

预处理方法和机器学习算法对玻璃材料性能预测的影响： 比较研究和优化策略

王思雨^{1,2}, 孙元成¹, 余 睿², 杜秀蓉¹, 周利生¹, 刘 星¹, 陈志威^{1,3}

(1. 中国建筑材料科学研究总院有限公司, 北京 100024; 2. 武汉理工大学材料与微电子学院, 武汉 430070;
3. 武汉理工大学材料科学与工程学院, 武汉 430070)

摘 要: 系统研究了不同预处理方法和机器学习算法对玻璃材料性能预测的影响。实验基于一组包含 4 种组分作为输入、7 种性能作为输出的数据集, 首先通过标准化与 3 种插值方法因不同顺序生成 6 组样本, 随后分别采用随机森林(RF)、极限梯度提升(XGB)、分类回归树(CART)、K 近邻(KNN)、支持向量机(SVM)和人工神经网络(ANN)对每组样本的每一列输出进行单目标实验, 并对所有输出进行综合实验。在此基础上, 选择 RF 与 SVM 进行超参数调优, 比较了随机搜索、网格搜索和贝叶斯优化的效果。结果表明, 预处理顺序对误差指标(MSE/MAE)的敏感性高于决定系数(R^2), 建议优先采用“B-RF 插值”组合以平衡精度与稳定性; 而数据缺失率 $>10\%$ 时, 插值方法对 R^2 的影响权重显著增加, 高缺失场景($>40\%$)需特别注意插值方法对 SHAP 特征排序的干扰; 算法性能呈 $RF=XGB>KNN=ANN>SVM\geq CART$ 的稳定层级, 建议首选集成算法构建预测模型; 组分分析显示 Na_2O 主导密度与折射率, B_2O_3 主导玻璃化转变温度 T_g , SiO_2 调控膨胀软化温度 M_g , 而热膨胀系数存在阶段性; 贝叶斯优化可有效提升 SVM, 而集成算法建议采用遗传算法强化参数搜索。

关键词: 机器学习算法比较; 性能预测; 超参数调优; 数据预处理

中图分类号: TQ171.1; TP181 文献标志码: A 文章编号: 0454-5648(2025)10-2870-12

网络出版时间: 2025-09-12



机器学习(ML)是一门通过统计学习从数据中自动发现规律并优化模型性能的交叉学科, 通常可实现预测或决策的目的^[1]。自 Arthur Samuel 于 1959 年提出第 1 个被应用的人工智能学习框架以来^[2], ML 已经逐渐成为各领域中取代需要传统高成本实验和计算密集型模拟的关键工具^[3-5]。除此之外, 从人类无法掌握的复杂数据模式中得出有意义的推理也是 ML 的重要功能, 这极大地改变了材料设计和制造领域^[6]。

以玻璃材料为例, 玻璃是一种复杂的无序物质, 因此, 难以通过准确预测玻璃的状态相关性(如液相线温度、溶解动力学、断裂韧性等)进行建模^[7]。基于预测目标的数据结构特征, 机器学习技术在玻璃研究中的应用主要涵盖 3 大类型: 回归分析、分类与聚类、概率估计, 分别对应经典算法包括: 支持向量机(SVM)、人工神经网络(ANN)、K 近邻(KNN)等回归模

型; 卷积神经网络(CNN)、K 均值(K-means)、决策树等分类器; 以及期望最大化(EM)、朴素贝叶斯等概率方法^[8]。其中, 回归分析因其连续量化的特性^[9], 常被用于探究输入-输出参数间的相互作用机制。

近年来, 学者们通过构建多样化机器学习模型, 在玻璃性能预测领域取得了显著进展。Tian 等^[10]对比了 LASSO 线性回归、支持向量机(SVR)、随机森林(RF)在预测高维氧化物玻璃密度、维氏硬度的预测效能, 并成功进行了系统的实验验证。Liao 等^[11]开发了基于 BP 神经网络模型的工艺参数优化框架, 实现了对石英玻璃表面锥形微通道形状的精准调控。Mastelini 等^[12]系统评估分类回归树(CART)、K 近邻(KNN)、多层感知器(MLP)、随机森林(RF)对硫化物玻璃在玻璃化转变温度、热膨胀系数、折射率的建模能力, 发现 RF 的 R^2 可达 0.92, 并揭示了

收稿日期: 2025-04-14。 修订日期: 2025-05-17。

基金项目: CNBM 关键技术研发计划(2023SJYL03); 中国建材总院自立科研项目(ZT-20)。

第一作者: 王思雨(2000—), 女, 硕士研究生。

通信作者: 孙元成(1984—), 男, 博士, 正高级工程师。

Received date: 2025-04-14. Revised date: 2025-05-17.

First author: WANG Siyu (2000—), female, Master candidate.

E-mail: wsiyu@whut.edu.cn

Correspondent author: SUN Yuancheng (1984—), male, Ph.D., Professor.

E-mail: sunych@foxmail.com

Ge/Si/Ga 元素对 T_g 的正向调控机制。Shi 等^[13]使用人工神经网络(ANN)、支持向量机(SVM)、随机森林(RF)对杨氏模量、剪切模量、温度依赖性玻璃成分之间的关系,证实模型选择需与目标特性的物理本质相匹配。值得注意的是,Liu 等^[14]探究了不同特征提取方法对于 RF、GBT、SVR、KNN 预测氧化物玻璃的折射率和阿贝数性能,证明特征提取方法在非集成模型(SVR、KNN)中的重要作用。

然而,现有研究在建模的关键环节仍存在局限性:首先是对数据预处理的系统性不足:多数文献仅采用基础数据清洗(如删除极值与重复条目),缺乏对缺失值填补方法与特征标准化顺序的探讨。其次是超参数优化策略单一:尽管有学者在模型中引入了随机搜索^[9]或贝叶斯优化^[15]等先进方法,但普遍忽视算法特性与调优方法的适配性研究,更缺乏跨模型超参数优化方式的对比分析。这种局限性可

