information of a chromosome can be represented by $Chr = \{JPlan, TkPlan, MPlan, TmPlan\}$.

4.2.2. Active schedule-based decoding

Bierwirth and Mattfeld (1999) suggested the decoding procedures for translating encoding permutations into semi-active, active, non-delay, and hybrid schedules. In this study, active schedule-based decoding is adopted to generate high-quality solutions. This decoding process revolves around ensuring that the decoded operation simultaneously satisfies the completion time constraints of both the machine predecessor and the job predecessor. However, for the GFJSP_PBPM, it is essential to comprehensively consider the completion times of all predecessors for each job in the PBPO. The constraints considered in the decoding process become more intricate. The designed active schedule-based decoding for GFJSP_PBPM is illustrated in Algorithm 1.

Algorithm 1 Active schedule-based decoding

Input: $Chr = \{JPlan, TkPlan, MPlan, TmPlan\}$// *A chromosome for decoding*
$EMP = zeros(|M|)$//*The completion time of the immediate machine predecessor for each machine*
$EJP = zeros(|J|)$//*The completion time of the immediate job predecessor for each job*
$[t\_s, t\_e]_m = \emptyset, m \in M$//*Idle time intervals of machine m*
$SPlan = \emptyset$// *Scheduling scheme (result)*
$i = 1$
while $i \leq Length(TkPlan)$ ***do***
$O \leftarrow TkPlan(i)$// *An operation or a PBPO*
Determine the job set J_O *of task O*
$m \leftarrow MPlan(i), Pt \leftarrow TmPlan(i), AS \leftarrow \max\{EJP[j], \forall j \in J_O\}$
if $\exists [t\_s, t\_e] \in [t\_s, t\_e]_m, \max\{AS, t\_s\} + Pt \leq t\_e$ ***then***
$S \leftarrow \max\{AS, t\_s\}, E \leftarrow S + Pt, [t\_s, t\_e]_m \leftarrow [t\_s, t\_e]_m / [t\_s, t\_e]$
if $S - t\_s \geq \min\{TmPlan(k) | MPlan(k) = m, i < k \leq Length(TmPlan)\}$ ***then***
$[t\_s, t\_e]_m \leftarrow [t\_s, t\_e]_m \cup [t\_s, S]$
end if
if $t\_e - E \geq \min\{TmPlan(k) | mPlan(k) = m, i < k \leq Length(TmPlan)\}$ ***then***
$[t\_s, t\_e]_m \leftarrow [t\_s, t\_e]_m \cup [E, t\_s]$
end if
$EJP(j) \leftarrow E, \forall j \in J_O, SPlan \leftarrow SPlan \cup [O, m, S, E]$
end if
if $\forall [t\_s, t\_e] \in [t\_s, t\_e]_m, \max\{AS, t\_s\} + Pt > t\_e$ ***then***
$S \leftarrow \max\{EMP(m), EJP(j), j \in J_O\}, E \leftarrow S + Pt$
if $AS > EMP(m)$ ***then***
$[t\_s, t\_e]_m \leftarrow [t\_s, t\_e]_m \cup [EMP(m), AS]$

(*continued on next column*)

Algorithm 1 Active schedule-based decoding (*continued*)

end if
$EMP(m) \leftarrow E, EJP(j) \leftarrow E, \forall j \in J_O, SPlan \leftarrow SPlan \cup [O, m, S, E]$
end if
$i \leftarrow i + 1$
end while
$C_{max} \leftarrow \max\{EMP\}$
return(output) $SPlan, C_{max}$

4.3. Hybrid initialization strategy.

During the initialization of the population, it is essential to generate individuals that satisfy all constraints, thereby creating feasible initial populations. Simultaneously, maximizing diversity and high-quality within the population is essential. To achieve this, a hybrid initialization strategy is proposed for population generation.

In this strategy, *JPlan* and corresponding *TkPlan* is generated using a classification random strategy, incorporating specific conditions tailored for non PBPO task, MPBPO and FPBPO. The generation of the *JPlan and TkPlan* ensures that all job predecessors corresponding to the task have been placed in the encoding sequence. It also ensures that MPBPO must undergo compulsory parallel operations on the same machine, whereas operations from FPBPO can be randomly combined into one or more PBPOs or processed individually. The assignment of machine is designed using the hybrid random and greedy strategy. The population hybrid initialization strategy is outlined in Algorithm 2. The classification random strategy for the generation of *JPlan and TkPlan* and the hybrid random and greedy strategy for machine assignment is given in Algorithm 3 and Algorithm 4 respectively. Fig. 4 represents an instance of the generated chromosome, and it can be observed that all constraints have been satisfied. The FPBPO formed by $O_{2,2}$, $O_{3,3}$, $O_{4,5}$ and $O_{5,5}$ have been decomposed into two PBPOs for processing. In which the first one consists of $O_{2,2}$, $O_{3,3}$, and the second one comprises $O_{4,5}$ and $O_{5,5}$.

Algorithm 2 Hybrid initialization strategy

Input: *PPlan// The process plans of all the N jobs*
$JPlan = \emptyset, TkPlan = \emptyset, MPlan = \emptyset, TmPlan = \emptyset$, *chromosome* $Chr = \emptyset$, *population* $P_0 = \emptyset$, *population size PSize*, $np = 0$
while $np \leq PSize$ ***do***
$JPlan, TkPlan \leftarrow$ *Classification random strategy*$(PPlan)$
$MPlan, TmPlan \leftarrow$ *Hybrid random and greedy strategy*$(PPlan, TkPlan)$
$Chr \leftarrow \{JPlan, TkPlan, MPlan, TmPlan\}, P_0 \leftarrow P_0 \cup Chr, Chr \leftarrow \emptyset$
end while
return(output) P_0

Algorithm 3. Classification random strategy

Input: *PPlan//* The process plans of all the *N* jobs
Determine the ordered operations $OPSet_j$ of job j and the alternative machines $aM_{j,o}$ for $O_{j,o}, 1 \leq j \leq N, O_{j,o} \in OPSet_j$
Determine the weight $w^m_{j,o}$ of $O_{j,o}$ on machine m and the capacity Cap^m of machine m
Determine the non PBPO tasks *NOPSet*, the MPBPO set *MBOSet* and FPBPO set *FBOSet*

(*continued on next page*)

Algorithm 3. Classification random strategy (*continued*)

Input: *PPlan*// The process plans of all the *N* jobs
Determine the job sequence *JPlan* and task sequence *TkPlan*
while $\exists OPSet_j \neq \emptyset, 1 \leq j \leq N$ **do**
$j \leftarrow Random(1,N)$, Select the first operation $O_{j,o}$ from $OPSet_j$
if $O_{j,o} \in NOPSet$ **then**
Append $O_{j,o}$ after *TkPlan*,$OPSet_j = OPSet_j/\{O_{j,o}\}$
end if
if $(O_{j,o} \in MBOSet_{k'}) \wedge (JP[MBOSet_{k'}] \in TkPlan)$ **then**
$//MBOSet_{k'} \in MBOSet$,$JP[MBOSet_{k'}]$ is the immediate job predecessor (s) of $MBOSet_{k'}$
Append $MBOSet_{k'}$ after *TkPlan*,$OPSet_j \leftarrow OPSet_j/\{O_{j,o}\}$
end if
if $O_{j,o} \in FBOSet_k$ **then** $//FBOSet_k \in FBOSet$
Randomly select t operations as $FBPO_k$ from $FBOSet_k$,$1 \leq t \leq |FBOSet_k|$
Select a machine m from aM_{FBPO_k} satisfying $\sum_{O_{j,o} \in FBPO_k} w_{j,o}^m \leq \max\{Cap^m, m \in aM_{FBPO_k}\}$
if $(O_{j,o} \in FBPO_k) \wedge (JP[FBPO_k] \in TkPlan) \wedge (m \neq \emptyset)$ **then**
Append $FBPO_k$ after *TkPlan*,$OPSet_j \leftarrow OPSet_j/\{O_{j,o}\}$
end if
end if
end while
Determine the job sequence *JPlan* of *TkPlan* (with padded 0 s)
return(output) *JPlan*, *TkPlan*// Return the job sequence and task sequence

Algorithm 4 Hybrid random and greedy strategy

Input: *PPlan*, *TkPlan*// *The process plans of all the N jobs and the task sequence*
MPlan $= \emptyset$,*TmPlan* $= \emptyset$, *load of the M machines load* $= zeros(|M|)$,$i = 1$
while $i \leq length(TkPlan)$ ***do***
Obtain the alternative machines aM_O *for task O and the processing time* $Pt_{j,o}^m$ *of* $O_{j,o}$, $O_{j,o} \in O$ *on* $m, m \in aM_O$
if $Rand(0,1) < 0.5$
Randomly select a machine m from aM_O
end if
if $Rand(0,1) \geq 0.5$
$m \leftarrow \text{argmin}\{load(m) + \max(Pt_{j,o}^m), m \in aM_O, O_{j,o} \in O\}$
end if
Append m after MPlan, append $\max\{Pt_{j,o}^m\}, \forall O_{j,o} \in O$ *after TmPlan*
$load(m) \leftarrow load(m) + TmPlan(i)$
$i = i + 1$
end while
return(output) *MPlan*, *TmPlan*// *Return the machine sequence and processing time sequence*

4.4. Roulette wheel selection

During the iteration process, genetic operations, including selection, crossover, mutation, and elite preservation, are conducted. The selection operation aims to choose high-quality individuals from the population for subsequent crossover and mutation. Roulette wheel selection method is adopted as the selection operator in this study. Roulette wheel selection is a probability-based method in which the selection probability of each individual is associated with its fitness. Individuals with higher fitness are more likely to be chosen, but those with lower fitness still have a certain probability of being selected. First, the reciprocal of the makespan is calculated, which will serve as the fitness function f. Then, probability of each individual and corresponding cumulative probabilities are calculated for a population of size *PSize* by $PR_i = f_i / \sum f_i, 1 \leq i \leq PSize, Q_q = \sum_{i=1}^{q} PR_i, 1 \leq q \leq PSize$ respectively. Finally, a random number r is generated between (0,1), and when $r \leq Q_1$, the first individual is selected; when $Q_{q-1} \leq r \leq Q_q$, the individual q is selected. This process is repeated until *PSize* individuals are selected to construct the new population *Pop1*.

4.5. Split-One by one order crossover

Crossover of GA is designed to explore the potential solution space, promote diversity, and systematically improve individual performance. The global search capability of GA depends on the effectiveness of the crossover operator, which significantly impacts its overall performance. The commonly used crossover operators include JOX (Job-based Order Crossover), SXX (Subsequence Exchange Crossover), PPX (Precedence Preservation Crossover), SPX (Set-Partition Crossover), and POX (Precedence Operation Crossover), and so forth (Gong et al., 2019). Due to the involvement of multi job predecessors and successors for a PBPO task, the position movement of each task needs to consider the coupled effects of the jobs associated with PBPO. Directly applying existing crossover operators to GFJSP_PBPM will inevitably result in infeasible solutions. Therefore, a novel operator named Split-OOOX is designed, as outlined in Algorithm 5. This operator ensures that the crossover not only generates feasible solutions but also guarantees that the newly generated solutions exhibit diversity.

Algorithm 5 Split-OOOX

Input: *Chr*1 $= \{JPlan1, TkPlan1, MPlan1, TmPlan1\}$,*Chr*2 $= \{JPlan2, TkPlan2, MPlan2, TmPlan2\}$
CJPlan $= \emptyset$,*CMPlan* $= \emptyset$,*CTmPlan* $= \emptyset$
$\{EJPlan2, ETkPlan2, EMPlan2, ETmPlan2\} \leftarrow Split(JPlan2, TkPlan2, MPlan2, TmPlan2)$
while *JPlan*1 $\neq \emptyset$ ***do***
if $Rand(0,1) < 0.5$ ***then***
*Obtain the job IDs JobIDs, task Ops, machine MID and the processing time Tm from the first gene of Chr*1
Append JobIDs,MID and Tm after CJPlan,CMPlan and CTmPlan respectively
*Chr*1 *deletes the first gene*
*Find the positions Pos of Ops in ETkPlan*2
*Chr*2 *deletes their genes at the position Pos*
end if
if $Rand(0,1) \geq 0.5$ ***then***
*Obtain the job IDs JobIDs, task Op, machine MID and the processing time Tm from the first gene of Chr*2
if *Op* $\notin$ *PBPOs* in *TkPlan*1 ***then***
Append JobID,MID and Tm after CJPlan,CMPlan and CTmPlan respectively
*Chr*2 *deletes their first genes.*
*Find the positions Pos of Op in TkPlan*1
*Chr*1 *deletes their genes at the position Pos*
end if
end if
end while
return(output) *CJPlan,CMPlan,CTmPlan//Return an offspring*

Algorithm 6. Split

Input: $Chr = \{JPlan, TkPlan, MPlan, TmPlan\}$
$EJPlan = \emptyset, ETkPlan = \emptyset, EMPlan = \emptyset, ETmPlan = \emptyset$
Find the position *Pos* of the first PBPO in *TkPlan*
while *JPlan* $\neq \emptyset$ **do**
Obtain the job IDs *JobIDs*, task *Ops*, machine *MID* and the processing time *Tm* from the first gene of *Chr*
if $|JobIDs| = 1$ **then**
Append *JobIDs*,*OPs*[*i*],*MID* and *Tm* after *EJPlan*,*ETkPlan*,*CTmPlan* and *ETmPlan* respectively
end if
if $|JobIDs| > 1$ **then**
$i = 1$
while $i \leq |JobIDs|$ **do**
Append *JobIDs*[*i*],*OPs*[*i*] and *MID* after *EJPlan*,*ETkPlan* and *EMPlan* respectively
$O_{j,o} \leftarrow OPs[i], m \leftarrow MID$
Determine the processing time $Pt_{j,o}^m$ of $O_{j,o}$ on machine *m*, and append $Pt_{j,o}^m$ after *ETmPlan*
$i = i + 1$
end while
end if
Chr deletes the first gene respectively
end while
return(output) $EChr = \{EJPlan, ETkPlan, EMPlan, ETmPlan\}$

Fig. 5 demonstrates the procedure of Split-OOOX. Initially, parent *P2* {*JPlan2*, *MPlan2*, *TmPlan2*} undergoes splitting according to Algorithm 6, yielding EP2 {*EJPlan2*, *EMPlan2* and *ETmPlan2*}. Subsequently, the OOOX crossover is executed for *P1* {*JPlan1*, *MPlan1*, *TmPlan1*} and EP2.

During this process, the first gene with job ID 5 in *JPlan1* corresponding to $O_{5,1}$ in *the TkPlan (*omitted from the figure) is randomly selected and appended to *CJPlan*. The machine and processing time for this task are selected from *MPlan1* and *TmPlan1*, respectively, and appended to *CMPlan* and *CTmPlan*. Following this, the first gene in *JPlan1*, *MPlan1*, and *TmPlan1* are deleted, as well as those corresponding to task $O_{5,1}$ in *EJPlan2*, *EMPlan2*, and *ETmPlan2*. The dashed line in the figure indicates the deleted genes, while red labeling highlights the genes scheduled for deletion in the current step. This iterative process continues when the randomly selected gene is from *JPlan1*. However, if the randomly selected gene is from *EJPlan2* and the operation corresponding to the gene belongs to a PBPO in the *TkPlan1*, then selection is limited to *JPlan1*. For example, the first gene with job ID 2 in *EJPlan2* corresponds to operation $O_{2,2}$, which belongs to PBPO ($O_{2,2}$, $O_{3,3}$) in P1′s *TkPlan1*. Hence, to preserve PBPO integrity in *P1*, the first gene with job ID 4 in *JPlan1* is reselected. This process repeats until all genes in *P1* and *EP2* are deleted, ultimately resulting in the generation of offspring *CP* {*CJPlan*, *CMPlan*, *CTmPlan*}. Based on *CJPlan*, the corresponding task sequence *CTkPlan* can be determined, and thus the information of final CP is {*CJPlan*, *CTkPlan*, *CMPlan*, *CTmPlan*}.

4.6. Hybrid mutation operator

Mutation is achieved by introducing random changes in the genome to enhance the diversity of the population, thereby mitigating the risk of premature convergence in the algorithm, and strengthening the GA's local search capability (Sun,et al., 2023). Common mutation operators include single-point mutation, multiple-point mutation, swap mutation, insert mutation, inversion mutation, and so on (Wu and Sun, 2018). However, directly applying existing mutation operators by randomly changing the positions of job IDs without considering the coupled effects of PBPO will inevitably yield infeasible solutions for GFJSP_PBPM. Therefore, a hybrid mutation operator is developed to generate diverse feasible individuals. In this approach, *JPlan* undergoes segment swap mutation, while the *MPlan* experiences single-point mutation. The overall mutation process is outlined in Algorithm 7.

Algorithm 7 Hybrid mutation operator

1:	***Input***:$Chr = \{JPlan, TkPlan, MPlan, TmPlan\}$
2:	$newJPlan = \emptyset, newTkPlan = \emptyset, newMPlan = \emptyset, newTmPlan = \emptyset$
3:	*Using PBPO to partition the JPlan into different segments, and storing these segments in subJPlans*
4:	*Randomly select a segment subJPlan from subJPlans and two positions* $[l_1, l_2]$ *from subJPlan*
5:	*Swap the two positions of JPlan, MPlan, TmPlan to generate newJPlan, newMPlan, newTmPlan respectively*
6:	*Determine the new task sequence newTkPlan corresponding to newJPlan*
7:	*Randomly select a position mp in newMPlan*
8:	$O \leftarrow newTkPlan(mp)$
9:	*Randomly select a machine m from its alternative machines*,$newMPlan(mp) \leftarrow m$
10:	$newTmPlan(mp) \leftarrow \max\{Pt^m_{j,o}\}, \forall O_{j,o} \in O // Pt^m_{j,o}$ *is the processing time of* $O_{j,o}$
11:	***return(output)***$newChr = \{newJPlan, newTkPlan, newMPlan, newTmPlan\}$

Fig. 6 provides an instance of the hybrid mutation. In this instance, the JPlan consists of three PBPOs, and during the mutation for JPlan, the chromosome is divided into three segments corresponding to these PBPOs. Assuming a random interval of r = 3 is generated, the positions for mutation can only be chosen from the third segment. Subsequently,

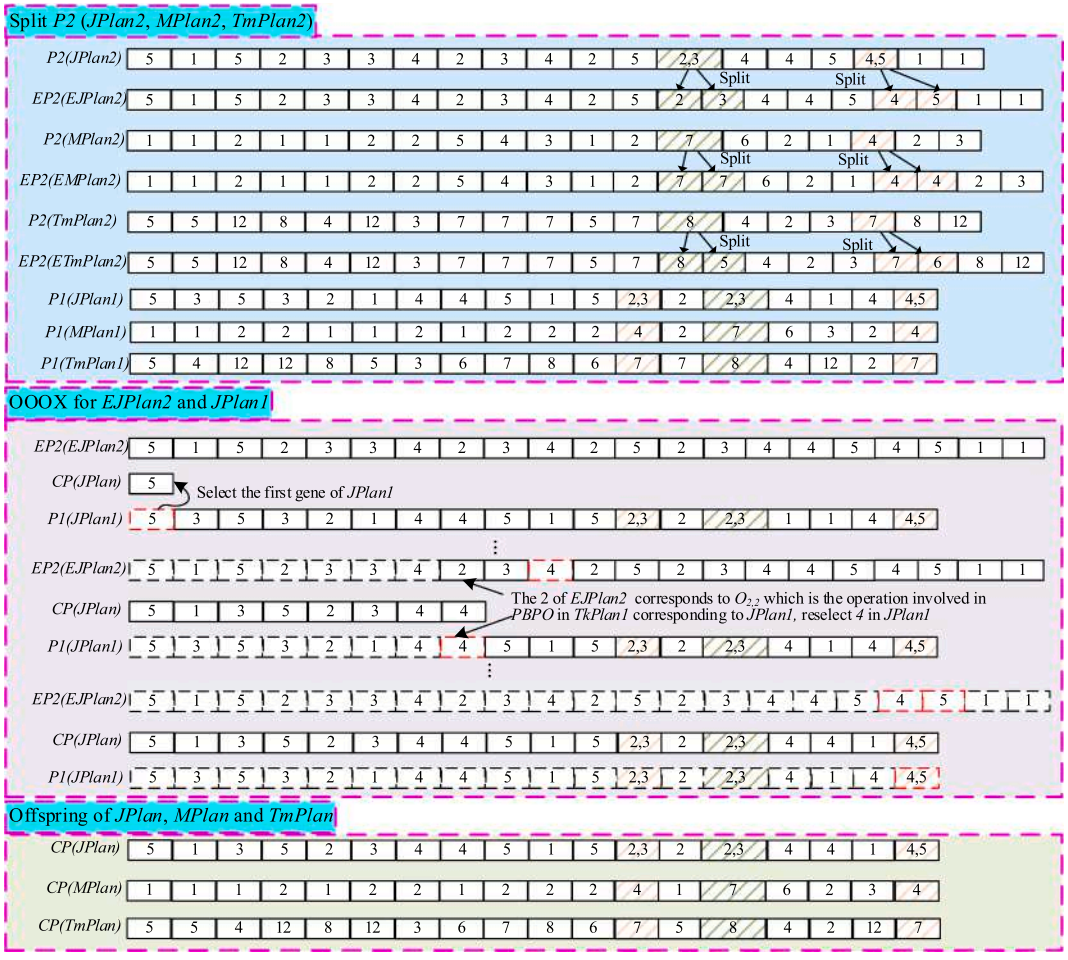


Fig. 5. Split-OOOX operator.

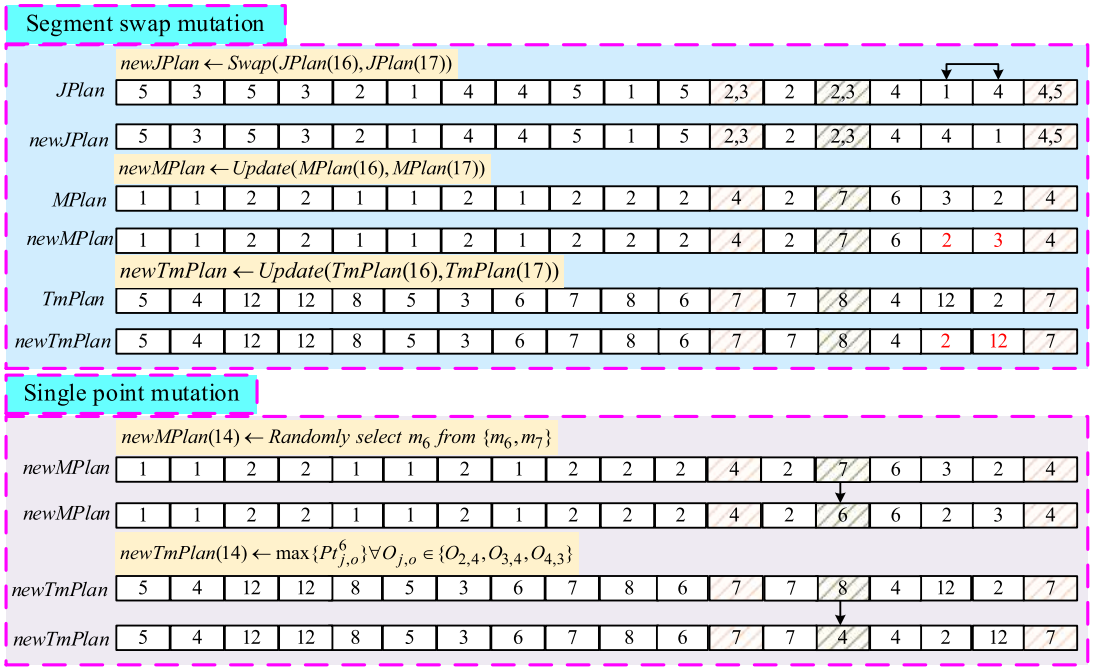


Fig. 6. Hybrid mutation operator.

two positions 16, 17 are randomly selected within this segment, and the two genes in the *JPlan*,*MPlan* and *TmPlan* are swapped to generate the *NewJPlanNewMPlan* and *NewTmPlan*. Based on the *NewJPlan*, the corresponding task sequence newTkPlan can be determined. For machine mutation, let 14 be the randomly selected position. The corresponding task ($O_{2,4}$, $O_{3,4}$) can be determined based on *NewTmPlan*(14) (omitted from the figure). Then, a machine m_6 is randomly selected from its alternative machines {m_6, m_7}, and *NewMPlan*(14) is updated as 6. The processing time is updated by *newTmPlan*(14)←max{$Pt^6_{j,o}$}$\forall O_{j,o} \in \{O_{2,4}, O_{3,4}\}$ accordingly.

4.7. Hybrid N5/7 neighborhood structure

Balas (1969) asserted that, for any feasible solution, altering the processing sequence of adjacent critical operations will not lead to an infeasible solution. This implies that the makespan can be enhanced by reducing the length of the critical path. The critical path is defined as the longest path without a time interval between operations in a feasible schedule, the duration of this path represents the makespan of the scheduling scheme. Operations positioned on the critical path are called critical operations. When two or more neighboring critical operations are processed on the same machine, they collectively constitute a critical block. In this study, tasks, whether operations or PBPOs, constitute elements of the critical path and critical block. Therefore, tasks positioned on the critical path are referred to as critical tasks. The algorithm for identifying critical tasks (operations or PBPOs) within a chromosome is outlined in Algorithm 8.

Algorithm 8 Find critical tasks

Input: $Chr = \{JPlan, TkPlan, MPlan, TmPlan\}$
$COPSet = \varnothing$
$SPlan \leftarrow$ *Active schedule-based decoding* (Chr)
Find a task O with maximum completion time according to SPlan
Find the start time S_O *of task O*
$COPSet \leftarrow O$
while $S_O > 0$ ***do***
Determine the completion time EMP[*O*] *of the immediate machine predecessor MP*[*O*] *of O*

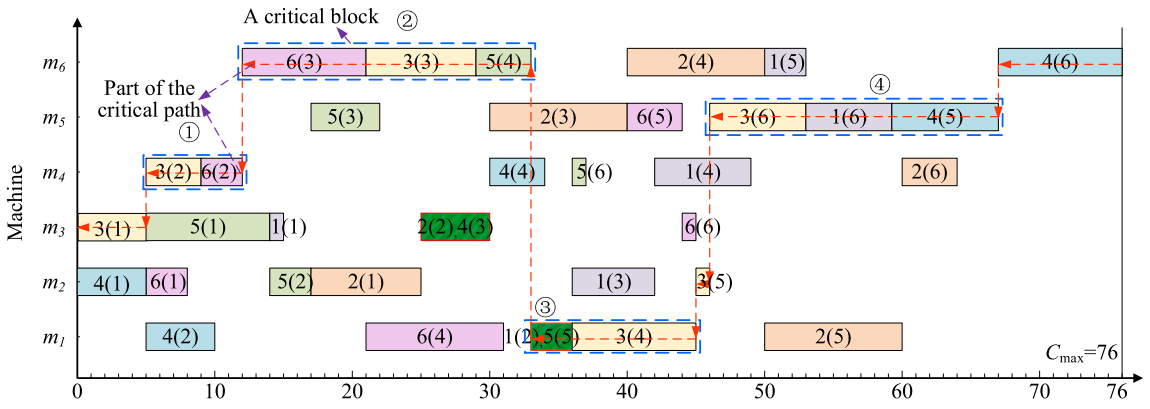
(*continued on next column*)

Algorithm 8 Find critical tasks (*continued*)

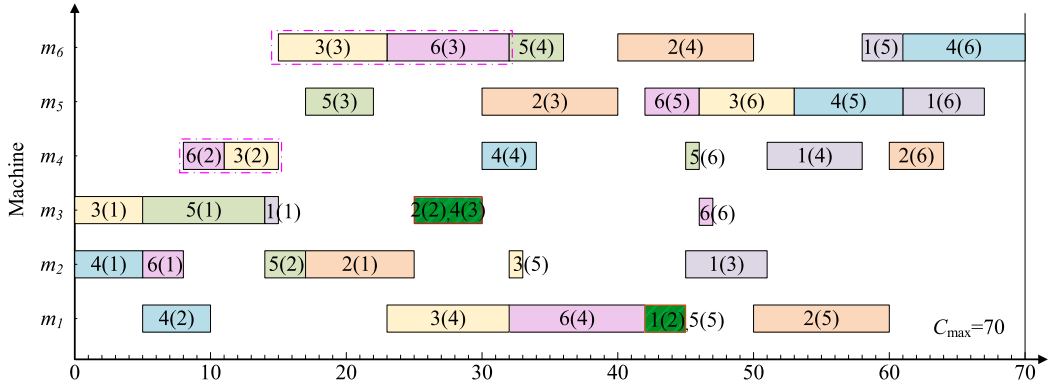
Input: $Chr = \{JPlan, TkPlan, MPlan, TmPlan\}$
Determine the completion time EJP[*O*] *of the immediate job predecessor*(*s*)*JP*[*O*] *of O*
if $EMP[O] = S_O$ ***then***
$COPSet \leftarrow COPSet \cup MP[O], O \leftarrow MP[O]$
end if
if $EMP[O] \neq S_O$ ***then***
$O_{j,o} \leftarrow \text{argmax}\{EJP[O], O_{j,o} \in JP[O]\}, COPSet \leftarrow COPSet \cup O_{j,o}, O \leftarrow O_{j,o}$
end if
Find the start time S_O *of task O*
end while
return(output) $COPSet$

Researchers have explored and innovated a series of neighborhood structures denoted as N1 to N8 (Zhang et al., 2007; Xie et al., 2023b). These structures can be classified into two primary categories: insertion-type and exchange- type. The N5 and N7 are respectively the most important and famous neighborhood structures of exchange-type and insertion-type. The N5 neighborhood structure, initially introduced by Nowicki and Smutnicki (1996), involves reversing the order of the two front or rear operations within the critical block. The N7 neighborhood structures, first proposed by Zhang et al. (2007), comprise backward insertion and forward insertion (Zhang et al., 2007). A backward insertion involves relocating the last operation of the critical block either within the block or inserting an inner operation at the beginning of the block. Conversely, a forward insertion involves moving the first operation of the critical block either within the block or inserting an inner operation at the end of the block. Many algorithms have achieved good results for solving (F)JSP using either N5 (Tang et al., 2019; Abedi et al., 2020) or N7 (Caldeira and Gnanavelbabu, 2019).

When the problem size is small and there are few operations on a critical block (e.g., fewer than 4), the reversing operation of N5 is simple and efficient. Conversely, when the problem size is large and there are many operations on a critical block (e.g., more than 4), the N7 neighborhood structure leads to a considerably larger neighborhood, exploring a much wider space. However, N5 and N7 primarily concentrate on only one randomly selected critical block to generate their neighborhoods. This limited disturbance scale impedes their capacity to fully utilize neighborhoods to solve large-scale problems. Moving the



(a) Gantt chart of a scheduling scheme with critical path (makespan=76)



(b) Gantt chart of the new scheduling scheme after exchanging $O_{3,2}$ and $O_{6,2}$, and $O_{3,3}$ and $O_{6,3}$ (makespan=70)

Fig. 7. Schematic diagram of the movement of critical operations.

positions of multiple tasks of multiple blocks may be more advantageous for fully utilizing the idle time intervals of machines involved in the critical path. Taking Fig. 7 as an example, if only $O_{3,2}$ and $O_{6,2}$ of Block ① are exchanged by N5, the makespan of the neighborhood is 76. If only $O_{3,3}$ and $O_{6,3}$ of Block ② are exchanged either by N5 or N7, the makespan of corresponding neighborhoods become 78. By sequentially exchanging $O_{3,2}$ and $O_{6,2}$ of Block ① and $O_{3,3}$ and $O_{6,3}$ of Block ②, the corresponding makespan decreases from 76 in Fig. 7(a) to 70 in Fig. 7 (b).

To better accommodate problems of different scales, leverage the advantages of N5 and N7, expand their high-quality neighborhoods space, and address constraints associated with both FJSP and PBPO, we propose a hybrid N5/7 neighborhood structure. The specific details are illustrated in Algorithm 9 below. To minimize ineffective moves that cannot reduce the total makespan, neither the first task of the first block nor the last task of the last block on the critical path are moved in both N5 and N7.

Algorithm 9 Hybrid N5/7

Input: scale factor:λ, population Pop
$P_{Nsm} \leftarrow \lceil \lambda \times |Pop| \rceil$ chromosomes in Pop that rank high in fitness,$P_N = \varnothing$
while $i < size(P_{Nsm}, 1)$ **do**
$j = 1$
$COPSet \leftarrow$ Find critical tasks $(P_{Nsm}(i))$
Determine the critical blocks CBSet containing more than 1 tasks in COPSet
$\xi = \lceil length(CBSet)/4 \rceil + 1$ // More than one critical block will be selected for neighborhood operation
while $j \le \xi$ **do**

(continued on next column)

Algorithm 9 Hybrid N5/7 (*continued*)

Input: scale factor:λ, population Pop
Randomly select a block CBlock from CBSet,$ls \leftarrow length(CBlock)$
Randomly select the first or the last two tasks O_1, O_2 of CBlock
$O_1 \leftarrow CBlock[1]$ or $CBlock[ls]$ and randomly select a task as O_2 from $CBlock[2], \ldots CBlock[ls-1]$ if $ls > 4$
Determine the job set J_1, J_2 related to O_1,O_2 respectively
Determine the tasks subOSet between O_1 and O_2
Determine the PBPOs subBPSet in $\{O_1 \cup subOSet \cup O_2\}$ and job set J_3 related to these PBPOs
Determine the start time S_{O_1} of task O_1 and the start time S_{O_2} of task O_2
if $(subBPSet \neq \varnothing) \wedge (J_1 \cup J_2 \cap J_3 = \varnothing)$ or $(subBPSet = \varnothing) \wedge (J_1 \cap J_2 = \varnothing)$ or $(J_1 \cap J_2 = \varnothing) \wedge (J_1 \cup J_2 \cap J_3 = \varnothing)$ **then**
if $(ls \ge 2) \wedge (ls \le 4) \wedge (S_{O_1} \neq 0) \wedge (S_{O_2} \neq C_{max})$ **then** // Not the first block and the last block
$P_{Nsm}(i) \leftarrow$ N5 Reverse$(P_{Nsm}(i), O_1, O_2)$// Reversing O_1, O_2 based on N5
end if
if $(ls > 4) \wedge (S_{O_1} \neq 0) \wedge (S_{O_2} \neq C_{max}) \wedge (S_{O_2} \neq 0) \wedge (S_{O_1} \neq C_{max})$ **then**
$P_{Nsm}(i) \leftarrow$ N7 Insertion $(P_{Nsm}(i), O_1, O_2)$// Forward or backward insertion O_1 or O_2 based on N7
end if
end if
$CBSet \leftarrow CBSet/\{CBlock\}, j \leftarrow j + 1$
end while
$P_N \leftarrow P_{Nsm}(i) \cup P_N, i \leftarrow i + 1$
end while
return(output) P_N

Fig. 8 (a) illustrates the hybrid N5/7 neighborhood structure based on the Gantt chart, with arrows highlighting a critical path. During the process of identifying critical tasks, the $O_{4,4}$ is determined as the predecessor critical task of PBPO ($O_{4,5}$, $O_{5,5}$) based on the maximum

completion time of $JP_{4,5}, JP_{5,5}$, while the $O_{5,4}$ can be established as the predecessor critical task of PBPO ($O_{2,3}$, $O_{3,4}$, $O_{4,3}$), as there is no time interval between $O_{5,4}$ and ($O_{2,3}$, $O_{3,4}$, $O_{4,3}$) on the same machine m_2. According to Algorithm 9, it is evident that only Block ① is eligible for neighborhood operation, and it can undergo N7 insertion movement since it comprises 6 tasks within this block. Assuming a N7 backward insertion involves the two critical tasks ($O_{2,3}$, $O_{3,4}$, $O_{4,3}$) and $O_{5,3}$, the job sets corresponding to the two tasks are labeled as $J_1 = \{2, 3, 4\}$ and $J_2 = \{5\}$ respectively. The job set linked to the tasks positioned between $O_{5,3}$ and ($O_{2,3}$, $O_{3,4}$, $O_{4,3}$) in the TkPlan fulfills the condition ($subBPOSet \neq \varnothing$) $\wedge(J_1 \cup J_2 \cap J_3 = \varnothing)$ in Algorithm 9, where *subBPOSet* represents the PBPO ($O_{2,3}$, $O_{3,4}$, $O_{4,3}$). Consequently, N7 backward insertion can be executed. The scheduling scheme following the insertion by moving PBPO ($O_{2,3}$, $O_{3,4}$, $O_{4,3}$) before $O_{5,3}$ is shown in Fig. 8(b).

4.8. Elite retention strategy

The makespan individuals in the Pop3 are calculated using active schedule-based decoding. Then, Pop3, the selected populations (Pop1), and the populations after genetic operations (Pop2) are merged to create a new population. An elite retention strategy is implemented to preserve individuals with outstanding genes in the merged population, whereby the top preset *PSize* chromosomes with the highest fitness values are chosen as the initial population for the subsequent generation.

4.9. Pseudocode of GANS for GFJSP_PBPM

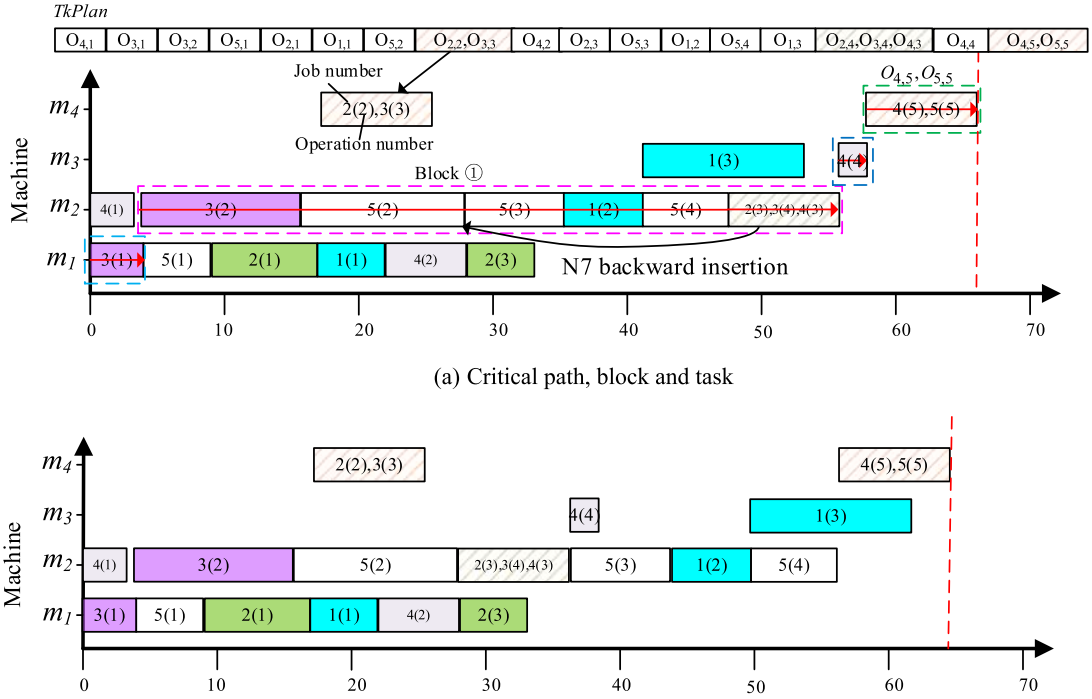
Based on the customized task-based encoding and active schedule-based decoding, hybrid initialization strategy, roulette wheel selection, Split-OOOX crossover, hybrid mutation, Hybrid N5/7, and elite retention strategy, the pseudo-code of the GANS for the GFJSP_PBPO is outlined in Algorithm 10. In this algorithm, the termination condition is set as the maximum iterations. Correspondingly, to use the preset time (*PT*) as the termination condition, only the condition judgment in the while loop needs to be changed to the elapsed time $t \leq PT$.

Algorithm 10 GANS for GFJSP_PBPM

Input *PPlan // The flexible process plans of all the N jobs*
Set the population size PSize, crossover rate Pc, mutation rate Pm, max iterations maxT, scale factor λ,$k = 0$
Pop←Hybrid initialization strategy (*PPlan*)*//Generate PSize chromosomes as initial population using* Algorithm 2
$C_{\max,i}$←Active schedule - based decoding $(Chr_i), Chr_i \in Pop, 1 \leq i \leq PSize$*// Decode each chromosome according to* Algorithm 1
while $k \leq \max T$ ***do***
Pop1 ← *Roulette wheel selection* (*Pop*) *//According to description in Section 4.4 and the decoded* C_{max} *for each Chr*
$Pop2 = \varnothing, Pop3 = \varnothing$
for $j = 1{:}2{:}PSize\text{-}1$
if $Rand(0,1) \leq Pc$ ***then*** *// Perform crossover operation according to* Algorithm 5
$newChr_1, newChr_2$←Split - OOOX $(Chr_j, Chr_{j+1}), Chr_j, Chr_{j+1} \in Pop1$
$Pop2 \leftarrow Pop2 \cup \{newChr_1, newChr_2\}$
end if
end for
for $r = 1{:}PSize$
if $Rand(0,1) \leq Pm$ ***then*** *// Perform mutation according to* Algorithm 7
$newChr_r$←Hybird mutation$(Chr_r), Chr_r \in Pop2$
$Pop2(r) \leftarrow newChr_r$
end if
end for
Pop3←*Hybrid* N5/7 $(\lambda, Pop2)$*// Perform neighborhood search according to* Algorithm 9
Pop←Elite rentention $(Pop1, Pop2, Pop3)$*// According to description in Section 4.8*
$k = k + 1$
end while
output *the best Chr and its objective function value* C_{max}

5. Computational experiments and case study

In this section, experiments are conducted to validate the effectiveness of the constructed model, hybrid initialization strategy, and hybrid N5/7 neighborhood structure. Comparative analyses are performed with GRASP (Boyer et al., 2021; Knopp et al., 2017) and VND-hGA (Liu et al.,



(a) Critical path, block and task

(b) Gantt chart of the new scheduling scheme after inserting PBPO ($O_{2,3}$, $O_{3,4}$, $O_{4,3}$) before $O_{5,3}$

Fig. 8. Instance of hybrid N5/7 neighborhood structure.

2021) to verify the superiority of GANS. Additionally, a case study is carried out to verify its practical application. To guarantee the stability and reliability of the results and minimize the impact of randomness, we execute each test instance and the engineering case ten times for each algorithm. The GANS, GRASP, and VND-hGA algorithms are implemented using MATLAB R2019a. The experiments are run on a laptop equipped with an AMD Ryzen 5–5600 h processor and 16 GB of RAM, running the Windows 10 operating system.

5.1. Instance generation

Due to the absence of relevant benchmarks for the GFJSP_PBPM study, test instances are generated based on the BRdata (Brandimarte, 1993) and Fdata (Bagheri et al., 2010). This is achieved by randomly selecting different job operations to create 1 to 3 MPBPOs and 1 to 3 FPBPOs, resulting in instances tailored for the GFJSP_PBPM in this study. Each MPBPO comprises 2 to 4 operations, and each FPBPO consists of 2 to 5 operations. The BRdata includes 15 problems: mk01 ~ mk15. Details of the newly generated benchmark, GBRdata, based on BRdata, are presented in Table 1 where the problems are renamed as gmk01 ~ gmk15. The Fdata consists of 20 problems categorized into two classes: 10 small-sized problems (sfjs1-sfjs10) and 10 medium and large-sized problems (mfjs1-mfjs10). Details of the newly generated benchmarks, GFdata, based on Fdata, are provided in Table 2 where the problems are renamed as gsfjs1-gsfjs10 for small-sized problems and gmfjs1-gmfjs10 for medium and large-sized problems. The capacity of each machine is set to 10 units, and the weight of each operation, representing the number of units for a job, ranges from 2 to 5.

5.2. Parameter setting and notations

The performance of an algorithm is affected by its parameters. The parameters selected for the GANS are determined through empirical experimentation to achieve satisfactory results within a reasonable amount of time. The specific parameter settings are as follows: population size (*Psize*) is 500; maximum number of iterations is 300; scale factor for neighborhood structure is 0.3. The crossover rate (P_c) and mutation rate (P_m) are set at 0.8 and 0.3 respectively.

The notations used as performance evaluation metrics in this section are as follows:

$B(C_{max})$: best makespan out of ten runs.

$Av(C_{max})$: average makespan across ten runs.

$Sd(C_{max})$: standard deviation of C_{max} across ten runs.

$B(Cov)$:fewest iterations until the algorithm converges to the best makespan out of ten runs.

$Av(Cov)$: average convergence (iterations) of the ten runs.

Table 1
Description of the instances in GBRdata.

Instance	$n \times m$	Os	m/p	MPBPOs	FPBPOs
gmk01	10 × 6	55	2	1	1
gmk02	10 × 6	58	3.5	1	2
gmk03	15 × 8	150	3	2	2
gmk04	15 × 8	90	2	2	1
gmk05	15 × 4	106	1.5	2	2
gmk06	10 × 15	150	3	2	2
gmk07	20 × 5	100	3	1	2
gmk08	20 × 10	225	1.5	3	2
gmk09	20 × 10	240	3	2	2
gmk10	20 × 15	240	3	3	2
gmk11	30 × 5	179	1.5	2	2
gmk12	30 × 10	193	1.5	2	2
gmk13	30 × 10	231	3	2	3
gmk14	30 × 15	277	1.5	1	2
gmk15	30 × 15	284	3	2	3

Os: total operations; *m/p*: average number of machines per operation; MPBPOs: number of MPBPO; FPBPOs: number of FPBPO.

Table 2
Description of the instances in GFdata.

Instance	$n \times m$	Os	m/p	MPBPOs	FPBPOs
gsfjs1	2 × 2	4	2	1	1
gsfjs2	2 × 2	4	1.5	1	1
gsfjs3	3 × 2	6	1.6	1	1
gsfjs4	3 × 2	6	1.6	1	1
gsfjs5	3 × 2	6	2	1	1
gsfjs6	3 × 3	9	1.6	1	1
gsfjs7	3 × 5	9	2	1	1
gsfjs8	3 × 4	9	2	1	1
gsfjs9	3 × 3	9	2	1	1
gsfjs10	3 × 5	12	1.6	2	1
gmfjs1	5 × 6	15	2	1	1
gmfjs2	5 × 7	15	2.6	1	1
gmfjs3	6 × 7	18	2.6	1	2
gmfjs4	7 × 7	21	2.6	2	1
gmfjs5	7 × 7	21	2.6	1	2
gmfjs6	8 × 7	24	2.5	1	2
gmfjs7	8 × 7	32	2.4	2	2
gmfjs8	9 × 8	36	2.3	2	1
gmfjs9	11 × 8	44	2.3	2	2
gmfjs10	12 × 8	48	2.3	2	2

Dev(%):relative deviation between $B(C_{max})$ obtained by compared algorithm *A* and our GANS, given by

$$Dev = \frac{B(C_{\max})ofA - B(C_{\max})ofGANS}{B(C_{\max})ofA} \times 100\%$$

5.3. Effectiveness of optimization model

To validate the effectiveness of the developed MIP model for GFJSP_PBPM, IBM ILOG CPLEX 12.10 is employed for solving, with a time limit set at 3600 s. The experimental results indicate that only scheduling solutions for small-scale problems from gsfjs1 to gsfjs10 can be obtained through CPLEX, as detailed in Table 3, which includes their computational times (CPU in seconds). Although CPLEX can only provide feasible solutions for small-scale GFJSP_PBPM, it is verified that the constructed model accurately represents the constraints of GFJSP_PBPM and the feasibility of the model is confirmed. However, as the scale of GFJSP_PBPM increases, CPLEX fails to obtain feasible solutions, indicating that CPLEX is difficult to adapt to the engineering application of GFJSP_PBPM.

5.4. Effectiveness of hybrid initialization strategy

Two experiments are designed to evaluate the effectiveness and performance of the proposed hybrid initialization strategy. In the first experiment, 500 initial individuals are generated for each test problem in GBRdata and GFdata using both the hybrid initialization and random initialization strategy. The quality of the initial population is compared from the perspective of $B(C_{max})$ and $Av(C_{max})$ to validate the enhancement brought by the hybrid initialization strategy. The quality of the initial solution can also be measured by the speed at which the algorithm converges to the optimal solution within a given time. If the initial solution can accelerate the convergence process of the algorithm, then it is considered effective. Therefore, we conduct the second experiment, a standard GA is applied to address the gmk05, and utilizing both the

Table 3
Results achieved by CPLEX.

Instance	$B(C_{max})$	CPUs	Instance	$B(C_{max})$	CPUs
gsfjs1	51	0.20	gsfjs6	440	0.37
gsfjs2	107	0.28	gsfjs7	407	0.45
gsfjs3	208	0.24	gsfjs8	430	0.46
gsfjs4	272	0.32	gsfjs9	494	0.32
gsfjs5	100	0.45	gsfjs10	520	0.47

hybrid and random initialization strategies while keeping other parameters constant. The evolutionary process are documented to examine whether the hybrid initialization strategy expedited the convergence of the GA.

The results of the first experiment confirm that both the deliberately designed random initialization and the hybrid initialization strategies, tailored to the GFJSP_PBPM, are capable of generating feasible scheduling schemes. This ensures that that all operations for each job are arranged in the order specified by their process plans. Additionally, it guarantees that MPBPOs undergo compulsory parallel operations on the same machine, while operations from FPBPOs can be randomly grouped into one or more PBPOs or processed individually. The statistical results are outlined in Table 4. It can be seen that both random initialization and hybrid initialization strategies can achieve the same initial $B(C_{max})$ and $Av(C_{max})$ for small-sized instances (gsfjs1-gsfjs10). As the size of the instances increases, the hybrid initialization strategy demonstrates significant advantages. For the 25 small-medium (gmfjs1-gmfjs10) and medium-large (gmk01 ~ gmk15) scale instances, both $B(C_{max})$ and $Av(C_{max})$ obtained through hybrid initialization are superior to those obtained through random initialization. This indicates that the hybrid initialization method enhances the approximation (smaller gap between the initial solution and the optimal solution) and stability across different scenarios or runs. These findings lead to the following conclusions: both the designed random initialization and hybrid initialization strategies are capable of generating initial feasible scheduling plans in accordance with the constraints of the GFJSP_PBPM. The hybrid initialization strategy outperforms the random strategy in terms of $B(C_{max})$ and $Av(C_{max})$. Clearly, the hybrid initialization strategy proves to be valuable in generating a high-quality initial population.

The results of the second experiment are illustrated in Fig. 9. When addressing the gmk05, the population initialized using the hybrid strategy demonstrates faster and more efficient convergence towards a superior makespan, as compared to the population initialized using the random method.

5.5. Effectiveness of hybrid N5/7

An experiment is also conducted to assess the performance of the developed hybrid N5/7 neighborhood structure using GBRdata and gmfjs1 ~ gmfjs10 from GFdata. To verify the effectiveness of hybrid N5/7, the individuals generated by hybrid initialization strategy are used as the initial individuals for this experiment. Subsequently, the hybrid N5/7 is incorporated into the GA (GANS) for optimization. Additionally, comparative experiments are conducted using GA without employing neighborhood search (GA), GA with the N5 (GAN5), and GA with the N7 (GAN7). Finally, the $B(C_{max})$, $Av(C_{max})$, $Sd(C_{max})$, $B(Cov)$ and $Av(Cov)$ are calculated.

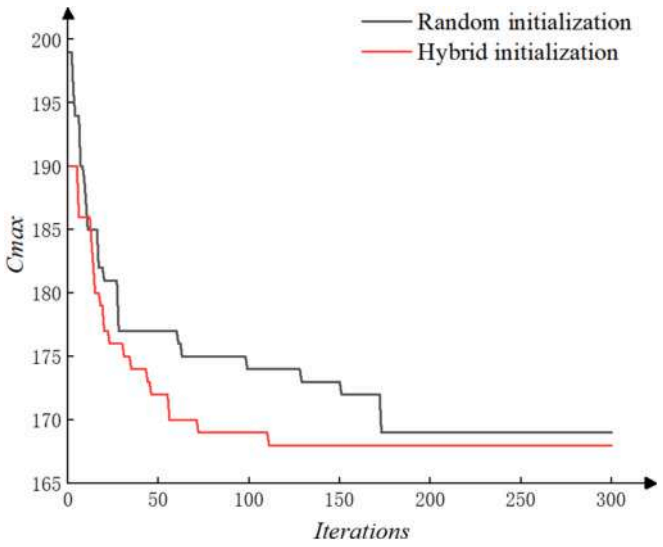


Fig. 9. Convergence illustration of GA with two different initialization strategies for gmk05.

The outcomes of the experiment are outlined in Table 5, Table 6 and Fig. 10. Table 5 demonstrates that none of the $B(C_{max})$ and $Av(C_{max})$ obtained by GANS are inferior to those obtained by GA, GAN5 and GAN7. In particular, GA, GAN5, and GAN7 achieve the same $B(C_{max})$ as GANS in only 11, 14, and 14 cases, respectively. Similarly, GA, GAN5, and GAN7 yield identical $Av(C_{max})$ to GANS in only 3, 8, and 7 instances, respectively. This indicates that the hybrid N5/7 further enhances GA's ability to search for superior makespan, both in terms of best and average values. Table 6 shows that the $Sd(C_{max})$ of C_{max} achieved by GANS is superior to GA in 13 instances and inferior in only 5 instances. Furthermore, the $Sd(C_{max})$ attained by GANS outperforms GAN5 in 15 instances while being worse in only 4 instances. Similarly, compared to GAN7, the $Sd(C_{max})$ achieved by GANS is better in 15 instances but worse in only 7 instances. This suggests that the hybrid N5/7 improves the stability of GA in solving the GFJSP_PBPM addressed in this study.

Fig. 10(a) and Fig. 10(b) illustrate the best convergence $B(Cov)$ and average convergence $Av(Cov)$, respectively, for each instance. The $B(Cov)$ achieved by GANS is superior to GA in 24 instances, with only 1 instance showing inferior performance. Similarly, the $Av(Cov)$ obtained by GANS surpasses GA in 22 instances and falls behind in only 3

Table 4
Comparison of hybrid and random initialization strategies.

Instance	Random initialization		Hybrid initialization		Instance	Random initialization		Hybrid initialization	
	$B(C_{max})$	$Av(C_{max})$	$B(C_{max})$	$Av(C_{max})$		$B(C_{max})$	$Av(C_{max})$	$B(C_{max})$	$Av(C_{max})$
gmk01	52	56.9	**50**	**54.0**	gsfjs4	**272**	**272**	**272**	**272**
gmk02	45	46.5	**31**	**32.3**	gsfjs5	**100**	**100**	**100**	**100**
gmk03	262	273.4	**253**	**265.5**	gsfjs6	**440**	**440**	**440**	**440**
gmk04	84	86.7	**80**	**85.4**	gsfjs7	**407**	**407**	**407**	**407**
gmk05	199	202.9	**190**	**201.5**	gsfjs8	**430**	**430**	**430**	**430**
gmk06	131	137.4	**99**	**99.6**	gsfjs9	**494**	**494**	**494**	**494**
gmk07	217	223.0	**208**	**208.0**	gsfjs10	**520**	**520**	**520**	**520**
gmk08	561	573.3	**553**	**570.1**	gmfjs1	477	529.6	**469**	**509.9**
gmk09	438	448.6	**430**	**444.0**	gmfjs2	477	503.3	**472**	**485.4**
gmk10	362	374.2	**304**	**309.6**	gmfjs3	635	653.3	**625**	**636.4**
gmk11	699	713.7	**696**	**711.5**	gmfjs4	684	714.0	**627**	**712.7**
gmk12	613	633.9	**605**	**631.4**	gmfjs5	685	716.2	**661**	**715.2**
gmk13	635	661.3	**613**	**620.9**	gmfjs6	782	826.6	**766**	**818.0**
gmk14	915	957.1	**905**	**943.5**	gmfjs7	1114	1183.0	**1051**	**1171.3**
gmk15	542	563.3	**529**	**551.7**	gmfjs8	1150	1217.4	**1069**	**1099.8**
gsfjs1	**51**	**51**	**51**	**51**	gmfjs9	1446	1496.6	**1393**	**1433.5**
gsfjs2	**107**	**107**	**107**	**107**	gmfjs10	1653	1760.6	**1567**	**1613.4**
gsfjs3	**208**	**208**	**208**	**208**					

Table 5
$B(C_{max})$ and $Av(C_{max})$ obtained by different method.

Instance	GA		GAN5		GAN7		GANS	
	$B(C_{max})$	$Av(C_{max})$	$B(C_{max})$	$Av(C_{max})$	$B(C_{max})$	$Av(C_{max})$	$B(C_{max})$	$Av(C_{max})$
gmk01	42	42.0	42	42.0	42	42.0	**41**	**41.8**
gmk02	**27**	27.8	**27**	27.2	**27**	27.8	**27**	**27.0**
gmk03	**204**	204.2	**204**	204.2	**204**	204.2	**204**	**204.0**
gmk04	**67**	69.8	**67**	69.6	**67**	**68.4**	**67**	**68.4**
gmk05	168	169.6	168	170.4	168	170.4	**166**	**168.2**
gmk06	**76**	78.6	**76**	79.0	**76**	79.6	**76**	**77.6**
gmk07	**141**	142.2	**141**	144.6	**141**	145.6	**141**	**141.6**
gmk08	**523**	**523.0**	**523**	**523.0**	**523**	**523.0**	**523**	**523.0**
gmk09	331	337.0	331	338.6	330	333.8	**327**	**332.8**
gmk10	255	260.2	**243**	**255.2**	254	261.4	**243**	258.2
gmk11	613	622.9	612	619.8	**608**	614.2	**608**	**613.0**
gmk12	**508**	516.2	**508**	511.2	**508**	511.4	**508**	**509.8**
gmk13	456	469.2	447	457.8	451	460.0	**419**	**448.6**
gmk14	**694**	**694.0**	**694**	**694.0**	**694**	**694.0**	**694**	**694.0**
gmk15	425	428.8	**403**	**423.2**	423	427.8	**403**	**423.2**
gmfjs1	**468**	468.6	**468**	**468.0**	**468**	**468.0**	**468**	**468.0**
gmfjs2	448	448.0	448	448.0	448	448.0	**433**	**434.8**
gmfjs3	**538**	**538.0**	**538**	**538.0**	**538**	**538.0**	**538**	**538.0**
gmfjs4	568	581.8	**553**	577.2	**553**	**562.0**	**553**	**562.0**
gmfjs5	549	560.2	549	561.4	**547**	556.2	**547**	**548.2**
gmfjs6	643	653.6	643	656.4	643	654.6	**634**	**643.0**
gmfjs7	879	895.0	879	**879.0**	879	**879.0**	**875**	891.0
gmfjs8	889	927.4	889	905.2	889	907.8	**874**	**897.2**
gmfjs9	**1055**	1120.3	**1055**	**1092.4**	**1055**	1109.0	**1055**	1108.5
gmfjs10	1282	1308.8	1224	1264.4	1234	1277.0	**1198**	**1245.4**

Table 6
Standard deviation of C_{max} obtained by different method.

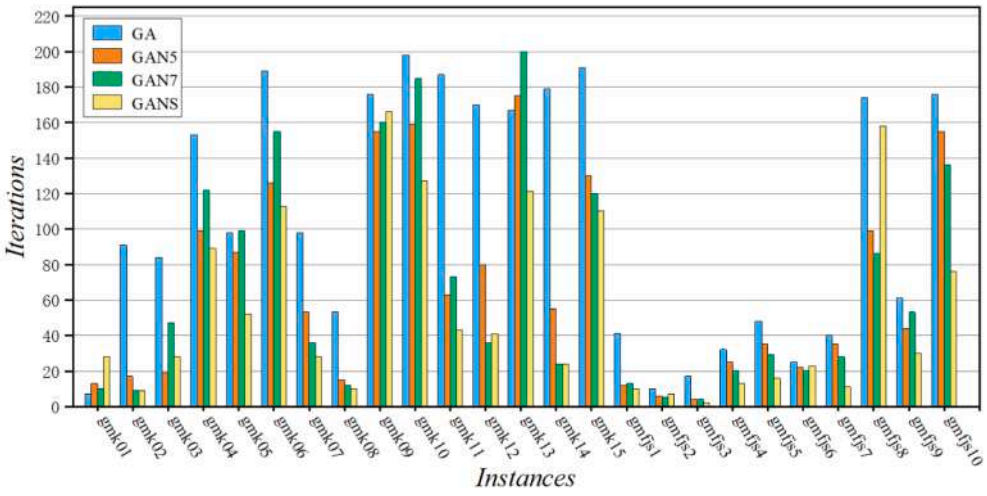
Instance	GA	GAN5	GAN7	GANS	Instance	GA	GAN5	GAN7	GANS
gmk01	**0**	**0**	**0**	0.42	gmk14	**0**	**0**	**0**	**0**
gmk02	0.42	0.42	0.42	**0**	gmk15	**4.29**	17.00	4.52	17.00
gmk03	0.42	0.42	0.42	**0**	gmfjs1	0.52	**0**	**0**	**0**
gmk04	2.15	2.22	1.96	**1.08**	gmfjs2	**0**	**0**	**0**	1.55
gmk05	1.96	1.71	1.71	**1.23**	gmfjs3	**0**	**0**	**0**	**0**
gmk06	1.71	2.54	2.46	**1.58**	gmfjs4	11.93	21.22	13.93	**7.75**
gmk07	1.03	2.55	2.55	**0.84**	gmfjs5	5.90	15.88	12.42	**1.03**
gmk08	**0**	**0**	**0**	**0**	gmfjs6	9.70	18.00	13.65	**7.75**
gmk09	8.82	6.96	**3.68**	8.12	gmfjs7	**8.43**	11.09	9.69	10.67
gmk10	6.20	6.58	6.22	**4.34**	gmfjs8	34.67	**15.35**	16.15	19.25
gmk11	6.26	5.27	5.03	**3.13**	gmfjs9	43.44	52.52	41.28	**41.03**
gmk12	5.98	3.39	3.24	**1.55**	gmfjs10	**21.58**	36.97	41.14	31.02
gmk13	11.38	8.59	**7.77**	17.00					

instances. Moreover, compared to GAN5, the $B(Cov)$ and $Av(Cov)$ attained by GANS outperforms it in 20 instances while being worse in only 5 instances. Furthermore, in comparison to GAN7, the $B(Cov)$ obtained by GANS excels in 17 instances but lags behind in 6 instances. Additionally, the $Av(Cov)$ achieved by GANS outperforms GAN7 in 20 instances while being worse in only 5 instances. This findings indicate that GANS outperforms GA, GAN5 and GAN7 in terms of both best convergences $B(Cov)$ and average convergences $Av(Cov)$. Thus, utilizing the hybrid N5/7 proves effective in improving the convergence performance of the GA for the GFJSP_PBPM. Overall, we can conclude that the hybrid N5/7 improves the GA's search capability for superior makespan while also enhancing its stability and convergence performance.

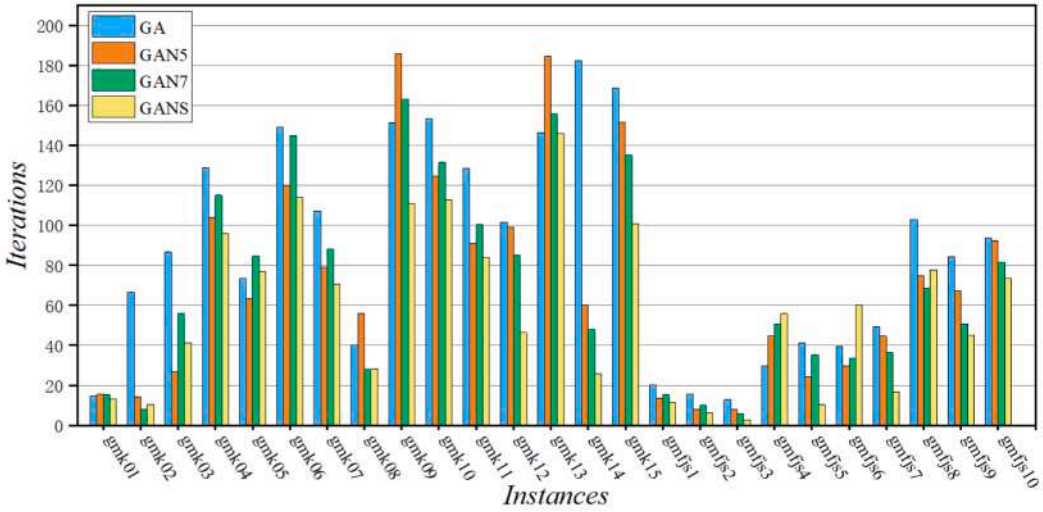
5.6. Comparison with other algorithms

Given the novelty of the GFJSP_PBPM addressed in this study and the absence of existing research results for comparison, two algorithms closely related to our study are selected for comparison. The first algorithm is the latest GRASP, specifically designed by Knopp et al. (2017) and Boyer et al. (2021) to address the GFJSP_PBPM. The distinction between the GFJSP_PBPM we are investigating and the one tackled by GRASP lies in the fact that GRASP solely focuses on PBPM with FPBPO. Two neighborhood structures are embedded in GRASP for local search, the first neighborhood structure reassigns each task on the critical path to an optional machine, and the second neighborhood structure exchanges a task on the critical path with a task that precedes that task on the same machine. The second algorithm is VND-hGA proposed by Liu et al. (2021). This method is a recent algorithm designed specifically for solving traditional FJSP, combining a GA with a variable neighborhood structure. This variable neighborhood structure involves three types of neighborhood operations, the first one replaces operations on the critical path to other optional machines, the second one inserts operations on the critical path into other non-critical path blocks, and the third one is the N6 neighborhood structure. In this study, adjustments are made to GRASP and VND-hGA to ensure that the constraints of GFJSP_PBPM could be satisfied. Especially for VND-hGA, it's imperative to replace the initialization, crossover, and mutation steps with the hybrid initialization strategy, Split-OOOX, and hybrid mutation operators, respectively. This ensures that VND-hGA can effectively tackle the specific problem addressed in this research.

In the FJSP literature, $B(C_{max})$, $Av(C_{\max})$ and Av(CPU) are mainly considered when comparing the effectiveness and efficiency of algorithms (Liu et al., 2021). However, GRASP terminates its algorithm based on the number of iterations and preset runtime, making it challenging to compare its efficiency with other algorithms. Therefore, in this comparative experiment, each algorithm adopts the conditions of



(a) Best convergence for each instance



(b) Average convergence for each instance

Fig. 10. Convergence of the GA, GAN5, GAN7 and GANS.

setting the maximum iterations 300 or a uniform runtime 500 s to evaluate the results obtained by each algorithm. Meanwhile, the population size of VND-hGA are also set to 500, which is the same as GANS.

The $B(C_{\max})$ and $Av(C_{max})$ values for each instance are compared across these algorithms, with the results presented in Table 7. It can be observed that, among all 35 instances in GBRdata and GFdata, GRASP attains the minimum $B(C_{\max})$ in 11 instances and the minimum $Av(C_{max})$ in 10 instances; VND-hGA achieves the minimum $B(C_{\max})$ in 24 instances and the minimum $Av(C_{max})$ in 22 instances; GANS attains the minimum $B(C_{\max})$ in 33 instances and the minimum $Av(C_{max})$ in 26 instances. Only in gmk01 and gmk05, the $B(C_{\max})$ metric of GANS is slightly worse than that of VND-hGA. The VND-hGA has a strong local search capability, but it is time-consuming for it integrates three types of neighborhood operations. Consequently, the number of iterations conducted by VND-hGA within the same time is reduced, resulting in solutions that are not as effective as those generated by GANS. The relative deviation (*Dev*) obtained by GANS compared to GRASP ranges from 0 % to 36.67 %, with a corresponding average improvement of 10.41 %. The *Dev* obtained by GANS compared to VND-hGA ranges from −8.00 % to 13.96 %, with a corresponding average improvement of 0.99 %. This indicates that GANS has better optimization capability compared to GRASP and VND-hGA under the same stopping conditions. In the instances from gmfjs1 to gmfjs10, except for gmfjs5, there is an overall trend where, as the problem scale increases, relative deviation between GRASP and GANS become larger. Meanwhile, the difference in $B(C_{\max})$ and $Av(C_{max})$ obtained by GRASP is significantly larger, while the difference in $B(C_{\max})$ and $Av(C_{max})$ obtained by GANS is smaller comparing to GRASP and VND-hGA, indicating that GANS has relatively less randomness in different runs and good stability. In summary, the following conclusions can be drawn: GANS can achieve high-quality solutions and has good stability. Fig. 11, Fig. 12 present the Gantt charts of gmk15 and gmfjs10 obtained by GANS, respectively.

5.7. Case study

This study aims to further validate the model and the proposed GANS through practical example involving three distinct categories of products at the testing lab of CEPREI. The electronic products under corresponding performance testing include in-vehicle navigators, mobile phones, and unmanned aerial vehicles (UAVs). The overall performance testing process plan for these three types of products are shown in Fig. 1, Fig. 13, and Fig. 14. Each figure includes the number of prototypes for each category of product (weight of the job), the process plan of a job, and MPBPOs formed by different jobs of the same product. Each type of product has 2 different models with the same testing standard and overall process plans waiting to be scheduled. Flexible concurrent

Table 7
Comparison of different algorithms.

Instances	GRASP			VND-hGA			GANS	
	$B(C_{max})$	$Av(C_{max})$	$Dev(\%)$	$B(C_{max})$	$Av(C_{max})$	$Dev(\%)$	$B(C_{max})$	$Av(C_{max})$
gmk01	43	45.2	4.65	42	**42.0**	2.38	**41**	42.6
gmk02	40	41.6	32.50	**25**	**26.4**	−8.00	27	27.6
gmk03	238	249.7	14.29	**204**	**204.0**	0	**204**	**204.0**
gmk04	77	78.7	12.99	**65**	**66.8**	−3.08	67	70.8
gmk05	185	192.2	10.27	167	**168.7**	0.60	**166**	169.4
gmk06	120	125.1	36.67	**76**	**77.8**	0	**76**	78.8
gmk07	199	206.3	29.15	**141**	143.3	0	**141**	**143.0**
gmk08	534	542.6	2.06	**523**	524.0	0	**523**	**523.0**
gmk09	392	402.6	16.58	342	350.6	4.39	**327**	**335.4**
gmk10	341	345.6	28.74	263	269.2	7.60	**243**	**252.0**
gmk11	666	677.8	8.71	633	640.2	3.95	**608**	**617.8**
gmk12	572	589.2	11.12	525	529.9	3.24	**508**	**514.0**
gmk13	561	587.1	25.31	487	495.9	13.96	**419**	**464.4**
gmk14	834	869.2	16.89	720	743.5	3.61	**694**	**694.0**
gmk15	517	523.0	21.93	439	453.7	3.54	**424**	**444.8**
gsfjs1	**51**	**51.0**	0	**51**	**51.0**	0	**51**	**51.0**
gsfjs2	**107**	**107.0**	0	**107**	**107.0**	0	**107**	**107.0**
gsfjs3	**208**	**208.0**	0	**208**	**208.0**	0	**208**	**208.0**
gsfjs4	**272**	**272.0**	0	**272**	**272.0**	0	**272**	**272.0**
gsfjs5	**100**	**100.0**	0	**100**	**100.0**	0	**100**	**100.0**
gsfjs6	**440**	**440.0**	0	**440**	**440.0**	0	**440**	**440.0**
gsfjs7	**407**	**407.0**	0	**407**	**407.0**	0	**407**	**407.0**
gsfjs8	**430**	**430.0**	0	**430**	**430.0**	0	**430**	**430.0**
gsfjs9	**494**	**494.0**	0	**494**	**494.0**	0	**494**	**494.0**
gsfjs10	**520**	**520.0**	0	**520**	**520.0**	0	**520**	**520.0**
gmfjs1	**468**	471.7	0	**468**	468.1	0	**468**	**468.0**
gmfjs2	448	450.8	3.35	**433**	**433.0**	0	**433**	**433.0**
gmfjs3	562	582.7	4.27	**538**	**538.0**	0	**538**	**538.0**
gmfjs4	597	639.3	7.87	**550**	**560.5**	0	**550**	563.8
gmfjs5	559	617.5	2.15	**547**	**553.3**	0	**547**	556.2
gmfjs6	718	748.9	11.70	**634**	**644.3**	0	**634**	649.0
gmfjs7	976	1062.1	10.35	**875**	892.0	0	**875**	**886.2**
gmfjs8	1071	1093.4	18.39	**874**	908.6	0	**874**	**896.5**
gmfjs9	1265	1324.0	16.60	1060	**1109.2**	0.47	**1055**	1131.6
gmfjs10	1455	1514.0	17.66	1224	1266.1	2.12	**1198**	**1240.2**
Average improvement	10.41			0.99				

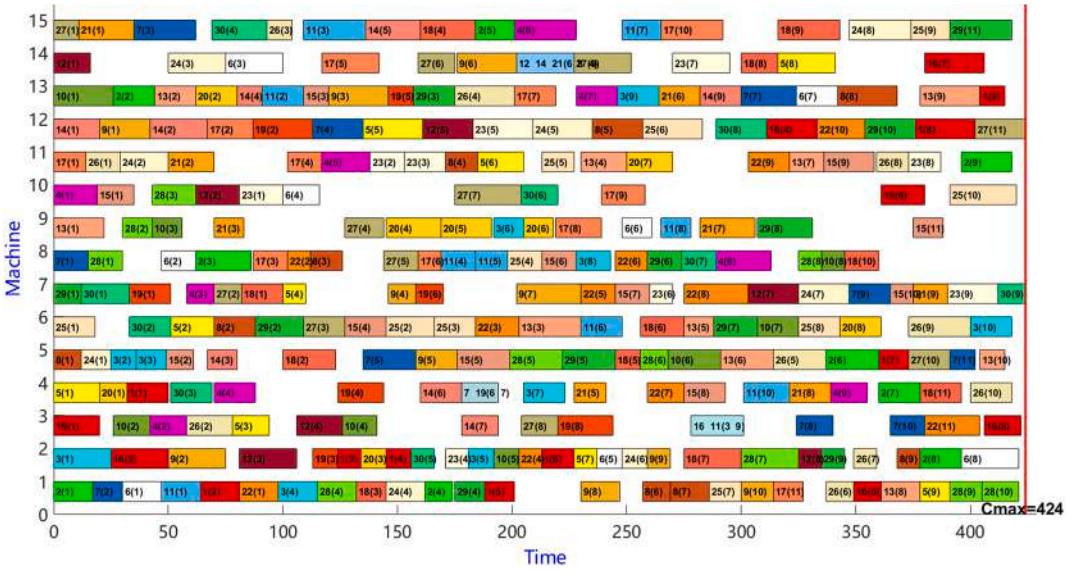


Fig. 11. Gantt chart of a scheduling scheme for gmk15.

processing of temperature cycling test of the two types of navigators, thermal cycling test of the two types of mobile phones, and the high-low temperature charge–discharge test of two types of UAVs, thereby forming FPBPOs. The capacity of machines performing PBPOs are set to 12, 20, 6 while testing navigator, mobile phone and UAVs, respectively.

Similarly, utilizing GRASP, VND-hGA and GANS to solve the problem. The statistical results are shown in Table 7. In the "Rule-based" column of Table 8, it presents the actual result obtained from the scheduling rules implemented in the testing workshop. In the workshop, the Shortest Processing Time (SPT) scheduling rule is applied to non

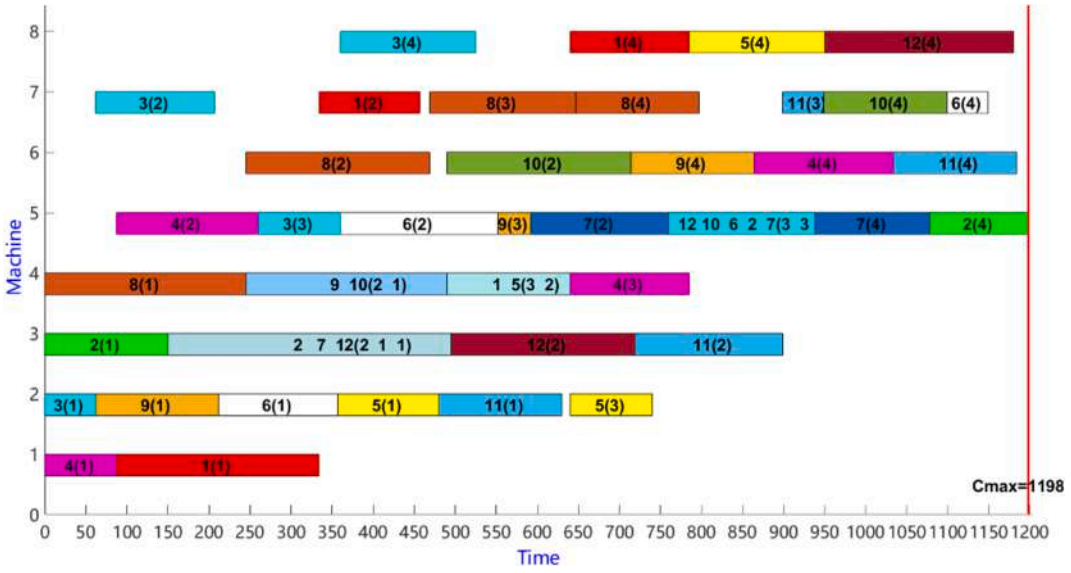


Fig. 12. Gantt chart of a scheduling scheme for gmfjs10.

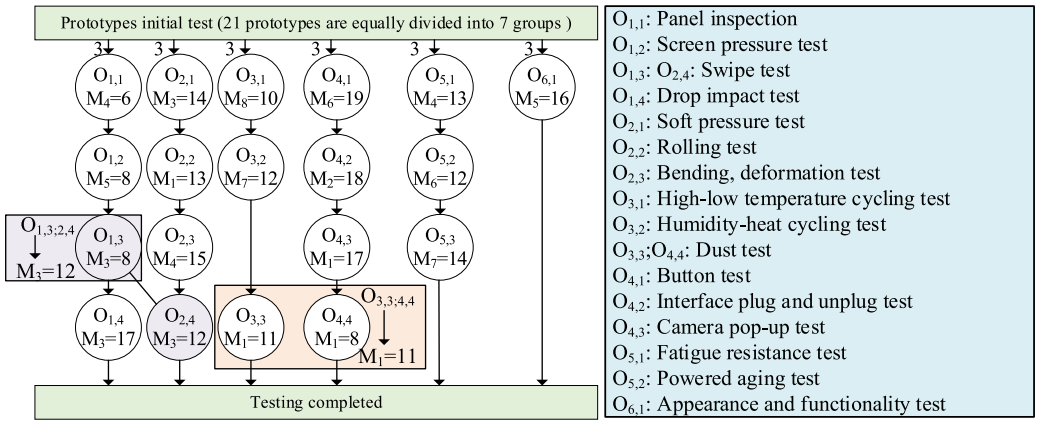


Fig. 13. Performance testing process plan of the mobile phone.

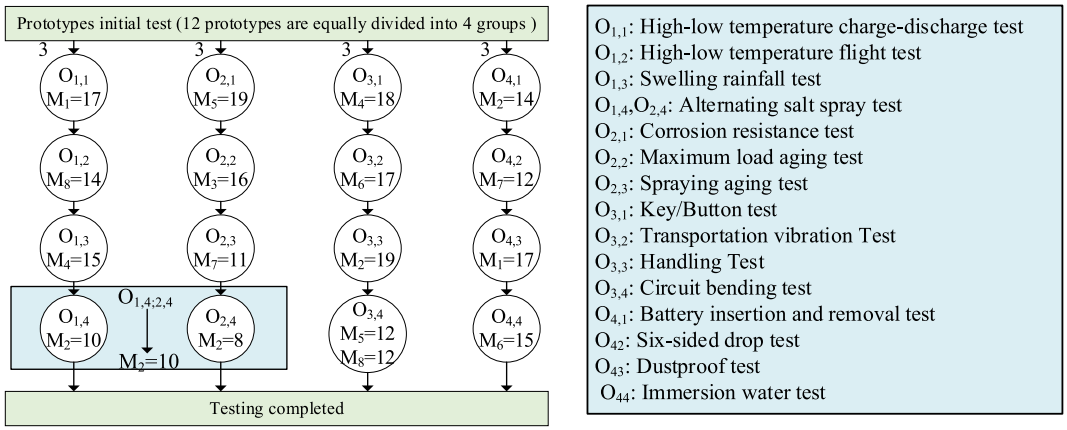


Fig. 14. Performance testing process plan of the UAVs.

PBPO tasks, while a greedy strategy is employed for each PBPO task. Under this strategy, operations suitable for batch processing are conducted in parallel on the same machine whenever the capacity permits. It can be seen that both GANS and VND-hGA achieved the best $B(C_{max})$ and $Av(C_{max})$, and the $Av(C_{max})$ of the 10 runs is the same as the $B(C_{max})$. This further illustrates that GANS can achieve the minimum $B(C_{max})$, and it also has a clear advantage in terms of stability and consistency. Comparing to the actual result in the workshop, the GANS optimized $B(C_{max})$ by 27.0 %. Fig. 15 presents the Gantt chart of the optimal scheduling scheme achieved by GANS. In which, Job1-Job7 represents

Table 8
Comparison of scheduling results for instances from electronic product testing workshop.

Rule based	GRASP		VND-hGA		GANS	
$B(C_{max})$	$B(C_{max})$	$Av(C_{max})$	$B(C_{max})$	$Av(C_{max})$	$B(C_{max})$	$Av(C_{max})$
185	136	137.6	**135**	**135.0**	**135**	**135.0**

the 7 groups of the in-vehicle navigator, Job8-Job13 represents the 6 groups of the mobile phone, Job14-Job17 represents the 4 groups of the UAVs.

5.8. Discussion

The optimization model, validated with the CPLEX solver, although it can only solve feasible solutions for small-scale problems, still proves that the model effectively captures the complex constraints of GFJSP_PBPM and provides a foundational understanding and articulation of the problem, which is essential for designing other algorithms. The customized task-based encoding and active schedule-based decoding proposed in this study have been applied to the corresponding N5, N7 and hybrid N5/7 neighborhood structures, as well as the GRASP, VND-hGA, and GANS algorithms, all of which yield feasible solutions without violating the relevant constraints. This encoding and decoding methods clarify the structure of GFJSP_PBPM, thereby laying the groundwork for developing additional algorithms for this problem.

The hybrid initialization strategy integrates random and greedy approaches for machine assignment while considering the distinctions among non-PBPO, MPBPO, and FPBPO in job sequence. Results indicate that this strategy ensures the feasibility of the initial solutions, enhances the quality of the initial population, and accelerates convergence toward the optimal makespan. The Split-OOOX integrates information from two chromosomes and verifies that the predecessor tasks of operations in the PBPO are correctly positioned in the offspring during the generation of new individuals, enabling the GA to maintain solution feasibility and diversity throughout the iterations. The segment swap mutation for job sequences swaps two tasks within a randomly selected segment divided by two PBPOs, preventing constraint violations and maintaining diversity during iterations. Additionally, single-point mutation for machine sequence further enhance diversity during the iterations. These strategies and operators have been implemented in both VND-hGA and GANS, with results demonstrating their feasibility and ensuring the GA's robust global search capability. These methods can serve as fundamental operators for other GAs addressing GFJSP_PBPM.

The N5/7 neighborhood structure enhances the GA search capability for superior makespan while simultaneously improving its stability and convergence performance. This enhancement can be attributed to several key factors: addressing the constraints of both FJSP and PBPO, which ensures the feasibility of neighborhood solutions; selecting varying numbers of critical blocks for neighborhood operations according to the problem scale, thereby flexibly expanding the high-quality neighborhood space; and choosing either N5 or N7 for neighborhood operations depending on the number of tasks within those blocks, thus making better use of the simplicity and efficiency of N5 as well as the broader search range of N7. This methodologies and principles underlying the N5/7 also offer valuable insights for developing adaptive neighborhood structures for FJSP and GFJSP_PBPM.

However, each strategy or operator in the GANS designed for GFJSP_PBPM—including hybrid initialization, the Split-OOOX operator, hybrid mutation, and the N5/7 neighborhood structure—requires searching all predecessor tasks and determining whether they are completed to establish the placement of the current task, ensuring that predecessor-successor constraints are not violated. This search and verification process reduces the efficiency of GANS, highlighting the need for a more effective judgment and constraint assurance mechanism.

