

文章编号: 1000-7032(2025)11-2138-12

加速钙钛矿太阳能电池开发: 基于机器学习驱动框架的SHAP分析

梁锐权¹, 刘 谦¹, 胡春华¹, 郑建楂², 李 阳^{1*}, 王亦文^{3*}, 麦耀华³

(1. 五邑大学 电子与信息工程学院, 广东 江门 529020;

2. 澳门大学 应用物理与材料工程学院, 中国 澳门 999078;

3. 暨南大学 物理与光电工程学院, 广东 广州 510632)

摘要: 钙钛矿太阳能电池(Perovskite solar cells, PSCs)以其高性能在新型太阳能电池技术中占据了重要地位。针对传统实验试错法在提升PSCs的光电转换效率(Power conversion efficiency, PCE)方面存在研究周期长和效率低下的不足,提出了一种基于机器学习(Machine learning, ML)的PSCs制备工艺的智能优化方法。通过应用多种ML算法构建PCE预测模型,最终选择性能卓越的梯度提升(Gradient boosting, GB)模型进行夏普利加性解释(Shapley additive explanations, SHAP)可视化分析和实验验证。实验结果表明,基于模型输出指导的实验使PSCs的PCE达到了21.81%。本工作不仅有效解决了传统实验试错法的局限性,还攻克了ML在PSCs领域应用中预测精度低的难题,为快速开发高PCE的PSCs提供了新视角和科学依据,也为其他新型太阳能电池技术的开发提供了参考。

关键词: 钙钛矿太阳能电池; 机器学习; 夏普利加性解释(SHAP)分析; 光电转换效率预测

中图分类号: TM914.4 文献标识码: A

DOI: 10.37188/CJL.20250149 CSTR: 32170.14.CJL.20250149

Accelerating Perovskite Solar Cell Development: A Machine Learning-driven Framework with SHAP Explainability

LIANG Ruiquan¹, LIU Qian¹, HU Chunhua¹, ZHENG Jianzha², LI Yang^{1*},WANG Yiwen^{3*}, MAI Yaohua³

(1. School of Electronic and Information Engineering, Wuyi University, Jiangmen 529020, China;

2. Institute of Applied Physics and Materials Engineering, University of Macao, Macao 999078, China;

3. College of Physics and Optoelectronic Engineering, Jinan University, Guangzhou 510632, China)

* Corresponding Authors, E-mail: insidesun51@163.com; wangyiwen@jnu.edu.cn

Abstract: Perovskite solar cells (PSCs) have garnered significant attention in the realm of innovative photovoltaic technologies due to their impressive performance. Traditional trial-and-error experimental methods often result in lengthy research cycles to enhance the power conversion efficiency (PCE) of PSCs. We propose a machine learning (ML)-based intelligent optimization strategy to accelerate research cycles in PSC fabrication. By applying various ML algorithms to develop PCE prediction models, the gradient boosting (Gradient boosting, GB) model was chosen for Shapley additive explanations (SHAP) visualization analysis and experimental validation. The experimental results revealed that the design guided by model predictions and SHAP analysis achieved a PCE of 21.81% for wide bandgap (1.65 eV) PSCs. This study effectively addresses the limitations of conventional trial-and-error approaches and overcomes the challenge of low predictive accuracy in ML applications within the PSC domain. It provides a new

收稿日期: 2025-05-13; 修订日期: 2025-05-29

基金项目: 国家自然科学基金青年科学基金(62405113); 广州市科技计划项目(SL2024A04J00418); 广东省重点领域研发计划项目(2019B010132004)

Supported by National Natural Science Foundation of China (NSFC) Young Scientist Fund (62405113); Guangzhou Science and Technology Program Project (SL2024A04J00418); Key Realm R&D Program of Guangdong Province (2019B010132004)

perspective and scientific basis for the rapid development of high-PCE PSCs, and also offers a reference for the development of other new solar cell technologies.

Key words: perovskite solar cells; machine learning; Shapley additive explanations (SHAP) analysis; power conversion efficiency (PCE) prediction

1 引 言

自 2009 年以来,钙钛矿太阳能电池(Perovskite solar cells, PSCs)已成为新型光伏技术领域的重要研究方向^[1]。钙钛矿材料凭借其可调节的带隙、优异的光吸收特性和简单的制备工艺,使其特别适用于光伏器件^[2-7]。如今,单结 PSCs 的最高光电转换效率(Power conversion efficiency, PCE)已突破 27%^[8],展现出巨大的商业化潜力^[9]。尽管单结 PSCs 在提高 PCE 方面表现出色,其 PCE 仍然受到光谱吸收范围的限制。为了突破这一瓶颈,叠层太阳能电池(Tandem solar cells, TSCs)近年来得到了广泛关注^[10]。与单结太阳能电池相比, TSCs 通过堆叠具有不同带隙的多层吸光层,充分利用太阳光谱。宽带隙(Wide band gap, WBG) PSCs (带隙大于 1.63 eV) 作为 TSCs 的顶电池,吸收高能紫外-可见光;而其他类型的电池,如硅、铜铟镓硒(Copper indium gallium selenide, CIGS)、碲化镉(Cadmium telluride, CdTe)和有机太阳能电池,则作为底电池,吸收低能红外光。这些电池的结合实现了更高效的 PCE^[11]。尽管 WBG PSCs 在实际应用方面展现出显著的商业潜力,但其实现高 PCE 且可重复的生产工艺仍面临技术挑战^[12]。

机器学习(Machine learning, ML)技术的发展为材料设计带来了革命性的变化, ML 不仅能加速材料设计的进程^[13-15],通过分析大规模数据集,还能加速材料的筛选及探索^[16-17],进行 PCE 预测^[18]等,从而节省试验时间和节约成本^[19]。2019 年, Li 及其团队^[20]利用 ML 技术研究了各种钙钛矿材料的带隙及其对 PSCs PCE 的影响。研究团队构建了一个 ML 模型,根据材料组成预测带隙、能级,成功预测了 PSCs 的 PCE,并与 Shockley-Queisser 极限的理论^[21]预测相符,为新型钙钛矿材料的合成提供了思路。2020 年, Stoddard 等^[22]提出一种基于光学透射和反射暗场成像的预测方法,以评估 PSCs 的 PCE 衰退。研究发现, MAPbI₃ 材料的载流子扩散长度在准费米能级分裂之前就开始下降。通过 ML 算法,研究者能够以 12.8% 的平均误差预测载流子扩散长度降至初始值的 85% 所

需时间,并揭示了透射率变化速率与传输损失时间之间的强相关性。这项研究为钙钛矿材料和器件的稳定性评估提供了有效工具,展示了其在预测中的潜力。2022 年, Liu 等^[23]介绍了一个 ML 框架,通过顺序学习方法对 PSCs 的制备过程进行优化,成功地在有限的实验条件下实现了 18.5% 的 PCE,再次证明 ML 应用的巨大潜力。