能导致以 SVM 为代表的超参数敏感模型的潜力未被充分挖掘。

针对上述挑战,本工作建立了一套如图 1 所示的面向玻璃性能预测的全流程优化框架:1) 在数据预处理阶段,对数据集 A 系统评估 3 种缺失值填补策略(中位数插值、多项式插值、随机森林插值)与两种标准化顺序(先插值后标准化;先标准化后插值)的组合效应;2) 在模型构建阶段,采用 6 类具有不同归纳偏置的算法(RF、XGB、CART、KNN、SVM、ANN),结合三重评估指标(MSE、MAE、 R^2)量化其性能预测效能,并采用新数据集(A)对模型结果过进行验证;3) 采用 SHAP 可解释性框架解析关键组分的作用路径;4) 在超参数优化层面,对 SVM 和 RF 进行随机搜索、网格搜索与贝叶斯优化对比实验,建立超参数调优策略与模型结构的适配准则。

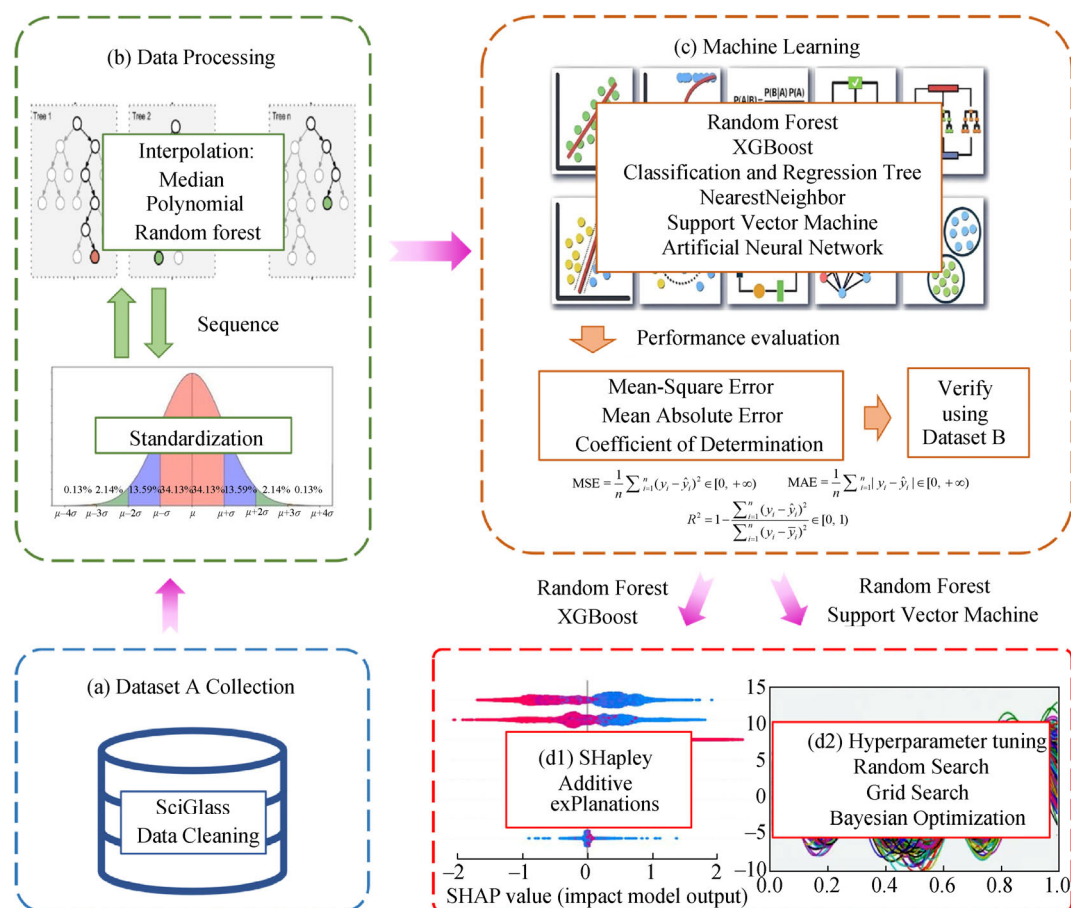


图 1 玻璃性能预测的全流程优化框架

Fig. 1 A whole-process optimization for glass performance prediction

1 材料及方法

1.1 数据准备与预处理

初始数据均为从 SciGlass 数据库

(<https://github.com/epam/SciGlass>)中提取,这是一个专门收集文献专利中报道的玻璃性质和化学组成的数据库。挑选数据时选择共同性能多的氧化物玻璃。

得到数据后对其进行数据清洗与预处理 2 步。

数据清洗包含删除极值与重复条目,对热膨胀系数取对数。前者是为了避免相同组分被划分为不同的集合而引起数据泄露问题^[10]。后者是源于未经处理的热膨胀系数将严重影响模型的准确性^[16]。为平衡材料科学原理与工程实践需求,本工作直接建立组分与性能之间的映射关系以保证工艺可控性。数据包含 4 种组分输入(SiO_2 、 B_2O_3 、 Na_2O 和 K_2O , 输入数值为该组分的摩尔质量占比), 7 种性能输出(密度 Den 、20~300 °C 的热膨胀系数 T_3 、20~400 °C 的热膨胀系数 T_4 、折射率 R_i 、玻璃化转变温度 T_g 、膨胀软化温度 M_g 和硬度 H)。将清洗结束后的数据进行统计得到描述性统计量结果如表 1。

表 1 各特征数据集的描述性统计信息

Table 1 Descriptive statistics of the used datasets for each property

Group	Count	Missing	Mean	Std	Min	50%	Max
		measure rate/%					
SiO_2	299	0	56.568	21.360	0.000	62.200	85.000
B_2O_3	299	0	24.265	19.182	0.000	20.000	90.000
Na_2O	299	0	12.550	10.565	0.000	13.200	40.000
K_2O	299	0	6.618	9.379	0.000	0.000	35.000
Den	194	35.1	2.449	0.088	2.146	2.477	2.550
T_3	273	8.7	1.931	0.115	1.505	1.940	2.210
T_4	273	8.7	1.955	0.135	1.533	1.951	2.352
R_i	205	31.4	1.512	0.013	1.470	1.516	1.530
T_g	271	9.4	537.458	66.078	347.000	547.000	668.000
M_g	271	9.4	577.845	66.630	373.000	589.000	725.000
H	179	40.1	472.492	83.902	260.000	469.000	680.000

而后进行缺失值处理与标准化,为比较不同缺失值处理的效果以及标准化在预处理过程中的顺序是否对模型结果有影响,本工作选用 3 种插值方法(中位数插值、多项式插值、随机森林插值),同时区分了插值与标准化的先后顺序共获得了 6 组样本: B-Z(先标准化后中位数插值)、B-E(先标准化后多项式插值)、B-RF(先标准化后随机森林插值); Z-B(先中位数插值后标准化)、E-B(先多项式插值后标准化)、RF-B(先随机森林插值后标准化)。