6. Conclusion and future research

To optimize the makespan of GFJSP_PBPM, this study constructs an optimization model based on MIP and introduces a hybrid mechanism, GANS. In GANS, global search is performed using GA, while local refinement is achieved through a hybrid N5/7 neighborhood structure. The GA incorporates a hybrid initialization strategy, novel encoding and decoding schemes, a Split-OOOX operator, and a hybrid mutation, all specifically designed to address the constraints of GFJSP_PBPM. The hybrid N5/7 structure refines N5 and N7 to adapt to GFJSP_PBPM constraints, selects critical blocks for neighborhood operations based on problem scale, and ultimately decides between N5 or N7 for neighborhood operations based on the number of tasks within those blocks.

The effectiveness test of the optimization model based on CPLEX solver confirms that it accurately represents the constraints of GFJSP_PBPM and demonstrates its feasibility. The performance of the proposed hybrid initialization strategy and hybrid N5/7 structure is

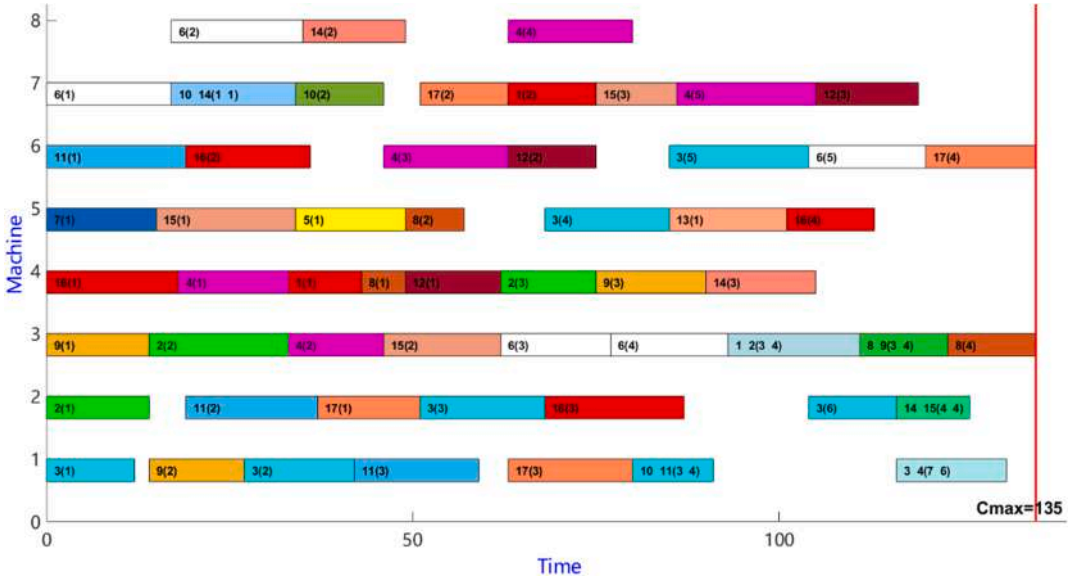


Fig. 15. Gantt chart of a scheduling scheme obtained by GANS for the three types of products.

empirically validated using 35 newly created benchmark instances. The results indicate that the hybrid initialization strategy effectively generates feasible solutions, enhances the quality of initial solutions, and accelerates convergence speed. The hybrid N5/7 neighborhood structure is proven effective in enhancing GA's ability to search for superior makespan while simultaneously improving its stability and convergence performance. GANS is evaluated through comparative analysis with GRASP and VND-hGA, demonstrating superior makespan, as well as advantages in stability and efficiency. Furthermore, GANS is applied to optimize GFJSP_PBPM in an electronic product testing workshop, validating its superior performance in both optimal and average makespan.

In practical applications, various constraints are encountered in electronic product performance testing and mold manufacturing workshops, including MPBPO, FPBPO, multiple resource constraints, and diverse time constraints. The upcoming phase of research will concentrate on exploring GFJSP with the aforementioned multiple constraints, aiming to better optimize real-world production scenarios.

CRediT authorship contribution statement

Hucheng Zhang: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Shengping Lv:** Investigation, Conceptualization, Data curation, Formal analysis, Methodology, Resources, Writing – original draft, Writing – review & editing, Project administration. **Dequan Xin:** Data curation, Formal analysis, Methodology, Software, Validation, Visualization. **Hong Jin:** Investigation, Conceptualization, Formal analysis, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Shengping Lv reports financial support was provided by National Natural Science Foundation of China (Grant No. 52275487). Shengping Lv reports financial support was provided by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2021A1515012395). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Grant No. 52275487), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2021A1515012395). The authors thank the anonymous reviewers for their valuable and constructive comments that greatly helped improve the quality and completeness of the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2024.125888.

Data availability

Data will be made available on request.

References

Abedi, M., Chiong, R., Noman, N., & Zhang, R. (2020). A multi-population, multi-objective memetic algorithm for energy-efficient job-shop scheduling with deteriorating machines. *Expert Systems with Applications, 157*, Article 113348. https://doi.org/10.1016/j.eswa.2020.113348

Bagheri, A., Zandieh, M., Mahdavi, I., & Yazdani, M. (2010). An artificial immune algorithm for the flexible job-shop scheduling problem. *Future Generation Computer Systems, 26*(4), 533–541. https://doi.org/10.1016/j.future.2009.10.004

Balas, E. (1969). Machine sequencing via disjunctive graphs: An implicit enumeration algorithm. *Operation Research, 17*(6), 941–957. https://doi.org/10.2307/168317

Bierwirth, C., & Mattfeld, D. C. (1999). Production scheduling and rescheduling with genetic algorithms. *Evolutionary Computation, 7*, 1–17. https://doi.org/10.1162/evco.1999.7.1.1

Boyer, V., Vallikavungal, J., Rodríguez, X. C., & Salazar-Aguilar, M. A. (2021). The generalized flexible job shop scheduling problem. *Computers & Industrial Engineering, 160*, Article 107542. https://doi.org/10.1016/j.cie.2021.107542

Brandimarte, P. (1993) Routing and scheduling in a flexible job shop by tabu search. *Annals of Operations Research*, 41, 157-183. https:// 10.1007/BF02023073.

Brucker, P., & Schlie, R. (1990). Job-shop scheduling with multi-purpose machines. *Computing*, 45(4), 369-375. https:// 10.1007/BF02238804.

Caldeira, R. H., & Gnanavelbabu, A. (2019). Solving the flexible job shop scheduling problem using an improved Jaya algorithm. *Computers & Industrial Engineering, 137*, Article 106064. https://doi.org/10.1016/j.cie.2019.106064

Chen, N.-L., Xie, N.-M., & Wang, Y.-Q. (2022). An elite genetic algorithm for flexible job shop scheduling problem with extracted grey processing time. *Applied Soft Computing, 131*, Article 109783. https://doi.org/10.1016/j.asoc.2022.109783

Chen, R.-H., Yang, B., Li, S., & Wang, S.-L. (2020). A self-learning genetic algorithm based on reinforcement learning for flexible job-shop scheduling problem. *Computers & Industrial Engineering, 149*, Article 106778. https://doi.org/10.1016/j.cie.2020.106778

Dauzère-Pérès, S., Ding, J. W., Shen, L.-J., & Tamssaouet, K. (2023). The flexible job shop scheduling problem: A review. *European Journal of Operational Research, 314*(2), 409–432. https://doi.org/10.1016/j.ejor.2023.05.017

Fan, H.-L., & Su, R. (2022). Mathematical modelling and heuristic approaches to job-shop scheduling problem with conveyor-based continuous flow transporters. *Computers & Operations Research, 148*, Article 105998. https://doi.org/10.1016/j.cor.2022.105998

Fowler, J. W., & Mönch, L. (2022). A survey of scheduling with parallel batch (p-batch) processing. *European Journal of Operational Research, 298*(1), 1–24. https://doi.org/10.1016/j.ejor.2021.06.012

Fontes, D. B. M. M., Homayouni, S. M., & Gonçalves, J. F. (2023). A hybrid particle swarm optimization and simulated annealing algorithm for the job shop scheduling problem with transport resources. *European Journal of Operational Research, 306*(3), 1140–1157. https://doi.org/10.1016/j.ejor.2022.09.006

Gao, J., Sun, L.-Y., & Gen, M. (2008). A hybrid genetic and variable neighborhood descent algorithm for flexible job shop scheduling problems. *Computers & Operations Research, 35*(9), 2892–2907. https://doi.org/10.1016/j.cor.2007.01.001

Gao, K.-Z., Suganthan, P. N., Pan, Q.-K., Chua, T.-Y., Chong, C.-S., & Cai, T.-X. (2016). An improved artificial bee colony algorithm for flexible job-shop scheduling problem with fuzzy processing time. *Expert Systems With Applications, 65*, 52–67. https://doi.org/10.1016/j.eswa.2016.07.046

Gong, G.-L., Deng, Q.-W., Chiong, R., Gong, X.-R., & Huang, H.-Z.-Y. (2019). An effective memetic algorithm for multi-objective job-shop scheduling. *Knowledge-Based Systems, 182*, Article 104840. https://doi.org/10.1016/j.knosys.2019.07.011

Ham, A. M., & Cakici, E. (2016). Flexible job shop scheduling problem with parallel batch processing machines: MIP and CP approaches. *Computers & Industrial Engineering, 102*, 160–165. https://doi.org/10.1016/j.cie.2016.11.001

Ham, A. (2017). Flexible job shop scheduling problem for parallel batch processing machine with compatible job families. *Applied Mathematical Modelling, 45*, 551–562. https://doi.org/10.1016/j.apm.2016.12.034

Ham, A., Fowler, J. W., & Cakici, E. (2017). Constraint programming approach for scheduling jobs with release times, non-identical sizes, and incompatible Families on parallel batching machines. *IEEE Transactions on Semiconductor Manufacturing, 30*(4), 500–507.

Hu, C. M., Zheng, R., Lu, S.-J., Liu, X.-B., & Cheng, H. (2023). Integrated optimization of production scheduling and maintenance planning with dynamic job arrivals and mold constraints. *Computers & Industrial Engineering, 186*, Article 109708. https://doi.org/10.1016/j.cie.2023.109708

Huang, J. P., Gao, L., Li, X.-Y., & Zhang, C.-J. (2023). A cooperative hierarchical deep reinforcement learning based multi-agent method for distributed job shop scheduling problem with random job arrivals. *Computers & Industrial Engineering, 185*, Article 109650. https://doi.org/10.1016/j.cie.2023.109650

Huang, X.-B., & Yang, L.-X. (2019). A hybrid genetic algorithm for multi-objective flexible job shop scheduling problem considering transportation time. *International Journal of Intelligent Computing and Cybernetics, 12*(2), 154–174. https://doi.org/10.1108/IJICC-10-2018-0136

Knopp, S., Dauzère-Pérès, S., & Yugma, C. (2017). A batch-oblivious approach for complex job-shop scheduling problems. *European Journal of Operational Research, 263*(1), 50–61. https://doi.org/10.1016/j.ejor.2017.04.050

Li, X.-Y., & Gao, L. (2016). An effective hybrid genetic algorithm and tabu search for flexible job shop scheduling problem. *International Journal of Production Economics, 174*, 93–110. https://doi.org/10.1016/j.ijpe.2016.01.016

Liang, X., Huang, M., & Ning, T. (2018). Flexible job shop scheduling based on improved hybrid immune algorithm. *Journal of Ambient Intelligence and Humanized Computing, 9*(1), 165–171. https://doi.org/10.1007/s12652-016-0425-9

Lin, W.-H., Deng, Q.-W., Han, W.-W., Gong, G.-L., & Li, K.-X. (2022). An effective algorithm for flexible assembly job-shop scheduling with tight job constraints. *International Transactions in Operational Research, 29*(1), 496–525. https://doi.org/10.1111/itor.12767

Liu, Z.-F., Wang, J.-L., Zhang, C.-X., Chu, H.-Y., Ding, G.-L., & Zhang, L. (2021). A hybrid genetic-particle swarm algorithm based on multilevel neighbourhood structure for

flexible job shop scheduling problem. *Computers & Operations Research, 135*, Article 105431. https://doi.org/10.1016/j.cor.2021.105431

Meng, L.-L., Zhang, C.-Y., Zhang, B., & Ren, Y.-P. (2019). Mathematical modeling and optimization of energy-conscious flexible job shop scheduling problem with worker flexibility. *IEEE Access, 7*, 68043–68059. https://doi.org/10.1109/ACCESS.2019.2916468

Nowicki, E., & Smutnicki, C. (1996). A fast taboo search algorithm for the job shop problem. *Management Science, 42*(6), 797–813. https://doi.org/10.1287/mnsc.42.6.797

Ozguven, C., Ozbaklr, L., & Yavuz, Y. (2010). Mathematical models for job-shop scheduling problems with routing and process plan flexibility. *Applied Mathematical Modelling, 34*(6), 1539–1548. https://doi.org/10.1016/j.apm.2009.09.002

Palacios, J. J., González, M. A., Vela, C. R., González-Rodríguez, I., & Puente, J. (2015). Genetic tabu search for the fuzzy flexible job shop problem. *Computers & Operations Research, 54*, 74–89. https://doi.org/10.1016/j.cor.2014.08.023

Park, J. S., Ng, H. Y., Chua, T. J., Ng, Y. T., & Kim, J. W. (2021). Unified genetic algorithm approach for solving flexible job-shop scheduling problem. *Applied Sciences, 11*(14), 6454. https://doi.org/10.3390/app11146454

Sotskov, Y. N. (1991). The complexity of shop-scheduling problems with two or three jobs. *European Journal of Operational Research, 53*(3), 326–336. https://doi.org/10.1016/0377-2217(91)90066-5

Sun, K.-X., Zheng, D.-B., Song, H.-H., Cheng, Z.-W., Lang, X.-D., Yuan, W.-D., & Wang, J.-Q. (2023). Hybrid genetic algorithm with variable neighborhood search for flexible job shop scheduling problem in a machining system. *Expert Systems with Applications, 215*, 19359. https://doi.org/10.1016/j.eswa.2022.119359

Tang, H.-T., Chen, R., Li, Y.-B., Peng, Z., Peng, S.-S., & Du, Y.-Z. (2019). Flexible job-shop scheduling with tolerated time interval and limited starting time interval based on hybrid discrete PSO-SA: An application from a casting workshop. *Applied Soft Computing, 78*, 176–194. https://doi.org/10.1016/j.asoc.2019.02.011

Türkyılmaz, A., & Bulkan, S. (2014). hybrid algorithm for total tardiness minimisation in flexible job shop: Genetic algorithm with parallel VNS execution. *International Journal of Production Research, 53*(6), 1832–1848. https://doi.org/10.1080/00207543.2014.962113

Wu, X.-L., & Sun, Y.-J. (2018). A green scheduling algorithm for flexible job shop with energy-saving measures. *Journal of Cleaner Production, 172*, 3249–3264. https://doi.org/10.1016/j.jclepro.2017.10.342

Wolpert, D. H., & Macready, W. G. (2005). Coevolutionary free lunches. *IEEE Transactions on Evolutionary Computation, 9*(6), 721–735. https://doi.org/10.1109/TEVC.2005.856205

Xie, J., Gao, L., Peng, K.-K., Li, X.-Y., & Li, H.-R. (2019). Review on flexible job shop scheduling. *The Institution of Engineering and Technology, 1*(3), 66–77. https://doi.org/10.1049/iet-cim.2018.0009

Xie, J., Li, X.-Y., Gao, L., & Gui, L. (2023a). A hybrid genetic tabu search algorithm for distributed flexible job shop scheduling problems. *Journal of Manufacturing Systems, 71*, 82–94. https://doi.org/10.1016/j.jmsy.2023.09.002

Xie, J., Li, X.-Y., Gao, L., & Gui, L. (2023b). A new neighbourhood structure for job shop scheduling problems. *International Journal of Production Research, 61*(7), 2147–2161. https://doi.org/10.1080/00207543.2022.2060772

Zhang, C.-Y., Li, P. G., Guan, Z. L., & Rao, Y.-Q. (2007). A tabu search algorithm with a new neighborhood structure for the job shop scheduling problem. *Computers & Operations Research, 34*(11), 3229–3242. https://doi.org/10.1016/j.cor.2005.12.002

Zhang, F.-Y., Li, R., & Gong, W.-Y. (2024). Deep reinforcement learning-based memetic algorithm for energy-aware flexible job shop scheduling with multi-AGV. *Computers & Industrial Engineering, 189*, Article 109917. https://doi.org/10.1016/j.cie.2024.109917

Zhang, G.-H., Zhang, L.-J., Song, X., Wang, Y.-C., & Zhou, C. (2019). A variable neighborhood search based genetic algorithm for flexible job shop scheduling problem. *Cluster Computing, 22*, 11561–11572. https://doi.org/10.1007/s10586-017-1420-4

Zhang, H.-L., Xu, G.-J., Pan, R.-L., & Ge, H.-J. (2022). A novel heuristic method for the energy-efficient flexible job-shop scheduling problem with sequence-dependent set-up and transportation time. *Engineering Optimization, 54*(10), 1646–1667. https://doi.org/10.1080/0305215X.2021.1949007

Zhang, S.-J., Du, H.-T., Borucki, S., Jin, S.-F., Hou, T.-T., & Li, Z.-X. (2021). Dual resource constrained flexible job shop scheduling based on improved quantum genetic algorithm. *Machines, 9*(6), 108. https://doi.org/10.3390/machines9060108

Zhong, H.-Y., Liu, J. J., Chen, Q.-X., Mao, N., & Yang, X. J. (2020). Performance assessment of dynamic flexible assembly job shop control methods. *IEEE Acess, 8*, 226042–226058. https://doi.org/10.1109/ACCESS.2020.3043880

Zhou, Y., Yang, J.-J., & Zheng, L.-Y. (2019). Multi-agent based hyper-heuristics for multi-objective flexible job shop scheduling: A case study in an aero-engine blade manufacturing plant. *IEEE Acess, 7*, 21147–21176. https://doi.org/10.1109/ACCESS.2019.2897603

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 November 2023
DOI 10.3389/fpls.2023.1283230

OPEN ACCESS

EDITED BY
Qiang Lyu,
Southwest University, China

REVIEWED BY
Yunchao Tang,
Guangxi University, China
Liantao Liu,
Hebei Agricultural University, China

*CORRESPONDENCE
Shengping Lv
lvshengping@scau.edu.cn

RECEIVED 25 August 2023
ACCEPTED 20 October 2023
PUBLISHED 08 November 2023

CITATION
Li X, Zhang Z, Lv S, Liang T, Zou J, Ning T and Jiang C (2023) Detection of breakage and impurity ratios for raw sugarcane based on estimation model and MDSC-DeepLabv3+. *Front. Plant Sci.* 14:1283230. doi: 10.3389/fpls.2023.1283230

Detection of breakage and impurity ratios for raw sugarcane based on estimation model and MDSC-DeepLabv3+

Xin Li, Zhigang Zhang, Shengping Lv*, Tairan Liang, Jianmin Zou, Taotao Ning and Chunyu Jiang

College of Engineering, South China Agricultural University, Guangzhou, China

Broken cane and impurities such as top, leaf in harvested raw sugarcane significantly influence the yield of the sugar manufacturing process. It is crucial to determine the breakage and impurity ratios for assessing the quality and price of raw sugarcane in sugar refineries. However, the traditional manual sampling approach for detecting breakage and impurity ratios suffers from subjectivity, low efficiency, and result discrepancies. To address this problem, a novel approach combining an estimation model and semantic segmentation method for breakage and impurity ratios detection was developed. A machine vision-based image acquisition platform was designed, and custom image and mass datasets of cane, broken cane, top, and leaf were created. For cane, broken cane, top, and leaf, normal fitting of mean surface densities based on pixel information and measured mass was conducted. An estimation model for the mass of each class and the breakage and impurity ratios was established using the mean surface density and pixels. Furthermore, the MDSC-DeepLabv3+ model was developed to accurately and efficiently segment pixels of the four classes of objects. This model integrates improved MobileNetv2, atrous spatial pyramid pooling with deepwise separable convolution and strip pooling module, and coordinate attention mechanism to achieve high segmentation accuracy, deployability, and efficiency simultaneously. Experimental results based on the custom image and mass datasets showed that the estimation model achieved high accuracy for breakage and impurity ratios between estimated and measured value with R^2 values of 0.976 and 0.968, respectively. MDSC-DeepLabv3+ outperformed the compared models with mPA and mIoU of 97.55% and 94.84%, respectively. Compared to the baseline DeepLabv3+, MDSC-DeepLabv3+ demonstrated significant improvements in mPA and mIoU and reduced Params, FLOPs, and inference time, making it suitable for deployment on edge devices and real-time inference. The average relative errors of breakage and impurity ratios between estimated and measured values were 11.3% and 6.5%, respectively. Overall, this novel approach enables high-precision, efficient, and intelligent detection of breakage and impurity ratios for raw sugarcane.

KEYWORDS

raw sugarcane, breakage ratio, impurity ratio, estimation model, MDSC-DeepLabv3+

1 Introduction

Sugarcane is an important raw material for the sugar industry worldwide. In China, sugarcane-based sugar production reached 4.6 million tons in 2022, which is 4.3 times that of beet sugar (National Development and Reform Commission, 2023). In recent years, the use of machine-harvested sugarcane has been steadily increasing, with plans to reach 30% of total sugarcane harvest in China by 2025 (Chinese government website, 2018). Machine harvesting significantly improves efficiency and reduces labor intensity; however, it also leads to higher ratios of broken cane and impurities such as top, leaf, which can negatively impact the yield of the sugar manufacturing process. As a result, the breakage and impurity ratios are crucial indicators for assessing the quality and pricing of raw sugarcane in practice, and determining these two ratios is indispensable for sugar refineries. Unfortunately, the commonly used manual sampling approach for detecting breakage and impurity ratios brings several issues, including strong subjectivity, low efficiency, and significant result discrepancies.

To address the aforementioned problem, an estimation model was established, and machine vision technology was employed to provide a more objective, efficient, accurate, and intelligent approach for quantifying the cane, broken cane, and impurities, as well as the ratios of breakage and impurity. This enables seamless integration with the sugarcane harvesting and sugar processing stages. Both cane and broken cane can be used as raw materials, but broken cane is considered in mass deduction by sugar refineries because it results in the loss of sugar content and impacts the quality of the final sugar product. The sugarcane top, leaf, root, sand, gravel, and soil and so forth are collectively referred to as impurities (Guedes and Pereira, 2018). Adjusting the height between the harvester's cutting device and the ridge surface will reduce the introduction of sand, gravel, and soil during sugarcane harvesting. Furthermore, when the mechanical harvester operates smoothly and adheres to specifications, it noticeably decreases the levels of mud, stone, and cane root (Xie et al., 2018). Mechanical removal methods, such as vibration, can often be used to screen out the sand, gravel, and soil (Martins and Ruiz, 2020). However, the top, leaf and cane root are unavoidable impurities as they are naturally part of each sugarcane stem (de Mello et al., 2022). Regarding cane root, object detection can be utilized to count its quantities. Combining this with the average weight of the cane root helps predict the mass of root impurity after excluding sand, gravel and soil. Based on the quality detection practice of sugar refineries, the four categories of cane, broken cane, top, and leaf are selected as the detection objects in this study.

Estimation models and machine vision technology have been widely used for the detection and monitoring of impurities in grain crops such as rice, wheat, and corn. For example, Chen et al. (2020) used morphological features and a decision tree for the classification of rice grains and impurities with 76% accuracy to optimize combine harvester parameters. Liu et al. (2023) proposed a NAM-EfficientNetv2 lightweight segmentation approach for rapid online detection of rice seed and impurities in harvesters, achieving high evaluation index F1 scores of 95.26% and 93.27% for rice grain and impurities, respectively. To improve accuracy in wheat and impurity recognition, Shen et al. (2019) constructed a dataset and trained a recognition model called WheNet based on Inception_v3, achieving a recall rate of 98% and an efficiency of 100ms per image. Chen et al. (2022) designed a vision system based on DeepLabv3+ to identify seeds and impurities in wheat, obtaining mean pixel accuracy (mPA) values of 86.86% and 89.91% for grains and impurities, and mean intersection over union (mIoU) scores of 0.7186 and 0.7457, respectively. For the detection of impurities in the corn deep-bed drying process, Li et al. (2022) employed a multi-scale color recovery algorithm to enhance images and eliminate noise. They used HSV color space parameter thresholds and morphological operations for segmentation and achieved F1 scores of 83.05%, 83.87%, and 87.43% for identifying broken corncob, broken bract, and crushed stone, respectively. Liu et al. (2022) developed a CPU-Net semantic segmentation model based on U-Net, incorporating the convolutional block attention module (CBAM) and pyramid pooling modules to improve segmentation accuracy for monitoring corn kernels and their impurities. They established a mass-pixel linear regression model to calculate the kernel impurity rate and experimental results demonstrated that CPU-Net outperforms other comparative approaches with average mIoU, mPA, and inference time scores of 97.31%, 98.71%, and 158.4ms per image, respectively. The average relative error between the impurity rate obtained by the model and manual statistics was 4.64%.

Detection of impurities in cash crops such as soybean, cotton, and walnut during harvesting or processing has also been extensively studied in recent years. Momin et al. (2017) used HSI to segment the image background of soybean with three categories of impurities. They employed various image processing techniques, such as median blur, morphological operations, watershed transformation, projection area-based analysis, and circle detection, for feature recognition of soybean and impurities. The experimental results showed pixel accuracy of 96%, 75%, and 98% for split bean, contaminated bean, and defective bean, and stem/pod, respectively. Jin et al. (2022) developed an improved UNet segmentation model to address issues of soybean sticking, stacking, and complex semantics in images. The experimental results demonstrated comprehensive evaluation index values of 95.50%, 91.88%, and 94.34% for complete grain, broken grain, and impurity segmentation, respectively, with a mIoU of 86.83%. The field experiment indicated mean absolute errors of 0.18 and 0.10 percentage points for fragmentation and impurity rate between the model-based value and the measured value, respectively. For real-time detection of impurity ratio in cotton processing, Zhang et al. (2022) utilized the enhanced Canny algorithm to segment cotton and its impurities. They employed YOLOv5 to identify the segmented objects and determine their respective categories. They also developed an estimation model for the impurity ratio based on segmented volume and estimated mass and utilized a multithread technique to shorten the processing time, achieving a 43.65% reduction compared to that of a single thread. To improve the recognition accuracy of white and near-cotton-colored impurities in raw cotton, Xu et al. (2023) proposed a weighted feature fusion module and a decoupled detection strategy to enhance the detection head of YOLOv4-tiny. The proposed method decreased

computation during the inference process, boosted the speed of inference, and enhanced the accuracy of cotton impurity localization. Experimental results showed a respective increase of 10.35% and 6.9% in mAP and frames per second (FPS) compared to the baseline YOLOv4-tiny. The detection accuracy of white and near cotton-colored impurities in raw cotton reached 98.78% and 98%, respectively. To achieve real-time segmentation of juglans impurity, Rong et al. (2020) proposed a hybrid approach by combining a segmentation model based on a multi-scale residual full convolutional network and a classification method based on a convolutional network. The proposed method accurately segmented 99.4% and 96.5% of the object regions in the test and validation images, respectively, with a segmentation time of within 60ms for each image. Yu L. et al. (2023) presented an improved YOLOv5 with lower parameters and quicker speed for walnut kernel impurity detection by incorporating target detection layers, CBAM, transformer-encoder, and GhostNet. The results indicated a mAP of 88.9%, which outperformed the baseline YOLOv5 by 6.7%.

In recent years, researchers have also achieved notable progress in the field of impurity detection in sugarcane. Guedes and Pereira (2019) constructed an image dataset comprising 122 different combinations of sugarcane stalk, vegetal plant part, and soil to evaluate the impurity amount. They converted color samples into color histograms with ten color scales and employed three classifiers, namely soft independent modeling of class analogy, partial least squares discriminant analysis (PLS-DA), and *k* nearest neighbors (KNN), to classify cane and its impurities. Guedes et al. (2020) further proposed an analytical method using artificial neural networks (ANNs) combined with the ten color histograms to predict the content of sugarcane in the presence of impurities. The experimental results demonstrated correlation coefficients of 0.98, 0.93, and 0.91 for the training, validation, and test sets, respectively. Aparatana et al. (2020) employed principal component analysis (PCA), PLS-DA, and support vector machine (SVM) to classify and differentiate sugarcane and impurities, including green leaf, dry leaf, stone, and soil, based on their spectral information. The research findings indicated that PCA, PLS-DA, and SVM achieved classification rates of 90%, 92.9%, and 98.2%, respectively. Dos Santos et al. (2021) used a similar mechanism by combining ten color histograms and ANNs to classify raw sugarcane. They achieved 100% accurate classification for two ranges of raw sugarcane in the samples, from 90 to 100 wt% and from 41 to 87 wt%. However, these studies mentioned above recognize raw sugarcane and impurities based on their color features, making it difficult to differentiate objects with inter-class similarity, such as sugarcane top and leaf, which have similar color features at the pixel level. Additionally, these methods may not be suitable for practical situation with multiple combinations of impurities in arbitrary proportions, which present significant challenges in building samples with a vast combination of weight percentages of impurities.

From the perspective of recognition tasks, the aforementioned studies can be categorized into three types: image classification, object detection, and semantic segmentation. Image classification-based approaches (Momin et al., 2017; Guedes and Pereira, 2019; Shen et al., 2019; Aparatana et al., 2020; Chen et al., 2020; Guedes et al., 2020; Dos Santos et al., 2021; Li et al., 2022) cannot capture pixel-level information for subsequent construction of a mass-pixel fitting model. Object detection can be utilized for real-time classification and localization of crops and impurities (Zhang et al., 2022; Xu et al., 2023; Yu J. et al., 2023), but they still cannot support subsequent mass estimation based on pixels of detected objects. Semantic segmentation, on the other hand, enables pixel-wise classification of an image and facilitates the precise determination of the number of pixels and their respective categories in a specific region. Mass-pixel fitting models can be established by combining the number of pixels and the actual mass of each category of object (Rong et al., 2020; Chen et al., 2022; Jin et al., 2022; Liu et al., 2022; Liu et al., 2023), thus supporting the quantitative analysis of the quality of the detected objects. In order to quantify the ratio of breakage and impurity in raw sugarcane, semantic segmentation technology was utilized to abstract the of raw sugarcane and impurities in this study. However, the aforementioned approaches and findings are difficult to be directly applied to the detection of sugarcane and impurities in this study. Firstly, there is currently a lack of image databases that include raw sugarcane and impurities. Secondly, the estimation models developed in the above studies are only suitable for relatively stable scenarios of surface density (mass/pixel) for each detection category. However, the surface density of broken cane varies significantly due to different degrees of breakage, and the residual leaf at the top of the cane is scattered, resulting in a more varied surface density. Therefore, it is necessary to establish a corresponding image dataset and segmentation model for the detection of raw sugarcane and impurities and build new estimation model for quality evaluation based on segmented pixels.

Popular and widely applied deep learning (DL)-based semantic segmentation approaches have achieved excellent results in image processing in agriculture (Luo et al., 2023). Among these approaches, end-to-end semantic segmentation models like FCN, UNet, PSPNet, and DeepLabv3+ have demonstrated good performance with simple structures. DeepLabv3+ in particular has gained significant popularity and has been extensively enhanced due to its exceptional segmentation accuracy, making it a widely practiced and verified model in agricultural applications. For instance, Wu et al. (2021) developed an enhanced version of DeepLabv3+ to segment abnormal leaves in hydroponic lettuce. Peng et al. (2023) constructed an RDF-DeepLabv3+ for segmenting lychee *stem*. Zhu et al. (2023) proposed a two-stage DeepLabv3+ with adaptive loss for the segmentation of apple leaf disease images in complex scenes. Wu et al. (2023) utilized Deeplabv3+ and post-processing image analysis techniques for precise segmentation and counting of banana bunches. Their findings indicated that DeepLabv3+-based segmentation models can effectively perform pixel-level segmentation of crop objects, and the segmentation effects were superior to those of compared approaches. In this study, DeepLabv3+ was adopted for the semantic segmentation of raw sugarcane and impurities, and efforts were made to further improve its segmentation accuracy, reduce parameters, and optimize inference time.

This study aims to address the detection of breakage and impurity ratios in raw sugarcane. The specific research content of this study includes: (1) Designing a machine vision-based acquisition platform for online image collection of raw sugarcane (cane, broken cane) and impurities (top, leaf). Custom datasets of masses and corresponding images were constructed. (2) Establishing a normal fitting model to determine the mean surface density of each class based on measured masses and extracted pixels. Additionally, an estimation model was developed to assess the ratios of breakage and impurity using the estimated mass of each class, along with their pixels and fitted mean surface density. (3) Developing a MDSC-DeepLabv3+ model for accurate segmentation of raw sugarcane and impurity pixels based on DeepLabv3+. The model was further improved by incorporating improved MobileNetv2, atrous spatial pyramid pooling (ASPP) with deepwise separable convolution (DSC) and strip pooling (SP) named ASPP_DS, and coordinate attention (CA) mechanism to enhance segmentation accuracy, reduce parameters, and optimize inference time. (4) Conducting experiments to verify the accuracy of the proposed estimation model in assessing breakage and impurity ratios, and evaluate the capability of MDSC-DeepLabv3+ in rapidly and accurately identifying the pixels of cane, broken cane, top, and leaf. Comprehensive experimental results show that the average relative errors of breakage and impurity ratio between predicted values and measured values are low. These findings have significant implications for the development of intelligent detection and cleaning system for sugarcane impurity.

2 Materials and methods

2.1 Raw sugarcane and impurity dataset construction

2.1.1 Detection device design

In order to provide a stable environment and meet the continuous image acquisition requirements that align with the raw sugarcane convey process in the sugar refinery, a dedicated platform for image acquisition of raw sugarcane and impurities was designed, as shown in Figure 1A. The platform mainly consists of portable energy storage, an acquisition room, a light source, an image acquisition module, a computer, and a motion assistance module.

The portable energy storage is used to supply power to the platform, especially in situations where electricity supply is limited. The interior of the image acquisition room, as depicted in Figure 1B, is covered with black matte paper to create a diffused lighting environment. Additionally, four magnetic base LED light bars are strategically placed around the room to ensure consistent illumination for the image acquisition module. The image acquisition module comprises an industrial camera and an industrial lens. The computer is connected to the image acquisition module via a USB 3.0 interface, which facilitates image storage and processing. The motion assistance module is composed of a conveyor, a cross beam guide rail, and a pair of vertical slider guide rails with self-locking function. The conveyor simulates the transmission of raw sugarcane before entering the pressing workshop. The vertical slider guide rails, equipped with scale markings, support and allow for adjustment of the cross beam guide rail where the camera is mounted. This feature enables easy adjustment of the camera's field of view and ensures the stability of the image acquisition module.

Table 1 shows the model parameters of the main components of the acquisition platform. The conveyor belt speed is determined based on sugar refinery practice and is measured in meters per second (m/s). The dimensions of the indoor acquisition room are set according to the requirements, with horizontal (H_{FOV}) and vertical (V_{FOV}) dimensions are set to the belt width of 450mm and indoor length of 600mm, respectively. The selected industrial camera has a horizontal (H_{CMOS}) and vertical (V_{CMOS}) size of the image sensor as 7.6×5.7mm, and the working distance (W_D) is set to 490mm considering the inner height of the acquisition room. The imaging principle of this acquisition platform is illustrated in Figure 2. Using the imaging principle and the dimensions of H_{CMOS}, V_{CMOS} and W_D, the field of view can be determined using Eq.(1).

$$f/W_D = V_{CMOS}/V_{FOV} = H_{CMOS}/H_{FOV} \quad (1)$$

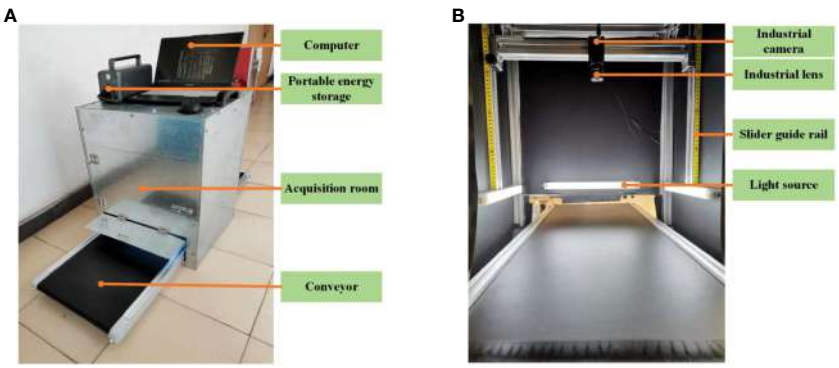


FIGURE 1
Machine vision acquisition platform. **(A)** Acquisition device structure. **(B)** Acquisition room.

TABLE 1 Main components of the acquisition platform.

Components	Parameters	Components	Parameters
Acquisition room	Indoor space 600mm×500mm×700 mm	Slider guide rail	SGR15N-500mm×2
Industry camera	MV-CA020-10UC with 89.1fps@1624×1240, image sensor size 7.6×5.7mm	Computer	AMD Ryzen7 5800H GeForce GTX 1650
Industry lens	MVL-MF0828M-8MP	Portable energy storage	72000mAh/3.2V
Light source	3600Lux×4	Conveyor	2000mm×450mm×100mm,1.5m/s, ≤20kg

As a result, the focal length is determined by $f = W_D \times (V_{CMOS}/V_{FOV}) = 490 \times (7.6/3450) = 8.27$mm, and the MVL-MF0828M-8MP industry lens is selected.

2.1.2 Image and mass data acquisition

The image and mass acquisition of raw sugarcane and impurities took place in the sugarcane unloading workshop of Junshi sugar refinery in Jijia Town, Leizhou City, Guangdong Province. The data collection period started from the middle of February to the end of the month in 2023, coinciding with the local sugarcane harvesting season. For this study, large-scale cultivated sugarcane variety "Yuetang 159" was selected. The raw sugarcane samples were randomly collected from different machine-harvested vehicles at various time intervals throughout the day using a loader. These samples were then manually placed on the conveyor belt of the acquisition platform for image collection. In total, 910 RGB 8-bit photos with jpg format and a resolution of 1624×1240 were captured. Each image contains four categories: cane, broken cane, top, and leaf, as shown in Figure 3. Following the image capturing process, 300 samples of raw sugarcane and impurities were randomly selected from the collected images. Each category of material in these samples was weighed using a calibrated electronic scale with a precision of 0.01g, and their masses were measured in grams (g).

2.1.3 Image labeling and dataset augmentation

The original dataset consists of 910 images containing cane, broken cane, top, leaf, and the background. These images were manually labeled and colored using the image annotation tool Labelme. The labeled regions of the five classes of objects were used to evaluate the training loss of intersection over union (IoU) between predicted bounding boxes and ground truth. The RGB values for cane, broken cane, top, and leaf were set to [128,0,0], [0,0,128], [0,128,0], and [128,128,0], respectively, while the background was set to [0,0,0]. To ensure model performance validation and testing, the dataset was randomly divided into training (546 images), validation (182 images), and test sets (182 images) with a ratio of 6:2:2.

In order to improve the generalization of the model, data augmentation techniques were applied to the training, validation, and test sets separately. Techniques such as random rotation, affine transformation, fogging, Gaussian noise, median filtering, and cutout were used to enhance the original images. After augmentation, the images were checked and corrected using Labelme to ensure accurate labeling of each class in every image. The annotated images were stored in the PASCAL VOC format and named Raw Sugarcane and Impurity (RSI). The label counting algorithm was used to calculate the number of labels in the RSI images, and the corresponding statistics are shown in Table 2. The dataset demonstrates a relatively balanced distribution of samples across each class. Examples of the original annotated images and augmented images can be observed in Figure 4.

2.2 Estimation model establishment

2.2.1 Surface density distribution analysis

In general, previous estimation models that are based on image pixels for assessing the mass of crops (such as wheat, corn, and soybean) often assume that the surface density (mass/pixel) of each crop category remains stable across different images (Chen et al., 2022; Jin et al., 2022; Liu et al., 2022). However, when it comes to broken cane and impurities, their surface density can vary significantly in different images. Therefore, before building the estimation model, it is essential to analyze the surface density distributions of cane, broken cane, top, and leaf separately. This analysis will help to account for the variation in surface density and ensure more accurate estimation for breakage and impurity ratios in raw sugarcane.

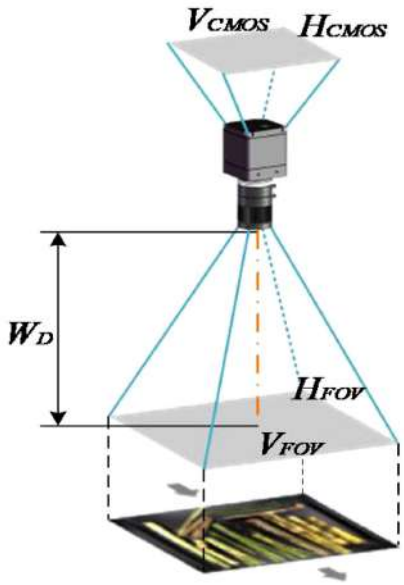


FIGURE 2
Imaging principle in this acquisition platform.

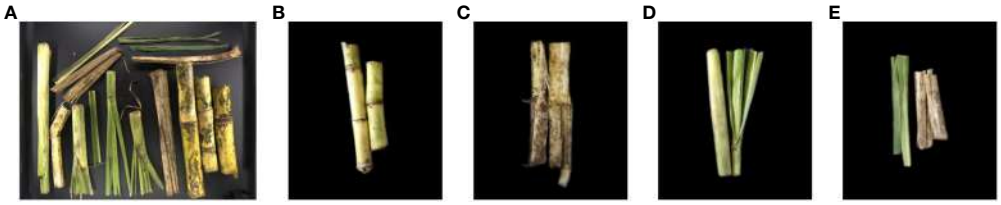


FIGURE 3
Acquisition materials and segmentation classes. **(A)** Original image, **(B)** Cane, **(C)** Broken cane, **(D)** Top, **(E)** Leaf.

The analysis of surface density distribution was conducted using 300 samples of mass data and the corresponding images for each category. The OpenCV threshold function was utilized to count the number of pixels in each category. Let P_C, P_B, P_T andP_L represent the number of pixels of cane, broken cane, top, and leaf in each image sample, respectively, and their corresponding masses are denoted as M_C, M_B, M_T and M_L, respectively. The spatial distribution of the surface density for raw sugarcane, including cane and broken cane, as well as the top and leaf, is presented in Figure 5. Based on the surface density distribution of raw sugarcane in Figure 5A, it can be observed that the surface density of cane fluctuates less and is more concentrated. The surface density of broken cane is approximately half of that of cane, and the data is scattered. Figures 5B, C illustrate that the surface density distribution of top and leaf is more scattered compared to broken cane.

To address the scattered surface density of broken cane, top, and leaf, a Gaussian distribution probability density function was used to fit the frequency histograms of surface density for each category. The mean surface density μ for each category was then obtained through the fitting process, and the results are demonstrated in Figure 6. It can be observed that all fitting coefficients R^2 are greater than 0.95, indicating high fitting accuracy.

The fitting results showed that the mean surface density of cane, broken cane, top, and leaf are μ_c = 1.52E-3, μ_B 7.4E-4, μ_T = 8.8E-4 and μ_L = 3E-5 with unit g/pix, respectively. Moreover, it is evident that the mean value of cane μ_c is approximately twice the mean value of broken cane μ_B and top surface density μ_T, and μ_c is more

TABLE 2 Statistic of Raw Sugarcane and Impurity (RSI) dataset.

Dataset	Training dataset	Validation dataset	Test dataset	Complete dataset
Images	5460	1820	1820	9100
Cane labels	16882	4151	3850	24883
Broken cane labels	13735	3310	3410	17045
Top labels	15903	4071	4390	24364
Leaf labels	17234	4015	3830	25079

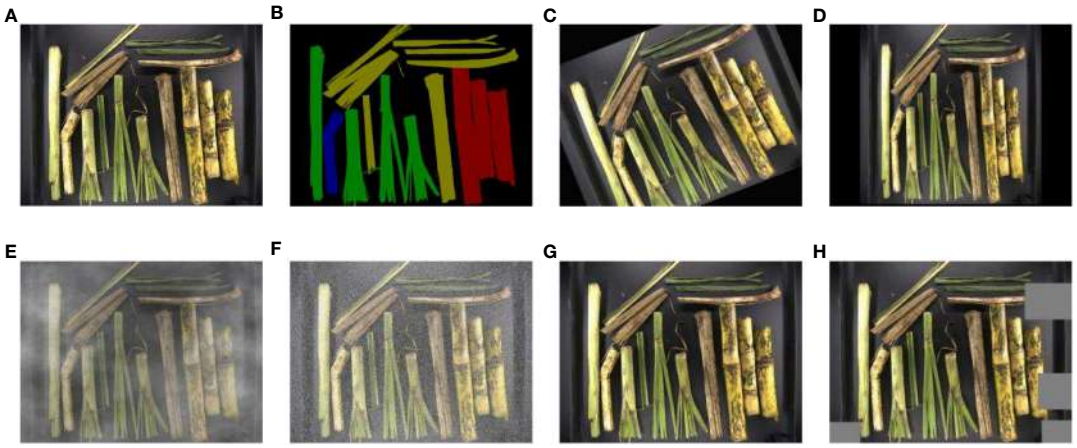


FIGURE 4
Augmented image samples and image label. **(A)** Original image, **(B)** Ground truth, **(C)** Random rotation, **(D)** Affine transformation, **(E)** Fogging, **(F)** Gaussian noise, **(G)** Median filtering, **(H)** Cutout.

10.3389/fpls.2023.1283230

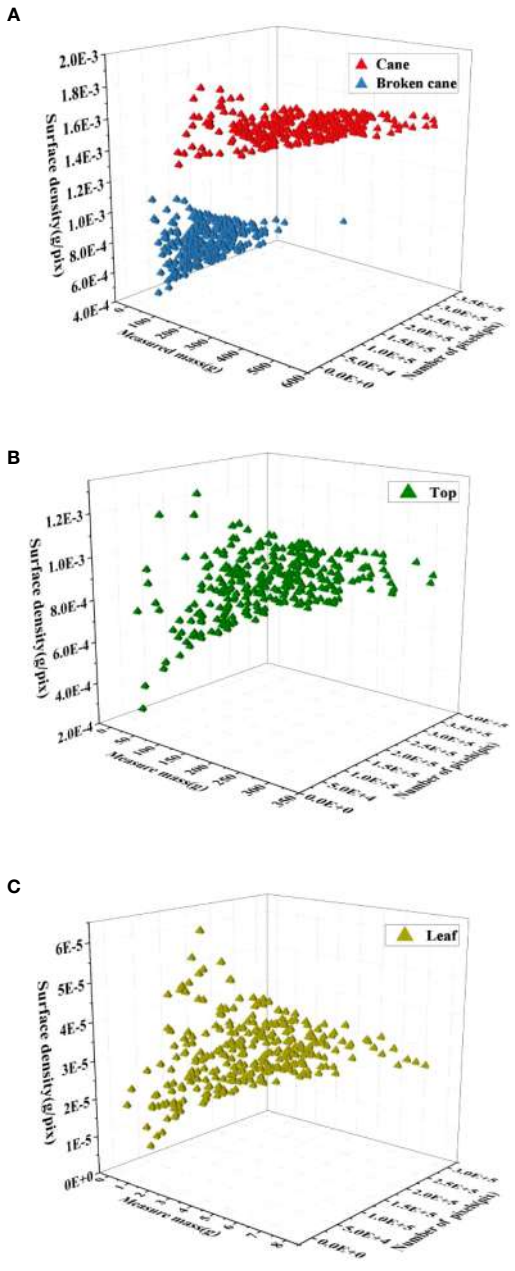


FIGURE 5
Spatial distribution of surface density for the 4 classes in RSI. **(A)** Raw sugarcane, **(B)** Top, **(C)** Leaf.

than fifty times *of* μ_L. The mass error of leaf has little effect on the overall mass error. Therefore, when establishing the estimation model, the accuracy of the estimated mass of cane should be ensured first, followed by broken cane, top, and finally leaf. This approach is consistent with the low deduction percentage setting (as low as 0.2%) employed by sugar refineries for leaf impurities.

2.2.2 Fitting and estimation model establishment

On the basis of the mean values of surface density given in Figure 6, the estimated mass of cane M'_C, broken cane M'_B, top M'_T, and leaf M'_L based on their pixels can be expressed as follows:

$$M'_C = \mu_C \times P_C = 1.52\text{E} - 3P_C \quad (1)$$

$$M'_B = \mu_B \times P_B = 7.4\text{E} - 4P_B \quad (2)$$

$$M'_T = \mu_T \times P_T = 8.8\text{E} - 4P_T \quad (3)$$

$$M'_L = \mu_L \times P_L = 3\text{E} - 5P_L \quad (4)$$

Furthermore, a linear regression of the estimated and measured mass was conducted to validate the accuracy of the mass estimation model defined by Eq.(1)-(4). Based on the distribution characteristics shown in Figure 6, a total of 285 mass data of cane, broken cane, top, and leaf within a 95% confidence interval were selected for fitting, and the fitting results were presented in Figure 7 and Table 3. It can be seen that the measured mass of the cane is highly correlated with the estimated mass with an R^2 value of 0.983. This indicates that the linear regression model is capable of explaining the numerical relationship between the measured mass and the estimated mass of the cane. The R^2 value for broken cane and top are 0.894 and 0.88, respectively, demonstrating the regression model's good fitting capability. The R^2 value for the leaf is 0.764 suggesting that the model can still adequately fit the relationship between the measured mass and the estimated mass. In addition, the results of ANOVA in Table 3 indicate that the significance $F<0.01$ between estimated cane, broken cane, top, and leaf and their measured values proves a high correlation.

Based on the mass of each category, the ratios of breakage (R_B) and impurity (R_I) *is* defined as:

$$\begin{aligned} R_B &= \frac{M_B}{M_C+M_B} \times 100\,\% \\ &= \frac{7.4E{-}4 \times P_E}{1.52E{-}3\times P_c+7.4E{-}4\times P_E} \times 100\,\% \end{aligned} \quad (5)$$

$$\begin{aligned} R_I &= \frac{M_T + M_L}{M_C+M_B+M_T+M_L} \times 100\,\% \\ &= \frac{8.8E{-}4\times P_T + 3E{-}5\times P_L}{1.52E{-}3\times P_c+7.4E{-}4 \times P_E+8.8E{-}4\times P_T + 3E{-}5\times P_L} \times 100\,\% \end{aligned} \quad (6)$$

Where M_C, M_B, M_T and M_L is the mass of cane, broken cane, top and leaf in an image sample. The estimated breakage and impurity ratios R'_B and R'_I can also be determined by replacing M_C, M_B, M_T and M_L in Eq.(5)-(6) with estimated mass M'_C, M'_B, M'_T and M'_L. Thereby Eq.(5)-(6) can be taken as the estimation model for breakage and impurity ratios.

2.3 Raw sugarcane and impurity segmentation model development

2.3.1 MDSC-DeepLabv3+ framework

In order to facilitate the M'_C, M'_B, M'_T, M'_L, R'_B and R'_I calculation, a segmentation model, MDSC-DeepLabv3+, was developed for the intelligent extraction of pixels of cane P_C, broken cane P_B, top P_T, and leaf P_L in each image sample. MDSC-DeepLabv3+ is an improvement upon the DeepLabv3+. The DeepLabv3+ comprises two modules: an encoder and a decoder (Chen et al., 2018). In the encoder, the Xception

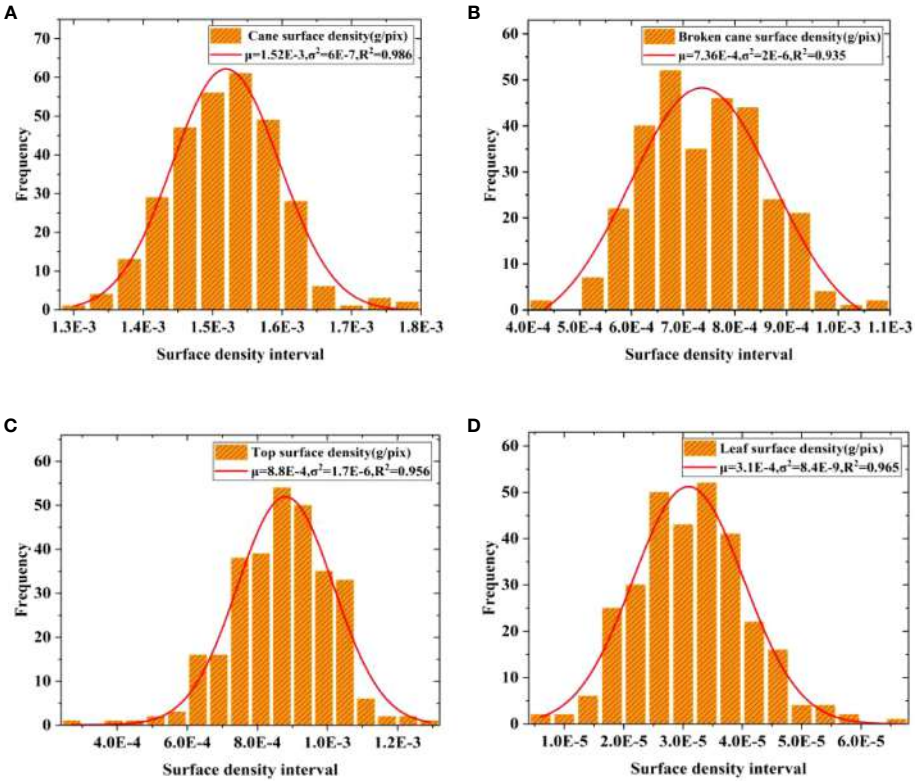


FIGURE 6
Gaussian distribution fitting of surface density. **(A)** Cane, **(B)** Broken cane, **(C)** Top, **(D)** Leaf.

backbone is used to extract input image features, resulting in two effective feature maps. One of the feature map undergoes processing through atrous spatial pyramid pooling named ASPP, and is then using a 1×1 standardization convolution for the fused features from ASPP. This produces high-level features that are subsequently fed into the decoder. The other feature map directly outputs to the decoder. The ASPP is composed of a 1×1 standardization convolution, three 3×3 depthwise separable convolutions named DSC with varying dilation rates (6, 12, and 18), and an average pooling layer. These convolutions generate feature maps at four different scales, which are stacked along the channel dimension.

In the decoder, the low-level features obtained from the Xception backbone first undergo 1×1 convolution to reduce the number of channels. Meanwhile, the high-level features from the encoder are bilinearly upsampled by a factor 4 to improve the image resolution. Afterwards, the 1×1 convoluted low-level features are fused with the upsampled high-level features, and a 3×3 DSC is utilized to extract information from the fused features, followed by another bilinear upsampling by a factor 4. Previous studies have demonstrated the effective use of DeepLabv3+ in agricultural fields, such as fruit picking, crop disease and pest, and field road scenes (Wu et al., 2021; Peng et al., 2023; Yu J. et al., 2023).

To enhance both the accuracy and deployability of the model, as well as reduce inference time, various improvements including improved MobileNetv2, ASPP_DS module and CA mechanism were introduced in this study. First, the atrous convolution was employed to optimize the MobileNetv2, and Xception was replaced by the improved MobileNetv2 in DeepLabv3+. In the MobileNetv2, dilated convolution was incorporated into the last two layers by increasing the kernel size, thus expanding the receptive field. This enhancement allows the network to better perceive surrounding information without significantly increasing computational complexity or compromising the resolution of the feature maps. Then, the dilation rates in the ASPP module were adjusted as 4, 8, and 12, and a strip pooling layer was added parallel to DSC to build a module named ASPP_DS. Module ASPP_DS can reduce the model parameters and establish long-range dependencies between regions distributed discretely, and focus on capturing local details. ASPP employs diverse padding and compact dilation strategies to extract receptive fields at various scales, effectively capturing information from both multi-scale contexts and small objects. Additionally, ASPP integrates a parallel strip pooling layer with elongated and narrow pooling kernels to grasp local contextual details in both horizontal and vertical spatial dimensions. This approach helps in reducing interference from unrelated regions in label prediction results. Finally, CA was appended to the output of MobileNetv2 and ASPP_DS separately, that allows the model to acquire weight information from the dimensions of feature channels and effectively leverage positional data. This incorporation enables the accurate capture of spatial relationships and contextual information of the target, thereby enhancing training efficiency. The enhanced version of DeepLabv3+ is denoted as MDSC-

10.3389/fpls.2023.1283230

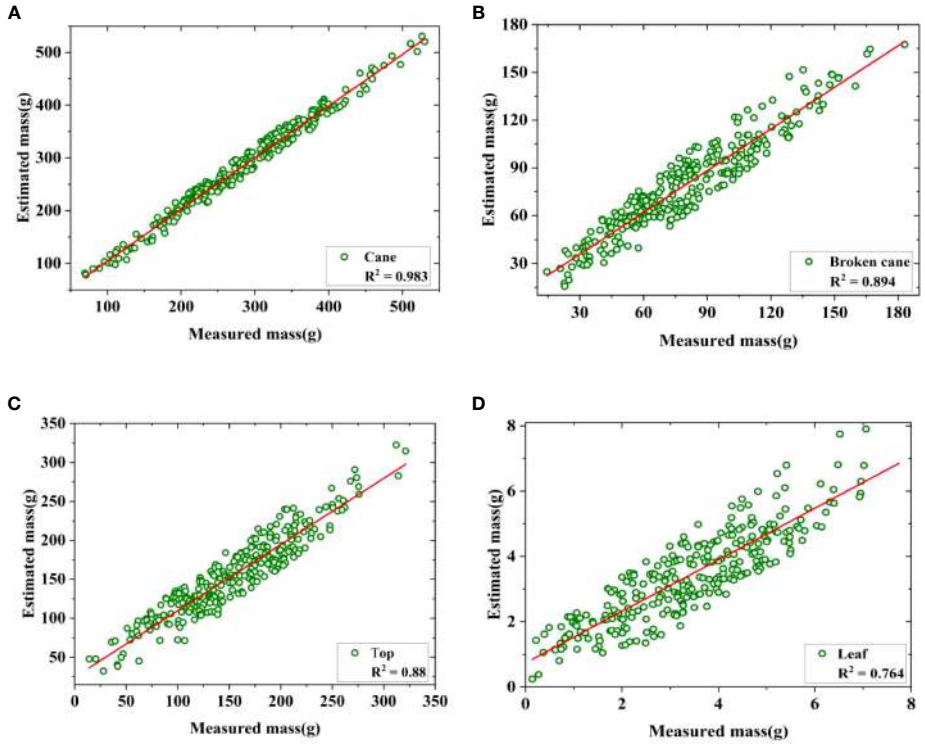


FIGURE 7
Regression of estimated and measured mass. **(A)** Cane, **(B)** Broken cane, **(C)** Top, **(D)** Leaf.

DeepLabv3+. The overall framework of MDSC-DeepLabv3+ is depicted in Figure 8.

2.3.2 Improved MobileNetv2

The basic structure unit of MobileNetv2 is the inverted residual block (IRB), which mainly consists of dimensionality expansion, feature extraction and dimensionality compress three main steps. The MobileNetv2 employs 3×3 depthwise convolution (Dwise) and 1×1 convolution to construct two IRBs with s= 1, s=2 (Sandler et al., 2018). In cases where the stride is equal to 1 and the shape of the input feature matrix matches that of the output feature matrix, a shortcut connection is employed, as shown in Figure 9. In addition, the dimensionality compression process in MobileNetv2 uses a linear activation function instead of the Relu activation function to reduce information loss caused by compression.

TABLE 3 Analysis of Variance (ANOVA) of estimated and measured mass. .

Category		DF	Square sums	Mean square	F	Significance F
Cane	Regression analysis	1	2340192.15697	2340192.15697	16820.25846	4.23041E-254
	Residual	283	39373.61497	139.12938		
	Total	284	2379565.77194			
Broken cane	Regression analysis	1	225202.9665	225202.9665	2390.43448	4.97988E-140
	Residual	283	26661.44583	94.21006		
	Total	284	251864.41233			
Top	Regression analysis	1	656015.70993	656015.70993	2055.21929	8.58987E-132
	Residual	283	90332.18347	319.19499		
	Total	284	746347.8934			
Leaf	Regression analysis	1	431.20971	431.20971	915.53104	1.07792E-90
	Residual	283	133.29133	0.47099		
	Total	284	564.50104			

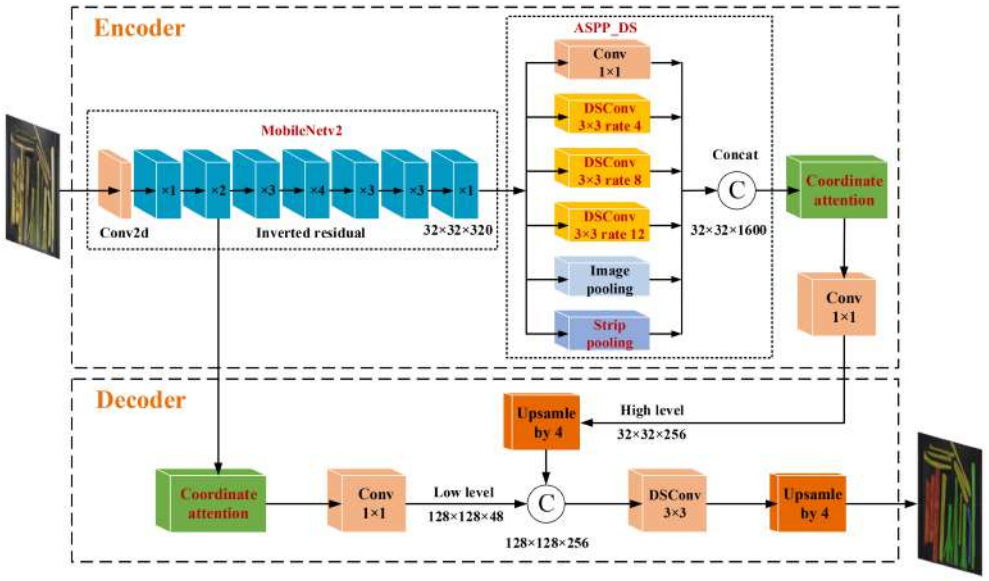


FIGURE 8
Framework of MDSC-DeepLabv3+.

To reduce computing costs and memory usage, this study utilizes the first 8 layers of the MobileNetv2 model. This choice is made because starting from the 9th layer, the number of output channel increases to 1280, leading to higher computing resource consumption. To minimize the loss of down-sampling information while increasing receptive field, the stride of the 7th layer is modified to 1 (Meng et al., 2020).

Furthermore, dilated convolutions with a factor not exceeding 1 are utilized to replace conventional convolutions. According to research by Wang et al. (2018), sparse concatenation of dilated convolution may introduce grid effects, hindering the lower layers of the network from fully leveraging features from the initial layer and causing the loss of fine-grained details. Therefore, dilation rates of 2 and 5 are applied in the 7th and 8th layer respectively, while the remaining layers maintain a dilation rate of 1, aiming to expand the receptive field and preserve edge detail information. The structure and hyperparameter of the improved MobileNetv2 are displayed in Table 4, in which t is the expansion factor, c is the output channel, n is the number of repetitions of bottleneck, s is the first module strid, and r is dilation rate. When dilation rate of 1 results in atrous convolution being equivalent to a regular convolution. This design achieves a balance between computational resource consumption and network performance requirements.

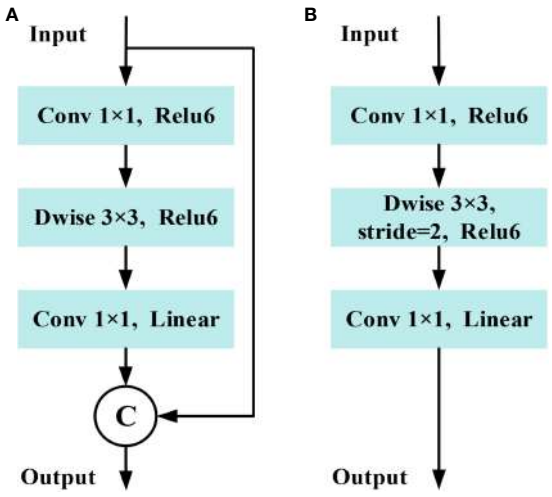


FIGURE 9
Structure of inverted residual block in MobileNetv2. **(A)** Stride=1 block. **(B)** Stride=2 block.

2.3.3 Strip pooling

To better handle the segmentation of broken cane and top with irregular and complex shapes, a lightweight strip pooling layer was added in parallel to DSC in the ASPP. This allows for more efficient acquisition of information from a large receptive field, facilitating the collection of remote contextual information from different spatial dimensions by ASPP. Strip pooling utilizes a pooling kernel (rectangular area) that performs pooling operations along the horizontal and vertical dimensions. The structure of strip pooling (Hou et al., 2020) is shown in Figure 10, where X ϵ $R^{C\times H\times W}$ is the input tensor, C denotes the number of channels, H denotes the height, and W denotes the width. First, the input X is pooled horizontally and vertically to obtain y^h ϵ $R^{C\times H\times 1}$ and y^v ϵ $R^{C\times 1\times W}$, respectively. Then, the feature maps are expanded to the same resolution C×H×W as the input X using a 1D convolution with a kernel size of 3×3 to obtain the expanded y^h, y^v. Next, the expanded feature maps are fused to obtain a final representation.

$$y_{c,i,j} = y^h_{c,i} + y^v_{c,j}, 1 \le c \le C, 1 \le i \le H, 1 \le j \le W$$

Finally, after a 1 × 1 standard convolution and a sigmoid layer, the final output Z of strip pooling is obtained by multiplying the corresponding elements with the original input.

$$Z = Scale(X, \sigma(f(y)))$$

where Scale (-, -) is the element-level multiplication, σ is the sigmoid function, and f is the 1×1 convolution, y is feature fusion results.

The element of specified location in the output tensor (i, j),$1 \le i \le H$,$1 \le j \le W$ corresponds to the result of strip pooling of the horizontal and the vertical pooling window in the input tensor. By repeatedly applying the aggregation process using long and narrow pooling kernels, the ASPP_DS module can efficiently capture information from a wide receptive field throughout the entire scene. Due to the design of the elongated and narrow shape of the pooling kernel, it not only establishes remote dependency relationships between regions distributed discretely but also focuses on capturing local detailed features.

2.3.4 Coordinate attention

Inspired by the prominence of the region-of-interest search in the human visual system, attention mechanisms aim to simulate this process by dynamically adjusting the weights based on the input image features. Attention mechanisms can be categorized into various types, such as channel attention (e.g. SE), hybrid attention (e.g. CBAM), temporal attention (e.g. GLTR), branch attention (e.g. SKNet), and position attention mechanisms (e.g. CA). These attention mechanisms have been widely applied in fields such as object detection (Yu J. et al., 2023) and image segmentation (Zhu et al., 2023).

TABLE 4 Hyperparameters of MobileNetv2.

Input size	Operator	t	c	n	s	r
512×512×3	conv2d	–	32	1	2	1
256×256×32	bottleneck	1	16	1	1	1
256×256×16	bottleneck	6	24	2	2	1
128×128×24	bottleneck	6	32	3	2	1
64×64×32	bottleneck	6	64	4	2	1
32×32×64	bottleneck	6	96	3	1	1
32×32×96	bottleneck	6	160	3	1	2
32×32×160	bottleneck	6	320	1	1	5

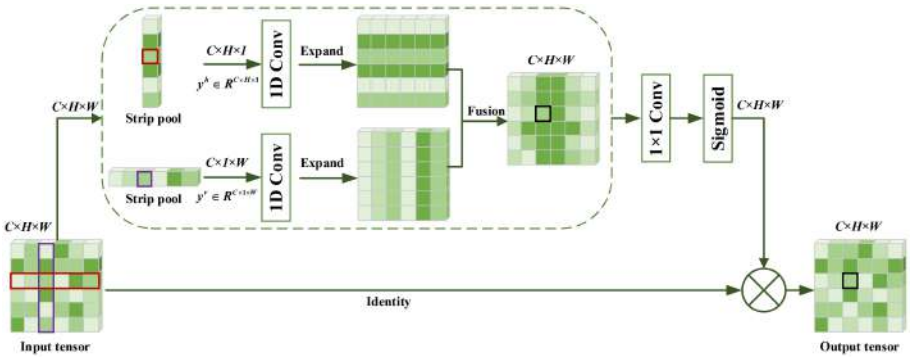


FIGURE 10
Structure of strip pooling.

The CA not only models channel relationships but also utilizes positional information to capture long-range dependencies (Hou et al., 2021). Therefore, CA was selected in the MDSC-DeepLabv3+ to highlight the regions of interest. The CA consists of coordinate information embedding (CIE) and coordinate attention generation (CAG) two main operation, as shown in Figure 11. CIE introduces two global average pooling to encode each channel along the horizontal and vertical coordinate on the input feature map, respectively, hence aggregates features along the two spatial directions. These two pairs of global average pooling operation enable CA to capture long-range dependencies along one spatial direction and preserve precise positional information along other one, which allows the network to more precisely locate the objects of interest. CAG first conducts concatenation (Concat) and Conv2d for the feature maps obtained from CIE followed by batch normalization and non-linear activation operation. Then, the intermediate feature map is split into two separate tensors along the spatial dimension. Next, 1×1 Conv2d and sigmoid activation are utilized to separately transform the output tensors to tensors with the same channel number as the input feature maps. Finally, the output tensors are then expanded into elements and used as attention weights. The final output of CA is the element-wise multiplication of original input of CIE and the attention weights.

Introduction of CA before low feature processing and after the features fusion of ASPP_DS is beneficial in fully utilizing positional information. This allows the model to accurately capture the spatial relationships and contextual information of the target, thus improving the accuracy of sugarcane and impurity phenotype segmentation in denser images.

3 Experiments and results

3.1 Analyzing of estimation model

The effectiveness of estimation model for breakage and impurity ratios defined in Section 2.2.2 was validated by fitting estimated and measured value. First, the measured mass of cane, broken cane, top, and cane leaf M_C, M_B, M_T and M_L, along with the number of pixels for each category manually labeled in the selected 285 images (95% confidence interval of samples) were obtained. Then, estimated masses of M'_C, M'_B, M'_T and M'_L for the four categories were determined based on the mean surface density μ_C, μ_B, μ_T and μ_L according to Eq.(1)-(4). Next, the measured and estimated ratios of breakage and impurity were obtained according to Eq.(5)-(6) based on the measured and estimated masses. Finally, the measured breakage and impurity ratios were linearly fitted with the estimated breakage and impurity ratios, and the fitting results are shown in Figure 12 and Table 5, respectively.

It can be observed that the fitting R^2 values are as high as 0.976 and 0.968, respectively. In addition, the results of the ANOVA presented in Table 5 indicate a high correlation between the estimated breakage and impurity ratios and their measured values, with a significance level of $F<0.01$. Therefore, it is feasible to utilize the fitted surface density to estimate mass for each category and furthermore predict the breakage and impurity ratios for raw sugarcane.

3.2 Analyzing of segmentation model

3.2.1 Training environment and evaluation metrics

The semantic segmentation categories considered in this study are background, cane, broken cane, top, and leaf. In the process of sugarcane harvesting, raw sugarcane is primarily composed of cane, with cane tops and leaves present as impurities to a lesser extent. Broken cane represents the category with the lowest representation, leading to an extreme class imbalance. Consequently, this often leads to imbalanced positive and negative samples, along with varying sample difficulties. Therefore, this study utilizes the Focal Loss function as the primary loss function to address the imbalance between easy and difficult samples, facilitating better parameter optimization during the backpropagation process (Lin et al., 2017). In addition, the model incorporates the multi-class Dice Loss as an auxiliary loss function to enhance segmentation accuracy and address class imbalance scenarios (Milletari et al., 2016). The combination of Focal Loss and multi-class Dice Loss as the loss function enhances the model's predictive capability. The Focal loss for multi-objective segmentation is defined as.

$$L_F = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

Where p_t is the confidence value of the sample category prediction. γ is an adjustable parameter, and the default is 2.

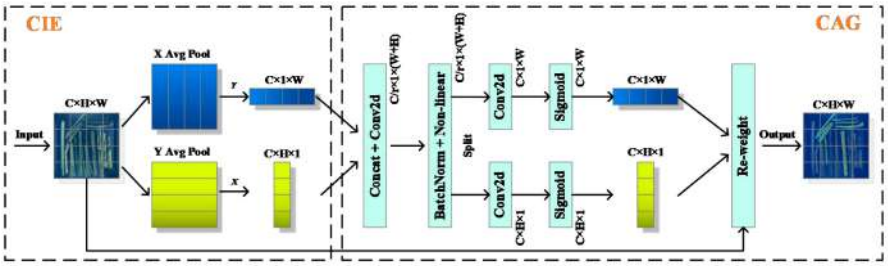


FIGURE 11
Structure of coordinate attention.

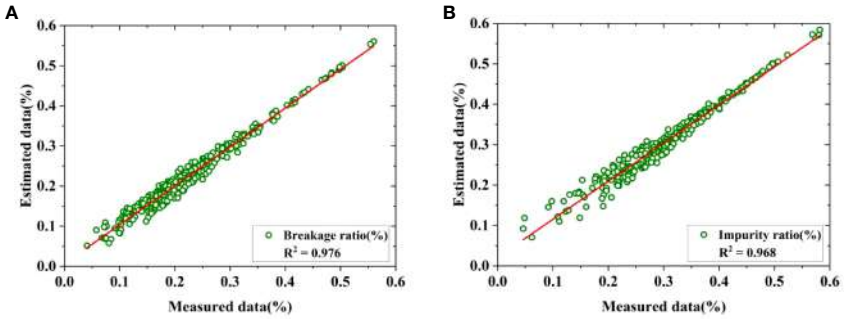


FIGURE 12
Fitting of estimated and measured ratio. **(A)** Breakage ratio, **(B)** Impurity ratio.

The Dice loss for multi-objective segmentation is defined as.

$$L_D = 1 - \sum_{j=1}^{c} \frac{2W_j \sum_{i=1}^{N} gt(j,i)\log(p_{i,j})}{\sum_{i=1}^{N}\left(gt(j,i)^2 + \log(p_{i,j})^2\right)}$$

Where, N is the number of samples, c is the target class, and $p_{i,j}$ is the softmax output of class j target class; $gt(j,i)$ is the ground-truth label of class j target, and W_j is the weight of the objective of class j, $W_j = 1/j$.

The experiments were conducted on a server in the lab with the configuration shown in Table 6. The MDSC-DeepLabv3+ used the Adam optimizer to compute the gradient of the loss function in each epoch to perform parameter updates. The initial learning rate was set to E-4. The batch size was set to 6. The training process consists of 100 epochs. In each epoch, the image dataset was randomly shuffled and fed into the model to ensure a different order of dataset used in different epochs. This technique enhances the convergence speed of the model and improves the prediction results on the test set.

In order to comprehensively evaluate the performance of the proposed and comparative semantic segmentation models, three aspects of each model, namely accuracy, deployability, and efficiency, are comprehensively evaluated. The commonly used mIoU and mPA were utilized as accuracy evaluation metrics. And the model deployability was evaluated using model parameter quantity (Param) and model computation volume floating point operations (FLOPs). Efficiency was evaluated using inference time for each image. The metrics of IoU, mIoU and mPA which is represented by the following Eq. (7)-(9), respectively.

$$IoU_i = \frac{P_{ii}}{\sum_{i=0}^{c-1} P_{ij} + \sum_{j=0}^{c-1} P_{ji} - P_{ii}} \times 100\,\% \tag{7}$$

$$mIoU = \frac{1}{c}\sum_{i=0}^{c-1} IoU_i \tag{8}$$

$$mPA = \frac{1}{c}\sum_{i=0}^{c-1} \frac{p_{ii}}{\sum_{j=0}^{c-1} p_{ij}} \tag{9}$$

Where c denotes the number of categories, so c=4 (cane, broken cane, top and leaf), P_{ij} or P_{ji} denotes the number of category prediction that is incorrect, while P_{ii} denotes the number of correct predictions made by categories.

3.2.2 Model training

The size of the input image is a crucial factor affecting the model's performance. Increasing the image size enhances accuracy by preserving semantic information for small targets and preventing information loss caused by low-resolution feature maps. However, excessively large image sizes can lead to reduced detection accuracy due to the limited receptive field imposed by the fixed network structure. This, in turn, diminishes the network's ability to accurately predict targets of various scales (Lin et al., 2022). In practical applications, there is a trade-off between accuracy and speed that requires careful consideration. For this study, the input image was resized to three different dimensions: 256×256, 512×512, and 768×768. The proposed MDSC-DeepLabv3 + model was trained accordingly, and the results obtained are presented in Table 7. It can be observed that reducing the input

TABLE 5 ANOVA of breakage and impurity ratios.

Ratio		DF	Square sums	Mean square	F	Significance F
Breakage ratio	Regression analysis	1	2.58018	2.58018	11405.03085	1.05518E-230
	Residual	283	0.06402	2.26232E-4		
	Total	284	2.64421			
Impurity ratio	Regression analysis	1	2.41267	2.41267	8470.24579	6.21725E-213
	Residual	283	0.08061	2.84841E-4		
	Total	284	2.49328			

TABLE 6 Experimental environment.

Parameter	Configuration	Parameter	Configuration
Operating system	Ubuntu 18.04	Operating environment	CUDA 11.2
Deep learning framework	PyTorch 1.8	CPU	Intel(R) Xeon(R) Silver 4214 CPU @2.20GHz
Programming Language	Python 3.7	GPU	NVIDIA GeForce RTX 3080 12G @1260-1710MHz

image size to 512×512 achieves an optimal balance between speed and accuracy.

The segmentation results of models using different loss functions are displayed in Figure 13. The MDSC-DeepLabv3+ using only the Dice loss function exhibits the highest fluctuations in mPA and mIoU, leading to inferior segmentation results. Similarly, the MDSC-DeepLabv3+ using only Focal Loss demonstrates notable fluctuations during the early stages of the validation process, with slow growth in mPA and mIoU values in later stages. In contrast, the MDSC-DeepLabv3+ which combines Focal Loss and multi-class Dice Loss exhibits lesser sawtooth fluctuations during the increase in mPA and mIoU values, ultimately reaching their peak during the validation process. Consequently, the integration of Focal Loss and multi-class Dice Loss yields optimal outcomes in the segmentation of raw sugarcane and impurities.

3.2.3 Ablation experiment

To verify the effectiveness of the three improvements, including improved MobileNetv2, ASPP_DS and CA presented in Section 2.3, the following 7 models were constructed according to the control variable method, with a downsampling factor of 8.

TABLE 7 mPA and inference time obtained with different input image sizes.

Resize of image/pixels	mPA/%	Inference time/ms
256×256	94.68	10.69
512×512	97.55	13.85
768×768	97.07	24.19

1. DeepLabv3+_base: MobileNetv2 replaced the backbone Xception in DeepLabv3+.
2. M-DeepLabv3+: MobileNetv2 in DeepLabv3+_base was enhanced with atrous convolution operation.
3. MDS-DeepLabv3+: ASPP_DS replaced ASPP module in M-DeepLabv3+.
4. MC1-DeepLabv3+: CA was applied independently before 1×1 Conv of low-level features by the decoder in M-DeepLabv3+.
5. MC2-DeepLabv3+: CA was applied independently after the fusion of ASPP in M-DeepLabv3+.
6. MC-DeepLabv3+: CA was added separately before 1×1 Conv the low-level features and after the fusion of ASPP features in M-DeepLabv3+.
7. MDSC-DeepLabv3+: CA was added separately before processing the low-level features and after the fusion of ASPP_DS features in MDS-DeepLabv3+.

Table 8 presents the results of the ablation experiment for the seven aforementioned models. It can be observed that the MDSC-DeepLabv3+ outperforms the baseline DeepLabv3+_base, with an improvement of 1.25 in mPA and 1.8 in mIoU. Additionally, it achieves a reduction of 16.42% in Params and 31.46% in FLOPs, however, the inference time per image has slightly increased from 13.48ms to 13.85ms. These results demonstrate that the MDSC-DeepLabv3+ surpasses the DeepLabv3+_base in terms of segmentation accuracy and deployability metrics, while still maintaining comparable efficiency. Furthermore, it can be seen that the MDSC-DeepLabv3+ achieves the highest segmentation accuracy (mPA and mIoU) compared to other models, while exhibiting minimal differences in terms of deployability (Params, FLOPs) and efficiency (inference time) metrics.

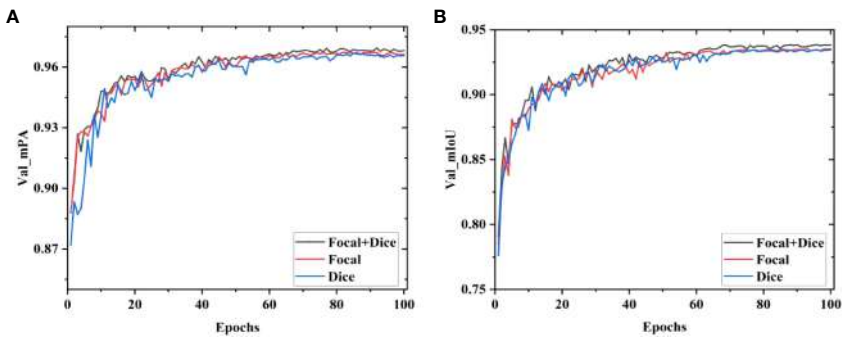


FIGURE 13
Results of mPA and mIoU with different loss functions. **(A)** Valid mPA, **(B)** Valid mIoU.

In order to visually demonstrate the improvement of the models, Grad-CAM (Selvaraju et al., 2020) was used to visualize the channels of the feature maps of DeepLabv3+ and MDSC-DeepLabv3+. The visualization segmentation instances of top were illustrated in Figure 14. In group (a), the two feature maps are extracted by the Xception in DeepLabv3+ and the enhanced MobileNetv2 in MDSC-DeepLabv3+, respectively. In group (b), two feature maps are the output of ASPP in DeepLabv3+ and ASPP_DS in MDSC-DeepLabv3+, respectively. In group (c), the two feature maps are the output of DeepLabv3+ and MDSC-DeepLabv3+, respectively.

In Figure 14A, it can be observed that Xception in DeepLabv3+ achieves clearer pixel segmentation than that obtained by MobileNetv2 in MDSC-DeepLabv3+. The reason is that MobileNetv2 is a lightweight and shallow model compared to Xception, and its depthwise convolution can lead to information loss and limit the number of channels, thereby resulting in a lower-level feature map with fewer information. However, the two heat maps in group (b) indicate that there is pixels misfocus at the top-right corner in the first line of the feature map extracted by ASPP, while ASPP_DS results in more complete pixel segmentation, enhances preservation of details, and eliminates the top-right misfocus. The heat map illustrates that the introduced strip pooling in ASPP_DS rectifies the shortage of MobileNetv2, and the dense and compact dilation rates (4, 8, 12) improve its capability of focusing on capturing local detailed features. Heat map of final outputs of MDSC-DeepLabv3+ and DeepLabv3+ given in Figure 14C demonstrates that the CA in MDSC-DeepLabv3+ further enhances the color intensity in heat map, indicating that the inclusion of CA allows the model to focus more on the features of the categories, thereby enhancing its distinguishability of cane, broken cane, top and leaf.

3.2.4 Comparative experiment

To further validate the superiority of the proposed model MDSC-DeepLabv3+, comparative experiments were conducted using the RSI dataset under the same experimental conditions. The compared models include UNet, PSPNet, SegFormer-B0, and the baseline DeepLabv3+. Previous research results have shown that UNet (Ronneberger et al., 2015) and PSPNet (Zhao et al., 2017) perform well in terms of accuracy in segmentation tasks with challenges like cell tracking ISBI and Cityscapes. SegFormer-B0 is

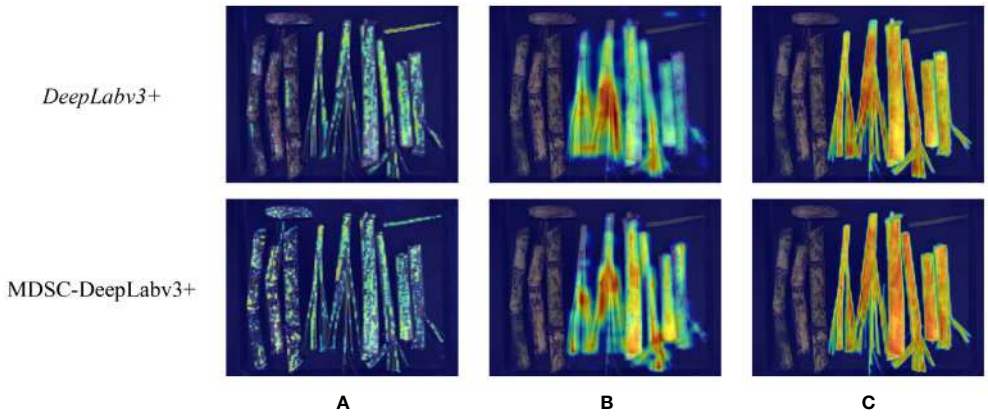


FIGURE 14
(A) Backbone output. **(B)** Encode output. **(C)** Decode output.