2024 年,诺贝尔物理学奖首次授予人工智能领域的科学家,标志着人工智能在科学研究中的重要地位和影响^[24]。构建 PSCs 实验数据库后结合人工智能中的 ML 方法,为分析实验数据和评估不同制备条件的优劣提供了重要支持。随着实验次数和实验参数维度的增加,系统优化所有工艺条件的难度显著提升。然而,利用 ML 可以有效地应对这一挑战,通过对大量实验数据进行智能化分析,使得能够发现潜在的规律和优化的方向。随着实验次数的不断积累,以及实验条件的不断完善,更多尚未探索的制备工艺维度将被揭示,从而为进一步提升 PCE 提供新的方向。尽管 ML 在 PSCs 领域的应用已有一定进展,但仍存在许多尚未解决的问题,例如,将 ML 方法与实验数据紧密结合,开发新的模型以提高预测精度等。

本文以 WBG PSCs (钙钛矿材料主要成分为 Cs_{0.05}FA_{0.80}MA_{0.15}PbI_{2.25}Br_{0.75}, 带隙 1.65 eV) 为例,系统地阐述了一套通过 ML 优化 PSCs 制备工艺的流程。在数据采集阶段,制定了严格的工艺参数标准和操作规程,建立了详尽的合成工艺数据库,以确保数据的全面性和一致性。在此基础上,通过特征工程技术识别出钙钛矿材料合成过程中潜在的优化点,包括离散型特征和连续型特征,构建了一个详尽的合成工艺数据库。随后,进一步应用多种 ML 算法来构建预测 PCE 的模型,并选择最优的 ML 模型。最后,通过解释模型输出并设计实验来验证模型的预测。实验结果表明,基于模型输出指导的实验设计显著提升了 PSCs 的 PCE。该工作通过优化数据收集与预处理流程,有效提升了数据质量,同时应对 ML 在 PSCs 领域中预测精度不足以及传统实验方法研发周期长等

关键问题。本研究不仅为开发高 PCE 的 PSCs 提供了新的视角和科学依据,也为该领域的研究和开发工作带来了创新思路。

2 方 法

本研究工作流程图如图 1 所示。为解决实验的不确定性问题导致的数据质量问题,初步筛选出不符合条件的数据。这样做有助于提高数据质量,确保所分析数据的可靠性和代表性,排除人为因素和环境干扰对实验数据的影响,提高数据的质量。筛选完成后,对保留下来的数据进行了深入分析,并对异常值进行进一步检查,删除了异常值。随后,对实验数据进行了标准化处理。对于离散型变量,我们对每个分类特征进行了编码操作,采用了独热编码^[25](One-hot encoding)的方式。独热编码是一种将每个类别转换为一组二进制向量的技术,其原理是为每一个可能的类别创建一个新的特征列,类别出现的位置为 1,其余为 0,从而有效避免了类别间的虚假顺序关系。在初步实验阶段,我们还尝试了标签编码(Label encoding)和目标编码(Target encoding)等方法。标签编码将每个类别映射为一个整数,虽然简单高效,但会引入类别之间的顺序信息,不适合无序分类变量;目标编码则是将每个类别映射为其对应的目标变量的统计值,如均值,适合类别数较多的情况,但容易产生目标泄漏风险。相比之下,独热编码在模型的性能、稳定性以及特征可解释性方面表现更优,因此我们最终选用了该方法。对于连续型数据,采用了最大最小归一化方法^[26]。该方法可以消除不同特征量纲和范围的差异,使其在分析和建模中具有同等的权重,尤其适用于距离度量敏感的算法(如 KNN 和 SVM 等)。经过独热编码和归一化处理后,数据的结构更加清晰且具备一致性,为后续模型的构建和特征工程奠定了坚实基础。同时,为探究数据不平衡问题对模型性能的影响,采用 SMOTE-Tomek 方法^[27]进行了研究。SMOTE-Tomek 是一种结合 SMOTE(合成少数类过采样技术)和 Tomek 链接的数据平衡方法,旨在通过生成新样本和清理噪声样本同时解决类别不平衡问题。SMOTE 通过在少数类样本之间合成新的数据点来扩展少数类表示,而不是简单地复制现有样本,它通过计算每个少数类样本的若干邻居,并在选定样本和其邻居之间的连线上随机生成合成样本。Tomek 链接则识别出类别边界

附近的样本对,并移除这些可能导致模型混淆的噪声样本。SMOTE-Tomek 在生成少数类样本的同时清理了数据集中的冗余和噪声点,非常适合需要增强少数类并提高数据质量的场景。

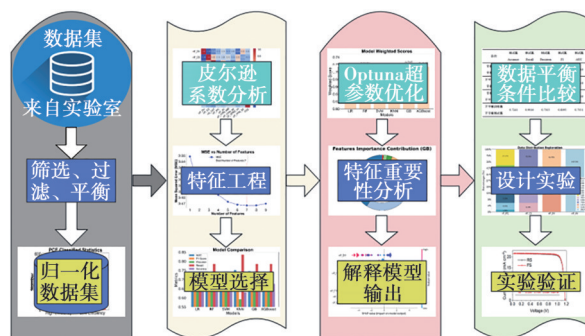


图 1 基于 ML 技术优化制备工艺提升 PCE 的工作流程

Fig.1 Workflow for optimizing preparation processes to enhance PCE based on ML techniques

为了识别对 PCE 具有显著影响的关键特征,首先应用随机森林算法进行特征重要性评估。我们选择该方法主要基于其鲁棒性强、计算效率高、对特征间非线性关系处理能力强等优势,适合用于特征的初步筛选。该算法通过计算每个特征在决策树分裂过程中的重要性,并利用 Gini 不纯度^[28]来衡量特征对模型预测的贡献,能够量化每个特征对模型决策的影响。这一方法可帮助识别出最具预测能力的特征,从而为后续的模型优化和工艺改进提供了科学依据。接下来,为了进一步优化模型性能,基于皮尔逊相关系数(Pearson product-moment correlation coefficient, PCC)评估了特征间的相关性^[29]。计算公式如下:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}, \quad (1)$$

其中, X_i 和 Y_i 分别表示两个变量的第 i 个观测值, $\bar{X}$ 和 $\bar{Y}$ 分别表示两个变量的均值, n 为观测值的数量。PCC 是一种衡量两个变量之间线性相关性强度的统计指标,取值范围为 $[-1, 1]$, 其中正值表示正相关,负值表示负相关,绝对值越接近 1 表示相关性越强。考虑到高度相关的特征可能导致多重共线性问题,进而影响模型的解释性和泛化能力,同时冗余特征可能增加模型复杂度并引发过拟合风险,通过去除高相关性的特征来避免冗余信息的干扰。通过这一过程,优先删除了那些与其他特征高度相关和重要性较低的特征。为了进一步提

升模型性能并精确优化制备工艺, 采用了递归特征消除 (Recursive feature elimination, RFE) 方法^[30]。RFE 是一种迭代式的特征选择方法, 通过递归地训练模型并评估特征对模型性能的贡献, 逐步删除对预测性能影响较小的特征。在每轮迭代中, RFE 通过评估模型的表现 (如均方误差, MSE) 来确定哪些特征对目标变量的预测最为重要。借助这一过程, 能够精确地识别出对 PCE 最具影响力的制备条件, 从而更有效地调整制备工艺参数, 优化实验设计并提升 PSCs 的性能。最后, 借助了 Scikit-learn 工具包对筛选出的特征进行建模与训练, 应用逻辑回归 (Logistic regression, LR)、随机森林 (Random forest, RF)、支持向量机 (Support vector machine, SVM)、K 最近邻 (K-nearest neighbors, KNN)、梯度提升 (Gradient boosting, GB) 和极端梯度提升 (Extreme gradient boosting, XGBoost) 六种 ML 模型进行预测任务, 评估不同模型的性能和准确性。本文相关评价指标如下^[31]。