1.2 模型训练

为尽可能对比不同的机器学习模型在玻璃性能中的应用优势点与预测准确性,本工作选择了树模型:分类回归树(CART);实例学习: K 近邻(KNN);核方法:支持向量机(SVM);集成算法:极限梯度提升(XGB)与随机森林(RF);深度学习:人工神经网络(ANN)。采用每种算法分别针对 7 项离散性能指标及综合性能评价体系进行实验,共获得 8 组结

果。在训练过程中将数据集划分为训练集、测试集,比例为 0.8:0.2。表 2 则展示了 6 种模型在默认情况下的参数选择。

表 2 各模型的参数细节
Table 2 Parameter details of each model

Model	Hyper-parameter	Value	Structure parameter	Value
CART	Max-depth	None	Splitter	best
	Min-samples-leaf	1	criterion	Squared-error
KNN	n-neighbors	5	metric	minkowski
	weights	Uniform		
SVM	C	1	Kernel	rbf
	gamma	Scale		
XGB	Learning-rate	0	Objective	9.379
	Max-depth		booster	
	subsample			
RF	n-estimators	100	Criterion	Squared-error
	Max-depth	None	Bootstrap	true
	Min-samples-split			
ANN			Structure	[64,32,16]
			activation	ReLU

1.3 评估指标

对于模型的性能,采用均方误差(MSE)、平均绝对误差(MAE)和决定系数(R^2)作为模型性能评价体系。其中, MSE 量化预测值与真实值的平方偏差均值,反映模型预测精度; MAE 表征预测误差的绝对量级,提供对异常值的鲁棒性评估^[17]; R^2 解析模型对目标变量方差变化的解释能力,表征拟合优度^[10]。所用的公式如下:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \in [0, +\infty) \quad (1)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \in [0, +\infty) \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \in [0, 1) \quad (3)$$

式中: n 为样本数量; i 为数据点的索引(即第 i 个样本); y_i 表示第 i 个样本的真实值; $\hat{y}_i$ 表示第 i 个样本的预测值。

1.4 可解释性

尽管机器学习通常被认为是“黑盒子”,但目前存在一些技术帮助理解机器学习的结果,对其结果进行解释。SHAP 就是其中发展的最为完备的 1 个, SHAP 的值通常基于博弈论进行计算^[18],可以表示每个类别预测概率的贡献,进而确定每个特征对最终预测的贡献^[12,15]。

2 结果与讨论

2.1 数据分析

图 2 为各特征统计信息与分布。由图 2 可见, 每个性能的 6 幅图中上方是先标准化后进行插值, 下方则是先插值后进行标准化的结果。对比可以看出, 标准化和插值的顺序并不影响分布形状, 但影响分布范围: 先标准化后插值的样本分布范围更窄, 而使用 RF 进行插值时先后顺序还会影响纵值取值

范围。

当数据缺失率达 30% 时(如性能 H 、 Den 和 R_f), RF 与其他两者的插值结果差异显著。这种现象源于简单的插值方式(尤其是中位数)无法在数据缺失率较大时有效捕捉数据规律, 此时应选择具备非线性建模能力的 RF。从插值后的分布可以发现, RF 插值后的数据通常有更好的分布形态, 这极有可能更利于后续模型从中学习到有效的规律进而实现更准确的预测。