TABLE 8 Results of ablation experiment.

Number	ASPP_DS	Coordinate Attention		mPA/ %	mIoU/ %	Param/ M	FLOPs/ G	Inference time/ ms
		Before decoder	After ASPP (_DS)					
(1)				96.3	93.04	4.81	69.29	13.48
(2)				96.67	93.36	3.35	45.49	12.13
(3)	√			97.16	94.37	3.36	45.41	12.76
(4)		√		96.88	93.66	3.55	46.83	13.51
(5)			√	97.05	94.48	3.63	46.88	13.56
(6)		√	√	97.22	94.57	3.68	46.88	13.67
(7)	√	√	√	97.55	94.84	4.02	47.49	13.85

a lightweight model that combines transformers with a lightweight multilayer perceptron decoder (Xie et al., 2021). The comparative results are given in Table 9.

It can be seen that the accuracy of MDSC-DeepLabv3+ surpasses that of the aforementioned four models with significant improvements. Specifically, the mIoU of MDSC-DeepLabv3+ is higher by 0.81, 5.22, 12.47, and 0.28 compared to UNet, PSPNet, SegFormer-B0, and DeepLabv3+, respectively. Moreover, the mPA of MDSC-DeepLabv3 + reaches an impressive 97.55%, which outperforms UNet, PSPNet, SegFormer-B0, and DeepLabv3+ by 0.69, 2.7, 7.76, and 0.34, respectively. These remarkable improvements can be attributed to the adoption of the advanced DeepLabv3+ as the basic model, coupled with the enhancements introduced through strip pooling and CA. Strip pooling plays a crucial role in collecting remote contextual information from different spatial dimensions and addressing the issue of information loss resulting from the atrous convolution operation in DeepLabv3. On the other hand, CA efficiently utilizes positional information, enabling accurate capturing of the spatial relationships and contextual information of the detected cane, broken cane, top, and leaf.

In terms of deployability, MDSC-DeepLabv3+ demonstrates remarkable reductions in Params and FLOPs when compared to UNet, PSPNet, and DeepLabv3+. Specifically, it reduces Params by 83.65%, 91.29%, and 90.35%, and FLOPs by 89.49%, 59.9%, and 66.37% compared to UNet, PSPNet, and DeepLabv3+ respectively. This significant reduction in model size and computational complexity makes MDSC-DeepLabv3+ highly efficient and resource-friendly. Moreover, MDSC-DeepLabv3+ achieves impressive segmentation efficiency, with a recognition speed of only 13.85ms per image. This inference time per image is far less than the above three models, with reductions of 48.97%, 10.18%, and 43.31%, respectively. This indicates that MDSC-DeepLabv3+ is able to perform fast and accurate segmentation, making it highly suitable for real-time applications. Although SegFormer-B0 may have some advantages in terms of deployability, its accuracy is much lower compared to MDSC-DeepLabv3+ (89.79% vs. 97.55%). The reason for this superior performance is the utilization of the improved lightweight MobileNetv2, which replaces Xception in DeepLabv3+, leading to an efficient and accurate model overall. In summary, the proposed MDSC-DeepLabv3+ outperforms the compared four models in the task of segmenting sugarcane and impurities, offering a winning combination of high segmentation accuracy, deployability, and recognition speed.

Instances of the results obtained using the aforementioned segmentation models are illustrated in Figure 15. In which, red [128,0,0] represents cane, blue [0,0,128] represents broken cane, green [0,128,0] represents top, yellow [128,128,0] represents leaf, and black [0,0,0] represents the background. From the visualization of test results, it is evident that all five models perform well in most cases. However, the segmentation obtained by MDSC-DeepLabv3+ stands out as more complete, with clearer preservation of details in general. Upon closer observation, it can be seen that UNet, PSPNet, and SegFormer-B0 misclassify their categories, for instance, misclassifying broken cane as leaf, and vice versa. This indicates inaccuracies in pixel differentiation for these models. Additionally, the compared four models result in fuzzy segmentation and ambiguous boundaries between objects. On the other hand, the proposed MDSC-DeepLabv3+ demonstrates superior performance in addressing the issue of detail adhesion. This can be observed in the instances marked out in the line of MDSC-DeepLabv3+ where the model is capable of better distinguishing object boundaries and preserving fine details.

3.3 Analyzing of comprehensive experiment

The breakage and impurity ratios of raw sugarcane were estimated using the estimation model presented in Section 2.2 and the MDSC-DeepLabv3+ segmentation model presented in Section 2.3. These estimated values were then compared with the measured breakage and impurity ratios obtained through manual weighing to assess the effectiveness of the proposed method.

First, a subset of 25% (70) of the images was randomly selected from the mass dataset with 300 samples. The MDSC-DeepLabv3+ model was applied to semantically segment the selected 70 images and determine the number of cane, broken cane, top, and leaf pixels for each image. Then, corresponding masses were estimated using Eq.(1)-(4), based on the mean values of the surface density for each category obtained through normal fitting. The ratios of breakage and impurity were calculated according to the estimation model defined in Eq.(5)-(6) based on the estimated masses. Finally, the measured breakage and impurity ratios were determined using the measured mass and the relative errors between the estimated and measured results were calculated. Tables 10. 11 document and analyze the relative errors in the breakage ratio and impurity ratio for each sample, as well as the

TABLE 9 Test results of different recognition models.

Segmentation models	IoU/%					mIoU/%	mPA%	Param/M	FLOPs/G	Inference time/ms
	Background	Cane	Broken cane	Top	Leaf					
UNet	98.13	94.18	91.01	93.11	93.73	94.03	96.86	24.89	451.77	27.14
PSPNet	95.45	90.38	87.89	86.48	87.89	89.62	94.85	46.71	118.43	15.42
SegFormer-B0	95.6	82.98	72.38	81.12	79.79	82.37	89.79	3.72	13.56	16.78
DeepLabv3+	97.78	95.18	91.83	93.38	94.62	94.56	97.21	42.19	141.22	24.43
MDSC-DeepLabv3+	97.94	95.13	91.85	94.27	95.03	94.84	97.55	4.07	47.49	13.85

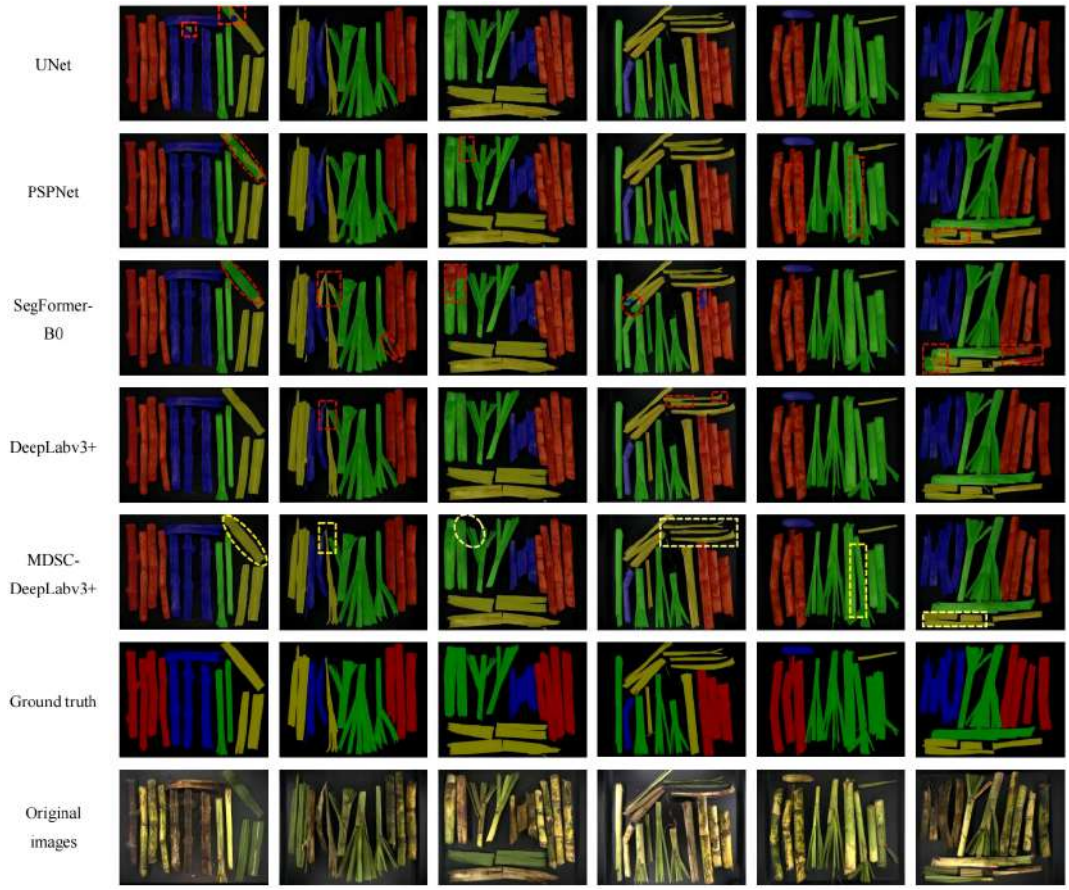


FIGURE 15
Test results of each detection model.

average relative error of the overall samples. The average relative errors were found to be 11.3% and 6.5% for breakage and impurity ratios, respectively. These results indicated that the proposed method exhibits strong reliability.

Additionally, the visualization of measured and estimated ratios of the 70 samples is depicted in Figure 16. This aids in the intuitive observation and analysis of the relationship and differences between predicted and manual measured results. It can be observed that the results obtained using the proposed method exhibit only slight deviations compared to the results obtained through manual weighing measurements, and the fluctuations are minimal. This suggests that the estimated breakage and impurity ratios can maintain their stability. Consequently, the proposed method based on estimation model and MDSC-DeepLabv3+ offers an efficient, accurate, and intelligent means of quantitatively estimating the breakage and impurity ratios of raw sugarcane.

4 Conclusions

In practice, objective, efficient, accurate, and intelligent detection of breakage and impurity ratios is an urgent requirement in the sugar refinery. Therefore, this study developed a novel approach combining the estimation model and MDSC-DeepLabv3+ segmentation network to tackle this problem. First, a machine vision-based acquisition platform was designed, and custom image and mass datasets of raw sugarcane and impurities were constructed. Then, estimation model was built to assess the ratios of breakage and impurity, considering the variation of surface density for the four categories of objects. Finally, the MDSC-DeepLabv3+ segmentation network dedicated to the detection of cane, broken cane, top, and leaf was developed. It effectively incorporated improved MobileNetv2, ASPP_DS, and CA based on DeepLabv3+ to enhance segmentation accuracy, reduce parameters and inference time. The analysis of the experimental results leads to the following conclusions:

1. The breakage and impurity ratios obtained through estimation model based on normal fitted surface density exhibit high accuracy, with corresponding R^2 of 0.976 and 0.968, respectively.
2. The proposed MDSC-DeepLabv3+ achieved superiority considering segmentation accuracy, deployability, and efficiency simultaneously. The mPA and mIoU achieved by MDSC-DeepLabv3+ were as high as 97.55% and 94.84%, respectively, surpassing the baseline DeepLabv3+ by 0.34

TABLE 10 Breakage ratios of 70 samples.

Sample number	Breakage ratio/%			Sample number	Breakage ratio/%		
	Measured	Estimated	Relative errors		Measured	Estimated	Relative errors
1	0.393	0.433	0.103	36	0.163	0.146	0.103
2	0.244	0.215	0.119	37	0.067	0.077	0.148
3	0.328	0.319	0.027	38	0.075	0.077	0.017
4	0.122	0.096	0.218	39	0.117	0.127	0.082
5	0.259	0.272	0.049	40	0.253	0.242	0.047
6	0.486	0.562	0.156	41	0.381	0.418	0.097
7	0.319	0.290	0.090	42	0.145	0.125	0.137
8	0.165	0.174	0.057	43	0.268	0.259	0.036
9	0.173	0.201	0.162	44	0.272	0.247	0.091
10	0.298	0.269	0.097	45	0.060	0.046	0.231
11	0.389	0.416	0.069	46	0.298	0.323	0.087
12	0.235	0.284	0.208	47	0.192	0.168	0.126
13	0.225	0.222	0.012	48	0.209	0.209	0.001
14	0.102	0.131	0.282	49	0.361	0.343	0.049
15	0.105	0.141	0.340	50	0.112	0.150	0.344
16	0.152	0.163	0.077	51	0.233	0.193	0.171
17	0.403	0.340	0.157	52	0.226	0.215	0.048
18	0.144	0.154	0.071	53	0.253	0.281	0.110
19	0.108	0.124	0.150	54	0.299	0.271	0.093
20	0.273	0.267	0.025	55	0.056	0.071	0.262
21	0.388	0.404	0.042	56	0.138	0.168	0.218
22	0.371	0.387	0.045	57	0.141	0.167	0.188
23	0.456	0.480	0.052	58	0.109	0.106	0.035
24	0.264	0.247	0.064	59	0.201	0.207	0.028
25	0.348	0.330	0.053	60	0.385	0.425	0.105
26	0.257	0.240	0.065	61	0.314	0.289	0.079
27	0.170	0.136	0.198	62	0.120	0.130	0.089
28	0.184	0.149	0.191	63	0.227	0.201	0.113
29	0.353	0.337	0.044	64	0.125	0.120	0.044
30	0.351	0.343	0.023	65	0.416	0.451	0.084
31	0.296	0.255	0.138	66	0.160	0.195	0.219
32	0.356	0.342	0.039	67	0.281	0.278	0.011
33	0.214	0.233	0.088	68	0.162	0.186	0.149
34	0.215	0.277	0.286	69	0.322	0.277	0.141
35	0.172	0.150	0.132	70	0.195	0.231	0.184
				Average			0.113

10.3389/fpls.2023.1283230

TABLE 11 Impurity ratios of 70 samples.

Sample number	Impurity ratio/%			Sample number	Impurity ratio/%		
	Measured	Estimated	Relative errors		Measured	Estimated	Relative errors
1	0.473	0.499	0.055	36	0.328	0.319	0.026
2	0.237	0.254	0.071	37	0.313	0.322	0.032
3	0.423	0.424	0.004	38	0.091	0.104	0.142
4	0.292	0.298	0.021	39	0.217	0.240	0.103
5	0.303	0.263	0.133	40	0.380	0.381	0.001
6	0.602	0.570	0.053	41	0.241	0.240	0.005
7	0.445	0.445	0.002	42	0.369	0.369	0.000
8	0.372	0.341	0.082	43	0.292	0.282	0.035
9	0.393	0.352	0.104	44	0.328	0.337	0.026
10	0.294	0.280	0.048	45	0.146	0.167	0.148
11	0.529	0.554	0.046	46	0.274	0.316	0.156
12	0.277	0.272	0.018	47	0.310	0.273	0.118
13	0.378	0.388	0.028	48	0.410	0.382	0.068
14	0.206	0.199	0.034	49	0.254	0.269	0.063
15	0.332	0.314	0.055	50	0.320	0.319	0.004
16	0.240	0.217	0.098	51	0.343	0.386	0.124
17	0.482	0.452	0.062	52	0.328	0.325	0.009
18	0.277	0.298	0.073	53	0.385	0.355	0.077
19	0.331	0.317	0.043	54	0.211	0.232	0.102
20	0.274	0.265	0.034	55	0.228	0.248	0.088
21	0.358	0.322	0.102	56	0.420	0.389	0.073
22	0.491	0.470	0.042	57	0.268	0.267	0.007
23	0.417	0.439	0.054	58	0.209	0.200	0.043
24	0.286	0.318	0.110	59	0.239	0.245	0.023
25	0.273	0.241	0.119	60	0.427	0.421	0.014
26	0.316	0.337	0.066	61	0.332	0.320	0.036
27	0.267	0.265	0.006	62	0.319	0.315	0.011
28	0.251	0.272	0.082	63	0.253	0.239	0.054
29	0.208	0.249	0.196	64	0.313	0.339	0.082
30	0.375	0.334	0.109	65	0.500	0.483	0.034
31	0.296	0.347	0.173	66	0.418	0.378	0.095
32	0.229	0.290	0.265	67	0.297	0.320	0.077
33	0.283	0.279	0.014	68	0.299	0.337	0.128
34	0.357	0.332	0.072	69	0.475	0.465	0.021
35	0.350	0.342	0.024	70	0.301	0.302	0.002
				Average			0.65

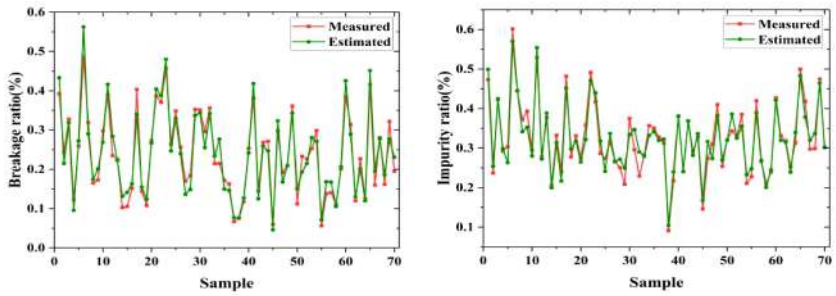


FIGURE 16
Instances of estimation and measured breakage and impurity ratio.

and 0.28. This improvement in accuracy was accomplished with 38.12M, 93.73G, and 10.58ms reduction in Params, FLOPs, and inference time, respectively, making it advantageous for deployment on edge devices and real-time inference.

3. The estimated data obtained according to the approach developed in this study fit the manually obtained breakage and impurity ratios with average relative errors of 11.3% and 6.5%, respectively. The lower segmentation accuracy of broken *cane* is due to their burr and ambiguous boundaries, resulting in a higher average relative error of the breakage ratio.

The raw sugarcane not only includes top and leaf impurities but also contains other impurities like dispersed root whiskers. The upcoming research will emphasize mechanical cleaning of sand, gravel, soil, and similar substances. Additionally, a pivotal aspect of the forthcoming study will involve counting sugarcane roots and estimating their quality through object detection.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

XL: Methodology, Writing – original draft, Writing – review & editing. ZZ: Funding acquisition, Methodology, Writing – review & editing. SL: Methodology, Writing – review & editing. TL: Data curation, Investigation, Visualization, Writing – original draft. JZ: Data curation, Visualization, Writing – original draft. TN: Visualization, Writing – original draft. CJ: Investigation, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work is supported in part by earmarked fund for China Agriculture Research System, grant number CARS-17. number CARS-17.

Acknowledgments

We are very grateful to Junshi Sugar Refinery for providing us with the site for collecting experimental data.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Publisher's note

References

Aparatana, K., Saengprachatanarug, K., Izumikawa, Y., Nakamura, S., and Taira, E. (2020). Development of sugarcane and trash identification system in sugar production using hyperspectral imaging. *J. Infrared Spectrosc.* 28, 133–139. doi: 10.1177/0967033520905369

Chen, J., Lian, Y., and Li, Y. (2020). Real-time grain impurity sensing for rice combine harvesters using image processing and decision-tree algorithm. *Comput. Electron. Agric.* 175, 105591. doi: 10.1016/j.compag.2020.105591

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*. (Germany: ECCV) 801–818. doi: 10.1007/978-3-030-01234-249

Chen, M., Jin, C., Ni, Y., Xu, J., and Yang, T. (2022). Online detection system for wheat machine harvesting impurity rate based on DeepLabV3+. *Sensors* 22 (19), 7627. doi: 10.3390/s22197627

Chinese government website (2018). Available at: http://www.gov.cn/zhengce/content/2018-12/29/content_5353308.htm?ivk_sa=1024320u (Accessed May 24, 2023).

de Mello, M. L., Barros, N. Z., Sperança, M. A., and Pereira, F. M. V. (2022). Impurities in raw sugarcane before and after biorefinery processing. *Food. Anal. Methods* 15, 96–103. doi: 10.1007/s12161-021-02105-1

Dos Santos, L. J., Filletti, É.R., and Pereira, F. M. V. (2021). Artificial intelligence method developed for classifying raw sugarcane in the presence of the solid impurity. *Eclética Quím.* 46 (3), 49–54. doi: 10.26850/1678-4618eqj.v46.3.2021.p49-54

Guedes, W. N., dos Santos, L. J., Filletti, É.R., and Pereira, F. M. V. (2020). Sugarcane stalk content prediction in the presence of a solid impurity using an artificial intelligence method focused on sugar manufacturing. *Food. Anal. Methods* 13, 140–144. doi: 10.1007/s12161-019-01551-2

Guedes, W. N., and Pereira, F. M. V. (2018). Classifying impurity ranges in raw sugarcane using laser-induced breakdown spectroscopy (LIBS) and sum fusion across a tuning parameter window. *Microchem. J.* 143, 331–336. doi: 10.1016/j.microc.2018.08.030

Guedes, W. N., and Pereira, F. M. V. (2019). Raw sugarcane classification in the presence of small solid impurity amounts using a simple and effective digital imaging system. *Comput. Electron. Agric.* 156, 307–311. doi: 10.1016/j.compag.2018.11.039

Hou, Q., Zhang, L., Cheng, M.-M., and Feng, J. (2020). "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (Seattle, WA, USA), 4003–4012. doi: 10.1109/CVPR42600.2020.00406

Hou, Q., Zhou, D., and Feng, J. (2021). "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (Nashville, TN, USA), 13713–13722. doi: 10.1109/CVPR46437.2021.01350

Jin, C., Liu, S., Chen, M., Yang, T., and Xu, J. (2022). Online quality detection of machine-harvested soybean based on improved U-Net network. *Trans. Chin. Soc Agric. Eng.* 38, 70–80. doi: 10.11975/j.issn.1002-6819.2022.16.008

Li, T., Tong, J., Liu, M., Yao, M., Xiao, Z., and Li, C. (2022). Online detection of impurities in corn deep-bed drying process utilizing machine vision. *Foods* 11, 4009. doi: 10.3390/foods11244009

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal loss for dense object detection," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 42 (2), 318–327. doi: 10.1109/TPAMI.2018.2858826

Lin, P., Li, D., Jia, Y., Chen, Y., Huang, G., Elkhouchlaa, H., et al. (2022). A novel approach for estimating the flowering rate of litchi based on deep learning and UAV images. *Front. Plant Sci.*, 3001. doi: 10.3389/fpls.2022.966639

Liu, L., Du, Y., Chen, D., Li, Y., Li, X., Zhao, X., et al. (2022). Impurity monitoring study for corn kernel harvesting based on machine vision and CPU-Net. *Comput. Electron. Agric.* 202, 107436. doi: 10.1016/j.compag.2022.107436

Liu, Q., Liu, W., Liu, Y., Zhe, T., Ding, B., and Liang, Z. (2023). Rice grains and grain impurity segmentation method based on a deep learning algorithm-NAM-EfficientNetv2. *Comput. Electron. Agric.* 209, 107824. doi: 10.1016/j.compag.2023.107824

Luo, Z., Yang, W., Yuan, Y., Gou, R., and Li, X. (2023). Semantic segmentation of agricultural images: A survey. *Inf. Process. Agric.* doi: 10.1016/j.inpa.2023.02.001

Martins, M. B., and Ruiz, D. (2020). Influence of operational conditions of mechanized harvesting on sugarcane losses and impurities. *Eng. Agric.* 40, 352–355. doi: 10.1590/1809-4430-Eng.Agric.v40n3p352-355/2020

Meng, L., Xu, L., and Guo, J. (2020). Semantic segmentation algorithm based on improved MobileNetv2. *Acta Electonica Sin.* 48, 1769. doi: 10.3969/j.issn.0372-2112.2020.09.015

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV). IEEE*. (Stanford, CA, USA: IEEE) 2016, 565–571. doi: 10.1109/3DV.2016.79

Momin, M. A., Yamamoto, K., Miyamoto, M., Kondo, N., and Grift, T. (2017). Machine vision based soybean quality evaluation. *Comput. Electron. Agric.* 140, 452–460. doi: 10.1016/j.compag.2017.06.023

National Development and Reform Commission (2023). Available at: https://www.ndrc.gov.cn/fgsj/tjsj/jjmy/zyspqk/202302/t20230217_1348905_ext.html (Accessed May 24 2023).

Peng, H., Zhong, J., Liu, H., Li, J., Yao, M., and Zhang, X. (2023). ResDense-focal-DeepLabV3+ enabled litchi branch semantic segmentation for robotic harvesting. *Comput. Electron. Agric.* 206, 107691. doi: 10.1016/j.compag.2023.107691

Rong, D., Wang, H., Xie, L., Ying, Y., and Zhang, Y. (2020). Impurity detection of juglans using deep learning and machine vision. *Comput. Electron. Agric.* 178, 105764. doi: 10.1016/j.compag.2020.105764

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, October 5-9, 2015, Proceedings, Part III 18* (Cham: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (Salt Lake City, UT, USA). 2018, 4510–4520. doi: 10.1109/CVPR.2018.00474

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* 128, 336–359. doi: 10.1109/ICCV.2017.74

Shen, Y., Yin, Y., Zhao, C., Li, B., Wang, J., Li, G., et al. (2019). Image recognition method based on an improved convolutional neural network to detect impurities in wheat. *IEEE Access* 7, 162206–162218. doi: 10.1109/ACCESS.2019.2946589

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., et al. (2018). "Understanding convolution for semantic segmentation," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. (Lake Tahoe, NV, USA), 2018, 1451–1460. doi: 10.1109/WACV.2018.00163

Wu, Z., Yang, R., Gao, F., Wang, W., Fu, L., and Li, R. (2021). Segmentation of abnormal leaves of hydroponic lettuce based on DeepLabV3+ for robotic sorting. *Comput. Electron. Agric.* 190, 106443. doi: 10.1016/j.compag.2021.106443

Wu, F., Yang, Z., Mo, X., Wu, Z., Tang, W., Duan, J., et al. (2023). Detection and counting of banana bunches by integrating deep learning and classic image-processing algorithms. *Comput. Electron. Agric.* 209, 107827. doi: 10.1016/j.compag.2023.107827

Xie, L., Wang, J., Cheng, S., Zeng, B., and Yang, Z. (2018). Optimisation and finite element simulation of the chopping process for chopper sugarcane harvesting. *Biosyst. Eng.* 175, 16–26. doi: 10.1016/j.biosystemseng.2018.08.004

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *rXiv [Preprint]*. doi: 10.48550/arXiv.2105.15203

Xu, T., Ma, A., Lv, H., Dai, Y., Lin, S., and Tan, H. (2023). A lightweight network of near cotton-coloured impurity detection method in raw cotton based on weighted feature fusion. *IET Image Process* 17 (9), ipr2.12788. doi: 10.1049/ipr2.12788

Yu, L., Qian, M., Chen, Q., Sun, F., and Pan, J. (2023). An improved YOLOv5 model: application to mixed impurities detection for walnut kernels. *Foods* 12, 624. doi: 10.3390/foods12030624

Yu, J., Zhang, J., Shu, A., Chen, Y., Chen, J., Yang, Y., et al. (2023). Study of convolutional neural network-based semantic segmentation methods on edge intelligence devices for field agricultural robot navigation line extraction. *Comput. Electron. Agric.* 209, 107811. doi: 10.1016/j.compag.2023.107811

Zhang, C., Li, T., and Li, J. (2022). Detection of impurity rate of machine-picked cotton based on improved canny operator. *Electronics* 11, 974. doi: 10.3390/electronics11070974

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (Honolulu, HI, USA). 2017, 2881–2890. doi: 10.1109/CVPR.2017.660

Zhu, S., Ma, W., Lu, J., Ren, B., Wang, C., and Wang, J. (2023). A novel approach for apple leaf disease image segmentation in complex scenes based on two-stage DeepLabv3+ with adaptive loss. *Comput. Electron. Agric.* 204, 107539. doi: 10.1016/j.compag.2022.107539

applied sciences

MDPI

Article

YOLO-DSD: A YOLO-Based Detector Optimized for Better Balance between Accuracy, Deployability and Inference Time in Optical Remote Sensing Object Detection

Hengxu Chen, Hong Jin and Shengping Lv *

College of Engineering, South China Agricultural University, Guangzhou 510642, China; hengxuchen@stu.scau.edu.cn (H.C.); hjin@scau.edu.cn (H.J.)
* Correspondence: lvshengping@scau.edu.cn; Tel.: +86-187-1937-3880

Abstract: Many deep learning (DL)-based detectors have been developed for optical remote sensing object detection in recent years. However, most of the recent detectors are developed toward the pursuit of a higher accuracy, but little toward a balance between accuracy, deployability and inference time, which hinders the practical application for these detectors, especially in embedded devices. In order to achieve a higher detection accuracy and reduce the computational consumption and inference time simultaneously, a novel convolutional network named YOLO-DSD was developed based on YOLOv4. Firstly, a new feature extraction module, a dense residual (DenseRes) block, was proposed in a backbone network by utilizing a series-connected residual structure with the same topology for improving feature extraction while reducing the computational consumption and inference time. Secondly, convolution layer–batch normalization layer–leaky ReLu (CBL) ×5 modules in the neck, named S-CBL×5, were improved with a short-cut connection in order to mitigate feature loss. Finally, a low-cost novel attention mechanism called a dual channel attention (DCA) block was introduced to each S-CBL×5 for a better representation of features. The experimental results in the DIOR dataset indicate that YOLO-DSD outperforms YOLOv4 by increasing $mAP^{0.5}$ from 71.3% to 73.0%, with a 23.9% and 29.7% reduction in Params and Flops, respectively, but a 50.2% improvement in FPS. In the RSOD dataset, the $mAP^{0.5}$ of YOLO-DSD is increased from 90.0~94.0% to 92.6~95.5% under different input sizes. Compared with the SOTA detectors, YOLO-DSD achieves a better balance between the accuracy, deployability and inference time.

Keywords: optical remote sensing; object detection; feature extraction; attention mechanism

Citation: Chen, H.; Jin, H.; Lv, S. YOLO-DSD: A YOLO-Based Detector Optimized for Better Balance between Accuracy, Deployability and Inference Time in Optical Remote Sensing Object Detection. *Appl. Sci.* **2022**, *12*, 7622. https://doi.org/10.3390/app12157622

Academic Editors: Weitao Chen, Ailong Ma and Guohua Wu

Received: 22 June 2022
Accepted: 25 July 2022
Published: 28 July 2022

1. Introduction

Object detection in optical remote sensing images (ORSIs) is a crucial but challenging task for remote sensing technology and has been widely applied in many fields, such as military, natural resources exploration, urban construction, agriculture and mapping [1,2]. The development of a cost-effective detector considering the characteristic of ORSIs is the persistently pursued direction, and has attracted a large amount of attention from scholars and practitioners.

The approaches for object detection can be roughly divided into traditional detectors and deep learning (DL)-based detectors. DL-based detectors, especially convolutional neural network (CNN) detectors, have gradually replaced traditional detectors since they possess better adaptability and generalization in different application scenarios. There are two categories of DL-based detectors: one-stage [3–9] and two-stage [10–13]. The one-stage detectors directly regress bounding boxes and probabilities for each object simultaneously without region proposals; thus, they perform well regarding inference speed. Two-stage detectors employ the region proposals to improve the location and detection accuracy, with the sacrifice of the inference speed. With the emergence of large-scale natural scene images (NSIs) datasets for object detection tasks such as Pascal VOC [14] and MS COCO [15],

DL-based detectors have been further developed for a better tradeoff between accuracy and cost, including Faster-RCNN [12], single shot multibox detector (SSD) [3], the series of You Only Look Once (YOLO) [4–6,8], CenterNet [7], EfficientDet [9] and RetinaNet [16]. These detectors with continuous improvement have been widely applied in various natural scene visual detection tasks.

Since ORSIs are photographed from an overhead perspective at different heights, whereas NSIs are shot from a horizontal perspective at relatively close distance, three main differences have emerged as follows: first, the available feature of most detected objects in ORSIs is less obvious than that in NSIs and leads to greater inter-class similarity. Second, the intra-class difference is more prominent since object scales of the same category in ORSIs usually vary greater. Third, the background in ORSIs is more complex and abundant than that in NSIs. Differences between ORSIs and NSIs with instances are shown in Figure 1. These differences make object detection in ORSIs more difficult, and most of the well-designed detectors for NSIs are not elaborately optimized for ORSIs. For the problems of a greater intra-class difference and inter-class similarity caused by the characteristic of objects in ORSIs, the detector needs to extract more abundant object features with high-level semantics to overcome it. However, the feature of objects in ORSIs are easily submerged by the redundant and complex background information and thus will decrease or even disappear when transmitted in the detector. Thus, DL-based detectors also require a stronger feature extraction and transmission ability.

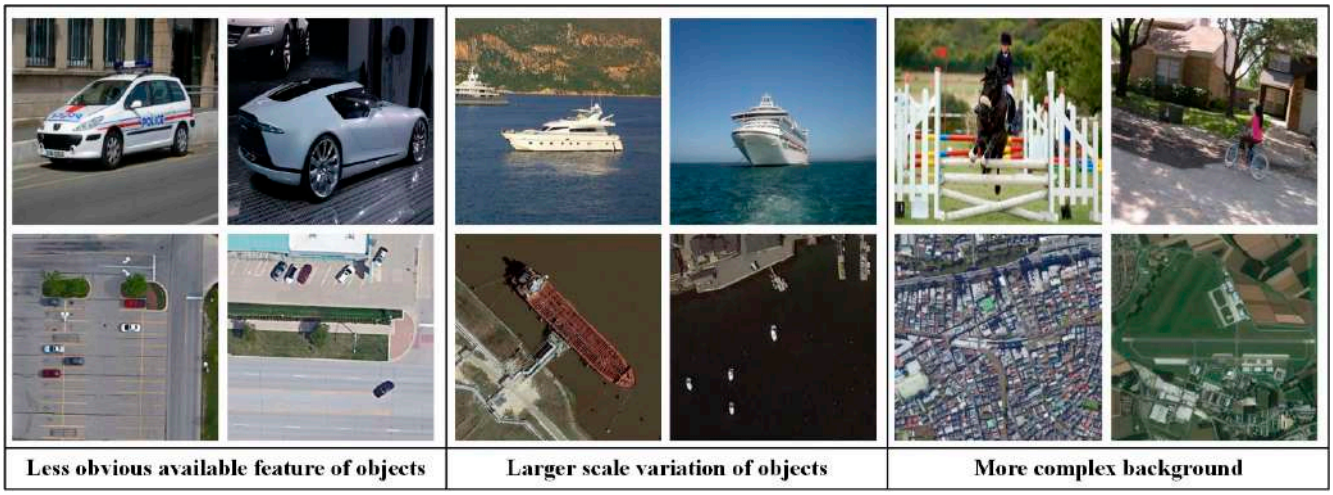


Figure 1. Three main differences between RSIs and NSIs. The first and second lines show instances from NSIs and RSIs, respectively.

With the popularity and wide application of embedded devices such as unmanned aerial vehicles (UAVs), the demand for real-time optical remote sensing object detection deployed on edge devices has increased rapidly. UAVs with far less computing resource and storage space than computers involve wide application scenarios such as rescue, military and surveying tasks, which require a high detection accuracy, flexible equipment deployment and less inference time for detectors [17].

In recent years, several outstanding achievements have been made by researchers in fields related to ORSIs, and can be roughly divided into heavyweight [18–21] and lightweight detectors [22–25]. Most of the heavyweight detectors usually have a high accuracy but require a large computational cost, and thus hinder their real-time response and the deployment on UAVs, whereas lightweight detectors have practical deployability and a fast inference speed but it is difficult for them to achieve as high a competitive accuracy as heavyweight detectors, especially for large multi-category object detection tasks [23,24,26]. Therefore, optimizing the structure of heavyweight detectors toward a better balance between accuracy, deployability and inference time is an issue well worth

investigating. To establish a detector with a better balance between accuracy, deployability and inference time, a novel detector called YOLO-DSD for real-time optical remote sensing object detection based on YOLOv4 was developed in this study. The main contributions are as follows: (1) a new feature extraction module named a dense residual (DenseRes) Block was designed for better feature extraction and to reduce the computational cost and inference time in the backbone network. (2) Convolution layer–batch normalization layer–leaky ReLu (CBL) ×5 modules in the neck were improved with a short-cut connection and named S-CBL×5 to strengthen the transmission of object features. (3) A novel low-cost attention mechanism called a dual channel attention (DCA) Block was proposed to enhance the representation of the object feature. The experimental results in the DIOR dataset indicate that YOLO-DSD outperforms YOLOv4 by increasing $mAP^{0.5}$ from 71.3% to 73.0%, with a 23.9% and 29.7% reduction in Params and Flops, respectively, but a 50.2% improvement in FPS. In the RSOD dataset, the $mAP^{0.5}$ of YOLO-DSD is increased from 90.0~94.0% to 92.6~95.5% under different input sizes. Compared with the SOTA detectors, YOLO-DSD achieves a better balance between accuracy, deployability and inference time.

2. Related Works

2.1. DL-Based Detectors for Optical Remote Sensing Object Detection

DL-based detectors have been widely applied in natural sense visual tasks. However, detectors established on NSIs need to further improve their feature extraction ability for optical remote sensing object detection tasks due to the problems of a greater intra-class difference, inter-class similarity and feature loss in ORSIs. Therefore, some heavyweight detectors have been improved and applied in ORSIs by many scholars. Xu et al. [18] modified YOLOv3 with a multi-receptive field to take full advantage of the feature information and to detect optical remote sensing objects effectively. Cheng et al. [19] designed an end-to-end cross-scale feature fusion framework for ORSIs object detection based on Faster R-CNN with a feature pyramid network (FPN) [16]. Yin et al. [20] proposed a multi-scale feature extraction network based on RetinaNet, which strengthens the detection performance of irregular objects in ORSIs. Yuan et al. [21] established a multi-FPN that performs well in object detection with a complex background. The above research has successfully made obvious improvements in detection accuracy, but come with the non-ignorable sacrifice of the deployability or inference speed, and thus further hinder the application of detectors in edge devices. As a consequence, some lightweight DL-based detectors have been elaborately designed and improved to facilitate the application in edge devices. Li et al. [22] designed a lightweight detector by taking advantage of YOLOv3 and DenseNet [27]. Lang et al. [23] employed the backbone network of ThunerNet [28] and constructed a six-layer feature fusion pyramid to enhance the detection performance. The improved YOLOv4-tiny proposed by Lei et al. [24] was constructed with an efficient channel attention mechanism to enhance the information sensitivity in each channel. Li et al. [25] established a lightweight detector for vehicle and ship detection through using a semantic transfer block and the distillation loss. Although these lightweight detectors have a better accuracy after improvement, there is still an obvious gap in the detection accuracy compared with heavyweight detectors.

Our motivation is to propose an end-to-end detector that can achieve a higher detection accuracy, better deployability and less inference time in order to meet the requirements of edge device real-time detection. YOLOv4 [8] is one of the widely used one-stage detectors, with an impressive performance in accuracy, deployability and inference time. It has been improved and applied in various fields, such as agriculture, industry and transportation [29–32], which verify its excellent generalization. In this study, YOLOv4 was utilized as the basic framework, while it was optimized from feature extraction modules, structures of the neck and the attention mechanism for a better application in optical remote sensing object detection.

2.2. Feature Extraction Modules in Backbone

The backbone that is utilized to extract high-level semantic features of images is the first part of the DL-based detector. It comprises several feature extraction modules. VGG [33] is one of the earliest backbones for object detection and utilizes 3×3 convolution layers as the feature extraction module. However, its heavy computation burden and shallow depth hinder the deployability and performance of detectors.

To solve this problem, He et al. [34] introduced a new feature extraction module named a Res Block to deepen the depth of backbones by adding short-cut connections. ResNet based on the Res Block achieves a better accuracy than VGG in the natural scene dataset, with a lower computation burden and deeper depth. The backbone DarkNet53 of YOLOv3 [6] also uses the Res Block as the main feature extraction module. Since then, many feature extraction modules based on the Res Block, such as a ResNeXt Block [35], Res2 Block [36], Dense Block [27] and CSP Block [37], have been improved and developed. The trunk of the ResNeXt Block is split into 32 paths that transform the input from high to low dimensions and back to high dimensions using the same topology, and aggregates them through element-wise addition. Although the ResNeXt Block outperforms the Res Block with fewer parameters and a higher detection accuracy in the natural scene dataset, since the semantic relevance between background and detected objects in ORSIs is stronger than that in NSIs [38], the operation of the ResNeXt Block easily breaks this relevance, and is thus not conducive to the detection performance in ORSIs. The Res2 Block can generate multi-scale features through a hierarchical short-cut connection and increase in receptive fields, thus improving the detection accuracy and reducing the computational consumption. However, its structure with parallel convolution and interactive operations significantly increases the inference time. The Dense Block contains several dense layers. The output of the dense layer is concatenated with its input, and the concatenated feature map serves as the input of the next dense layer. This structure takes full advantage of the short-cut that can better retain the feature and reduce the computation burden. However, the Dense Block will deteriorate in the situation where the background submerges features of detected objects in ORSIs, since the background information is more redundant and complex. Meanwhile, the structure of the Dense Block will reduce its inference speed due to the asymmetry of the input channels number and output channels number for a convolution operation. The CSP Block is the feature extraction module of the backbone CSP DarkNet in YOLOv4. It is mainly composed of several Res units based on a short-cut and a cross-stage part containing a 1×1 convolution layer. Although this structure can double the number of gradient paths and improve the detection accuracy through a splitting and merging strategy, there is parallel convolution and the problem of the trunk of the CSP Block being stacked alternately by an excessive convolution layer, which significantly increase the degree of network fragmentation and thus decrease the inference speed [39].

In order to alleviate the shortcomings of the above Blocks, a novel feature extraction module DenseRes Block is proposed in this study to improve the backbone in YOLOv4. Firstly, the input feature map of the DenseRes Block was compressed in order to increase the proportion of object feature information. Then, the series-connected residual structure with the same topology was utilized not only to obtain the high-level semantics of the object feature but also to reduce the computational consumption and inference time. Finally, the feature map output from the residual structure was combined with the input of the DenseRes Block to enhance the semantic relevance between background and detected objects.

2.3. Structure of the Neck

In the neck, feature maps output from the backbone will be processed and transmitted to the prediction part of the detector. The neck of the early DL-based detectors only directly transmits the last feature map of the backbone to the prediction part. The shallow feature map contains rich location information but low-level semantic information, whereas the deep feature map is the opposite; thus, this structure is not conducive to object detection,

especially for small objects. In order to improve the detection performance of detectors for small objects, Liu et al. [3] proposed a neck structure that directly transfers the feature maps of different levels from the backbone to the prediction part of the detector for multi-scale detection, and proves that the utilization of a shallow feature map is beneficial for small object detection. However, shallow feature maps still lack high-level semantic information, while deep feature maps are still short of location information. FPN [16] is designed to transfer the high-level semantic information to the shallow feature map through the bottom-up structure to further improve the detection performance of the detector for small objects. In order to make the deep feature map possess rich location information and high-level semantic information, BFPN [40] has been developed to fuse the penultimate feature map and the last feature map based on FPN, while PANet [41] adds a top-down structure based on FPN to transmit location information to the deep feature map. Both BFPN and PANet can improve the detection performance for middle and large objects while maintaining a high detection accuracy for small objects.

YOLOv4 adopts the PANet as the framework in the neck. However, YOLOv4 suffers from the problem of feature loss in ORSIs due to many convolution operations in the neck. Therefore, a short-cut connection based on a residual is introduced to each CBL$\times$5 in the neck for strengthening the transmission of object features without an increase in the computational burden and inference time.

2.4. Attention Mechanism

The attention mechanism assigns different weights to the pixel according to the spatial or channel relationship between the pixels in the feature map to enhance the representation of the feature, and it mainly includes three categories: a channel attention mechanism (e.g., an SE Block [42] and ECA Block [43]), spatial attention mechanism (e.g., a CA Block [44]) and hybrid attention mechanism (e.g., a CBAM Block [45]). The attention mechanism can improve the detection accuracy in NSIs with a few parameters and computation burden increase for detectors. The SE Block squeezes and then extends channel information through two full connection layers in order to learn the relationship of global channel information and effectively improve the detection performance, but the relationship between local channel information is not considered. The ECA Block learns the relationship between local channels through 1-D convolution with an adaptive convolution kernel, which promotes the detection performance but ignores the relationship of global channel information. In the CA Block, the information is extracted by average pooling in horizontal and vertical directions, respectively, and then concatenated and fused by 2-D convolution. The fused information is split into two parts and each part is further extracted by the convolution layer, respectively. The hybrid attention mechanism CBAM Block combines the channel and spatial attention mechanism. Both the CA Block and CBAM Block bring an obvious improvement in the detection accuracy in the natural scene dataset, but their complex structure increase the inference time. Meanwhile, it is difficult for them to use a few parameters to extract the spatial information of ORSIs due to their more complex background and less spatial feature information for detected objects in ORSIs.

In order to more efficiently highlight the feature related to the detection task in ORSIs with a better robustness, a novel channel attention mechanism named a DCA Block was proposed to enhance the representation of the object feature in ORSIs through combing global and local channel information with a slight inference time increase.

3. Proposed Methods

3.1. Method Overview

The structure of YOLOv4 is given in Figure 2. YOLOv4 consists of a backbone, neck and prediction. YOLOv4 is established for NSIs and not practical enough to be adopted in ORSIs directly. Specifically, the backbone CSP DarkNet in YOLOv4 utilizes the CSP Block [37] as the feature extraction module and performs well in detection accuracy, but its model complexity and computational burden can be further reduced to improve its

deployability and inference speed for ORSIs. The neck PANet [41] employed in YOLOv4 can strengthen the integration of a shallow and deep feature map, but its CBL×5 modules will easily cause the problem of feature loss, which is not conducive to information transmission for objects in ORSIs. Moreover, attention mechanisms that can enhance the feature representation are not utilized in YOLOv4.

The proposed detector YOLO-DSD based on YOLOv4 is shown in Figure 3. Three new modules are presented to improve the performance of YOLOv4. In the backbone, we developed a DenseRes Block as the main module for a better feature extraction and reduction in computational cost. In the neck, S-CBL×5 was proposed to handle the information loss problem, and the proposed attention mechanism, the DCA Block, was added after each S-CBL×5 module to enhance the representation of features.

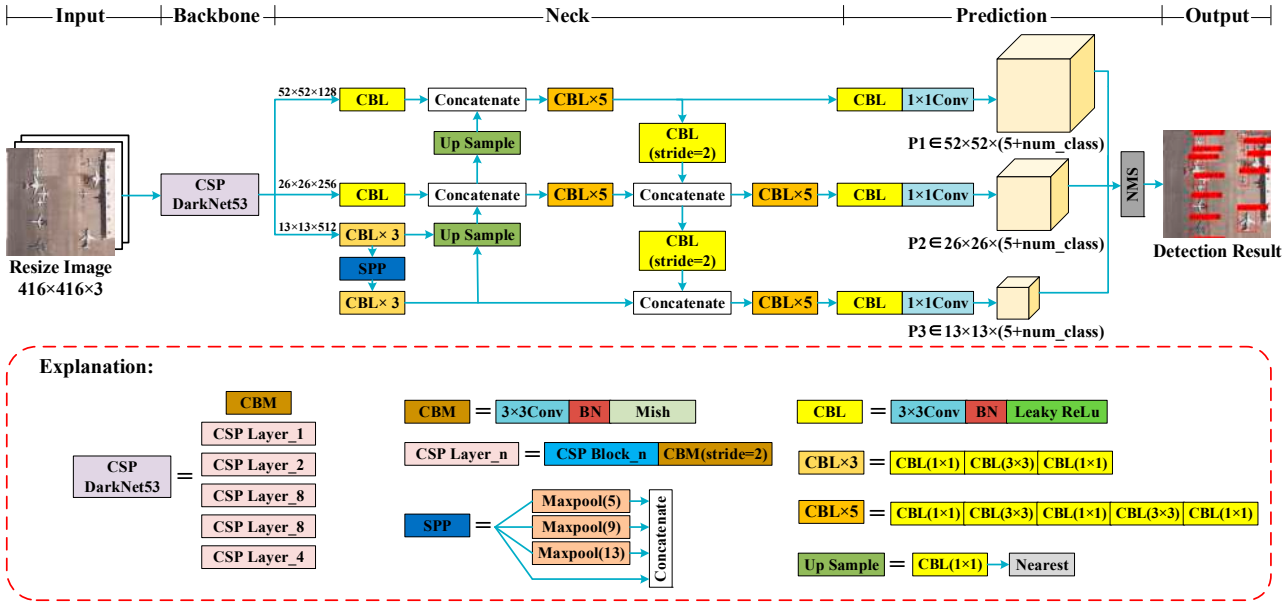


Figure 2. The architecture of YOLOv4.

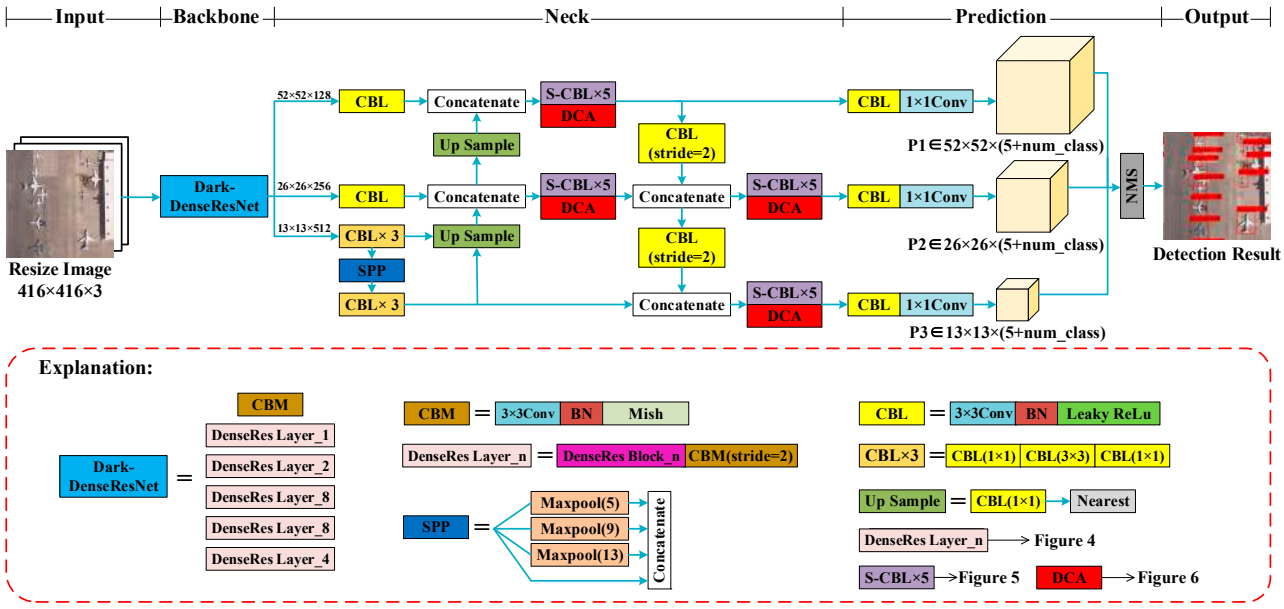


Figure 3. The architecture of YOLO-DSD.

3.2. Improvement in the Backbone

YOLOv4 adopts a CSP Block, shown in Figure 4a, to extract features of images in the backbone. Although the CSP Block performs well in detection accuracy, the structure of the CSP Block containing a parallel convolution operation for reusing the feature of the 'Input' and excessive convolution layers caused by 'Res Unit' takes up a large amount of computing resources and inference time [39]. Aiming at this problem of the CSP Block, we proposed a DenseRes Block, shown in Figure 4b, and employed it in the backbone for feature extraction.

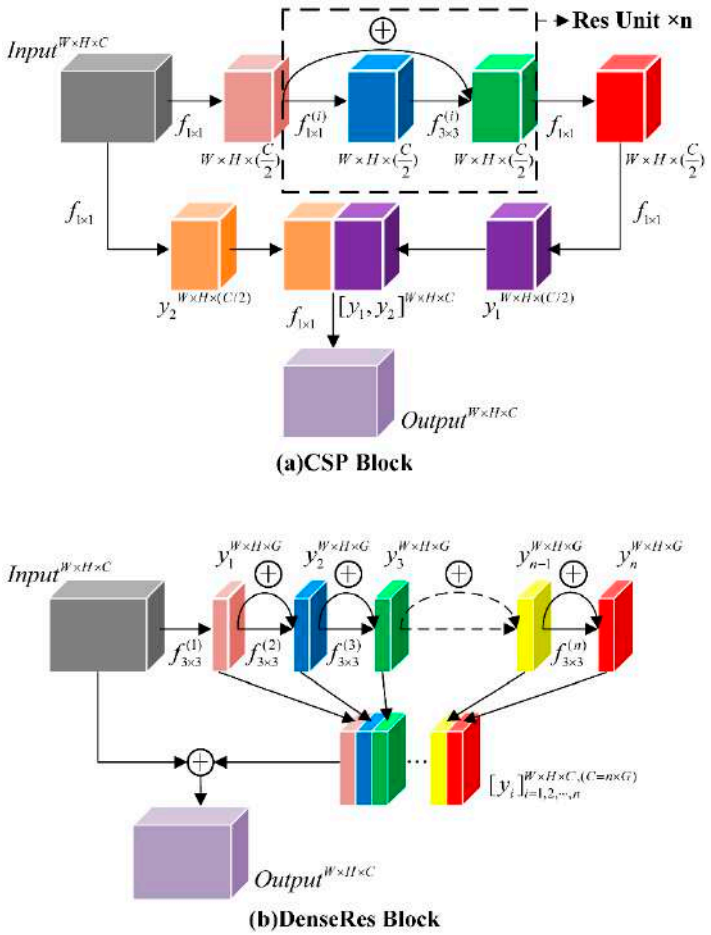


Figure 4. The structure comparison between CSP Block (**a**) and DenseRes Block (**b**).

The DenseRes Block is only composed of several series-connected 3 × 3 convolution operations $f_{3\times3}^{(i)}$ $(i = 1, 2, \cdots, n)$ and short-cut connections based on residual learning. y_i is the output feature map of $f_{3\times3}^{(i)}$. For the feature map $Input \in \mathbb{R}^{W\times H\times C}$, W, H and C indicate the height, width and channel number of the map, respectively. Since the feature of detected objects in ORSIs is easily overwhelmed by that of the background when transmitted, we utilized a feature map with fewer channels as the output of the first convolution operation to compress the '*Input*' to focus on object features and reduce the proportion of background information. Therefore, the feature map $y_1 \in \mathbb{R}^{W\times H\times G}$ was computed by

$$y_1^{W\times H\times G} = f_{3\times3}^{(1)}\left(Input^{W\times H\times C}\right) \tag{1}$$

where $C = n \times G$, $f_{3\times3}^{(1)}$ contains the 3 × 3 convolution layer that compacts the number of channels from C to G, the BN layer and the leaky ReLu activation function. If $n = 1$, the DenseRes Block is the same as the Res Block. When $n > 1$, the DenseRes Block will compress the '*Input*' and make a feature extraction. It was proven in Ref. [39] that the following

operations can effectively reduce the memory access cost and the inference time of the model: (1) the input channel and output channel of the convolution layer should be equal as much as possible; (2) the number of fragmented operators (i.e., the number of individual convolution or parallel operations in one building block) should be reduced. Therefore, $y_j(1 < j \leq n) \in \mathbb{R}^{W \times H \times G}$ could be designed as

$$y_{j(1<j\leq n)}^{W\times H\times G} = y_{j-1}^{W\times H\times G} \oplus f_{3\times 3}^{(j)}\left(y_{j-1}^{W\times H\times G}\right) \tag{2}$$

where $f_{3\times 3}^{(j)}(1 < j \leq n)$ contains the 3×3 convolution layer with the same number G of input and output channels, the BN layer and the leaky ReLu activation function. $\oplus$ indicates the element-wise addition. From the comparison between the CSP Block and the DenseRes Block shown in Figure 4, the output of each 'Res Unit' in the CSP Block will go through two convolution layers with different kernel sizes, whereas that of each 'y_i' in the DenseRes Block only goes through one 3×3 convolution layer. Therefore, the fragment degree can be decreased. Moreover, we used a short-cut based on residual learning to connect $y_j(1 < j \leq n)$ and $y_{j-1}(1 < j \leq n)$ for the problem of feature loss in the process of feature extraction.

In ORSIs, there will be potential semantic relevance between the object and the background [21,38]. For example, cars and airplanes tend to park on land whereas ships tend to sail on the sea, and bridges are built over water whereas overpasses are built over land. In order to make the network better learn high-level semantic relevance, the $Output \in \mathbb{R}^{W \times H \times C}$ was designed as

$$Output^{W\times H\times C} = Input^{W\times H\times C} \oplus [y_i]_{i=1,2,\cdots,n}^{W\times H\times C(C=n\times G)} \tag{3}$$

where $[y_i]_{i=1,2,\cdots,n}^{W\times H\times C(C=n\times G)}$ concatenates $y_1, y_2, \ldots, y_n$ in the channel dimension to a feature map with the same size as $Input \in \mathbb{R}^{W \times H \times C}$. $[y_i]_{i=1,2,\cdots,n}^{W\times H\times C(C=n\times G)}$ possessing more object information was combined with the $Input \in \mathbb{R}^{W \times H \times C}$ holding more background information by element-wise addition directly to improve the detection accuracy. Compared with the CSP Block, such a designed structure in the DenseRes Block not only reuses the feature of '*Input*' but also omits a parallel convolution operation, which can further reduce the degree of the fragment in the backbone.

The DenseRes Block was utilized in order to replace the original module, the CSP Block, in the backbone. The architecture and complexity of the restructured backbone, named DarkNet-DenseRes, is shown in Table A1, Appendix A.

3.3. Improvement in the Neck

YOLOv4 uses the feature pyramid structure of PANet in the neck to fuse feature maps of different levels and extract a feature, which performs well in object detection in natural scenes. However, the feature information of objects in ORSIs is usually far less obvious than that of objects in natural scenes, and information loss caused by excessive convolutional operations in PANet limits the detection performance of the network for the objects in ORSIs. In order to solve this problem, S-CBL×5 was utilized to replace each CBL×5 in the original neck as shown in Figure 3. The structure comparison between CBL×5 and S-CBL×5 is given in Figure 5. S-CBL×5 adds two short-cuts based on CBL×5 and does not add additional parameters and inference time.

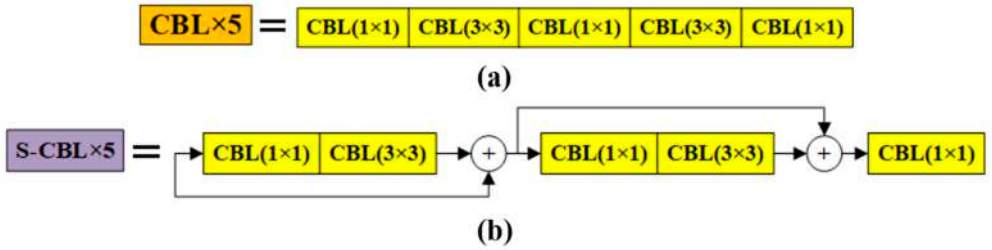


Figure 5. The structure comparison between CBL×5 (**a**) and the proposed S-CBL×5 (**b**).

To highlight significant features related to the detection task, the DCA Block was proposed to optimize the weight distribution of each feature map in the channel dimension by combining the local and global relationship between channels with a slight increase in computations cost and inference time. The structure of the DCA Block is shown in Figure 6.

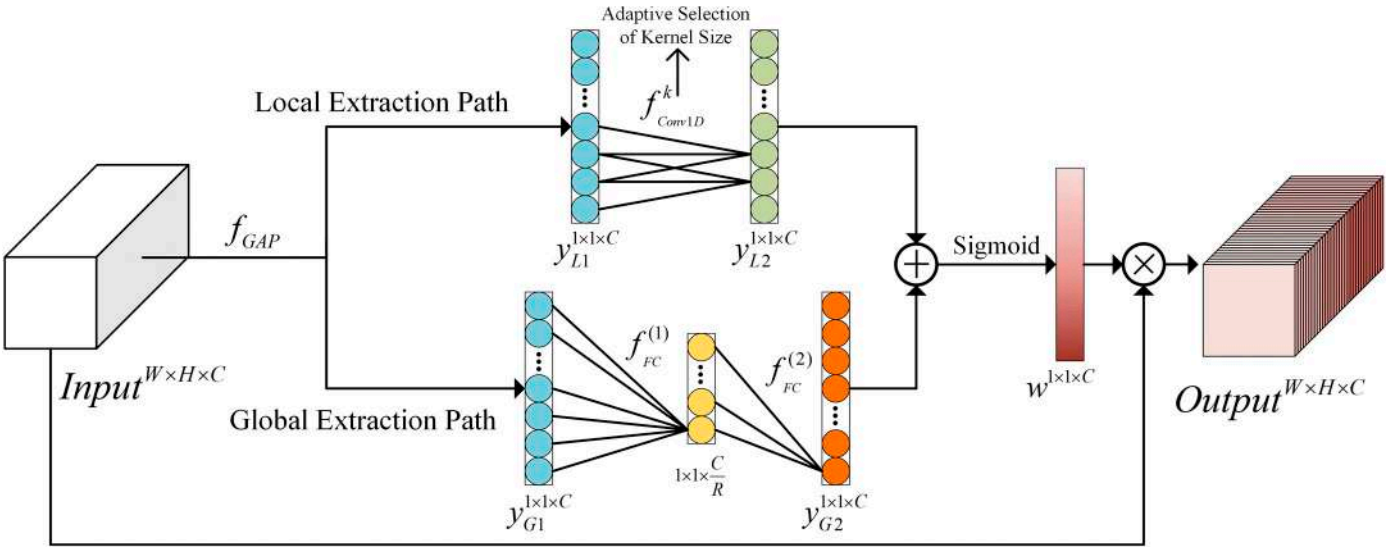


Figure 6. The structure of the proposed DCA Block.

The DCA Block is composed of a 'Local Extraction Path' and 'Global Extraction Path' in parallel. The 'Global Extraction Path' is used to learn the global relationship between channels, whereas the 'Local Extraction Path' is employed to extract the local channel relationship.

Firstly, global average pooling was employed to obtain the integrated information in the space dimension of each channel $y_{L1}^{1\times1\times C}$ and $y_{G1}^{1\times1\times C}$, where $y_{L1}^{1\times1\times C}$ and $y_{G1}^{1\times1\times C}$ indicate the input of the 'Local Extraction Path' and 'Global Extraction Path', respectively, and $y_{L1}^{1\times1\times C} = y_{G1}^{1\times1\times C}$.

Secondly, $y_{L2}^{1\times1\times C}$ in the 'Local Extraction Path' could be computed by

$$y_{L2}^{1\times1\times C} = f_{Conv1D}^{k}\left(y_{L1}^{1\times1\times C}\right) \tag{4}$$

$$k = \frac{\log_2 C + 1}{2} \tag{5}$$

where f_{Conv1D}^{k} represents the 1-dimension convolution layer. Since each feature map has a different number of channels and the kernel size of the convolution layer is proportional to the number of the channels [43], the mapping between its kernel size (k) and the number of input channels (C) is given in Equation (5). f_{Conv1D}^{k} could adaptively select the kernel size according to non-linearly mapping Equation (5); thus, it can extract the local relationship between covered channels more effectively than the convolution layer with a hand-given convolution kernel size.

At the same time, two full connection layers were used as a bottleneck in the 'Global Extraction Path' to build the global relationship of each channel:

$$y_{G2}^{1\times1\times C} = f_{FC}^{(2)}\left(f_{FC}^{(1)}(y_{G1}^{1\times1\times C})\right) \tag{6}$$

where $f_{FC}^{(1)}$ is the first full connection layer that compresses the channel number from C to C/R, and $f_{FC}^{(2)}$ is the second full connection layer that extends the channel number from C/R to C. The value of the zoom factor R that could reduce the complexity of the structure was set to 32 according to the experimental results in Section 4.4.1. The structure of the 'Global Extraction Path' with two full connection layers has a stronger non-linearity and can fit better with the complex global relationship between each channel.

Thirdly, the output of the 'Global Extraction Path' and 'Local Extraction Path' were combined by element-wise addition, and the sigmoid function was applied to generate the weight $w \in \mathbb{R}^{1\times1\times C}$. Finally, the output of the DCA Block was calculated as:

$$w^{1\times1\times C} = Sigmoid\left(y_{G3}^{1\times1\times C} \oplus y_{L2}^{1\times1\times C}\right) \tag{7}$$

$$Output^{W\times H\times C} = w^{1\times1\times C} \otimes Input^{W\times H\times C} \tag{8}$$

where $\otimes$ represents the operation of the element-wise product. As shown in Figure 3, we added the proposed DCA Block after each S-CBL×5 to generate an improved PANet (shown in Figure 7) with a structure that is more suitable for optical remote sensing object detection and has a nearly equal computational cost compared to the original structure.

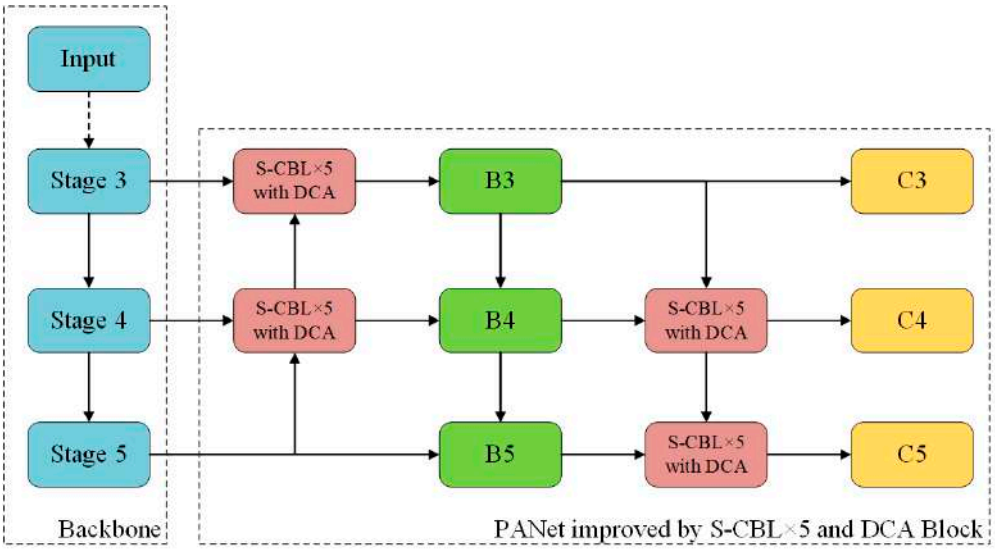


Figure 7. The structure of the improved PANet.

3.4. Prediction

Decoding and obtaining the detection result were processed in the prediction. As shown in Figure 3, each output of the neck went through with a CBL module and a 1×1 convolution layer, and three feature maps, $P_1 \in \mathbb{R}^{52\times52\times num\_class}$, $P_2 \in \mathbb{R}^{26\times26\times num\_class}$ and $P_3 \in \mathbb{R}^{13\times13\times num\_class}$, were generated. Then, as shown in Figure 8, P_1, P_2 and P_3 were mapped back to the original image and the image was divided into 52×52, 26×26 and 13×13 sizes of grids. Each grid corresponding to a feature map contains the information of three anchors. In each anchor, (x, y) and (w, h) are the offset coefficient and size coefficient, respectively, Cf is the confidence of the grid containing the object and $C_1, C_2, C_3, \ldots, C_n$ are the confidence of each object class, respectively.

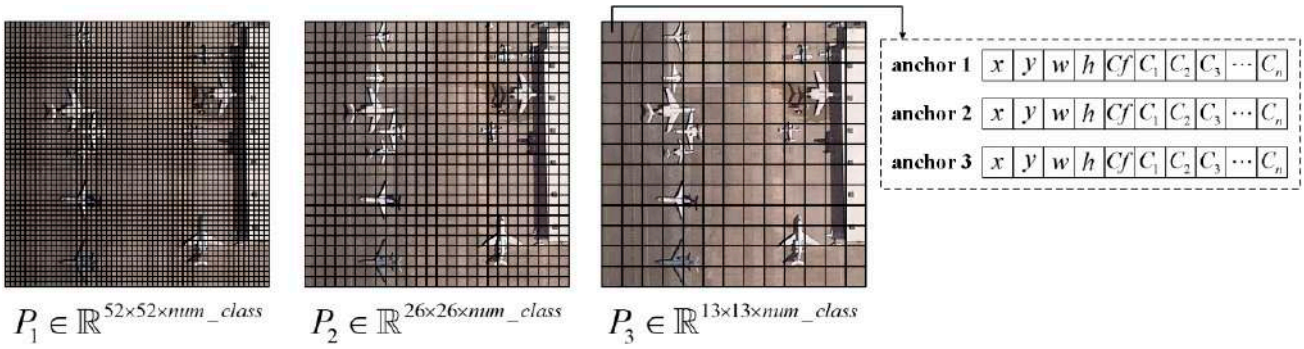


Figure 8. The image is divided into 52 × 52, 26 × 26 and 13 × 13 by P_1, P_2 and P_3, respectively.

Then, each grid generated three bounding boxes according to the information combined with anchors, and the process of converting the anchor to the bounding box is illustrated in Figure 9. (C_x, C_y) are the upper left corner position of the current grid and the center of each grid anchor. ($\sigma(x)$, $\sigma(y)$) is the offset of the bounding box relative to the anchor. The width b_w and height b_h of the bounding box were obtained through multiplying the width p_w and height p_h of the anchor by scaling factors e^w and e^h, respectively. Finally, the detection results were obtained after redundant bounding boxes were removed through NMS.

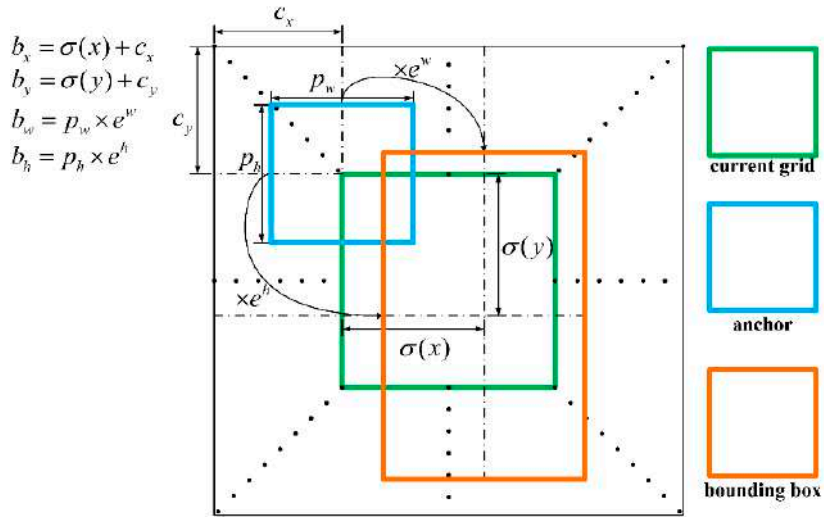


Figure 9. The process of converting anchor to bounding box.

3.5. Loss Function

The loss function of YOLOv4 includes three parts: confidence, classification and bounding box regression loss. YOLOv4 employs the complete intersection over union (*IoU*) loss (*CIoU*) [46], replacing the mean squared error loss adopted in YOLOv3 with the bounding box regression loss. *CIoU* takes the overlap area, center point distance and aspect ratio into consideration simultaneously, and the convergence speed and detection accuracy were improved. *CIoU* introduces a penalty item $\alpha\nu$ based on the distance *IoU* loss to impose the consistency of the aspect ratio for the ground truth (bb^{gt}) and bounding box (bb^b). The loss of *CIoU* can be defined as Equation (9).

$$\begin{aligned} Loss_{CIoU} &= 1 - \left(IoU - \frac{\rho^2\left(b^{gt}, b^b\right)}{c^2} - \alpha\nu \right) \\ \alpha &= \frac{\nu}{1 - IoU + \nu}, \ \nu = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w^b}{h^b}\right)^2 \end{aligned} \tag{9}$$

where b^{gt}, b^{b} are the center of bb^{gt}, bb^{b}, respectively, ρ denotes the Euclidean distance, c represents the diagonal length of the smallest enclosing rectangle covering bb^{gt}, bb^{b}, α is a positive trade-off value and ν means the consistency of the aspect ratio. w^{gt}, w^{b} are the width of the bb^{gt}, bb^{b}, respectively. h^{gt}, h^{b} are the height of the bb^{gt}, bb^{b}, respectively.

CIOU can directly minimize the distance between the bounding box and ground truth and accelerate the model convergence. Previous works [47–49] have proved that *CIOU* can perform better in detecting objects with diverse sizes, which can match well with the characteristics of remote sensing object detection tasks.

4. Experiments and Discussion

In this section, we conduct ablation and comparative experiments on a public optical remote sensing dataset DIOR [2] with 20 categories to validate the proposed YOLO-DSD, considering the accuracy, deployability and speed indictors. Another optical remote sensing dataset RSOD [50] with 4 categories was utilized to further verify the effectiveness of the proposed YOLO-DSD compared with YOLOv4.

4.1. Datasets

4.1.1. DIOR Dataset

DIOR [2] is a large ORSIs dataset that was established in 2020 to develop and validate data-driven methods. It contains 23,463 images and 192,472 objects in total, covering 20 categories in optical remote sensing field. Images in this benchmark dataset have been clipped into 800 × 800 pixels. There are vast scale variations across objects in DIOR because it contains images with spatial resolutions ranging from 0.5 m to 30 m. According to the definition of COCO [15], objects with area of ground truth less than 32 × 32 pixels, between 32 × 32 pixels and 96 × 96 pixels and larger than 96 × 96 pixels are taken as small, middle and large-sized objects, respectively. Each category and the size distribution of objects in DIOR is shown in Figure 10. It can be seen that objects in DIOR possess great size difference and are concentrated in small and middle-sized.

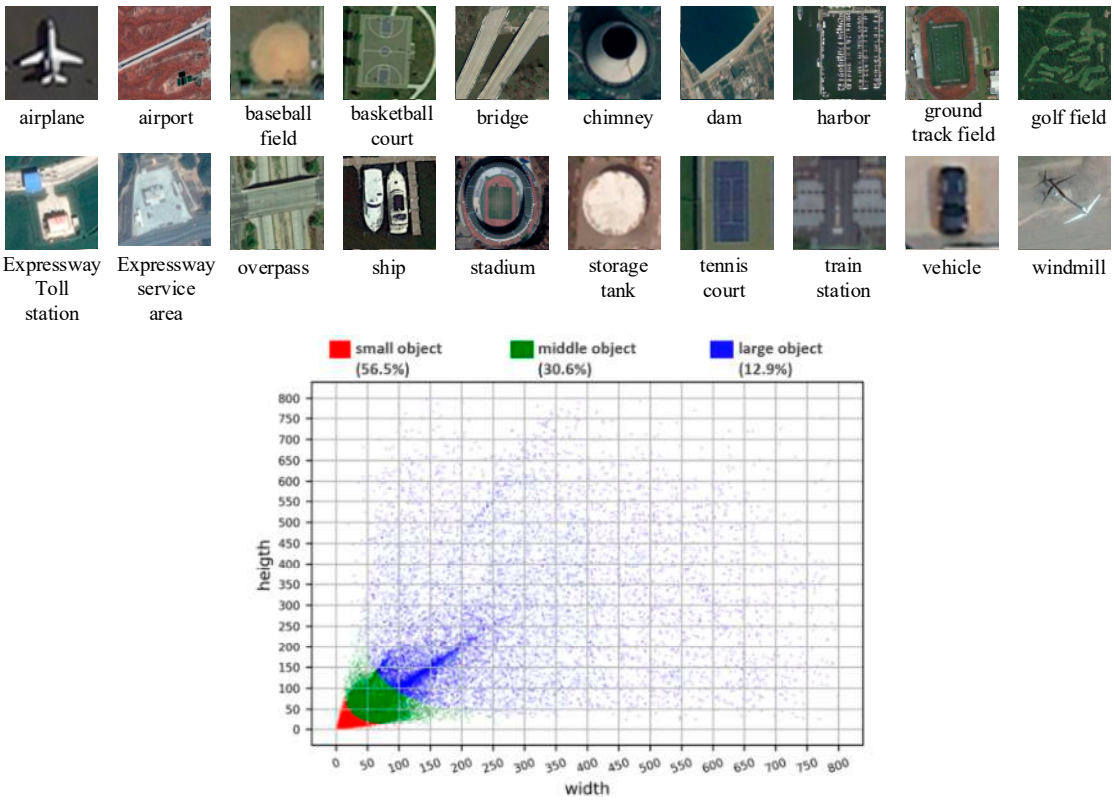


Figure 10. Each category and the size distributions of objects in DIOR.

Moreover, since images in DIOR are carefully collected under various environment conditions, such as different weathers and seasons, these images possess richer variations in viewpoint, background, occlusion, etc. Problems of intra-class diversity and intra-class similarity are more laborious due to the above characteristics. The main difficulties in real-world tasks can be well reflected by DIOR; thus, ablation experiments of YOLO-DSD and comparative experiments with SOTA detectors were conducted in DIOR dataset.

4.1.2. RSOD Dataset

RSOD [50] contains 976 images that have been clipped into approximately 1000 × 1000 pixels, and the spatial resolution of these images ranges from 0.3 m to 3 m. There are 6950 object instances in this dataset in total, covered by 4 common classes in ORSIs, including 4993 aircraft, 1586 oil tanks, 180 overpasses and 191 playgrounds. Each instance of classes is shown in Figure 11.

In addition, instances in RSOD dataset are under various scenes, including urban, grasslands, mountains, lakes, airport, etc. Although the scale of RSOD is not as large as that of DIOR, the characteristics of images in optical remote sensing object detection task can also be reflected by RSOD dataset. Therefore, we further analyzed the effectiveness of YOLO-DSD compared with YOLOv4 in RSOD dataset.

Figure 11. Each category of objects in RSOD.

4.2. Evaluation Indicator

Detectors in this study were analyzed from three perspectives, including detection accuracy, deployability and speed. The evaluation indicators of each performance are shown in Table 1. The higher the mAP and FPS, but the lower Params and Flops, the better the detector.

Table 1. The evaluation indicators.

Indicator Class	Indicator	Description
Accuracy	$mAP^{0.5}$ (%)	Average precision when IOU = 0.5. It is the most used indicator in remote sensing object detection.
	$mAP^{0.5:0.95}$ (%)	Mean values of mAPs under each IOU, which are taken at an interval of 0.05 between 0.5 and 0.95.
	mAP^{S}, mAP^{M}, mAP^{L} (%)	The $mAP^{0.5:0.95}$ of small, middle and large-sized object defined in MS COCO.
Deployability	Params	Number of detector parameters.
	Flops	Floating point operations.
Speed	FPS (img/s)	Frames transmitted per second.

4.3. Experiment Setting

In this study, the deep learning framework PyTorch1.7.1 was utilized to implement all of the detectors in this study. The experimental environment was ubuntu18.04, CUDA11.1, CUDNN8.0.5 and NVIDIA GeForce RTX 3080. In order to ensure enough training samples and to make the test set reflect the characteristics of each dataset well, training and test sets in DIOR were split by 1:1, whereas those in RSOD were split by 4:1 randomly. A total of 90% of the training set was utilized for training detectors, and 10% was used for monitoring to avoid overfitting. The input size and batch size of detectors was set to

416 × 416 and 7, respectively. Adam optimizer was employed to update the parameters, with a weight decay of 2×10^{-4}. The relationship between learning rate and epoch is shown in Figure 12. For anchor-based detectors, K-means was utilized to optimize the size of anchors before training.

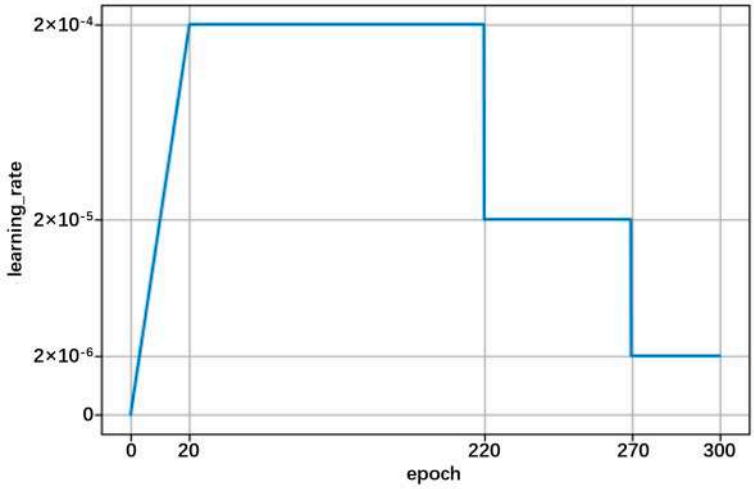


Figure 12. The relationship between learning rate and epoch.

4.4. Experiment Results and Discussion in DIOR Dataset

4.4.1. Ablation Experiment

Ablation experiments were conducted to verify the effectiveness of each improved module in YOLO-DSD, and the results are shown in Table 2. The detector improved with the DenseRes Block reduces Params by 23.9% ($\frac{48.81-64.17}{64.17} \times 100\%$) and Flops by 29.7% ($\frac{21.12-30.07}{30.07} \times 100\%$), and increases FPS by 63.4% ($\frac{65.7-40.2}{40.2} \times 100\%$), while achieving a 0.2% higher $mAP^{0.5}$ and almost the same $mAP^{0.5:0.95}$ compared with YOLOv4 as the baseline. The detector improved by S-CBL×5 in the neck based on "+DenseRes Block" is beneficial for $mAP^{0.5}$ and $mAP^{0.5:0.95}$, which are brought about by the increase in mAP^{M} and mAP^{L} without affecting the deployability and inference speed. However, the mAP^{S} slightly decreased by 0.3% because the short-cut utilized in S-CBL×5 strengthened the transmitting of the feature, and thus introduced background features additionally, which attenuated the representation of the feature for small-sized objects. The detector further improved by the DCA Block achieved a significant increase in mAP due to the enhancement of feature expression, and made up for the loss of mAP^{S} caused by the short-cut with the same Params and Flops, while the FPS was only slightly reduced by 5.3 img/s.

In summary, YOLO-DSD outperforms YOLOv4 both in the detection accuracy, deployability and speed evaluation indicator. YOLO-DSD based on YOLOv4 increases the commonly used indicator $mAP^{0.5}$ by 1.7% and the more rigorous indicator $mAP^{0.5:0.95}$ by 0.9%. Specifically, YOLO-DSD has a greater advantage in mAP^{M} and mAP^{L}, while it achieves a similar and competitive mAP^{S} compared with YOLOv4. In terms of deployability performance, the Params and Flops of YOLO-DSD decreased by 23.9% and 29.7% more than those of YOLOv4, respectively. YOLO-DSD also performs well in inference speed: it is 50.2% faster than YOLOv4 in FPS.

Table 2. The ablation results of YOLO-DSD.

Detectors	Params	Flops	FPS	$mAP^{0.5}$	$mAP^{0.5:0.95}$	mAP^{S}	mAP^{M}	mAP^{L}
YOLOv4(Baseline)	64.17 M	30.07 G	40.2	71.3	39.1	10.1	30.2	55.1
+DenseRes Block	48.81 M	21.12 G	65.7	71.5	38.8	9.4	30.4	54.9
+S-CBL×5	48.81 M	21.12 G	65.7	71.9	39.2	9.1	30.9	55.7
+DCA(YOLO-DSD)	48.81 M	21.12 G	60.4	73.0	40.0	9.6	31.6	56.4

We further analyzed the performance of the DenseRes Block. The ablation results of the DenseRes Block are shown in Table 3. The structure of the DenseRes Block in each detector is shown in Figure 13. Model 1 is the detector improved by the DenseRes Block without the structure of the 'Short-cut' and 'Combine'. 'Short-cut' and 'Combine' are introduced to the DenseRes Block in Model 2 and Model 3, respectively. Model 4 utilizes the complete DenseRes Block to improve the backbone of YOLOv4. From the comparison between Model 1 and Model 2, the 'Short-cut' introduced to DenseRes Block for the mitigation of feature loss can improve the mAP of objects in each size. After adding the 'Combine' to DenseRes Block, Model 3 performs better on the middle and large-sized object, while the mAP^S decreases slightly by 0.1%. The possible reason for this is that the feature of the middle and large-sized object is obvious enough to build high-level semantic relevance with the background feature, while the feature of the small object is not obvious enough and thus it is easy for it to be overwhelmed. Model 4 improved by the complete DenseRes Block achieves the highest mAP and a significant increase in mAP^S, mAP^M and mAP^L. It is probable that, on the basis of the 'Short-cut', the feature of each size object can be better retained when transmitting in the DenseRes Block, and can thus benefit the building of high-level semantic relevance with a background feature through 'Combine'.