准确率 (Accuracy, A_{cc}) 是模型正确预测的样本占总样本的比例。公式 (2) 给出了 A_{cc} 的计算方法:

$$A_{cc} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}, \quad (2)$$

其中, N_{TP} 为模型正确预测为正类的样本数量 (True positives), N_{TN} 为模型错误地将负类预测为正类的样本数量 (True negatives), N_{FP} 为模型错误地将负类预测为正类的样本数量 (False positives), N_{FN} 为模型错误地将正类预测为负类的样本数量 (False negatives)。 A_{cc} 是一种直观且易于计算的评估指标, 适用于类别平衡的数据集, 因为它能够直接反映出模型正确分类的整体比例。

召回率 (Recall, R_{ec}) 也叫真正率, 是模型识别出所有正类样本的比例。公式 (3) 提供了 R_{ec} 的详细描述:

$$R_{ec} = \frac{N_{TP}}{N_{TP} + N_{FN}}, \quad (3)$$

R_{ec} 能够有效评估模型对正类的识别能力, 尤其在漏检正类样本代价较高时 (如疾病诊断) 非常重要。然而, R_{ec} 过高时往往会导致假正例增多, 降低模型的精确度, 甚至可能将大量负类样本错误地预测为正类, 从而影响模型的稳定性和实际应用效果。

精确率 (Precision, P_{re}) 是模型预测为正类的

样本中, 真正是正类的比例。根据公式 (4), P_{re} 的计算方法如下:

$$P_{re} = \frac{N_{TP}}{N_{TP} + N_{FP}}, \quad (4)$$

P_{re} 强调模型在预测为正类时的准确性, 适用于假正例代价较高的应用场景, 如垃圾邮件过滤和欺诈检测。然而, P_{re} 单独使用时并不能反映模型对正类的全面识别能力, 过度优化 P_{re} 可能导致漏掉正类样本, 进而影响 R_{ec} 。

F_1 值 (F_1 score, F_1) 是 P_{re} 和 R_{ec} 的调和平均值, 考虑到了二者的权衡, F_1 计算公式为:

$$F_1 = 2 \times \frac{P_{re} \times R_{ec}}{P_{re} + R_{ec}}, \quad (5)$$

F_1 值在考虑两者的同时, 避免了过度优化单一指标的问题, 在二分类问题中, 它有助于找到一个既不过度偏向 P_{re} 也不过度偏向 R_{ec} 的模型。

曲线下面积 (Area under the curve, A_{uc}) 指受试者工作特征曲线 (Receiver operating characteristic, ROC) 下的面积, 表示模型区分正负类样本的能力。 A_{uc} 越接近 1, 说明模型越优秀; A_{uc} 越接近 0.5, 说明模型的预测能力接近随机猜测。Scikit-learn 实现的 A_{uc} 计算基于梯形法则对 ROC 曲线进行数值积分。计算 ROC 曲线下的面积, 可以通过积分得到, 通常使用数值方法来计算。公式 (6) 给出了 A_{uc} 计算的具体表达式:

$$A_{uc} = \frac{1}{N_{pos} \cdot N_{neg}} \sum_{i=1}^{N_{pos}} \sum_{j=1}^{N_{neg}} \theta(y_i > y_j), \quad (6)$$

其中, N_{pos} 和 N_{neg} 分别代表正类和负类样本的数量; y_i 和 y_j 分别指第 i 个正类样本和第 j 个负类样本的预测分数; $\theta(y_i > y_j)$ 是一个指示函数, 如果 $y_i > y_j$, 即正类样本的得分大于负类样本的得分, 则该函数值为 1, 否则为 0。 A_{uc} 提供了一个更全面的评估, 能够反映模型在不同分类阈值下的性能。它通过衡量模型区分正负类的能力, 避免了对单一阈值的依赖。

为了优化模型性能, 采用了 Optuna 工具^[32]进行超参数调优。Optuna 是一种基于贝叶斯优化算法的自动化超参数优化框架, 它通过定义超参数空间并逐步探索不同的超参数组合, 在每次实验结果的基础上调整搜索策略, 从而高效地提升模型的表现。Optuna 通过建立概率模型来预测哪些超参数可能带来更好的结果, 并基于这些预测引导后续的搜索过程, 这使得搜索策略更加高效, 避免了无效的探索; 但由于优化过程受到初始

参数、搜索策略和计算时间限制的影响, Optuna 找到的解未必是全局最优解。此外, 为了减少因训练集划分不同而可能导致的模型性能波动, 对每个模型进行了 5 折交叉验证 (Cross-validation, 将数据集划分为 5 个子集, 每次使用其中 4 个子集进行训练, 剩下的 1 个子集用于验证。这个过程重复 5 次, 每个子集都作为一次验证集。最终, 模型的评估结果是所有 5 次验证结果的平均值), 以进一步确保模型的泛化能力。最后, 模型的性能通过 A_{cc} 、 R_{ec} 、 P_{re} 、 F_1 值和 A_{uc} 多项指标进行综合评估, 以全面衡量其分类效果和稳定性。

在确定适用于 PCE 预测的最优 ML 算法后, 进一步利用夏普利加性解释 (Shapley additive explanations, SHAP) 方法^[33]对这些模型进行深入解析。SHAP 方法基于合作博弈论中的 Shapley 值理论, 能够量化每个输入特征对模型输出的贡献, 为复杂的 ML 模型提供透明度和可解释性。通过应用 SHAP, 不仅可以识别关键特征及其相对重要性, 还能够深入分析特征与预测目标之间的相互作用和非线性关系。这种方法为优化模型设计和理解数据驱动的科学问题提供了有力支持, 是探索复杂系统规律的重要工具。

3 数据处理与建模

3.1 数据集

在本研究中, 实验数据来源于实验室, 包含 869 条钙钛矿材料的实验记录。钙钛矿光伏器件的电流密度-电压 (J - V) 曲线通过辐射面积为 $100\text{ mm} \times 100\text{ mm}$ 、功率大于 450 W 且与 AM1.5 光谱匹配的太阳能模拟器测量, 以确保数据的可靠性。器件制备采用一套多靶磁控溅射系统, 具备极限真空度为 $4.2 \times 10^{-6}\text{ Pa}$, 支持最高加热温度达 $800\text{ }^{\circ}\text{C}$, 配置有 4 个磁控靶, 具有高精度控制与稳定性, 能够实现多材料的高质量沉积。数据涵盖了大部分可调节的材料合成工艺参数, 建立了一个针对钙钛矿材料的较小型数据库。每条记录由一组 66 个涵盖离散型和连续型参数的特征向量及对应的 PCE 值组成。经分析, 数据集集中的高 PCE 和低 PCE 样本在特征空间中具有广泛覆盖 (表 S1), 充分反映了实验工艺的多样性和可变性, 为模型训练提供了可靠基础。随后, 剔除已知显著影响 PCE 的光电性能参数: 开路电压 (Open-circuit voltage, V_{oc})、短路电流 (Short-circuit current density, J_{sc})、填充因子 (Fill factor, FF)、光照强度等, 然后对异常