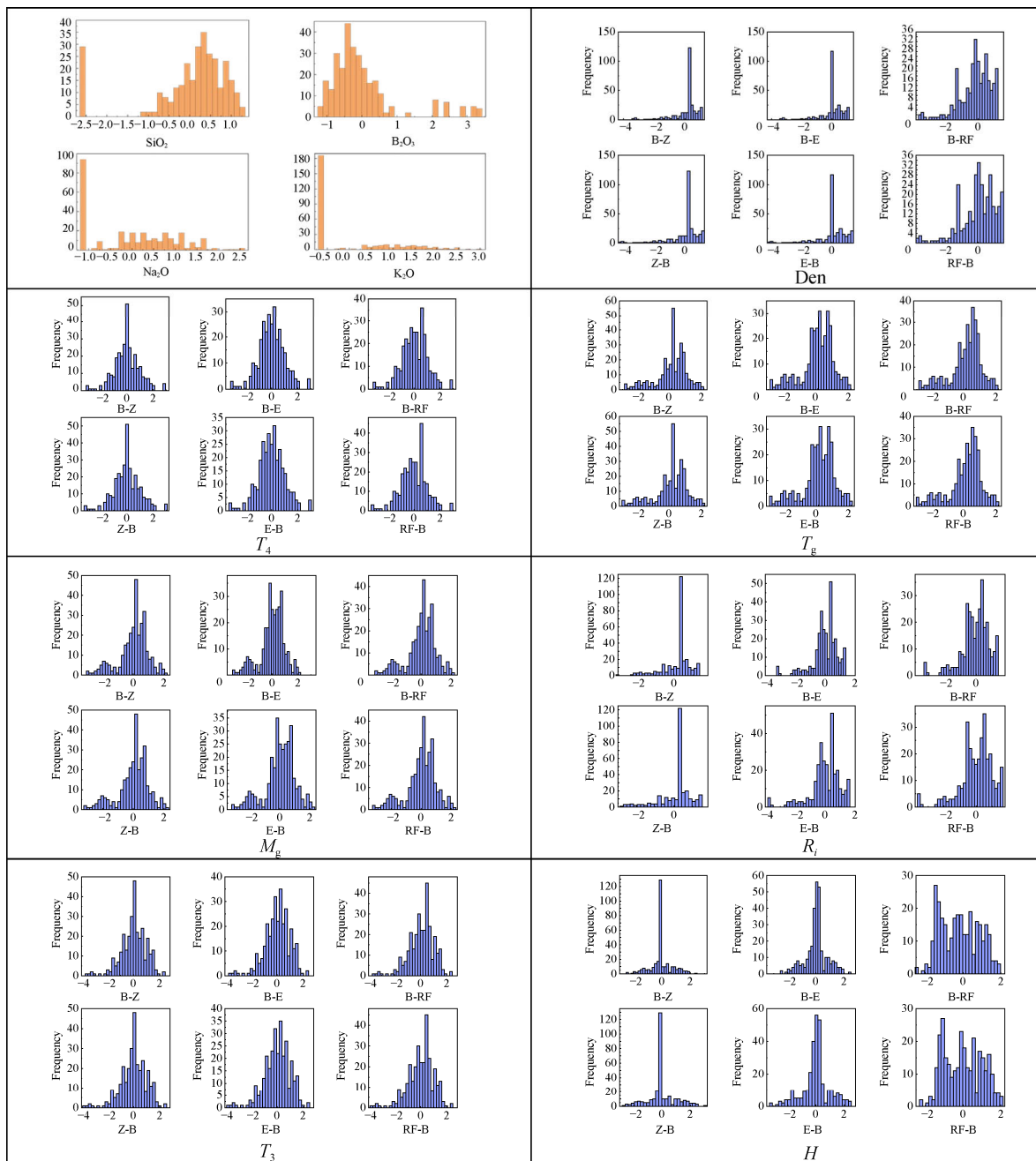


图 2 各特征统计信息与分布
Fig. 2 Statistical information and distribution of each feature

2.2 模型训练评估结果

2.2.1 模型性能对比 为对比模型架构的固有特性与玻璃材料之间的匹配程度，同时方便精准选择

模型进行后续的超参数调优，本节选用各模型默认参数对数据进行试验。图 3 分别展示了各模型针对测试集的评估结果，模型性能总体呈现 $\text{RF}=\text{XGB}>$

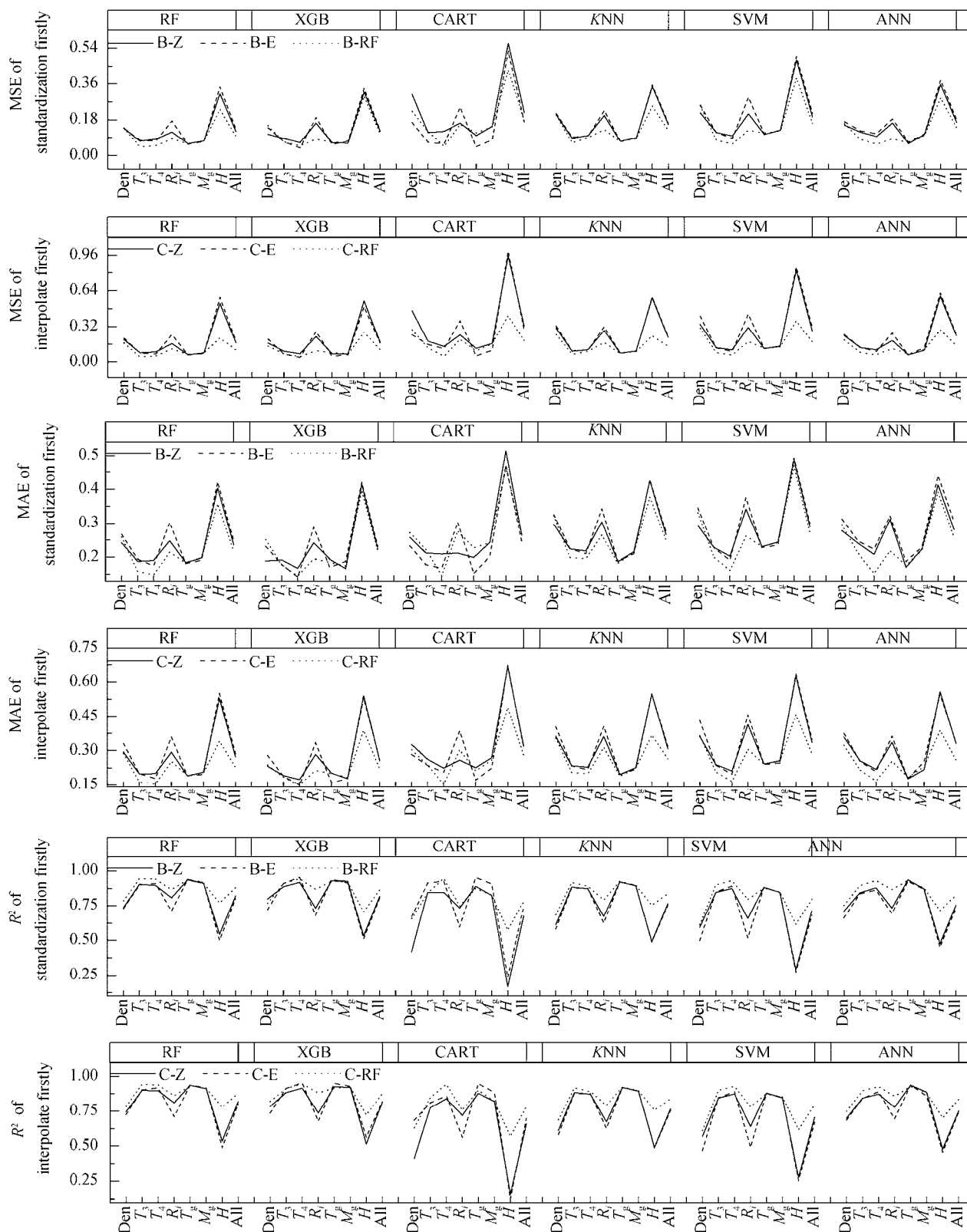


图 3 各模型的实验结果

Fig. 3 Experimental results of each model

KNN=ANN>SVM $\geq$ CART 的趋势。RF 与 XGB 作为树模型的集成算法, 组合了多个弱学习器的预测, 可以在模型复杂性和准确性之间取得平衡^[19-20], 故训练结果最好, 在各种数据中均保持最优的性能, 与现有文献一致^[21-22]。较为意外的是 KNN 在这种低维数据下表现得并不优异, 这与其对噪声和特征缩放敏感有关^[23-24]。而 ANN 结果与 KNN 类似, 这与数据样本量^[25]密切相关: 本数据集虽然特征少, 但样本量仅为 299, ANN 容易过拟合, 导致泛化能力差于集成方法。过少的数据量需要结合物理化学原理进行预测可以达到更好的效果^[26]。CART 的局限性源于在低维数据下, 单棵决策树容易过拟合或欠拟合^[27]。而 SVM 是因为对超参数敏感^[28], 而本模型选择了不合适的超参数。