Table 3. The ablation results of DenseRes Block.

Detectors	DenseRes Block		FPS	$mAP^{0.5}$	$mAP^{0.5:0.95}$	mAP^S	mAP^M	mAP^L
	+Short-Cut*	+Combine*						
Model 1			65.7	70.8	38.2	9.1	29.9	54.4
Model 2	√		65.7	71.3	38.7	9.5	30.3	54.7
Model 3		√	65.7	71.2	38.5	9.0	30.2	54.9
Model 4	√	√	65.7	71.5	38.8	9.4	30.4	54.9

Note: 'Short-cut*' indicates a short-cut to connect y_i $(1 < i \leq n)$ and $y_{(i-1)}$ $(1 < i \leq n)$ in DenseRes Block; 'Combine*' means $[y_i]$ $(1 \leq i \leq n) \oplus$ *Input* in DenseRes Block.

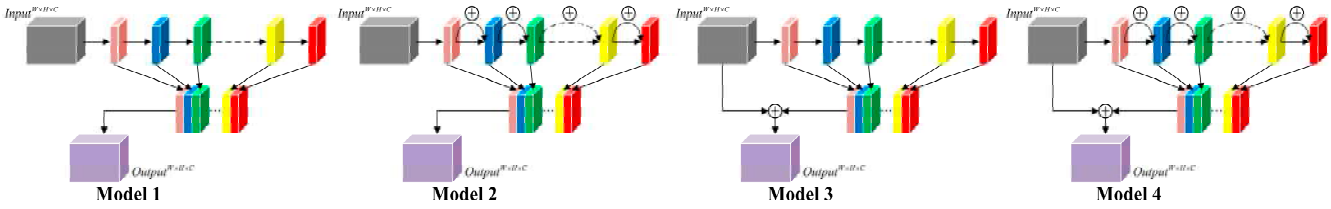


Figure 13. The structure of DenseRes Block in each detector in Table 3.

The experimental results of DCA module ablation are shown in Tables 4 and 5. Table 4 shows the influence of scaling factor R on the performance of the DCA Block. The results show that, when R = 32, DCA can achieve the best performance. Table 5 exhibits the influence of three different fusion methods shown in Figure 14 on the performance of the DCA Block. The results show that the DCA Block with a different fusion method can effectively improve the detection accuracy. Specifically, compared with DCA in series, DCA in parallel has a more obvious advantage in small and middle-sized objects, while the FPS is slightly reduced by 0.7 img/s. This may be due to the fact that, when employing the same number of operation layers in one building block, although the structure designed in parallel has a higher fragment, it can keep the integrity of the feature better compared with that in series. For the proposed DCA Block, which has a small structure complexity, utilizing the structure in parallel makes it perform better in the enhancement of feature expression without an obvious sacrifice of inference time.

Table 4. Results of different zoom factor 'R's in DCA Block.

Detectors	DCA Block	FPS	$mAP^{0.5}$	$mAP^{0.5:0.95}$	mAP^{S}	mAP^{M}	mAP^{L}
	Zoom Factor 'R'						
Model 1	R = 1	60.3	72.4	39.3	9.2	31.1	55.8
Model 2	R = 2	60.3	72.2	39.2	9.1	30.3	55.9
Model 3	R = 4	60.3	71.8	39.0	9.2	30.1	55.9
Model 4	R = 8	60.4	72.8	39.8	9.7	31.1	56.4
Model 5	R = 16	60.4	72.3	39.5	9.4	31.2	56.2
Model 6	R = 32	60.4	73.0	40.0	9.6	31.6	56.4
Model 7	R = 64	60.4	72.1	39.6	9.2	31.4	56.1

Table 5. Results of different fusion forms in DCA Block.

Detectors	DCA Block	FPS	$mAP^{0.5}$	$mAP^{0.5:0.95}$	mAP^{S}	mAP^{M}	mAP^{L}
	Fusion Form						
Model 1	'Global Path' + 'Local Path' (In series)	61.1	72.2	39.7	9.3	31.2	56.5
Model 2	'Local Path' + 'Global Path' (In series)	61.1	72.0	39.5	9.3	30.5	56.6
Model 3	'Global Path' + 'Local Path' (In parallel)	60.4	73.0	40.0	9.6	31.6	56.4

Note: 'Global Path' and 'Local Path' refer to 'Global Extraction Path' and 'Local Extraction Path', respectively.

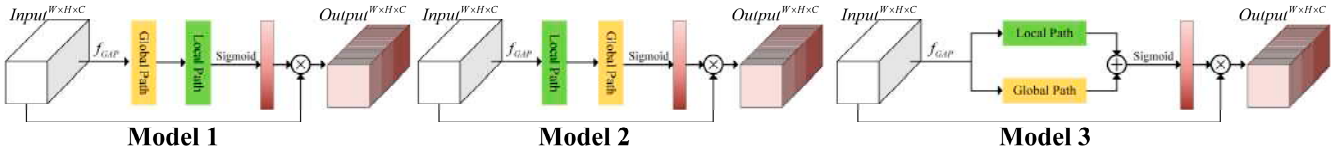


Figure 14. The structure of DCA Block in each detector in Table 5.

4.4.2. Comparative Experiment

Four experiments were conducted in this study to verify the superiority of the proposed method. (1) ResNet50 [34], VGG16 [33] and the backbones that were established based on the CSP DarkNet framework with different feature extraction modules, including the Res Block [34], ResNeXt Block [35], Res2 Block [36], Dense Block [27], CSP Block [37] and DenseRes Block, were compared. (2) A comparison of different neck structures, including FPN [16], BFPN [40], PANet [41], S-PANet (PANet improved with the proposed S-CBL×5) and none (without feature pyramid structure), was conducted. (3) The performance of different attention mechanisms, including the SE Block [42], ECA Block [43], CA Block [44], CBAM Block [45] and DCA Block(R = 32), was compared and analyzed. (4) YOLO-DSD was compared with eight SOTA detectors, including Faster-RCNN, SSD, RetinaNet, YOLOv3, YOLOv4, YOLO-Lite (MobileNetV2 [51]—YOLOv4), CenterNet [7] and EfficientDet [9], which have been widely applied in various natural scene visual detection tasks due to their acceptable tradeoff between accuracy, deployability and inference time.

Comparative experiment for different backbones: The performances of the CSP DarkNet, which is improved by the proposed DenseRes Block (DarkNet-DenseRes) and other backbones, are demonstrated in Table 6. Based on the CSP DarkNet framework, the proposed DenseRes Block outperforms the ResNeXt Block and Dense Block in all indicators. Although the $mAP^{0.5}$ and $mAP^{0.5:0.95}$ of DarkNet-DenseRes are slightly lower than those of DarkNet-Res by 0.1% and 1.3%, the Params and Flops of DarkNet-DenseRes are only approximately 1/3 and 1/4 of DarkNet-Res, while the FPS of DarkNet-DenseRes is approximately 1/4 higher than that of DarkNet-Res. Similarly, the $mAP^{0.5}$ and $mAP^{0.5:0.95}$ of DarkNet-DenseRes are 0.9% and 1.1% lower than those of DarkNet-Res2; however, the Params and Flops of DarkNet-DenseRes are only approximately 1/3 and 1/2 of those of DarkNet-Res2, while the interference speed is 2.3 times that of DarkNet-Res2 according to FPS. The superiority of DarkNet-DenseRes compared with CSP DarkNet was analyzed and

proved in ablation experiments. DarkNet-DenseRes also has obvious advantage in all indicators compared with ResNet50. Although DarkNet-DenseRes has a similar accuracy and speed to VGG16, VGG16 has seven times as much Flops than that of DarkNet-DenseRes. Therefore, DarkNet-DenseRes achieves the optimal balance of accuracy, deployability and speed.

Table 6. Results of comparative experiment for different backbones.

Backbone	Params	Flops	FPS	$mAP^{0.5}$	$mAP^{0.5:0.95}$	mAP^{S}	mAP^{M}	mAP^{L}
CSP DarkNet (Baseline)	26.61 M	17.34 G	40.2	71.3	39.1	10.1	30.2	55.1
DarkNet-Res [1]	40.58 M	24.61 G	52.8	71.6	40.1	9.9	31.7	56.1
DarkNet-ResNeXt [2]	20.55 M	12.71 G	39.1	68.4	36.4	8.2	28.3	52.6
DarkNet-Res2 [3]	31.65 M	19.33 G	28.4	72.4	39.9	10.2	31.6	55.4
DarkNet-Dense [4]	14.06 M	8.16 G	50.9	69.6	37.5	7.8	29.7	54.5
DarkNet-DenseRes [5](Ours)	11.26 M	8.42 G	65.7	71.5	38.8	9.4	30.4	54.9
ResNet50	23.51 M	13.41 G	47.4	68.5	36.5	7.8	28.7	53.2
VGG16	17.07 M	54.64 G	70.1	71.1	38.9	9.9	29.8	55.2

Note: [1, 2, 3, 4, 5] means CSP DarkNet utilizes Res Block, ResNeXt Block, Res2 Block, Dense Block and DenseRes Block as the main feature extraction module, respectively.

Comparative experiment for different necks: Table 7 shows the performance of each neck structure that was tested by applying a no-feature pyramid structure (None), FPN, BFPN, PANet (Baseline) and S-PANet to the modified YOLOv4, with the DenseRes Block in the backbone. 'None' has the lowest Params (18.83 M) and Flops (4.89 G) and the highest FPS (85.5 img/s), but it does not perform well in detection accuracy, and, in particular, its mAP^{S} is only 8.1%, whereas that of the other four necks ranges from 9.1% to 9.5%. Therefore, the feature pyramid structure is vital for detection accuracy and, in particular, for small size objects, which occupy more than 50% in DIOR. Although FPN and BFPN are slightly better than PANet in deployability and inference speed, they have more than a 2.6% inferiority in mAP of middle and large-sized objects, which, in total, account for approximately 50% of objects in DIOR. It was proven that the structure of PANet is important to the detection accuracy in YOLOv4 for ORSIs. PANet and S-PANet have almost the same Params, Flops and FPS, but our S-PANet performs better than PANet in $mAP^{0.5}$ and $mAP^{0.5:0.95}$. In conclusion, S-PANet is more suitable for optical remote sensing object detection than other necks.

Table 7. Results of comparative experiment for different necks.

Neck	Params	Flops	FPS	$mAP^{0.5}$	$mAP^{0.5:0.95}$	mAP^{S}	mAP^{M}	mAP^{L}
None	18.83 M	4.89 G	85.5	68.3	35.5	8.1	27.4	51.6
FPN	27.22 M	8.50 G	71.8	69.1	36.0	9.2	27.2	51.4
BFPN	35.68 M	10.84 G	68.1	69.9	36.8	9.5	27.8	52.1
PANet (Baseline)	37.55 M	12.73 G	65.7	71.5	38.8	9.4	30.4	54.9
S-PANet(ours)	37.55 M	12.73 G	65.7	71.9	39.2	9.1	30.9	55.7

Comparative experiments for different attention mechanisms: Taking modified YOLOv4 with the DenseRes Block in the backbone and S-PANet in the neck as the baseline (None), the indicator values of different attention mechanisms are exhibited and compared in Table 8. The CA Block and CBAM Block containing the spatial attention mechanism fail to improve the detection accuracy, and the FPS decreases significantly due to those complex structures. Most channel attention mechanisms, including the SE Block, ECA Block and DCA Block, can improve the detection accuracy. The DCA Block improves the detection accuracy for small, medium and large sizes of objects, and achieves the highest $mAP^{0.5}$ = 73.0% and $mAP^{0.5:0.95}$ = 40.0%, with an increase of 1.1% and 0.8% compared with 'None', respectively, when R = 32, and the FPS only decreases by 5.3 img/s. In the case of the SE Block, $mAP^{0.5}$ and $mAP^{0.5:0.95}$ increases by 0.2% and 0.1%, and the FPS decreases by 3.4 img/s. The ECA Block improves both $mAP^{0.5}$ and $mAP^{0.5:0.95}$ by 0.1%, and decreases

the FPS by 2.8 img/s. Therefore, the proposed DCA Block can achieve the best balance between accuracy and speed.

Table 8. Results of comparative experiment for different attention mechanisms.

Attention Mechanism	Params	Flops	FPS	$mAP^{0.5}$	$mAP^{0.5:0.95}$	mAP^{S}	mAP^{M}	mAP^{L}
None (Baseline)	0	0	65.7	71.9	39.2	9.1	30.9	55.7
CA	42.36 K	1126.41 K	57.2	71.6	39.0	9.1	30.4	55.9
CBAM	102.79 K	516.59 K	52.8	70.6	38.1	8.4	29.3	55.7
SE	51.20 K	51.272 K	62.3	72.1	39.3	9.0	30.8	56.5
ECA	0.02 K	0.02 K	62.9	72.0	39.3	9.2	30.7	56.1
DCA (R = 32)	51.22 K	830.24 K	60.4	73.0	40.0	9.6	31.6	56.4

Comparative experiments for different detectors: The performances of the proposed YOLO-DSD and eight SOTA detectors are demonstrated in Table 9. RetinaNet and EfficientDet have a better deployability than YOLO-DSD, but their detection accuracy, especially for small-sized objects and speed, are far behind that of YOLO-DSD, so this hinders the application of these detectors in optical remote sensing object detection. The large Flops of SSD and Faster-RCNN require a huge amount of computing resources, which greatly increases the difficulty in deploying them on edge devices. Although the Params and Flops of CenterNet are 67% and 69% that of YOLO-DSD, and the FPS is 46% faster, the detection accuracy of CenterNet is significantly lower than that of YOLO-DSD ($mAP^{0.5:0.95}$:35.8% vs. 40.0%), and the mAP^{S} is only 62.5% that of YOLO-DSD. YOLO-Lite has an obvious disadvantage in detection accuracy for small and large-sized objects, even though it has a better deployability compared with YOLO-DSD. The inference speed of YOLOv3 is nearly the same as that of YOLO-DSD, but the deployability and detection accuracy of YOLOv3 are obviously inferior to that of YOLO-DSD. The superiority of YOLO-DSD compared with YOLOv4 was analyzed and proved in ablation experiments. Therefore, YOLO-DSD outperforms other SOTA detectors in the balance of accuracy, deployability and speed.

Table 9. Results of comparative experiment for different detectors.

Detector	Params	Flops	FPS	$mAP^{0.5}$	$mAP^{0.5:0.95}$	mAP^{S}	mAP^{M}	mAP^{L}
RetinaNet	36.72 M	17.24 G	44.6	62.7	37.6	4.8	30.9	57.5
EfficientDet	3.60 M	1.30 G	18.1	50.4	29.4	2.4	24.9	46.0
SSD	26.15 M	59.59 G	87.3	61.9	37.8	4.6	31.0	58.2
CenterNet	32.67 M	14.62 G	88.5	61.4	35.8	6.0	27.3	55.3
Faster-RCNN	28.47 M	364.14 G	21.9	56.1	31.8	2.8	23.7	53.2
YOLO-Lite	10.48 M	3.89 G	54.1	64.5	33.1	6.5	26.1	48.7
YOLOv3	61.63 M	32.83 G	61.4	69.2	34.7	7.8	27.4	49.8
YOLOv4	64.17 M	30.07 G	40.2	71.3	39.1	10.1	30.2	55.1
YOLO-DSD (ours)	48.81 M	21.12 G	60.4	73.0	40.0	9.6	31.6	56.4

Figures 15–17 exhibit the detection performance of Faster-RCNN, CenterNet, YOLOv4 and YOLO-DSD on DIOR. The detection result of the small-sized instance in Figure 15 indicates that both Faster-RCNN and CenterNet obviously miss detection. Although YOLOv4 could completely detect airplanes, it incorrectly detected a storage tank. Our YOLO-DSD can correctly detect all airplanes without any false detection. Figure 16 presents the detection results of an instance in the complex urban background. We can see that Faster-RCNN only detects one ground track field, and that CenterNet misses two bridges and two ground track fields and misdetects an overpass. YOLOv4 misses one bridge and one ground track field, whereas YOLO-DSD detects all objects correctly. The detection results of instances in a complex suburban background are given in Figure 17. It can be seen that Faster-RCNN detects only one Expressway-Service-Area, CenterNet has two false detections of an overpass and windmill, YOLOv4 detects two Expressway-Service-

Areas as one, and YOLO-DSD correctly detects all objects. The above instances verify that YOLO-DSD can handle object detection under different complex backgrounds well.

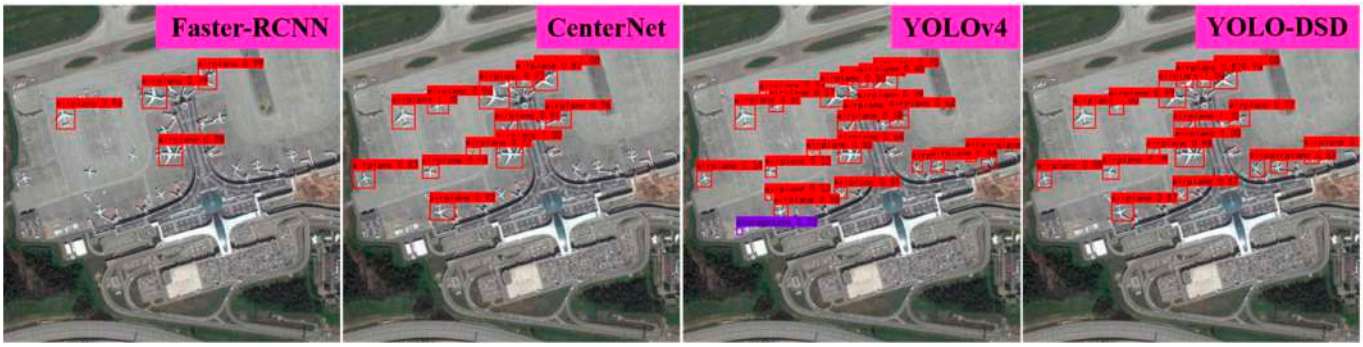


Figure 15. The detection result of small-sized instance.

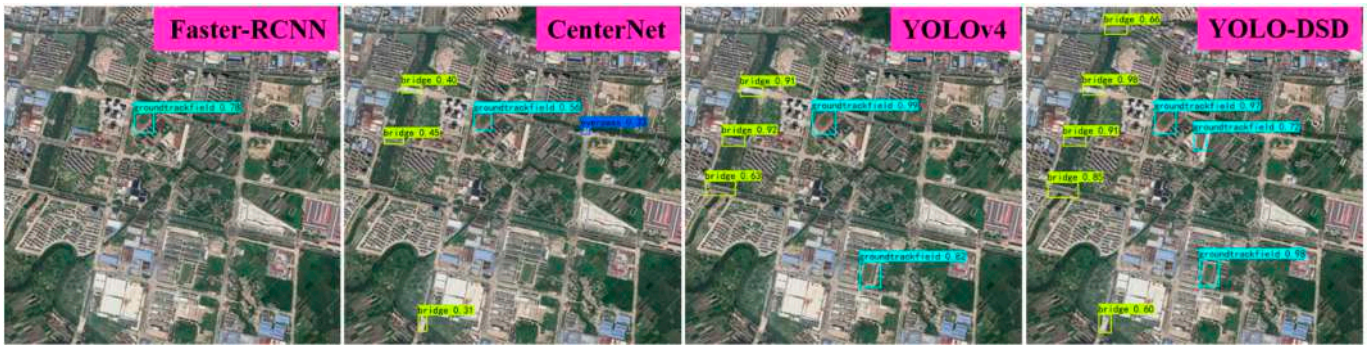


Figure 16. The detection result of instance in complex suburban background.

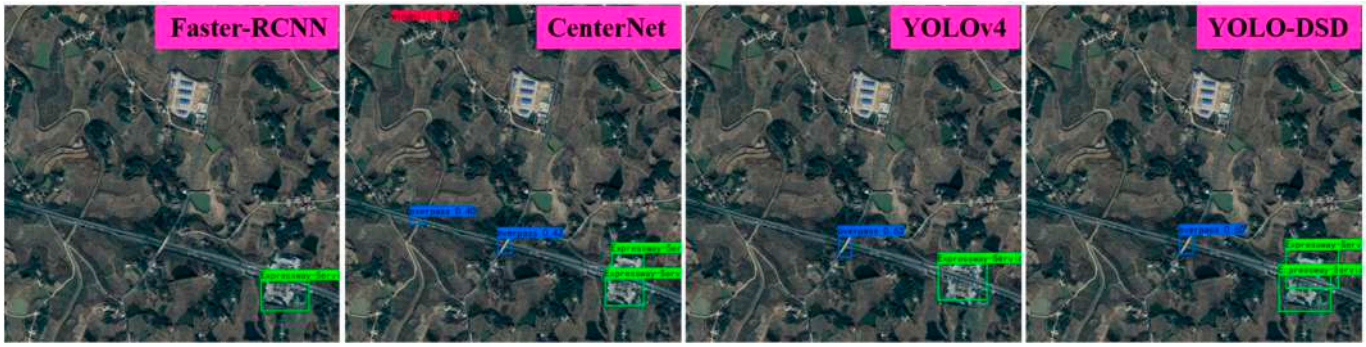


Figure 17. The detection result of instance in complex urban background.

The precision–recall curves and AP (IOU = 0.5) of YOLOv4 and YOLO-DSD in each category are given in Figure 18 for a better illustration of the difference in detection accuracy. It can be seen that YOLO-DSD detects better than YOLOv4 in 11 categories, including airplane, airport, baseball field, chimney, dam, Expressway-Service-Area, golffield, groundtrackfield, stadium, storagetank and transtation. In particular, the AP of YOLO-DSD in airport, baseballfield, Expressway-Service-Area and groundtrackfield is over 2% higher than that of YOLOv4. The AP of YOLO-DSD in airplane, transtation and stadium significantly increase by 6.63%, 5.21% and 17.02%, respectively. For the other nine categories, YOLO-DSD only slightly decreases by 0.35~1.78% compared with YOLOv4 in AP, but still has a competitive accuracy. Therefore, YOLO-DSD has a better accuracy performance than YOLOv4 in the large-scale ORSIs dataset DIOR in total.

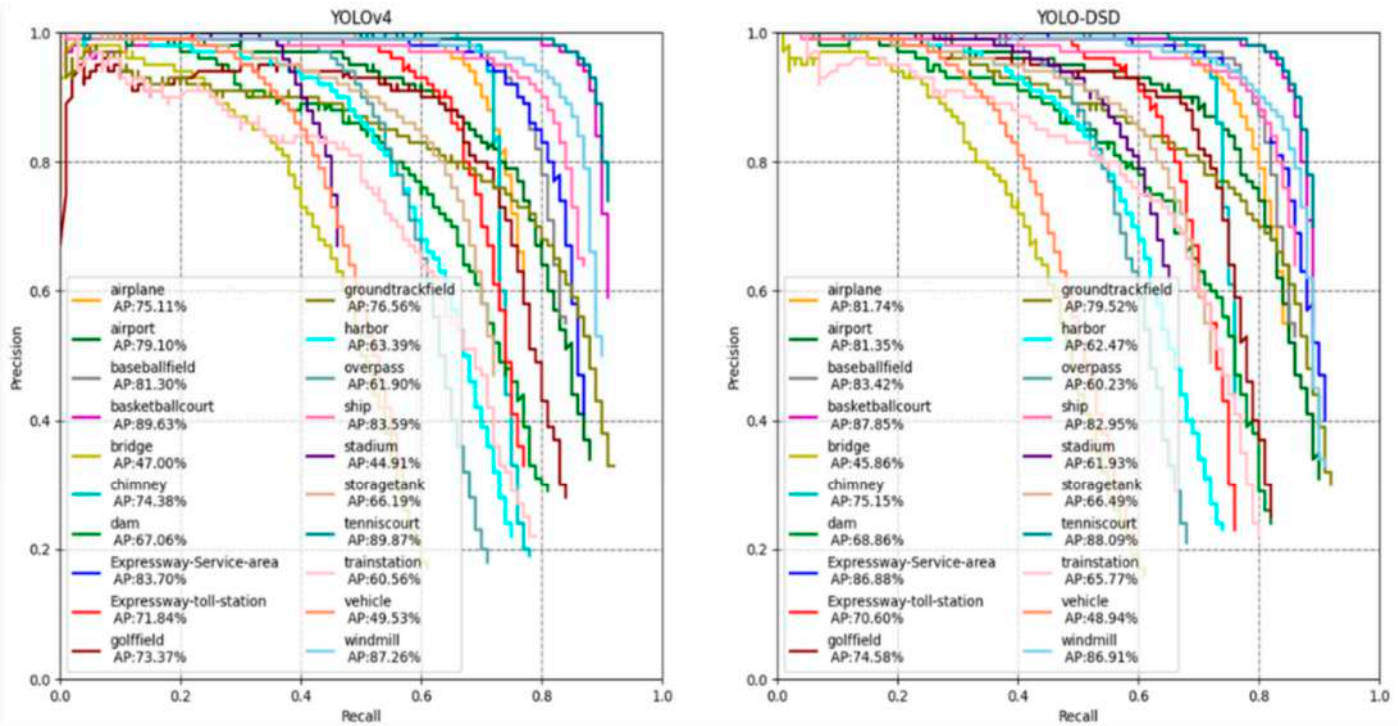


Figure 18. The precision–recall curves and AP (IOU = 0.5) of YOLOv4 and YOLO-DSD in each category.

4.5. Experiment Results and Discussion in RSOD Dataset

In order to further exhibit the superiority of the proposed YOLO-DSD based on YOLOv4 in optical remote sensing object detection application, another comparison experiment between YOLO-DSD and YOLOv4 was conducted in a four-category dataset, RSOD [50], which contained aircraft, oil tank, playground and overpass. The experiment result is shown in Table 10. It can be seen that YOLO-DSD outperforms in accuracy and inference time compared with YOLOv4 under different input sizes, including 416 × 416, 512 × 512 and 608 × 608. Specifically, YOLO-DSD increases $mAP^{0.5}$, $mAP^{0.5:0.95}$ and FPS by 2.6%, 0.8% and 50.2%, respectively, under the input size 416 × 416, while, under the input size 512 × 512, $mAP^{0.5}$, $mAP^{0.5:0.95}$ and FPS are improved by 2.1%, 1.9% and 54.9%, respectively. In terms of the input size 608 × 608, the $mAP^{0.5}$, $mAP^{0.5:0.95}$ and FPS of YOLO-DSD are 1.5%, 1.2% and 59.3% higher than those of YOLOv4.

However, it is noteworthy that the overpass AP of YOLO-DSD is higher than that of YOLOv4 in RSOD, whereas it is the opposite in DIOR. One possible reason for this is that 'bridge' and 'overpass' possess a significant inter-class similarity and thus interfere with the detection performance of YOLO-DSD in these two categories in DIOR. Therefore, how to overcome the inter-class similarity between 'bridge' and 'overpass' for a better detection accuracy while keeping its deployability and inference speed is one of our future works.

Table 10. Results of comparative experiment for YOLOv4 and YOLO-DSD in RSOD.

Detector	Input Size	FPS	$AP^{0.5}$				$mAP^{0.5}$	$mAP^{0.5:0.95}$
			Aircraft	Oil Tank	Playground	Overpass		
YOLOv4	416 × 416	40.2	97.8	95.7	99.4	67.2	90.0	52.1
YOLO-DSD		60.7	98.0	98.2	99.6	74.4	92.6	52.9
YOLOv4	512 × 512	37.1	98.1	97.5	99.5	73.5	92.2	55.1
YOLO-DSD		57.5	98.5	98.6	99.8	80.1	94.3	57.0
YOLOv4	608 × 608	34.9	98.2	98.5	99.9	79.2	94.0	58.5
YOLO-DSD		55.6	99.1	98.9	99.9	84.2	95.5	59.7

5. Conclusions

In this study, a new detector, YOLO-DSD, based on YOLOv4, was proposed to balance the accuracy, deployability and inference time for remote sensing object detection. Three main improvements were utilized in YOLO-DSD, including the DenseRes Block, S-CBL×5 and DCA Block. Firstly, the DenseRes Block improves the backbone, which can better compress and extract the object feature with a high accuracy but less computational consumption. Secondly, S-CBL×5 introduced in the neck can mitigate feature loss without increasing the consumption and inference time. Finally, a new channel attention mechanism, the DCA Block, added to S-CBL×5 better highlights the important features in the channel dimension.

Experiments on a large dataset, DIOR, were conducted to analyze the detection performance from the accuracy (mAP), deployability (Params and Flops) and speed (FPS). The results of the experiments indicate that the proposed DenseRes Block is superior to other feature extraction modules, such as the Res Block, ResNeXt Block, Res2 Block, Dense Block and CSP Block. Moreover, S-CBL×5 performs better than currently widely used FPN, BFPN and PANet. In addition, the proposed DCA Block outperforms other attention mechanisms, including the SE Block, ECA Block, CA Block and CBAM Block. Compared with YOLOv4, YOLO-DSD reduces Params by 23.9% and Flops by 29.7%, but increases FPS by 50.2%, while $mAP^{0.5}$ and $mAP^{0.5:0.95}$ increased from 71.3% to 73.0% and 39.1% to 40.0%, respectively. Compared with other SOTA detectors, including Faster-RCNN, SSD, RetinaNet, YOLOv3, YOLOv4, CenterNet, YOLO-Lite and EfficientDet, YOLO-DSD achieves the optimal balance of accuracy, deployability and inference time. In terms of the RSOD dataset, compared with YOLOv4, YOLO-DSD achieves 1.5~2.6%, 0.8~1.2% and 50.2~59.3% increases in $mAP^{0.5}$, $mAP^{0.5:0.95}$ and FPS under different input sizes, including 416 × 416, 512 × 512 and 608 × 608.

However, YOLO-DSD has a limitation in processing a serious inter-class similarity, such as 'bridge' and 'overpass', compared with YOLOv4. In order to further improve the performance of the proposed detector, we will try to combine depthwise separable convolution [52] with the proposed DenseRes Block for a better feature extraction and deployability reduction. Moreover, other non-consumption methods, such as image preprocessing and anchor optimization, will be considered to improve the detector.

Author Contributions: Conceptualization, S.L. and H.C.; methodology, S.L. and H.C.; software, H.C.; validation, H.C. and H.J.; formal analysis, S.L. and H.C.; investigation, S.L. and H.C.; resources, S.L. and H.C.; data curation, H.C.; writing—original draft preparation, S.L. and H.C.; writing—review and editing, H.J.; visualization, H.C.; supervision, S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Natural Science Foundation of Guangdong, China with grant number 2021A1515012395, and was supported by earmarked fund for China Agriculture Research System, grant number CARS-17.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used during the study have been uploaded at: https://gcheng-nwpu.github.io/#Datasets (last accessed on 27 July 2022) and https://github.com/RSIA-LIESMARS-WHU/RSOD-Dataset- (last accessed on 27 July 2022).

Acknowledgments: We gratefully appreciate the editor and anonymous reviewers for their efforts and constructive comments, which have greatly improved the technical quality and presentation of this study.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. The architecture and complexity of DarkNet-DenseRes.

Stage	Output Size	Operation	Number	Params	Flops
CBM	416 × 416 × 32	Conv-BN-Mish (k = 3 × 3, c = 32, s = 1) *	1	928	166,133,760
DenseRes Layer_1	208 × 208 × 64	Conv-BN-Mish (k = 3 × 3, c = 64, s = 2)	1	18,560	805,748,736
	208 × 208 × 64	Conv-BN- Leaky ReLu (k = 3 × 3, c = 64, s = 1)	1	36,992	1,603,190,784
	208 × 208 × 64	Concatenation	1	/	/
DenseRes Layer_2	104 × 104 × 128	Conv-BN-Mish (k = 3 × 3, c = 128, s = 2)	1	73,984	801,595,392
	104 × 104 × 64	Conv-BN- Leaky ReLu (k = 3×3, c = 64, s = 1)	2	110,848	1,200,316,416
	104 × 104 × 128	Concatenation	1	/	/
DenseRes Layer_3	52 × 52 × 256	Conv-BN-Mish (k = 3 × 3, c = 256, s = 2)	1	295,424	799,518,720
	52 × 52 × 32	Conv-BN- Leaky ReLu (k = 3 × 3, c = 32, s = 1)	8	138,752	375,861,632
	52 × 52 × 256	Concatenation	1	/	/
DenseRes Layer_4	26 × 26 × 512	Conv-BN-Mish (k = 3 × 3, c = 512, s = 2)	1	1,180,672	798,480,384
	26 × 26 × 64	Conv-BN- Leaky ReLu (k = 3 × 3, c = 64, s = 1)	8	553,984	374,839,296
	26 × 26 × 512	Concatenation	1	/	/
DenseRes Layer_5	13 × 13 × 1024	Conv-BN-Mish (k = 3 × 3, c = 1024, s = 2)	1	4,720,640	797,961,216
	13 × 13 × 256	Conv-BN- Leaky ReLu (k = 3 × 3, c = 256, s = 1)	4	4,130,816	698,280,960
	13 × 13 × 1024	Concatenation	1	/	/
Total Params		11,261,600			
Total Flops		8,421,927,296			

Note: * k, c, and s mean kernel size, output channels and stride of the convolution layer, respectively.

References

1. Cheng, G.; Han, J.W. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm* **2016**, *117*, 11–28. [CrossRef]
2. Li, K.; Wan, G.; Cheng, G.; Meng, L.Q.; Han, J.W. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm* **2020**, *159*, 296–307. [CrossRef]
3. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
4. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
5. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
6. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
7. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
8. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
9. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
10. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
11. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Boston, MA, USA, 7–12 June 2015; pp. 1440–1448.
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Processing Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
13. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2961–2969.
14. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
15. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
16. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.

17. Al Ridhawi, I.; Bouachir, O.; Aloqaily, M.; Boukerche, A. Design Guidelines for Cooperative UAV-supported Services and Applications. *ACM Comput. Surv.* **2022**, *54*, 1–35. [CrossRef]
18. Xu, D.Q.; Wu, Y.Q. MRFF-YOLO: A Multi-Receptive Fields Fusion Network for Remote Sensing Target Detection. *Remote Sens.* **2020**, *12*, 3118. [CrossRef]
19. Cheng, G.; Si, Y.J.; Hong, H.L.; Yao, X.W.; Guo, L. Cross-Scale Feature Fusion for Object Detection in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 431–435. [CrossRef]
20. Yin, W.; Diao, W.; Wang, P.; Gao, X.; Li, Y.; Sun, X. PCAN—Part-based context attention network for thermal power plant detection in remote sensing imagery. *Remote Sens.* **2021**, *13*, 1243. [CrossRef]
21. Yuan, Z.C.; Liu, Z.M.; Zhu, C.B.; Qi, J.; Zhao, D.P. Object Detection in Remote Sensing Images via Multi-Feature Pyramid Network with Receptive Field Block. *Remote Sens.* **2021**, *13*, 862. [CrossRef]
22. Li, Z.L.; Zhao, L.N.; Han, X.; Pan, M.Y. Lightweight Ship Detection Methods Based on YOLOv3 and DenseNet. *Math. Probl. Eng.* **2020**, *2020*, 4813183. [CrossRef]
23. Huyan, L.; Bai, Y.P.; Li, Y.; Jiang, D.M.; Zhang, Y.N.; Zhou, Q.; Wei, J.Y.; Liu, J.N.; Zhang, Y.; Cui, T. A Lightweight Object Detection Framework for Remote Sensing Images. *Remote Sens.* **2021**, *13*, 683. [CrossRef]
24. Lang, L.; Xu, K.; Zhang, Q.; Wang, D. Fast and Accurate Object Detection in Remote Sensing Images Based on Lightweight Deep Neural Network. *Sensors* **2021**, *21*, 5460. [CrossRef]
25. Li, Y.Y.; Mao, H.T.; Liu, R.J.; Pei, X.; Jiao, L.C.; Shang, R.H. A Lightweight Keypoint-Based Oriented Object Detection of Remote Sensing Images. *Remote Sens.* **2021**, *13*, 2459. [CrossRef]
26. Huang, W.; Li, G.Y.; Chen, Q.Q.; Ju, M.; Qu, J.T. CF2PN: A Cross-Scale Feature Fusion Pyramid Network Based Remote Sensing Target Detection. *Remote Sens.* **2021**, *13*, 847. [CrossRef]
27. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
28. Qin, Z.; Li, Z.; Zhang, Z.; Bao, Y.; Yu, G.; Peng, Y.; Sun, J. ThunderNet: Towards real-time generic object detection on mobile devices. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 6718–6727.
29. He, H.; Huang, X.; Song, Y.; Zhang, Z.; Wang, M.; Chen, B.; Yan, G. An insulator self-blast detection method based on YOLOv4 with aerial images. *Energy Rep.* **2022**, *8*, 448–454. [CrossRef]
30. Roy, A.M.; Bose, R.; Bhaduri, J. A fast accurate fine-grain object detection model based on YOLOv4 deep neural network. *Neural Comput. Appl.* **2022**, *34*, 3895–3921. [CrossRef]
31. Song, W.; Fu, C.; Zheng, Y.; Cao, L.; Tie, M.; Sham, C.W. Protection of image ROI using chaos-based encryption and DCNN-based object detection. *Neural Comput. Appl.* **2022**, *34*, 5743–5756. [CrossRef]
32. Gu, Y.; Si, B.J.E. A novel lightweight real-time traffic sign detection integration framework based on YOLOv4. *Entropy* **2022**, *24*, 487. [CrossRef]
33. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
35. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
36. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef]
37. Wang, C.-Y.; Liao, H.-Y.M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; Yeh, I.-H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 390–391.
38. Xu, C.; Li, C.; Cui, Z.; Zhang, T.; Yang, J. Hierarchical Semantic Propagation for Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4353–4364. [CrossRef]
39. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Salt Lake City, UT, USA, 18–23 June 2018; pp. 116–131.
40. Zhang, X.; Wan, T.; Wu, Z.; Du, B. Real-time detector design for small targets based on bi-channel feature fusion mechanism. *Appl. Intell.* **2022**, *52*, 2775–2784. [CrossRef]
41. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
42. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
43. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
44. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.

45. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3–19.
46. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, Seattle, WA, USA, 13–19 June 2020; pp. 12993–13000.
47. Dai, W.; Li, D.; Tang, D.; Jiang, Q.; Wang, D.; Wang, H.; Peng, Y. Deep learning assisted vision inspection of resistance spot welds. *J. Manuf. Processes* **2021**, *62*, 262–274. [CrossRef]
48. Tian, R.; Jia, M. DCC-CenterNet: A rapid detection method for steel surface defects. *Measurement* **2022**, *187*, 110211. [CrossRef]
49. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Trans. Cybern.* **2021**, *52*, 8574–8586. [CrossRef]
50. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
51. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
52. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.

Computing (2022) 104:2489–2509
https://doi.org/10.1007/s00607-022-01083-4

REGULAR PAPER

A hybrid teaching-learning-based optimization algorithm for QoS-aware manufacturing cloud service composition

Hong Jin[1] · Cheng Jiang[1] · Shengping Lv[1] · Haiping He[1] · Xinting Liao[1]

Received: 15 September 2021 / Accepted: 12 April 2022 / Published online: 23 June 2022

Abstract
Quality of service (QoS)-aware manufacturing cloud service composition (QoS-MCSC) is one of the key issues in Cloud manufacturing (CMfg). More and more manufacturing cloud services offering the same or similar functionality but different QoS attributes are provided in the CMfg platform. It is a challenging issue to construct an optimal composite service satisfying customers' requirements. In this study, a novel hybrid teaching-learning-based optimization algorithm is proposed to solve QoS-MCSC problems. It integrates the advantages of uniform mutation, adaptive flower pollination algorithm, and teaching-learning-based optimization algorithm. The experimental results show that the proposed algorithm finds higher quality results than other compared algorithms.

Keywords Cloud manufacturing · Service composition · Quality of service · Hybrid teaching-learning-based optimization

Mathematics Subject Classification 90-08 · 90C27 · 68T20

Cheng Jiang and Shengping Lv have contributed equally to this work.

✉ Shengping Lv
lvshengping@scau.edu.cn

Hong Jin
hjin@scau.edu.cn

Cheng Jiang
20193142013@stu.scau.edu.cn

Haiping He
hehaiping@stu.scau.edu.cn

Xinting Liao
295246971@qq.com

[1] College of Engineering, South China Agricultural University, Guangzhou 510642, China

1 Introduction

As a new service-oriented networked manufacturing model, Cloud manufacturing (CMfg) aims to provide convenient, on-demand, safe and reliable, cheap, and high-quality manufacturing services to users for the whole life cycle of manufacturing [1, 2]. The manufacturing resources are virtualized and encapsulated into manufacturing cloud services (MCSs), and then MCSs are published in the CMfg platform [3]. Each cloud service is described by some functional and non-functional attributes [4]. For instance, a functional attribute of an MCS may be the machining accuracy of a computer numerical control machine and a non-functional attribute can be the task execution time. The non-functional attribute is also named as Quality of Service (QoS). According to the functional requirements and QoS attributes, the service users select appropriate MCSs to complete manufacturing tasks.

A lots of MCSs with the same or similar functionality but different QoS attributes are provided in the CMfg platform. Some QoS attributes cannot be optimized at the same time [4]: a MCS may have a shorter execution time but a higher cost whereas another one might have a lower price but a longer execution time. The QoS value of a composite manufacturing cloud service (CMCS) depends on the MCSs selected in the composite service. Meanwhile, in the process of service composition, MCSs may have certain correlations with other services, and service correlations may affect the global QoS of the CMCS [5]. Therefore, how to select services to be integrated into the composite service is a challenging and complex issue.

The QoS-aware manufacturing cloud service composition (QoS-MCSC) problem is an NP-hard combinatorial optimization problem [6]. It is difficult to find the global (even local) optimal solution for this type of problem. Intelligent optimization algorithms, as a kind of stochastic optimization algorithm, shows a good performance in solving these kinds of problems. Some intelligent optimization algorithms, such as Genetic Algorithm (GA) [5], Particle Swarm Optimization (PSO) [7], Artificial Bee Colony (ABC) [8], Flower Pollination Algorithm (FPA) [9], and so on, are widely used to solve QoS-MCSC problems. However, proper selection of the initial parameters would be very difficult for searching the optimal solutions when these algorithms are utilized to solve different kinds of optimization problems.

Teaching-Learning-Based Optimization (TLBO) is a novel meta-heuristic swarm intelligence optimization algorithm that has been proposed by Rao [10]. It mimics the teaching and learning behavior between the teacher and students in the class, which does not demand any algorithm-specific parameters [11]. When using TLBO to solve optimization problems, the influence of improper algorithm-specific parameters on the results can be avoided. Thus, TLBO has been successfully utilized to solve various optimization problems, such as discrete routing problems [12], hybrid flow shop scheduling problems [13], clustering problems [14], constrained engineering design problems [15], and so on. However, not all optimization problems can be effectively solved by the TLBO algorithm [16]. It is necessary to improve the performance of TLBO for its application in other fields, especially when it is used for solving large scale QoS-MCSC problems.

In this study, a hybrid teaching-learning-based optimization algorithm (HTLBO) is proposed to solve QoS-MCSC problems, combining the uniform mutation of GA,

the adaptive FPA, and the TLBO algorithm. The uniform mutation of GA is utilized to preserve the diversification of the population. Then, to avoid TLBO premature convergence, the population is divided into two parts according to the fitness of each individual. The top parts are updated by the TLBO algorithm which has a high local search ability and a fast convergence speed. And the others are updated by the adaptive FPA which has a high global search ability. The adaptive parameters are utilized to enhance the global search ability of standard FPA. The experimental results show that the proposed approach outperforms the algorithms like GA, PSO, TLBO, EWOA (Eagle strategy with Whale Optimization Algorithm), and EFPA (Extended FPA).

The remainder of this study is structured as follows. The related works on QoS-MCSC problems are discussed in Sect. 2. Section 3 describes the QoS-MCSC model. An HTLBO algorithm using uniform mutation, adaptive FPA, and TLBO is proposed in Sect. 4. Section 5 analysis the applicability of HTLBO in different kinds of scales QoS-MCSC problems. Section 6 concludes the whole study and discusses future works.

2 Related work

The existing studies about service composition mainly focused on four main areas: business flow-based [17], graph-based [18], agent-based [19], and QoS-based service composition methods [20]. QoS-based service composition methods are the research hotspots of all the service composition methods mentioned above. Existing approaches for solving QoS-aware service composition problems can be classified into three categories: scalarization-based, Pareto dominance-based, and other approaches.

2.1 Scalarization-based approaches

In these approaches, a global evaluation function is utilized to aggregate several different QoS parameters into a real number. This aggregation procedure is named as scalarization. The simple additive weighting (SAW) technique is widely utilized to transform the QoS vector of a composite service into a global QoS value. Based on SAW, Akbaripour et al. [21] proposed an approach based on mixed integer programming (MIP) for solving manufacturing cloud service selection problems. Mohamed Essaid et al. [22] introduced a novel approach employing the k-means clustering method, lexicographic optimization method, and a search tree to obtain near-optimal solutions. Compared with non-heuristic algorithms and heuristic algorithms, meta-heuristic algorithms can obtain near-optimal solutions more effectively and efficiently [23]. Huang et al. [24] adopted a novel chaos control optimal algorithm to obtain near-optimal solutions for QoS-MCSC problems. The experiment results showed that the proposed approach could find higher quality solutions than the algorithms like GA and chaotic genetic algorithm (CGA).

Considering the correlations among cloud services, Jin et al. [5] introduced a service correlation description model and a correlation mapping model. The description model was used to describe the attributes the associated services should have. And the mapping model was applied to automatically obtain the correlation QoS values among

services. Zhang et al. [9] proposed an extended FPA (EFPA) to solve correlation-aware QoS-MCSC problems. In this approach, the adaptive switch probability and a scaling factor of FPA were utilized to improve the performance of standard TLBO. And the crossover strategy and mutation strategy of GA were applied to avoid the algorithm falling into the local optimum.

To balance the exploration and exploitation abilities of algorithms, Fateh et al. [25] introduced a hybrid approach combining GA and fruit fly optimization algorithm for QoS-MCSC problems. In this approach, the selection operator, crossover operator, and mutation operator of GA were utilized to do global exploration. The fruit fly optimization algorithm was applied to do local exploitation, which was used to obtain better solutions after the GA operators. Gavvala et al. [26] presented a novel approach using eagle strategy and WAO to find high quality results for QoS-MCSC problems. In this approach, a simple randomization method was employed to do the global search, which improved the exploration ability in the process of algorithm optimization.

2.2 Pareto dominance-based approaches

In this category of approaches, the concept of Pareto dominance was utilized to obtain a set of non-dominated solutions establishing different trade-offs among all the objectives for QoS-MCSC problems [27]. To solve this kind of multi-objective problem, Li et al. [28] utilized Strength Pareto Evolutionary Algorithm 2 (SPEA2) to find the Pareto optimal solutions for QoS-aware service composition problems. Yao et al. [29] presented Non-dominated Sorting Genetic Algorithm II (NSGA-II) to obtain the optimal solution set. It has been proved that SPEA2 and NSGA-II have a good performance when two or three objectives are considered to be optimized [27]. Chattopadhyay et al. [30] presented an approach based on NSGA to find the Pareto optimal solutions for QoS-aware automatic web service composition problems. Tao et al. [7] discussed an approach based on PSO and non-dominated sorting technique to find the Pareto front for the correlation-aware manufacturing resource service composition problems. However, with the increasing number of objectives, the Pareto dominance-based approaches may lose the efficiency to find the optimal solutions.

Zhang et al. [31] presented a Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) to separate the proximate Pareto front into a series of scalar optimization sub-problems. The computational complexity of MOEA/D is lower than NSGA-II. And the experimental results show that the solution quality obtained by MOEA/D is outperformed NSGA-II. Based on MOEA/D, Zhou et al. [32] proposed a multi-objective differential ABC algorithm for service composition and optimal selection, which used the concept of Pareto dominance to guide the searching of a bee swarm. Considering the energy consumption and the QoS, Yang et al. [33] introduced an enhanced multi-objective grey wolf optimizer (EMOGWO) for QoS-MCSC problems. In this approach, the backward learning method is utilized to enhance the diversity of the initial population. And an enhanced search strategy is applied to increase the search areas and improve the exploration of the leaders.

2.3 Other approaches

There are some approaches based on Artificial Intelligence (AI) techniques or hybrid approaches. Liang et al. [34] proposed a Deep Reinforcement Learning (DRL) algorithm for QoS-MCSC problems. The proposed approach is an AI technique. The experimental results revealed that the proposed approach is outperformed Deep Q-Network DQN and Q-Learning algorithm. Liu et al. [35] discussed an approach based on Deep Reinforcement Learning (DRL) for adaptive service composition problems. In this method, the recurrent neural network (RNN) is applied to predict the objective function. Reinforcement learning and deep learning are combined to obtain service composition solutions. Zhou et al. [36] presented a hybrid approach using both Scalarization-based and Pareto dominance-based approaches. There are two objectives considered in this approach: QoS and energy consumption. The QoS aggregated value is calculated by scalarization-based approaches. The final Pareto optimal solutions are found by a multi-objective hybrid ABC algorithm.

In sum, with the increasing number of objectives, the Pareto dominance-based approaches will lose the efficiency to obtain the optimal solution set for QoS-MCSC problems. Meta-heuristic algorithms used in Scalarization-based approaches could find near-optimal solutions more effectively and efficiently. However, the above-mentioned algorithms still have some drawbacks, such as slow convergence rate and global search ability in the later process of evolution. Thus, a novel HTLBO algorithm is presented to solve QoS-MCSC problems in this study.

3 QoS-aware manufacturing cloud service composition model

The framework of QoS-MCSC model is shown in Fig. 1. A complex CMfg Task can be performed on a CMfg platform by the following steps:

(1) Task decomposition: according to manufacturing features and process requirements of the complex task, it is decomposed into multiple subtasks as Task $T = \{ST_1, ST_2, \ldots, ST_i, \ldots, ST_I\}$, where I denotes the number of subtasks.

(2) Service discovery: for each subtask, all MCSs that can meet the functional requirements are collected to form its candidate MCS set (CMCSS), as $CMCSS_i = \{MCS_{i,1}, MCS_{i,2}, \ldots, MCS_{i,j}, \ldots, MCS_{i,Ki}\}$, where K_i is the total number of available MCSs in the ith CMCSS.

(3) Service optimal selection: a single candidate MCS or a combination of multiple candidates MCS is selected for each subtask from its CMCSS to generate a composite manufacturing service. There are $\prod_{i=1}^{I} K_i$ possible composite manufacturing services for Task T. So the QoS-MCSC problem is to select the optimal composite service from $\prod_{i=1}^{I} K_i$ possible composite manufacturing services to accomplish Task T.

The four widely adopted QoS attributes are selected to establish the QoS evaluation model for MCSs in this study; they are time (q_{time}), cost (q_{cost}), reliability (q_{rel}) and availability (q_{avail}). These QoS attributes for each MCS is represented as $Q(MCS_{i,j}) = \{q_{time}(MCS_{i,j}), q_{cost}(MCS_{i,j}), q_{rel}(MCS_{i,j}), q_{avail}(MCS_{i,j})\}$, where $q_{cost}(MCS_{i,j})$ is the cost attributes of the jth available MCSs for the ith subtask in the ith CMCSS. The aggregation QoS value of a CMCS is done based on the struc-

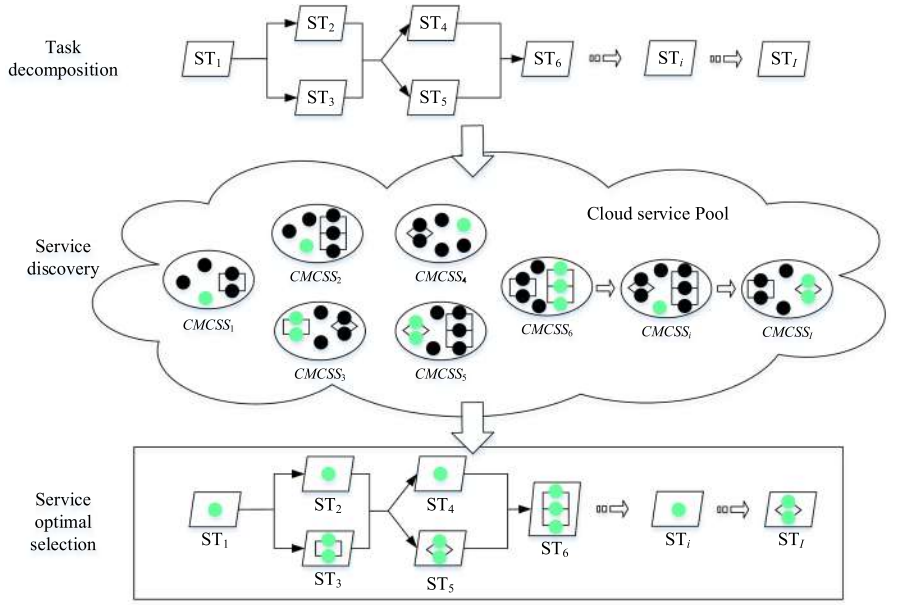


Fig. 1 The framework of QoS- MCSC model

ture of a workflow. Four basic structures, i.e., sequential structure, parallel structure, selective structure, and loop structure, are taken into consideration. Before calculating the QoS value of a CMCS, other structures should be transformed into the sequential structure. Then the QoS value of a CMCS can be aggregated according to the aggregation function of the sequential structure. The QoS aggregation functions for different composition structures are given in Table 1 [25].

As the range of the value of QoS attributes and the unit of measurement of QoS attributes are quite different, it requires transforming the values of QoS attributes into an evaluable attribute value. The QoS attributes can be classified into two categories: positive attributes (Q^+) and negative attributes (Q^-). The Q^+ means that a high value of a QoS attribute is desirable, i.e., reliability and availability. While the Q^- means that a low value is desirable, i.e., time and cost. The SAW technique is utilized to transform the aggregated QoS vector of CMCS into a global QoS value. The global QoS value of CMCS_k can be defined as follows:

$$Q(CMCS_k) = \sum_{q_t \in Q^-} \frac{q_{t,\max} - q_t(CMCS_k)}{q_{t,\max} - q_{t,\min}} * w_t + \sum_{q_t \in Q^+} \frac{q_t(CMCS_k) - q_{t,\min}}{q_{t,\max} - q_{t,\min}} * w_t \tag{1}$$

where w_t indicates the weight of the tth QoS attribute, $q_{t\ ,\ max}$ and $q_{t\ ,\ min}$ denote the maximum and minimum aggregate values of the tth QoS parameter, respectively.

Table 1 QoS aggregation functions

Structure	Time	Cost	Reliability	Availability
Sequential	$\sum_{i=1}^{n} q_{time}(MCS_i)$	$\sum_{i=1}^{n} q_{cost}(MCS_i)$	$\prod_{i=1}^{n} q_{rel}(MCS_i)$	$\prod_{i=1}^{n} q_{avail}(MCS_i)$
Parallel	$Max\{q_{time}(MCS_i)\}$	$\sum_{i=1}^{n} q_{cost}(MCS_i)$	$\prod_{i=1}^{n} q_{rel}(MCS_i)$	$\prod_{i=1}^{n} q_{avail}(MCS_i)$
Selective	$\sum_{i=1}^{n}(q_{time}(MCS_i) * \gamma_i)$	$\sum_{i=1}^{n}(q_{cost}(MCS_i) * \gamma_i)$	$\sum_{i=1}^{n}(q_{rel}(MCS_i) * \gamma_i)$	$\sum_{i=1}^{n}(q_{avail}(MCS_i) * \gamma_i)$
Loop	$k_l * \sum_{i=1}^{n} q_{time}(MCS_i)$	$k_l * \sum_{i=1}^{n} q_{cost}(MCS_i)$	$\prod_{i=1}^{n} q_{rel}(MCS_i)$	$\prod_{i=1}^{n} q_{avail}(MCS_i)$

Note:$\gamma_i$represents the probability that MCS_i will be selected and $\sum_{i=1}^{n} \gamma_i = 1$. k_l is the cycle time

4 The proposed HTLBO algorithm for QoS-MCSC problems

4.1 Encoding

In this paper, an I–dimension vector $X_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,j}, \ldots, x_{i,I}]$ represents a solution for the QoS-MCSC problems, where $x_{i,j}$ is represents the index of the MCS in the jth CMCSS. The value of $x_{i,j}$ is bounded to be in the interval $[lb_j, ub_j]$, where lb_j is 1 and ub_j is the number of MCS in the jth CMCSS. The dimension of vector I is equal to the number of subtasks.

4.2 The uniform mutation

The mutation operator in GA is utilized to prevent the population to become too similar. There are various mutation strategies used in GA, such as exchange mutation, inversion mutation, scramble mutation, uniform mutation, and so on. The uniform mutation is suitable for real and integer representation. Thus, in this paper, the uniform mutation strategy of GA is utilized to preserve the diversity of solutions. The uniform mutation strategy can be formulated as follows:

$$x_{i,j}(t+1) = \begin{cases} lb_j + r_{i,j} \cdot (ub_j - lb_j) & \text{if } k_{i,j} \leq p_m \\ x_{i,j}(t) & \text{if } k_{i,j} > p_m \end{cases} \tag{2}$$

where t denotes the current number of iterations, lb_j is 1 and ub_j is the number of MCS in the jth CMCSS, $r_{i,j}$ is a number randomly generated in the range [0, 1], p_m is the probability of mutation. p_m can be calculated as follows:

$$p_m = 0.01 + 0.1(1 - \frac{t}{MaxIter}) \tag{3}$$

where t and $MaxIter$ denote the current and maximum iteration number, respectively.

4.3 The TLBO

TLBO mimics the teaching and learning behavior between the teacher and students in the class. The best individual in the class population is considered as the teacher. Besides learning knowledge from the teacher, students can also learn from each other. There are two stages for individuals to update their location: the teaching phase and the learning phase. The mathematical model of each operator is discussed in detail in the following subsections.

4.3.1 Teaching phase

In this phase, the individuals learn from the teacher. The teaching phase can be considered as a global search phase. The best individual in the class is selected as the teacher X_T. For QoS-MCSC problems, the individual with the highest global QoS value in

the population is regarded as X_T. The teacher X_T will improve students' performance in various subjects to a certain extent through teaching and try to improve the class's average score close to his or her level. The students update their position according to the difference between the teacher and the average value of the class. This process is represented by the following formulas:

$$X_i(t+1) = X_i(t) + Difffference_t \tag{4}$$

$$Difffference_t = r_t(X_T - T_f * Mean) \tag{5}$$

where $X_i(t+1)$ and $X_i(t)$ represent the new and old position of the ith student, respectively, t denotes the current number of iterations, Mean = $(\sum_{i=1}^{P} X_i(t))/N_P$ is the mean state of the class, N_P denotes the population size, r_t is the learning factor which is randomly generated in the interval [0, 1], T_f is the teaching factor. The formula of T_f can be expressed as:

$$T_f = round[1 + rand(0, 1)] \tag{6}$$

For each student, the new position is calculated by Eq. (2). If $X_{i,new}$ is better than the previous value, the new position is accepted. Each student represents a feasible solution for the QoS-MCSC problem.

4.3.2 Learning phase

The students improve their performance by analyzing the differences between themselves and other students. The learning phase can be considered as a local search phase. The learning method is similar to the mutation strategy in DE. But in TLBO, different students have different learning factors. For the ith students X_i in the class, the updating mechanism can be formulated as follows:

$$X_i(t+1) = \begin{cases} X_i(t) + r_i * (X_i(t) - X_j(t)), & \text{if } f(X_i(t)) < f(X_j(t)) \\ X_i(t) + r_i * (X_j(t) - X_i(t)), & \text{if } f(X_j(t)) < f(X_i(t)) \end{cases} \tag{7}$$

where $X_j(t)$ is a student randomly selected from the current population, $f(X_i(t))$ and $f(X_j(t))$ denote the fitness value of $X_i(t)$ and $X_j(t)$, respectively, r_i is the learning factor in this phase which is randomly generated in the interval [0, 1].

Similar to the teaching phase, the new position is accepted if its fitness value is better than the previous value.

4.4 The adaptive FPA

FPA is a bio-inspired meta-heuristic optimization algorithm presented by Yang [37]. It mimics the process of flowers pollination. The pollens are transferred from a flower to other flowers by various pollinators such as bees, wind, birds, etc. FPA includes two pollination methods for pollens to update their location: biotic pollination and abiotic pollination. For biotic pollination, pollens are transferred by pollinators over long

distances. Nevertheless, for abiotic pollination, pollens are transferred to other flowers nearby. The standard PFA has the possibility of converging toward local optima. An adaptive FPA is proposed to overcome this shortcoming and improve the performance of the standard FPA. The mathematical representation of adaptive FPA is discussed in the following subsections.

4.4.1 Biotic pollination

In biotic pollination, the current best individual X_{best} needs to be defined first. For QoS-MCSC problems, the pollen with the highest global QoS value in the population is regarded as X_{best}. After the current best pollen is confirmed, the other pollens will update their positions toward the current best pollen. This mechanism can be represented by the following formulas:

$$X_i(t+1) = X_i(t) + \gamma L(\lambda)(X_{best}(t) - X_i(t)) \tag{8}$$

where $X_i(t+1)$ and $X_i(t)$ denote the new and old position of the ith pollen, γ is a scaling factor for controlling the step size during biotic pollination, $L(\lambda)$ is an I–dimension vector, which obeys the Lévy distribution. The Lésvy distribution can be defined as:

$$L(\lambda) \sim \frac{\varphi \times \mu}{|v|^{1/\lambda}} \tag{9}$$

where λ is a real number in the interval [1, 2] and taken to be 1.5, μ and v follow the standard normal distributions. φ is defined as follows:

$$\varphi = \left[\frac{\Gamma(1+\lambda) \times \sin(\pi\lambda/2)}{\Gamma((1+\lambda)/2) \times \lambda \times 2^{(\lambda-1)/2}} \right]^{1/\lambda} \tag{10}$$

where Γ denotes the standard Gamma function.

4.4.2 Abiotic pollination

In abiotic pollination, each pollen updates its position according to the position of the other two pollens instead of the positions of the current best pollen. This process can be represented by the following formulas:

$$X_i(t+1) = X_i(t) + r(X_p(t) - X_q(t)) \tag{11}$$

where $X_p(t)$ and $X_q(t)$ denote two pollens randomly selected from the current population, r is a random number that obeys the standard normal distribution in the interval [0, 1].

4.4.3 Adaptive parameters

In the standard PFA, each pollen has a probability of p to select the switch between biotic pollination and abiotic pollination to renew the position. The switch probability p is usually set to a fixed value of 0.8. However, a fixed switch probability cannot solve different scales of QoS-MCSC problems well. Thus, an adaptive switch probability is proposed to improve the performance of the PFA. The value of p can be calculated as follows:

$$p = e^{-\frac{MaxIter-t}{MaxIter}} \tag{12}$$

where t and $MaxIter$ denote the current and maximum iteration number, respectively. It can be found that more pollens perform biotic pollination in the later process of evolution.

In addition, the scaling factor γ is set to a fixed value of 1 in the standard PFA. To avoid the algorithm falling into local optimal solutions, an adaptive scaling factor is proposed to change the step length in evolution. The scaling factor γ can be calculated as follows:

$$\gamma = \frac{f(X_{best}(t)) - f(X_{worst}(t))}{f(X_{best}(t))} \tag{13}$$

where $f(X_{best}(t))$ and $f(X_{worst}(t))$ denote the best and worst fitness value at the tth iteration, respectively.

4.5 The structure and process of the proposed hybrid approach

The framework of the proposed HTLBO algorithm is shown in Fig. 2, where the elaborated explanations of uniform mutation, TLBO, and adaptive FPA are presented in the previous subsections. The pseudo-code of HTLBO is shown in Algorithm 1. First, it initializes the related parameters and population $\boldsymbol{P}$. Second, it updates individuals in population $\boldsymbol{P}$ using the uniform mutation strategy proposed in Section 4.2. After that, according to the fitness of each individual, the population $\boldsymbol{P}$ is divided into two parts: the top part $\boldsymbol{P}_1$ and the others $\boldsymbol{P}_2$. The size of the subpopulation $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are the same which are 50% of P. For individuals in subpopulation $\boldsymbol{P}_1$, they are updated by the TLBO algorithm proposed in Section 4.3. And for individuals in subpopulation $\boldsymbol{P}_2$, they are updated by the adaptive FPA algorithm proposed in Section 4.4. Later, subpopulation $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are combined to form $\boldsymbol{P}$. On last, it repeats the uniform mutation strategy, the TLBO algorithm, and the adaptive FPA algorithm until the end conditions are satisfied.

The termination conditions for the intelligent optimization algorithm can be defined as follows:

(1) Set a fixed large number as the maximum number of evolutions.
(2) Set a specific CPU time as the maximum running time of the algorithm.

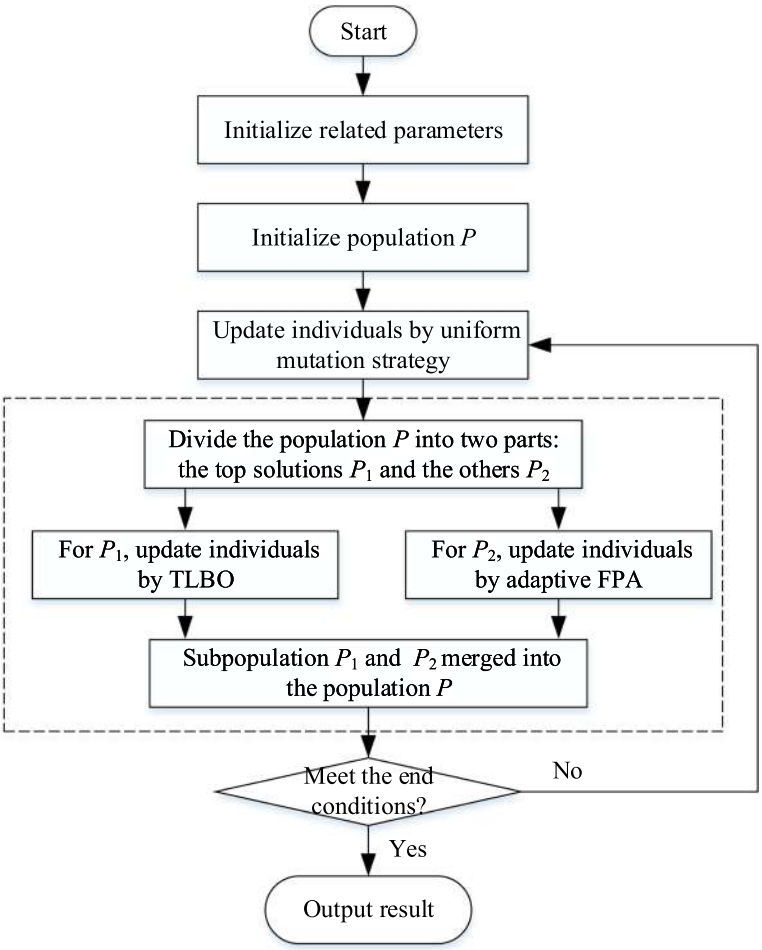


Fig. 2 The framework of the proposed HTLBO

(3) Set a fixed number of generation: max_gen_rem. If the best solution remains unchanged over max_gen_rem generations, the algorithm stops and outputs the result.
(4) The algorithm converges to a specific value of the fitness function.

Among the four termination conditions for the intelligent optimization algorithm, the first criterion is the most widely applied. In this study, the first criterion is utilized as the termination conditions of the proposed HTLBO. According to [25, 26], the maximum number of iterations is set to1000.

Algorithm 1: QoS-MCSC based on the proposed HTLBO

1: Generate initial population $\boldsymbol{P}$ randomly
2: Determine the fitness of all individuals by Eq. (1)
3: Store $\boldsymbol{X}^*$ as the best solution
4: **while** ($t <$ max number of iterations)
5: ***// The uniform mutation*(Presented in Section 4.2)**
6: **for** each individual $\boldsymbol{X}_i$ in population $\boldsymbol{P}$
7: Calculate the probability of mutation p_m by Eq. (3)
8: Generate a new individual $\boldsymbol{X}_{i,new}$ by Eq. (2)
9: Calculate the fitness of the new individual $\boldsymbol{X}_{i,new}$
10: Replace $\boldsymbol{X}_i$ with $\boldsymbol{X}_{i,new}$ if $\boldsymbol{X}_{i,new}$ is better than $\boldsymbol{X}_i$
11: **end for**
12: Rank the population $\boldsymbol{P}$ according to the fitness of each individual
13: Divide the $\boldsymbol{P}$ into two parts: the top part $\boldsymbol{P}_1$ and the others $\boldsymbol{P}_2$
14: ***// TLBO for subpopulation* $\boldsymbol{P}_1$**
15: Identify the teacher $\mathbf{X}_T$ in $\boldsymbol{P}_1$
16: **for** each individual $\boldsymbol{X}_i$ in $\boldsymbol{P}_1$
17: ***// Teaching phase(Presented in Section 4.3.1)***
18: Generate a new individual $\boldsymbol{X}_{i,new}$ by Eq. (4)
19: Calculate the fitness of the new individual $\boldsymbol{X}_{i,new}$
20: Replace $\boldsymbol{X}_i$ with $\boldsymbol{X}_{i,new}$ if $\boldsymbol{X}_{i,new}$ is better than $\boldsymbol{X}_i$
21: ***// Learning phase(Presented in Section 4.3.2)***
22: Generate a new individual $\boldsymbol{X}_{i,new}$ by Eq. (7)
23: Calculate the fitness of the new individual $\boldsymbol{X}_{i,new}$
24: Replace $\boldsymbol{X}_i$ with $\boldsymbol{X}_{i,new}$ if $\boldsymbol{X}_{i,new}$ is better than $\boldsymbol{X}_i$
25: **end for**
26: ***// Adaptive FPA for subpopulation* $\boldsymbol{P}_2$ *(Presented in Section 4.4)***
27: Identify the best solution in $\boldsymbol{P}_2$
28: Update p and γ by Eq. (12) and Eq. (13), respectively
29: **for** each individual $\boldsymbol{X}_i$ in $\boldsymbol{P}_2$
30: **if** $rand < p$
31: Determine new individual in biotic pollination by Eq. (8)
32: **else**
33: Determine new individual in abiotic pollination by Eq. (11)
34: **end if**
35: if new individuals are better, update them in $\boldsymbol{P}_2$
36: **end for**
37: subpopulation $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are combined to form the population $\boldsymbol{P}$
34: Update $\boldsymbol{X}^*$ if there is a better solution
35: t=t+1
36: **end while**
37: **return** X^*

5 Experiments

To verify the effectiveness of the proposed approach in solving the QoS-MCSC problems, the proposed HTLBO algorithm is compared with some other intelligence optimization algorithms that have recently been applied in QoS-MCSC, such as GA [5], PSO [7], TLBO [11], EWOA [26], and EFPA [9]. The parameter settings used in these algorithms are stated in Table 2. It is noteworthy that TLBO does not require any algorithm-specific controlling parameters. The CMCSs considered in the experiments have a sequential structure. Since other structures can be transformed into the

Table 2 Parameter settings

Algorithm	Parameter	Value
	Crossover probability	0.7
GA	Mutation probability	0.1
	Generation gap	0.9
	Acceleration factors c_1	Linearly decrease from 2.5 to 0.5
PSO	Acceleration factors c_2	Linearly increase from 0.5 to 2.5
	Inertia weight	Linearly decrease from 0.9 to 0.4
EWOA	Switch control parameter p_e	0.2
EFPA	Switch probability p	Exponential increase from e^{-1} to 1
	Scaling factor γ	Related to the fitness values
	Crossover probability	0.2
	Mutation probability	0.1

GA Genetic Algorithm [5], *PSO* Particle Swarm Optimization [7], *TLBO* Teaching-Learning-Based Optimization [11], *EWOA* Eagle strategy with Whale Optimization Algorithm [26], *EFPA* Extended Flower Pollination Algorithm [9]

sequential structure, the validity and relevance of the experimental results will not be affected by this choice [22].

We assume that the QoS values of each MCS were generated randomly. The values of q_{time}, q_{cost}, q_{rel}, and q_{avail} of each MCS were randomly generated in the range [0.7, 0.98]. The customer's preference for these QoS attributes were chosen as w_{time}=0.35, w_{cost}=0.3, w_{rel}=0.2, and w_{avail}=0.15. We also suppose that 40% of MCSs had potential quality correlations among them.

Twenty scales of QoS-MCSC problems are considered to verify the effectiveness of the proposed HTLBO algorithm. A complex CMfg Task includes 10, 20, 30, 40, and 50 subtasks, while the amount of candidate MCSs ranges as 50, 100, 150, and 200. For example, T-40-150 denotes that the number of subtasks is 40 and the number of candidate MCSs for each subtask is 150.

In these experiments, the population size of all the algorithms is 50 and the maximum number of iterations is 1000. Each algorithm independently runs 30 times for each problem, and the experimental results are based on the average performance of the 30 runs. The experiments are implemented on a PC with Intel i7-6500U 2.50GHz, 8GB RAM, Windows 10 (64 bit), and MATLAB R2011a.

Table 3 shows the average, standard deviation, and best solutions obtained by GA, PSO, TLBO, EWOA, EFPA, and HTLBO. We observed that the average QoS fitness values for these problems obtained by HTLBO are better than other compared algorithms except for the problem of T-10-50. For T-10-50, the average solution obtained by EFPA is the same as HTLBO. From Table 3, the best solutions of thirty runs obtained by HTLBO are better than other compared algorithms except for the problem of T-30-150, T-40-100, T-50-100, and T-50-150. For these four scales, the best solutions obtained by EWOA are better than HTLBO. But the average solutions and standard deviation obtained by HTLBO are better than EWOA. Thus, the proposed HTLBO algorithm has enough competitive with other compared algorithms.

Table 3 Results of GA, PSO, TLBO, EWOA, EFPA, and HTLBO on 20 test problems

Problems	Index	GA	PSO	TLBO	EWOA	EFPA	HTLBO
	Mean	0.666227	0.662499	0.671974	0.671517	**0.692080**	**0.692080**
T-10-50	Std	0.020148	0.018964	0.001738	0.000000	0.000000	0.000000
	Best	0.671517	0.678367	0.678367	0.671517	**0.692080**	**0.692080**
	Mean	0.597272	0.609502	0.609593	0.692363	0.685640	**0.703274**
T-10-100	Std	0.014219	0.022385	0.022914	0.014786	0.000700	0.012739
	Best	0.624616	0.668677	0.664253	0.725269	0.689348	**0.735145**
	Mean	0.599401	0.670220	0.669744	0.709673	0.698156	**0.711171**
T-10-150	Std	0.037666	0.020194	0.007894	0.011878	0.003783	0.011156
	Best	0.705326	0.709683	0.691499	0.725064	0.700372	**0.725064**
	Mean	0.591283	0.617351	0.619632	0.702823	0.715979	**0.751139**
T-10-200	Std	0.017836	0.014776	0.003756	0.025928	0.002463	0.003299
	Best	0.623465	0.646494	0.627889	0.753216	0.720240	**0.754433**
	Mean	0.592141	0.559537	0.587863	0.592141	0.609828	**0.610521**
T-20-50	Std	0.000000	0.020319	0.002401	0.000000	0.002639	0.000000
	Best	0.592141	0.586561	0.592141	0.592141	**0.610521**	**0.610521**
	Mean	0.514483	0.534129	0.548299	0.597655	0.601282	**0.614009**
T-20-100	Std	0.026076	0.009530	0.003320	0.011378	0.000995	0.001518
	Best	0.557301	0.551797	0.556506	0.611138	0.602331	**0.621134**
	Mean	0.498876	0.564324	0.597600	0.597623	0.620584	**0.622346**
T-20-150	Std	0.026047	0.021525	0.000216	0.000221	0.006078	0.000850
	Best	0.593685	0.608547	0.598742	0.598790	0.622181	**0.626839**
	Mean	0.498702	0.531682	0.554150	0.610018	0.623314	**0.641583**
T-20-200	Std	0.010930	0.012280	0.001773	0.020643	0.003145	0.003564
	Best	0.523751	0.566272	0.559752	0.637713	0.627159	**0.648169**
	Mean	0.543052	0.514584	0.541867	0.558969	0.557486	**0.559316**
T-30-50	Std	0.016475	0.010290	0.001832	0.003861	0.000000	0.004472
	Best	0.547943	0.533770	0.546105	0.567369	0.557486	**0.580210**
	Mean	0.488898	0.519451	0.541953	0.563360	0.570770	**0.579466**
T-30-100	Std	0.037484	0.010330	0.002672	0.006697	0.000836	0.001339
	Best	0.548647	0.541636	0.547507	0.575346	0.572936	**0.582433**
	Mean	0.461195	0.511823	0.533214	0.560652	0.569051	**0.577685**
T-30-150	Std	0.010617	0.011577	0.004737	0.010350	0.002212	0.002127
	Best	0.482307	0.534795	0.545046	**0.589143**	0.570125	0.581750
	Mean	0.466688	0.520989	0.543838	0.566288	0.566207	**0.573281**
T-30-200	Std	0.011917	0.010408	0.004696	0.004077	0.002728	0.002991
	Best	0.500318	0.542458	0.552911	0.573323	0.570244	**0.579766**
	Mean	0.532002	0.491929	0.529415	0.546483	0.550603	**0.561261**
T-40-50	Std	0.000984	0.013187	0.001574	0.003262	0.000959	0.003960
	Best	0.533342	0.523986	0.531942	0.551626	0.551681	**0.566239**
	Mean	0.491730	0.496357	0.531215	0.553236	0.542261	**0.557904**

Table 3 continued

Problems	Index	GA	PSO	TLBO	EWOA	EFPA	HTLBO
T-40-100	Std	0.047271	0.014020	0.002568	0.008413	0.001378	0.004220
	Best	0.537378	0.516861	0.536744	**0.570352**	0.545351	0.568424
	Mean	0.448523	0.493966	0.539187	0.552538	0.549421	**0.560773**
T-40-150	Std	0.009589	0.013455	0.002244	0.003457	0.001989	0.003966
	Best	0.471557	0.519079	0.542763	0.562592	0.555576	**0.568463**
	Mean	0.450524	0.497732	0.542426	0.555923	0.547725	**0.558929**
T-40-200	Std	0.011563	0.015741	0.002758	0.002673	0.001510	0.002178
	Best	0.473049	0.531298	0.547884	0.561096	0.550960	**0.564473**
	Mean	0.516393	0.489800	0.511323	0.551626	0.542516	**0.557402**
T-50-50	Std	0.000859	0.009188	0.002322	0.005320	0.000974	0.003598
	Best	0.517613	0.505669	0.515351	0.562627	0.544336	**0.564358**
	Mean	0.452379	0.481838	0.519916	0.551797	0.542351	**0.556407**
T-50-100	Std	0.036335	0.012383	0.002140	0.004535	0.000619	0.002474
	Best	0.523401	0.503337	0.524224	**0.561897**	0.543466	0.560433
	Mean	0.430470	0.471352	0.520178	0.539421	0.547497	**0.552132**
T-50-150	Std	0.010304	0.010494	0.001807	0.007136	0.000539	0.001341
	Best	0.452598	0.492971	0.524136	**0.558496**	0.548445	0.556487
	Mean	0.429626	0.470927	0.530118	0.546643	0.541730	**0.551460**
T-50-200	Std	0.008989	0.008750	0.002552	0.003113	0.001246	0.003726
	Best	0.449952	0.484923	0.535759	0.553655	0.544195	**0.559434**

Bold values in the row of 'Mean' represents the maximum value of the average solution obtained in the process of running 30 times of the six intelligent algorithms. And the bold values in the row of 'Best' represents the maximum value of the optimal solution obtained in the process of running 30 times of the six intelligent algorithms

Figure 3, 4, 5 shows the convergence curves for the twenty scales of QoS-MCSC problems. We observed that the standard TLBO converges faster than GA, PSO, EWOA, and EFPA. But this fast convergence rate might not help to maintain the diversity of the population, which might cause the algorithm to converge to a local optimum. Although the convergence speed of the proposed HTLBO algorithm is worse than TLBO and EFPA, it performs better global search ability than other compared algorithms. Even in the later process of evolution, it has an excellent exploration ability. We also observed that the proposed algorithm, the same as compared algorithms, shows similar trends on the problems of different scales. It means that the convergence effect of the proposed algorithm may not be affected by the scales of the problem.

Table 4 shows the time consumption of different algorithms for 10, 20, 30, 40, and 50 subtasks with respect to 150 candidate MCSs. From Table 4, HTLBO takes more time than other compared algorithms due to more efforts are put into the global search. Although the proposed algorithm takes more execution time, it can find better solutions for QoS-MCSC problems. Moreover, the execution time spent by the algorithm is insignificant compared to the hundreds of hours of completing the manufacturing task.

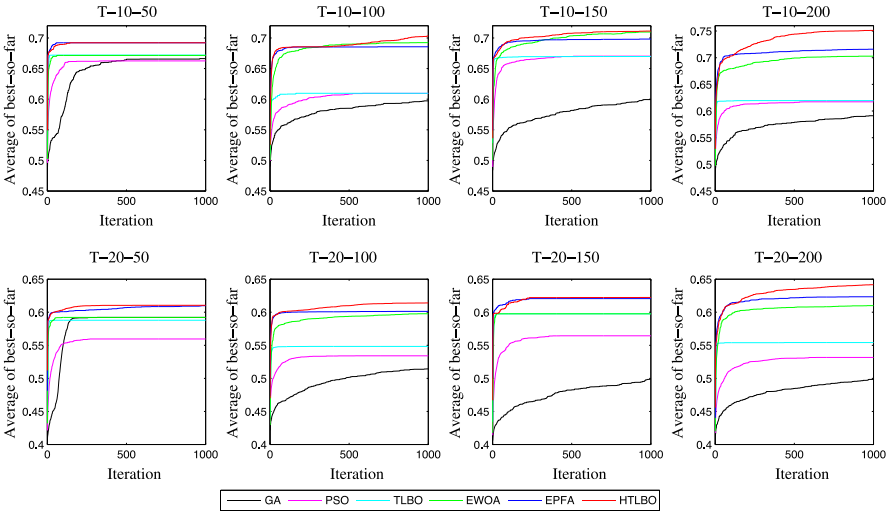


Fig. 3 Convergence curves for the problems include 10 and 20 subtasks

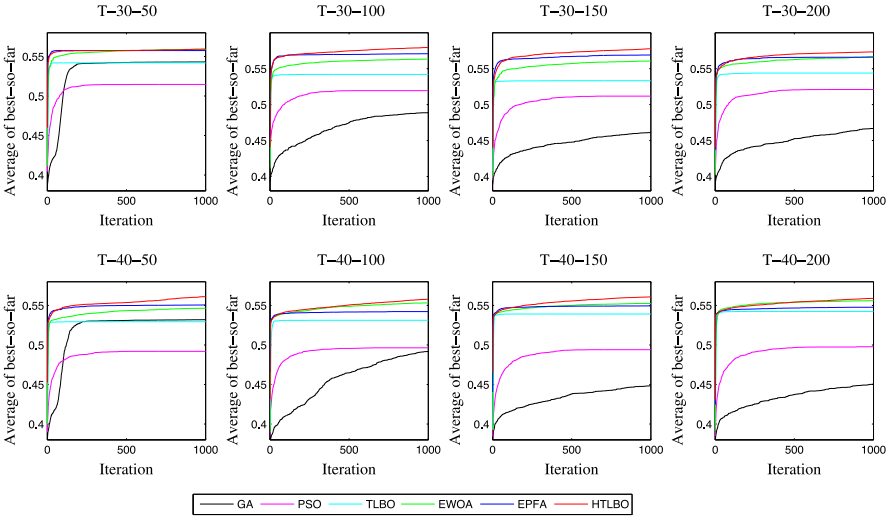


Fig. 4 Convergence curves for the problems include 30 and 40 subtasks

It is valuable to spend more computing time to find better solutions for manufacturing tasks.