数据进行处理。基于实验经验判断, 设定 $V_{oc} > 1.1\text{ V}$ 、 $\text{FF} > 75\%$ 、 $15\text{ mA/cm}^2 < J_{sc} < 23\text{ mA/cm}^2$ 、 $\text{PCE} > 15\%$ 的条件, 筛选出符合要求的数据, 以提高数据的质量和可靠性。同时, 为了确保不同特征在模型训练中具有相同的量纲, 对数据进行了归一化处理。通过这些数据预处理步骤, 确保了数据质量的提升和后续模型训练的有效性。

阈值的选择对模型性能有显著影响。在不平衡数据集的二分类问题中, 模型通常会输出每个样本属于某一类别的概率值, 而不是直接给出类别标签。通过设置不同的 PCE 阈值, 可以调整哪些样本被划分为高 PCE 类别, 从而影响模型的分类边界。若阈值过高, 可能导致大量高 PCE 值的样本被错误地归为低 PCE 类别, 从而降低 R_{ec} ; 若阈值过低, 可能会误将低 PCE 值的样本划分为高 PCE 类别, 导致假阳性增加, P_{re} 下降。因此, 合适的阈值选择可以在 P_{re} 和 R_{ec} 之间找到最佳平衡, 优化模型对不同类别的识别能力。

从图 2 中可见, 在类别极端不平衡的情况下, 模型往往表现出过拟合现象, 即模型对某一类别的特征学习过度, 导致性能评估结果 (如 A_{uc} 值) 虚高。因此, 这些情况在实际应用中并不可取。在本研究中, $\text{PCE} > 18.5$ 被划分为高 PCE 类别 (Class 1), $\text{PCE} \leq 18.5$ 则归为低 PCE 类别 (Class 0), 原因在于该点处低 PCE 和高 PCE 类别样本数量差异较小, 类别间数据分布较为平衡; 同时, A_{uc} 曲线在该处达到下降速率的临界点, 表明该阈值对模型预测性能具有较高的敏感性和代表性。为了更精确量化这一现象, 同时避免在分类过程中因某一类别样本过少而导致数据失衡, 本研究设置了 PCE 阈值筛选范围为 $17.4 \sim 20.1$, 步长为 0.1 。该

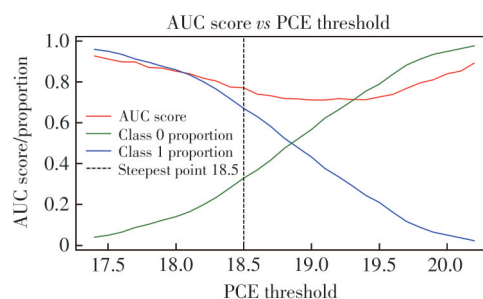


图 2 基于 A_{uc} 斜率最负点的 PCE 二分类阈值筛选与综合评估

Fig.2 Selection and comprehensive evaluation of PCE binary classification threshold based on the steepest A_{uc} slope

范围确保每次划分后两个类别均具有足够样本数量,从而避免了因样本稀缺而无法进行过采样的问题,也确保了数据平衡处理的有效性。针对每一个 PCE 阈值均单独训练模型并计算对应的 A_{uc} 值,进而分析不同 PCE 阈值下 A_{uc} 值的变化速率。通过这一过程,能够定位到 A_{uc} 下降最为剧烈的位置,其前一个阈值所对应的 PCE 值即本研究选择的最优二分类阈值。

图 S1 展示了 PCE 阈值为 18.5 时, SMOTE 平衡前后训练集类别分布的变化。在处理不平衡数据集的二分类问题时,首先通过分层抽样方法将数据集划分为 80% 的训练集和 20% 的测试集,以确保训练集和测试集中的类别分布与原始数据一致。接着,如表 S1 中方法 1 所示,使用 SMOTE-Tomek 方法对训练集进行了平衡处理,使得正负类样本的比例接近 1:1,以确保训练集和测试集中的类别分布与原始数据集一致,从而消除由于数据划分不均匀带来的偏差。

3.2 特征工程

首先,使用 RF 算法进行特征重要性分析,确定哪些特征对提升 PCE 有较大贡献。随后,利用 PCC 分析,剔除了与 PCE 相关性绝对值大于 0.65 (设定的 PCC 阈值) 的特征,优先删除贡献值较小的特征,以简化模型并提高预测效果,避免冗余信息对模型性能产生负面影响。图 3 呈现了各特征之间以及每个特征与 PCE 的 PCC 矩阵,表明在进行 PCC 筛选后,所有特征之间的相关性均低于 0.65,有效避免了多重共线性问题,确保了具有较高独立性和预测能力的特征被保留。基于 PCC

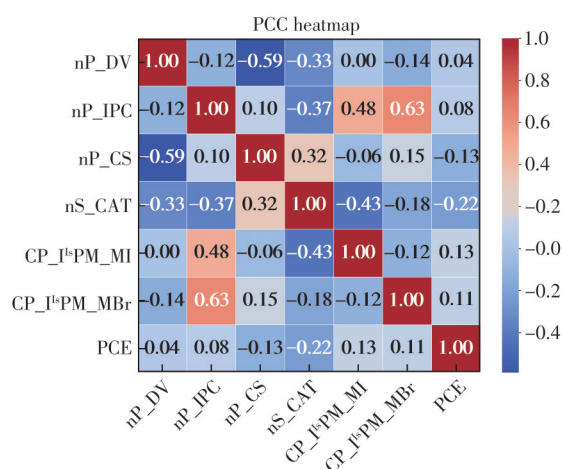


图 3 特征之间及特征与 PCE 的 PCC 热图

Fig.3 The PCC heatmap between features and between feature and PCE

热图筛选出的关键特征及其缩写如表 1 所示(具体含义详见补充文件附录 A)。其中,前缀为“num”的特征对应连续型变量,已通过最大最小归一化处理;前缀为“cat”的特征则为离散型变量,已采用 One-Hot 编码进行处理。图 3 及后文中使用“n”和“c”是对应“num”和“cat”的缩写形式,旨在简化图示标注,二者含义保持一致。此外,还计算并可视化了各特征与 PCE 的相关性,结果显示这些特征与 PCE 之间的相关性较弱,表明它们在统计上并未表现出显著的线性或非线性关系,从而避免了冗余信息的干扰,提高了模型捕捉每个特征独立作用的能力。

表 1 基于 PCC 热图筛选出的关键特征及其缩写

Tab. 1 The key features selected based on the PCC heatmap and its abbreviations

No.	特征	缩写
1	num__PVK_DilutionVolume	nP_DV
2	num__PVK_IntrinsicPassivationConcentration	nP_IPC
3	num__PVK_CoatingSpeed	nP_CS
4	num__SAM_CleanAnnealTime	nS_CAT
5	cat__PVK_IntrinsicPassivationMaterial_MDAI ₂	cP_I ^h PM_MI
6	cat__PVK_IntrinsicPassivationMaterial_MDABr ₂	cP_I ^h PM_MBr
7	cat__PVK_InterfacePassivationMaterial_MDAI ₂	cP_I ^h PM_MI
8	num__PVK_SubmergedPassivationAnnealTime2	nP_SPAT2