为评估各模型的泛化能力, 此处使用 1 个之前未被使用过的样本量为 30 的小数据集 B 对模型进行测试。表 3 则展示了训练后各模型对未知数据集的泛化能力, 通过 MSE、MAE、 R^2 对其进行评估发现: RF、XGB 依然保持最佳的整体性能, 然后

依次是 ANN、KNN、SVM、CART。该排序与上述的结论相同, 但显然模型在新数据集的性能差于测试集, 这是机器学习中的典型情况^[15]。

表 3 各模型在数据集 B 上的性能评估

Table 3 Performance evaluation of each model on dataset B

Model	Dataset B-MSE	Dataset B-MAE	Dataset B- R^2
RF	0.141 1	0.259 9	0.780 8
XGB	0.132 6	0.234 6	0.794 1
CART	0.305 8	0.380 2	0.525 0
KNN	0.203 4	0.321 1	0.683 9
ANN	0.170 5	0.295 5	0.735 1
SVM	0.230 4	0.307 3	0.642 1

2.2.2 预处理方式对模型效果的影响 本工作采用方差分析(ANOVA)分别计算顺序(Sequence)、插值方法(Method)对每个指标的独立影响, 并判断这两者是否存在协同作用。计算结果包含 Sum-sq、显著性 p 值(PR)及偏 η^2 , 并将 p 值 <0.05 (即为该指标对结果有影响)的项展示在表 4。

表 4 预处理方式对不同性能的影响评价

Table 4 Evaluation of the influence of pretreatment methods on different properties

Index	Sum-sq	PR(>F)	Properties of materials	Performance of model	Partial eta squared/%
Sequence	0.072 356	0.000 006 76	H	MAE	49.64
	0.292 757	0.000 198	H	MSE	37.45
	0.042 125	0.000 203	R_i	MSE	37.35
	0.021 34	0.002 167	R_i	MAE	27.28
	0.053 7	0.003 146	Den	MSE	25.58
	0.023 057	0.003 642	Den	MAE	24.91
Method	0.217 131	0.000 000 252	R_i	R^2	63.68
	0.114 08	0.000 000 549	R_i	MSE	61.75
	0.638 734	0.000 005 84	H	R^2	55.22
	0.087 726	0.000 000 755	H	MAE	54.44
	0.056 407	0.000 032 5	R_i	MAE	49.79
	0.435 094	0.000 071 4	H	MSE	47.08
	0.011 466	0.000 186	T_4	MAE	43.59
	0.014 988	0.000 386	T_4	R^2	40.78
	0.009 593	0.000 619	T_4	MSE	38.89
	0.010 78	0.024 389	T_3	R^2	21.93
	0.006 42	0.030 606	T_3	MSE	20.74
	0.039 162	0.001 64	H	MAE	34.79
Sequence-method	0.171 59	0.010 981	H	MSE	25.98

由表 4 可以看出, 预处理顺序会显著影响 Den、 R_i 、 H 的 MSE 与 MAE, 偏 η^2 最高可达 49.64%, 且缺失值越多影响越强烈, 然而对 R^2 并无显著影响。插值方式则会在缺失值较多时(H 、 R_i)对 R^2 的影响显

著性多于对 MSE 与 MAE, 差距最高可达 13.89%。插值方式对 T_3 、 T_4 同样影响较大, 但对 Den 算不上影响显著。推测这与数据清洗时对热膨胀系数取对有关。至于两者间的交互作用则仅存在于 H 。

按照式(4)量化标准化优先(B)对于插值优先(C)策略的提升效果可以得到图 4。

$$\Delta I_{ijk} = \frac{B_{ijk} - C_{ijk}}{C_{ijk}} \quad (4)$$

式中： i 代表模型类型； j 代表性能类型； k 代表评价标准类型。结果表明，标准化优先，中位数及多项式插值下各模型的 MSE 与 MAE 性能普遍更好，尤其对于高缺失率的情况(如 H 、 Den 、 R_i)。这对 2.1 节中的猜测进行了论证与扩展：传统插值方法在处理高缺失率数据时，其处理效果对预处理顺序具有显著敏感性，而随机森林(RF)等基于机器学习的插值方法通过特征交互建模^[29]，可降低数据对顺序的敏感性。

2.3 SHAP 值结果

2.3.1 对预处理流程及模型分析 图 5 为缺失率分别为 35.1%、8.7%和 40.1%的 Den 、 T_3 和 H 数据在表现最好的模型(RF、XGB)与表现较差模型(KNN)的 SHAP 值雷达图。独立观察每个雷达图可以发现，预处理顺序对特征重要性排序的影响较弱，但会引发特征贡献值的偏移。然而对于缺失值过多的 H ，采用 RF 进行插值明显不同于中位数及多项式，结合 2.2 节中 RF 插值后的样本拥有更好的性能这一点，认为在数据缺失率较多时，RF 插值是优于中位数及多项式，可以帮助机器学习模型有效捕捉数据规律的插值方式。