In summary, the results of this section reveal that HTLBO is highly competitive for QoS-MCSC problems. The strategies of uniform mutation and adaptive FPA are effective for improving the exploration and exploitation abilities of TLBO. There are two reasons for HTLBO outperforming other compared algorithms. The first is that the uniform mutation is utilized to preserve the diversification of the population. Individuals always have a probability of p_m to do the global search in this strategy,

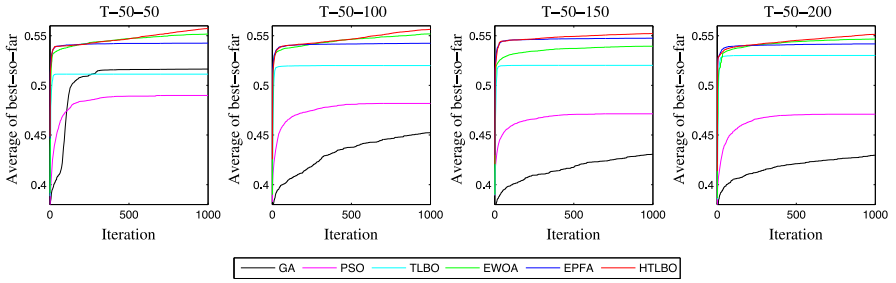


Fig. 5 Convergence curves for the problems include 50 subtasks

Table 4 Comparison of time consumptions of the algorithms (in seconds)

Problem	GA	PSO	TLBO	EWOA	EFPA	HTLBO
T-10-150	1.82	2.18	6.05	4.65	7.81	7.87
T-20-150	2.88	3.56	9.30	8.07	10.70	12.37
T-30-150	3.06	3.94	10.06	9.14	11.55	14.55
T-40-150	3.88	4.98	12.69	11.72	14.11	18.64
T-50-150	4.94	6.26	15.68	15.32	17.18	22.71

which improves the global search ability of HTLBO. The second is that the population is divided into two subpopulations according to the fitness of each individual. The top parts are updated by the standard TLBO which has a high local search ability and a fast convergence speed. And the others are updated by the adaptive FPA which has a high global search ability. The adaptive parameters are utilized to enhance the global search ability of standard PFA. The two parts are evolved with different mechanisms and then exchange information with each other. Therefore, the proposed HTLBO integrates the advantages of the uniform mutation, the adaptive FPA, and the TLBO algorithm, which can preserve the diversity of the population and find the high quality results.

6 Conclusions

Manufacturing cloud service composition is an effective method to improve the utilization of manufacturing resources and realize the value-added of manufacturing resources. In CMfg platform, there are many available MCSs with similar functional attributes but different non-functional attributes. Users often find it very difficult to select an optimal composite service to meet their complex requirements. To address QoS-MCSC problems, a hybrid HTLBO algorithm, combining the uniform mutation of GA, the adaptive FPA, and the TLBO algorithm, is proposed. The main contributions of this work are summarized as follows:

(1) The uniform mutation strategy of GA is utilized to preserve the diversification of the population. Individuals always have a probability of p_m to do the global search, which improves the global search ability of the proposed HTLBO algorithm.

(2) To integrate the advantages of the adaptive FPA and the TLBO algorithm, the population is divided into two parts according to the fitness of each individual. The two parts are evolved with different mechanisms and then exchange information with each other.
(3) The proposed approach is used to solve different scales of QoS-MCSC problems. The results are compared with the results of its competitors presented in the literatures. Experimental results show that the proposed approach performs better global search ability than other compared algorithms. Even in the later process of evolution, it has an excellent exploration ability.

However, the manufacturing cloud service composition considered in this paper is operated in static and certain service environments. There are several areas where the research can be carried on in the future. First, the service composition model can be proposed in dynamic and uncertain service environments. Many dynamic changes will affect the results of the manufacturing cloud service composition, such as the addition of new services, cloud service withdrawal, and so on. Second, there are some small cloud services with small manufacturing capabilities combined into one candidate service with strong manufacturing capabilities for users to select. Finally, other optimization methods might be combined to improve the efficiency of the proposed approach.

Author Contributions H. J. implemented the algorithm and wrote the paper. S. L. edited the paper and improved the quality of the article. C. J., H. H. and X. L. conducted the experiments and analyzed the data. H. J. and S. L. received funding. All authors have read and approved the final manuscript.

Funding The work is supported by the Natural Science Foundation of Guangdong Province under Grant Nos. 2018A030310216 and 2021A1515012395, the National Natural Science Foundation of China under Grant No. 51605169.

Availability of data and materials We confirm that data is open and transparent.

Declarations

Conflicts of interest The authors declare no competing interests.

Code availability Not applicable.

Consent for publication All authors agree to the publication of the paper.

References

1. Tao F, Zhang L, Venkatesh VC, Luo Y, Cheng Y (2011) Cloud manufacturing: a computing and service-oriented manufacturing model. Proc Inst Mech Eng Part B J Eng Manuf 225(10):1969–1976
2. Mourad MH, Nassehi A, Schaefer D, Newman ST (2020) Assessment of interoperability in cloud manufacturing. Robot Comput-Integr Manuf 61:101832
3. Zhang Y, Zhang G, Liu Y, Hu D (2017) Research on services encapsulation and virtualization access model of machine for cloud manufacturing. J Intell Manuf 28:1109–1123
4. Cremene M, Suciu M, Pallez D, Dumitrescu D (2016) Comparative analysis of multi-objective evolutionary algorithms for QoS-aware web service composition. Appl Soft Comput 39:124–139
5. Jin H, Yao X, Chen Y (2017) Correlation-aware QoS modeling and manufacturing cloud service composition. J Intell Manuf 28(8):1947–1960

6. Liu Y, Wang L, Wang XV, Xu X, Zhang L (2019) Scheduling in cloud manufacturing: state-of-the-art and research challenges. Int J Prod Res 57(15–16):4854–4879
7. Tao F, Zhao D, Hu Y, Zhou Z (2010) Correlation-aware resource service composition and optimal-selection in manufacturing grid. Eur J Oper Res 201(1):129–143
8. Xu X, Liu Z, Wang Z, Sheng QZ, Yu J, Wang X (2017) S-ABC: a paradigm of service domain-oriented artificial bee colony algorithms for service selection and composition. Futur Gener Comp Syst 68:304–319
9. Zhang W, Yang Y, Zhang S, Yu D, Li Y (2018) Correlation-aware manufacturing service composition model using an extended flower pollination algorithm. Int J Prod Res 56(14):4676–4691
10. Rao RV, Savsani VJ, Vakharia DP (2011) Teaching-learning-based optimization: a novel method for constrained mechanical design optimization problems. Comput Aided Des 43(3):303–315
11. Rao RV, Savsani VJ, Vakharia DP (2012) Teaching-learning-based optimization: an optimization method for continuous non-linear large scale problems. Inf Sci 183(1):1–15
12. El Ghazi A, Ahiod B (2018) Energy efficient teaching-learning-based optimization for the discrete routing problem in wireless sensor networks. Appl Intell 48:2755–2769
13. Tsai HC (2019) Confined teaching-learning-based optimization with variable search strategies for continuous optimization. Inf Sci 500:34–47
14. Kumar Y, Singh PK (2019) A chaotic teaching learning based optimization algorithm for clustering problems. Appl Intell 49:1036–1062
15. Li Z, Zhang X, Qin J, He J (2020) A reformative teaching-learning-based optimization algorithm for solving numerical and engineering design optimization problems. Soft Comput 24:15889–15906
16. Ji X, Ye H, Zhou J, Yin Y, Shen X (2017) An improved teaching-learning-based optimization algorithm and its application to a combinatorial optimization problem in foundry industry. Appl Soft Comput 57:504–516
17. Sun C, Zhao Y, Pan L, Liu H, Chen TY (2018) Automated testing of WS-BPEL service compositions: a scenario-oriented approach. IEEE Trans Serv Comput 11(4):616–629
18. Lu J, Zhou H, Zhu H, Zhang Y, Liang Q, Xiao G (2020) DCEM: a data cell evolution model for service composition based on bigraph theory. Futur Gener Comp Syst 112:330–347
19. Yu L, Zhang JX (2017) Service composition based on multi-agent in the cooperative game. Futur Gener Comp Syst 68:128–135
20. Yang Y, Yang B, Wang S, Jin T, Li S (2020) An enhanced multi-objective grey wolf optimizer for service composition in cloud manufacturing. Appl Soft Comput 87:106003
21. Akbaripour H, Houshmand M, van Woensel T, Mutlu N (2018) Cloud manufacturing service selection optimization and scheduling with transportation considerations: mixed-integer programming models. Int J Adv Manuf Tech 95:43–70
22. Khanouche ME, Attal F, Amirat Y, Chibani A, Kerkar M (2019) Clustering-based and QoS-aware services composition algorithm for ambient intelligence. Inf Sci 482:419–439
23. Jatoth C, Gangadharan GR, Buyya R (2017) Computational intelligence based QoS-aware web service composition: a systematic literature review. IEEE Trans Serv Comput 10(3):475–492
24. Huang BQ, Li CH, Tao F (2014) A chaos control optimal algorithm for QoS-based service composition selection in cloud manufacturing system. Enterp Inf Syst 8(4):445–463
25. Seghir F, Khababa A (2018) A hybrid approach using genetic and fruit fly optimization algorithms for QoS-aware cloud service composition. J Intell Manuf 29:1773–1792
26. Gavvala SK, Jatoth C, Gangadharan GR, Buyya R (2019) QoS-aware cloud service composition using eagle strategy. Futur Gener Comp Syst 90:273–290
27. Ramírez A, Parejo JA, Romero JR, Segura S, Ruiz-Cortés A (2017) Evolutionary composition of QoS-aware web services: a many-objective perspective. Expert Syst Appl 72:357–370
28. Li L, Cheng P, Ou L, Zhang Z (2010) Applying multi-objective evolutionary algorithms to QoS-aware web service composition. In: Cao L, Zhong J, Feng Y (eds.) Proceedings of the 6th international conference on advanced data mining and applications, Chongqing, China. Springer, Berlin, pp 270–281 (2010)
29. Yao Y, Chen H (2009) QoS-aware service composition using NSGA-II. In: Proceedings of proceedings of the 2nd international conference on interaction sciences: information technology, culture and human, Seoul, Korea, pp 358–363 (2009)
30. Chattopadhyay S, Banerjee A (2020) QoS-aware automatic web service composition with multiple objectives. ACM Trans Web 14:12

31. Zhang Q, Li H (2007) MOEA/D: a multiobjective evolutionary algorithm based on decomposition. IEEE Trans Evol Comput 11(6):712–731
32. Zhou J, Yao X, Lin Y, Chan FTS, Li Y (2018) An adaptive multi-population differential artificial bee colony algorithm for many-objective service composition in cloud manufacturing. Inf Sci 456:50–82
33. Yang YF, Yang B, Wang SL, Jin TG, Li S (2020) An enhanced multi-objective grey wolf optimizer for service composition in cloud manufacturing. Appl Soft Comput 87:106003
34. Liang H, Wen X, Liu Y, Zhang H, Zhang L, Wang L (2021) Logistics-involved QoS-aware service composition in cloud manufacturing with deep reinforcement learning. Robot Comput Integr Manuf 67:101991
35. Liu JW, Hu LQ, Cai ZQ, Xing LN, Tan X (2019) Large-scale and adaptive service composition based on deep reinforcement learning. J Vis Commun Image Represent 65:102687
36. Zhou J, Yao X (2017) A hybrid approach combining modified artificial bee colony and cuckoo search algorithms for multi-objective cloud manufacturing service composition. Int J Prod Res 55(16):4765–4784
37. Yang X-S (2012) Flower pollination algorithm for global optimization. In: Durand-Lose J and Jonoska N (eds.) proceedings of unconventional computation and natural computation, Orléans, France, Springer, Berlin, pp 240–249 (2012)

applied sciences

MDPI

Article

YOLOv4-MN3 for PCB Surface Defect Detection

Xinting Liao, Shengping Lv *, Denghui Li, Yong Luo, Zichun Zhu and Cheng Jiang

College of Engineering, South China Agricultural University, Guangzhou 510642, China; 20193079006@stu.scau.edu.cn (X.L.); lidh@stu.scau.edu.cn (D.L.); taffy@stu.scau.edu.cn (Y.L.); 20202157001@stu.scau.edu.cn (Z.Z.); 20193142013@stu.scau.edu.cn (C.J.)

* Correspondence: lvshengping@scau.edu.cn; Tel.: +86-187-1937-3880

Featured Application: An improved YOLOv4 algorithm for PCB surface defect detection can achieve higher detection accuracy and faster detection speed with lower memory consumption and fewer multiply–accumulate operations compared with the cutting-edge YOLOv4.

Abstract: Surface defect detection for printed circuit board (PCB) is indispensable for managing PCB production quality. However, automatic detection of PCB surface defects is still a challenging task because, even within the same category of surface defect, defects present great differences in morphology and pattern. Although many computer vision-based detectors have been established to handle these problems, current detectors struggle to achieve high detection accuracy, fast detection speed and low memory consumption simultaneously. To address those issues, we propose a cost-effective deep learning (DL)-based detector based on the cutting-edge YOLOv4 to detect PCB surface defect quickly and efficiently. The YOLOv4 is improved upon with respect to its backbone network and the activation function in its neck/prediction network. The improved YOLOv4 is evaluated with a customized dataset, collected from a PCB factory. The experimental results show that the improved detector achieved a high performance, scoring 98.64% on mean average precision (*mAP*) at 56.98 frames per second (*FPS*), outperforming the other compared SOTA detectors. Furthermore, the improved YOLOv4 reduced the parameter space of YOLOv4 from 63.96 M to 39.59 M and the number of multiply-accumulate operations (*Madds*) from 59.75 G to 26.15 G.

Keywords: printed circuit board; surface defect detection; YOLOv4; MobileNetV3

Citation: Liao, X.; Lv, S.; Li, D.; Luo, Y.; Zhu, Z.; Jiang, C. YOLOv4-MN3 for PCB Surface Defect Detection. *Appl. Sci.* **2021**, *11*, 11701. https://doi.org/10.3390/app112411701

Academic Editor: Alfio Dario Grasso

Received: 18 November 2021
Accepted: 6 December 2021
Published: 9 December 2021

1. Introduction

Printed circuit board (PCB) surface defects are local areas of surface that do not meet design or manufacturing requirements. However, PCBs' surface defects not only affect their aesthetics but also their performances and functionalities. Therefore, surface defect detection is indispensable in managing PCB production quality. In modern factories, basic electric tests, such as manual visual inspection, are still commonly used [1]. However, manual detection relies heavily on many experienced inspectors exercising extensive concentration under strong illumination, while electric tests, as a contact-detection technique, may themselves cause defects in the board. Therefore, more attention is being paid to non-contact automated optical inspection (AOI), and practice indicates that it has greatly improved detection accuracy and efficiency [2].

As the core of AOI software, detection algorithms can be divided into traditional detectors and deep learning (DL)-based detectors according to their feature extraction methodology. Traditional detectors utilize traditional machine learning methods, an image-processing approach and prior knowledge to extract low-level defect feature. Wang et al. [3] extracted the center, aperture, roundness and area of holes as features, and compared the feature information between testing images and standard images to detect the defects in PCB holes. Gaidhane et al. [4] utilized companion matrices of testing images and standard images to construct a symmetrical matrix and adopted its rank as a similarity metric to

determine the defects detected on PCB boards. Eun et al. [5] combined the speeded-up robust features and random forest algorithms to extract PCB fault patterns and drew a weighted kernel density estimation map for defect detection based on probability values. Fonseka et al. [6] integrated color transformation, graph-cut based segmentation and k-means color clustering to detect solder bridging, solder voids and excess solder. Liu et al. [7] used the mathematical morphology method to obtain standard images, and then an image aberration detection algorithm was introduced to detect PCB defects. These traditional detectors highly rely on prior knowledge to determine previously seen features, or to store a large number of standard images and then precisely align testing images to them for element matching. Therefore, traditional detectors are not conducive to generalization between different application scenarios.

A DL-based detector utilizes a convolutional neural network (CNN) to extract and defect features and learn inherent patterns of defects, automatically, without standard images or manual design rules [8,9], which greatly improves detection accuracy, efficiency and model generalization. DL-based detectors can be roughly divided into two categories, two-stage detectors and one-stage detectors. Two-stage detectors divide the training process into candidate boxes (region proposals), where extraction and feature classification are based on region proposals over two steps. R-CNN [10], Fast R-CNN [11] and Faster RCNN [12] are classical two-stage detectors. These models can achieve high detection accuracies and precisions of location but are trapped by their detection speeds. One-stage detectors directly regress bounding boxes and probabilities for each object in an input image, simultaneously, without region proposals. YOLO series models (YOLO, YOLOv2, YOLOv3, YOLOv4) are widely used one-stage detectors [13]. These detectors can speed up detection but suffer from a drop in accuracy.

Both one-stage and two-stage DL-based detectors are constituted by backbone, neck and prediction networks (also called a head network), and improvements for the three networks are continuously emerging for better matching within different application scenarios [8]. DL-based detectors, in surface defect detection, have attracted much attention in recent years. Many improved YOLO and R-CNN detectors [14–16] have been developed for surface defect detection with the available labeled surface defect datasets in the industry, such as from steel, metro tunnel and commutator train production. Many studies and practitioners have also introduced this mechanism into PCB defect detection.

Ding et al. [17] developed a tiny defect detection network (TDD-Net) based on fast-RCNN for PCB defect detection and employed an online example of a hard mining technique in their training phase to alleviate the adverse effects of small datasets and sample imbalance. Hu et al. [18] improved the two-stage Faster RCNN for PCB defect detection. First, they replaced the backbone and neck network of Faster RCNN with ResNet50 and feature pyramid networks (FPN) respectively. Second, they introduced a guided anchor region proposal network as a substitute of the original region proposal network for better anchor generation. Additionally, the residual module of ShuffleNetV2 was adopted in their backbone to reduce the model parameter and operation. Dai et al. [1] employed YOLO to locate hundreds of small and dense solder joints automatically in PCB images before defect detection. Zhang et al. [19] combined a dual attention mechanism and Path Aggregation Feature Pyramid Network in MobileNetV2 to build PCB defect detector. The above DL-based models exhibited high detection accuracy but suffered from a large number of parameters and high computational cost. Meanwhile, inter-class diversity of surface defect was not considered in these studies, whereas the same category of detected defect possesses great difference in their morphologies and patterns in practice. Collecting real defect samples from PCB production factory, and establishing detectors with high detection accuracy, fast detection speed, low memory consumption and low multiply-accumulate operations (*Maccs* or *Madds*) is a promising tendency.

To solve the above-mentioned problems, this study proposes a cost-effective DL-based detector called as YOLOv4-MN3 based on cutting-edge YOLOv4 and MobileNetV3 lightweight network. First, we design a PCB image acquisition device. Then, surface defect

images are collected into a dataset with 2008 samples, in which were contained bumpy or broken line, clutter, scratch, line repair damage, hole loss and over oil-filling—the six categories of the most common defects. Second, we utilize the MobileNetV3 lightweight network with small number of parameters and lower *Madds*, replacing CSPDarknet53 as the backbone network. Third, the influence of different activation functions in the neck and prediction networks are tested and compared, for which the Mish activation function is selected. The experimental results show that the proposed YOLOv4-MN3 for PCB surface defect detection achieves a higher detection accuracy, faster detection speed and lower *Madds* compared to SOTA detectors.

The contents of this paper are organized as follows. Section 2 presents the workflow of the proposed detection approach and the architecture of the proposed detector YOLOv4-MN3. Section 3 introduces the building of a customized dataset, which includes a surface defect image acquisition device, defect image collection, data augmentation and labeling. Section 4 gives the training and detection results with comprehensive comparisons. Section 5 contains our conclusions.

2. Methodology

2.1. Framework of the Methodology

The proposed YOLOv4-MN3 for PCB surface defect detection is established based on a cutting edge one-stage detector, YOLOv4, and it consists of dataset building, model training and performance evaluation in three steps, as described in Figure 1. First, all the PCB defect images are collected by a specially designed image acquisition device, after which the augmentation and annotation labeling are conducted. Then, YOLOv4 is modified and trained based on a customized dataset. Finally, the performance of the proposed YOLOv4-MN3 and other SOTA detectors are evaluated and compared.

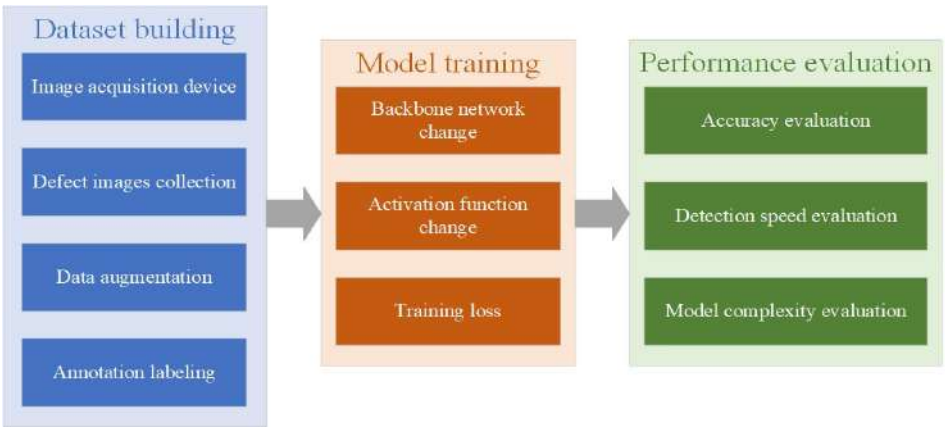


Figure 1. Framework of the proposed methodology.

Dataset building: The PCB surface defect images were collected by an image acquisition device specially designed and manually labeled by the program labelImg to generate training and image-defect testing datasets with corresponding annotation files.

Model training: Bumps, broken lines, clutter, scratched, line repair damage, hole losses and over oil-filling, the six categories of the most common surface defects, were selected for the training. The original backbone network CSPDarknet53 of YOLOv4 was replaced by VGG16, Resnet50, Darknet53, MobileNetV2 and MobileNetV3 for the purposes of selecting an appropriate backbone that could decrease memory consumption and computational cost. To better fit to customized defect dataset, the activation functions of the neck and prediction networks in YOLOv4 were replaced with five different activation functions; thus, six different YOLOv4-MN3 detectors were constructed. Finally, fifteen detectors, including ten backbone or activation function-modified YOLOv4 detectors—original YOLOv4, YOLOv3, Faster RCNN, Retinanet and SDD—were trained with PyTorch.

Performance evaluation: The average precision (*AP*), mean average precision (*mAP*) and F_1 score were adopted as metrics of detection accuracy, and frames per second (*FPS*) as a speed metric. The models' parameters (*Params*) and *Madds* were collected to evaluate model complexity.

2.2. Proposed YOLOv4-MN3

YOLOv4 [20], as one of the cutting-edging one stage DL-based models for object detection, makes many improvements on YOLOv3 [21], including its network architecture, activation function, loss function etc., and integrates many training tricks. The framework of YOLOv4 can also been divided into its backbone, neck and prediction networks. In YOLOv4, a cross-stage partial darknet53 network (CSPDarknet53) is used in the backbone to extract features from the input images. Spatial pyramid pooling (SPP) [22] and path aggregation networks (PANet) [23] were employed as the neck networks to generate a feature pyramid. SPP + PANet, in neck networks, fuse low-level spatial features with accurate location information and high-level semantic features with high semantic information bi-directionally [20]. The prediction network applies anchor boxes to multiscale feature maps of neck network to generate detection boxes.

2.2.1. YOLOv4-MN3 Architecture

YOLOv4 employs the cross-stage partial network (CSPNet) in Darknet53 to construct a CSPDarknet53 backbone. CSPNet partitions the feature map of the input layer into two parts and then merges them through the proposed cross-stage hierarchy for the purpose of enriching gradient combination [24]. However, CSPDarknet53 suffers from high memory consumption, with 29 convolution layers and 27.6 million parameters. Replacing CSPDarknet53 by a lightweight model with fewer parameters while preserving its detection accuracy is a worthy attempt.

For this study, a cost-effective detector YOLOv4-MN3 was developed, in which the CSPDarknet53 in YOLOv4 was replaced by the lightweight network MobileNetV3. MobileNetV3 utilizes depthwise separable convolution to construct feature maps for each layer. The main convolution process is composed of two parts, as given in Figure 2. The first part is depthwise convolution, and it introduces a filter for each input channel and conducts convolutions for each pair of filter and feature map separately. Its second part is pointwise convolution, and it convolutes a 1×1 filter to channels output from the depthwise convolution for the purposes of increasing or decreasing the depth of a feature map. If the convolution input and output are 16 and 32, respectively and the filter size is 3×3, then the parameters of standard convolution and depth separable convolution are $16 \times 32 \times 3 \times 3 = 4608$, $16 \times 3 \times 3 + 32 \times 16 \times 1 \times 1 = 656$ respectively, which indicates that depthwise separable convolution can successfully minimize the number of parameters.

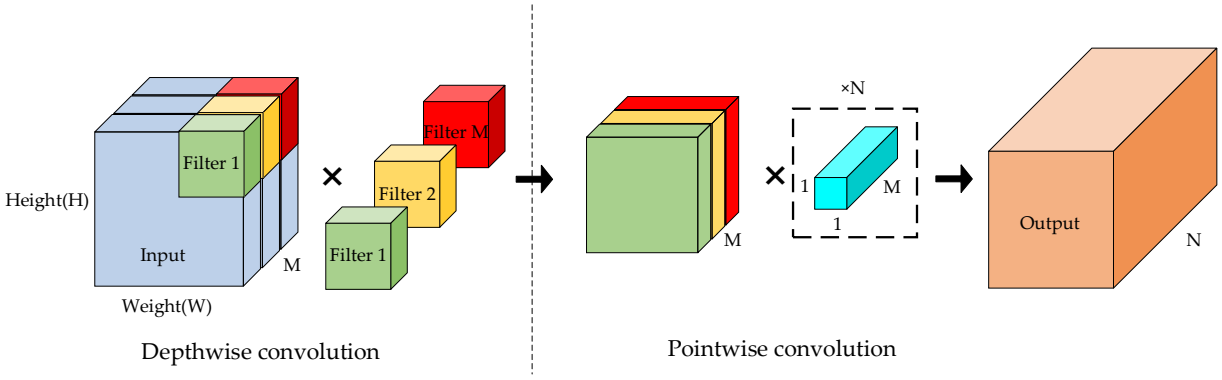


Figure 2. Depthwise separable convolution.

In order to improve the detection accuracy, MobileNetV3 introduces the squeeze and excitation attention module into the bottleneck of MobileNetV2, the basic unit (Bneck) of MobileNetV3, given in Figure 3. Meanwhile, MobileNetV3 modifies the Swish activation function to improve detection accuracy. Experimental results verify the effectiveness and superiority of MobileNetV3 because it can achieve high detection speed and accuracy simultaneously [25]. The MobileNetV3 has MobileNetv3-Small and MobileNetV3-Large—two versions, used according to the depth of its layers. MobileNetV3-Large is employed in YOLOv4-MN3 for a balance of accuracy and speed, based on some initial experimentation. The architecture of YOLOv4-MN3 is given in Figure 4, and only the first 15 Bnecks of MobileNetV3-Large are used to extract three different (52×52, 26×26, 13×13) spatial resolution feature maps, which are directly matched to the input size of the YOLOv4 neck network.

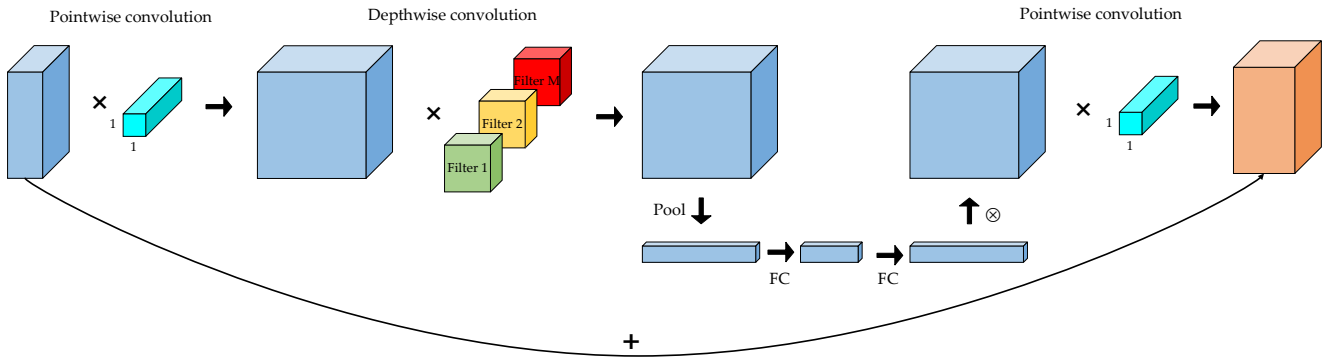


Figure 3. Bneck of MobileNetV3.

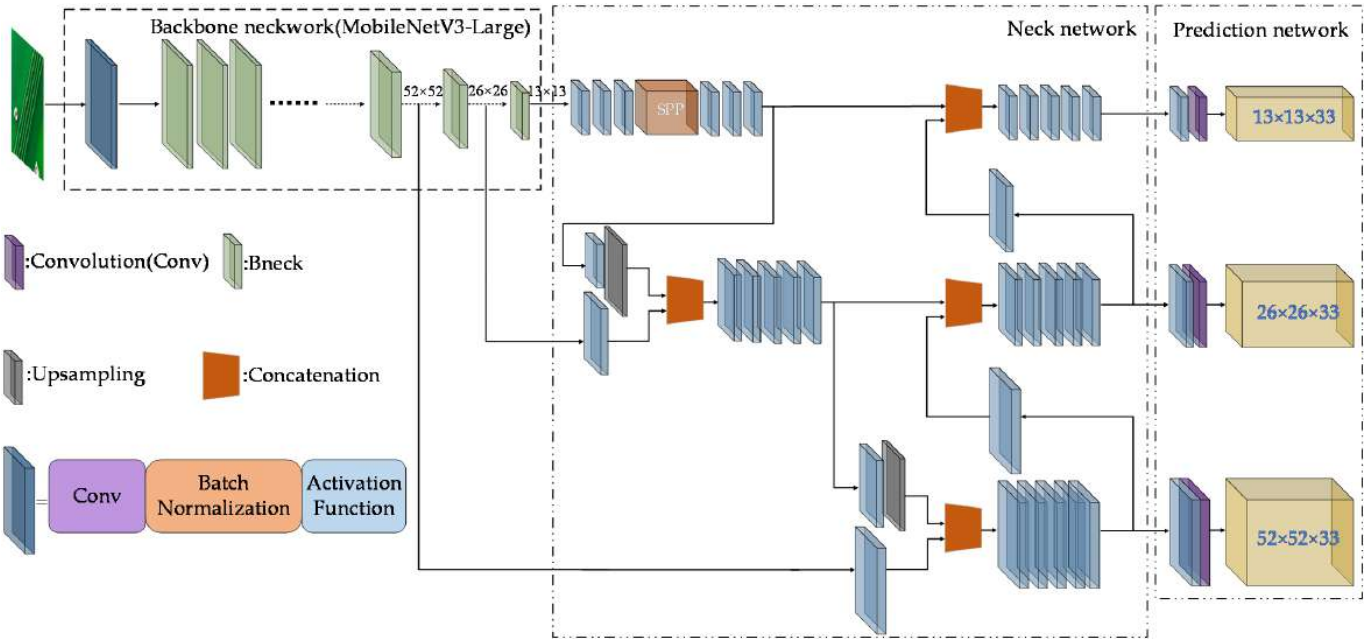


Figure 4. YOLOv4-MN3 architecture.

2.2.2. Activation Functions

It is worth noting that few studies use YOLOv4 to detect PCB surface defect and discuss the influence of different activation functions. Previous research and initial training experiments [26] have shown that activation functions with different properties influence detection performance. Therefore, we analyze the characteristics of different activation functions and conduct comparison experiments between them to select most suitable one

according to their training and detection performances on the customized dataset. Since MobileNetV3 has optimized its activation function, this study only selects the activation function for the neck/prediction network.

Neural network' activation functions greatly influence their customized training convergence procedures because of their derivative, monotonicity properties, among others [27]. The selection of activation functions that perform will in training converging and detection accuracy is an essential step for model establishment [28]. Therefore, the sigmoid, tan hyperbolic (Tanh), rectified linear unit (ReLU), leaky ReLU, Swish, and Mish activation functions were implemented and configured for the training files, based on our customized dataset, to optimize the selection of an activation function.

Figure 5 and Table 1 show the plots and expressions of the six activation functions, respectively. Sigmoid and Tanh, as traditional sigmoid-like units, have dominated neural network practice for several decades. However, they are computationally expensive, and easily lead to gradient vanishing during training. The activation functions of ReLU and Leaky ReLU are widely used in deep CNN. ReLU and Leaky ReLU are not symmetric functions and are unlike the symmetric functions of Tanh and Sigmoid. Thus, they can deal with the problem of gradient vanishing and can update weights continuously during their entire training processes [29,30]. Leaky ReLU introduces an alpha parameter to ensure the gradient of each node would not be zero during the propagation process so that the training loss is easily be trapped into local optima. Swish is a non-monotonic and smooth activation function [31], in which the non-monotonicity property is designed to handle gradient vanishing, while its smoothness is beneficial for model generation and optimization.

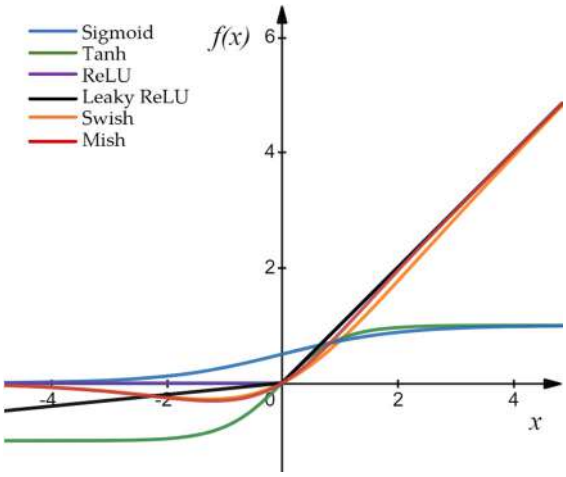


Figure 5. Plots of different activation function.

Table 1. Expression of different activation functions.

Activation Function	Expression	Activation Function	Expression
Sigmoid	$f(x) = 1/(1+1+e^{-x})$	Tanh	$f(x) = (e^{x} - e^{-x})/(e^{x} + e^{-x})$
ReLU	$f(x) = \max(0, x)$	Leaky ReLU	$f(x) = \max(\alpha x, x)$
Swish	$f(x) = x \times 1/(1 + e^{-x})$	Mish	$f(x) = x \times \tanh(\ln(1 + e^{x}))$

Similar to Swish, Mish is a non-monotonic and smooth function with a range of [≈−0.31,∞). Mish outperforms other activation functions in many DL-based detectors across challenging datasets, and we can easily define a Mish activation layer in any standard DL framework for its implementation [26].

2.2.3. Loss Function

The loss function of YOLOv4 includes three parts, confidence, classification, and bounding box regression loss. YOLOv4 employs a novel complete-intersection over union (*IoU*) loss (*CIoU*), replacing the mean–squared–error loss adopted in YOLOv3 with bounding box regression loss [20]. *CIoU* takes the overlap area, center point distance and aspect ratio into consideration simultaneously, improving detection speed and accuracy. *CIoU* introduces a penalty item αv on the basis of distance-*IoU* loss to impose a consistency of aspect ratio for the ground truth bounding box (bb^{gt}) and the prediction bounding box (bb). *CIoU* loss can be defined as in Equation (1)

$$\begin{gathered} Loss_{CIoU} = 1 - (IoU - \frac{\rho^2(b,b^{gt})}{c^2} - \alpha v) \\ \alpha = \frac{v}{1-IoU+v}, v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \end{gathered} \tag{1}$$

where b, b^{gt} are the centers of bb, bb^{gt} respectively, $\rho(\cdot)$ denotes Euclidean distance, c represents the diagonal length of the smallest enclosing rectangle covering bb, bb^{gt} and α is a positive trade-off value, v means the consistency of aspect ratio. w,w^{gt} are the widths of bb, bb^{gt} respectively. h, h^{gt} are the heights of bb, bb^{gt}, respectively.

3. Dataset Building

3.1. Image Acquisition Device

A specially designed image acquisition device is depicted in Figure 6, and it consists of an auxiliary module, an illumination module and an image acquisition module. The auxiliary module provides a physical framework for the installation of the illumination and image acquisition modules, and is composed of a dark box (a), motion control parts (b) and a movable loading platform (c). The illumination module (d) provides suitable lights for the camera's image acquisition module to take photographs. The image acquisition module (e) is responsible for collecting PCB images and connects to a computer for image storage and preparation, and consists of a camera support framework, Hikvision MV-CE120-10GC industrial camera with 12 million pixels and an external computer. The camera captures PCB images sequentially under the stable light source provided by the illumination module. The maximum field of view is 120 mm × 90 mm, and multi-point shooting was used for PCB s larger than this size, in which the camera is moved by motion control components to locate its position for each shooting.

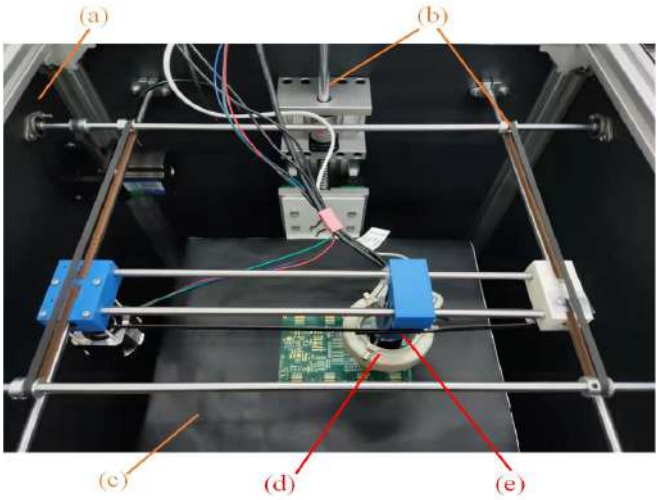
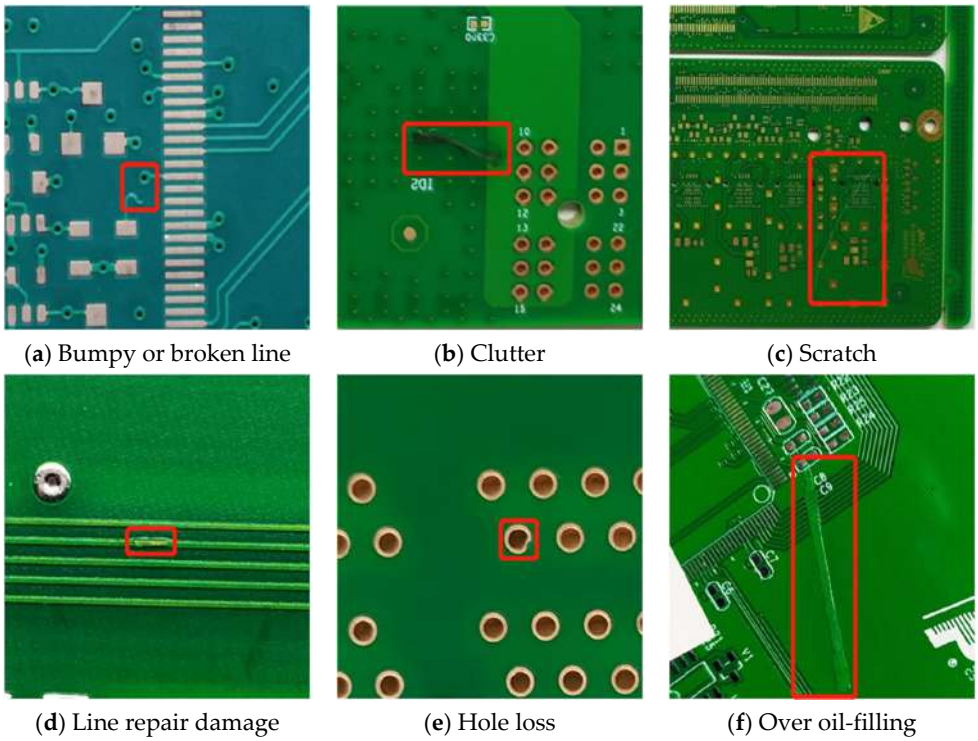


Figure 6. PCB image acquisition device. (a) dark box, (b) motion control parts, (c) platform, (d) illumination module and (e) image acquisition module.

3.2. Defect Images Collection

The original 2008 surface defect images, with one surface defect in each, were collected from a PCB production factory in Guangzhou, China, with the device given in Figure 6, and the size of each defect image was 4024 × 3036 pixels. Six categories of defects, including bumpy or broken line, clutter, scratch, line repair damage, hole loss and over oil-filling, were selected, as these six categories of defect account for more than 80% of surface defects in PCB factories, according to historical statistical data. The instances of each category are given in Figure 7. Each type of surface defect possesses several different morphologies and patterns. As shown in Figure 8, there are five typical morphologies of bumpy or broken lines.



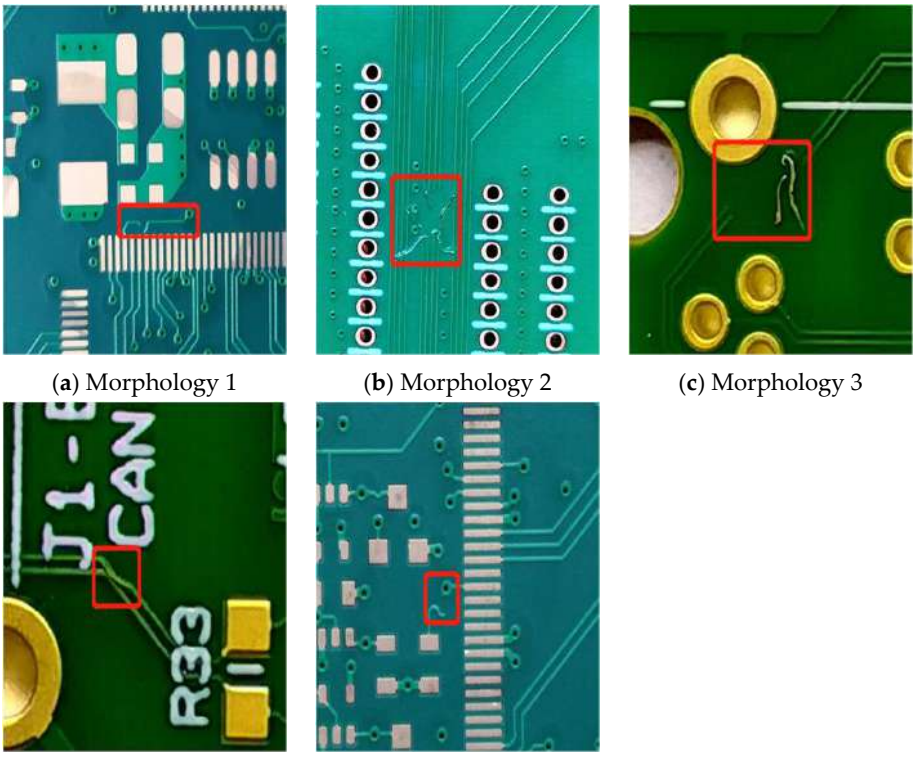
(**a**) Bumpy or broken line (**b**) Clutter (**c**) Scratch

(**d**) Line repair damage (**e**) Hole loss (**f**) Over oil-filling

Figure 7. PCB surface defect instances.

3.3. Data Augmentation and Labeling

Data augmentation can improve the sample diversity and enhance model generalization. It has been widely used in DL-based model development, especially for industrial applications for which it is difficult to obtain large, labeled samples [27]. Random rotation, cropping, translation, horizontal and vertical flipping, luminance balance, etc. are the commonly used data augmentation techniques. Random clipping, image rotation and luminance changes were conducted to augment the original images in this study. Random clipping crops the image randomly, with a mouse click positioned as the operation's center, to get different sizes of image. Rotation augmentation rotates the image at 90°, 180° and 270° angles, such that four different angles of the same defect can be obtained. Luminance changes specify brightness values, adjusting the brightness of the original image. Rotation and randomly clipping images aid detection performance and the robustness of improvement. Luminance changes simulate the deviating brightness of different environmental lighting and improves models' adaptability to different lighting [32]. Some instances of these augmentations are given in Figure 9.

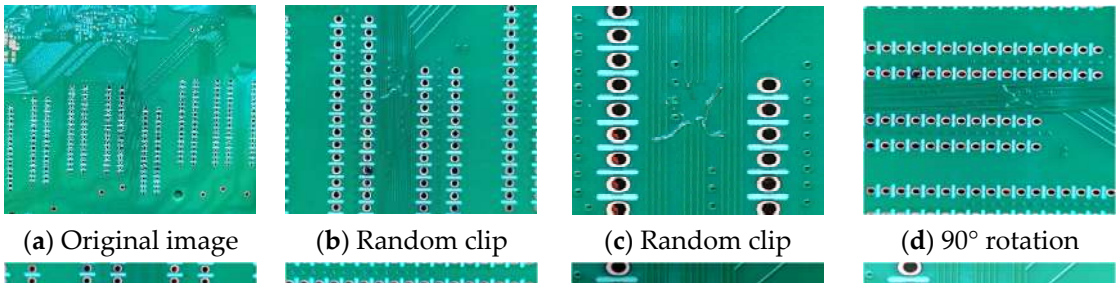


(**a**) Morphology 1 (**b**) Morphology 2 (**c**) Morphology 3

(**d**) Morphology 4 (**e**) Morphology 5

Figure 8. Five different morphologies of bumpy or broken line.

Taking the original high-resolution images as input is conducive to improving detection accuracy. However, large input sizes greatly increase model burden and computing resources. Therefore, all images of the 3018 × 4096-pixels dataset were resized to 416 × 416 pixels in this study. The data augmentation was performed in Python and 19,029 images were obtained, and the number of images belonging to each category of defect are given in Table 2.



(**a**) Original image (**b**) Random clip (**c**) Random clip (**d**) 90° rotation

(**e**) 180° rotation (**f**) 270° rotation (**g**) Luminance change (**h**) Luminance change

Figure 9. Data augmentation instances.

Table 2. PCB surface defect dataset.

Defect Category	Original Images	Augmented Images
bumpy or broken line	345	3090
clutter	332	3458
scratch	443	3463
line repair damage	298	2816
Hole loss	263	3132
over oil-filling	327	3070

Each of the 19,029 defect images were manually marked with a rectangle and labelled with their category. The annotated images mark the baseline truth for each defect, and they can be utilized to evaluate the training loss of *IoU* when combined with the predicted bounding box. The surface defect in each image was labeled by the program LabelImg and stored in PASCAL VOC format. Finally, we randomly split the dataset into training and testing sets, which included 90% and 10% of the images, respectively.

4. Experiment

We conduct three experiments based on the customized PCB dataset to validate the proposed YOLOv4-MN3 in this section. First, the accuracy, parameters, operators and detection-speed performance of different backbone networks are compared for the selection of MobileNetV3. Second, the accuracy performances of different activation functions in the neck/prediction network are compared to facilitate Mish selection. Third, the performance of YOLOv4-MN3 is compared with Faster R-CNN, RetinaNet, SSD, YOLOv3 and YOLOv4 to verify the superiority of the proposed approach. We use the deep learning framework PyTorch1.7 to implement YOLOv4-MN3 and all the compared SOTA models. The experimental environment was ubuntu18.04, CUDA11.0, CUDNN8.0.5 and an NVIDIA GeForce RTX 3080.

4.1. Evaluation Metrics

AP, *mAP*, and F_1 score are taken to evaluate the detection accuracy, and they can be defined as follows:

$$AP = \int_0^1 P(R)dR \tag{2}$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{3}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{4}$$

where $P(R)$ is the precision of a class when recall is R and C is the number of all categories in the image dataset. Recall and precision can be defined as follows:

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$P = \frac{TP}{TP + FP} \tag{6}$$

where *TP*, *FN* and *FP* refer to true positive, false negative and false positive respectively. The prediction anchor box will be categorized into true positive (*TP*) if the IoU is greater than or equal to a pre-setting threshold T, and it is set as 0.5 in this study.

The number of model parameters (*Params*), *Madds* are used to quantitatively evaluate the computational burden of model [25]. And *FPS* is used to evaluate the detection speed. The higher *mAP*, F_1 score and *FPS* but lower *Params* and *Madds*, the better the detector.

4.2. Training Details for YOLOv4-MN3

Pre-training on large, natural-image datasets is a common strategy for DL-based detector training. Therefore, we pre-trained the proposed YOLOv4-MN3 on VOC2007 first. Then, the model training is split into two stages. In the first stage, pre-trained parameters are taken as the initial weights and the weights of the backbone network are frozen, then the parameters of the neck and head networks are trained and optimized. In the second stage, all the parameters are trained based on the first stage network weights. An Adam optimizer is employed to update the parameters with a weight decay of 5×10^{-4}. A total of 100 epochs are performed in the model training, and both first and second stage last 50 epochs. The memory occupancy of the first stage is smaller than the second stage because of the pre-training. Accordingly, the batch sizes were set to eight and two for the two training stages, respectively, and their learning rates were set to 0.001 and 0.0001, respectively.

Optimizing for anchor size that can match well with the size of detected defect facilitates the model's achieving better detection performance. The default anchor size of YOLOv4 is generated based on natural image objects that were not optimized for PCB surface defect detection. Therefore, K-means is employed to cluster and reset the anchor size based on the labeled data to better match the size of defect. Nine anchor boxes of sizes (26,23), (38,38), (52,48), (60,69), (94,83), (116,131), (140,41), (157,284), (239,149) were obtained after clustering. Different three-scale prediction output layers are employed to detect defects of different scales in YOLOv4-MN3. The three smallest anchor boxes were allocated to the 52×52 prediction output layers, the three middle-sized anchor boxes were assigned to 26×26 prediction output layers, and the 13×13 prediction output layers used the three largest anchor boxes.

4.3. Impacts of Different Backbone Networks

The CSPDarknet53 backbone network in the original YOLOv4 suffers from a large number of parameters and a huge computational burden. We tried replacing it with lightweight networks; however, not without a loss of detection performance. Six backbone networks—VGG16, Resnet50, Darknet53 (backbone network of YOLOv3), CSPDarknet53 (backbone network of YOLOv4), MobileNetV2 and MobileNetV3—were selected and comparison experiments were conducted to verify their impact on YOLOv4. VGG16 and Resnet50 are commonly used backbone networks, while Darknet53 and CSPDarknet53 are the SOTA backbone networks for one-stage detectors. The low-power and low-latency of parameterized MobileNet is commonly used for small models [33]. The comparison results are given in Table 3 according to the evaluation metrics given in Section 4.1.

Table 3. Performance of different backbone network.

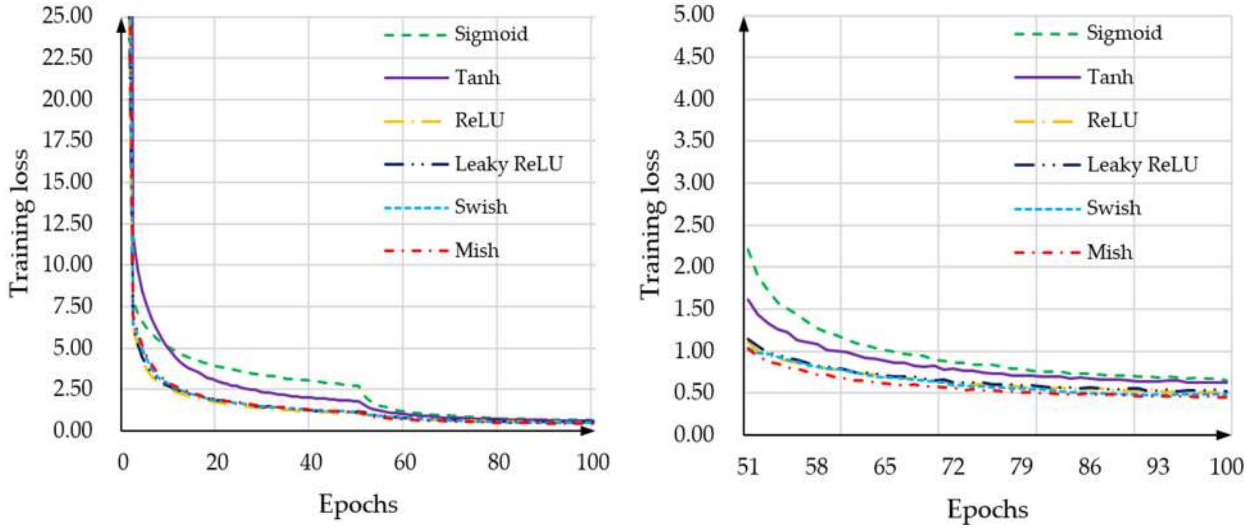
Backbone Network	*mAP* (%)	F_1 (%)	*Params* (M)	*Madds* (G)	*FPS*
VGG16	96.00	93.17	51.80	206.03	55.58
Resnet50	95.10	90.67	61.54	53.98	56.74
Darknet53	95.61	93.00	77.93	74.24	54.93
CSPDarknet53	96.49	95.17	63.96	59.75	51.64
MobileNetV2	94.95	91.33	38.66	26.71	56.99
MobileNetV3	**97.26**	**95.83**	**39.59**	**26.15**	**57.12**

It can be seen that MobileNetV3 obtained the highest *mAP* among all backbone networks. Compared with VGG16, Resnet50, Darknet53, CSPDarknet53, and MobileNetV2, the value of *mAP* obtained by MobileNetV3 improves by 1.26%, 2.16%, 1.65%, 0.77% and 2.31% respectively. Meanwhile, MobileNetV3 achieved the highest F_1 score among the six backbone networks with 2.66%, 5.16%, 2.83%, 0.66% and 4.50% greater improvement over VGG16, Resnet50, Darknet53, CSPDarknet53, and MobileNetV2 respectively. The results of the *mAP* and F_1 score presented in Table 3 indicate that MobileNetV3 outperforms other compared backbone networks in prediction accuracy. Comparing the results from

Params and *Madds* shows that MobileNetV3 greatly simplified the backbone network, requiring only 76.43%, 64.33%, 50.80% and 61.90% of the model parameters used by VGG16, Resnet50, Darknet53 and CSPDarknet53, respectively, and it achieved the lowest *Madds* among these networks as well. The results also show that MobileNetV3 exhibited no significant difference between either MobileNetV2 or MobileNetV3 with respect to *Params* and *Madds*. The *FPS* comparison indicated that MobileNetV3 achieved the highest detection speed, with 57.12 *FPS*. Notably, the *Madds* of MobileNetV3 reduced by 33.60 G, whereas its *FPS* increased by 5.48 as compared with CSPDarknet53 in the original YOLOv4. In summary, MobileNetV3 achieved the highest detection accuracy, low model parameters, lowest *Madds* and fastest detection speed, making it distinctly advantageous for detection.

4.4. Impacts of Different Activation Functions

Popular activation functions in neck and prediction networks, including Sigmoid, Tanh, ReLU, Leaky ReLU, Swish, and Mish were implemented and experimentally compared based on the customized dataset. The detection performance of the six activation functions is shown in Figure 10 and Table 4. In Figure 10a, the training losses of Sigmoid and Tanh are higher than other four activation functions at the first 50 epochs, and they tend to stop converging at the end of the first training stage. However, the training losses of all activation functions decreased again, at the 51st epoch, when all the parameters were unfrozen at the second training stage. Since the training losses were close to each other in the second stage, we selected the training loss from epoch 51–100, given in Figure 10b, to amplify the loss difference between different activation functions clearly. The detailed comparison result given in Figure 10b shows that Mish obtained the lowest training loss, which indicates it had the best training result, generally. It can be roughly concluded that Mish outperformed the other activation functions.



(**a**) Training loss of 100 epochs (**b**) Training loss from epoch 51 to 100

Figure 10. Training loss of different activation functions.

As shown in Table 4, the detection performance varies for different activation functions. Mish's activation function obtained the best *mAP* value, of 98.64%, a 2.95%, 2.02%, 1.62%, 1.38% and 1.36% increase over Sigmoid, Tanh, ReLU, Leaky ReLU and Swish, respectively. Meanwhile, Mish achieved the highest F_1 score, with a 1.5–3.3% increase as compared with the other five activation functions. The *AP* value obtained by Mish outperformed the compared activation functions in the surface defect categories bumpy or broken line, scratch and line repair damage, and showed only slight inferiority to the best of the five activation functions for the defects of clutter and hole loss. One interesting finding is that the detection accuracy for scratches was worse than other defects for all

activation functions. The reason may be that the color of the scratches was similar to their backgrounds, and they are characteristically tiny lines occupying a small number of pixels relative to the overall board, which is not conducive to distinguishing them from their backgrounds. Meanwhile, the diversity of scratch patterns and scales may have hindered the detection performance. However, Mish still achieved 94.20% *AP* and exhibited obvious superiority over the compared activation functions, which further shows that it is suitable for improving detection accuracy.

Table 4. Performance of different activation functions on YOLOv4-MN3.

Activation Function	*AP* (%)						*mAP* (%)	F_1 (%)
	Bumpy or Broken Line	Clutter	Scratch	Line Repair Damage	Hole Loss	Over Oil-Filling		
Sigmoid	98.12	96.80	87.18	97.44	95.64	98.95	95.69	94.50
Tanh	98.56	95.94	88.33	98.43	99.15	99.32	96.62	95.67
ReLU	98.89	99.41	88.53	96.35	98.95	99.99	97.02	95.75
Leaky ReLU	99.10	98.98	90.92	95.73	99.51	99.32	97.26	95.83
Swish	99.66	98.41	86.80	98.87	99.91	100.00	97.28	96.33
Mish	**100.00**	**98.69**	**94.20**	**99.36**	**99.61**	**100.00**	**98.64**	**97.83**

4.5. Comparison of Different Detectors

The detection accuracy, speed and model complexity of YOLOv4-MN3 were tested and compared with the other five SOTA detectors, Faster RCNN, RetinaNet, SSD, YOLOv3, and original YOLOv4. The experimental results of the different indicators are illustrated in Figure 11. It can be seen that the proposed YOLOv4-MN3 achieve the highest detection accuracy, with a 98.64% *mAP* value, which is 5.69%, 4.58%, 3.28%, 1.14% and 2.15% superior to Faster-RCNN, RetinaNet, SSD, YOLOv3 and YOLOv4, respectively. Meanwhile, YOLOv4-MN3 outperformed the other five compared models in F_1, at 20.66%, 14.00%, 10.66%, 2.33% and 2.66% higher than Faster-RCNN, RetinaNet, SSD, YOLOv3 and YOLOv4 respectively. In terms of model complexity, YOLOv4-MN3 greatly reduced its parameters and *Madds*. The number of *Params* needed by YOLOv4-MN3 reduced by 28.94%, 64.32% and 61.90% those of Faster RCNN, YOLOv3 and YOLOv4, respectively, and the *Madds* of YOLOv4-MN3 was 90.45 G, 39.29 G and 33.60 G lower than of Faster RCNN, YOLOv3 and YOLOv4, respectively. The reason for this is that MobileNetV3 uses depthwise separable convolution, replacing the standard convolution, allowing the convolution weights and operations to be reduced greatly. In addition, RetinaNet and SSD had small numbers of parameters but large *Madds*. The reason for this is that the number of parameters indicates the space complexity or storage consumption, and *Madds* is the metric of time complexity. Detectors with small *Params* cannot ensure lower *Madds* or higher detection speed, and vice versa. For example, the activation function and pooling operations increased *Madds* while *Params* remained unchanged. Meanwhile, we can see that the *Madds* of YOLOv4-MN3 is 43.72 G and 90.26 G lower than RetinaNet and SSD, although with a slight increase in parameters. YOLOv4-MN3 also had significant advantage in detection speed, and its *FPS* value was greater by 10.60, 3.01, 1.98, 11.49 and 5.34 than those of Faster RCNN, RetinaNet, SSD, YOLOv3 and YOLOv4. Based on the comparison of SOTA detectors, the YOLOv4-MN3 model delivered significant advantage in terms of detection accuracy, model parameters, operations and detection speed.

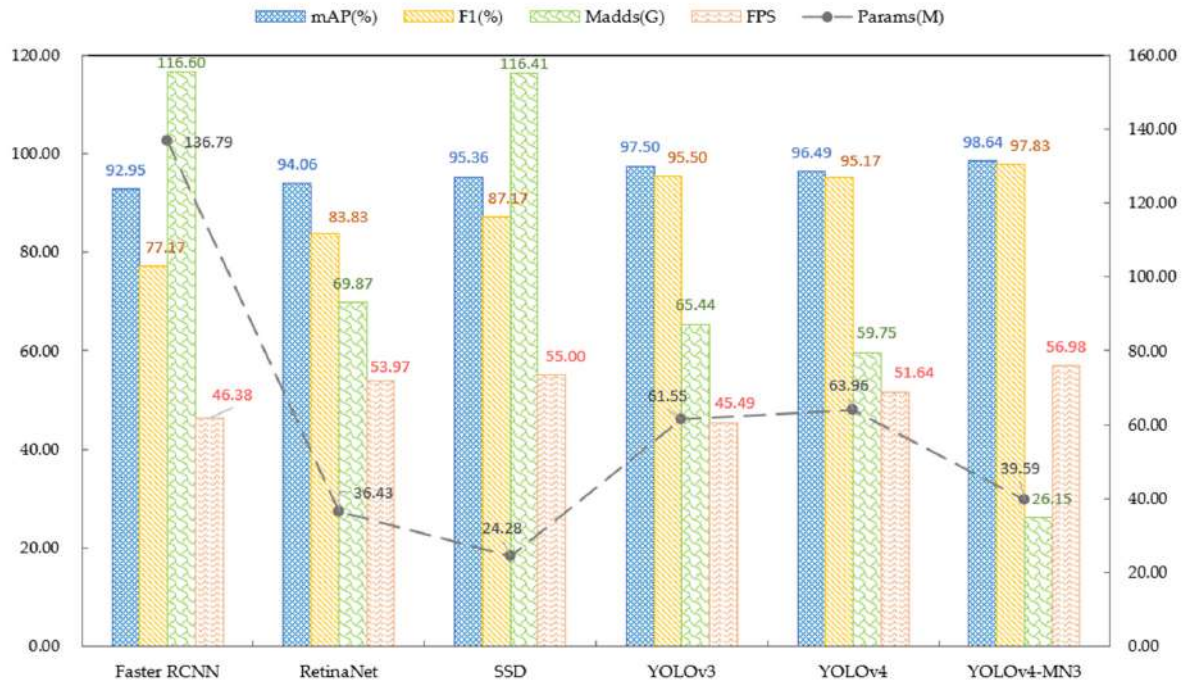
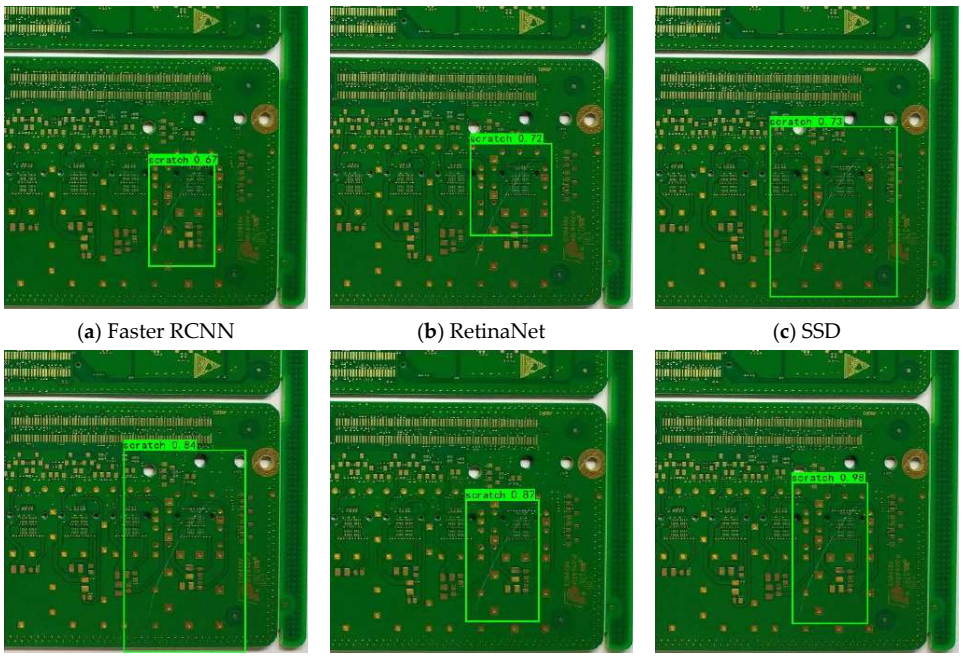


Figure 11. Performance of different SOTA models.

Figure 12 gives some detection instances obtained by YOLOv4-MN3 and the compared detectors. The detection instances given in Figure 12a,b show that Faster RCNN and RetinaNet only detected a part of the defect with low confidence. SSD and YOLOv3 defect the whole defect but also with low confidence, as shown in Figure 12c,d. Figure 12e,f indicates that both YOLOv4 and YOLOv4-MN3 could detect the whole defect. However, YOLOv4-MN3 achieved the highest confidence 0.98, while the confidence of YOLOv4 was only 0.87, in this instance.

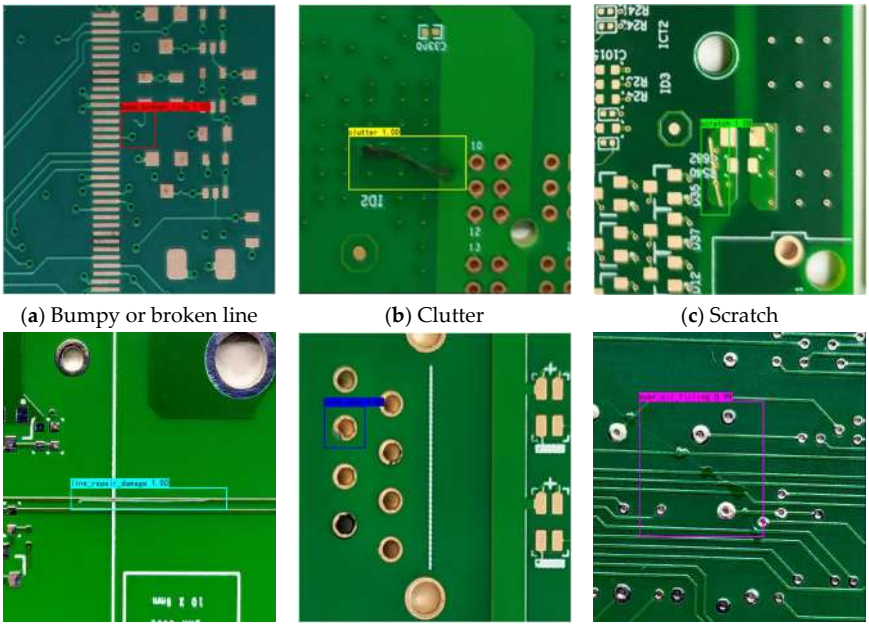


(**a**) Faster RCNN (**b**) RetinaNet (**c**) SSD

(**d**) YOLOv3 (**e**) YOLOv4 (**f**) YOLOv4-MN3

Figure 12. Detection instance obtained by different detectors.

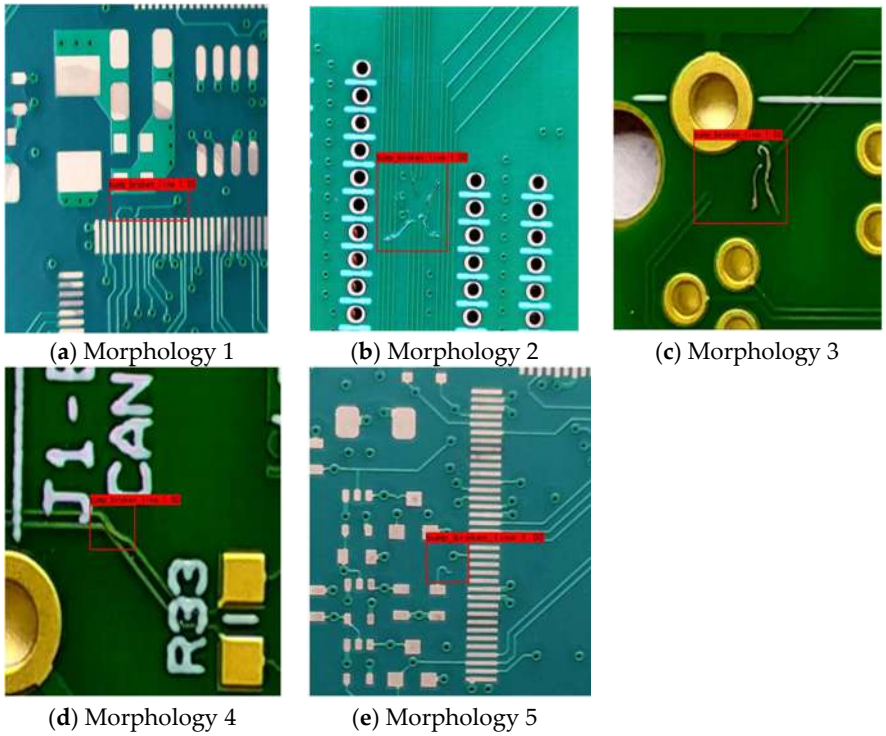
Figure 13 shows some detection instances for the aforementioned six categories of defects. Figure 14 gives an instance of detecting "bumpy or broken line" defect category, with its five different morphologies. The instances detected with high confidence, given in Figures 13 and 14, also indicate that the proposed YOLOv4-MN3 can adapt to different categories of surface defect, and it can handle the difficult problem of a diversity of defect morphologies.



(**a**) Bumpy or broken line (**b**) Clutter (**c**) Scratch

(**d**) Line repair damage (**e**) Hole loss (**f**) Over oil-filling

Figure 13. Detection instance of different defects obtained by YOLOv4-MN3.



(**a**) Morphology 1 (**b**) Morphology 2 (**c**) Morphology 3

(**d**) Morphology 4 (**e**) Morphology 5

Figure 14. Detection instance of bumpy or broken line with different morphologies obtained by YOLOv4-MN3.

5. Conclusions

In order to improve the performance of PCB surface defect detection, a DL-based detector, YOLOv4-MN3, was proposed. YOLOv4-MN3 replaces the CSPDarknet53 backbone network of YOLOv4 with a lightweight one, MobileNetV3, and the original activation function in the neck/prediction network was optimized. To verify the efficiency and the effectiveness the proposed detector, a customized dataset was built using a specially designed image-acquisition device. Experimental results on the customized dataset showed that YOLOv4-MN3 achieved the highest detection accuracy, with 98.64% *mAP* and 97.83% F_1, the fastest detection speed, at 56.98 *FPS*, and the lowest *Madds* as compared with the SOTA models. Generalization experiments of YOLOv4-MN3 indicated that it can adapt to different categories of surface defect and handle the difficult problem of defect morphology diversity, which has promising application prospects.

Although the detector proposed in this paper achieved good results, there are some problems to explore. Firstly, PCB surfaces' backgrounds, defect categories and morphology diversity, in real industry, are complex and diverse. Thus, it is necessary to update PCB surface defect samples continuously and facilitate the training of practicability-oriented DL-based models. Secondly, YOLOv4-MN3, as a supervised learning method, requires a large amount of manually labeled samples, and exploring semi-supervised or unsupervised mechanism for this problem is a worthy attempt.

Author Contributions: Conceptualization, X.L. and S.L.; methodology, X.L. and S.L.; software, X.L. and D.L.; validation, X.L. and Y.L.; formal analysis, X.L. and Z.Z.; investigation, X.L. and D.L.; re-sources, X.L. and D.L.; data curation, C.J. and Y.L.; writing—original draft preparation, X.L. and S.L.; writing—review and editing, X.L. and S.L.; visualization, X.L.; supervision, S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Natural Science Foundation of Guangdong, China with grant number 2021A1515012395.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: The authors wish to thank the editor and reviewers for their suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dai, W.; Mujeeb, A.; Erdt, M.; Sourin, A. Soldering defect detection in automatic optical inspection. *Adv. Eng. Inform.* **2020**, *43*, 101004.
2. Ebayyeh, A.; Mousavi, A. A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry. *IEEE Access.* **2020**, *8*, 183192–183271. [CrossRef]
3. Wang, W.C.; Chen, L.B.; Chang, W.J.; Chen, S.L.; Li, S.M. A machine vision based automatic optical inspection system for measuring drilling quality of printed circuit boards. *IEEE Access.* **2017**, *5*, 10817–10833. [CrossRef]
4. Gaidhane, V.H.; Hote, Y.V.; Singh, V. An efficient similarity measure approach for PCB surface defect detection. *Pattern Anal. Appl.* **2018**, *21*, 277–289. [CrossRef]
5. Yuk, E.H.; Park, S.H.; Park, C.-S.; Baek, J.-G. Feature-Learning-Based Printed Circuit Board Inspection via Speeded-Up Robust Features and Random Forest. *Appl. Sci.* **2018**, *8*, 932. [CrossRef]
6. Fonseka, C.; Jayasinghe, J. Implementation of an automatic optical inspection system for solder quality classification of THT solder joints. *IEEE Trans. Compon. Packag. Manufact. Tech.* **2018**, *9*, 353–366. [CrossRef]
7. Liu, Z.; Qu, B. Machine vision based online detection of PCB defect. *Microprocess. Microsyst.* **2021**, *82*, 103807. [CrossRef]
8. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *arXiv preprint* **2019**, arXiv:1905.05055. Available online: https://arxiv.org/abs/1905.05055 (accessed on 16 May 2019).
9. Rida, I.; Al-Maadeed, N.; Al-Maadeed, S.; Bakshi, S. A comprehensive overview of feature representation for biometric recognition. *Multimed. Tools Appl.* **2020**, *79*, 4867–4890. [CrossRef]

10. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
11. Girshick, R. Fast R-CNN. *arXiv preprints* **2015**, arXiv:1504.08083. Available online: https://arxiv.org/abs/1504.08083 (accessed on 30 April 2015).
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
13. Sharma, V.; Mir, R.N. A comprehensive and systematic look up into deep learning based object detection techniques: A review. *Comput. Sci. Rev.* **2020**, *38*, 100301. [CrossRef]
14. Kou, X.; Liu, S.; Cheng, K.; Qian, Y. Development of a YOLO-V3-based model for detecting defects on steel strip surface. *Measurement* **2021**, *182*, 109454. [CrossRef]
15. Li, D.; Xie, Q.; Gong, X.; Yu, Z.; Xu, J.; Sun, Y.; Wang, J. Automatic defect detection of metro tunnel surfaces using a vision-based inspection system. *Adv. Eng. Inform.* **2021**, *47*, 101206.
16. Shu, Y.F.; Li, B.; Li, X.; Xiong, C.; Cao, S.; Wen, X.Y. Deep learning-based fast recognition of commutator surface defects. *Measurement* **2021**, *178*, 109324.
17. Ding, R.; Dai, L.; Li, G.; Liu, H. TDD-net: A tiny defect detection network for printed circuit boards. *CAAI Trans. Intell. Technol.* **2019**, *4*, 110–116. [CrossRef]
18. Hu, B.; Wang, J. Detection of PCB Surface Defects with Improved Faster-RCNN and Feature Pyramid Network. *IEEE Access* **2020**, *8*, 108335–108345. [CrossRef]
19. Zhang, Y.; Xie, F.; Huang, L.; Shi, J.; Yang, J.; Li, Z. A Lightweight One-Stage Defect Detection Network for Small Object Based on Dual Attention Mechanism and PAFPN. *Front. Physics.* **2021**, *9*, 491. [CrossRef]
20. Bochkovskiy, A.; Wang, C.Y.; Liao, H. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint* **2020**, arXiv:2004.10934. Available online: https://arxiv.org/abs/2004.10934 (accessed on 23 April 2020).
21. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint* **2018**, arXiv:1804.02767. Available online: https://arxiv.org/abs/1804.02767 (accessed on 8 April 2018).
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal. Mach. Intell.* **2014**, *37*, 1904–1916. [CrossRef]
23. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768. Available online: https://ieeexplore.ieee.org/document/8579011 (accessed on 17 December 2018).
24. Wang, C.Y.; Liao, H.; Wu, Y.H.; Chen, P.Y.; Yeh, I.H. CSPNet: A New Backbone that can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 1571–1580.
25. Howard, A.; Sandler, M.; Chen, B. Searching for mobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324.
26. Misra, D. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint* **2019**, arXiv:1908.08681. Available online: https://arxiv.org/abs/1908.08681 (accessed on 23 August 2019).
27. Guo, F.; Qian, Y.; Shi, Y. Real-time railroad track components inspection based on the improved yolov4 framework. *Autom. Constr.* **2021**, *125*, 103596. [CrossRef]
28. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for activation functions. *arXiv preprint* **2017**, arXiv:1710.05941. Available online: https://arxiv.org/abs/1710.05941 (accessed on 16 October 2017).
29. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. *J. Mach. Learn. Res.* **2011**, *15*, 315–323. Available online: https://www.researchgate.net/publication/215616967 (accessed on 14 January 2010).
30. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Proc. IEEE Int. Conf. Comput. Vision* **2015**, *1*, 1026–1034.
31. Eger, S.; Youssef, P.; Gurevych, I. Is it Time to Swish? Comparing Deep Learning Activation Functions across NLP tasks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 4415–4424.
32. Zhao, J.; Zhang, X.; Yan, J.; Qiu, X.; Yao, X.; Tian, Y.; Zhu, Y.; Cao, W. A Wheat Spike Detection Method in UAV Images Based on Improved YOLOv5. *Remote Sens.* **2021**, *13*, 3095. [CrossRef]
33. Kulkarni, U.; Meena, S.M.; Gurlahosur, S.V.; Bhogar, G. Quantization friendly mobilenet (qf-mobilenet) architecture for vision based applications on embedded platforms. *Neural Netw.* **2021**, *136*, 28–39. [CrossRef]

中国计算机学会会刊
中国科技核心期刊
中 文 核 心 期 刊

ISSN 1007-130X
CODEN JGKEF6

计算机工程与科学

Jisuanji Gongcheng yu Kexue
Computer
Engineering & Science

2
2025

第47卷 第2期 2025年2月
Vol.47 No.2 Feb. 2025

ISSN 1007-130X
02>
9 771007 130250

国防科技大学计算机学院 主办
《计算机工程与科学》杂志社 出版
中国 长沙 Changsha China

计算机工程与科学
(月刊)
2025 年第 2 期
总 362 期
CN 43-1258/TP
ISSN 1007-130X

目 次

出版日期:2025 年 2 月
本刊电话:(0731)87002567
邮 箱:jsjgcykx@vip.163.com
杂志主页:joces.nudt.edu.cn

期刊基本参数:CN43-1258/TP * 1973 * m * A4 * 192 * zh * P * ¥58.00 * 3000 * 19 * 2025-02

CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 47 卷第 2 期 2025 年 2 月
Vol. 47, No. 2, Feb. 2025

文章编号:1007-130X(2025)02-0276-12

PCB 表面缺陷数据集与基于 YOLOv5s-P6SE 的检测*

梁泰然[1],蒋诗新[2,3],李泉洲[2,3],欧阳斌[1],吕盛坪[1]

(1. 华南农业大学工程学院,广东 广州 510642;
2. 中国赛宝实验室(工业和信息化部电子第五研究所),广东 广州 511370;
3. 工业装备质量大数据工业和信息化部重点实验室,广东 广州 511370)

摘　要:针对 PCB 生产中表面缺陷检测的需求,结合车间实际制定一个包含 11 种类别的缺陷分类标准,采集真实 PCB 表面缺陷图像,最终构建一个包含 3 239 幅图像 4 672 个缺陷目标的数据集 Dataset_PCBSD。基于 YOLOv5s 改进得到一种新的 PCB 表面缺陷检测模型 YOLOv5s-P6SE。为提高检测精度,在 YOLOv5s 中增加用于检测特大目标的 P6 检测层,引入了 SE 注意力模块和柔性非极大抑制后处理。实验结果显示,相较于基准模型 YOLOv5s,YOLOv5s-P6SE 在均值平均精度上提升了 5.5%。同时,YOLOv5s-P6SE 在 *mAP* 和模型大小上均优于 Faster R-CNN、SSD、PCB 缺陷检测模型 YOLOv4-MN3 以及 DETR 模型 RT-DETR-L,且在平衡 *mAP* 和模型大小方面优于 YOLOv8s。

关键词:印制电路板;表面缺陷检测;YOLOv5s-P6SE;SE 注意力模块;柔性非极大抑制

中图分类号:TP391.41　　**文献标志码**:A

doi:10.3969/j.issn.1007-130X.2025.02.010

PCB surface defect dataset and detection based on YOLOv5s-P6SE

LIANG Tairan[1], JIANG Shixin[2,3], LI Quanzhou[2,3], OUYANG Bin[1], LÜ Shengping[1]

(1. College of Engineering, South China Agricultural University, Guangzhou 510642;
2. CEPREI, Guangzhou 511370;
3. Key Laboratory of Industrial Equipment Quality Big Data, Guangzhou 511370, China)

Abstract: To address the demand for surface defect detection in PCB production, a defect classification standard encompassing 11 categories was established based on actual workshop conditions, images of real PCB surface defects were collected, and finally a dataset named Dataset_PCBSD was constructed, containing 3 239 images with 4 672 defective objects. A new PCB surface defect detection model, YOLOv5s-P6SE, was developed based on improvements to YOLOv5s. To enhance detection accuracy, a P6 detection layer for detecting extremely large objects was added to YOLOv5s, along with the introduction of the SE attention module and soft non-maximum suppression post-processing. Experimental results show that YOLOv5s-P6SE achieves a 5.5% improvement in mean average precision (*mAP*) compared to the baseline model YOLOv5s. Additionally, YOLOv5s-P6SE outperforms Faster R-CNN, SSD, the PCB defect detection model YOLOv4-MN3, and the DETR model RT-DETR-L in terms of both *mAP* and model size. It also excels in balancing *mAP* and model size compared to YOLOv8s.