图 S2 展示了通过 RFE 方法,依据均方误差 (Mean squared error, MSE) 筛选出的最佳特征数量。通过这一过程,确定了对模型预测性能贡献最佳的特征数量,从而优化了模型性能,提高了 PCE 预测的准确度。最后,通过 RF 算法计算特征重要性,图 S3 展示了按特征贡献度排序后 8 个对 PCE 具有显著影响的关键制备条件,结合 RFE 的结果,进一步剔除了贡献度较小的两个特征 cP_I^hPM_MI 和 nP_SPAT2。通过这一系列步骤,成功减少了冗余特征,提升了模型的准确性,并为优化 PSCs 的制备工艺提供了有力支持。

3.3 预测模型

本研究分别比较了 LR、RF、SVM、KNN、GB、XGBoost 六种 ML 模型,使用的实验数据已进行归一化、分层抽样、数据平衡和五折交叉验证处理。评价指标如图 4 所示。

本文选择模型的评估标准按照 $A_{uc} > F_1 > P_{re} > R_{cc} > A_{cc}$ 的顺序,是因为 A_{uc} 能够全面衡量模型在不同阈值下区分正负类样本的能力,这对于识别高 PCE 样本至关重要。 F_1 值紧随其后,因为它

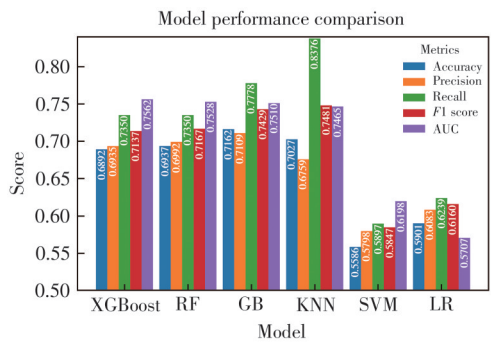


图4 六种ML模型的5个性能评价指标

Fig.4 Visual comparison of the six ML models' performance across five evaluation metrics

结合了 P_{re} 和 R_{ec} ,能够平衡这两个指标,确保既不漏掉高 PCE 样本,又能减少假正例的出现。 P_{re} 和 R_{ec} 在评估模型的表现时也很重要, P_{re} 保证了模型预测为高 PCE 的样本确实具备高 PCE; 而 A_{cc} 在类别不平衡的任务中容易产生误导,因此作为最后的评估指标。在本研究中,为选择最合适的模型,综合考虑多个评估指标,并为每个指标分配不同的权重。具体而言, A_{uc} 被赋予最高权重(0.3),因为它能够全面评估模型在不同阈值下区分正负类样本的能力,对于识别高 PCE 样本尤其重要。 F_1 值作为第二重要的评估标准(权重 0.25),它兼顾了 P_{re} 与 R_{ec} ,能够在减少假正例的同时,确保高 PCE 样本的准确识别。 P_{re} 的权重为 0.2,突出其在确保预测为高 PCE 样本的可靠性方面的作用。 R_{ec} 的权重为 0.15,确保不会漏掉真正的高 PCE 样本。而 A_{cc} 的权重为 0.1,主要考虑到在类别不平衡的情况下 A_{cc} 可能带来的误导性。如图 S4 所示,通过这一加权方法,能够全面评估模型的性能,并得出选择最优的模型为 GB(总得分 0.7415,高于其他模型)。为进一步探讨数据平衡对模型性能的影响,基于原始训练集探究 SMOTE-Tomek 采样对模型性能的影响。结合 GB 算法,分析了在不同数据平衡条件下模型的表现。

如表 2 所示,方法 1 采用 SMOTE-Tomek 对训练集进行平衡采样,测试集保持原始不平衡分布;方法 2 则在训练集和测试集均保留原始不平衡状态。对比结果显示,在测试集上,方法 1 在 A_{cc} 和 A_{uc} 方面优于方法 2,说明其整体判断能力和泛化性能更强。尽管方法 2 在 R_{ec} 和 F_1 分数上略高,但也伴随着精度下降,提示其可能偏向于过度识别多数类,从而增加假阳性风险。在训练集上,方法 2 的各项指标均高于方法 1,尤其是 F_1 分数和 R_{ec} ,

这种显著的性能差距提示其可能存在过拟合风险。相比之下,方法 1 在训练集与测试集之间的表现更加一致,模型表现更加稳健,泛化能力更可靠。

表 2 基于 GB 的不同数据平衡条件下的训练集和测试集性能指标

Tab.2 Trainsets and testsets performance metrics of GB under different data balancing conditions

数据集	方法	A_{cc}	P_{re}	R_{ec}	F_1	A_{uc}
训练集	1	0.710 4	0.710 8	0.759 7	0.734 4	0.789 9
测试集	1	0.716 2	0.710 9	0.777 8	0.742 9	0.751 0
训练集	2	0.746 8	0.747 4	0.939 9	0.832 7	0.772 9
测试集	2	0.695 4	0.707 8	0.931 6	0.804 4	0.701 5

为验证 SMOTE-Tomek 是否对特征分布造成影响,我们对关键特征进行了核密度可视化分析(图 S5)。结果显示,采样前后主要特征的分布基本一致,未观察到明显偏移或异常峰值,说明该方法未引入显著的分布扭曲。综合考虑 A_{cc} 、 A_{uc} 及训练-测试指标一致性,方法 1 在平衡性能与泛化能力方面表现更优,适合实际材料筛选中对模型稳定性的需求。

与以往的研究一致^[18,34-36],本研究确认平衡训练集可有效缓解类别不均带来的偏差,显著提升模型的泛化能力。虽然过采样存在过拟合风险,但通过合理采样和模型调优,本研究成功控制了该风险。相较之下,保持不平衡训练集虽提升了 R_{ec} 和 F_1 分数,但伴随精度下降,可能增加假阳性率,影响模型可靠性。部分算法通过类别权重调整或损失函数优化,进一步增强对不平衡数据的适应性。鉴于高低 PCE 类别差异不大,适度的数据平衡有助于减少类别偏斜引发的识别误差,提升模型稳定性。综上,采用平衡训练集的方法在本研究中表现更优,验证了数据平衡策略在提升模型性能和泛化能力方面的必要性与有效性。

4 分 析

4.1 特征重要性分析

由于 GB 模型综合的优秀性能,在对 GB 模型进行 Optuna 超参数优化后,基于 GB 算法训练模型并使用特征重要性评估来衡量单个特征对高 PCE 样本识别和预测准确性的影响。通过特征重要性评估,能够量化每个特征在模型中所起的作用,识别出哪些特征对区分高 PCE 与低 PCE 样本至关重要。这不仅有助于优化模型性能,还为深