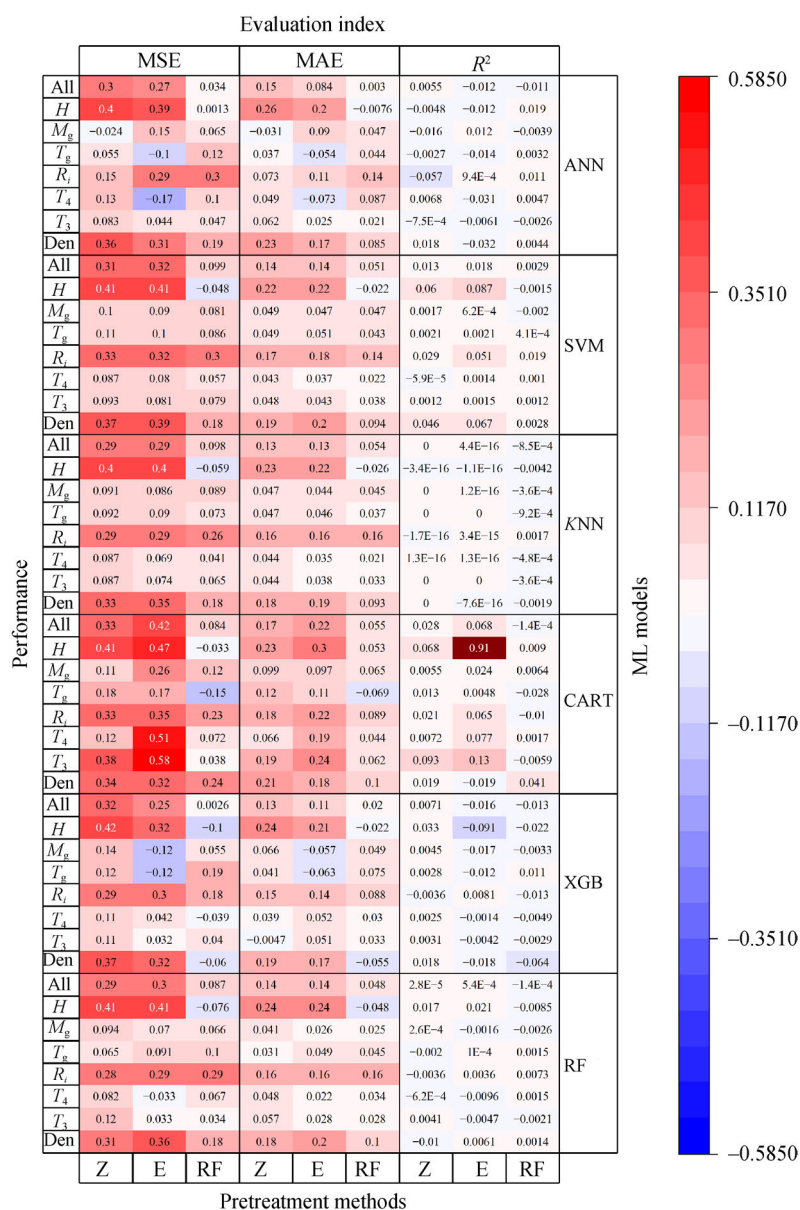


图 4 标准化优先下的性能评估结果相较于插值优先的相对变化百分比

Fig. 4 Relative percentage change in the performance evaluation results under normalized priority compared to interpolation priority

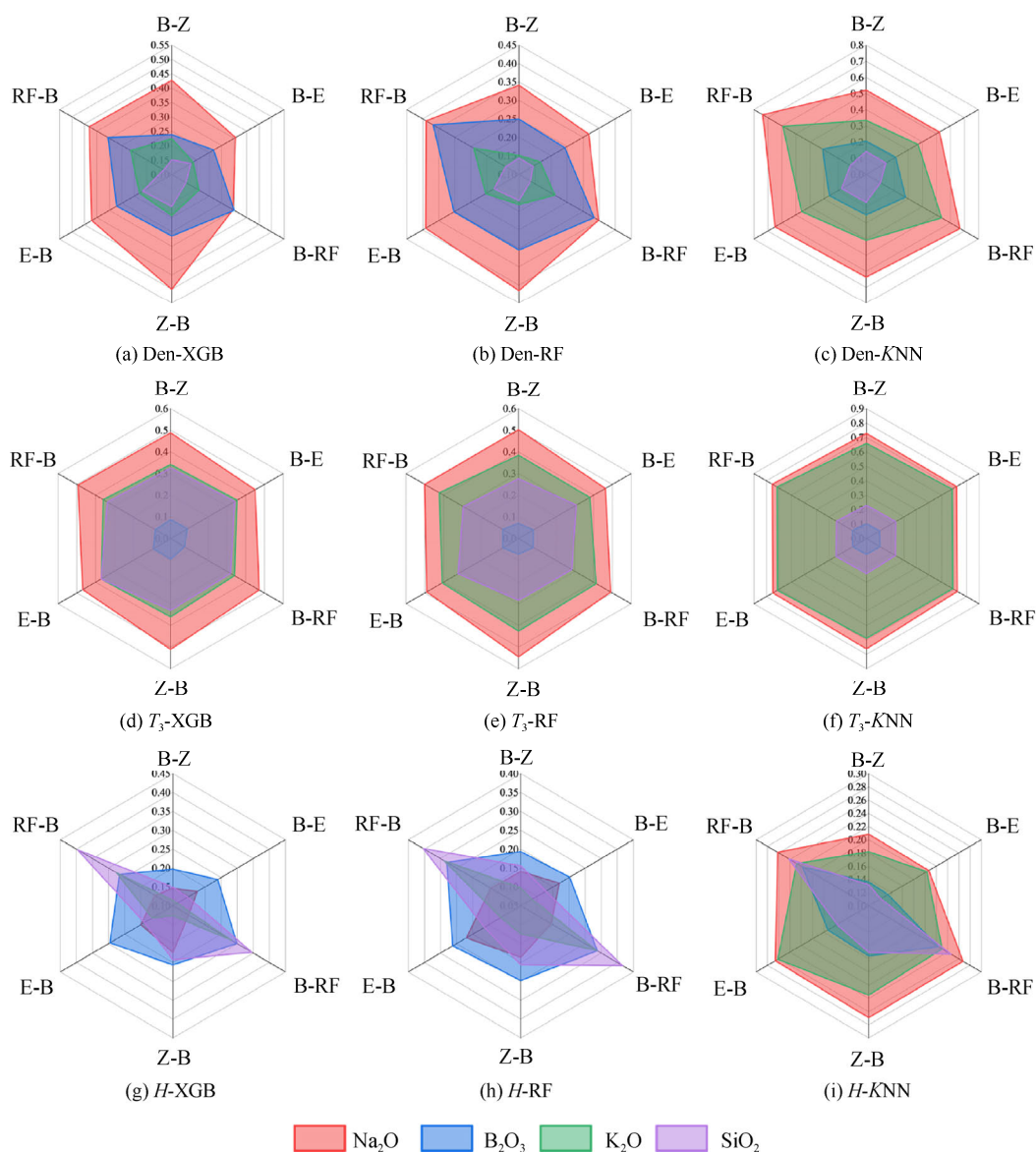


图 5 3 种典型算法在 3 类材料性能预测中的特征贡献度分布

Fig. 5 Characteristic contribution distribution of three typical algorithms in performance prediction of three kinds of materials performance

就模型性能,性能最好的 RF 与 XGB 基本上都保持一致的解析,鉴于其运算逻辑与 SHAP 的匹配性与其良好的预测能力^[30],认为这两者的解释拥有最大的可信度。相比之下 KNN 的可靠性较低,在 T_3 中还可以保持与 RF、XGB 一致的组分影响力排序,但随缺失值变多,排序差异越大,这与其为实例学习有关,使其无法再缺失值较多的情况下有效捕捉规律,进而模型性能较差。结合参考文献发现,此时对 KNN 应选择 SHAP 的变式^[31]或者 LIME 进行分析。