Key words: printed circuit board; surface defect detection; YOLOv5s-P6SE; SE attention module; soft non-maximum suppression

* **收稿日期**:2023-10-07;**修回日期**:2023-12-27
基金项目:广东省自然科学基金(2021A1515012395)
通信作者:吕盛坪(lvshengping@scau.edu.cn)
通信地址:510642 广东省广州市华南农业大学工程学院
Address: College of Engineering, South China Agricultural University, Guangzhou 510642, Guangdong, P. R. China

1 引言

印制电路板PCB(Printed Circuit Board)作为现代电子产品不可或缺的组成部件,在各个领域得到了广泛应用。PCB的质量对于电子产品的性能至关重要。然而,在PCB的生产制造过程中,由于复杂工艺要求、设备运行状况以及人为和环境因素等多重影响,PCB在不同工序中可能出现各种表面缺陷,如开路、鼠咬、短路等[1],涉及到基材、铜面、线路和孔等不同部位。这些缺陷不仅会影响美观,还有可能降低PCB的性能甚至导致整板报废。因此,表面缺陷检测成为印刷电路板生产过程中质量控制的基本要求。

PCB表面缺陷检测方法主要包括人工目视法、自动机器视觉检测AVI(Automated Visual Inspection)和自动光学检测法AOI(Automated Optical Inspection)等[2]。AVI和AOI作为无损非接触式的缺陷在线检测技术,利用传统的图像处理技术和机器学习技术代替人工视觉进行检测、分析、判断和决策,可大幅度提高生产效率和自动化程度。但是,通过企业调研发现,国内外相关设备仍存在漏检率和召回率高、需投入大量人力成本进行目视复检等问题。

随着深度学习技术的快速发展,以深度学习为核心检测模型的AOI/AVI得到了深入研究[3]。HUANG等人[4]构建了一个PCB表面缺陷数据集PKU-Market-PCB,该数据集基于10幅独立的PCB图像人工合成了693个表面缺陷,包括漏孔、鼠咬、开路、短路、毛刺和伪铜共6类。在此数据集基础上,DING等人[5]通过几何与光学图像变换将该数据集扩增到10 668幅图像。相关研究人员基于该数据集提出了不同的检测模型[6-12]。

HU等人[13]构建了一个包含开路、短路、鼠咬、毛刺、针孔和焊球6类PCB表面缺陷的数据集,并基于Faster R-CNN(Faster Region-based Convolutional Neural Network)提出了一个有效的二阶段PCB表面缺陷检测模型。ADIBHATLA等人[14]构建了一个包含23 000幅图像的PCB表面缺陷二分类数据集,并使用YOLOv5l模型进行训练检测,取得了99.74%的精度。PHAM等人[15]构建了一个包含22 909幅图像的PCB二分类数据集,并提出了半监督目标检测模型PCB_SS(defect detection in Printed Circuit Boards using Semi-Supervised learning)。本课题组前期构建了一个包含线路不良、杂物、划伤、补线不良、孔损、补油超标6类表面缺陷的数据集,并提出一个改进的YOLOv4检测模型[16]。

然而,目前广泛使用的PKU-Market-PCB数据集中的缺陷类别和样本数量有限,缺陷图像仅6类693幅,需要扩展缺陷类型,并增加缺陷图像数量。同时,该数据集相关缺陷主要是通过人工合成,缺陷类内差异极小,与生产中出现的真实缺陷存在很大差异。此外,该数据集在划分前对原始图像进行了增强处理,导致验证(或测试)集和训练集中的样本高度相似,这使得验证深度学习模型的检测准确性和泛化性能更具挑战。其他类数据集也存在缺陷类别少[13-16]、数据集划分在扩增之前[16]等不足。

为此,本文将结合企业生产实际制定PCB表面缺陷分类标准,面向PCB检测环节需求,构建PCB表面缺陷图像数据集。同时基于YOLOv5s设计针对PCB表面缺陷检测的改进模型YOLOv5s-P6SE。

2 PCB表面缺陷数据集构建

2.1 表面缺陷类别划分

PCB表面缺陷种类多样,通过调研相关企业总结出表面缺陷类型有80多种。然而,这些分类方法主要基于人工经验和理解,缺乏从深度学习特征提取的角度进行科学定义和系统划分。因此,需要对缺陷的共性特征进行抽象提炼,重新从深度学习的视角定义缺陷类型,并制定相应的分类标准。

结合企业生产实际,PCB表面缺陷主要发生在基材、铜面、线路和孔上。为此,本文联合PCB生产企业,将在基材和铜面上出现的表面缺陷按照形状和尺寸划分为点状缺陷PD(Point Defect)、线状缺陷LD(Line Defect)和块状缺陷BD(Block Defect)3类。在线路上出现的表面缺陷,按照线路铜料、位置和颜色要求划分为线路余料WMS(Wiring Material Surplus)、线路缺料WMD(Wiring Material Deficiency)、线路断路OC(Open Circuit)、线路短路SH(Short)、线路歪斜WD(Wiring Deflection)和线路污染WC(Wiring Contamination)6类。孔上出现的缺陷按照孔的位置和外观要求划分为孔破HB(Hole Breakout)和孔异物HFO(Hole Foreign Object)2类。

上述划分较现有相关缺陷划分更全面系统地涵盖了PCB常见表面缺陷,相关企业反馈该11类缺陷占企业总表面缺陷的95%以上。同时,上述缺陷划分融入了形状、尺寸、颜色和部分位置语义,更利于深度学习提取相关特征。

2.2 样本获取与数据集构建

本文所使用的图像样本来自广州某PCB生产企业AOI设备所拍摄的带缺陷的外层蚀刻图像，其中像素分别有108×108和144×144这2种，最终得到3 239幅PCB表面缺陷样本图像，并通过LabelImg软件对样本进行标注，构建一个具有4 672个目标的PCB表面缺陷图像数据集Dataset_PCBSD。该数据集的各类目标数量和缺陷样本图像示例分别如表1和图1所示。

Table 1 Quantity of various defect objects

表1 各类缺陷目标数量

序号	目标类别	目标数量/个	序号	目标类别	目标数量/个
0	PD	1 003	6	OC	196
1	LD	367	7	WD	42
2	BD	945	8	WMD	693
3	HB	77	9	WMS	639
4	HFO	193	10	WC	252
5	SH	265	总计		4 672

另外，按照COCO(Common Objects in COntext)数据集中对大型(大于96×96像素)、中型(大于32×32像素且小于96×96像素)和小型(小于32×32像素)目标的定义[17]，本文对Dataset_PCBSD中各类缺陷目标大小进行了统计，结果如表2和图2所示。在Dataset_PCBSD中，小型目标占到了总数的52.5%，并且主要集中在PD、WMD和WMS上，而中型、大型目标也几乎占据了总目标数量的一半。因此，如何有效检测到各个大小的目标成为在Dataset_PCBSD上进行研究的关键。

3 基于YOLOv5s-P6SE的检测模型

3.1 YOLOv5

YOLOv5是一种一阶段目标检测模型，在YOLOv4[18]基础上引入了新的改进。在模型训练阶段，YOLOv5采用了Mosaic数据增强、自适应锚框计算和自适应图像缩放等多种方法。在模型的主干中，YOLOv5整合了其他检测模型的创新

Table 2 Statistical analysis of various defect objects sizes

表2 各类缺陷目标大小统计

序号	目标类别	小型目标/个	中型目标/个	大型目标/个	总计/个
0	PD	987	16	0	1 003
1	LD	158	201	8	367
2	BD	23	717	205	945
3	HB	12	60	5	77
4	HFO	7	170	16	193
5	SH	89	175	1	265
6	OC	139	57	0	196
7	WD	8	33	1	42
8	WMD	572	121	0	693
9	WMS	386	246	7	639
10	WC	72	169	11	252
总计		2 453	1 965	254	4 672

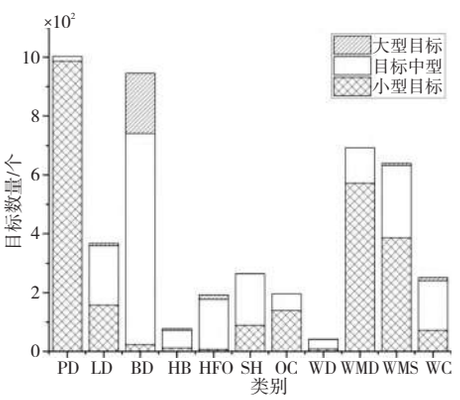


Figure 2 Distribution of various defect objects sizes

图2 各类缺陷目标大小分布情况

方法，如Focus结构、SPPF(Spatial Pyramid Pooling Fast)结构和CSP(Cross Stage Partial)结构。在模型的颈部中，YOLOv5引入了特征金字塔网络FPN(Feature Pyramid Network)[19]和路径聚合网络PAN(Path Aggregation Network)[20]，通过多尺度特征融合、特征增强与下采样以及目标定位和尺度匹配等操作，YOLOv5能够更好地检测不同大小和尺度的目标物体。在模型的头部输出层，YOLOv5改进了损失函数CIoU Loss[21]。上

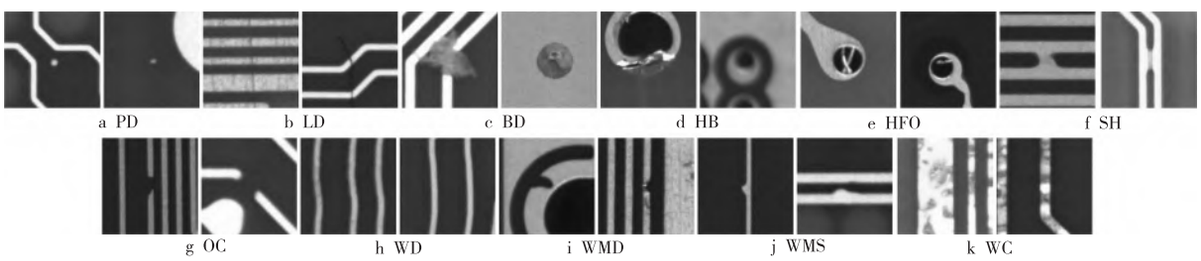


Figure 1 Image instances of various classes of defects

图1 各类缺陷图像实例

述改进提升了YOLOv5的检测速度和准确性,已在目标检测领域被广泛采用。

3.2 改进多尺度检测结构

YOLOv5s使用P3、P4和P5这3个不同尺度的特征层进行检测,分别对原图像的8,16和32倍下采样图像(小型目标、中型目标、大型目标)进行检测。而Dataset_PCBSD中存在一些特大目标,如图3所示,其大小几乎覆盖了整个图像。为解决数据集中的特大目标(主要缺陷类型为BD)检测精度低或难以检测的问题,本文设计了四尺度检测模型YOLOv5s-P6,增加了P6特征层用于检测特大目标,即对输入图像进行64倍下采样。

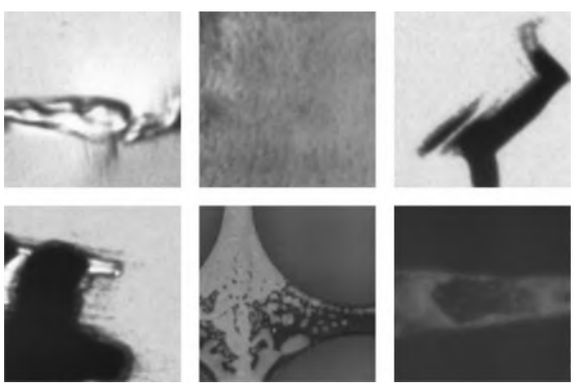


Figure 3 Image instances of x-large defect objects

图3 特大缺陷目标图像实例

四尺度检测模型YOLOv5s-P6结构如图4所示。相较于原始模型,改进如下:模型主干增加一层输出尺寸为4×4×512的残差结构$\boldsymbol{B}_6$;模型颈部增加1个尺度与$\boldsymbol{B}_6$相同的特征图$\boldsymbol{P}_6$进行特征融合,实现对4个尺度的特征图进行检测。

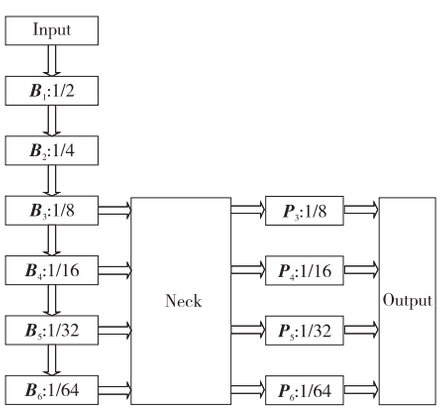


Figure 4 Structure of YOLOv5s-P6 model

图4 YOLOv5s-P6模型结构

首先,将256×256×3的图像通过模型主干的多阶段残差结构生成多种尺度的特征图,然后通过模型颈部融合4幅不同尺度的特征图,分别是32×32×128,16×16×256,8×8×384和4×4×512。最后,对FPN结构输出的特征图进行检测,即32×32×128的特征图用于检测小目标,16×16×256的特征图用于检测中目标,8×8×384的特征图用于检测大目标,4×4×512的特征图用于检测特大目标。

相较于原始的YOLOv5s模型,改进后的YOLOv5s-P6模型结构更深、更复杂,能够更好地捕捉特大目标的细节和形状特征,从而提高检测的准确性。改进后的YOLOv5s-P6在模型主干和头部使用了更多的CSP模块[22],在特征提取的过程中引入了更多的非线性变换,从而有助于更好地捕捉特定目标的语义信息。

3.3 增加SE注意力模块

在实际的卷积过程中,不同通道特征信息的重要性是有差异的。为了解决卷积过程中特征图在不同通道重要性不同的问题,本文在模型颈部的PAN结构中引入了SE(Squeeze-and-Excitation)注意力模块[23]。SE注意力模块结构如图5所示,其由压缩(Squeeze)部分和激励(Excitation)部分组成。在Squeeze部分,通过全局平均池化(Avgpool)操作,将高为H、宽为W和通道数为C的特征图$\boldsymbol{F}$转化为$1\times1\times C$的权重矩阵,提取每个通道的全局特征。在Excitation部分,通过使用2个全连接层FC(Fully Connected layer)与ReLU(Rectified Linear Unit)激活函数组成多层感知机MLP(Multi-Layer Perception)对通道进行加权,再利用Sigmoid激活函数进行归一化处理得到不同通道的权重。最后,将加权后得到的$1\times1\times C$权重矩阵按通道C与输入特征$H\times W\times C$相乘(图5中以⊗表示),从而将计算得到的注意力权重分配到每个通道的特征图上,得到输出特征$\boldsymbol{F}'$,实现了通道中的特征增强。

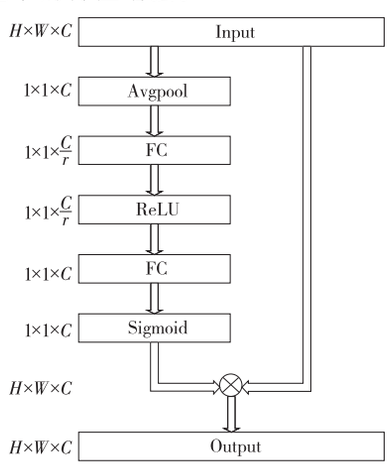


Figure 5 Structure of SE attention module

图5 SE注意力模块结构

上述过程可用式(1)和式(2)表示：

$$\boldsymbol{F}'=\frac{1}{1+\mathrm{e}^{-W_2(\max(0,W_1(\boldsymbol{x})))}}\otimes\boldsymbol{F} \tag{1}$$

$$x_c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}u_c(i,j) \tag{2}$$

其中，$\boldsymbol{F}\in\mathbf{R}^{H\times W\times C}$，$u_c(i,j)$ 表示输入特征图 $\boldsymbol{F}$ 的第 c 个通道在 (i,j) 坐标的像素值，$c\in\{1,2,\cdots,C\}$，$W_1(\cdot)$、$W_2(\cdot)$ 分别表示第一和第二层的全连接操作，$\boldsymbol{x}$ 为从输入特征中提取的每个通道的像素平均值组成的序列。

本文引入 SE 注意力模块后得到的最终模型 YOLOv5s-P6SE 的结构如图 6 所示。从图 6 中可以看出，SE 注意力模块被插入在模型颈部 PAN 结构的低阶特征图与高阶特征图转换过程中，并将每个等级中通过 SE 注意力模块的特征图输出到模型头部中进行检测。特征从模型主干输出到头部的过程可用式(3)～式(5)表示：

$$\boldsymbol{F}'_i=\begin{cases}C'_3(UPSAMPLE(CBS(\boldsymbol{F}_i))+\boldsymbol{B}_i),\\ \qquad i=3,4,5\\ C'_3(UPSAMPLE(CBS(\boldsymbol{B}_i))+\boldsymbol{B}_{i-1}),i=6\end{cases} \tag{3}$$

$$\boldsymbol{F}_i=\begin{cases}CBS(\boldsymbol{F}'_i),i=4,5\\ CBS(\boldsymbol{B}_i),i=6\end{cases} \tag{4}$$

$$\boldsymbol{P}_i=\begin{cases}SE(C'_3(DOWNSAMPLE(\boldsymbol{P}_{i-1})+\boldsymbol{F}_i)),\\ \qquad i=4,5,6\\ SE(\boldsymbol{F}'_i),i=3\end{cases} \tag{5}$$

其中，$\boldsymbol{B}_i$ 为图 6 模型主干第 i 阶段输出的特征图，$\boldsymbol{F}'_i$ 为图 6 中模型颈部自下而上第 i 阶段输出的特

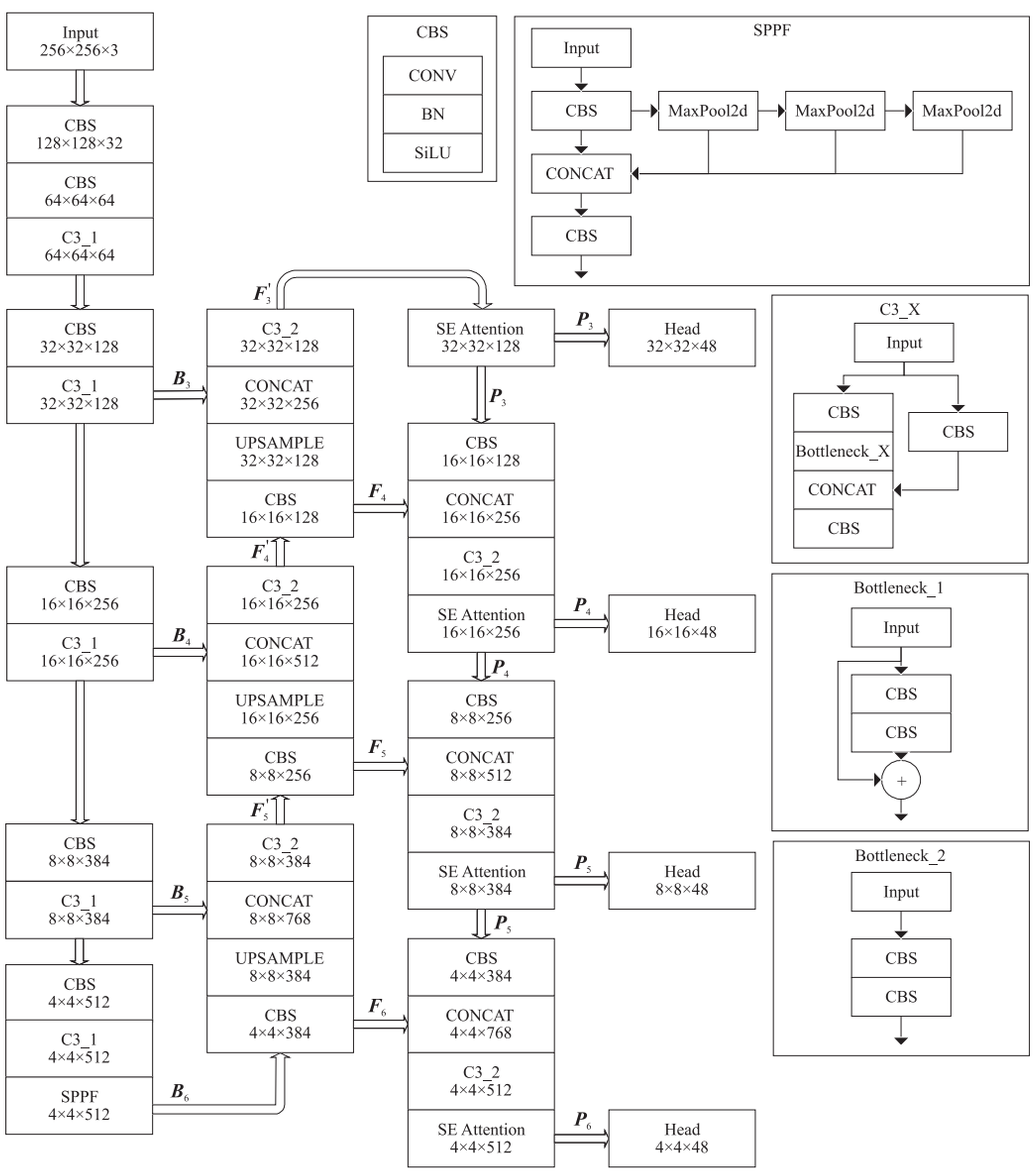


Figure 6 Structure of YOLOv5s-P6SE model

图 6 YOLOv5s-P6SE 模型结构

征图，$\boldsymbol{F}_i$ 为图6中模型颈部从左向右传递的中间特征图，$\boldsymbol{P}_i$ 为图6模型颈部自上而下输出的特征图，$C_3'(\cdot)$ 为非shortcut瓶颈卷积模块，CBS 为由普通卷积、正则化层、SiLU(Sigmoid Linear Unit)激活函数组合而成的卷积模块，$UPSAMPLE(\cdot)$、$DOWNSAMPLE(\cdot)$ 分别为上采样与下采样操作，$SE(\cdot)$ 为3.3节中提及的SE注意力机制。

由于PAN结构主要用于解决目标检测中不同尺度特征的融合问题，在PAN结构中插入SE注意力模块，能够进一步优化特征融合效果。SE注意力模块能够自适应地学习每个通道的权重，使得其在特征融合过程中，对重要的特征通道有更高的关注度，从而提升融合后特征的表达能力，最终提升模型整体的检测性能。

3.4 柔性非极大抑制后处理

Dataset_PCBSD中存在一些目标框重叠的图像，如图7所示。YOLOv5s默认使用的非极大抑制NMS(Non-Maximum Suppression)算法是将检测框按得分排序，然后保留得分最高的检测框，同时删除与该检测框重叠面积大于一定比例的其他检测框。NMS检测框得分的计算方式如式(6)所示：

$$s_i'=\begin{cases}s_i, I_{\mathrm{IoU}}(A',B')<N_{\mathrm{t}}\\ 0, I_{\mathrm{IoU}}(A',B')\geqslant N_{\mathrm{t}}\end{cases} \tag{6}$$

其中，s_i 为置信度得分；i 为除了得分最大的 A' 目标框以外，剩余目标框以置信度得分从高到低排序的序号；N_{t} 为置信度阈值。

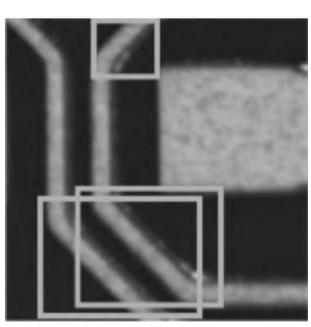


Figure 7 Image instances with overlapped detection objects

图7 检测目标重叠图像实例

NMS算法会将相邻检测框的分数强制归零，这可能会导致重叠目标框被删除，出现漏检，在一定程度上降低了目标检测精度。为解决该问题，本文在后处理阶段采用柔性非极大抑制(Soft-NMS)[24]算法。与NMS不同，Soft-NMS不会直接将目标框的分数置为0，而是使用较低的分数来代替原始分数，从而在一定程度上保留了部分重叠的目标框。Soft-NMS算法中，检测框得分的计算方式如式(7)所示：

$$s_i'=\begin{cases}s_i, I_{\mathrm{IoU}}(A',B')<N_{\mathrm{t}}\\ s_i\exp\left[-\dfrac{I_{\mathrm{IoU}}(A',B)^2}{\sigma}\right], \\ \qquad I_{\mathrm{IoU}}(A',B')\geqslant N_{\mathrm{t}}\end{cases} \tag{7}$$

其中，σ 为高斯函数惩罚项系数。

通过式(7)将大于阈值 N_{t} 的检测分数衰减作为关于检测框 A' 重叠度的高斯函数惩罚项系数，因此当待处理的检测框 B' 远离检测框 A' 时，将不受影响。2个检测框的重叠度越高时，s_i 越小，降低检测框 B' 的得分可以避免被强制删除而造成漏检的情况，从而提高检测的均值平均精度。

4 实验和结果分析

4.1 评价指标

本文使用均值平均精度 mAP(mean Average Precision)、查准率 P(Precision)和查全率 R(Recall)评价模型的检测性能。其计算方式分别如式(8)～式(11)所示：

$$mAP=\frac{\sum_{i=1}^{C}AP_i}{C} \tag{8}$$

$$AP=\int_0^1 P(R)\,\mathrm{d}R \tag{9}$$

$$R=\frac{TP}{TP+FN} \tag{10}$$

$$P=\frac{TP}{TP+FP} \tag{11}$$

其中，AP_i 为第 i 类的精确度，计算方式见式(9)，TP 为真正例的数量，FN 为假反例的数量，FP 为假正例的数量。并且，本文使用 $mAP@0.5$ 作为均值平均精度，即当预测框与标注真实值(ground truth)的交并比大于或等于0.5时被归类为真正例。除此之外，评价指标还使用了模型大小，用于反映模型的可部署性能。

4.2 实验环境与参数设置

本文实验所用平台为Ubuntu 20.04-64位操作系统，Intel® Xeon® Gold 6242R处理器，NVIDIA® GeForce RTX™ 3090图形处理器，PyCharm编译软件，PyTorch深度学习框架为CUDA 11.2+cuDNN 8.0.4的并行计算框架。

实验数据分别使用Dataset_PCBSD验证模型改进的有效性、使用PKU-Market-PCB与Deep-PCB[25]2个公开PCB缺陷数据集验证模型的泛化

性,并且统一采用随机划分方法将数据集中每个类别的缺陷图像按照8∶2的比例划分成训练集与测试集。其中,Dataset_PCBSD包含训练集图像2 591幅与测试集图像648幅,PKU-Market-PCB包含训练集图像8 534幅与测试集图像2 134幅,DeepPCB包含训练集图像1 200幅与测试集图像300幅。

模型训练采用分布式并行计算,使用8个图形处理器进行训练。训练过程中,每次训练批量大小为64,训练轮数(*epoch*)为300。损失函数优化方式采用随机梯度下降,其中学习率调节方式为线性减少,初始学习率为0.01,最终学习率为0.000 1。

由于样本图像输入模型时需要通过缩放来统一图像的尺寸,而不同的尺寸大小对模型的检测结果影响不同,因此需要设置不同的图像输入尺寸进行实验以确定其大小。表3展示了在YOLOv5s-P6上输入从128×128×3到640×640×3共9种不同尺寸大小的图像所得的结果对比。结果显示,当输入尺寸为512×512×3时模型的查准率最高,而当输入尺寸为576×576×3时模型的查全率最高,但当输入尺寸为256×256×3时,模型的mAP最高达73.1%,并且查准率和查全率相对其他两者更均衡。因此模型输入尺寸大小设定为256×256×3。

Table 3 Detection results with different input sizes

表3 不同输入尺寸所得检测结果 %

输入尺寸	P	R	mAP
128×128×3	59.9	63.3	62.9
192×192×3	74.9	65.3	70.8
256×256×3	72.8	70.7	**73.1**
320×320×3	76.5	64.7	71.7
384×384×3	79.5	67.0	71.8
448×448×3	75.7	66.2	70.9
512×512×3	**77.2**	65.9	72.1
576×576×3	70.1	**71.5**	72.1
640×640×3	70.4	66.1	69.3

4.3 实验结果与分析

4.3.1 训练过程

图8展示了YOLOv5s-P6SE模型在训练过程中的各项损失值曲线。纵坐标表示损失值,横坐标表示训练轮数。其中,定位损失用于计算预测框和标定框之间的误差,置信度损失用于计算网络的置信度,分类损失用于计算锚框对应的分类是否正确。从图8中可以看出,经过约250轮的模型训练,YOLOv5s-P6SE模型的各项损失值不断减少,随后不再具有下降趋势,最终稳定,模型收敛。

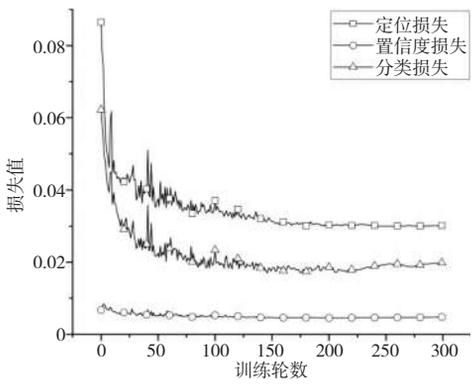


Figure 8 Loss curves of training process

图8 训练损失曲线

4.3.2 消融实验

本文基于Dataset_PCBSD,通过逐步增加所提出的改进方法开展消融实验,以验证各改进的有效性和所提出模型的性能。具体实验结果如表4所示。

Table 4 Results of ablation experiments

表4 消融实验结果 %

模型	改进			P	R	mAP
	P6	SE Attention	Soft-NMS			
模型1				73.9	65.2	70.8
模型2	√			72.8	70.7	73.1
模型3	√	√		80.5	66.8	73.3
模型4	√	√	√	81.4	66.4	76.3

从表4可以看出,每种改进都在一定程度上提升了模型的mAP。总体而言,YOLOv5s-P6SE(模型4)相对基准模型YOLOv5s(模型1)mAP提升了5.5%,查准率提升了7.5%,查全率提升了1.2%。图9展示了消融实验中模型1~模型4每个改进模型的P-R(Precision-Recall)曲线。从图中可以看到YOLOv5s-P6SE的P-R曲线更趋向于坐标轴的右上方,表明改进能够有效地提升模型的性能,满足了对PCB表面缺陷检测中对定位精确性和识别准确性的需求。

模型2在模型1的基础上增加了P6特征层用于检测特大目标。图10展示了模型1和模型2所得特大目标检测结果实例,可以看出模型2引入P6特征层后,在特大目标检测方面取得了更高的置信度,直观地验证了该改进的有效性。

模型3在模型2的PAN结构中引入了SE注意力模块,以增强各通道中的特征信息。图11展

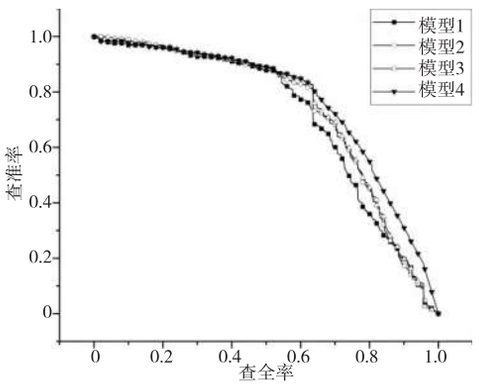


Figure 9 PR curves of different models

图 9 不同模型 PR 曲线

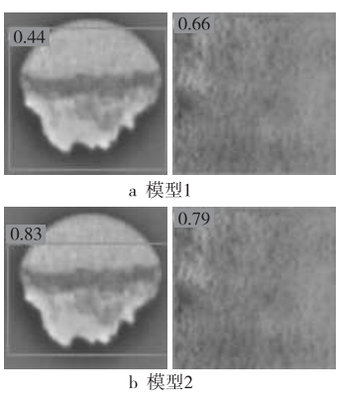


Figure 10 Instances of x-large-object detection results of model 1 and model 2

图 10 模型 1 与模型 2 所得特大目标检测结果示例

示了模型 2 与模型 3 的检测结果示例及其对应热力图。模型热力图用于呈现图像像素在模型中的权重分布,区域中灰度值越高权重越大。通过对比图 11a 和图 11b 可知,加入 SE 注意力模块后,模型提升了目标的置信度,并增强了缺陷在模型中的权重。从图 11 中可见加入 SE 注意力模块后缺陷检测更集中在真正的缺陷区域。综上所述,在模型中引入 SE 注意力模块能够有效提升性能。

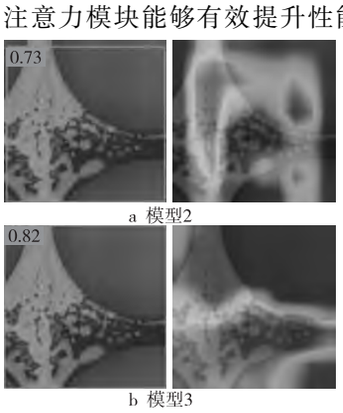


Figure 11 Instances of detection results and heatmaps of model 2 and model 3

图 11 模型 2 与模型 3 检测结果示例与其热力图

基于模型 3,模型 4 在后处理时采用 Soft-NMS 替代了传统的 NMS。图 12 展示了模型 3 与模型 4 在处理重叠目标样本时的检测结果实例。可以看到,在模型 3 中经过 NMS 过滤后模型 4 左下方存在缺陷的目标框被删除了。然而,在模型 4 中采用了 Soft-NMS 后,该目标框得以保留,从而进一步提升了模型 4 的 mAP。

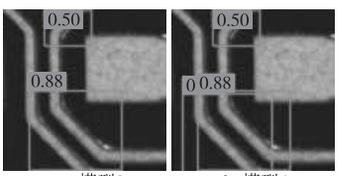


Figure 12 Instances of overlapped object detection results of model 3 and model 4

图 12 模型 3 与模型 4 的重叠目标检测结果示例

4.3.3 实验结果分析

表 5 展示了各类表面缺陷基于模型 4(YOLOv5s-P6SE)所得的均值平均精度,图 13 给出了均值平均精度与训练集目标分布情况。结合表 5 和图 13 可以得出,YOLOv5s-P6SE 在检测 HFO 这一类缺陷时 AP 高达 97.0%。其原因在于这类缺陷样本中的缺陷特征较为清晰,类间差异性较大且类内差异性较小。具体示例如图 14 所示,HFO 能够被准确检测,且具有较高置信度。

Table 5 *AP* of various defects using YOLOv5s-P6SE

表 5 使用 YOLOv5s-P6SE 检测出的各类缺陷 *AP* 值

序号	缺陷类型	AP/%	序号	缺陷类型	AP/%
0	PD	64.3	6	OC	76.0
1	LD	73.9	7	WD	77.7
2	BD	83.8	8	WMD	73.7
3	HB	83.6	9	WMS	76.0
4	HFO	**97.0**	10	WC	60.1
5	SH	73.3			

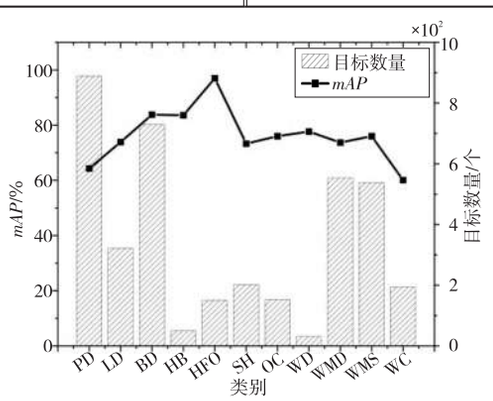


Figure 13 *AP* and numbers of training objects for various classes

图 13 各类别 *AP* 与训练目标数量

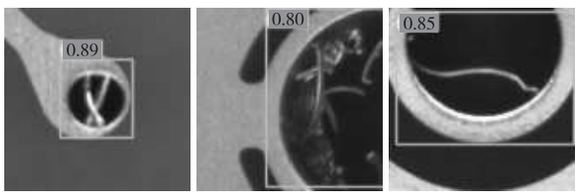
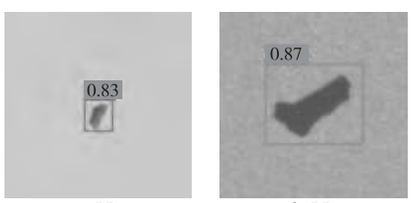


Figure 14 Instances of detection results of HFO

图 14 HFO 检测结果示例

然而,YOLOv5s-P6SE 在检测 DP 和 WC 这 2 类缺陷时 *AP* 值较低。原因是 PD 与 BD 之间相似性较高,它们的区别仅在于外形特征的大小。如图 15 所示,图 15a 显示了一个被误检为 BD 的 PD 缺陷,图 15b 是实际的 BD,这 2 类缺陷之间的区别几乎只在于外形特征的大小。

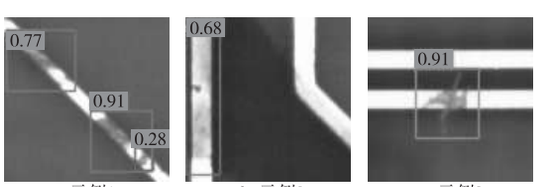


a PD b BD

Figure 15 Instances of detection results of PD and BD

图 15 PD 与 BD 的检测结果示例

WC 这一缺陷类型呈现较大类内差异。如图 16 所示,图 16a～图 16c 均展示了 WC 检测结果,可以看出,图 16a 有 WMD 检测框,表明模型将 WC 误检为 WMD,而图 16b 出现漏检,图 16c 将 WC 错误地归类为 WMS。



a 示例1 b 示例2 c 示例3

Figure 16 Instances of detection results of WC

图 16 WC 检测结果示例

总体来说,对于识别效果不佳的类别,可通过获取更多的样本来扩充数据集,以增强模型对这些类别的学习能力;同时,还需优化模型训练过程,提升其对类间相似性的区分度,从而进一步提升整体性能。

4.3.4 不同模型在 Dataset_PCBSD 上的对比

本文将 YOLOv5s-P6SE 与其他常见的目标检测模型以及本课题组前期提出的 PCB 缺陷检测模型 YOLOv4-MN3[16]在 Dataset_PCBSD 数据集上进行了对比。常见的目标检测模型包括二阶段模型 Faster R-CNN,一阶段模型 SSD(Single Shot multibox Detector)、YOLOv5s,最新的 YOLOv8s 与 DETR(DEtection TRansformer)模型 RT-DETR-L(Real-Time DEtection TRansformer-Large)[26]。其中 Faster R-CNN 模型使用了 VGG16(Visual Geometry Group 16)和 RES101(RESidual network-101)这 2 种模型作为主干进行了实验。对比实验结果如表 6 所示。

Table 6 Comparative results of multiple models

表 6 多模型对比结果

模型	*mAP*/%	模型大小/MiB
Faster R-CNN(VGG16)	63.46	1 094.1
Faster R-CNN(ResNet101)	63.14	360.8
SSD	70.55	95.7
YOLOv4-MN3	58.46	54.0
YOLOv5s	70.80	**13.7**
YOLOv8s	69.30	22.5
RT-DETR-L	64.90	66.2
YOLOv5s-P6SE(本文)	**76.30**	25.1

从表 6 中可以看出,与二阶段模型 Faster R-CNN、一阶段模型 SSD 和 YOLOv4-MN3 相比,YOLOv5s-P6SE 在 *mAP* 和模型大小方面都具有显著优势。与轻量化一阶段模型 YOLOv5s(基准模型)相比,虽然 YOLOv5s-P6SE 的模型大小增加了 11.4 MiB,但 *mAP* 提升了 5.5%,表现出更强的定位和识别能力。与最新 YOLOv8s 相比,虽然 YOLOv5s-P6SE 的模型大小增加了 2.6 MiB,但 *mAP* 提高了 7%。与新的 DETR 模型 RT-DETR-L 相比,YOLOv5s-P6SE 的 *mAP* 提升了 11.4%,模型大小少了 41.1MiB。

4.3.5 不同模型在公开 PCB 缺陷数据集上的对比

为验证 YOLOv5s-P6SE 模型的泛化性,本文分别将公开的 PCB 缺陷数据集 PKU-Market-PCB 与 DeepPCB[25]载入模型进行训练,并与其他 PCB 表面缺陷检测模型进行对比,结果如表 7 和表 8 所示。

Table 7 Results of different models on PKU-Market-PCB

表 7 不同模型在 PKU-Market-PCB 上的结果

模型来源	基准模型	*mAP*/%
文献[5](2019 年)	Faster R-CNN	98.90
文献[6](2022 年)	Faster R-CNN	98.91
文献[8](2023 年)	YOLOv5	97.40
文献[9](2023 年)	YOLOv5	95.30
文献[10](2023 年)	YOLOX	96.65
文献[12](2023 年)	YOLOv5	95.97
文献[27](2024 年)	YOLOv5	99.12
本文	YOLOv5	**99.20**

在训练时,针对数据集图像尺寸的大小将模型输入大小设置为 640×640,其他训练参数与本文模型训练参数相同。YOLOv5s-P6SE 在 PKU-Market-PCB 与 DeepPCB 上分别取得 99.20%和 99.10%的 mAP,模型得到的 PR 曲线分别如图 17 和图 18 所示。

Table 8 Results of different models on DeepPCB

表 8 不同模型在 DeepPCB 上的结果

模型来源	基准模型	mAP/%
文献[28](2022 年)	Faster R-CNN	98.83
文献[29](2023 年)	YOLOv4	98.90
文献[30](2023 年)	YOLOv5	98.30
文献[27](2024 年)	YOLOv5	97.63
本文	YOLOv5	**99.10**

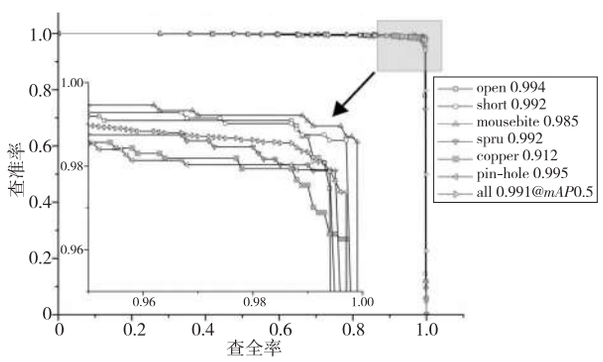


Figure 17 PR curve of YOLOv5s-P6SE training on PKU-Market-PCB

图 17 YOLOv5s-P6SE 在 PKU-Market-PCB 上训练所得的 PR 曲线

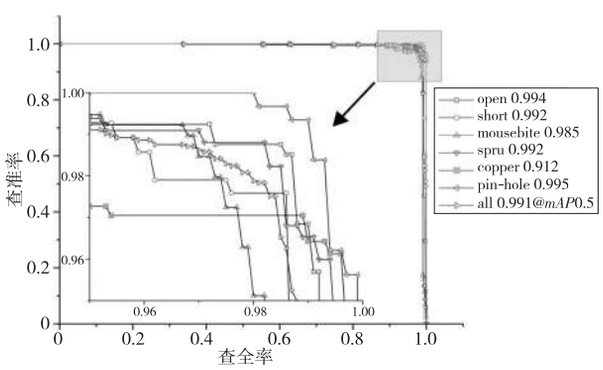


Figure 18 PR curve of YOLOv5s-P6SE training on DeepPCB

图 18 YOLOv5s-P6SE 在 DeepPCB 上训练所得的 PR 曲线

YOLOv5s-P6SE 在 2 个数据集上都取得了良好的 mAP 结果,并且其表现均优于表 7 与表 8 中对比的其他模型,表明 YOLOv5s-P6SE 具有较好的泛化能力,能够有效地进行 PCB 表面缺陷检测。

综上所述,YOLOv5s-P6SE 在综合性能和模型大小之间取得了平衡,且在不同的 PCB 缺陷数据集上均取得了良好的结果,可被认为是一种性能优越的 PCB 表面缺陷检测模型。

5 结束语

本文面向 PCB 生产实际制定了 PCB 表面缺陷类型分类标准,构建包含 3 239 幅图像 4 672 个缺陷目标的数据集 Dataset_PCBSD,可为相关模型训练和模型性能检测提供更符合实际的数据基准。

本文基于 YOLOv5s,通过引入四尺度检测模型结构、SE 注意力机制和 Soft-NMS 后处理构建了 YOLOv5s-P6SE 模型。基于所构建的 Dataset_PCBSD 进行验证,实验结果显示 YOLOv5s-P6SE 取得了 76.3%的 mAP,相对于 Faster R-CNN、SSD、YOLOv4-MN3、RT-DETR-L 在 mAP 和模型大小上具有明显优势,与最新的 YOLOv8s 相比,在保证模型大小相差仅 2.6 MiB 的情况下获得了明显更优的 mAP,表明了所提出 YOLOv5s-P6SE 的优越性。

后续将进一步扩充现有的数据集,并且平衡各类别样本。同时,深入研究新的检测机制,以提升检测精度,并在模型参数量和检测速度等方面保持平衡。

参考文献:

[1] ANITHA D B,RAO M. A survey on defect detection in bare PCB and assembled PCB using image processing techniques[C]//2017 International Conference on Wireless Communications,Signal Processing and Networking,2017:39-43.

[2] 吴一全,赵朗月,苑玉彬,等. 基于机器视觉的 PCB 缺陷检测算法研究现状及展望[J]. 仪器仪表学报,2022,43(8):1-17.
WU Yiquan,ZHAO Langyue,YUAN Yubin,et al. Research status and the prospect of PCB defect detection algorithm based on machine vision[J]. Chinese Journal of Scientific Instrument,2022,43(8):1-17.

[3] LING Q,ISA N A M. Printed circuit board defect detection methods based on image processing,machine learning and deep learning:A survey[J]. IEEE Access,2023,11:15921-15944.

[4] HUANG W B,WEI P. A PCB dataset for defects detection and classification[J]. arXiv: 1901.08204,2019.

[5] DING R W,DAI L H,LI G P,et al. TDD-net:A tiny defect detection network for printed circuit boards

[J]. CAAI Transactions on Intelligence Technology, 2019,4(2):110-116.

[6] 胡江宇,贾树林,马双宝.基于改进级联 Faster RCNN 的 PCB 表面缺陷检测算法[J].仪表技术与传感器,2022(7):106-110.
HU Jiangyu, JIA Shulin, MA Shuangbao, et al. PCB surface defect detection algorithm based on improved cascaded Faster RCNN[J]. Instrument Technique and Sensor, 2022(7):106-110.

[7] 李振华,张雷.改进 YOLOv5 的轻量级 PCB 缺陷检测算法[J].无线电工程,2023,53(6):1342-1350.
LI Zhenhua, ZHANG Lei. Lightweight PCB defect detection algorithm of improved YOLOv5[J]. Radio Engineering, 2023,53(6):1342-1350.

[8] 王淑青,张子言,朱文鑫,等.基于改进 YOLOv5 的 PCB 板表面缺陷检测[J].仪表技术与传感器,2023(5):106-111.
WANG Shuqing, ZHANG Ziyan, ZHU Wenxin, et al. Surface defect detection of PCB based on improved YOLOv5[J]. Instrument Technique and Sensor, 2023(5):106-111.

[9] DU B W, WAN F, LEI G B, et al. YOLO-MBBi: PCB surface defect detection method based on enhanced YOLOv5[J]. Electronics, 2023, 12(13): Article No.: 2821.

[10] 庹冰,黄丽雯,唐鑫,等.基于 YOLOX-WSC 的 PCB 缺陷检测算法研究[J].计算机工程与应用,2023,59(10):236-243.
TUO Bing, HUANG Liwen, TANG Xin, et al. Research on PCB defect detection algorithm based on YOLOX-WSC[J]. Computer Engineering and Applications, 2023, 59(10):236-243.

[11] 王淑青,鲁濠,鲁东林,等.基于轻量化人工神经网络的 PCB 板缺陷检测[J].仪表技术与传感器,2022(5):98-104.
WANG Shuqing, LU Hao, LU Donglin, et al. PCB board defect detection based on lightweight artificial neural network[J]. Instrument Technique and Sensor, 2022(5):98-104.

[12] TANG J L, LIU S B, ZHAO D X, et al. PCB-YOLO: An improved detection algorithm of PCB surface defects based on YOLOv5[J]. Sustainability, 2023, 15(7): Article No.: 5963.

[13] HU B, WANG J H. Detection of PCB surface defects with improved Faster R-CNN and feature pyramid network[J]. IEEE Access, 2020, 8: 108335-108345.

[14] ADIBHATLA V A, CHIH H-C, HSU C-C, et al. Applying deep learning to defect detection in printed circuit boards via a newest model of you-only-look-once[J]. Mathematical Biosciences and Engineering, 2021, 18(4): 4411-4428.

[15] PHAM T T A, THOI D K T, CHOI H, et al. Defect detection in printed circuit boards using semi-supervised learning[J]. Sensors, 2023, 23(6): Article No.: 3246.

[16] LIAO X T, LÜ S P, LI D H, et al. YOLOv4-MN3 for PCB surface defect detection[J]. Applied Sciences, 2021, 11(24): Article No.: 11701.

[17] LIN T-Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision, 2014: 740-755.

[18] BOCHKOVSKIY A, WANG C-Y, LIAO H. YOLOv4: Optimal speed and accuracy of object detection[J]. arXiv: 2004.10934, 2020.

[19] LIN T-Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 936-944.

[20] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 8759-8768.

[21] ZHENG Z H, WANG P, REN D W, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics, 2022, 52(8): 8574-8586.

[22] WANG C-Y, LIAO H-Y M, WU Y-H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 1571-1580.

[23] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[24] BODLA N, SINGH B, CHELLAPPA R, et al. Soft-NMS -- Improving object detection with one line of code[C]// 2017 IEEE International Conference on Computer Vision, 2017: 5562-5570.

[25] TANG S L, HE F, HUANG X L, et al. Online PCB defect detector on a new PCB defect dataset[J]. arXiv: 1902.06197, 2019.

[26] ZHAO Y, LÜ W Y, XU S L, et al. DETRs beat YOLOs on real-time object detection[C]//2024 IEEE/

CVF Conferenceon Computer Vision and Pattern Recognition, 2024:16965-16974.

[27] 廖鑫婷,张洁,吕盛坪.融合浅层特征和注意力机制的PCB缺陷检测方法[J].计算机集成制造系统,2024,30(3):1092-1104.
LIAO Xinting, ZHANG Jie, LÜ Shengping. Defect detection for PCB by combining shallow features and attention mechanism[J]. Computer Integrated Manufacturing Systems, 2024, 30(3):1092-1104.

[28] AN K, ZHANG Y P. LPViT:A transformer based model for PCB image classification and defect detection[J]. IEEE Access, 2022, 10:42542-42553.

[29] XIE Y F, HU W T, XIE S W, et al. Surface defect detection algorithm based on feature-enhanced YOLO[J]. Cognitive Computation, 2023, 15(2):565-579.

[30] 谢翔,肖金球,汪俞成,等.基于改进YOLOv5s的Deep PCB缺陷检测算法研究[J].微电子学与计算机,2023,40(7):1-9.
XIE Xiang, XIAO Jinqiu, WANG Yucheng, et al. Deep PCB defect detection based on improved YOLOv5s algorithm[J]. Microelectronics & Computer, 2023, 40(7):1-9.

作者简介:

梁泰然(　　—),男,广东佛山人,硕士生,研究方向为机器视觉和缺陷识别。**E-mail**:liangtairan@stu.scau.edu.cn

LIANG Tairan, born in 1999, MS candidate, his research interests include machine vision and defect recognition.

蒋诗新(　　—),男,湖北孝感人,硕士,工程师,研究方向为智能制造和质量大数据。**E-mail**:jiangshixin@ceprei.com

JIANG Shixin, born in 1989, MS, engineer, his research interests include intelligent manufacturing and quality big data.

李泉洲(　　—),男,湖北黄冈人,硕士,工程师,研究方向为智能制造和质量大数据。**E-mail**:liquanzhou@ceprei.com

LI Quanzhou, born in 1993, MS, engineer, his research interests include intelligent manufacturing and quality big data.

欧阳斌(　　—),男,江西赣州人,硕士生,研究方向为机器视觉和缺陷识别。**E-mail**:ouyangbin@stu.scau.edu.cn

OUYANG Bin, born in 1998, MS candidate, his research interests include machine vision and defect recognition.

吕盛坪(　　—),男,湖南新邵人,博士,副教授,研究方向为智能制造和智慧农业。**E-mail**:lvshengping@scau.edu.cn

LÜ Shengping, born in 1982, PhD, associate professor, his research interests include intelligent manufacturing and intelligent agriculture.

ISSN 1001-3695
CN 51-1196/TP
CODEN JYYIC7

计算机应用研究

Application Research of Computers

第41卷第8期 2024年8月
Vol. 41 No. 8 Aug. 2024

8

2024

四川省计算机研究院主办
中国计算机学会会刊

❖英国《科学文摘》(INSPEC)来源期刊
❖俄罗斯《文摘杂志》(AJ)来源期刊
❖美国《乌利希期刊指南(网络版)》(Ulrichsweb)收录期刊
❖2017—2019年中国国际影响力优秀学术期刊(自然科学与工程技术)
❖第二届国家期刊奖百种重点科技期刊
❖中国科技核心期刊 ❖全国中文核心期刊
❖中国科技论文统计源期刊
❖中国学术期刊综合评价数据库来源期刊
❖RCCSE核心学术期刊 ❖中国期刊方阵双效期刊
❖《日本科学技术振兴机构数据库》(JST)来源期刊
❖美国《文摘与引文数据库》(EBSCO)全文来源期刊
❖美国《剑桥科学文摘(自然科学)》(CSA(NS))核心期刊
❖波兰《哥白尼索引》(IC)来源期刊
❖中国科学引文数据库(CSCD)来源期刊
❖《中文科技期刊数据库》来源期刊
❖《中国期刊网》《中国学术期刊(光盘版)》来源期刊
❖中国精品科技期刊顶尖学术论文(F5000)项目来源期刊
❖《电子科技文献数据库》来源期刊
❖《中国工程技术电子信息网》来源期刊

计算机应用研究
Jisuanji Yingyong Yanjiu
第41卷 第8期 2024年8月

目 次

考虑强制同机并行作业的广义作业车间调度优化*

金 鸿，张胡成，信德全，吕盛坪†
（华南农业大学 工程学院，广州 510642）

摘 要： 模具组合加工、电子产品合检等带来不同工件强制同机并行作业，这打破了作业车间调度同一机器不能在同一时刻处理不同工件的约束。为解决该类作业车间调度问题，提出一种自适应混合初始化遗传算法对其进行求解。首先，将该问题定义为考虑强制同机并行作业的广义作业车间调度；利用混合整数规划法以最小化最大完工时间为优化目标建立优化模型。然后，新设计了相应的编码、解码以支持同机并行作业约束下可行调度方案的表达和约束解析；建立了种群混合初始化方法，以支持新约束下高质量可行解的生成；设计了新的交叉、变异操作方法，保证了同机并行作业约束下新生解的可行性；构建了交叉、变异自适应算子，实现了子代的自适应更新，提高了算法全局搜索能力。最后，基于作业车间调度基准算例构建了 40 个测试算例，对该测试算例和电子产品分组合检实例开展实验。结果表明，所构建模型和算法可以有效求解强制同机并行作业的广义作业车间调度问题，提出的改进策略均有效提升了解的质量，验证了模型的可行性和算法的优越性。

关键词： 强制同机并行作业；广义作业车间调度；自适应；混合初始化；遗传算法
中图分类号： TP18 **文献标志码：** A **文章编号：** 1001-3695(2024)08-014-2343-08
doi:10.19734/j.issn.1001-3695.2023.12.0609

General job shop scheduling optimization considering mandatory concurrent operations on same machine

Jin Hong, Zhang Hucheng, Xin Dequan, Lyu Shengping†
(*School of Engineering, South China Agricultural University, Guangzhou* 510642, *China*)

Abstract: The combination processing of molds and electronic product joint inspection introduce the mandatory concurrent operations on the same machine(MCDSM), which breaks the constraint that the same machine cannot process different jobs at the same time in the job shop scheduling(JSP). To solve this type of job shop scheduling problem, this paper proposed an adaptive hybrid initialization genetic algorithm(AHIGA). Firstly, this paper defined the problem as a generalized JSP considering mandatory concurrent operations on the same machine(GJSPMCO) problem, and established an optimization model using mixed integer programming with the objective of minimizing the maximum completion time. Next, it designed corresponding encoding and decoding techniques to support the representation and resolution of feasible scheduling plans within the constraints of MCOSM. Meanwhile, it established a population hybrid initialization method to generate high-quality feasible solutions. This paper also designed novel crossover and mutation operation methods to ensure the feasibility of newly generated solutions under the constraint of MCOSM. Additionally, it constructed an adaptive crossover and mutation operators to enable the adaptive updating of offspring, thereby enhancing the algorithm's global search capability. Finally, it constructed 40 test cases based on JSP benchmark, and used these test cases and an example of joint inspection of electronic products for experiments. The results show that the proposed model and algorithm can effectively solve GJSPMCO problem, and the improvement strategies can enhance the quality of solutions, demonstrating the feasibility of the model and the superiority of AHIGA.

Key words: mandatory concurrent operations on the same machine; generalized job shop scheduling; adaptive; hybrid initialization; genetic algorithm

0 引言

模具生产中的部分工序需要进行组合加工，使得车间运行优化时需要考虑不同工件强制同机并行作业的特殊约束。比如，模具型腔结构复杂，经常设计成由多个镶块组成的型腔，为保证模具型腔成型精度和外观的美观，在进行型腔加工之前应先将镶块正确地镶拼在相应位置，而各镶块上的其他特征单独进行加工。类似，在电子产品检测过程中，检测车间对同种产品样机设计总工艺路线，并将样机划分为不同组，各分组样机按照其工艺路线指定子路线串行完成各工序的检测。但部分检测样机需要跨组组合检测经历不同前置检测工序的多个组件或功能，形成强制同机并行作业约束。无论是型腔加工还是电子产品检测，都要确保强制同机并行作业工序的所有前驱工序全部完成，才能开始进行强制同机并行作业。

模具组合加工、电子产品检测车间同种产品不同样机合检等带来了不同工件强制同机并行作业，这打破了作业车间调度(job shop scheduling,JSP)同一机器不能在同一时刻处理多个不同工件的约束。这种新约束给车间的智能调度优化带来了新的挑战，解决带强制同机并行作业约束的作业调度是模具生产和电子产品检测车间亟待突破的瓶颈。为此，本研究将其定义为考虑强制同机并行作业的广义作业车间调度(general JSP considering mandatory concurrent operations on the same machine,

收稿日期： 2023-12-28；**修回日期：** 2024-02-18 **基金项目：** 国家自然科学基金资助项目(52275487)
作者简介： 金鸿(1988—)，男，湖北通城人，讲师，硕导，博士，主要研究方向为智能制造、智能优化算法；张胡成(1999—)，男，安徽芜湖人，硕士研究生，主要研究方向为车间调度、智能优化算法；信德全(1995—)，男，河南周口人，硕士研究生，主要研究方向为车间调度、绿色智能制造；吕盛坪(1982—)，男(通信作者)，湖南邵阳人，副教授，硕导，博士，主要研究方向为机器视觉、智能调度优化与工业大数据(lvshengping@scau.edu.cn)。

GJSPMCO)。

JSP 是车间运行优化的核心,是优化利用车间生产资源、提高生产效率、减少车间运行成本、缩短生产周期的重要手段[1]。国内外学者对其已开展了大量研究。从模型约束角度看,相关研究者主要考虑了车间不确定性[2]、搬运和货物托盘影响(如 AGV 联动)[3]、人机双资源约束[4]等。这些约束给 JSP 模型构建和优化求解带来了新的挑战,但其更符合车间生产实际。同时,JSP 研究扩展考虑了工艺柔性(如机器可选、工艺路线可选以及工序顺序无关)[5~7],形成了柔性作业车间调度(flexible job shop scheduling,FJSP)。无论是 JSP 还是 FJSP,其所考虑的核心约束为单一工件作业时独占其所选机器,未涉及本研究所考虑强制同机并行作业约束。

从(F)JSP 优化方法角度看,现有研究主要经历了三个主要阶段。20 世纪 60 年代,混合整数规划、动态规划等运筹学方法以及一些寻找近似优化解的启发式算法被提出来[8,9];70 年代,JSP 被相关学者证实为 NP-Complete 问题[10],难以在多项式时间内得到精确最优结果,故大量启发式规则不断涌现;80 年代以来,大量智能优化方法不断应用于(F)JSP[11~16],这些方法能快速获得近似最优甚至最优化结果,大大扩展了(F)JSP 的求解范围与规模,是当前深受欢迎的方法。近年来,候鸟算法[17]、蛙跳算法[18]、灰狼算法[19]、樽海鞘群算法[20]、强化学习算法[21]等智能优化算法被广泛应用于该类问题。

上述研究推动了(F)JSP 约束模型构建和优化求解,但是所建立的约束模型和相应优化机制难以直接应用于 GJSPMCO。因为 GJSPMCO 不同于传统的(F)JSP,(F)JSP 中的工件之间相互独立,作业遵循各自的加工路径,在可行方案表达、生成、解析、迭代操作时只需考虑各自工件的前驱工序约束;GJSPMCO 中的某些工件之间存在强制同机并行作业约束,导致工件之间并不相互独立,在可行方案表达、生成、解析、迭代操作每一步都需要联合考虑强制同机并行作业工序的前驱工序约束,从而保证每一步生成的个体都为可行个体。将(F)JSP 中的可行方案表达、生成、解析、迭代操作应用于 GJSPMCO,打破强制同机并行作业工序的前驱工序约束,从而生成不可行解。因此强制同机并行作业约束给问题建模、优化求解时可行方案的表达、生成、解析、迭代操作等带来挑战。遗传算法因自身具有并行性、灵活性、适应性、可拓展性等特点,在求解(F)JSP 时能取得不错的效果,所以当前大量学者在求解(F)JSP 时会选择在遗传算法的基础上进行改进[22]。而且遗传算法的求解过程是基于个体的基因编码,可以对问题进行灵活的建模和表达,所以可以根据实际问题的需要定义适当的编码方式和操作,以便更好地表达问题的约束和目标,并且遗传算法的可拓展性可以使遗传算法与其他优化方法结合,形成混合算法,从而进一步提高求解效果。为简单高效地求解 GJSPMCO,本文将在遗传算法的基础上对 GJSPMCO 开展研究。创新工作如下:a)定义了 GJSPMCO 新问题,并利用混合整数规划法建立了 GJSPMCO 优化模型;b)以最小化最大完工时间为优化目标,根据 GJSPMCO 问题特性提出一种自适应混合初始化遗传算法(adaptive hybrid initialization genetic algorithm,AHIGA),具体包括针对强制同机并行作业约束设计了相应编码、解码、混合初始化种群以及交叉、变异操作方法,并引入了自适应算子提高了算法全局搜索能力;c)基于 JSP 基准算例,通过随机引入强制组合作业构建了 GJSPMCO 测试算例,基于该测试算例和电子产品分组合检实例开展测试和对比分析,验证了 AHIGA 的可行性和优越性。

1 GJSPMCO 问题描述

GJSPMCO 问题描述如下:N 个工件$\{1,2,\cdots,N\}$在 M 台机器$\{1,2,\cdots,M\}$上作业;各工件具有确定的工艺路线,但存在不同工件的多个工序形成组合工序的同时,在同一机器上并行完成作业;为提高泛化性并满足车间生产实际,同时考虑机器柔性即每道工序可能在多台不同机器上作业,各工件工序的作业时间由所在机器确定。调度目标是在满足机器可用、工序顺序和强制同机并行作业约束条件下,为各工序选择最合适的机器,确定每台机器上各工序的最佳作业顺序,使系统的某些性能指标达到最优。表 1 给出了四个工件四台机器的 GJSPMCO 实例,机器下面对应的数字表示工序在该机器上的作业时间。

表 1 GJSPMCO 实例

Tab. 1 Example of GJSPMCO

工件	工序	可选择的作业机器			
		M1	M2	M3	M4
J_1	O_{11}	5	—	7	—
	O_{12}	5	—	—	—
	O_{13}	10	—	12	—
	O_{14}	14	13	—	14
J_2	O_{21}	8	—	7	10
	O_{23}	5	7	—	5
J_3	O_{31}	4	7	6	—
	O_{32}	8	12	—	—
	O_{34}	5	—	3	5
J_4	O_{41}	—	4	—	7
	O_{42}	6	—	7	—
	O_{44}	—	12	—	8
J_2, J_3	O_{22}, O_{33}	7	4	—	—
J_2, J_4	O_{24}, O_{43}	5	3	6	—

注:J_1 表示工件 1,O_{12}表示 J_1 的第 2 道工序,其余类似;J_2、J_3 表示工件 2 与工件 3 中存在工序进行强制同机并行作业,对应强制同机作业组合工序为 O_{22}、O_{33},其余类似。

为此,GJSPMCO 还需要满足如下约束条件:

a)一台机器在同一时刻只能处理一个工序或指定组合工序;

b)一个工件同一时刻只能在一台机器上作业;

c)同一工件的工序之间存在先后顺序约束,不同工件的工序之间受到强制同机并行作业约束影响;

d)所有工件的优先级相同;

e)某道(组合)工序一旦在某台机器上开始作业就不能中断,直到该工件作业完成;

f)所有工件在零时刻可以对其进行作业。

现有的(F)JSP 模型无法准确描述 GJSPMCO 问题:一是强制同机并行作业的多个工序形成组合工序,组合工序中组成元素对应工件紧前工序耦合影响该组合工序的可开始时间,该组合工序的结束时间影响组成元素的所有工件紧后工序的可开始时间,但现有(F)JSP 模型中并未引入该约束;二是现有(F)JSP模型无法约束构成组合工序的多个工序元素之间的关系。为解决上述问题,建模时先将各工件的工序 O_{ij}统一抽象为只包含一个元素的组合工序 O_{IJ},并将 O_{ij}和 O_{IJ}统称为工序。在此基础上构建优化模型,模型涉及的符号定义如表 2 所示。

在此,基于混合整数规划法构建其优化模型,以最小化最大完工时间为优化目标。

$$C_{\max}=\min(\max(C_m)) \quad 1\leqslant m\leqslant M \tag{1}$$

约束:

$$C_{IJ}-S_{IJ}=P_{IJm} \quad \forall I,J,m \tag{2}$$

$$\sum_{m=1}^{M}X_{IJm}=1 \quad \forall I,J \tag{3}$$

$$S_{IJ}+X_{IJm}P_{IJm}\leqslant C_{IJ} \quad \forall I,J,m \tag{4}$$

$$C_{IJ}\leqslant C_{\max} \quad \forall I,J \tag{5}$$

$$C_{IJ}+Q(Y_{IJPKm}-1)\leqslant S_{PK} \quad \forall I,J,P,K,m \tag{6}$$

$$C_{i(j-1)}\leqslant S_{IJ} \quad \forall I,J,O_{ij}\in O_{IJ} \tag{7}$$

$$S_{IJ}\geqslant 0, P_{IJm}>0, C_{IJ}>0 \quad \forall I,J,m \tag{8}$$

$$S_{ij}\times X_{IJm}=S_{pq}\times X_{IJm} \quad \forall O_{ij},O_{pq}\in O_{IJ} \tag{9}$$

$$C_{ij}\times X_{IJm}=C_{pq}\times X_{IJm} \quad \forall O_{ij},O_{pq}\in O_{IJ} \tag{10}$$

上述模型约束关系主体基于 O_{IJ} 构建,如无特别说明,这里的工序均指 O_{IJ};具体组成元素 O_{ij} 以工件工序进行说明。式(1)为优化目标函数;式(2)表示工序进行作业后中途不能中断;式(3)表示任意工序在机器上只作业一次;式(4)表示任意工序的开始作业时间都不大于其完工时间;式(5)表示任意工序完工时间都不大于最大完工时间;式(6)表示每台机器在同一时刻只能有一道工序在作业;式(7)表示任意工件工序开始作业时间不小于该工序工件紧前工序的完工时间,同时约束了组合工序的开始时间大于等于其所有组成元素的工件紧前工序的结束时间;式(8)表示任意工序的开始作业时间非负,作业时间和完工时间大于 0;式(9)和(10)表示组合工序 O_{IJ} 的各元素 $O_{ij}\in O_{IJ}$ 必须同时开始和同时完工。

表 2　符号说明

Tab. 2　Symbol description

符号	说明	符号	说明
N	工件的数量	$C_{IJ}(C_{ij})$	$O_{IJ}(O_{ij})$ 的完成时间
M	机器的数量	E_{IJm}	O_{IJ} 在机器 m 上完成时间
I,P	工件集编号	P_{IJm}	O_{IJ} 在机器 m 的作业时间
J,K	工序集编号	M_{IJ}	O_{IJ} 可选机器集
m	机器编号	C_m	机器 m 上最后一道工序的完工时间
O_{IJ}	组合工序 $O_{IJ}=\{O_{ij},O_{pq},\cdots,O_{yz}\}$	Q	一个无穷大的实数
N_i	工件 i 的工序数	X_{IJm}	O_{IJ} 是否在机器 m 上作业,在机器 m 作业为 1,否则为 0
O_{ij}	工件 i 的工序 $j,j\in[1,N_i]$	Y_{IJPKm}	如果 O_{IJ} 机器 m 上先于 O_{PK} 作业时为 1,反之为 0
$S_{IJ}(S_{ij})$	$O_{IJ}(O_{ij})$ 的开始作业时间		

2　自适应混合初始化遗传算法

为满足强制同机并行作业约束要求,设计新的编码、解码以及初始可行方案生成机制,在此基础上研究相应的优化迭代操作方法,为此提出 AHIGA 机制,具体流程如图 1 所示。

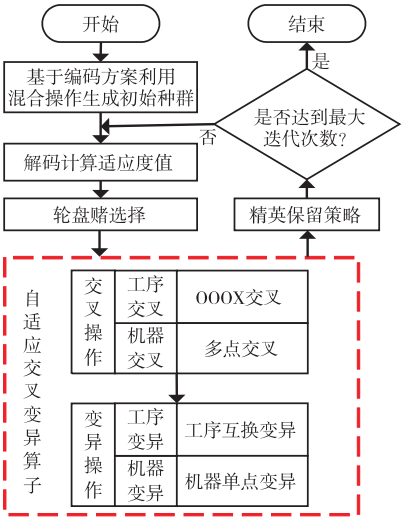


图 1　AHIGA 流程

Fig. 1　Flow of AHIGA

2.1　编码和解码

GJSPMCO 考虑工序排序和机器选择,在此采用工序编码,编码由三段整数编码序列组成,表 1 对应的编码实例如图 2 所示。

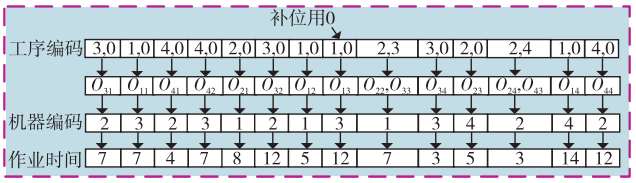


图 2　编码示意

Fig. 2　Schematic of encoding

第一段为工序编码序列,每个基因用工件集序号表示,为了保证各基因位的一致性,每个基因位的元素数量设为 $l=\max\{|O_{IJ}|\},\forall I,J$;对于 $|O_{IJ}|<l$,在该基因位用虚拟工件 0 进行补位,具体如图 2 第一行所示。每个基因位中非 0 元素 i,$i\in[1,N]$ 出现频次 $j\in[1,N_i]$ 代表工件 i 对应工序 O_{ij},如图 2 中第二行所示。第二段为机器编码,其可选范围为 $[1,M]$,机器编码顺序与工序编码相对应,表示该工序编码序列对应(组合)工序所指定作业机器。第三段为作业时间编码,表示其对应工序在指定机器上作业所需要时间。

基于编码方案设计的顺序解码方法如算法 1 所示。

算法 1　顺序解码方法

a)设置 $A_O(1,1:ON)$、$A_O(2,1:ON)$ 分别存储各工序的开始时间和完工时间,$ON=|OPlan|$;用 $JP=\{JP[1],JP[2],\cdots,JP[N]\}$、$MP=\{MP[1],MP[2],\cdots,MP[M]\}$ 分别记录各工件和机器紧前工序的工序结束时间。初始化 A_O、MP、JP 元素为 0,$k=1$。

b)从左往右顺序读取工序编码序列 $OPlan[k]$,$1\leqslant k\leqslant|OPlan|$ 中各基因位非 0 元素(工件编号),确定其工序 O_{IJ},并从 $MPlan$ 和 $TmPlan$ 中分别获取 O_{IJ} 作业机器 m 和作业时间 P_{IJm}。

c)O_{IJ} 的开始和完工时间分别为 $S_{IJ}=\max\limits_{i\in I}\{JP[i],MP[m]\}$,$C_{IJ}=S_{IJ}+P_{IJm}$。

d)更新 $JP[i]=C_{IJ},i\in I,MP[m]=C_{IJ},A_O(1,k)=S_{IJ},A_O(2,k)=C_{IJ}$。

e)若 $k<ON,k=k+1$,转 b),否则转 f)。

f)结束。

2.2　混合初始化种群

初始种群的质量决定遗传算法求解质量和收敛速度。为增强初始解的质量,在此采用混合初始化方式[23],其中一半种群采用贪婪初始化方式生成,另一半种群采用随机初始化方式生成。基于编码方式设计混合初始化种群生成算法,如算法 2 所示。

算法 2　混合初始化种群

a)初始化种群规模 NP,定义 $NP\times 3$ 的数组 Pop,计数器 $np=1,k=\max\{|O_{IJ}|\},\forall I,J$。

b)生成新的空的基于工件表示的工序序列 $OPlan$、机器序列 $MPlan$ 和作业时间序列 $TmPlan$,设置 $[1,N]$ 所有数为未标状态。

c)如果 $[1,N]$ 全部被标记,转 f);反之,随机选取 $[1,N]$ 中未标记的数 i,判断 $OPlan$ 中 i 出现频次 j,如果 $j=N_i$,标记工件 i,转 c);如果 $j<N_i$,从不同工件工艺路线 $PPlan_k$,$1\leqslant k\leqslant N$ 中获取 O_{ij},如果 O_{ij} 为常规工序,将 i 加入 $OPlan$ 尾部,并在该基因位补 $l-1$ 个 0;如果 O_{ij} 为组合工序元素,转 d)。

d)获取 $O_{ij}\in O_{IJ}=\{O_{ij},O_{pq},\cdots,O_{yz}\}$ 对应工件集 I,将 I 中各工件编号绑定放入到 $OPlan$ 尾部,并在该基因位补 $l-|O_{IJ}|$ 个 0。

e)获取 O_{IJ} 未放入调度序列的前驱工序集 $JP[O_{IJ}]$,判断 O_{IJ} 中 $i\in I$ 是否有前驱组合工序,如果工件 i 无前驱组合工序,依次将 $O_{ij}\in JP[O_{IJ}]$ 对应工件编号 i 随机插入在 $OPlan$ 中 I 之前;反之将其随机插入在最靠近 i 的前驱组合工序和 I 之间,转 c);并在插入的基因位补 $l-1$ 个 0。

f)生成(0,1)的随机数 r,若 $r<0.5$,对于 $OPlan[k]$,$1\leqslant k\leqslant|OPlan|$ 中工序 O_{IJ},根据顺序解码选择 $m=\text{argmin}\{E_{IJm_1},E_{IJm_2},\cdots,E_{IJm_j}\}$,$m_1,m_2,\cdots,m_j\in M_{IJ}$ 作为作业机器;反之从 M_{IJ} 中随机选择 m 作为 O_{IJ} 作业机

器;最后形成机器序列 *MPlan*。进一步确定各工序作业时间,生成作业时间序列 *TmPlan*。存储 $Pop(np,1)=OPlan$、$Pop(np,2)=MPlan$、$Pop(np,3)=TmPlan$。

g)如果 $np<NP$,$np=np+1$,转 b);反之结束。

利用算法 2 和表 1 中的数据生成一个初始可行方案的示意,如图 3 所示。假设算法 2 中步骤 c)随机生成了(3,0)(1,0)(4,0)(2,0)(1,0)(1,0)工件集序号表示的工序序列(如图 3 中第一行),对应工序分别为 O_{31}、O_{11}、O_{41}、O_{21}、O_{12}、O_{13}。然后基于步骤 c)随机选取工件 2,对应工序 O_{22}、O_{22}和 O_{33}形成组合工序,基于步骤 d)在序列尾部放入 O_{22}、O_{33},对应工件(2,3)(如图 3 中第三行),但 O_{33}前驱工序 O_{32}未插入工序序列且 O_{22}、O_{33}无前驱组合工序,基于步骤 e)将 O_{32}对应工件集编号 3 插入到(2,3)之前,并采用 0 补位(如图 3 中的第四行),网格框为工件 3 随机插入位置。在已生成工序序列基础上基于步骤 c)生成工件编号 3 的工序序列,对应工序为 O_{34}。继续基于步骤 c)随机生成工件编号 4,对应工序 O_{43},因为 O_{24}和 O_{43}为组合工序 O_{24}、O_{43},基于步骤 d)在序列尾部放入 O_{24}、O_{43},对应工件(2,4)。但 O_{24}前驱工序中 O_{23}未进行插入且 O_{24}、O_{43}工件 2 存在前驱组合工序 O_{22}、O_{33},基于步骤 e)将 O_{23}对应工件编号 2 插入(2,4)和(2,3)之间,并采用 0 补位。根据步骤 c)继续选取未标记的工件编号连同补位 0 一起加入工序序列尾部,直至 1 ~4 中所有数都被标记,此时工序序列编码完成,最终工序编码序列如图 3 中第五行所示。进一步基于步骤 f)生成对应机器和作业时间序列,如图中第六、七行所示。

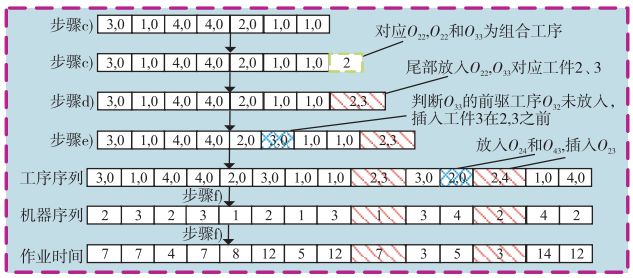


图 3 初始可行方案生成示意

Fig. 3 Schematic of initial feasible solution generation

2.3 自适应遗传操作

迭代过程中的遗传操作包括选择、交叉、变异和精英保留等。选择操作是为了挑选出种群中优秀的个体作为后续交叉变异的对象,本研究以目标函数的倒数为适应度函数 f,采用轮盘赌方式进行选择。对于种群规模为 NP 的种群,基于各个体概率 $P_i=f_i/\sum_{j=1}^{NP}f_j$,$1\leqslant i\leqslant NP$ 计算相应累计概率 $Q_q=\sum_{i=1}^{q}P_i$,$1\leqslant q\leqslant NP$,然后生成(0,1)之间的随机数 r,当 $r\leqslant Q_1$,选择第一个个体;当 $Q_{q-1}\leqslant r\leqslant Q_q$,选择第 q 个个体,当选择 NP 数量个体时选择操作结束。

交叉是产生具有更优基因组合后代的重要操作。常见的交叉操作有单点交叉、多点交叉、均匀交叉、次序交叉、工序顺序交叉和扩展顺序交叉等[24]。但这些交叉操作无法处理同机并行作业工序带来的耦合影响,无法避免在子代中产生不可行解。为满足同机并行作业约束,需设计新的交叉机制,对工序序列设计逐个顺序交叉方法(one by one order crossover, OOOX),具体步骤如下:

a)选择父代工序序列 *P1* 和 *P2*,设置工序序列子代 $CP=\varnothing$、设备序列子代 $CM=\varnothing$和作业时间序列子代 $CT=\varnothing$。

b)生成(0,1)的随机数 r,若 $r<0.5$,则选择 *P1* 工序序列第一个基因位元素,否则选择 *P2* 工序序列第一个基因位元素。

c)将所选元素添放在 *CP* 末尾,并将该元素对应机器和时间分别放入 *CM* 和 *CT* 末尾。在 *P1* 和 *P2* 中删除第一次出现该元素的基因。

d)重复 b)和 c),直到父代 *P1* 和 *P2* 中元素为空,则工序交叉完成。

以表 1 数据为例,OOOX 示意如图 4 所示。首先通过随机数 $r<0.5$ 选择 *P1* 工序序列第一个基因位元素 3,将元素 3 添放入子代 *CP* 首位,并删除 *P1* 和 *P2* 中第一次出现元素 3 的基因位;然后继续生成随机数 $r\geqslant 0.5$,选择 *P2* 第一个基因位元素 4,将元素 4 添放入子代 *CP* 末尾,并删除 *P1* 和 *P2* 中第一次出现元素 4 的基因位;依此类推,直到 *P1* 和 *P2* 中元素为空则交叉完成,生成子代工序序列 *CP*。

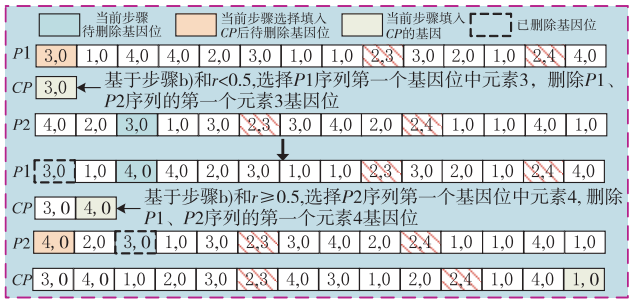


图 4 逐个顺序交叉示意

Fig. 4 Schematic of one by one order crossover

为使交叉更加充分,在完成工序交叉后进一步对机器编码序列进行多点交叉,具体步骤如下:

a)随机生成一段长度等于机器编码序列长度的 0-1 序列,序列中元素 1 对应机器编码位置即为进行机器交叉的位置。

b)将父代 *P1* 编码序列复制给子代 *C1*,父代 *P2* 编码序列复制给子代 *C2*。

c)将父代 *P1* 机器交叉位置的机器更新到子代 *C2* 对应工序的机器编码位置,将父代 *P2* 机器交叉位置的机器更新到子代 *C1* 对应工序的机器编码位置。

d)*P1* 机器交叉位置的机器对应的作业时间更新到子代 *C2* 对应工序的作业时间编码位,*P2* 机器交叉位的机器对应的作业时间更新到子代 *C1* 对应工序的作业时间编码位。

机器交叉示意如图 5 所示。首先将父代 *P1* 编码序列复制给子代 *C1*,父代 *P2* 编码序列复制给子代 *C2*,机器编码上方为该机器对应的工序;然后根据产生的随机序列中元素 1 的位置进行机器交叉,父代 *P1* 机器序列中工序 O_{11}、O_{41}、O_{32}、O_{22}、O_{33}、O_{34}、O_{14}、O_{44}对应机器 3、2、2、1、3、1、4 更新到子代 *C2* 工序 O_{11}、O_{41}、O_{32}、O_{22}、O_{33}、O_{34}、O_{14}、O_{44}对应的机器码位,父代 *P2* 机器序列中工序 O_{21}、O_{31}、O_{22}、O_{33}、O_{23}、O_{24}、O_{43}、O_{13}、O_{14}对应机器 3、1、2、1、1、4 更新到子代 *C1* 工序 O_{21}、O_{31}、O_{22}、O_{33}、O_{23}、O_{24}、O_{43}、O_{13}、O_{14}对应的机器码位。

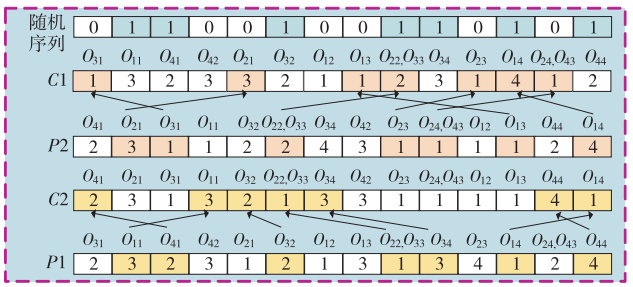


图 5 机器交叉示意

Fig. 5 Schematic of machine crossover

变异操作可增加种群的多样性并防止早熟,影响着算法的局部搜索能力。对于(F)JSP,常见的变异操作有互换变异、逆

序变异、插入变异和单点变异。但是这些变异操作应用于本研究问题的工序序列变异时,因其随机性,不可避免地打破了同机并行作业的工件前驱工序约束。为确保组合工序在变异后能够满足工序前后作业约束,对于工序变异,将以组合工序为节点进行分段变异。先以组合工序为分段点,依次判断各分段工序编码序列长度。若分段工序编码序列长度大于等于 2,则在分段工序编码序列中随机选择两个工序进行工序互换变异,否则不进行变异操作。为得到可行解,变异完成后需更新机器序列和作业时间序列,使得各工序对应的作业机器和作业时间与变异之前相同。工序变异操作如图 6 所示。

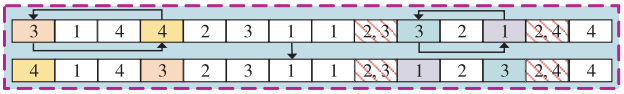


图 6　工序变异

Fig. 6　Process mutation

对于机器编码序列的变异操作采用单点变异方式进行变异,将该位置上的机器随机替换成该工序可选作业机器集中其他一台机器。为得到可行解,与机器编码变异位置对应的作业时间需更新为该工序在替换机器上的作业时间。

为获得更优种群并加速收敛,在此提出一种自适应交叉变异算子。将种群中的个体分为三类:适应度值 f 前 20% 的个体为优良个体,应降低交叉率和变异率来保留自身优良基因;f 倒数 20% 的个体为拙劣个体,应提高交叉率和变异率以增加生成优良个体的概率;其余个体称为普通个体,以一个适中的交叉率和变异率进行交叉和变异。分段函数描述的自适应交叉率和变异率如式(11)(12)所示。

$$P_c=\begin{cases}-\dfrac{\arctan\left(50\left(\dfrac{f_i-f_{\min}}{f_{\max}-f_{\min}+s}-0.2\right)\right)}{5\pi}+0.9 & f_{\min}\leqslant f_i\leqslant\dfrac{f_{\max}+4f_{\min}}{5}\\ -\dfrac{2\arctan\left(20\left(\dfrac{f_i-f_{\min}}{f_{\max}-f_{\min}+s}-0.2\right)\right)}{5\pi}+0.9 & \dfrac{f_{\max}+4f_{\min}}{5}<f_i\leqslant\dfrac{4f_{\max}+f_{\min}}{5}\\ -\dfrac{3\arctan\left(100\left(\dfrac{f_i-f_{\min}}{f_{\max}-f_{\min}+s}-0.8\right)\right)}{5\pi}+0.7 & \dfrac{4f_{\max}+f_{\min}}{5}<f_i\leqslant f_{\max}\end{cases}\tag{11}$$

$$P_m=\begin{cases}-\dfrac{\arctan\left(50\left(\dfrac{f_i-f_{\min}}{f_{\max}-f_{\min}+s}-0.2\right)\right)}{5\pi}+0.3 & f_{\min}\leqslant f_i\leqslant\dfrac{f_{\max}+4f_{\min}}{5}\\ -\dfrac{\arctan\left(20\left(\dfrac{f_i-f_{\min}}{f_{\max}-f_{\min}+s}-0.2\right)\right)}{5\pi}+0.3 & \dfrac{f_{\max}+4f_{\min}}{5}<f_i\leqslant\dfrac{4f_{\max}+f_{\min}}{5}\\ -\dfrac{\arctan\left(50\left(\dfrac{f_i-f_{\min}}{f_{\max}-f_{\min}+s}-0.8\right)\right)}{5\pi}+0.2 & \dfrac{4f_{\max}+f_{\min}}{5}<f_i\leqslant f_{\max}\end{cases}\tag{12}$$