入理解影响 PCE 的关键因素提供了依据。通过分析特征的重要性, 可以聚焦于那些对 PCE 具有显著影响的特征, 从而在材料合成过程中做出更加精准的决策。

从图 5 中可以观察到 GB 模型对特征的重要性分析。发现 nP_DV 是最重要的特征, 其重要性贡献为 39.16%, 表明本研究中的 nP_DV 对预测高 PCE 样本具有决定性影响。因为 nP_DV 直接影响钙钛矿前驱体溶液的浓度, 当 nP_DV 较大时, 钙钛矿溶液的浓度降低, 可能影响其结晶质量、薄膜的均匀性以及 PCE; 当 nP_DV 较小(即溶液浓度过高)时, 则可能导致钙钛矿薄膜的缺陷增加, 影响电子的传导性和光吸收能力, 从而降低 PCE。合理的 nP_DV 有助于优化钙钛矿溶液的稳定性及成膜质量, 进一步提高 PCE^[37]。nP_CS, 重要性贡献为 24.50%, 同样对 PCE 有着举足轻重的影响。较高的 nP_CS 通常能够形成较薄且均匀的钙钛矿薄膜, 有助于提高光吸收能力和载流子迁移效率, 进而提升 PCE。然而, nP_CS 过高同样也可能导致膜层过薄, 影响光吸收或引发不稳定的结晶结构, 降低电荷传输效率^[38]。因此, 合适的 nP_CS 不仅能够优化薄膜的均匀性和厚度, 还能促进钙钛矿材料的良好结晶, 从而提高 PCE。找到最优 nP_CS 亦是本研究提升 PSCs PCE 的关键之一。nP_IPC 的贡献为 17.45%, 亦是一个关键因素。这一特征的影响也源于浓度在钙钛矿材料合成中的关键作用。钙钛矿材料中的添加剂, 尤其是钝化剂和界面调节剂, 能够有效改善材料的表面性质和结晶质量^[39], 从而提高 PCE。适当的钝化浓度有助于减少钙钛矿材料表面缺陷, 这些缺陷常常成为载流子的复合中心, 导致 PCE 下

降。添加剂通过钝化表面缺陷, 降低复合率, 进而提高载流子迁移率和材料稳定性。然而, 钝化剂浓度过低可能导致表面缺陷未得到有效钝化, 从而影响 PCE; 而浓度过高则可能导致添加剂在材料中的不均匀分布, 形成新的缺陷, 反而降低材料的 PCE。合理地优化添加剂浓度不仅能提高光电导率和 PCE, 还能增强材料的稳定性, 是提升钙钛矿材料 PCE 的关键。适当的 nS_CAT (14.45%) 对于去除基底表面的有机残留物至关重要。通过控制 nS_CAT, 可以显著提高薄膜的清洁度和稳定性^[37]。适当的 nS_CAT 不仅有助于去除表面污染物, 还能促进表面原子或分子的重新排列, 从而增强 SAM 层的有序性, 确保在自组装过程中基底表面具有所需的清洁度和表面质量。良好的表面处理能够减少表面缺陷和不均匀性, 从而优化自分子组装的过程, 提高最终器件的 PCE 和稳定性。此外, cP_I^hPM_MI (2.42%) 和 cP_I^hPM_MBr (2.02%) 分别表示在钙钛矿溶液中加入 MDAl₂、MDABr₂ (某种特定的有机分子) 作为添加剂。模型预测该添加剂对改善钙钛矿材料的 PCE 有一定的影响。这些分子添加剂可能通过调节钙钛矿材料的结晶性、能带结构或界面特性, 进而增强材料的光吸收能力、载流子迁移率以及器件的稳定性, 从而提升最终器件的 PCE。这些特征对于本研究的电池制备起着重要作用, 通过优化这些特征的参数, 应可以进一步提升材料的 PCE 和稳定性。

4.2 模型可解释性分析

模型的可解释性是 ML 的重要研究方向, 在科学研究和实际应用中尤为关键, 因为它能够帮助理解模型的决策依据, 揭示特征之间的相互关系, 并识别优先优化的关键因素。这种透明性不仅提高了模型的信任度, 还能为实验设计和工艺优化提供有力支持。通过从复杂的高维数据中提取出关键特征, 可以探索这些特征与 PCE 之间的潜在关系, 从而更高效地提升材料的 PCE 和工艺稳定性。

在 SHAP 图中, 特征根据其对模型预测贡献的重要性排序, 排名靠前的特征对预测结果的影响最大。每个特征的 SHAP 值表示该特征对单个预测结果的具体贡献, 正值表明该特征推动预测结果增大, 负值则表示其推动预测结果减小。SHAP 图中的颜色深浅反映了特征值的大小, 红色通常代表特征值较大, 蓝色则表示特征值较小。

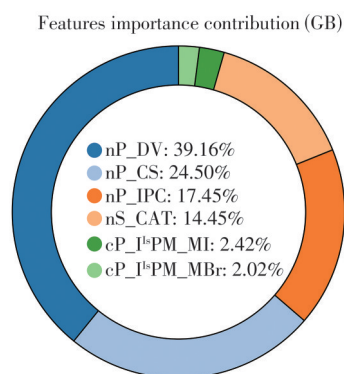


图 5 基于 GB 模型的特征重要性可视化

Fig.5 Visualization of feature importance based on the GB model

此外,点的宽度表示该特征在数据集中的分布范围,点越宽意味着该特征的取值范围更广,波动也越大,从而对预测结果产生更显著的影响。

图 6 的 SHAP 值分析揭示了不同特征对模型输出(PCE)的影响程度和方向。特征 nP_CS 是影响 PCE 的首要因素。SHAP 值分析显示,当旋涂速度较高时(对应特征值较大,红色数据点),其 SHAP 值主要分布在负区间,表明目前实验中较高的 nP_CS 可能不利于提高 PCE,后续可考虑使用中低的 nP_CS 进行实验。特征 nP_DV 是影响 PCE 的次要因素,其在左侧数据点与 nP_CS 的分布较为相似,代表高 nP_DV 同样不利于提高 PCE;不同的是 nP_DV 在 SHAP 图的正区间是紫色数据点居多,说明过高或过低的 nP_DV 可能都不利于提升 PCE,目前实验中 nP_DV 的值可能已经较为合适。特征 nS_CAT 的 SHAP 值分析显示, nS_CAT 越短越好。特征 nP_IPC 表明,低 nP_IPC 对 PCE 无提升作用,而高 nP_IPC 同样未必能够促进 PCE 提高。因为红色数据点在 SHAP 图左右两侧均存在,据此猜测高 nP_IPC 对于提升 PCE 存在不稳定的问题。因此,可以继续调控 nP_IPC,寻找中低 nP_IPC 以提升 PCE。特征 cP_I^hPM_MI 和 cP_I^hPM_MBr 的 SHAP 值分析显示,MDAI₂ 或者 MDABr₂ 的作用还需进一步探究,结合 nP_IPC 来看,很有可能是因为浓度的调配不当而导致的效果不佳。因此,需要进一步优化 MDAI₂ 及 MDABr₂ 的浓度和使用条件。通过类似的方式能够分析所有特征中可能存在的问题并探索有效提升 PCE 的途径,具体需要根据实际实验情况对具体问题进行分析。利用 SHAP 分析设计实验方案,进而确定可能提升 PCE 的实验条件。这种方法

有助于我们更精确地理解各个特征对模型输出的影响,从而为实验优化提供指导。

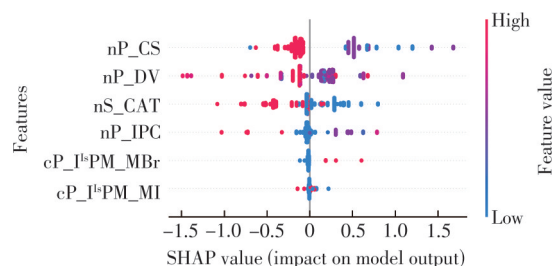


图 6 GB 模型的 SHAP 输出,显示了每个特征对模型预测的贡献

Fig.6 SHAP output of the GB model, showing the contribution of each feature to the model's prediction