2.3.2 对 SHAP 结果分析 基于上述的可靠性验证,分析将聚焦于标准化优先策略时 RF、XGB 模型下除硬度外的组分重要性顺序:

Na_2O 对密度与折射率均起决定性作用,这与 Natalia 等^[32]关于 $\text{Na}_2\text{O}-\text{B}_2\text{O}_3-\text{SiO}_2$ 体系的研究结论一致:其分子动力学模拟表明, Na_2O 的结构单元分布直接影响材料密度。与 B_2O_3 主导玻璃化转变温度 (T_g) 的特性不同,膨胀软化温度 (M_g) 的核心调控组分为 SiO_2 , 其次为 B_2O_3 。这一现象与 Mengguo Ren 等的研究相印证:当 $\text{B}_2\text{O}_3/\text{SiO}_2$ 发生取代时,玻璃熔体中 B 离子的高扩散系数(相较于 Si 离子)显著改变了离子扩散行为,进而影响 T_g 值演变规律^[33]。但值得注意的是,本工作进一步显示:热膨胀系数范围为 20~300 °C 时, Na_2O 十分重要,其次是 K_2O ;但在 20~400 °C 时重要特征变为 Na_2O 和 SiO_2 , 这暗示热膨胀行为的阶段性特征可能源于主导组分的动态更替。

2.4 超参数调优结果

选择了SVM与RF在标准化优先的样本实验中进行了超参数调优,并对比了3种超参数调优方法:随机搜索(RS)、网格搜索优化(GS)、贝叶斯优化(BO)。对于标准化优先随机森林插值样本的密度性能,SVM、RF在调优后的超参数调整如表5所示。各样本各性能调优后相比调优前的性能提升率如图6、图7所示(统一MSE、MAE、 R^2 结果为正值——提升,负值——降低)。

表5 不同调优下B-RF-密度的参数细节

Table 5 Parameter details of B-RF-density under different optimizations

Model	Parameter	Value/RS	Value/GS	Value/BO
SVM	C	10	10	11.083 1
	Gamma	Scale	Scale	0.335 1
	kernel	rbf	rbf	rbf
RF	n-estimators	100	100	297
	Max-depth	10	none	13
	Min-samples-split	2	2	2

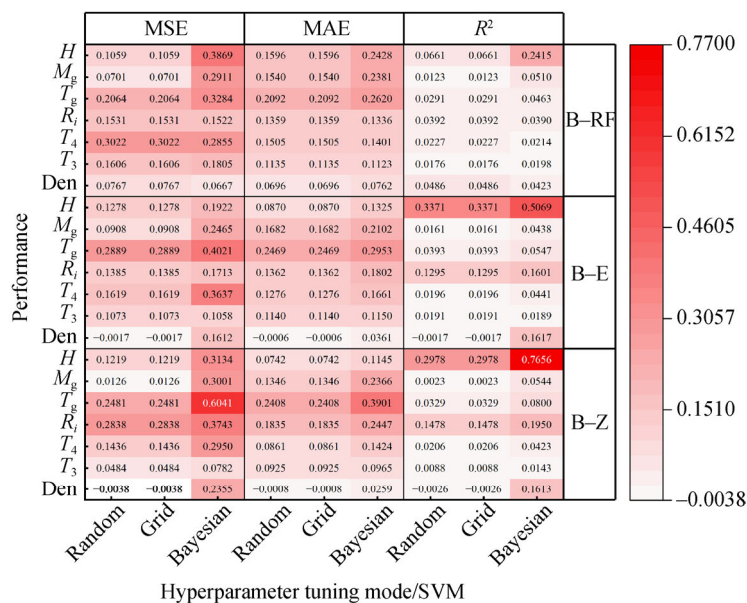


图6 超参数调优对SVM各指标性能的提升率

Fig. 6 Hyperparameter tuning improves the performance of each SVM index

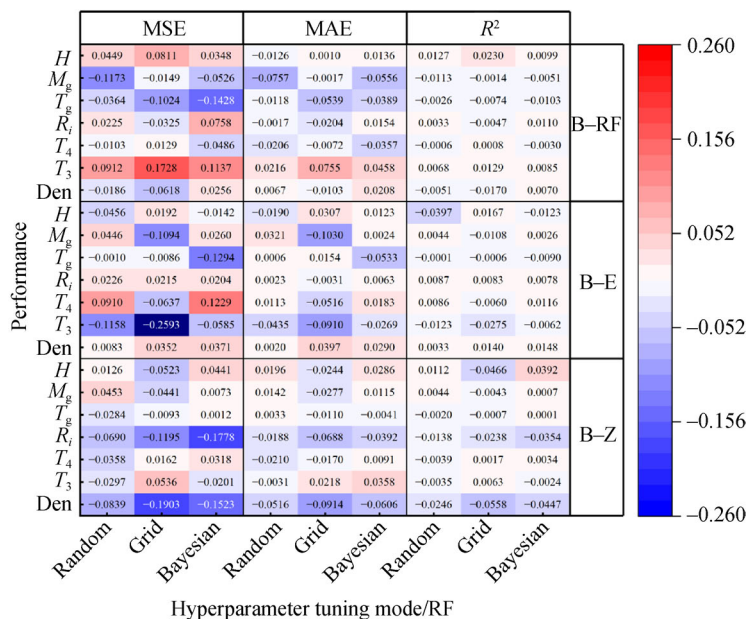


图7 超参数调优对RF各指标性能的提升率

Fig. 7 Hyperparameter tuning improves the performance of each RF index

可以看出超参数调优对于 SVM 的提升效果较好, MSE 、 MAE 、 R^2 的提升效果分别最高可达 60.41%、39.01%与 76.56%。优化后对 B-RF 处理下 T_3 、 T_4 、 T_g 、 M_g 的 R^2 分别从 90.1%、93.0%、87.6%、85.1%提升为 91.9%、95.0%、91.7%、89.4%。就超参数调优方法选择上, RS 与 GS 通常结果一致, 而 BO 对模型训练不同插值方式处理后的各种输出参数的各种性能均有提升, 且优于其他 2 种调优方法。这与 Wainer 等^[34]的研究结果一致, 其认为 BO 效果更好是源于其建模过程不仅允许优化器评估应该查询哪些数据点(因为它们可能产生良好的评估结果), 还允许优化器评估哪些数据点将改进其对响应函数本身的了解。