其中:$f_{\max}$、$f_{\min}$分别表示种群中最大和最小的适应度值;f_i 表示个体 i 的适应度值;s 为极小的正实数。

采取精英保留策略保留具有优良基因的个体到下一代。将经过遗传操作的种群与上一代种群合并形成新的种群,计算新种群中个体的适应度值,取新种群中适应度值大的 50% 个体作为下一代的初始种群,精英保留可以将历代种群中具有优良基因的个体保留下来,防止优良个体遗失。

设 N 为种群规模,L 表示一个染色体的位数,P 表示一个染色体中的组合工序数。上述混合初始化、轮盘赌选择、交叉、变异和精英保留等在最坏情况下操作的时间复杂度分别为 $O(N\times L)$、$O(N)$、$O(N\times L)$、$O(N\times(P+1))$、$O(2\times N)$,整体时间复杂度为 $O(N\times L)$,由此推断本文算法的时间复杂度并不高。

3　实验结果与分析

因缺乏 GJSPMCO 问题相关基准实例,本文在 MK01-MK15[25] 和 SFJS6-SFJS10[26] 基准算例基础上,通过随机引入一、两个组合工序生成测试算例,组合工序的作业时间信息如表 3 所示,其中由于 MK01-MK15 和 SFJS6-SFJS10 国际基准算例并未说明单位,所以本研究中基于这两个算例生成的 GJSPMCD 测试算例也无具体单位。以 SFJS6 为例:O_{12} 和 O_{22} 为第一个组合工序,可在机器 1 和 2 上作业,作业时间分别为 150 和 160;O_{13} 和 O_{23} 为第二个组合工序,可在机器 3 上作业,作业时间为 70。同时将对电子产品分组合检实例进行进一步验证。基于上述数据开展实验,所有实验均在 Windows 10 操作系统、主频 3.5 GHz、内存 16 GB 的个人计算机上完成。

表 3　组合工序作业机器和作业时间

Tab. 3　Machine and time of combination processes

测试算例	组合工序	可选机器	作业时间
SFJS6	O_{12},O_{22}	M1/M2	150/160
	O_{13},O_{23}	M3	70
SFJS7	O_{12},O_{32}	M4	140
	O_{13},O_{33}	M4/M5	190/160
SFJS8	O_{12},O_{22}	M2/M4	66/55
	O_{13},O_{33}	M3/M4	160/150
SFJS9	O_{11},O_{31}	M1/M2	50/60
	O_{13},O_{23}	M2/M3	70/60
SFJS10	O_{12},O_{22}	M2	130
	O_{33},O_{43}	M4	190
MK01	O_{33},O_{94}	M2/M4	3/5
	O_{34},O_{75}	M2/M4	3/5
MK02	O_{34},O_{93}	M2/M5/M6	4/2/3
	$O_{82},O_{10}(3)$	M4/M6	5/3
MK03	O_{47},O_{54}	M2/M3/M4/M5	18/13/5/10
	O_{57},O_{82}	M4/M7/M8	13/2/18
MK04	O_{34},O_{63}	M3/M4/M7	9/4/5
	$O_{15},O_{11}(3)$	M7/M8	5/9
MK05	O_{32},O_{62}	M2/M3/M4	6/9/5
	$O_{33},O_{12}(4)$	M1/M2/M3	8/6/7
MK06	O_{14},O_{37}	M2/M7	8/5
	O_{22},O_{94}	M1/M4/M6	6/6/2
MK07	O_{12},O_{33}	M1/M2	5/1
	O_{15},O_{92}	M2/M3	4/8
MK08	O_{74},O_{45}	M1/M10	10/19
	$O_{38},O_{17}(5)$	M3/M4	19/5
MK09	$O_{77},O_{12}(5)$	M2/M6/M8	16/10/17
	$O_{12}(8),O_{15}(6)$	M2/M3/M9	12/11/6
MK10	$O_{27},O_{8}(10)$	M2/M6/M7	5/5/15
	$O_{57},O_{13}(5)$	M2/M4/M7	16/13/14
MK11	O_{33},O_{54}	M4/M5	17/18
	$O_{65},O_{10}(4)$	M3/M4	28/22
MK12	O_{13},O_{57}	M5/M10	22/15
	O_{15},O_{71}	M5/M7	18/24
MK13	$O_{84},O_{10}(8)$	M1/M9	29/29
	$O_{15}(6),O_{16}(5)$	M2/M10	21/18
MK14	$O_{14}(2),O_{15}(7)$	M4/M13	16/10
	$O_{20}(1),O_{22}(4)$	M5/M9	25/28
MK15	O_{31},O_{68}	M2/M15	25/27
	O_{37},O_{44}	M3/M4	24/28

为验证所构建的 GJSPMCO 混合整数规划模型的有效性,基于引入组合工序的 SFJS6-SFJS10 算例,利用 IBM ILOG CPLEX 12.10 进行求解,运行时间上限设置为 360 s。CPLEX 所得结果如表 4 所示。

表 4　CPLEX 求解结果

Tab. 4　CPLEX solution results

测试算例	$n\times m$	$C_{\max}$	CPUs	测试算例	$n\times m$	$C_{\max}$	CPUs
SFJS6 单约束	3×3	327	0.37	SFJS8 双约束	3×4	276	0.28
SFJS6 双约束	3×3	317	0.46	SFJS9 单约束	3×3	210	0.37
SFJS7 单约束	3×5	407	0.59	SFJS9 双约束	3×3	210	0.26
SFJS7 双约束	3×5	417	0.45	SFJS10 单约束	4×5	558	0.36
SFJS8 单约束	3×4	253	0.30	SFJS10 双约束	4×5	618	0.32

表 4 中单约束为只考虑表 3 中对应算例的第一个组合工序,双约束为考虑各算例中两个组合工序,更多组合工序在原理上与双约束场景相似。$n \times m$ 表示测试算例问题规模,$C_{\max}$ 为最小化最大完工时间,CPUs 表示 CPLEX 实际求解时间,单位为秒(s)。可以看出,CPLEX 能获得这些小规模 GJSPMCO 问题的可行解,证明了模型的可行性。

进一步通过对比实验验证 OOOX、混合初始化策略以及自适应算子的效果。在此,将 GA + OOOX 与常规的 GA + 工序顺序交叉(POX)进行实验对比。因直接应用 POX 到 GJSPMCO 将得到不可行子代,所以在此限定 POX 只能对不涉及组合工序的工件进行交叉。在 GA + OOOX 基础上,引入混合初始化,形成 HIGA + OOOX;进一步增加自适应算子,形成最终算法 AHIGA。本研究选取文献中常见的实验参数进行初步实验并选取效果最好的一组实验参数:种群规模为 200,交叉率和变异率为 0.8 和 0.3,最大迭代次数为 300,并采用 MATLAB 2021a 软件实现。实验结果如表 5 所示,表中 C_M 为各算法运行 10 次所得 $C_{\max}$ 的最小值;sd 为算法 10 次运行所得 $C_{\max}$ 的标准差,用来评估算法在求解各算例的稳定性,sd 越小,算法的稳定性越高;sd_{mean} 为 30 个算例下各算法 sd 的平均值,用来评估算法面对不同规模算例时的综合稳定性;RPD 表示相对百分比差异,用于评估当前算法与最优算法之间的差距,计算公式为 $RPD = 100 \times (C_M - Min)/Min$,其中 Min 表示各算法最优结果的最小值,RPD 越小,算法的搜索能力越强;RPD_{mean} 为 30 个算例下所求解得到 RPD 的平均值,用来评估算法面对不同规模算例时的综合搜索能力。加粗数字为所有算法中的最优值。

表 5 改进算法有效性实验

Tab. 5 Improve algorithm effectiveness experiments

问题	GA + POX			GA + OOOX			HIGA + OOOX			AHIGA		
	C_M	sd	RPD	C_M	sd	RPD	C_M	sd	RPD	C_M	sd	RPD
MK01 单约束	**42**	0	0	**42**	0	0	**42**	0	0	**42**	0	0
MK01 双约束	43	1.82	2.38	**42**	1.15	0	**42**	1.52	0	**42**	0.45	0
MK02 单约束	29	1.22	3.57	**28**	0.84	0	**28**	1.14	0	**28**	0.45	0
MK02 双约束	34	1.14	25.93	28	0.45	3.70	28	0	3.70	**27**	1.10	0
MK03 单约束	**204**	0	0	**204**	0	0	**204**	0	0	**204**	0	0
MK03 双约束	192	5.55	2.67	**187**	4.02	0	**187**	4.02	0	**187**	0	0
MK04 单约束	**67**	0.89	0	**67**	0	0	**67**	0	0	**67**	0	0
MK04 双约束	70	3.03	6.06	67	**0.45**	1.52	**66**	**0.45**	0	**66**	**0.45**	0
MK05 单约束	176	1.30	2.33	173	0.84	0.58	**172**	1.14	0	**172**	**0.45**	0
MK05 双约束	176	1.52	3.53	172	**0.84**	1.18	171	0.89	0.59	**170**	0.89	0
MK06 单约束	83	6.28	16.90	72	2.49	1.41	72	1.92	1.41	**71**	**1.64**	0
MK06 双约束	103	9.42	39.19	75	**2.00**	1.35	75	2.88	1.35	**74**	3.83	0
MK07 单约束	141	2.51	4.44	140	1.87	3.70	**135**	**1.48**	0	**135**	2.17	0
MK07 双约束	147	3.91	8.89	139	3.56	2.96	136	**0.83**	0.74	**135**	1.64	0
MK08 单约束	**523**	**4.47**	0	**523**	**4.47**	0	**523**	**4.47**	0	**523**	**4.47**	0
MK08 双约束	**513**	5.48	0	**513**	**4.47**	0	**513**	**4.47**	0	**513**	**4.47**	0
MK09 单约束	342	8.29	6.54	325	5.72	1.25	**321**	5.72	0	**321**	**3.21**	0
MK09 双约束	334	11.62	6.03	331	5.76	5.08	328	**3.74**	4.13	**315**	**3.74**	0
MK10 单约束	266	6.61	17.70	240	5.18	6.19	231	6.18	2.21	**226**	**3.39**	0
MK10 双约束	277	6.77	20.43	237	6.30	3.04	235	5.98	2.17	**230**	**2.83**	0
MK11 单约束	617	4.09	1.48	616	3.67	1.32	611	**3.28**	0.49	**608**	4.10	0
MK11 双约束	622	4.22	2.81	613	3.05	1.32	609	**2.07**	0.66	**605**	**2.07**	0
MK12 单约束	**508**	7.60	0	**508**	**7.16**	0	**508**	**7.16**	0	**508**	8.76	0
MK12 双约束	**508**	7.47	0	**508**	0.45	0	**508**	0	0	**508**	0	0
MK13 单约束	464	12.12	10.21	447	7.82	6.18	423	7.17	0.48	**421**	**5.38**	0
MK13 双约束	462	9.82	10.79	446	11.63	6.95	**417**	9.36	0	**417**	**8.79**	0
MK14 单约束	**694**	0	0	**694**	0	0	**694**	0	0	**694**	0	0
MK14 双约束	**694**	0	0	**694**	0	0	**694**	0	0	**694**	0	0
MK15 单约束	419	12.97	13.55	396	5.27	7.32	377	**2.51**	2.17	**369**	8.06	0
MK15 双约束	413	13.29	8.12	394	7.17	3.14	**382**	**5.02**	0	**382**	5.89	0
sd_{mean}		5.10			3.22			2.78			**2.61**	
RPD_{mean}		7.12			1.94			0.67			0	

从表中可知,在所有 30 个测试算例中,引入 OOOX 的 GA + OOOX 得到的 C_M 和 RPD,有 21 个算例优于 GA + POX 所得结果,其余 9 个算例中,两者得到相同的 C_M 和 RPD,且 RPD_{mean} 降低了 72.8%,表明 OOOX 可以大幅度提升算法全局搜索能力。同时可以看出,GA + OOOX 得到的 sd 有 24 个算例优于 GA + POX 所得结果,5 个算例中,两者得到相同的 sd,仅 1 个算例 GA + OOOX 的 sd 劣于 GA + POX,且 GA + OOOX 与 GA + POX 相比将 sd_{mean} 降低了 36.9%,说明 OOOX 可以显著提高 GA 的稳定性。对比 HIGA + OOOX 和 GA + OOOX 可知,引入混合初始化策略的 HIGA + OOOX 在 GA + OOOX 基础上进一步改进了 15 个算例的 C_M 和 RPD,其余 15 个算例中,两者得到相同的 C_M 和 RPD,且 RPD_{mean} 降低了 65.5%,表明混合初始化策略进一步增强了算法的全局搜索能力。而 HIGA + OOOX 得到的 sd 有 13 个算例优于 GA + OOOX 所得结果,11 个算例中,两者得到相同的 sd,6 个算例 HIGA + OOOX 的 sd 劣于GA + OOOX,且 HIGA + OOOX 与 GA + OOOX 相比,将 sd_{mean} 降低了 13.7%,说明混合初始化策略也可以有效提高算法的稳定性。引入自适应算子的 AHIGA 在 HIGA + OOOX 基础上进一步改进了 12 个算例的 C_M 和 RPD,其余 18 个算例,两者得到相同的 C_M 和 RPD,且 RPD_{mean} 降低至 0,表现出自适应算子对提高算法全局搜索能力的有效性。而 AHIGA 得到的 sd 有 10 个算例优于 HIGA + OOX 所得结果,12 个算例中,两者得到相同的 sd,8 个算例中,AHIGA 的 sd 劣于 HIGA + OOOX,且 AHIGA 与 HIGA + OOOX 相比将 sd_{mean} 降低了 6.1%,表明自适应算子进一步提升了算法的稳定性。上述对比结果证明了 OOOX、混合初始化以及自适应算子均有利于提升算法全局搜索能力和维持算法稳

定性,所提出的优化策略具有可行性和有效性。

各优化策略算法求解 10 次 MK05 双约束问题获得最优结果的迭代过程收敛曲线对比如图 7 所示。可以看出,OOOX、混合初始化以及自适应算子均有利于减少 GA 的迭代次数,获得更好的优化结果,证明了本文方法能快速获得高质量解。

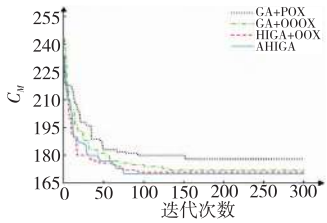


图 7　各算法求解 MK05 双约束问题迭代收敛曲线

Fig. 7　Iterative convergence curves for each algorithm to solve the MK05 doubly constrained problem

图 8 为针对 MK05 双约束问题 AHIGA 所得最优调度方案的甘特图。可以看出,图中所有工序均满足工序作业顺序约束、不同工序在同一机器上作业先后顺序与强制同机并行作业约束等要求。

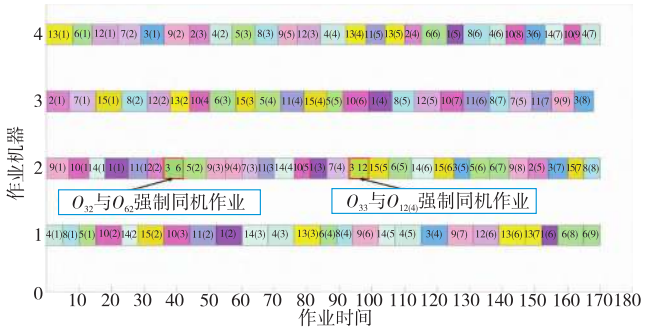


图 8　MK05 双约束调度结果甘特图

Fig. 8　MK05 Ghantt chart of double constrained scheduling results

进一步以电子产品检测车间中的无人机、手机和车载导航仪三类产品的检测实例来验证本文模型和所提出的 AHIGA。其中无人机、导航仪和手机相应的待检样机分别被划分为 4、7、7 组(每组视为一个工件),为各工件分别设计相应工艺路线,具体工艺路线及产品检测信息分别如图 9 ~ 11 所示,工序下方为该工序的作业机器及对应作业时间,作业时间单位为小时(h)。其中车载导航仪和手机的强制同机并行作业约束工序均为其分组后工件 1、2 的跌落冲击测试和工件 3、4 的灰尘实验;无人机的强制同机并行作业约束工序为工件 1、2 的交变盐雾测试。

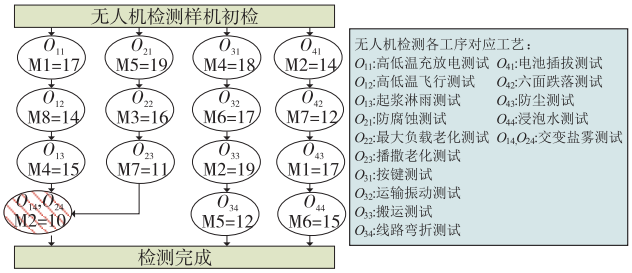


图 9　无人机品质检测工艺流程

Fig. 9　Quality inspection process plan of UAV

针对这三类产品检测作业的调度,表 6 给出了各优化策略运行 10 次的求解结果对比,其中 C_M 单位为小时(h)。可以看出,AHIGA 与 HIGA + OOOX 获得最优 C_M,AHIGA 与 GA + OOOX 获得最优 sd。进一步说明 AHIGA 具有较强的全局搜索能力和较好的解的稳定性。图 12 为 AHIGA 运行 10 次得到的最优调度结果甘特图,其中进行强制同机并行作业的组合工序用红色框图标注(参见电子版),横坐标单位为小时(h)。从图中可以看出,进行强制同机并行作业的组合工序依旧满足工序顺序、作业机器和作业时间约束,再一次验证了 AHIGA 求解 GJSPMCO 问题的可行性。

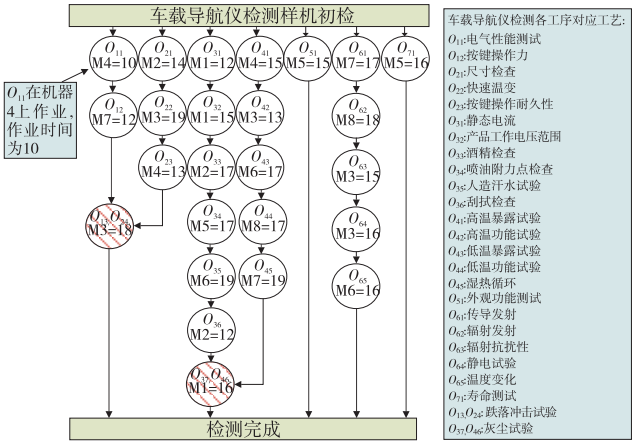


图 10　车载导航仪品质检测工艺流程

Fig. 10　Quality inspection process plan of car navigator

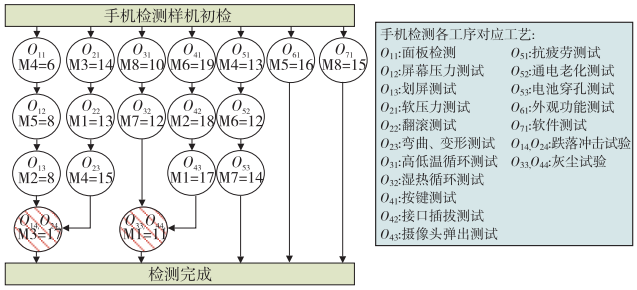


图 11　手机品质检测工艺流程

Fig. 11　Quality inspection process plan of mobile phone

表 6　检测车间实例调度优化结果对比

Tab. 6　Comparison of scheduling results for instances from inspection shop

问题	$n \times m$	GA + OOOX			HIGA + OOOX			AHIGA		
		C_M	sd	RPD	C_M	sd	RPD	C_M	sd	RPD
检测车间调度实例	18 × 8	129	0.55	0.80	128	0.71	0	128	0.55	0

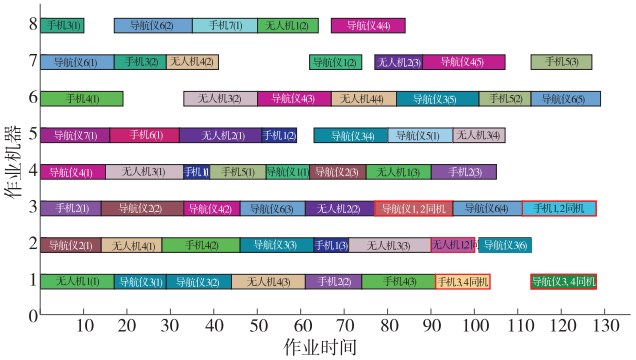


图 12　三类产品调度结果甘特图

Fig. 12　Gantt chart of scheduling of three types of products

4　结束语

本文探究利用自适应混合初始化遗传算法(AHIGA)求解考虑强制同机并行作业的广义作业车间调度(GJSPMCO)问题。但在实际生产应用中,存在强制同机并行作业、柔性同机并行作业和设备与工人耦合等多种约束影响,接下来将研究更符合车间实际情况的多约束广义作业车间调度问题。

参考文献:

[1]　张洁, 秦威. 智能制造调度为先—《制造系统智能调度方法与云

服务》导读[J]. 中国机械工程, 2019, 30(8): 1002-1007. (Zhang Jie, Qin Wei. Intelligent manufacturing scheduling first: a guide of manufacturing system intelligent scheduling method and cloud service[J]. China Mechanical Engineering, 2019, 30(8): 1002-1007.)

[2] Gao Liang, Pan Quanke. A shuffled multi-swarm micro-migrating birds optimizer for a multi-resource-constrained flexible job shop scheduling problem[J]. Information Sciences, 2016, 372: 655-676.

[3] Zhang Fayong, Li Rui, Gong Wenyin, *et al*. Deep reinforcement learning-based memetic algorithm for energy-aware flexible job shop scheduling with multi-AGV[J]. Computers & Industrial Engineering, 2024, 189: 109917.

[4] 郭鹏, 郝东辉, 郑鹏, 等. 考虑工人疲劳的双资源柔性作业车间调度优化[J]. 浙江大学学报: 工学版, 2023, 57(9): 1-10. (Guo Peng, Hao Donghui, Zheng Peng, *et al*. Scheduling optimization of dual resource-constrained flexible job shop considering worker fatigue[J]. Journal of Zhejiang University: Engineering Science, 2023, 57(9): 1-10.)

[5] 刘琼, 梅侦. 面向低碳的工艺规划与车间调度集成优化[J]. 机械工程学报, 2017, 53(11): 164-174. (Liu Qiong, Mei Zhen. Integrated optimization of process planning and shop scheduling for reducing manufacturing carbon emissions[J]. Journal of Mechanical Engineering, 2017, 53(11): 164-174.)

[6] Li Yufeng, He Yan, Wang Yulin, *et al*. An optimization method for energy-conscious production in flexible machining job shops with dynamic job arrivals and machine breakdowns[J]. Journal of Cleaner Production, 2020, 254: 120009.

[7] Gong Xu , Pessemier T D, Martens L, *et al*. Energy and labor-aware flexible job shop scheduling under dynamic electricity pricing: a many-objective optimization investigation[J]. Journal of Cleaner Production, 2019, 209: 1078-1094.

[8] Ku Wenyang, Beck J C. Mixed integer programming models for job shop scheduling: a computational analysis[J]. Computer & Operations Research, 2016, 73: 165-173.

[9] Ozolins A. Bounded dynamic programming algorithm for the job shop problem with sequence dependent setup times[J]. Operational Research, 2020, 20: 1701-1728.

[10] Meng Leilei, Duan Peng, Gao Kaizhou, *et al*. MIP modeling of energy-conscious FJSP and its extended problems: from simplicity to complexity[J]. Expert Systems with Applications, 2024, 241: 122594.

[11] 李佳磊, 顾幸生. 双种群混合遗传算法求解具有预防性维护的分布式柔性作业车间调度问题[J]. 控制与决策, 2023, 38(2): 475-482. (Li Jialei, Gu Xingsheng. Two-population hybrid genetic algorithm for distributed flexible job-shop scheduling problem with preventive maintenance[J]. Control and Decision, 2023, 38(2): 475-482.)

[12] 张国辉, 闫少峰, 陆熙熙, 等. 改进混合多目标蚁群算法求解带运输时间和调整时间的柔性作业车间调度问题[J]. 计算机应用研究, 2023, 40(12): 3690-3695. (Zhang Guohui, Yan Shaofeng, Lu Xixi, *et al*. Improved hybrid multi-objective ant colony optimization for flexible job-shop scheduling problem with transportation time and setup time[J]. Application Research of Computers, 2023, 40(12): 3690-3695.)

[13] 董君, 叶春明. 新型教与同伴学习粒子群算法求解作业车间调度问题[J]. 计算机应用研究, 2019, 36(12): 3764-3768. (Dong Jun, Ye Chunming. Novel teaching and peer-learning-based particle swarm optimization for job-shop scheduling problem[J]. Application Research of Computers, 2019, 36(12): 3764-3768.)

[14] Zhang Pengyu, Song Shiji, Niu Shengsheng, *et al*. A hybrid artificial immune-simulated annealing algorithm for multiroute job shop scheduling problem with continuous limited output buffers[J]. IEEE Trans on Cybernetics, 2022, 52(11): 12112-12125.

[15] 王雷, 邹新. 基于改进免疫克隆选择算法的柔性作业车间调度[J]. 南京理工大学学报, 2018, 42(3): 345-351. (Wang Lei, Zou Xin. Flexible job-shop scheduling based on improved immune clone selection algorithm[J]. Journal of Nanjing University of Science and Technology, 2018, 42(3): 345-351.)

[16] 吴锐, 郭顺生, 李益兵, 等. 改进人工蜂群算法求解分布式柔性作业车间调度问题[J]. 控制与决策, 2019, 34(12): 2527-2536. (Wu Rui, Guo Shunsheng, Li Yibing, *et al*. Improved artificial bee colony algorithm for distributed and flexible job-shop scheduling problem[J]. Control and Decision, 2019, 34(12): 2527-2536.)

[17] 杜凌浩, 向凤红. 改进多邻域候鸟优化算法的柔性作业车间调度研究[J]. 兵器装备工程学报, 2022, 43(12): 299-306. (Du Linghao, Xiang Fenghong. Research on flexible job shop scheduling based on improved multi-neighborhood migratory bird optimization algorithm[J]. Journal of Ordnance Equipment Engineering, 2022, 43(12): 299-306.)

[18] 杨冬婧, 雷德明. 新型蛙跳算法求解总能耗约束 FJSP[J]. 中国机械工程, 2018, 29(22): 2682-2689. (Yang Dongjing, Lei Deming. A novel shuffled frog-leaping algorithm for FJSP with total energy consumption constraints[J]. China Mechanical Engineering, 2018, 29(22): 2682-2689.)

[19] Jiang Tianhua, Zhang Chao. Application of grey wolf optimization for solving combinatorial problems: job shop and flexible job shop scheduling cases[J]. IEEE Access, 2018, 6: 26231-26240.

[20] 牛昊一, 吴维敏, 章庭棋, 等. 自适应樽海鞘群算法求解考虑运输时间的柔性作业车间调度[J]. 浙江大学学报: 工学版, 2023, 57(7): 1267-1277. (Niu Haoyi, Wu Weimin, Zhang Tingqi, *et al*. Adaptive salp swarm algorithm for solving flexible job shop scheduling problem with transportation time[J]. Journal of Zhejiang University: Engineering Science, 2023, 57(7): 1267-1277.)

[21] 王无双, 骆淑云. 基于强化学习的智能车间调度策略研究综述[J]. 计算机应用研究, 2022, 39(6): 1608-1614. (Wang Wushuang, Luo Shuyun. Research on intelligent shop scheduling strategies based on reinforcement learning[J]. Application Research of Computers, 2022, 39(6): 1608-1614.)

[22] Chen Ronghua, Yang Bo, Li Shi, *et al*. A self-learning genetic algorithm based on reinforcement learning for flexible job-shop scheduling problem[J]. Computers & Industrial Engineering, 2020, 149: 106778.

[23] 屈新怀, 王娇, 丁必荣, 等. 贪婪初始种群的遗传算法求解柔性作业车间调度[J]. 合肥工业大学学报: 自然科学版, 2021, 44(9): 1153-1156,1171. (Qu Xinhuai, Wang Jiao, Ding Birong, *et al*. Genetic algorithm of greedy initial population to solve flexible job-shop scheduling[J]. Journal of Hefei University of Technology: Natural Science, 2021, 44(9): 1153-1156,1171.)

[24] 黄学文, 陈绍芬, 周阅玉, 等. 求解柔性作业车间调度的遗传算法综述[J]. 计算机集成制造系统, 2022, 28(2): 536-551. (Huang Xuewen, Chen Shaofen, Zhou Tianyu, *et al*. Survey on genetic algorithms for solving flexible job-shop scheduling problem[J]. Computer Integrated Manufacturing Systems, 2022, 28(2): 536-551.)

[25] Brandimarte P. Routing and scheduling in a flexible job shop by Tabu search[J]. Annual Operation Research, 1993, 41: 157-183.

[26] Bagheri A, Zandieh M, Mahdavi I, *et al*. An artificial immune algorithm for the flexible job-shop scheduling problem[J]. Future Generation Computer Systems, 2010, 26(4): 533-541.

ISSN 1001-3695
CODEN JYYIC7

计算机应用研究

Application Research of Computers

第40卷第5期 2023年5月
Vol.40 No.5 May 2023

5

2023

四川省计算机研究院主办
中国计算机学会会刊

计算机应用研究
第四十卷第五期 二〇二三年五月

❖ 英国《科学文摘》(INSPEC)来源期刊 ❖《日本科学技术振兴机构数据库》(JST)来源期刊
❖ 俄罗斯《文摘杂志》(AJ)来源期刊 ❖ 美国《艾博思科学术数据库》(EBSCO)全文来源期刊
❖ 美国《乌利希期刊指南(网络版)》(Ulrichsweb)收录期刊 ❖ 美国《剑桥科学文摘(自然科学)》(CSA(NS))核心期刊
❖ 2017—2019年中国国际影响力优秀学术期刊(自然科学与工程技术) ❖ 波兰《哥白尼索引》(IC)来源期刊
❖ 第二届国家期刊奖百种重点科技期刊 ❖ 中国科学引文数据库(CSCD)来源期刊
❖ 中国科技核心期刊 ❖ 全国中文核心期刊 ❖《中文科技期刊数据库》来源期刊
❖ 中国科技论文统计源期刊 ❖《中国期刊网》《中国学术期刊(光盘版)》来源期刊
❖ 中国学术期刊综合评价数据库来源期刊 ❖ 中国精品科技期刊顶尖学术论文(F5000)项目来源期刊
❖ RCCSE核心学术期刊 ❖中国期刊方阵双效期刊 ❖《电子科技文献数据库》来源期刊
❖《中国工程技术电子信息网》来源期刊

计算机应用研究
Jisuanji Yingyong Yanjiu
第40卷 第5期 2023年5月

目 次

系统应用开发

网络与通信技术

信息安全技术

图形图像技术

信息集萃

期刊基本参数：CN51-1196/TP* 1984* m* A4* 320* zh* P* ￥50.00* 3840* 49* 2023-05* n

本期责任编辑　黄　莉，何　俐

基于 GENI-SD 的定制化印制电路板工序重要性评估 *

劳景春，金　鸿，吕盛坪†，李文强

(华南农业大学 工程学院，广州 510642)

摘　要：精准评估影响定制化印制电路板质量的关键工序可更好地指导企业精益管理产品品质，但企业常用方法难以利用工序关联实体间的深层语义。为解决上述问题，构建工序关联实体知识图谱并提出 GENI-SD 模型评估工序重要性。首先，使用知识图谱表征工序关联实体间的关系，并采用图神经网络模型 GENI 对工序节点进行重要性评估；然后，引入基于改进谓词感知注意力机制的采样模块和考虑邻边方向的分数聚合，改进 GENI 容易聚合到噪声邻节点且未考虑邻边方向的不足，建立工序重要性评估新模型 GENI-SD。最后，以 PCB 车间真实数据开展实验验证，结果显示 GENI-SD 优于其他对比模型，且得到的 Spearman 和 NDCG@10 指标值较 GENI 所得结果分别提高 6.78%和 0.71%。该研究为工序重要性评估提供了新的有效方法。

关键词：印制电路板；知识图谱；图神经网络；采样；分数聚合

中图分类号：TP183　　doi: 10.19734/j.issn.1001-3695.2022.10.0531

Estimating operation importance of customized printed circuit board based on GENI-SD

Lao Jingchun, Jin Hong, Lyu Shengping†, Li Wenqiang

(*College of Engineering*, *South China Agricultural University*, *Guangzhou* 510642, *China*)

Abstract: Accurately estimating the critical operations that affect the quality of customized printed circuit board (PCB) can better guide enterprises in lean management of product quality. However, the common methods of enterprises are difficult to use the deep semantics between operation-related entities. To solve this problem, this paper constructed a knowledge graph of operation-related entities and proposed GENI-SD to estimate the importance of PCB operation. Firstly, this paper employed the knowledge graph to represent the relationship between operation-related entities, and utilized the graph neural network model GENI to estimate the importance of operation nodes. Secondly, this paper improved GENI and established a novel model called as GENI-SD by introducing a sampling module based on an improved predicate-aware attention mechanism and a score aggregation considering the direction of adjacent edges, thus to resolve noisy neighbors and non-direction aggregation shortcoming of GENI. Finally, experiments were conducted on the real data of PCB workshop. The results demonstrate that GENI-SD outperforms SOTA and obtains 6.78% and 0.71% improvement in Spearman and NDCG@10 metrics comparing to that of GENI, respectively. This study provides a novel and effective method for estimating the importance of operation.

Key words: printed circuit board; knowledge graph; graph neural network; sampling; score aggregation

0　引言

印制电路板(printed circuit board，PCB)常被称为“电子产品之母”，是现代电子信息产品中不可或缺的电子元器件[1]。PCB 生产制造过程复杂、工艺流程长、涉及机械、电处理、光化学等多种加工工序。定制化 PCB 的生产更加复杂，具有个性化功能需求的样板常采用不同的工艺路线制造，工序多样且具有一定不确定性。确定影响定制化 PCB 质量的关键工序可更好地指导企业改进工艺、减少报废、提升品质。

当前，企业主要通过人工判断或统计各工序报废率评估 PCB 各工序重要性。但人工判断和各工序报废率统计时报废面积的责任工序指定都依赖于专家经验，且其难以利用工序关联实体间的深层语义。为此，本研究考虑构建工序关联实体知识图谱(knowledge graph，KG)表征工序与 PCB 之间复杂的语义关系，将工序重要性评估任务转换为 KG 中工序节点的重要性评估问题，利用图神经网络(graph neural network，GNN)模型在图数据处理上的优势[2]实现定制化 PCB 工序重要性评估。

KG 是由节点和边组成的语义图，节点是概念和实体的抽象，边表示节点之间的语义关系[3]。KG 因能够描述现实中实体以及实体间的关系，已被成功应用到农业[4~6]、生物医学[7~9]、推荐[10~14]和制造[15~17]等领域中。节点重要性评估是根据图结构和属性等信息推断图中节点的重要性，目前已经提出了多种评估节点重要性的方法。度中心性(Degree Centrality，

DC)[18]认为一个节点的邻节点数量越多，其重要性越大，这是描述节点重要性最简单的指标，但 DC 仅利用了节点最局部的邻居信息，没有考虑节点在图中的位置。与 DC 不同，介数中心性(betweenness centrality，BC)[19]和接近中心性(closeness centrality，CC)[20]从图全局角度描述节点的重要性，BC 考虑的是最短路径数量，认为经过一个节点的最短路径越多，该节点也就越重要;而 CC 考虑的是最短距离，认为一个节点与其他节点的平均最短距离越小，该节点越重要。BC 和 CC 作为全局中心性指标，充分考虑了节点在拓扑位置的重要性。上述三个方法只关注了节点的局部或全局中心性，忽略了重要性信息的传递在评估节点重要性的作用。PageRank (PR)[21]采用“随机游走”方式遍历图结构中每个节点，对于任意节点，其节点分数值以指定概率从邻近节点获取，并以另一指定概率从随机节点获取，PR 同时考虑了图的拓扑结构和节点重要性分数信息的传递，但难以利用某些节点先验重要性分数，Personalized PageRank(PPR)[22]引入节点先验重要性分数来解决 PR 的不足。然而，上述方法很难直接用于评估复杂 KG 中节点的重要性，因为这些方法重点都放在图的拓扑结构，而忽略了 KG 所包含丰富的语义信息。近年来，随着图深度学习的不断发展，一个基于 GNN 的模型 GENI(GNN for estimating node importance)[23]被开发出来，该模型在监督学习的框架下利用 KG 信息推断节点重要性，将节点特征映射为初始重要性分数，并使用注意力机制自适应地聚合信息，得益于监督学习框架和注意力机制，GENI 在节点重要性评估上有更好的性能。

GENI 可用于复杂 KG 中节点的重要性评估，但在工序重要性评估中发现该模型存在两个不足：第一，在实际中，各工序节点常关联成千上万的 PCB 节点，其中有不少噪声节点，GENI 在分数聚合过程使用当前节点所有邻节点分数，这使得 GENI 易聚合一些非重要的噪声邻节点信息。第二，工序与关联实体间的语义关联信息主要体现在 KG 边的关系类型和方向中，GENI 在分数聚合时未能考虑邻边的方向，会造成一定程度的信息丢失。针对 GENI 的两个不足，本研究提出 GENI-SD(GENI with sampling and direction)模型。首先，根据帕累托原理，重要的邻节点往往只占很小的比例[24]，GENI-SD 在 GENI 上添加了基于改进谓词感知注意力机制的采样模块，在每次分数聚合前对工序关联实体 KG 中节点进行采样，以便更好聚焦强关联邻节点、消除弱相关邻节点带来的噪声。其次，使用考虑邻边方向的分数聚合代替原分数聚合，聚合时通过引入不同向量描述 KG 中节点信息传播方向(连入、连出、自环)，使模型能感知工序与关联实体间边的方向信息。本研究的主要贡献如下：a)本研究提出使用 KG 表征工序关联实体间的关系，利用 GNN 模型对工序节点进行重要性评估，解决传统 PCB 工序重要性评估方法依赖专家经验且无法充分利用工序关联实体间深层语义信息的问题;b)针对节点重要性评估模型 GENI 的不足进行改进。引入了基于改进谓词感知注意力机制的采样模块，通过注意力机制为 KG 节点筛选重要邻节点，减少噪声邻节点信息对模型性能的影响；提出考虑邻边方向的分数聚合，使模型可以感知边方向，能够捕获 KG 中更丰富的信息。

1 知识图谱构建与 GENI-SD 模型

1.1 面向工序重要性评估的知识图谱构建

KG 可形式化表示为 $KG=\{V,E,P\}$，其中 V 代表 KG 中实体节点的集合，E 表示带有关系类型的边集合，P 是关系类型(谓词)的集合，每条边都属于一个唯一的谓词。在构建定制化 PCB 工序重要性评估 KG 时，其节点包括 PCB 节点和工序节点，其中每个 PCB 节点对应一个 PCB 订单记录，每个工序节点对应 PCB 一道工序。边上的关系类型指定工序(比如电镀、阻焊等)节点与 PCB 节点之间的不同关系。

图 1 为 PCB 工序重要性评估 KG 示意图，其中工序 A、B、C 等代表不同工序实体，$PCB1\sim PCB6$ 代表不同 PCB 实体；PCB 实体和不同工序实体之间不同类型的关系使用不同类型的边表示，图中还给出了部分 PCB 节点的重要性分数 s(图中节点左上角数值)。KG 中节点重要性分数 s 表示节点的重要程度，在此以定制化 PCB 订单相应报废率作为 KG 中 PCB 节点重要性分数；与工序关联的报废 PCB 订单数作为工序节点的重要性分数，这些分数可以相互比较以反映节点的重要性。

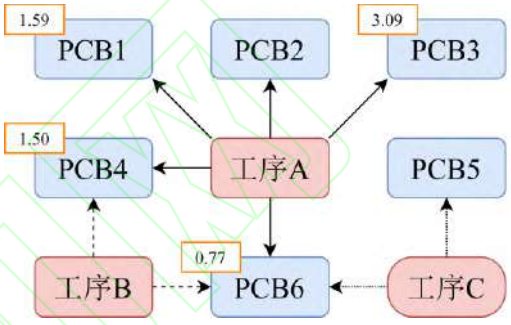


图 1 PCB 工序重要性评估知识图谱示意图

Fig. 1 Schematic of KG for PCB operation importance estimation

对于给定 KG 和节点集合子集 $V_s\subseteq V$ 的重要性分数 $\{s\}$，节点重要性评估的主要任务是学习一个函数 $F:V\rightarrow R$ 用来对 KG 中每个节点的重要性分数进行预测，其中 R 为实数。再按重要性分数评估各节点重要性，即重要性分数越大代表该节点越重要。KG 中有不同节点类型(如图 1 中有工序节点和 PCB 节点)，因此对节点的重要性评估也分为域内评估和域外评估。给定 T 类型节点 $V_s\subseteq V$ 的重要性分数，对 T 类型节点的重要性评估称为域内评估，而对非 T 类型节点重要性的评估称为域外评估。所构建的 PCB 工序重要性评估 KG 将作为所提出工序重要性评估模型 GENI-SD 的输入。

1.2 GENI-SD 模型

GENI 是 KG 中节点重要性评估的 GNN 模型，简单结构主要由节点嵌入、分数初始化、分数聚合头和中心调整四个部分构成。节点嵌入使用 node2vec[25]为 KG 的每个节点生成一个特征向量。分数初始化通过全连接神经网络将节点的特征向量映射为初始分数。分数聚合头基于谓词感知注意力机制对节点进行自适应分数聚合。中心调整综合考虑节点中心性和预测分数，输出最终节点的预测重要性分数。GENI 被扩展成更通用的结构，以(多层)分数聚合层替代分数聚合头，且在每个分数聚合层中设置多个分数聚合头，各聚合头彼此独立地进行注意力计算和分数聚合；最后通过中间聚合或最终聚合模块将层内分数聚合头输出的平均作为该聚合层的输出。

但 GENI 在工序重要性评估时容易引入噪声且未能考虑工序与 PCB 实体间关联的方向信息。为此，基于 GENI 通用结构提出 GENI-SD 模型，整体框架如图 2 所示。主要改进之处如图 2 中的基于谓词感知注意力机制的采样模块和考虑邻边方向的分数聚合层。采样模块基于改进谓词感知注意力机制为各分数聚合层中分数聚合头的节点筛选出重要邻节点。考虑邻边方向的分数聚合层在各分数聚合头的注意力系数计算中采用改进的谓词感知注意力机制以捕获 KG 中节点信息

传播方向。而节点嵌入、分数初始化、中间聚合、中心调整和最终聚合与 GENI 相同。

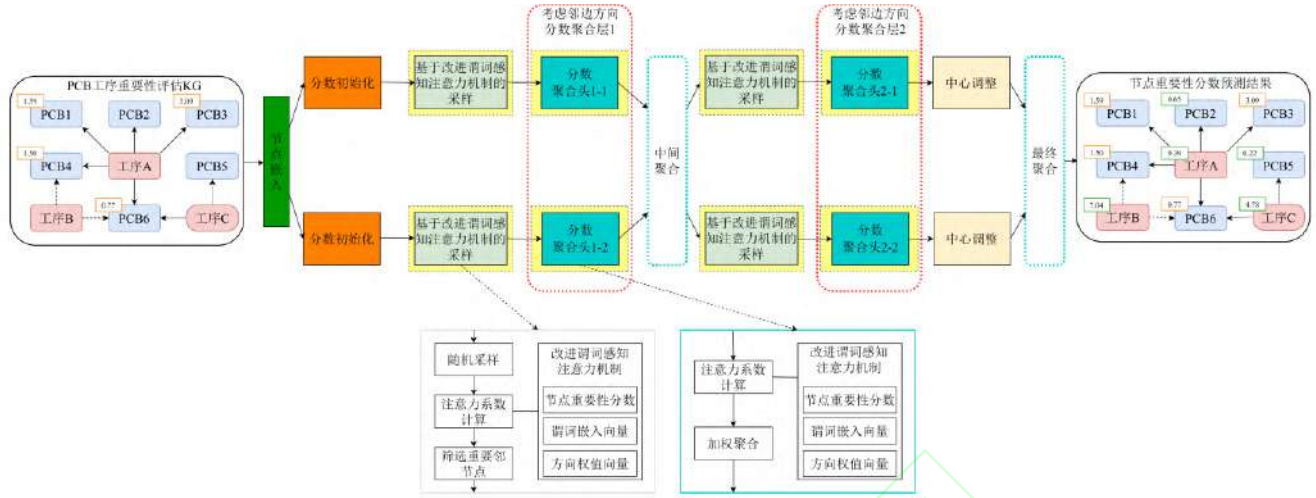


图 2　GENI-SD 框架

Fig. 2　Framework of GENI-SD

1.2.1　基于改进谓词感知注意力机制的采样

GENI 在分数聚合过程使用当前节点所有邻节点分数，这种不加区分地聚合所有邻节点信息的方式容易在分析工序重要性时引入噪声信息。为解决该问题，在分数聚合前添加采样模块，采样过程中更有针对性地选择相对重要的邻节点以使模型能聚合更有价值的邻节点。为筛选出重要邻节点，本文提出了一种基于改进谓词感知的注意力机制衡量节点的重要性。

本文提出的注意力机制考虑三个因素。首先考虑的是节点自身的信息，即节点的重要性分数。其次，要考虑节点间的谓词，不同的谓词可以在信息传播中扮演不同的角色，通过使用共享谓词嵌入将谓词合并到注意力计算中。最后要考虑的是节点间边的方向，KG 边的方向往往也带有意义，通过引入维度相同的不同权值向量来描述不同的方向信息，从而将信息传播的方向考虑进来。

使用 $s(i)$ 和 $s(j)$ 表示节点 i 与 j 的重要性分数，p_{ij}^{m} 表示节点 i 与邻节点 j 之间第 m 条边的谓词，$\phi(\cdot)$ 表示从谓词到其嵌入的映射，$\boldsymbol{a}_{m,i}^{T}$ 表示方向权值向量。节点 i 与邻节点 j 之间的注意力系数公式如式(1)所示。

$$e_{ij}=\sigma_a\left(\sum_m \boldsymbol{a}_{m,i}^{T}\left[s(i)\|\phi\left(p_{ij}^{m}\right)\|s(j)\right]\right) \tag{1}$$

其中，$\|$ 表示拼接操作，σ_a 为 Leaky ReLU 函数，$N(i)$ 为节点 i 的邻节点集合，$\boldsymbol{a}_{m,i}^{T}$ 根据式(2)进行取值。

$$\boldsymbol{a}_{m,i}^{T}=\begin{cases}\boldsymbol{a}_{in}^{T}, & \text{边}m\text{对节点}i\text{为连入}\\ \boldsymbol{a}_{out}^{T}, & \text{边}m\text{对节点}i\text{为连出}\\ \boldsymbol{a}_{self}^{T}, & \text{边}m\text{对节点}i\text{为自环}\end{cases} \tag{2}$$

式(2)中 $\boldsymbol{a}_{in}^{T}$、$\boldsymbol{a}_{out}^{T}$ 和 $\boldsymbol{a}_{self}^{T}$ 代表三个维度相同的不同可学习权值向量。为了使系数在不同邻节点之间容易比较，采用式(3)所示的 Softmax 函数对注意力系数进行归一化，注意力系数越高说明该邻节点对当前节点越重要。

$$\alpha_{ij}=\frac{\exp\left(e_{ij}\right)}{\sum_{k\in N(i)}\exp\left(e_{ik}\right)} \tag{3}$$

采样过程中，基于改进谓词感知的注意力机制充分利用节点信息、谓词类型和信息传播方向来衡量不同邻节点对当前节点的重要程度，为 KG 节点筛选重要邻节点，剔除非重要的噪声邻节点。在分析 PCB 工序重要性时，各工序节点常关联成千上万的 PCB 节点，其中有不少噪声节点，若在聚合时考虑所有邻节点，模型性能会受到影响，提出的采样模块可使模型聚焦于重要邻节点，有效避免聚合时易引入噪声邻节点的问题。

以节点 i 作为当前节点为例，采样具体过程如图 3 所示，主要包括随机抽样、注意力系数计算和样本邻节点选取三个主要步骤。首先，对每个节点的邻节点进行随机抽样，从邻节点集合 $N(i)$ 中随机抽样 b 个邻节点，若邻节点数量小于 b，则需要重复随机抽样，以确保每个节点所抽样的邻节点数量相同。然后，基于改进谓词感知注意力机制公式计算所选取的各邻节点的注意力系数，以判断邻节点对当前节点的重要性。最后，筛选出注意力系数最高的 u 个邻节点作为当前节点的样本邻节点集合 $N'(i)$。

$$N'(i)=\{t\,|\,t\in N(i)\},|\,N'(i)\,|=u \tag{4}$$

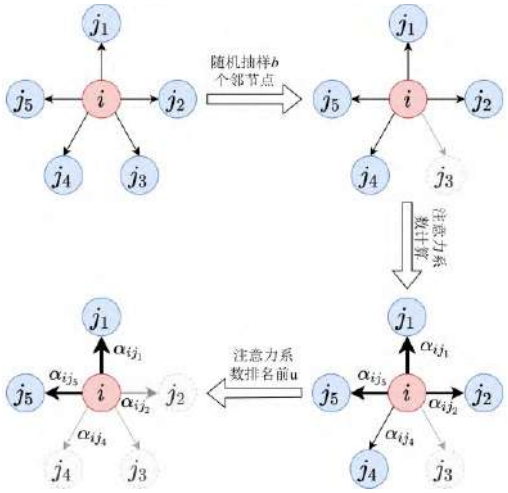


图 3　基于改进谓词感知注意力机制的采样过程

Fig. 3　Sampling based on improved predicate-aware attention mechanism

1.2.2　考虑邻边方向的分数聚合

GENI 在分数聚合时，通过基于谓词感知注意力机制自适应聚合邻节点分数，在计算注意力系数时只考虑了节点分数和谓词信息，忽略了 KG 节点间边的方向。受关系图卷积网络(Relational Graph Convolutional Networks, RGCN)[20]模型在信息聚合时为不同方向引入不同权值向量的启发，本文在

基于谓词感知注意力机制中通过引入维度相同的不同权值向量描述不同方向信息。具体来说，根据边方向的不同分别使用相应的权值向量对邻节点进行注意力系数计算，再通过 Softmax 函数对系数进行归一化，最后使用归一化的系数对邻节点分数进行加权聚合。对节点 i 进行分数聚合相应示意图如图 4 所示。

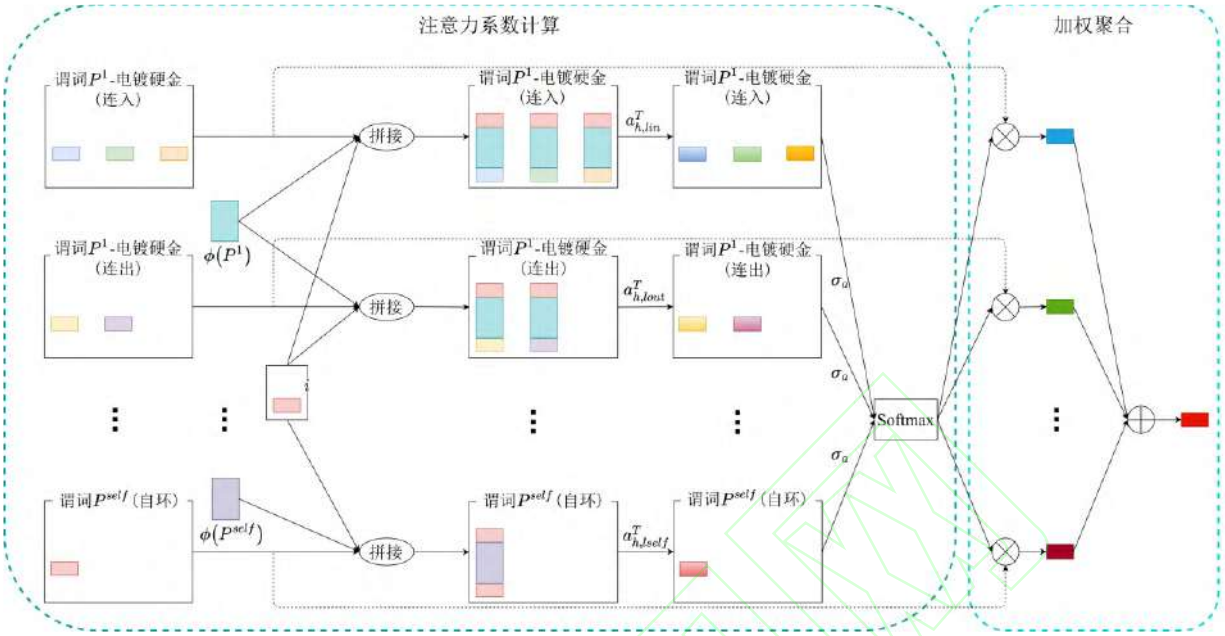


图 4　考虑邻边方向分数聚合示意图

Fig. 4　Schematic of score aggregation considering adjacent edge direction

在有多层分数聚合层，每层有多个分数聚合头的 GENI-SD 模型中，使用 h 表示分数聚合头索引，H^l 为第 l 层的分数聚合头数量，定义 $s_h^{l-1}(i)$ 为第 $l-1$ 层节点 i 的预测分数，对式(1)至式(3)进行扩展得到第 l 层分数聚合层第 h 个分数聚合头中节点 i 与 j 间的注意力系数计算公式：

$$\alpha_{ij}^{h,l}=\frac{\exp\left(\sigma_a\left(\sum_m \boldsymbol{a}_{h,l\_m,i}^T\left[s_h^{l-1}(i)\|\phi\left(p_{ij}^m\right)\|s_h^{l-1}(j)\right]\right)\right)}{\sum_{k\in N'(i)\cup\{i\}}\exp\left(\sigma_a\left(\sum_m \boldsymbol{a}_{h,l\_m,i}^T\left[s_h^{l-1}(i)\|\phi\left(p_{ik}^m\right)\|s_h^{l-1}(k)\right]\right)\right)} \tag{5}$$

其中：$N'(i)$ 表示采样后节点 i 的邻节点集合，$\boldsymbol{a}_{h,l\_m,i}^T\in\{\boldsymbol{a}_{h,lin}^T,\boldsymbol{a}_{h,lout}^T,\boldsymbol{a}_{h,lself}^T\}$ 根据边 m 对节点 i 方向的不同选择维度相同的不同可学习权值向量，第 l 层分数聚合层第 h 个分数聚合头节点 i 的分数可根据下面的公式加权聚合：

$$s_h^l(i)=\sum_{j\in N'(i)\cup\{i\}}\alpha_{ij}^{h,l}s_h^{l-1}(j) \tag{6}$$

在第一层分数聚合层中，每个分数聚合头单独从全连接神经网络的输出得到各节点的初始分数，对于后面分数聚合层的分数聚合头，都把来自前一层的输出作为输入，最后以层内分数聚合头输出的平均作为该聚合层的输出。第 l 层分数聚合层节点 i 的输出 $s_h^l(i)$ 计算公式如下：

$$s_h^l(i)=\begin{cases}FCNN_h(\boldsymbol{Z}_i), & l=0\\ AVERAGE(\{s_h^l(i)|h=1,\ldots,H^l\}), & l\geq 1\end{cases} \tag{7}$$

其中 $\boldsymbol{Z}_i$ 是节点 i 的特征向量。

在 KG 中，边方向表示信息从主体传播至客体，例如在构建的工序关联实体 KG 中，边由工序节点指向 PCB 节点，表示工序对 PCB 进行加工，边方向是不能忽视的重要信息，考虑边方向可带来更多有价值的信息。基于改进谓词感知的注意力机制在考虑节点分数信息和谓词类型的基础上，为不同方向引入不同权值向量，使其能够感知边方向，充分利用边方向信息也提高了模型可用的信息量，使模型的分数聚合更加准确可靠。

2　实验验证

2.1　实验数据

本研究实验数据来源于广州某 PCB 生产企业资源管理系统和制造执行系统，抽取下单日期在 2019 年 1 月 2 日~2020 年 4 月 1 日之间的 16926 条记录构建 KG，为每条 PCB 订单记录创建一个 PCB 节点，选取 15 个工序(如表 1 所示)并为每个工序创建一个工序节点，以工序对 PCB 加工方式作为工序节点与 PCB 节点间边的关系类型，用三元组(工序，加工方式，PCB)格式存储构建的 KG。

表 1　工序名称

Tab. 1　Operation name

序号	工序	序号	工序
1	沉银	9	沉金
2	电镀软金	10	字符
3	沉铜	11	沉锡
4	压合	12	阻焊塞孔
5	减薄铜工序	13	钻孔
6	树脂塞孔	14	喷锡表面处理
7	有机涂覆	15	化学镍钯金
8	电镀硬金		

将各 PCB 订单的报废率作为 PCB 节点重要性分数，与工序关联的报废 PCB 记录数作为工序节点的重要性分数。为减少节点重要性分数差异过大对模型构建的不利影响，对节点重要性分数进行 ln(加偏移量 1 以保证重要性分数为非负数)变换。

在实验中，将 PCB 节点重要性分数按 7:3 比例划分为训练集和测试集，使用 PCB 节点重要性分数进行模型训练和域内评估，所有工序节点重要性分数用于域外评估。数据集统计信息如表 2 所示。

2.2　评价指标

为评估所提出模型的预测精度，使用斯皮尔曼相关系数

(Spearman Correlation Coefficient, Spearman)和归一化折损累积增益(Normalized Discounted Cumulative Gain, NDCG)作为模型的评价指标。

表 2　实验数据集

Tab. 2　Experimental dataset

术语	数量	术语	数量
节点	16941	训练集	11848
关系类型(谓词)	15	测试集(域内评估)	5078
边	41567	测试集(域外评估)	15

Spearman 衡量的是节点真实重要性分数排名与预测重要性分数排名之间的相关性，即真实排名 $\boldsymbol{g}$ 与预测排名 $\boldsymbol{s}$ 之间的强度和方向，其取值范围为[-1,1]，值越接近 1 说明排序效果越好，计算公式为

$$Spearman=\frac{\sum_i(g_{ri}-\bar{g}_r)(s_{ri}-\bar{s}_r)}{\sqrt{\sum_i(g_{ri}-\bar{g}_r)^2}\sqrt{\sum_i(s_{ri}-\bar{s}_r)^2}} \tag{8}$$

其中 g_{ri} 和 s_{ri} 分别为节点 i 的真实重要性分数排名和预测重要性分数的排名，$\bar{g}_r$ 和 $\bar{s}_r$ 分别为 $\boldsymbol{g}$ 和 $\boldsymbol{s}$ 的平均值。

NDCG 是衡量排序质量的一个指标，其取值范围为[0,1]，值越大说明预测排序质量越好。给定一个按预测分数排序的节点列表以及节点对应的重要性分数，在位置 n 处的折损累积增益计算公式为

$$DCG@n=\sum_{i=1}^{n}\frac{r_i}{log_2(i+1)} \tag{9}$$

其中：r_i 表示排序位置 i 节点的重要性分数。将理想排序情况(即预测排序情况与真实排序情况相同)中前 n 位的理想折损累积增益记为 $IDCG@n$，则在位置 n 的归一化折损累积增益为

$$NDCG@n=\frac{DCG@n}{IDCG@n} \tag{10}$$

对于域内评估和域外评估，均使用 Spearman 和 NDCG@10 和作为评价指标，两评价指标之间相互补充，Spearman 考虑所有已知重要性分数的节点排名情况，而 NDCG@10 可评估预测排名前 10 实体的排名质量。本研究主要任务是对工序进行重要性评估，因此更关注域外评估的实验结果。

2.3　对比实验

为验证所提模型的有效性，与随机游走方法(PR、PPR)、非图监督方法(Random Forests[27], RF; Neural Networks, NN)以及基于图神经网络方法(Graph Attention Networks[28],GAT;RGCN [26];GENI[23])三类实现机制七种实现方法进行对比。PR 利用图的拓扑结构更新节点分数以确定各节点稳定分数值并将其作为节点预测分数值；PPR 在 PR 基础上引入节点先验重要性分数学习各节点分数值。RF 基于决策树的集成学习方法开展预测；NN 本研究使用一个多层全连接神经网络实现。GAT 使用图注意力机制计算节点间的注意力系数，根据注意力系数对邻节点信息进行加权聚合；RGCN 为不同关系类型引入专门的变换矩阵处理 KG 中复杂关系；GENI 将节点特征向量映射为初始分数，通过谓词感知注意力机制和中心调整对节点重要性分数进行预测。

节点特征向量维度设置为 128，在全连接神经网络结构、分数聚合层层数、每层的聚合头数和谓词向量维度上，GENI-SD 采用与 GENI 相同的设置，即用 128-96-1 结构的全连接神经网络初始化分数，分数聚合层其层数和聚合头数分别设置为 3 和 4，谓词向量维度设为 10。对于采样邻节点数(采样规模)，经过在数据集中的大量实验，最终设定为 8。模型迭代次数为 100，并使用均方误差作为损失函数。实验重复三次，取三次实验结果的平均值作为最终的实验结果。另外，观察到构建的 KG 中工序节点的邻节点从数百个到数万个不等，而 PCB 节点的邻节点只有少数几个，为避免采样过程中引入大量重复邻节点，GENI-SD 只对工序节点的邻节点进行采样处理。不同算法的实验结果如表 3 所示，最好的结果用粗体表示。

表 3　对比实验结果

Tab. 3　Comparative experimental results

模型	域外评估		域内评估	
	Spearman	NDCG@10	Spearman	NDCG@10
PR	-0.0929	0.8682	0.2202	0.7000
PPR	-0.0929	0.8682	0.1044	0.3952
RF	0.2893	0.9231	0.2541	0.6551
NN	-0.4667	0.8424	0.2524	0.5453
GAT	-0.1833	0.8415	0.3132	0.5712
RGCN	-0.3572	0.8294	0.3201	0.7064
GENI	0.4512	0.9254	0.3330	0.7075
GENI-SD	**0.5190**	**0.9325**	**0.3357**	**0.7159**

随机游走方法 PR 和 PPR 都仅在拓扑结构下考虑分数信息传递而忽略 KG 的语义信息，因此在各评价指标上都表现不好；PPR 虽然在 PR 基础上引入先验分数，效果却不如 PR，说明在 KG 信息不充分利用的情况下引入先验分数反而可能损害模型节点重要性评估性能。基于非图监督学习方法 RF 和 NN 仅利用节点特征向量来预测重要性分数，在实验结果上也无法达到好的效果，但 RF 总体表现要比 NN 好，原因可能是 RF 综合多个学习器预测结果。基于图神经网络方法 GENI 有效结合 KG 中节点和关系类型的信息进行推理，其性能在对比模型中表现最佳；另外两个基于图神经网络方法 GAT 和 RGCN 在各评价指标上都不如 GENI，原因是 GAT 和 RGCN 只关注 KG 中的部分信息，GAT 关注的是节点间的相关性，RGCN 关注的是关系类型，这表明 KG 信息使用不当也无法达到好的评估效果。本文提出的 GENI-SD 在所有评价指标中均表现最优，分别比 GENI 提升了 6.78%、0.71%、0.27%和 0.84%，GENI-SD 基于改进谓词感知注意力机制的采样模块使模型聚焦于重要的邻节点，考虑邻边方向的分数聚合能捕获 KG 更丰富的信息，这些改进使其具有更优的性能。

为更直观验证模型工序重要性评估的效果，以表 1 中所提的工序为案例，给出每个模型工序重要性的具体预测情况，通过具体的案例对模型性能进行分析。表 4 为各模型在一次实验中对工序重要性分数预测排名前 10 工序序号以及工序对应实际排名的情况。

可以看出，在预测的前 10 个重要工序中，NN 和 RGCN 只有 6 个工序在实际排名中也是前 10，预测工序实际排名前 10 的数量最少；而 PR、PPR、RF 和 GAT 有 7 个工序在实际排名中也是前 10，GENI-SD 和 GENI 有 9 个。虽然 GENI-SD 和 GENI 都有 9 个工序在实际排名中也是前 10，但 GENI-SD 预测前 10 个重要工序中实际排名最低为第 11 名，而 GENI 预测工序的实际排名最低为第 13 名，相对而言 GENI-SD 的评估结果比 GENI 更好。另一方面，在预测排名和实际排名的误差上，NN、GAT 和 RGCN 的平均绝对误差都超过 5.0，而 PR、PPR、RF、GENI 和 GENI-SD 平均绝对误差均不超过

4.0，其中 GENI-SD 的平均绝对误差为 3.2，在所有模型中表现最好。可以看出，GENI-SD 能预测到更多重要性排名靠前的工序，而且预测排名与实际排名的误差更小。总体而言，GENI-SD 在工序重要性评估任务上性能更优。

表 4　不同模型所得 PCB 工序重要性评估结果

Tab. 4　Estimation results of PCB operation importance for different models

模型		预测排名									
		1	2	3	4	5	6	7	8	9	10
PR	工序序号	1	2	3	4	5	6	7	8	9	10
	工序实际排名	9	11	5	3	14	6	10	8	7	12
	排名绝对误差	-8	-9	-2	1	-9	0	-3	0	2	-2
	平均绝对误差					3.6					
PPR	工序序号	1	2	3	4	5	6	7	8	9	10
	工序实际排名	9	11	5	3	14	6	10	8	7	12
	排名绝对误差	-8	-9	-2	1	-9	0	-3	0	2	-2
	平均绝对误差					3.6					
RF	工序序号	4	13	9	1	15	5	12	8	10	6
	工序实际排名	3	1	7	9	13	14	2	8	12	6
	排名绝对误差	-2	1	-4	-5	-8	-8	5	0	-3	4
	平均绝对误差					4.0					
NN	工序序号	5	2	10	1	4	9	15	7	8	12
	工序实际排名	14	11	12	9	3	7	13	10	8	2
	排名绝对误差	-13	-9	-9	-5	2	-1	-6	-2	1	8
	平均绝对误差					5.6					
GAT	工序序号	5	10	2	1	9	4	13	12	8	7
	工序实际排名	14	12	11	9	7	3	1	2	8	10
	排名绝对误差	-13	-10	-8	-5	-2	3	6	6	1	0
	平均绝对误差					5.4					
RGCN	工序序号	1	5	10	6	7	2	9	3	12	11
	工序实际排名	9	14	12	6	10	11	7	5	2	15
	排名绝对误差	-8	-12	-9	-2	-5	-5	0	3	7	-5
	平均绝对误差					5.6					
GENI	工序序号	7	14	6	12	15	3	8	9	13	1
	工序实际排名	10	4	6	2	13	5	8	7	1	9
	排名绝对误差	-9	-2	-3	2	-8	1	-1	1	8	1
	平均绝对误差					3.6					
GENI-SD	工序序号	7	6	3	12	14	1	9	8	2	13
	工序实际排名	10	6	5	2	4	9	7	8	11	1
	排名绝对误差	-9	-4	-2	2	1	-3	0	0	-2	9
	平均绝对误差					3.2					

2.4　消融实验

为检验 GENI-SD 采样和考虑邻边方向的分数聚合两个改进对模型性能均有正向作用，进行消融实验。用 GENI-S 表示 GENI 使用采样模块，GENI-D 表示 GENI 使用考虑邻边方向的分数聚合，实验结果如表 5 所示。

表 5　消融实验结果

Tab. 5　Ablation experiment results

模型	域外评估		域内评估	
	Spearman	NDCG@10	Spearman	NDCG@10
GENI	0.4512	0.9254	0.3330	0.7075
GENI-S	0.4738	0.9287	0.3343	0.7134
GENI-D	0.4655	0.9323	0.3343	0.7095
GENI-SD	**0.5190**	**0.9325**	**0.3357**	**0.7159**

可以看出，GENI-S 模型在域外评估中 Spearman 和 NDCG@10 与 GENI 相比分别提高了 2.26%和 0.33%，域内评估性能与 GENI 相比有所提升，这表明采样模块有针对性地选择相对重要的邻节点可为模型减少噪声数据，有效提高模型的准确性。GENI-D 模型在域外评估中 Spearman 和 NDCG@10 与 GENI 相比分别提高了 1.43%和 0.69%，域内评估性能与 GENI 相比也有所提升，说明考虑 KG 邻边方向能为模型提供更准确的 KG 信息，这有助于提高模型性能。综上所述，所提出的两个改进对模型性能的提升是有效的。

2.5　超参数敏感性实验

为探究采样规模和分数聚合层数对 GENI-SD 模型工序重要性评估效果的影响，通过调整相关超参数，控制其他超参数不变，进行超参数敏感性实验，分析不同超参数值对模型域外评估的影响。

采样规模在{2,4,8,16,32}中选择，实验结果如表 6 所示。通过实验分析，模型性能随采样规模增大先逐渐提升，当采样规模为 8 时，此时模型性能最优；当采样规模继续增大，模型性能反而下降，可能原因是其聚合到一些不重要的噪声邻节点。

分数聚合层数在{1,2,3,4,5}中选择，实验结果如表 7 所示。可以看出，随着层数的增加，模型性能也在不断提升，当分数聚合层层数为 3 时，模型达到最优性能；当层数继续增大时，模型的拟合能力增强了，但同时也增加了模型的复杂度，导致出现了过拟合现象，模型性能反而下降。

表 6　不同采样规模下 GENI-SD 实验结果

Tab. 6　Experimental results of GENI-SD under different sampling scales

采样规模	Spearman	NDCG@10
2	0.4643	0.9255
4	0.4714	0.9255
8	**0.5190**	**0.9325**
16	0.4976	0.9307
32	0.4583	0.9286

表 7　不同分数聚合层下 GENI-SD 实验结果

Tab. 7　Experimental results of GENI-SD under different score aggregation layers

分数聚合层数	Spearman	NDCG@10
1	0.3631	0.9231
2	0.5036	0.9317
3	**0.5190**	**0.9325**
4	0.4714	0.9282
5	0.4060	0.9267

3　结束语

本研究将 PCB 工序重要性评估任务转换为对 KG 中的工序节点重要性评估问题，针对 GENI 在 PCB 工序重要性评估中容易聚合到噪声邻节点且未考虑邻边方向的不足，分别引入基于改进谓词感知注意力机制的采样和考虑邻边方向的分数聚合对 GENI 进行改进。与 PR、PPR、RF、NN、GAT、RGCN、GENI 7 种方法进行了对比实验，结果显示 GENI-SD 在所有方法中表现最佳，且得到的 Spearman 和 NDCG@10 指标值较 GENI 所得结果分别提高 6.78%和 0.71%，证明了 GENI-SD 的可行性和优越性。此外，消融实验证明两个改进对模型性能均有正向作用，超参数敏感性实验为模型参数选择提供了指导。

下阶段将综合更多 PCB 样本并引入更多工序相关数据，构建更具泛化性的 KG；同时，将为各节点引入不同来源的重要性分数，综合考虑各节点多个重要性分数的情况。

参考文献：

[1] 郑彬彬, 吕盛坪, 李灯辉, 等. 基于自组织映射-反向传播网络的 PCB 样板投料预测 [J]. 计算机应用与软件, 2020, 37 (8) : 57-63. (Zheng Binbin, Lyu Shengping, Li Denghui, *et al.* PCB sample feeding prediction based on self-organizing maps and back propagation network [J]. Computer Applications and Software, 2020, 37 (8) : 57-63.)

[2] Zhou Jie, Cui Ganqu, Hu Shengding, *et al.* Graph neural networks: a review of methods and applications [J]. AI Open, 2020, 1: 57-81.

[3] Yan Jihong, Wang Chengyu, Cheng Wenliang, *et al.* A retrospective of knowledge graphs [J]. Frontiers of Computer Science, 2018, 12 (1): 55-74.

[4] 张善文, 王振, 王祖良. 结合知识图谱与双向长短时记忆网络的小麦条锈病预测 [J]. 农业工程学报, 2020, 36 (12): 172-178. (Zhang Shanwen, Wang Zhen, Wang Zuliang. Prediction of wheat stripe rust disease by combining knowledge graph and bidirectional long short term memory network [J]. Trans of the Chinese Society of Agricultural Engineering, 2020, 36 (12): 172-178.)

[5] 张海瑜, 陈庆龙, 张斯静, 等. 基于语义知识图谱的农业知识智能检索方法 [J]. 农业机械学报, 2021, 52 (S1): 156-163. (Zhang Haiyu, Chen Qinglong, Zhang Sijing, *et al.* Intelligent retrieval method of agricultural knowledge based on semantic knowledge graph [J]. Trans of the Chinese Society for Agricultural Machinery, 2021, 52 (S1): 156-163.)

[6] 于合龙, 沈金梦, 毕春光, 等. 基于知识图谱的水稻病虫害智能诊断系统 [J]. 华南农业大学学报, 2021, 42 (5) : 105-116. (Yu Helong, Shen Jinmeng, Bi Chunguang, *et al.* Intelligent diagnostic system for rice diseases and pests based on knowledge graph [J]. Journal of South China Agricultural University, 2021, 42 (5) : 105-116.)

[7] Li Linfeng, Wang Peng, Yan Jun, *et al.* Real-world data medical knowledge graph: construction and applications [J]. Artificial Intelligence in Medicine, 2020, 103 (19): 101817.

[8] Nicholson D N, Greene C S. Constructing knowledge graphs and their biomedical applications [J]. Computational and Structural Biotechnology Journal, 2020, 18: 1414-1428.

[9] Al-saleem J, Granet R, Ramakrishnan S, *et al.* Knowledge graph-based approaches to drug repurposing for covid-19 [J]. Journal of Chemical Information and Modeling, 2021, 61 (8): 4058-4067.

[10] Wang Hongwei, Zhang Fuzheng, Wang Jialin, *et al.* RippleNet: propagating user preferences on the knowledge graph for recommender systems [C]// Proc of the 27th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2018: 417-426.

[11] Wang Hongwei, Zhang Fuzheng, Xie Xing, *et al.* DKN: deep knowledge-aware network for news recommendation [C]// Proc of the 27th World Wide Web Conference. New York: ACM Press, 2018: 1835-1844.

[12] Wang Hongwei, Zhao Miao, Xie Xing, *et al.* Knowledge graph convolutional networks for recommender systems [C]// Proc of the 28th World Wide Web Conference. New York: ACM Press, 2019: 3307-3313.

[13] 陶天一, 王清钦, 付聿炜, 等. 基于知识图谱的金融新闻个性化推荐算法 [J]. 计算机工程, 2021, 47 (6) : 98-103. (Tao Tianyi, Wang Qingqin, Fu Yuwei, *et al.* Personalized recommendation algorithm for financial news based on knowledge graph [J]. Computer Engineering, 2021, 47 (6) : 98-103.)

[14] 汤伟韬, 余敦辉, 魏世伟. 融合知识图谱与用户评论的商品推荐算法 [J]. 计算机工程, 2020, 46 (8) : 93-100. (Tang Weitao, Yu Dunhui, Wei Shiwei. Commodity recommendation algorithm fusing with knowledge graph and user comment [J]. Computer Engineering, 2020, 46 (8) : 93-100.)

[15] 张栋豪, 刘振宇, 郑维强, 等. 知识图谱在智能制造领域的研究现状及其应用前景综述 [J]. 机械工程学报, 2021, 57 (5) : 90-113. (Zhang Donghao, Liu Zhenyu, Jia Weiqiang, *et al.* A review on knowledge graph and its application prospects to intelligent manufacturing [J]. Journal of Mechanical Engineering, 2021, 57 (5) : 90-113.)

[16] 陶家琦, 李心雨, 郑湃, 等. 制造领域知识图谱的应用研究现状与前沿 [J/OL]. 计算机集成制造系统: 1-32. (2022-03-29) [2022-09-11]. http://kns. cnki. net/kcms/detail/11. 5946. tp. 20220328. 1707. 013. html. (Tao Jiaqi, Li Xinyu, Zheng Pai, *et al.* State-of-the-art and frontier of manufacturing knowledge graph application [J/OL]. Computer Integrated Manufacturing Systems: 1-32. (2022-03-29) [2022-09-11]. http://kns. cnki. net/kcms/detail/11. 5946. tp. 20220328. 1707. 013. html.)

[17] 邱凌, 张安思, 李少波, 等. 航空制造知识图谱构建研究综述 [J].

计算机应用研究, 2022, 39 (4) : 968-977. (Qiu Ling, Zhang Ansi, Li Shaobo, *et al.* Survey on building knowledge graphs for aerospace manufacturing [J]. Application Research of Computers, 2022, 39 (4) : 968-977.)

[18] Freeman L C. Centrality in social networks conceptual clarification [J]. Social Networks, 1978, 1 (3): 215-239.

[19] Freeman L C. A set of measures of centrality based on betweenness [J]. Sociometry, 1977, 40 (1): 35-41.

[20] Sabidussi G. The centrality index of a graph [J]. Psychometrika, 1966, 31 (4): 581-603.

[21] Page L, Brin S, Motwani R, *et al.* The pagerank citation ranking: bringing order to the web [J]. Stanford Digital Libraries Working Paper, 1998.

[22] Haveliwala T H. Topic-sensitive pagerank [C]// Proc of the 11th International Conference on World Wide Web. New York: ACM Press, 2002: 517-526.

[23] Park N, Kan A, Dong X L, *et al.* Estimating node importance in knowledge graphs using graph neural networks [C]// Proc of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2019: 596–606.

[24] Zhang Zhenyu, Zhang Lei, Yang Dingqi, *et al.* KRAN: knowledge refining attention network for recommendation [J]. ACM Trans on Knowledge Discovery from Data, 2021, 16 (2): 39.

[25] Grover A, Leskovec J. Node2vec: scalable feature learning for networks [C]// Proc of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 855-864.

[26] Schlichtkrull M, Kipf T N, Bloem P, *et al.* Modeling relational data with graph convolutional networks [C]// Proc of the 15th European Semantic Web Conference. Berlin: Springer, 2018: 593-607.

[27] Breiman L. Random forests [J]. Machine Learning, 2001, 45 (1): 5-32.

[28] Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks [C]// Proc of the 6th International Conference on Learning Representations, 2018.

ISSN 1001-3695
CODEN JYYIC7

计算机应用研究

Application Research of Computers

第39卷 第4期 2022年4月
Vol.39 No.4 Apr. 2022

4

2022

四川省计算机研究院主办
中国 成都 Chengdu,China

计算机应用研究
第三十九卷第四期 二〇二二年四月

❖ 英国《科学文摘》(INSPEC)来源期刊
❖ 俄罗斯《文摘杂志》(AJ)来源期刊
❖ 美国《乌利希期刊指南(网络版)》(Ulrichsweb)收录期刊
❖ 2017—2019年中国国际影响力优秀学术期刊(自然科学与工程技术)
❖ 第二届国家期刊奖百种重点科技期刊
❖ 中国科技核心期刊 ❖ 全国中文核心期刊
❖ 中国科技论文统计源期刊
❖ 中国学术期刊综合评价数据库来源期刊
❖ RCCSE核心学术期刊 ❖中国期刊方阵双效期刊
❖《日本科学技术振兴机构数据库》(JST)来源期刊
❖ 美国《艾博思科学术数据库》(EBSCO)全文来源期刊
❖ 美国《剑桥科学文摘(自然科学)》(CSA(NS))核心期刊
❖ 波兰《哥白尼索引》(IC)来源期刊
❖ 中国科学引文数据库(CSCD)来源期刊
❖《中文科技期刊数据库》来源期刊
❖《中国期刊网》《中国学术期刊(光盘版)》来源期刊
❖ 中国精品科技期刊顶尖学术论文(F5000)项目来源期刊
❖《电子科技文献数据库》来源期刊
❖《中国工程技术电子信息网》来源期刊

计算机应用研究

Jisuanji Yingyong Yanjiu

第 39 卷 第 4 期 2022 年 4 月

目 次

综述评论

区块链技术

算法研究探讨

网络首发时间：2021-12-14 16:35:07
网络首发地址：https://kns.cnki.net/kcms/detail/51.1196.tp.20211213.1406.008.html

基于时间加权改进的 LDTW 算法 *

朱紫纯，吕盛坪†，廖鑫婷，江　城，罗　勇

(华南农业大学 工程学院，广州 510642)

摘　要：在时间序列相似性度量研究中，动态时间弯曲(dynamic time warping，DTW)是最为常用的算法之一，但其存在病态对齐问题且未考虑时间属性影响。限制对齐路径长度 DTW(DTW under limited warping path length，LDTW)和时间加权 DTW (time-weighed DTW，TDTW)分别尝试解决上述两个问题中的一个，但未能同时解决 DTW 两方面不足。本研究提出一种综合时间权重的 LDTW(time-weighting LDTW，TLDTW)算法：首先通过测量两个时间序列中时间点对的距离构建时间权值矩阵；然后在 LDTW 累计成本矩阵递归填充过程中融合对应的时间权值，以实现在考虑时间因素影响的同时保留有效抑制病态对齐特性。基于 UCR 数据集进行 1-NN 分类实验，实验结果显示基于 TLDTW 相似度量的分类准确率优于其他对比算法，且进一步对比验证了其可靠性。

关键词：时间序列；动态时间弯曲；病态对齐；时间加权；相似度度量

中图分类号：TP311　　doi: 10.19734/j.issn.1001-3695.2021.09.0401

Improved ldtw algorithm based on time-weighting

Zhu Zichun, LÜ Shengping†, Liao Xinting, Jiang Cheng, Luo Yong

(*College of Engineering*, *South China Agricultural University*, *Guangzhou* 510642, *China*)

Abstract: Dynamic time warping (DTW) is one of the commonly used algorithms in time series similarity measurement. However, DTW has the shortcoming of pathological alignment and ignores the influence of time attribute. Dynamic time warping under limited warping path length (LDTW) and time-weighed DTW (TDTW) have been proposed to handle that two shortcomings of DTW separately, however they cannot be solved simultaneously by LDTW or TDTW independently. In this paper, Time-Weighting LDTW (TLDTW) algorithm was proposed. Firstly, time weight matrix was constructed by measuring the distance between points in two series. Secondly, the corresponding time weights from time weight matrix were fused into the recursive filling procedure for cumulative cost matrix of LDTW; thus the time attribute was considered and the problem of pathological alignment can still be suppressed. Lastly, 1-NN classification experiment based on UCR dataset was conducted, and experimental results shown that the classification accuracy based on TLDTW is better than other compared algorithms, and the reliability of TLDTW was verified by further comparison.