4.3 优化制备工艺

图 S6 中清晰地体现了对 SHAP 图及对应具体数据的探索性分析。所有筛选出的 192 条实验条件均涉及 MDAI₂ 作为添加剂。进行数据探索性分析的原因在于帮助了解先前实验条件的分布,从而根据 SHAP 分析的结果做出更合理的决策。结合数据探索性分析结果,决定暂时不调控 nS_CAT 和 cP_I^hPM_MBr。具体实验设计可参考表 S2,通过系统地优化关键特征 nP_CS、nP_IPC 的值并控制其他特征,来寻找提升 PCE 的最佳实验条件。由于该设计基于模型预测结果,涉及多个参数的联合优化,因此难以设置仅改变单一变量的严格控制组。为体现模型优化带来的整体提升效果,我们在图 7(a)中加入了代表常规实验条件的对照组数据,作为与优化方案的性能对比。通过这些优化条件的实施,条件 2 成功制备出如图 7(a)所示的冠军 PCE 电池,突破了未曾达到的水平。我们测试了冠军 PCE 电池的外量子效率(External quantum efficiency, EQE),如图 7(b)所示。

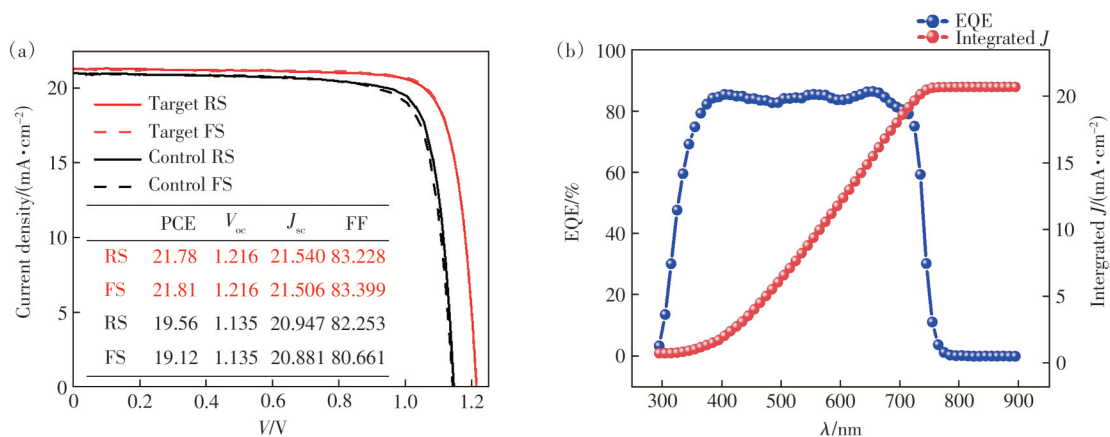


图 7 (a)冠军器件的电流-电压(J-V)曲线;(b)EQE 光谱及对应的电流密度响应

Fig.7 (a)Current-voltage (J-V) curve of the champion device. (b)EQE spectrum and corresponding current density response

尽管已有部分实验成功验证了模型推荐条件的有效性,但当前实验验证的覆盖范围仍有限,仅针对极少数预测结果进行了验证,尚无法全面评估模型的泛化能力和稳定性。为此,我们将在后续工作中进一步扩大实验样本数量,并对多个预测优良但尚未验证的条件进行系统测试,以构建更完整的“预测-验证”闭环体系,增强模型在实际器件开发中的实用性与可迁移性。

5 结 论

本研究通过收集实验室数据并剔除异常数据,随后选择表现最优的 GB 模型进行 PCE 预测,并使用 Optuna 超参数调优技术提高模型的预测精度。此外,结合 SHAP 可解释 ML 模型分析方法,对 nP_CS、nP_IPC 等关键特征进行深入分析,揭示其对 PCE 的影响机制。最终,对模型输出指导进行实验验证。实验结果表明, WBG PSCs 的 PCE 获得显著提升,充分验证了 ML 技术在工艺优化及 PCE 预测中的有效性,为高 PCE 的 PSCs 开发提供了全新思路。ML 不仅能够协助研究者设计并探索合

成工艺条件,还能在较短时间内实现高效优化,通过结合 SHAP 可解释性 ML 方法,可以揭示出传统实验方法难以发现的新规律,为 PSCs 的发展提供了新的理论依据与科学见解。然而,本研究仍存在一些局限性,如数据集适用性的限制以及特征选择方法的局限性,这可能在一定程度上影响本研究提出的方法框架的推广。未来研究可以围绕这两方面进一步优化框架:一方面,研究者们可尝试在不同实验数据集上验证框架的适用性,以增强其通用性;另一方面,亦可开发更高效的 ML 算法,以提升预测精度并增强实际应用的可操作性,实现跨领域的数据融合。本研究不仅为钙钛矿材料的优化设计提供了科学指导,也为实际应用中开发高 PCE 的 PSCs 奠定了坚实基础。随着 ML 技术的不断进步,期待其在材料科学领域催生更多创新与突破,为可再生能源的可持续发展贡献力量。

本文补充文件和专家审稿意见及作者回复内容的下载地址: <http://cjl.lightpublishing.cn/thesisDetails#10.37188/CJL.20250149>

参 考 文 献:

- [1] KOJIMA A, TESHIMA K, SHIRAI Y, *et al.* Organometal halide perovskites as visible-light sensitizers for photovoltaic cells [J]. *J. Am. Chem. Soc.*, 2009, 131(17): 6050-6051.
- [2] BATI A S R, ZHONG Y L, BURN P L, *et al.* Next-generation applications for integrated perovskite solar cells [J]. *Commun. Mater.*, 2023, 4(1): 2.
- [3] YANG C, HU W J, LIU J L, *et al.* Achievements, challenges, and future prospects for industrialization of perovskite solar cells [J]. *Light: Sci. Appl.*, 2024, 13(1): 227.
- [4] JIANG X, QIN S C, MENG L, *et al.* Isomeric diammonium passivation for perovskite-organic tandem solar cells [J]. *Nature*, 2024, 635(8040): 860-866.
- [5] WANG Y R, LIN R X, LIU C S Y, *et al.* Homogenized contact in all-perovskite tandems using tailored 2D perovskite [J]. *Nature*, 2024, 635(8040): 867-873.
- [6] CHEN Z L, DONG Q F, LIU Y, *et al.* Thin single crystal perovskite solar cells to harvest below-bandgap light absorption [J]. *Nat. Commun.*, 2017, 8(1): 1890.
- [7] WANG R, LIU X Y, YAN S, *et al.* Efficient wide-bandgap perovskite photovoltaics with homogeneous halogen-phase distribution [J]. *Nat. Commun.*, 2024, 15(1): 8899.
- [8] LI S S, JIANG Y Z, XU J, *et al.* High-efficiency and thermally stable FACsPbI₃ perovskite photovoltaics [J]. *Nature*, 2024, 635(8037): 82-88.
- [9] ZHANG L X, WANG Y S, MENG X C, *et al.* The issues on the commercialization of perovskite solar cells [J]. *Mater. Futures*, 2024, 3(2): 022101.
- [10] LIM J, PARK N G, SEOK SIL, *et al.* All-perovskite tandem solar cells; from fundamentals to technological progress [J]. *Energy Environ. Sci.*, 2024, 17(13): 4390-4425.
- [11] LIN R X, WANG Y R, LU Q W, *et al.* All-perovskite tandem solar cells with 3D/3D bilayer perovskite heterojunction [J]. *Nature*, 2023, 620(7976): 994-1000.
- [12] ZHANG D Y, LI D Y, HU Y, *et al.* Degradation pathways in perovskite solar cells and how to meet international standards [J].

- Commun. Mater.*, 2022, 3(1): 58.
- [13] CORREA-BAENA J P, HIPALGAONKAR K, VAN DUREN J, *et al.* Accelerating materials development *via* automation, machine learning, and high-performance computing [J]. *Joule*, 2018, 2(8): 1410-1420.
- [14] CHONG S S, NG Y S, WANG H Q, *et al.* Advances of machine learning in materials science: ideas and techniques [J]. *Front. Phys.*, 2024, 19(1): 13501.
- [15] MOBARAK M H, MIMONA M A, ISLAM M A, *et al.* Scope of machine learning in materials research—A review [J]. *Appl. Surf. Sci. Adv.*, 2023, 18: 100523.
- [16] DE A F. The impact of machine learning in energy materials research: The case of halide perovskites [J]. *ACS Energy Lett.*, 2023, 8(2): 1270-1272.
- [17] CHEN C, MAQSOOD A, JACOBSSON T J. The role of machine learning in perovskite solar cell research [J]. *J. Alloys Compd.*, 2023, 960: 170824.
- [18] HU J H, CHEN Z X, CHEN Y Z, *et al.* Interpretable machine learning predictions for efficient perovskite solar cell development [J]. *Sol. Energy Mater. Sol. Cells*, 2024, 271: 112826.
- [19] 弓箭, 陈谦, 李阳, 等. 机器学习辅助钙钛矿薄膜制备工艺优化及特征重要性评估 [J]. *发光学报*, 2024, 45(3): 399-406.
- GONG J, CHEN Q, LI Y, *et al.* Machine learning assisted optimization of perovskite thin film fabrication process and assessment of feature importance [J]. *Chin. J. Lumin.*, 2024, 45(3): 399-406. (in Chinese)
- [20] LI J X, PRADHAN B, GAUR S, *et al.* Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells [J]. *Adv. Energy Mater.*, 2019, 9(46): 1901891.
- [21] SHOCKLEY W, QUEISSER H J. Detailed balance limit of efficiency of *p-n* junction solar cells [J]. *J. Appl. Phys.*, 1961, 32(3): 510-519.
- [22] STODDARD R J, DUNLAP-SHOHL W A, QIAO H B, *et al.* Forecasting the decay of hybrid perovskite performance using optical transmittance or reflected dark-field imaging [J]. *ACS Energy Lett.*, 2020, 5(3): 946-954.
- [23] LIU Z, ROLSTON N, FLICK A C, *et al.* Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing [J]. *Joule*, 2022, 6(4): 834-849.
- [24] 唐乾元. 物理学与人工智能的连接: 2024年诺贝尔物理学奖解析 [J]. *科学通报*, 2025, 70(10): 1413-1420.
- TANG Q Y. The connections between physics and AI: a review of the 2024 Nobel Prize in Physics [J]. *Chin. Sci. Bull.*, 2025, 70(10): 1413-1420. (in Chinese)
- [25] 梁杰, 陈嘉豪, 张雪芹, 等. 基于独热编码和卷积神经网络的异常检测 [J]. *清华大学学报(自然科学版)*, 2019, 59(7): 523-529.
- LIANG J, CHEN J H, ZHANG X Q, *et al.* One-hot encoding and convolutional neural network based anomaly detection [J]. *J. Tsinghua Univ. (Sci. Technol.)*, 2019, 59(7): 523-529. (in Chinese)
- [26] SINGH D, SINGH B. Investigating the impact of data normalization on classification performance [J]. *Appl. Soft Comput.*, 2020, 97: 105524.
- [27] CHAWLA N V, BOWYER K W, HALL L O, *et al.* SMOTE: synthetic minority over-sampling technique [J]. *J. Artif. Intell. Res.*, 2002, 16: 321-357.
- [28] MENZE B H, KELM B M, MASUCH R, *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data [J]. *BMC Bioinformatics*, 2009, 10: 213.
- [29] SCHOBER P, BOER C, SCHWARTE L A. Correlation coefficients: appropriate use and interpretation [J]. *Anesth. Analg.*, 2018, 126(5): 1763-1768.
- [30] 吴辰文, 梁靖涵, 王伟, 等. 基于递归特征消除方法的随机森林算法 [J]. *统计与决策*, 2017, 33(21): 60-63.
- WU C W, LIANG J H, WANG W, *et al.* Random forest algorithm based on recursive feature elimination [J]. *Stat. Decis.*, 2017, 33(21): 60-63. (in Chinese)
- [31] VUJOVIĆ Ž Đ. Classification model evaluation metrics [J]. *Int. J. Adv. Comput. Sci. Appl.*, 2021, 12(6): 599-606.
- [32] AKIBA T, SANO S, YANASE T, *et al.* Optuna: a next-generation hyperparameter optimization framework [C]. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, USA*, 2019: 2623-2631.
- [33] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions [C]. *Proceedings of the 31st International*

- Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 4768-4777.*
- [34] KABIR M A, AHMED M U, BEGUM S, *et al.* Balancing fairness: unveiling the potential of SMOTE-driven oversampling in AI model enhancement [C]. *Proceedings of the 2024 9th International Conference on Machine Learning Technologies, Oslo, Norway, 2024: 21-29.*
- [35] GE J H, HUANG Y R, CHEN X X, *et al.* An intermediate-aided perovskite phase purification for high-performance solar cells [J]. *J. Am. Chem. Soc.*, 2025, 147(2): 1980-1990.
- [36] WU J C, TORRESI L, HU M M, *et al.* Inverse design workflow discovers hole-transport materials tailored for perovskite solar cells [J]. *Science*, 2024, 386(6727): 1256-1264.
- [37] LIU F W, XU J C, MA Y C, *et al.* Functional design and understanding of effective additives for achieving high-quality perovskite films and passivating surface defects [J]. *J. Energy Chem.*, 2025, 102: 597-608.
- [38] WEN J, ZHAO Y C, LIU Z, *et al.* Steric engineering enables efficient and photostable wide-bandgap perovskites for all-perovskite tandem solar cells [J]. *Adv. Mater.*, 2022, 34(26): 2110356.
- [39] ZHANG B Y, LIU D Y, CHEN P, *et al.* Improved perovskite crystallization *via* antisolvent-assisted processed using additive engineering for efficient perovskite solar cells [J]. *J. Alloys Compd.*, 2021, 855: 157396.



梁锐权(2001-),男,广东广州人,硕士研究生,2023年于广东石油化工学院获得学士学位,主要从事人工智能与新型钙钛矿太阳能电池的相关研究。

E-mail: 2681160546@qq.com



王亦文(1993-),女,山西运城人,博士,副研究员,2019年于中国香港浸会大学获得博士学位,主要从事新型有机光伏器件、钙钛矿太阳能电池及其稳定性的研究。

E-mail: wangyiwen@jnu.edu.cn



李阳(1969-),男,辽宁鞍山人,博士,教授,博士生导师,2000年于南开大学获得博士学位,主要从事光电材料与器件、运用人工智能优化光电材料与器件性能等方面的研究。

E-mail: insidesun51@163.com