Key words: time series; dynamic time warping; pathological alignment ; time-weight ; similarity measurement

0　引言

时间序列是一种常见且具有时间先后顺序的数据，它具有时间属性和其他变量属性[1]。这类数据广泛存在于各个领域，如医疗心电图[2]、气象温度气候变化[3]、客户行为和订单消费[4]、金融股票[5]等，时间序列数据的挖掘利用可以更好地把握研究对象的状态、更精准地预测变化趋势与规律，从而支持智能决策。近年来，相关学者对时间序列数据的挖掘利用进行了大量研究[6]，具体主要集中于相似性度量、聚类、分类、预测等[7]。

相似性度量分析是时间序列聚类、分类等的基础工作。基于模型的相似性度量、基于数据压缩的相似性度量和基于形状相似性度量等是其常用机制[8]。其中欧氏距离(euclidean distance，ED)和动态时间弯曲算法(dynamic time warping，DTW)是基于形状的时间序列相似性度量中常见方法。ED 相关算法对等长序列的度量效率高，但其“一对一”策略难以胜任非等长时间序列中缩放位移变化问题。DTW 通过弯曲拉伸收缩调整时间轴来计算两个序列的相似度，在应对时间序列相位偏移，振幅变化方面具有更强的鲁棒性[9,10]。

但 DTW 仍存在一些不足：如时间计算复杂度高[11]、序列匹配存在病态对齐[12]、忽视时间属性影响(匹配时间点对时未考虑时间间隔远近)[13]。随着计算机计算性能的快速提升，DTW 计算复杂度高问题得到了一定的缓解。针对病态对齐问题，Jeong[14]在 DTW 中的距离矩阵中引入相位差提出了权重动态弯曲算法(weighted DTW,WDTW)方法，WDTW 将相位差越高的元素赋予越高的惩罚权重，以避免时间序列过度弯曲和不合理匹配的问题。窗口限制法[10]限制了每一个时间点可以链接的数量，直接限制了病态对齐，但由于其刚性被过度修正，易导致正确链接被修改为错误链接。Zhang[15]等利用时间序列之间的对齐长度限制策略在全局上限制了对齐的总长度，在此基础上提出了 LDTW(DTW under limited warping path length)，从而抑制病态对齐现象。在 LDTW 基础上，夏寒松等[17]将对齐路径长度控制在某个区间改变为固定到某个具体值，并通过缩减 DTW 累计代价矩阵中元素的计算范围，以降低时间复杂度和开销。通常情况下，时间属性对相似性度量存在较大影响，即最近时间点的重要性远大于历史时间点的重要性，但是上述方法未考虑该影响因素。针对该问题，Li[17]等提出了一种基于时间加

权的 DTW 算法(time-weighed DTW, TDTW)，引入时间权重函数，给不同时间点赋予不同的权值，解决了时间点间隔远近贡献不同问题，提升相似度度量的效果。但 TDTW 在病态对齐的问题上仍然存在不足。

本文利用 TDTW 时间加权的思想，对 LDTW 进行改进，提出了时间加权改进的 LDTW 算法(LDTW based on time weighting, TLDTW)，TLDTW 继承 LDTW 算法对齐路径长度控制策略，并对序列的各时间点赋予时间权值，以期同时解决病态对齐和时间属性被忽略问题。

1 基于 DTW 的 LDTW 算法

DTW 最早由 Berndt 提出[10]，其核心思想是通过递归方式在两个序列之间找到一个最优的对齐方式，以实现最小的全局成本，即对齐路径中每对点之间成本的总和最小。

设时间序列 $A=\{a_1,a_2,a_3,\ldots,a_N\}$ 和 $B=\{b_1,b_2,b_3,\ldots,b_M\}$ 的长度分别为 N 和 M。定义时间点对之间的成本如(1)式所示。

$$d(a_i,b_j)=|a_i-b_j|^2 \tag{1}$$

两个序列任意两点之间的距离构成了 $N\times M$ 累计成本矩阵 $R_{N\times M}$，$R(i,j)$ 为矩阵第 i 行第 j 列对应单元的值，每个单元值所对应的是 DTW 中局部最优成本 $D(A,B_j)$，即 $R(i,j)=D(A,B_j)$，其中 A,B_j 分别表示 A，B 的子集(长度为 i 和 j)。其求解全局最优成本通过式(2)递归填充累计成本矩阵实现。DTW 的时间复杂度为 $O(NM)$。

$$D(A_i,B_j)=d(a_i,b_j)+\min\begin{cases}D(A_i,B_{j-1})\\D(A_{i-1},B_{j-1})\\D(A_{i-1},B_j)\end{cases} \tag{2}$$

DTW 通过弯曲时间序列的时域对时间序列的数据点进行匹配，不仅能够得到更好的形态度量效果，而且能够度量两条不等长的时间序列[18]。但在 DTW 递归求局部最优解时易出现病态对齐现象，如图 1 中红圈标注部分，出现大量“一对多”对齐的病态点(常称为奇点)，从而使得两个序列中不具有相似性的局部数据点进行匹配，以使得度量距离值极小，从而影响 DTW 相似性度量效果。

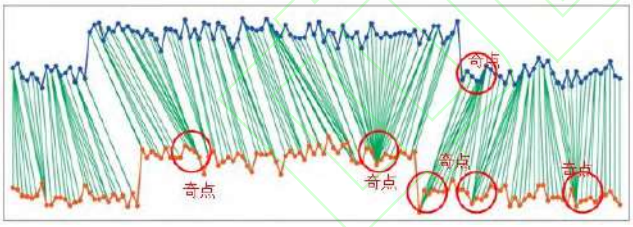


图 1 DTW 病态对齐现象

Fig. 1 DTW pathological alignment phenomenon

DTW 中子对齐路径 $p_k=R(i,j)$ 的长度为 l，$l\ni|i-j|$。当出现病态对齐现象时，序列间的链接总数会增加，所以对齐路径的总路径长度也会增加。基于 DTW 改进的 LDTW 其核心思想是通过限制两个时间序列之间的链接总数上限(而非每个点所涉及的链接数)来缓解病理对齐。LDTW 在 DTW 递归迭代过程中考虑其路径长度，$LD(A,B_j,l)$表式 LDTW 的局部成本，相应递归填充过程如式(3)所示。

$$LD(A_i,B_j,l)=R(i,j,l)=d(a_i,b_j)+\min\begin{cases}D(A_i,B_{j-1},l-1)\\D(A_{i-1},B_j,l-1)\\D(A_{i-1},B_{j-1},l-1)\end{cases} \tag{3}$$

l 表示局部最优成本所允许对齐路径的长度，L_{UB} 是对齐路径的上限，序列 A、B 之间最长对齐路径和最短对齐路径分别为 $MaxL$ 和 $MinL$，$MinL=max(N,M)$，$MaxL=N+M-1$。

图 2 表示对齐路径的最大和最小长度，L_{UB} 的范围在最大和最小长度之间；图中两条线分别表示最短的路径和其中一条最长路径(即两个序列长度之和减一)，最长路径具有的特点是只向当前点得右边或者上边搜索。LDTW 以第三维优先原则计算成本累计矩阵，并通过子序列间的对齐长度范围减少递归搜索空间[15]。基于 L_{UB} 范围约束，LDTW 搜寻符合约束的所有对齐路径并计算其对应距离值，选出其最小值作为度量值，并保留该最小值相应对齐路径。

图 3 为两条长度为 5 的时间序列构成的矩阵，每个单元格表示到达该点的路径长度，DTW 和 LDTW 选择路径示例如下所示。

DTW：$R[4][5]=d(a_4,b_5)+\min\{R[3][4],R[4][4],R[3][5]\}$

LDTW：

$R[4][5][4]=d(a_4,b_5)+\min\{R[3][4][3],\ R[4][4][3]\}$

$R[4][5][5]=d(a_4,b_5)+\min\{R[3][5][4],R[3][4][4],R[4][4][4]\}$

$R[4][5][6]=d(a_4,b_5)+\min\{R[3][5][5]\cdot R[4][4][5]\}$

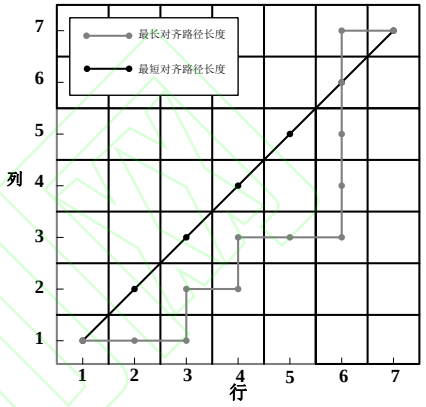


图 2 对齐路径的最大值和最小值

Fig. 2 Maximum and minimum values of alignment paths

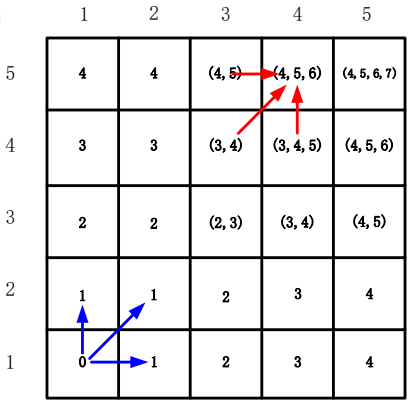


图 3 LDTW 算法的路径选择

Fig. 3 Path selection of ldtworithm

由此可见 LDTW 路径选择的长度限制，其到达点 R(4,5) 的可选路径长度有三种(4,5,6)，每一条路径选择都会有不同的对齐路径与之对应。

LDTW 相似性度量过程如算法 1 所示，相应时间复杂度从 $O(NM)$ 扩展到了三维 $O(NML_{UB})$。

图 4 为通过 LDTW 优化后的对齐效果，与图 1 不同之处在于红圈标注部分的病态对齐得到有效缓解，奇点大量减少。

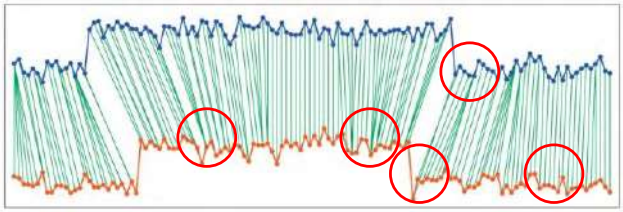


图 4 基于 LDTW 优化后的对齐效果

Fig. 4 Alignment effect optimized based on LDTW

算法 1　LDTW

说明：R(N, M, L_{UB})、L 分别为成本矩阵和对齐路径长度，$R(i,j,l)=LD(A_i,B_j,l)$，局部最优解可用累计成本矩阵的单元值表示

输入：时间序列 A, B;时间序列长度：N=length(A); M=length(B)；对齐路径距离上限/最大/最小路径长度：L_{UB} / $MinL$ / $MaxL$　。

输出：两序列累计距离总和和 $LD(A_N,B_M,L_{UB})$，即 $R(N,M,L_{UB})$，L_{UB}

a)初始化：$R(1,1,0)=D(a_1,b_1)$

/*初始化填充第一行*/

for n=2 to N do

$R(n,1,n-1)=R(n-1,1,n-2)+d(a_n,b_1)$

end for

/*初始化填充第一列*/

for m=2 to M do

$R(1,m,m-1)=R(1,m-1,m-2)\cdot+d(a_1,b_m)$

end for

b)递归填充累计成本矩阵:

for n=2 to N do

for m=2 to M do

$min\_l=\max(n,m)$

$max\_l=(n,m,L_{UB},N,M)$

for l= min_l to max_l do

$$R(i,j,l)=d(a_i,b_j)+\min\begin{cases}R(i,j-1,l-1)\\R(i-1,j-1,l-1)\\R(i-1,j,l-1)\end{cases}$$

end for

end for

end for

c)最佳路径选择：

$min\_L=\max(N,M)$

$max\_L=(L_{UB}-1)$

$LDTW=+\infty, L=0$

for l= min_L to max_L do

if $R(N,M,l)<LDTW$ then

$LDTW=R(N,M,l),\ L=l+1$

end if

end for

2　时间加权改进的 LDTW

LDTW 一定程度上解决了病态对齐问题，但仍忽略了时间属性，未考虑远近时间点时间权重不平衡现象，即在一般情况下，接近当前时间的时间点比远距离时间点具有更强影响力，应被赋予更大的时间权值。基于权重的 DTW(WDTW[18]) 和时间权重扩展的 DTW 算法(TWDTW[19])综合考虑了时间权重影响，但其基于相位差设计的权重系数对应逻辑函数高度依赖于时间序列类型，较适合于相位差较大的序列，泛化性较差。Li 提出的时间加权 DTW(TDTW)更关注时间序列中不同时间点对序列相似度度量的影响，但又忽略了病态对齐问题。为此，本研究融合限制对齐路径总长度和时间加权机制以解决上述不足。

时间加权的基本思想就是将越靠近当前时间点赋予更大的权重。设时间权重为 $W(i,j)$,并与成本矩阵中的每个单元 $R(i,j)$ 一一对应；在此，$W(i,j)$ 定义为(4)式所示。

$$W(i,j)=\frac{\sqrt{(i/\mathrm{N})^2+(j/M)^2}}{\sqrt{2}} \tag{4}$$

N,M 分别表示序列 A,B 的长度。令 $i/N, j/M$ 分别表示 A、B 两个序列中第 i 个和第 j 个点所对应的权重。下面给出了长度均为 10 的 A、B 两个序列时间权重矩阵：

$$W_{10\times10}=\begin{pmatrix}0.1000&0.1581&0.2236&0.2915&0.3606&0.4301&0.5000&0.5701&0.6403&0.7106\\0.1581&0.2000&0.2550&0.3162&0.3808&0.4472&0.5148&0.5831&0.6519&0.7211\\0.2236&0.2550&0.3000&0.3536&0.4123&0.4743&0.5385&0.6042&0.6708&0.7382\\0.2915&0.3162&0.3536&0.4000&0.4528&0.5099&0.5701&0.6325&0.6964&0.7616\\0.3606&0.3808&0.4123&0.4528&0.5000&0.5523&0.6083&0.6671&0.7280&0.7906\\0.4301&0.4472&0.4743&0.5099&0.5523&0.6000&0.6519&0.7071&0.7649&0.8246\\0.5000&0.5148&0.5385&0.5701&0.6083&0.6519&0.7000&0.7517&0.8062&0.8631\\0.5701&0.5831&0.6042&0.6325&0.6671&0.7071&0.7517&0.8000&0.8515&0.9055\\0.6403&0.6519&0.6708&0.6964&0.7280&0.7649&0.8062&0.8515&0.9000&0.9513\\0.7106&0.7211&0.7382&0.7616&0.7906&0.8246&0.8631&0.9055&0.9513&1.0000\end{pmatrix}$$

可以看出随着 i,j 值的增大，该时间点获得时间权重逐渐增大；可以保证越靠近近时间点获得的权值就越大，即 $W(i,j)\geq W(p,q), p\leq i,\ q\leq j$ 。由于 DTW 算法的特性，递归填充过程是正向传递，由图 2 可以看出，是由矩阵的左下角向右上角递归填充(最远时间点向最近时间点递归填充)，而最优路径选择则是反向从矩阵的右上角向左下角查找最小值(最近时间点向最远时间点查找)。如果在此先对原始时间序列 $A=\{a_1,a_2,a_3,\ldots,a_n\}$，$B=\{b_1,b_2,b_3,\ldots,b_m\}$ 进行反转操纵得到反转后时间序列 $\bar{A}=\{\bar{a}_1,\bar{a}_2,\ldots,\bar{a}_n\}=\{a_n,a_{n-1},\ldots,a_1\}$ 、$\bar{B}=\{\bar{b}_1,\bar{b}_2,\ldots,\bar{b}_m\}=\{b_m,b_{m-1},\ldots,b_1\}$，递归填充的过程便转换为从最近时间点向最远时间点填充(更加符合算法侧重近时间点重要性的特性)，且查找路径是从时间序列的首向尾查找(远时间点向近时间点查找)符合查找路径的逻辑，此时的权重也要作出适当的修改，反转序列之前 A 序列第 i 个点对应的是反转之后第 n-i+1 个点，B 序列同理，则时间权值的变化由 i/N 变化为 (N−i+1)/N 即 1−(i-1/N)，所并修正原有权重 $W(i,j)$ 为(5)式所示。

$$W(i,j)=\frac{\sqrt{(1-(i-1/n))^2+(1-(j-1/m))^2}}{\sqrt{2}} \tag{5}$$

修正 LDTW 累计成本矩阵初始化和相应递归填充过程如(6)式所示。

$$TLD(\bar{A}_i,\bar{B}_j,l)=R(i,j,l)=W(i,j)\times d(\bar{a}_i,\bar{b}_j)+\min\begin{cases}D(\bar{A}_i,\bar{B}_{j-1},l-1)\\D(\bar{A}_{i-1},\bar{B}_j,l-1)\\D(\bar{A}_{i-1},\bar{B}_{j-1},l-1)\end{cases} \tag{6}$$

$TLD(\bar{A},\bar{B}_j,l)$ 表示 TLDTW 中局部最优成本，它与矩阵的单元值一一对应，其他符号与式(3)同。

TLDTW 相似性度量过程如算法 2 所示。TLDTW 计算了累计成本矩阵每一个单元对应时间点对的时间权值，在填充迭代过程中每一步均加入了时间权重系数，其时间复杂度亦为 $O(nmL_{UB})$ 。

算法 2　TLDTW

说明：R(n,m,LUB)、L 分别为成本矩阵和对齐路径长度，$R(i,j,l)=TLD(A_i,B_j,l)$，局部最优解可用累计成本矩阵的单元值表示

输入：时间序列 A, B 反转后序列 $\bar{A},\bar{B}_j$，时间序列长度：N=length(A); M=length(B)，对齐路径距离上限/最大/最小路径长度：L_{UB} / $MinL$ / $MaxL$

输出：两个序列累计距离总和和 $TLD(A_N,B_M,L_{UB})$，即 $R(N,M,L_{UB})$，LUB

a)时间加权矩阵构建：

for i=1 to N do

$X(i)=1-\frac{i-1}{N}$

for j=1 to M do

$Y(i)=1-\frac{j-1}{M}$

$W(i,j)=\frac{\sqrt{(X(i))^2+(Y(j))^2}}{\sqrt{2}}$

end for

end for

b)初始化：

$R(1,1,0)=d(a_1,b_1,0)\times W(1,1)$

/*初始化填充第一行*/

```
for n=2 to N do
   R(n,1,n-1) = R(n-1,1,n-2) + d(a_n,b_1)×W(n,1)
end for
/*初始化填充第一列*/
for m=2 to M do
   R(1,m,m-1) = R(1,m-1,m-2) + d(a_1,b_m)×W(m,1)
end for
```

c)递归填充累计成本矩阵：

```
for n=2 to N do
  for m=2 to M do
     min_l = max(n,m)
     max_l = (n,m,L_UB,N,M)
    for l= min_l to max_l do
```

$$R(i,j,l)=d(a_i,b_j)\times W(i,j)+\min\begin{cases}R(i,j-1,l-1)\\R(i-1,j-1,l-1)\\R(i-1,j,l-1)\end{cases}$$

```
    end for
  end for
end for
```

d)最佳路径选择：

$min\_L=\max(N,M)$

$max\_L=(L_{UB}-1)$

$TLDTW=+\infty, L=0$

```
for l= min_L to max_L do
  if R(N,M,l)<TLDTW then
     TLDTW = R(N,M,l), L = l+1
  end if
end for
```

3 实验验证

在此以 UCR 时间序列分类文档为数据集[20]，采用近邻分类实验方法进行实验并开展对比分析，利用分类准确率(Accuracy)和可靠性指标(通过准确率增益 Gain 度量)进行评价，验证所提出 TLDTW 的有效性与优越性。

3.1 UCR 数据集

UCR 共包含 128 个时间序列数据集，每个数据集均带有类标签，并被划分为训练集和测试集。数据集中序列长度分布于[60,637]，在此随机选取 25 个数据集作为本实验用数据集，被选数据集序列长度分布于[96,637]之间。表 1 给出各数据集中的训练集和测试集的大小、类别和时间序列长度等属性。

3.2 1—NN 分类验证

1—NN 分类先利用相似性度量算法计算各时间序列之间的距离值，然后选定一个时间序列，并优选与该确定序列距离最短的序列，将其归为一类。未确定类别的时间序列采用上述相同操纵方式直到所有时间序列确定其所属类别。1—NN 不需要设置任何参数，精度完全取决于相似性距离度量的效果[21]。

以各数据集中的训练集训练 TLDTW，并与 ED、DTW、LDTW、TDTW 四种算法进行对比。表 2 给出了在训练集上分类准确率对比结果；其中 LDTW 和 TLDTW 需设置对齐路径上限 L_{UB} 参数值，该参数由留一验证实验(Leave-one-out cross-validation，LOOCV)选取(详见 3.4 节)。可以看出 TLDTW 在选取的 25 个数据集上有 20 个取得最优效果(最高分类准确率)，是所给出对比算法中准确率最高的。

图 5 给出了 TLDTW 和四种对比算法分类准确率的可视化结果。图中的数据点表示一个数据集，横轴表示对比算法所得准确率，纵坐标轴表示 TLDTW 准确率。数据点落在斜线以上区域说明 TLDTW 的效果更好，落在斜线下面说明对比算法的效果更好，若落在斜线之上说明效果相当。由图 5(a、b)可以看出，所提出的 TLDTW 相对于 ED 和 DTW 具有明显优势，相对于 ED 和 DTW 分别在 3 和 4 个数据集上取得相同的效果，在 22 和 20 个数据集上 TLDTW 取得了更好的准确率；由图 5(c、d)可以看出，TLDTW 相对于 TDTW 和 LDTW 分别在 17 个和 14 个数据集上取得更高的准确率，另外分别有 7 个和 6 个数据集上准确率一样。

表 1 UCR 部分数据集

Tab. 1 UCR partial dataset

序号	数据集名称	训练集	测试集	类别数	长度
1	Adiac	390	391	37	176
2	Beef	30	30	5	470
3	BirdChicken	20	20	2	512
4	car	60	60	4	577
5	CBF	30	900	3	128
6	coffee	28	28	2	286
7	ECG200	100	100	2	96
8	Face(all)	560	1690	14	131
9	FaceFour	24	88	4	350
10	Fish	175	175	7	463
11	Gun-Point	50	150	2	150
12	Ham	109	105	2	431
13	Lightning-2	60	61	2	637
14	Lightning-7	70	73	7	319
15	Meat	60	60	3	448
16	OliveOil	30	30	4	570
17	OSU Leaf	200	242	6	427
18	Plane	105	105	7	144
19	Swedish Leaf	500	625	15	128
20	Symbols	25	995	6	398
21	Trace	100	100	4	275
22	Two Patterns	1000	4000	4	128
23	wine	57	54	2	234
24	Yoga	300	3000	2	426
25	50Words	450	455	50	270

表 2 五种对比的分类准确率

Tab. 2 Classification accuracy of five comparison algorithms

序号	数据集	ED	DTW	TDTW	LDTW(LUB)	TLDTW(LUB)
1	Adiac	0.611	0.604	0.637	0.627(232)	0.643(232)
2	Beef	0.667	0.633	0.677	0.667(471)	0.692(471)
3	BirdChicken	0.550	0.750	0.726	0.659(522)	0.625(522)
4	Car	0.733	0.733	0.768	0.867(598)	0.884(583)
5	CBF	0.852	0.997	0.997	0.997(198)	0.999(198)
6	coffee	1.000	1.000	1.000	1.000(307)	1.000(307)
7	ECG200	0.88	0.77	0.883	0.880(98)	0.907(98)
8	Face(all)	0.714	0.808	0.727	0.823(173)	0.806(173)
9	FaceFour	0.784	0.83	0.806	0.898(386)	0.828(386)
10	Fish	0.783	0.823	0.769	0.914(474)	0.857(474)
11	Gun-Point	0.913	0.907	0.988	0.980(158)	0.988(158)
12	Ham	0.6	0.467	0.524	0.598(425)	0.611(425)
13	Lightning-2	0.754	0.869	0.907	0.902(778)	0.907(778)
14	Lightning-7	0.575	0.726	0.803	0.795(455)	0.803(455)
15	Meat	0.933	0.933	0.933	0.933(479)	0.933(479)
16	OliveOil	0.867	0.833	0.903	0.867(581)	0.911(581)
17	OSU Leaf	0.521	0.591	0.623	0.699(468)	0.723(468)
18	Plane	0.962	1.000	1.000	1.000(218)	1.000(218)
19	Swedish Leaf	0.789	0.792	0.827	0.869(134)	0.873(134)
20	Symbols	0.899	0.95	0.957	0.931(412)	0.965(412)
21	Trace	0.76	1.000	1.000	1.000(416)	1.000(416)
22	Two Patterns	0.907	1.000	1.000	1.000(194)	1.000(194)
23	wine	0.611	0.574	0.608	0.618(425)	0.618(425)
24	Yoga	0.83	0.836	0.867	0.851(467)	0.883(467)
25	50Words	0.631	0.69	0.538	0.820(581)	0.737(578)
	Aver.	0.765	0.805	0.829	0.847	0.847
	最优结果数	2	6	8	10	20

注：LDTW(LUB)和 TLDTW(LUB)中的 LUB 表示 LDTW 和 TLDTW 的最优对齐路径上限

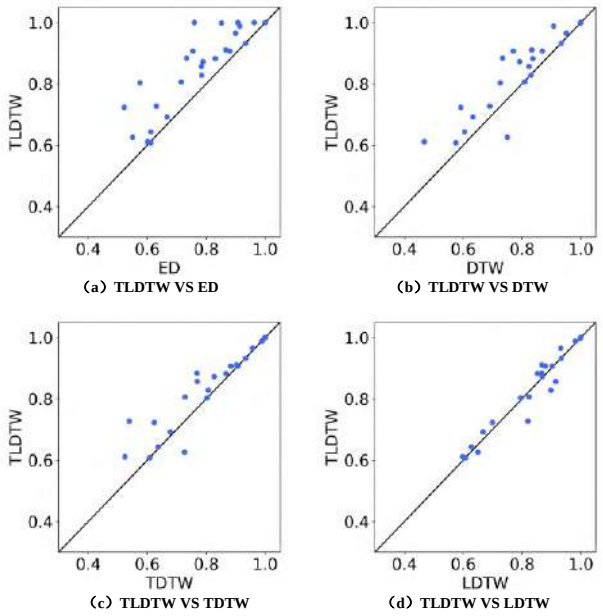


图 5　TLDTW 与其他四种算法在 1-NN 分类上的正确率对比

Fig. 5　Comparison of accuracy between TLDTW and other four algorithms in 1-NN classification

3.3　神枪手逻辑谬误

为避免德州神枪手逻辑谬论，证明算法的可靠性，须进一步验证算法在所取得准确率高的训练集对应测试集上依旧能取得好的效果。在此引入增益混淆矩阵评估所提出算法的可靠性，先通过式(7)计算 TLDTW 准确率增益(gain)，具体包括预期准确率增益(训练集中取得的最好结果)和实际准确率增益(测试集中取得的结果)：

$$\text{Gain}=\frac{\text{Accuracy}_{\text{TLDTW}}}{\text{Accuracy}_{\text{对比算法}}} \tag{7}$$

其中 $\text{Accuracy}_{\text{TLDTW}}$ 表示 TLDTW 的准确率，$\text{Accuracy}_{\text{对比算法}}$ 表示对比算法(ED、DTW、LDTW、TDTW)的准确率。*Gain* 大于 1 时表明 TLDTW 在给定数据集上优于对比算法，反之亦然。其中，TLDTW 和 LDTW 的 L_{UB} 最优值通过 LOOCV 确定。

图 6 给出了 TLDTW 对比于其他四种算法的预期准确率增益和实际准确率增益。图中的每个点表示一个数据集(共 25 个)，每个点落在图中四个区域中的一个，这四个区域分别是：

a) TP(真阳性)区：预测 TLDTW 相对对比算法将提高分类准确率，实际结果与预测一致。落在该区域的数点越多，证明该方法可靠性越高。

b) TN(真阴性)区：正确预测出 TLDTW 相对对比算法会降低分类准确性。应避免在这类数据集上使用 TLDTW。

c) FN(假阴性)区：预测 TLDTW 相对对比算法会降低准确率，但准确率实际上有所提高。

d) FP(假阳性)区：预测 LDTW 相对对比算法将提高准确率，但实际准确率实际上降低了。这个区域的点越多说明该算法的可靠性越高

从图 6 中的四幅图中可以看出数据集对应点大部分落在 TP 区(最多 22，最少 17)，说明所提出改进算法的分类正确率提高是可靠的。虽然落入 TP 区域的数据点有所减少，但减少的这些数据点并没有落入其他区域，而是处于横轴边缘，这意味着在测试集上的分类准确率并没有降低。

3.4　LDTW 和 TLDTW LUB 优化

LDTW 和 TLDTW 有一个对齐路径长度上界的参数 L_{UB}。在进行 1-NN 分类实验时，需要确定 LDTW 和 TLDTW 对齐路径长度上界参数 L_{UB}。找到最合适的 L_{UB}，LDTW 和 TLDTW 才能更好地缓解病态对齐。根据 LDTW 定义规则，长度分别为 N 和 M 的两条时间序列 A、B，L_{UB} 应满足 $\max(N,M)+1<L_{UB}<(N+M-2)$。以第 11 个数据集 Gun_Point 为例，其长度为 150，则 L_{UB} 应在 151-298 之间。在此，开展 LOOCV 交叉验证优化确定各数据集对应的 L_{UB}；对于数据集的序列长度变大，相应 L_{UB} 的区间范围也大(Lightning-2 范围在(638-1274)，这时候 LOOCV 搜索步长也适当增大。

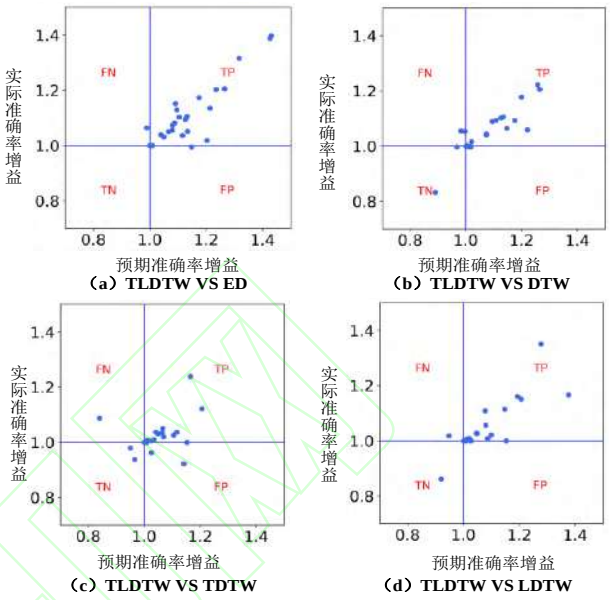


图 6　预期增益和实际增益对比

Fig. 6　Comparison between expected gain and actual gain

LOOCV 可以得到每一个候选 L_{UB} 值对应的 1-NN 分类错误率。图 7 给出了 3 个数据集在不同候选 L_{UB} 值对应的分类的错误率。通过时间序列分类的错误率可计算在该 L_{UB} 值之下准确率，计算出随着 L_{UB} 的变化分类准确率的变化，从而找到最优解。当只有一个最优点时，直接选取该点对应的 L_{UB} 长度(如图 7(a)所示)；当出现多个最优解时，由 LDTW 和 TLDTW 算法的时间复杂度描述，L_{UB} 选取就需要考虑算法计算量的因素，L_{UB} 的值越小算法的计算量越小，故一般取值最小值作为 L_{UB}。如图 7(b、c)所示。最终确定各数据集 LDTW 和 TLDTW 的 L_{UB} 如表 3 所示

表 3　交叉验证优选 L_{UB} 值

Tab. 3　Cross validation preferred L_{UB} value

数据集	LDTW	TLDTW	数据集	LDTW	TLDTW	数据集	LDTW	TLDTW
Adiac	232	232	Fish	474	474	Plane	218	218
Beef	471	471	Gun-Point	158	158	Swedish Leaf	134	134
BirdChicken	522	522	Ham	425	425	Symbols	412	412
Car	598	583	Lightning-2	778	778	Trace	416	416
CBF	198	198	Lightning-7	455	455	Two Patterns	194	194
coffee	307	307	Meat	479	479	wine	425	425
ECG200	98	98	OliveOil	581	581	Yoga	467	467
Face(all)	173	173	OSU Leaf	468	468	50Words	581	578
FaceFour	386	386						

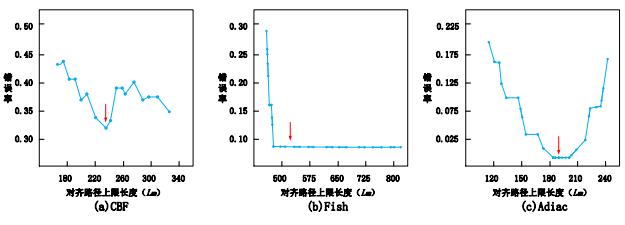


图 7　交叉验证 LUB 曲线

Fig. 7　Cross validation LUB curve

3.5 算法时间复杂度与时间开销

在本研究中，TLDTW 算法的时间复杂度为 $O(n^3)$,和 LDTW 算法的时间复杂度一致，由于时间加权的影响，本文可以进一步对比计算出 TLDTW 和 LDTW 算法的计算量。从实验结果参数相同的 10 个数据集其中四个为表现性能持平，三个在改进算法表现更优秀，三个在 LDTW 算法上分类正确率更高。表 4 展示了两种算法之间的时间花费的差值。为了更加直观的表现出对比算法的分类时间开销，本文将采用条形图对比，如图 8 所示。在图中看出由于时间加权的加入，在提升分类准确率的同时，TLDTW 算法时间开销都大于 LDTW 算法。表现在数据集中序列长度越短，则时间开销增加幅度越小，序列越长，时间开销增加越大。

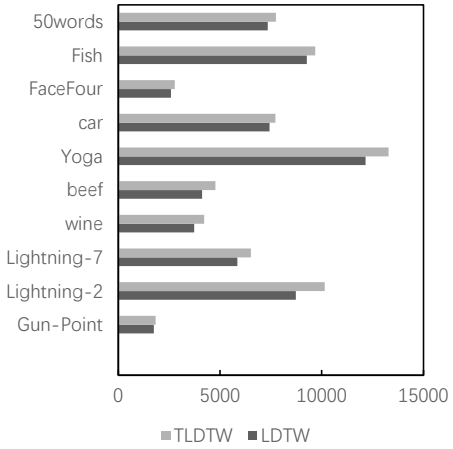


图 8 两种算法在相同数据集上时间开销

Fig. 8 The two algorithms spend time on the same data set

表 4 LDTW 和 TLDTW 算法的时间开销对比

Tab. 4 Comparison of time overhead between LDTW and TLDTW algorithms

数据集	LDTW(s)	TLDTW(s)	时间开销差(s)
Gun-Point	1747	1833	86
Lightning-2	8725	10143	1418
Lightning-7	5847	6521	674
wine	3725	4227	502
beef	4117	4773	656
Yoga	12159	13287	1128
car	7431	7721	290
FaceFour	2586	2774	188
Fish	9271	9687	416
50words	7341	7745	404

4 结束语

本研究在 LDTW 和 TDTW 的基础上提出 TLDTW 算法，并开展了相应实验验证，所提出算法的特点及其效果主要体现在如下几方面：

(1) TLDTW 利用 LDTW 限制对齐路径长度思想和 TDTW 时间加权机制，协同解决了 DTW 病态对齐和未考虑时间属性影响问题。

(2) 基于 UCR 数据集开展 1-NN 分类实验并与 ED、DTW、TDTW、LDTW 进行了对比，结果显示基于 TLDTW 相似度量的分类准确率相对于对比算法更高，在选取的 25 个数据集中，TLDTW 分类准确率最高的数据集数占 19 个，验证了所提出 TLDTW 的优越性；同时通过准确率增益量化分析了 TLDTW 的可靠性。

后续可从两方面开展研究：一是降低 LDTW 算法开销；二是尝试非线性时间权值(TLDTW 时间权值为基于时间序列长度的线性组合模型)以提高相似性度量效果。

参考文献：

[1] Li H. Multivariate time series clustering based on common principal component analysis [J]. Neurocomputing, 2019, 349: 239-247.

[2] Anguera A, Barreiro J M, Lara J A, *et al.* Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry [J]. Computational and structural biotechnology journal, 2016, 14: 185-199.

[3] Zhai Y, Wang J, Teng Y, *et al.* Water demand forecasting of Beijing using the time series forecasting method [J]. Journal of Geographical Sciences, 2012, 22 (5): 919-932.

[4] Ruan G, Hanson P C, Dugan H A, *et al.* Mining lake time series using symbolic representation [J]. Ecological Informatics, 2017, 39: 10-22.

[5] ZHANG Q F, He W M. Financial time series similarity search based on exponential smoothing and WKNN [J]. Modern Computer, 2019 (29): 21-25.

[6] Fu T. A review on time series data mining [J]. Engineering Applications of Artificial Intelligence, 2011, 24 (1): 164-181.

[7] Chen Y, Hu X, Fan W, *et al.* Fast density peak clustering for large scale data based on kNN [J]. Knowledge-Based Systems, 2020, 187: 104824.

[8] 陈海燕, 刘晨晖, 孙博. 时间序列数据挖掘的相似性度量综述 [J]. 控制与决策, 2017, 32 (001): 1-11. (Chen Haiyan, Liu Chenhui, Sun Bo. Survey on similarity measurement of time series data mining [J]. Control and Decision, 2017, 32 (001): 1-11.)

[9] 李正欣, 郭建胜, 毛红保, 等. 多元时间序列相似性度量方法 [J]. 控制与决策, 2017, 032 (002): 368-372. (Li Zhengxin, Guo Jiansheng, Mao Hongbao. Similarity measure for multivariate time series. [J]. Control and Decision, 2017, 032 (002): 368-372.)

[10] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series [C]// KDD workshop. 1994, 10 (16): 359-370.

[11] 李海林, 梁叶, 王少春. 时间序列数据挖掘中的动态时间弯曲研究综述 [J]. 控制与决策, 2018, 33 (8): 1345-1353. (Li Hailin, Liang Ye, Wang Shaochun. Review on dynamic time warping in time series data mining. [J]. Control and Decision, 2018, 33 (8): 1345-1353.)

[12] Keogh E J, Pazzani M J. Derivative dynamic time warping [C]// Proceedings of the 2001 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2001: 1-11.

[13] Toshniwal D, Joshi R C. Similarity search in time series data using time weighted slopes [J]. Informatica, 2005, 29 (1) .

[14] Jeong Y S, Jeong M K, Omitaomu O A. Weighted dynamic time warping for time series classification [J]. Pattern recognition, 2011, 44 (9): 2231-2240.

[15] Zhang Z, Tavenard R, Bailly A, *et al.* Dynamic time warping under limited warping path length [J]. Information Sciences, 2017, 393: 91-107.

[16] 夏寒松, 张力生, 桑春艳. 基于 LDTW 的动态时间规整改进算法 [J/OL]. 计算机工程: 1-14 [2021-06-27]. https://doi. org/10. 19678/j. issn. 1000-3428. 0059468. (Xia Hansong Zhang Lisheng Sang Chunyan. Improved Algorithm of Dynamic Time Warping Based on LDTW [J/OL]. Computer Engineering, 1-14 [2021-06-27]. https://doi. org/10. 19678/j. issn. 1000-3428. 0059468.)

[17] Li H. Time works well: Dynamic time warping based on time weighting for time series data mining [J]. Information Sciences, 2021, 547: 592-608.

[18] Anantasech P, Ratanamahatana C A. Enhanced weighted dynamic time warping for time series classification [C]// Third international congress on information and communication technology. Springer, Singapore, 2019: 655-664.

[19] Maus V, Câmara G, Cartaxo R, *et al.* A time-weighted dynamic time

warping method for land-use and land-cover mapping [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2016, 9 (8): 3729-3739.

[20] Dau, Hoang Anh, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping Chen, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen and Gustavo Batista (2018) . "The UCR Time Series Classification Archive. ".

[21] Ding H, Trajcevski G, Scheuermann P, *et al.* Querying and mining of time series data: experimental comparison of representations and distance measures [J]. Proceedings of the VLDB Endowment, 2008, 1 (2): 1542-1552.

上海市计算机学会会刊

ISSN 1000-386X
CN31 - 1260 /TP

8

Vol.37 No.8

计算机应用与软件

上海市著名商标

COMPUTER APPLICATIONS AND SOFTWARE

全国中文核心期刊（2017）
中国学术期刊综合评价数据库来源期刊
中文科技期刊数据库（全文版）收录期刊
美国《剑桥科学文摘》收录期刊

中国科技论文统计源期刊（中国科技核心期刊）
中国科学引文数据库（CSCD）来源期刊（2015-2016）
万方数据—数字化期刊群全文收录期刊
美国《乌利希国际期刊指南》收录期刊

internet and software

主　办

上海市计算技术研究所
上海计算机软件技术开发中心

2020

计算机应用与软件(月刊)

第 37 卷　第 8 期　2020 年 8 月

目　　次

基于自组织映射 - 反向传播网络的 PCB 样板投料预测

郑彬彬 吕盛坪* 李灯辉 冼荣亨
(华南农业大学南方农业机械与装备关键技术教育部重点实验室 广东 广州 510642)

摘 要 精准预测印制电路板样板物料投入将减少超投浪费和补投成本,为此提出结合自组织映射(Self-organizing maps, SOM) -反向传播(Back propagation network, BPN)网络的预测机制。基于 SOM 对样本进行聚类分组;采用特征选择机制优选各分组样板报废率关键影响属性;对各分组构建基于 BPN 的报废率预测模型;将其转换为预测投入生产面板数,并开展模型训练与性能评估。与多种模型进行对比分析,结果表明该模型在降低均方误差、绝对平均误差、平均绝对百分比误差、车间余数入库率和补投率等方面具有较明显优势。

关键词 印制电路板 投料预测 自组织映射 反向传播网络

中图分类号 TP391 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2020.08.011

PCB SAMPLE FEEDING PREDICTION BASED ON SELF-ORGANIZING MAPS AND BACK PROPAGATION NETWORK

Zheng Binbin Lü Shengping* Li Denghui Xian Rongheng
(*Key Laboratory of Key Technology on Agricultural Machine and Equipment, Ministry of Education, South China Agricultural University, Guangzhou 510642, Guangdong, China*)

Abstract More accurate prediction of PCB template material input reduces over-investment waste and supplemental feeding cost. This paper proposes a prediction mechanism based on self-organizing maps(SOM) and back propagation network(SOM-BPN). The samples were clustered and grouped based on the SOM, and then the feature selection mechanism was used to optimize the key impact attributes of the scrap rate of each sample category; the BPN prediction model of scrap rate based on BPN was constructed for each group; it was converted to the number of predicted production panels, and model training and performance evaluation were carried out. Compared with various models, and the results indicate that our model has obvious advantages in reducing the mean square error, the absolute average error, the average absolute percentage error, the workshop inventory storage rate and the supplementary investment rate.

Keywords PCB Material feeding prediction Self-organizing maps Back propagation network

0 引 言

印制电路板(Printed circuit board,PCB)是电子元器件的支柱,常被称为“电子产品之母”。随着计算机、通信、消费电子、5G、汽车电子、人工智能等行业的快速发展及其产品的迭代更新,具有不同设计特点和制造要求的多样个性化 PCB 订单(企业常称之为样板)快速增加,针对样板的生产模式也从传统的大规模批量生产转化为面向客户的小批量生产,相应的生产管控面临一系列新的挑战,生产前更准确预测每个订单的投料是关键问题之一。

目前,大部分 PCB 样板生产基本上依靠人工经验估算投料面积并转换计算相应生产面板(Panel)数。但人工投料常导致车间超投和补投均较高且波动较大。超投剩余个性化 PCB 样板只能置于库存或直接销毁。通过补投可以减少样板剩余,但会增加生产成本和造成交货拖期,影响企业信誉。生产前更合理地

收稿日期:2019-06-16。国家自然科学基金青年科学基金项目(51605169)。郑彬彬,硕士生,主研领域:工业大数据。吕盛坪,副教授。李灯辉,硕士生。冼荣亨,硕士生。

确定各订单投料面积和投入 Panel 数，可以降低物料、生产、库存和销毁等综合成本，减少投料人力投入[1-2]。同时，减少冗余生产可以降低因生产和销毁带来的化学药品和重金属污染。

数据驱动的智能制造框架[3-4]、范式[5-6]、分析方法和体系[7]等被大量研究。相应成果已广泛应用于支持产品设计、生产制造、销售、服务和回收等产品全生命周期不同阶段[8-9]。同时，数据挖掘为把握产品质量规律和改进质量提供了更精益智能化手段，相应研究主要集中在质量描述、预测、分类和参数优化四个方面[10]。具体到 PCB 质量规律挖掘主要集中在 PCB 贴装相关工艺，所采用理论方法主要集中在支持向量机(Support vector machine，SVM)[11]、人工神经网络(Artificial neural networks，ANN)[12-13]、ANN 与遗传算法结合[14-15]、模糊 ANN[16]、自组织映射[17-18]等。所涉及业务对象及其任务主要集中在 PCB 贴装相关工艺的质量描述、预测和参数优化。但是上述研究较少涉及 PCB 生产质量特别是样板质量规律挖掘研究。

结合企业需求，吕盛坪等[1]利用多元线性回归、卡方自动交互检测器、SVM 和 ANN 构建了报废率预测模型。随后，提出了考虑单属性变结构人工神经网络(multiple structural change ANN，MSC-ANN)预测模型[2]。但是订单结构及其报废率影响因素可能存在较大差异，综合考虑样板不同属性对订单进行分组，继而优选各分组订单质量影响关键因素，在此分组构建相应预测模型将有利于进一步提高预测模型的精准度。本文提出先基于 SOM 对样本进行聚类分组；继而采用特征选择机制，优选各分组数据报废率关键影响属性；在每个分组的基础上构建 BPN 报废率预测模型；综合 PCB 生产特点，将其转换为对应预测投入生产面板数；最后以生产车间样板训练上述模型并以不同评价指标验证所提出模型的可行性和优越性。

1 属性与样本

综合企业资源管理数据库中属性，利用继承、派生、转换等方式，共梳理影响样板报废率和统计分析属性 56 个，具体如表 1 所示[2]。编号 1－35 是可能影响每个样板报废率的属性；36－56 是统计变量，其中生产拼板数、要求生产数量、向上圆整 Panel 数、成品单元面积、要求生产面积等变量(编号分别为 36、38、39、46 和 47)，不仅可以作为统计参数，还可以作为预测模型建立的候选属性。

表 1 PCB 样板属性

编号	名称	符号	编号	名称	符号
1	板厚	Pt	29	是否无铅喷锡	Lfhasl
2	层数	Ln	30	是否 OSP	Osp
3	是否罗杰斯材料	Ro	31	是否图镀铜镍金	Cnapp
4	板镀次数	Plfr	32	是否镀金手指	Gfig
5	工序数	Noo	33	是否电镀硬金	Godp
6	半固化片数	NPP	34	是否软金镍钯金	Snap
7	每 SET 最多允许报废单元数	Sus	35	是否沉金沉银沉锡	Iasa
8	是否光电板	Photb	36	生产拼板数	Duap
9	是否高频板	Highfb	37	补投频次	Supff
10	是否半导体测试板	Semictb	38	要求生产数量	Reqq
11	是否有负片电镀	Nflp	39	向上圆整 Panel 数	Reqp
12	是否有减薄铜	Tinc	40	投入 Panel 数	Fedq
13	是否 III 级验收标准	IPCIII	41	至少投入 Panel 数	Lfp
14	是否华为验收标准	Huawei	42	投入 Panel 数	Fedp
15	内层最小线宽	Mwil	43	报废数量	Scraq
16	内层最小间距	Mlsil	44	合格入库数量	Qualq
17	外层最小线宽	Mwol	45	余数入库数量	Surpq
18	外层最小间距	Mlsol	46	成品单元面积	Dunita
19	芯板残铜率均值	Arcr	47	要求生产面积	Reqa
20	是否有阻焊塞孔	Srph	48	投料面积	Feda
21	是否树脂塞孔	Phwr	49	报废面积	Scrap area
22	是否有二钻	Secd	50	入库面积	Quala
23	是否有背钻	Bcdr	51	余数入库面积	Surpa
24	是否字符-打印机生产	Chaprt	52	补投率	Supfr
25	油墨颜色是否白色	White	53	报废率	Scrar
26	油墨颜色是否蓝色	Blue	54	合格率	Qualr
27	油墨颜色是否黑色	Black	55	余数入库率	Surpr
28	是否有铅喷锡	Hasl	56	历史良率	Hquar

在此基础上，从企业资源管理数据库中抽取一个厂 2013 年 10 月至 2016 年 10 月期间累计的共计 30 117 条有效数据，进一步采用多变量箱线图[2]筛除异常数据，最后得到 29 157 条样板数据作为本研究模型构建和测试分析样本。

2 SOM-BPN 预测模型

PCB 样板结构和报废率影响关键属性存在一定差异，将所有样本集中于单一模型之中易降低模型预测

精度，增大预测偏差，降低泛化能力。

本文先基于 SOM 对样板进行聚类分组，进一步优选各类样板报废关键影响属性并构建基于 BP 网络的预测模型。SOM 网络能将任意维的输入在输出层映射成一维或二维图形，并保持其拓扑结构不变。网络通过对输入数据的反复学习可以使权重向量空间与输入数据的概率分布趋于一致，使得输入属性相近的数据可以聚合在一起。SOM-BPN 模型框架如图 1 所示。

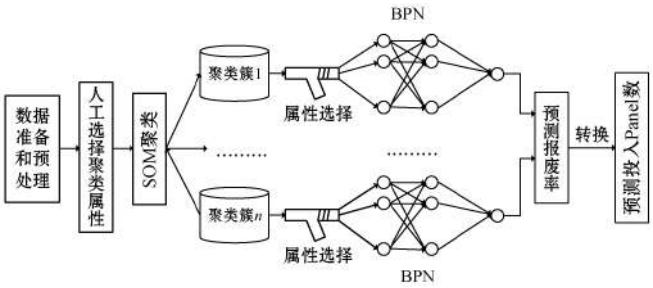


图 1　SOM-BPN 预测模型框架

具体步骤如下：

(1) 数据准备和预处理：基于表 1 给定属性及其抽取的历史数据，对各变量数据开展 0 - 1 归一化处理，以降低不同属性取值范围差异影响。

(2) 聚类属性选择：结合车间专家经验选取聚类输出属性，具体包括层数、工序数、内层最小线宽/间距、外层最小线宽/间距、要求生产数量、向上圆整 Panel 数、要求生产面积、是否有铅喷锡/无铅喷锡/OSP/图镀铜镍金/镀金手指/电镀硬金/软金镍钯金/沉金沉银沉锡。其中层数、工序数代表了样板整体特征；内外层最小线宽/间距是孔线加工代表性特征；要求生产数量、向上圆整 Panel 数、要求生产面积代表订单规模。

(3) 基于 SOM 的样本聚类分组：SOM 是一种只有输入层-竞争层的神经网络。在此使用的 SOM 输入层为上述 17 个属性；竞争层在此设置为由 2 × 3 神经元组成的二维平面离散网络，并且与输入层之间全连接。竞争层为 2 × 3 的二维平面将聚类数控制在 2 ~ 4 个之间，以便降低车间训练、测试和后续维护模型数量并保持较好的预测精准度。

(4) 聚类样本关键影响属性优选：模型构建输入属性过多将增加数据准备、预处理、模型构建、预测分析的复杂度和时间，且更容易导致模型过拟合、降低模型泛化能力。本文采用线性相关性、最大信息系数、递归特征消除、线性回归、Lasso 回归、Ridge 回归和随机森林回归等[19-20]，计算各属性对报废率影响得分，优选平均得分大于一定阈值(比如 0.15) 的属性为预测模型输入。

(5) BPN 预测模型构建：基于聚类样本及其优选属性，设置训练样本和测试样本，以相应训练样本开展模型训练。BPN 网络模型设置如下。

输入和输出：输入为归一化后各分组优选属性数据；输出为归一化处理后各样本预测报废率。

隐藏层设置：单一隐藏层，相应节点数采用较为常见的(输入节点数 + 输出节点数) /2 计算。可以看出其隐层节点数取决于各分组所选属性，较大的隐藏节点数一般能提高模型的非线性适应能力和预测精准度，在此将各分组 BPN 隐层节点统一设置为 15(各分组所选属性最多的一组为 28 个)。

激活函数：研究表明，BP 神经网络在其隐藏层采用 Sigmoid 函数即 $f(x) = 1/(1 + e^{-x})$，输出层采用线性函数 $f(x) = x$，只要隐含层中有足够的神经元，就几乎可以以任意精度拟合任何函数[21]。

学习率：0.05。

终止条件：最大迭代次数大于 25 000。

(6) 预测投入 Panel 数转换：反归一化计算确定预测报废率($Scrar\_Pd$)，然后采用 $Fedq\_Pd = \frac{100 \times Reqq}{(100 - Scrar\_Pd)}$、$Fedp\_Pd = \left\lceil \frac{Fedq\_Pd}{Duap} \right\rceil$ 转换计算预测投入 Panel 数，其中 $Scrar\_Pd$ 为预测报废率、$Reqq$ 为要求生产数量、$Duap$ 为生产拼板数、$Fedp\_Pd$ 为预测投入 $Panel$ 数。

SOM 通过(欧氏) 距离判断样本之间的相似性。学习过程中，输入样本找到与之距离最短的竞争层单元(获胜神经元)，并对其更新。同时，将邻近区域的权值更新。具体聚类流程如下：

(1) 网络初始化：用 0 ~ 1 之间随机数初始化输入层与竞争层之间权值矩阵 $w_{ij}(i = 1,2,\cdots,17, j = 1,2,\cdots,6$，表示竞争层第 i 个神经元与输入层第 j 个神经元之间的连接权重)。设定初始邻域 $N_c(0) = 2$，学习速率 $\eta(0) = 1/3$，最大迭代次数 $T = 500$，当前迭代次数 $t = 1$。

(2) 输入：从样本集中随机选取一个样本 $X = (x_1, x_2, \cdots, x_{17})$ 并计算 $\hat{X} = \frac{X}{\|X\|} = \left(\frac{x_1}{\sqrt{\sum_{i=1}^{17} x_i^2}}, \frac{x_2}{\sqrt{\sum_{i=1}^{17} x_i^2}}, \cdots, \frac{x_{17}}{\sqrt{\sum_{i=1}^{17} x_i^2}} \right)$。

(3) 计算获胜神经元：竞争层所有神经元对应权向量均与输入向量进行比较，最短距离的权向量为获胜神经元，即 $\|\hat{X} - W_{j^*}\| = \min_{j \in \{1,2,\cdots,6\}} \sqrt{\sum_{i=1}^{17} (x_i - w_{ij})^2}$。

(4) 权值、邻域和学习率更新：以 j^* 为中心确定 t 时刻的权值调整域。

$$\begin{cases}\hat{W}_j(t+1)=\hat{W}_j(t)+\Delta W_j=\hat{W}_j(t)+\eta(t)(\hat{X}-\hat{W}_j) & j\in N_c(t)\\ W_j(t+1)=\hat{W}_j(t) & j\notin N_c(t)\end{cases}$$

更新 $\eta(t)=e^{-N_c(t-1)}/(t+2)$，$N_c(t)=\left\lceil N_c(t-1)\times\left(1-\frac{t}{T}\right)\right\rceil$，其中 $\lceil\rceil$ 表示向上取整。

(5) 基于样本的学习: 随机抽取新样本，返回步骤 2，完成步骤 2 - 步骤 4，直至全部样本完成上述迭代。

(6) 终止条件判断: 若 $t<T$，则 $t=t+1$，返回步骤 2; 否则迭代结束。

3 结果分析

SOM-BPN 模型采用 Python 3.6 开发实现。基于 SOM 聚类后的 29 157 条样本被划分为 3 组，分别以 C1、C2 和 C3 标识，各聚类分组中样本规模分别为 12 992、6 674 和 9 491。因各分组内向上圆整 Panel 数、要求生产数量和外层最小线宽取值差异较大，绘制样本在上述三维空间的分布如图 2 所示。图 3 给出了不同聚类分组输入属性均值。

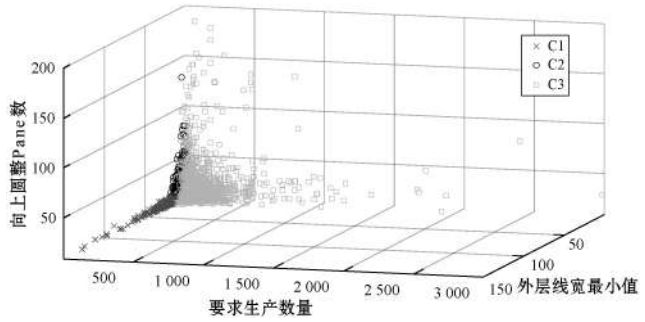
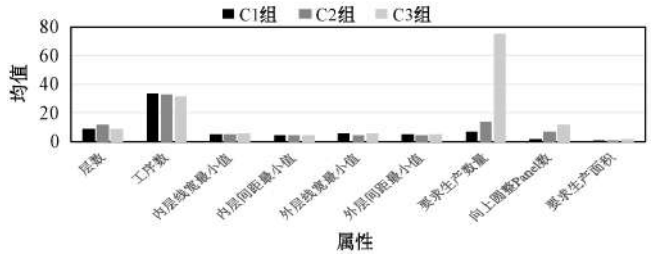
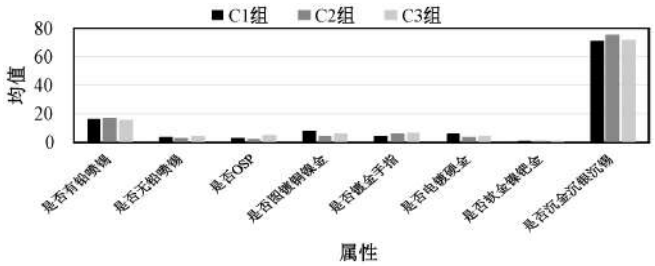


图 2 聚类结果分布



(a) 非二分类型属性均值对比



(b) 二分类型属性均值对比

图 3 各聚类分组中相应属性均值比较

三组样本的订单规模(要求生产数量、向上圆整 Panel 数和要求生产面积)均值差异较大，是区分和识别每个分组内样本差异的主要属性，与工厂实践一致，车间也是将订单规模视为重要变量。C2 中外层最小值线宽/间距均低于 C1 和 C3 中样本相应值，但层数均值更高，说明层数越高相应线路越密，这与实际一致。

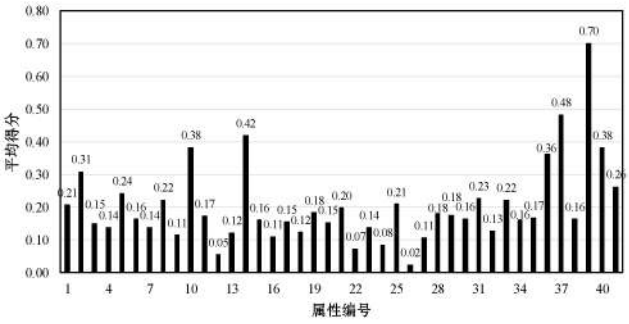
基于聚类分组样本，以表 2 给出的 41 个属性为输入，以报废率为预测目标，基于前述特征选择机制计算各属性对报废率影响得分，计算出各分组样本和所有样本各属性重要性得分均值，如图 4 所示，其中对应编号同表 2 给定编号。可以看出，不同样本组关键影响属性存在较大差异，原因之一是模型可能存在多个复杂分布[2]。

表 2 不同样本分组优选属性

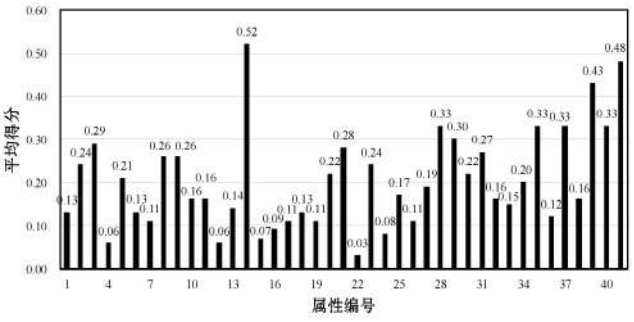
编号	属性	C1	C2	C3	全体
1	板厚	▲		▲	
2	层数	▲	▲	▲	▲
3	是否罗杰斯材料	▲	▲	▲	▲
4	板镀次数				
5	工序数	▲	▲	▲	▲
6	半固化片数	▲			
7	每 SET 最多允许报废单元数				
8	是否光电板	▲	▲	▲	
9	是否高频板		▲		
10	是否半导体测试板	▲	▲		▲
11	是否有负片电镀	▲	▲	▲	
12	是否有减薄铜			▲	
13	是否 III 级验收标准			▲	
14	是否华为验收标准	▲	▲	▲	▲
15	内层最小线宽	▲			
16	内层最小间距				
17	外层最小线宽	▲			
18	外层最小间距				
19	芯板残铜率均值	▲			
20	是否有阻焊塞孔	▲	▲	▲	▲
21	是否树脂塞孔	▲	▲	▲	▲
22	是否有二钻			▲	
23	是否有背钻		▲		▲
24	是否字符-打印机生产				
25	油墨颜色是否白色	▲	▲		
26	油墨颜色是否蓝色				

续表 2

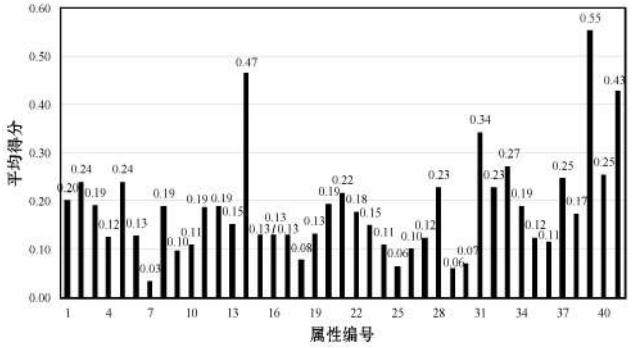
编号	属性	C1	C2	C3	全体
27	油墨颜色是否黑色		▲		
28	是否有铅喷锡	▲	▲	▲	
29	是否无铅喷锡	▲	▲		
30	是否 OSP	▲	▲		
31	是否图镀铜镍金	▲	▲	▲	▲
32	是否镀金手指		▲	▲	
33	是否电镀硬金	▲	▲	▲	
34	是否软金镍钯金	▲	▲	▲	
35	是否沉金沉银沉锡	▲	▲		
36	生产拼板数	▲			▲
37	要求生产数量	▲	▲	▲	▲
38	向上圆整 Panel 数	▲	▲	▲	▲
39	成品单元面积	▲	▲	▲	▲
40	要求生产面积	▲	▲	▲	▲
41	历史良率	▲	▲	▲	▲



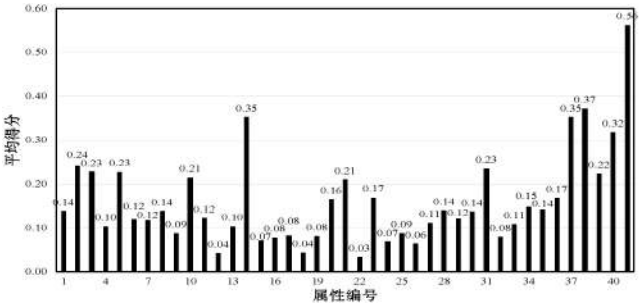
(a) C1 组样本属性重要性得分均值



(b) C2 组样本属性重要性得分均值



(c) C3 组样本属性重要性得分均值

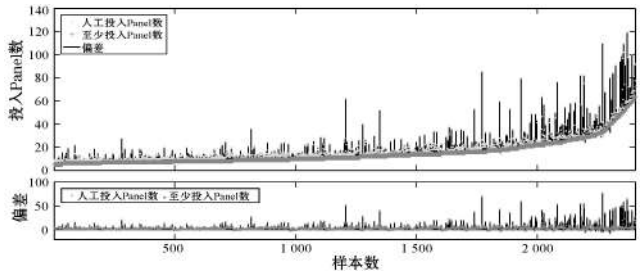


(d) 全体样本属性重要性得分均值

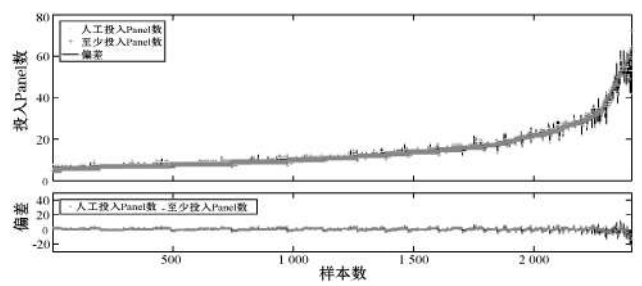
图 4　样本属性重要性得分均值

在此优选其重要性得分均值大于 0.15 的属性作为各 BP 网络预测模型的输入，各分组样本相应预测模型所优选属性在表 2 中以“▲”标识。可以看出，C1、C2、C3 组和全体样本选择属性数分别为 28、26、22 和 16。不同聚类组所选属性存在一定的差异，但层数、罗杰斯材料、工序数、华为验收标准、树脂塞孔、阻焊塞孔、背钻、图镀铜镍金、软金镍钯金、成品单元面积、要求生产数量/面积、向上圆整 Panel 数、历史良率等对各分组报废率均具有关键影响。

随机选择每组中 70% 样本用以训练相应 BPN 网络，剩余 30% 样本作为测试样本。其中 C1、C2、C3 以及全体数据中相应训练样本规模分别为 9 094、4 672、6 644 和 20 410，测试样本规模分别为 3 898、2 002、2 847 和 8 747。基于优选属性分组训练报废率预测模型，并将其转换为预测投入 Panel 数。图 5、图 6 分别为针对测试样本人工投人 Panel 数(车间实际投料方式)和基于 SOM-BP 预测投入 Panel 数与至少投入 Panel 数的偏差对比和回归图。可以看出，人工投料存在明显超投，基于 SOM-BP 预测机制能进一步降低车间因超投 Panel 导致的余数入库，从而降低车间因冗余带来的物料、生产、库存等浪费。

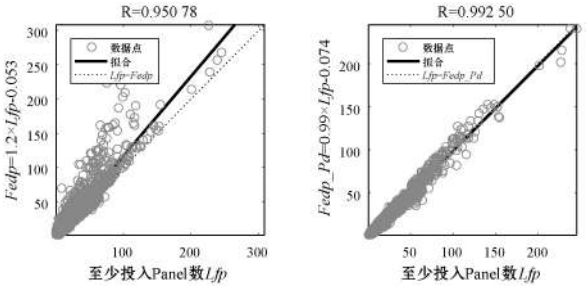


(a) 人工投入 Panel 数与至少投入 Panel 数偏差



(b) SOM-BP 预测投入 Panel 数与至少投入 Panel 数偏差

图 5　测试样本 Panel 数偏差



(a) 人工投入 Panel 数与至少投入 Panel 数拟合　(b) SOM-BP 预测投入 Panel 数与至少投入 Panel 数拟合

图 6　测试样本 Panel 数回归结果

进一步以预测投入 Panel 数与至少投入 Panel 数的均方误差(MSE)、绝对平均误差(MAE)和平均绝对百分比误差(MAPE)为评价指标判断相应预测效果，指标定义参考文献[2]。同时，在此以全部训练样板为输入，以表 2 中“全体”列中所优选属性为输入构建单一 BPN 预测模型。同时与基于单一参数(要求生产数量)划分样本后分组构建 BP 预测模型的 MSC-ANN[2]进行对比，对比结果如表 3 所示。可以看出 SOM-BPN 能明显降低 MSE、MAE 和 MAPE 的误差。其原因可能是基于分类划分和分组属性优选，尽可能地降低了同组内样本分布差异，从而提高模型预测精准度。

表 3　不同预测模型 MSE、MAE、MAPE 对比

预测模型	样本	MSE	MAE	MAPE
手工投料	训练样本	22.120	1.403	27.504
	测试样本	22.862	1.467	29.161
	所有样本	22.342	1.422	28.538
单一 BP 预测模型	训练样本	1.913	0.740	16.844
	测试样本	2.143	0.759	17.962
	全体样本	2.101	0.751	17.238
MSC-ANN	训练样本	0.719	0.330	5.402
	测试样本	1.272	0.396	5.687
	全体样本	0.872	0.349	5.522
SOM-BPN	训练样本	0.661	0.347	5.473
	测试样本	0.901	0.403	5.586
	所有样本	0.733	0.364	5.507

SOM-BPN 与 MSC-ANN 对比结果显示 SOM-BPN 所得 MSE 和 MAPE 指标优于 MSC-ANN，其中 MAE 稍高于 MSC-ANN 所得对应值。但是 MSC-ANN 将样本划分为 6 组，分别构建了 6 个预测模型，在前期模型训练构建、后续实施维护等方面均需要投入更多的人力物力，而 SOM-BPN 只需要分组构建三个预测模型，所以基于 SOM-BPN 的模型在优化人力投入上具有明显优势。

结合车间具体需求，最终考核指标一般为余数入库率和补投率，基于文献[2]中式(4)－式(11)转换计算余数入库率(Surpr_Pd)和补投率(Supfr Pd)，不同算法对比结果如表 4 所示。与投料相比，SOM-BPN 预测模型可同时降低余数入库率和补投率；其中前者从 27.44% 下降到 10.13%，后者从 17.91% 下降到 9.37%。另外，未经聚类的单一 BP 预测模型余数入库率和加投率明显高于 SOM-BPN 模型。同时，SOM-BPN 优于 MSC-ANN 所得结果，进一步证明本文模型在减少模型数量的同时可进一步优化车间投料，降低余数入库和补投带来的损失。

表 4　不同预测模型余数入库率和补投率的对比

预测模型	样本	余数入库率	补投率
手工投料	训练样本	26.57	17.06
	测试样本	28.49	18.53
	所有样本	27.95	17.91
单一 BP 预测模型	训练样本	16.21	12.52
	测试样本	16.85	13.02
	全体样本	16.53	12.89
MSC-ANN	训练样本	11.80	11.68
	测试样本	12.11	12.03
	全体样本	11.96	11.91
SOM-BPN	训练样本	9.26	9.01
	测试样本	10.83	9.68
	所有样本	10.13	9.37

4　结　语

本文结合 SOM 和 BPN 建立了基于 SOM-BPN 的 PCB 投料分组预测模型。SOM-BPN 较手工投料、单一 BP 预测模型能获取更低的 MSE、MAE、MAPE 以及与预测余数入库率(Surpr_Pd)和补投率(Supfr_Pd)；与 MSC-ANN 比较，SOM-BP 能获得更低的 MSE、Surpr_Pd 和 Supfr_Pd，且 MSC-ANN 需要训练、构建和维护 6 个预测模型，而 SOM-BP 只需维护 3 个。单一 BPN 预测模型将 Surpr_Pd 和 Supfr_Pd 从 27.95% 和 17.91% 分别降低至 16.53% 和 12.89%；MSC-ANN 将其降低至 11.96% 和 11.91%；而 SOM-BPN 分别将其降低至 10.13% 和 9.37%。这表明 SOM-BPN 可进一步降低因超/补投带来的损失。综合样本不同分布特点的分

组、优选的关键属性、基于优选属性的分组预测模型构建及其转换可为其他 PCB 样板厂投料优化提供参考。

直接基于多样样板和影响质量全因素自动分组划分样本，提取组内共享特征并训练相应预测模型，实施应用时能自动优选各样板最合适预测模型仍有待进一步深入研究。

参 考 文 献

[1] 吕盛坪，乐强生，刘涛. 基于数据挖掘的印制电路样板投料优化[J]. 系统仿真学报，2018，30(7)：2656－2665.

[2] Lü S P，Zheng B B，Kim H，et al. Data mining for material feeding optimization of printed circuit board template production[J]. Journal of Electrical and Computer Engineering，2018，2018：1852938.1－1852938.17.

[3] 吕佑龙，张洁. 基于大数据的智慧工厂技术框架[J]. 计算机集成制造系统，2016，22(11)：2691－2697.

[4] 张卫，丁金福，纪杨建，等. 工业大数据环境下的智能服务模块化设计[J]. 中国机械工程，2019，30(2)：167－173.

[5] 张洁，汪俊亮，吕佑龙，等. 大数据驱动的智能制造[J]. 中国机械工程，2019，30(2)：127－133.

[6] 姚锡凡，周佳军，张存吉，等. 主动制造—大数据驱动的新兴制造范式[J]. 计算机集成制造系统，2017，23(1)：172－185.

[7] 张洁，高亮，秦威，等. 大数据驱动的智能车间运行分析与决策方法体系[J]. 计算机集成制造系统，2016，22(5)：1221－1229.

[8] Li J，Tao F，Cheng Y，et al. Big data in product lifecycle management[J]. Journal of Advanced Manufacturing Technology，2015，81(5)：667－684.

[9] Ren S，Zhang Y，Liu Y，et al. A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing：A framework，challenges and future research directions[J]. Journal of Cleaner Production，2019，210：1343－1365.

[10] Koksal G，Batmaz I，Testik M C. A review of data mining applications for quality improvement in manufacturing industry[J]. Expert Systems with Applications，2011，38(10)：13448－13467.

[11] Khader N，Yoon S W，Li D B. Stencil printing optimization using a hybrid of support vector regression and mixed-integer linear programming[J]. Procedia manufacturing，2017，11：1809－1817.

[12] Liukkonen M，Hiltunen T，Havia E，et al. Modeling of soldering quality by using artificial neural networks[J]. IEEE Transaction on Electronics Packaging Manufacturing，2009，32(2)：89－96.

[13] Barajas L G，Egerstedt M B，Kamen E W，et al. Stencil printing process modeling and control using statistical neural networks[J]. IEEE Transaction on Electronics Packaging Manufacturing，2008，31(1)：9－18.

[14] Tsai T，Liukkonen M. Robust parameter design for the micro-BGA stencil printing process using a fuzzy logic-based Taguchi method[J]. Applied Soft Computing，2016，48(6)：124－136.

[15] Tsai T. Thermal parameters optimization of a reflow soldering profile in printed circuit board assembly：A comparative study[J]. Applied Soft Computing，2012，12(8)：2601－2613.

[16] Chan K Y，Kwong C K，Tsim Y C. Modelling and optimization of fluid dispensing for electronic packaging using neural fuzzy networks and genetic algorithms[J]. Engineering Applications of Artificial Intelligence，2010，23(1)：18－26.

[17] Stoyanov S，Bailey C，Tourloukis G. Similarity approach for reducing qualification tests of electronic components[J]. Microelectronics Reliability，2016，67：111－119.

[18] Liukkonen M，Havia E，Leinonen H，et al. Quality-oriented optimization of wave soldering process by using self-organizing maps[J]. Applied Soft Computing，2011，11(1)：214－220.

[19] Li Y，Li T，Liu H. Recent advances in feature selection and its applications[J]. Knowledge and Information Systems，2017，53(3)：551－577.

[20] Aldehim G，Wang W J. Determining appropriate approaches for using data in feature selection[J]. International Journal of Machine Learning and Cybernetics，2017，8：915－928.

[21] Sonoda S，Murata N. Neural network with unbounded activation functions is universal approximator[J]. Applied and Computational Harmonic Analysis，2017，43(2)：233－268.

(上接第 32 页)

[6] 张同杨. 基于 CXF 的 SOA 应用设计与实现[J]. 价值工程，2017，36(32)：83－87. [7] 徐金红. 用 BCP 程序解决新老校区 MELINETS 系统的数据平衡[J]. 情报杂志，2004，23(7)：21－23.

[8] 覃雄派，王会举，杜小勇，等. 大数据分析——RDBMS 与 MapReduce 的竞争与共生[J]. 软件学报，2012，23(1)：55－56.

[9] 卢小宾，徐超. 面向风险管理的银行大数据分析系统架构研究[J]. 信息资源管理学报，2018，29(2)：6－14.

[10] Krnac L. Pivotal certified spring enterprise integration specialist exam[M]. Berkeley：Apress，2015.

[11] 王德俊，黄林鹏，徐小辉，等. 事务控制的面向服务系统的动态更新协调[J]. 软件学报，2011，22(11)：2652－2667.

三、科研成果---授权发明专利清单

证书号第1758441号

北京市海淀区地方税务局 印花税代扣专用章 国家知识产权局

发明专利证书

发明名称：一种浸泡喷淋复合型果蔬预冷装置

发 明 人：吕盛坪;方思贞;陆华忠;吕恩利;岑康华

专 利 号：ZL 2014 1 0025923.2

专利申请日：2014年01月20日

专利权人：华南农业大学

授权公告日：2015年08月19日

本发明经过本局依照中华人民共和国专利法进行审查，决定授予专利权，颁发本证书并在专利登记簿上予以登记。专利权自授权公告之日起生效。

本专利的专利权期限为二十年，自申请日起算。专利权人应当依照专利法及其实施细则规定缴纳年费。本专利的年费应当在每年01月20日前缴纳。未按照规定缴纳年费的，专利权自应当缴纳年费期满之日起终止。

专利证书记载专利权登记时的法律状况。专利权的转移、质押、无效、终止、恢复和专利权人的姓名或名称、国籍、地址变更等事项记载在专利登记簿上。

局长
申长雨

中华人民共和国国家知识产权局

2015年08月19日

证书号第1876137号

发明专利证书

发明名称：一种基于流化冰的果蔬用预冷装置及其预冷方法

发明人：吕盛坪;方思贞;陆华忠;吕恩利;岑康华

专利号：ZL 2014 1 0025922.8

专利申请日：2014年01月20日

专利权人：华南农业大学

授权公告日：2015年12月09日

本发明经过本局依照中华人民共和国专利法进行审查，决定授予专利权，颁发本证书并在专利登记簿上予以登记。专利权自授权公告之日起生效。

本专利的专利权期限为二十年，自申请日起算。专利权人应当依照专利法及其实施细则规定缴纳年费。本专利的年费应当在每年01月20日前缴纳。未按照规定缴纳年费的，专利权自应当缴纳年费期满之日起终止。

专利证书记载专利权登记时的法律状况。专利权的转移、质押、无效、终止、恢复和专利权人的姓名或名称、国籍、地址变更等事项记载在专利登记簿上。

局长
申长雨

中华人民共和国国家知识产权局
2015年12月09日

证书号第3446775号

发明专利证书

发明名称：应用于柔性工艺过程规划的约束关系描述与可行工艺方案解析生成方法

发 明 人：吕盛坪;方思贞;杨径;王飞仁;徐岩

专 利 号：ZL 2016 1 0571834.7

专利申请日：2016年07月19日

专利权人：华南农业大学

地 址：510642 广东省广州市天河区五山路483号

授权公告日：2019年07月09日　　授权公告号：CN 106251003 B

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，颁发发明专利证书并在专利登记簿上予以登记。专利权自授权公告之日起生效。专利权期限为二十年，自申请日起算。

专利证书记载专利权登记时的法律状况。专利权的转移、质押、无效、终止、恢复和专利权人的姓名或名称、国籍、地址变更等事项记载在专利登记簿上。

局长
申长雨

国家知识产权局
2019年07月09日

其他事项参见背面

证书号第4847985号

发明专利证书

发明名称：数据挖掘课程教学实践系统和基于系统的教学实践方法

发明人：吕盛坪;罗勇;廖鑫婷;江城;朱紫纯;李灯辉;冼荣亨

专利号：ZL 2020 1 0150693.8

专利申请日：2020年03月06日

专利权人：华南农业大学

地址：510642 广东省广州市天河区五山路483号

授权公告日：2021年12月14日　　授权公告号：CN 111260969 B

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，颁发发明专利证书并在专利登记簿上予以登记。专利权自授权公告之日起生效。专利权期限为二十年，自申请日起算。

专利证书记载专利权登记时的法律状况。专利权的转移、质押、无效、终止、恢复和专利权人的姓名或名称、国籍、地址变更等事项记载在专利登记簿上。

局长

申长雨

国家知识产权局

2021年12月14日

其他事项参见续页

证书号第6238938号

发明专利证书

发明名称：一种分离编带的元器件的收纳装置

发明人：吕盛坪;李鑫;熊伟;信德全;朱紫纯;陈恒旭;劳景春
李文强;欧阳斌;赵贺杰;张胡成;江城;何海平

专利号：ZL 2021 1 1344764.9

专利申请日：2021年11月15日

专利权人：华南农业大学

地址：510642 广东省广州市天河区五山路483号

授权公告日：2023年08月15日　　授权公告号：CN 114021687 B

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，颁发发明专利证书并在专利登记簿上予以登记。专利权自授权公告之日起生效。专利权期限为二十年，自申请日起算。

专利证书记载专利权登记时的法律状况。专利权的转移、质押、无效、终止、恢复和专利权人的姓名或名称、国籍、地址变更等事项记载在专利登记簿上。

局长
申长雨

国家知识产权局
2023年08月15日

其他事项参见续页

证书号第8229354号

专利公告信息

发明专利证书

发明名称：一种高反光物体图像修复方法

专利权人：华南农业大学;佛山显扬科技有限公司

地　　址：510642 广东省广州市天河区五山路483号

发 明 人：吕盛坪;熊伟;李鑫;金鸿;丁克;信德全;赵贺杰;欧阳斌
张胡成

专 利 号：ZL 2022 1 1363259.3　　授权公告号：CN 115829860 B

专利申请日：2022年11月02日　　授权公告日：2025年09月05日

申请日时申请人：华南农业大学;佛山显扬科技有限公司

申请日时发明人：吕盛坪;熊伟;李鑫;金鸿;丁克;信德全;赵贺杰;欧阳斌
张胡成

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，并予以公告。

专利权自授权公告之日起生效。专利权有效性及专利权人变更等法律信息以专利登记簿记载为准。

局长
申长雨

国家知识产权局
2025年09月05日

证书号第8512916号

专利公告信息

发明专利证书

发明名称：一种基于改进NSGA-III的广义作业车间调度方法

专利权人：华南农业大学

地　　址：510642 广东省广州市天河区五山路483号

发 明 人：吕盛坪;赵贺杰;邹建民;宁韬涛;庄剑威;梁泰然;黎焯辉
张凯彬;陈健宇

专　利　号：ZL 2024 1 0141376.8　　授权公告号：CN 117875669 B

专利申请日：2024年01月31日　　授权公告日：2025年11月25日

申请日时申请人：华南农业大学

申请日时发明人：吕盛坪;赵贺杰;邹建民;宁韬涛;庄剑威;梁泰然;黎焯辉
张凯彬;陈健宇

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，并予以公告。

专利权自授权公告之日起生效。专利权有效性及专利权人变更等法律信息以专利登记簿记载为准。

局长
申长雨

国家知识产权局
2025年11月25日

三、科研成果---软件著作权清单

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2461160号

软 件 名 称：工艺规划与车间调度紧耦合集成系统 V1.0

著 作 权 人：华南农业大学

开发完成日期：2017年09月15日

首次发表日期：2017年09月23日

权利取得方式：原始取得

权 利 范 围：全部权利

登 记 号：2018SR132065

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权
登记专用章
2018年02月28日

No. 02351646

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第2461167号

软 件 名 称：电子产品测试车间调度系统 V1.0

著 作 权 人：华南农业大学

开发完成日期：2017年08月11日

首次发表日期：2017年09月01日

权利取得方式：原始取得

权 利 范 围：全部权利

登 记 号：2018SR132072

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权登记专用章
2018年02月28日

No. 02351647

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第3419033号

软 件 名 称：PCB样板投料优化软件 V1.0

著 作 权 人：华南农业大学

开发完成日期：2018年05月03日

首次发表日期：2018年09月10日

权利取得方式：原始取得

权 利 范 围：全部权利

登 记 号：2018SR1089938

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局 计算机软件著作权登记专用章

No. 03373011

2018年12月28日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号： 软著登字第4637439号

软 件 名 称： 数据挖掘教学示范与学生实践系统
[简称：SDMTDSP]
V1.0

著 作 权 人： 华南农业大学

开发完成日期： 2018年08月25日

首次发表日期： 2018年08月25日

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2019SR1216682

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权
登记专用章

No. 04854491

2019年11月26日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号： 软著登字第9532865号

软 件 名 称： 基于空间力系对三维点云模型的力学分析软件 V1.0

著 作 权 人： 华南农业大学

开发完成日期： 2022年03月24日

首次发表日期： 未发表

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2022SR0578666

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局 计算机软件著作权登记专用章

No. 10699244

2022年05月12日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号： 软著登字第9532866号

软 件 名 称： 面向机器视觉的工业相机智能匹配软件 V1.0

著 作 权 人： 华南农业大学

开发完成日期： 2022年03月24日

首次发表日期： 未发表

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2022SR0578667

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局 计算机软件著作权登记专用章

No. 10699245

2022年05月12日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号： 软著登字第10389569号

软 件 名 称： 基于改进Knn算法的邮政编码快速识别软件 V1.0

著 作 权 人： 华南农业大学

开发完成日期： 2022年08月17日

首次发表日期： 未发表

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2022SR1435370

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局 计算机软件著作权登记专用章

No. 11780948

2022年10月31日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号： 软著登字第10389522号

软 件 名 称： 基于深度卷积神经网络的形状识别软件 V1.0

著 作 权 人： 华南农业大学

开发完成日期： 2022年08月17日

首次发表日期： 未发表

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2022SR1435323

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

No. 11780947

中华人民共和国国家版权局 计算机软件著作权登记专用章

2022年10月31日

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号： 软著登字第10396808号

软 件 名 称： 计算机视觉技术教学示范与学生实践平台软件 V1.0

著 作 权 人： 华南农业大学

开发完成日期： 2022年08月17日

首次发表日期： 未发表

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2022SR1442609

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局 计算机软件著作权登记专用章

No. 11790147

2022年11月01日

中华人民共和国国家版权局
计算机软件著作权登记证书

证书号： 软著登字第17089662号

软 件 名 称： 考虑工人与并行机的柔性作业车间调度处理软件V1.0

著 作 权 人： 华南农业大学

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2025SR2433464

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局 计算机软件著作权登记专用章

2025年12月17日

中华人民共和国国家版权局
计算机软件著作权登记证书

证书号： 软著登字第16808397号

软 件 名 称： 带批处理的绿色柔性作业车间调度系统软件 V1.0

著 作 权 人： 华南农业大学

权利取得方式： 原始取得

权 利 范 围： 全部权利

登 记 号： 2025SR2152199

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局 计算机软件著作权登记专用章

2025年11月05日