相比之下, 超参数调优对于 RF 的提升效果并不明显, 有时甚至还会降低。鉴于 Wang 等^[35]通过遗传算法(GA)优化超参数将 RF 的 R^2 从 0.844 提升到 0.972, 推测本工作中所用的超参数调优方式并不适合 RF。对比可以发现, GA 探索未知超参数空间的能力更强, 这是网格搜索等调优方法缺乏的。

3 结论

1) 实验表明预处理顺序对模型 MSE 、 MAE 的影响大于 R^2 , 最大可达 49.64%; 而插值方式在缺失值较多(高于 10%)环境下对 R^2 的影响更大, 最高可达 63.68%。其中 B-RF 插值的组合效果最佳。对于 SHAP, 预处理顺序不改变特征间关联趋势, 但会显著改变数值分布形态, 使结果产生系统性偏差; 而插值方式在缺失率较高(>40%)时直接影响贡献度排序。

2) 6 种算法整体呈现 $RF=XGB>KNN=ANN>SVM\geq CART$ 的趋势。其中 RF 与 XGB 凭借其集成算法的优势表现最好。模型表现并不受缺失率影响, 但对模型的解释(SHAP)会在缺失率较高时表现不同。

3) 对于组分与性能之间的关系, 与 T_g 由 B_2O_3 主导不同, M_g 的关键成分首先是 SiO_2 , 其次才是 B_2O_3 ; Na_2O 在密度与折射率中均占主导地位。然而热膨胀系数的主导组分会随温度的范围而变化: 范围为 20~300 °C 时, 主次主导组分分别为 Na_2O 、 K_2O ; 20~400 °C 时则分别为 SiO_2 、 Na_2O 。

4) 本工作选择的超参数调优方式(尤其是贝叶斯优化)在对超参数敏感的低准确率模型 SVM 上有很好的调优效果, 对 R^2 提升效果最佳时可达 76.56%, 但对于集成算法 RF 改善不佳, 应采用搜

索能力更强的 GA 等算法。

(第二十七届中国科协年会学术论文)

参考文献:

- [1] MITCHELL T M. Machine Learning[M]. New York: McGraw-Hill, 1997
- [2] WIEDERHOLD G, MCCARTHY J. Arthur Samuel: Pioneer in machine learning[J]. IBM J Res Dev, 1992, 36(3): 329–331.
- [3] GUO K, YANG Z Z, YU C H, et al. Artificial intelligence and machine learning in design of mechanical materials[J]. Mater Horiz, 2021, 8(4): 1153–1172.
- [4] HASSANIEN A E, DARWISH A, EL-ASKARY H. Machine Learning and Data Mining in Aerospace Technology[M]. Cham: Springer International Publishing, 2020.
- [5] ZHONG S F, ZHANG K, BAGHERI M, et al. Machine learning: New ideas and tools in environmental science and engineering[J]. Environ Sci Technol, 2021, 55(19): 12741–12754.
- [6] LEE J, PARK D, LEE M Y, et al. Machine learning-based inverse design methods considering data characteristics and design space size in materials design and manufacturing: A review[J]. Mater Horiz, 2023, 10(12): 5436–5456.
- [7] MAURO J C. Decoding the glass genome[J]. Curr Opin Solid State Mater Sci, 2018, 22(2): 58–64.
- [8] SINGH J, SINGH S. A review on Machine learning aspect in physics and mechanics of glasses[J]. Mater Sci Eng B, 2022, 284: 115858.
- [9] BROOK R J, ARNOLD G C. Applied Regression Analysis and Experimental Design[M]. Boca Raton: CRC Press, 1985.
- [10] TIAN J, ZHAO Y X, HUANG Y P, et al. Theoretical prediction of vickers hardness for oxide glasses: Machine learning model, interpretability analysis, and experimental validation[J]. Materialia, 2024, 33: 102006.
- [11] LIAO K, WANG W J, MEI X S, et al. Shape regulation of tapered microchannels in silica glass ablated by femtosecond laser with theoretical modeling and machine learning[J]. J Intell Manuf, 2023, 34(7): 2907–2924.
- [12] MASTELINI S M, CASSAR D R, ALCOBACA E, et al. Machine learning unveils composition-property relationships in chalcogenide glasses[J]. Acta Mater, 2022, 240: 118302.
- [13] SHIH Y T, SHI Y F, HUANG L P. Predicting glass properties by using physics- and chemistry-informed machine learning models[J]. J Non Cryst Solids, 2022, 584: 121511.
- [14] LIU C C, SU H. Prediction of optical properties of oxide glass combined with autoencoder and machine learning[J]. J Non Cryst Solids, 2024, 642: 123166.
- [15] ZHU J Q, SUN G H, DING L F, et al. Advancing the prediction of crystalline phases in glass-ceramics via machine learning[J]. J Non Cryst Solids, 2024, 646: 123251.
- [16] CASSAR D R, MASTELINI S M, BOTARI T, et al. Predicting and interpreting oxide glass properties by machine learning using large datasets[J]. Ceram Int, 2021, 47(17): 23958–23972.
- [17] LIU C C, SU H. Prediction of glass transition temperature of oxide glasses based on interpretable machine learning and sparse data sets[J]. Mater Today Commun, 2024, 40: 109691.
- [18] LUNDBERG S M, NAIR B, VAVILALA M S, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery[J]. Nat Biomed Eng, 2018, 2(10): 749–760.

- [19] GHANAVATI M A, AHMADI S, ROHANI S. A machine learning approach for the prediction of aqueous solubility of pharmaceuticals: A comparative model and dataset analysis[J]. *Digit Discov*, 2024, 3(10): 2085–2104.
- [20] CHEN T Q, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA. ACM, 2016: 785–794.
- [21] ZOUNEMAT-KERMANI M, STEPHAN D, BARJENBRUCH M, et al. Ensemble data mining modeling in corrosion of concrete sewer: A comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models[J]. *Adv Eng Inform*, 2020, 43: 101030.
- [22] IMRAN M M H, JAMALUDIN S, MOHAMAD AYOB A F. A critical review of machine learning algorithms in maritime, offshore, and oil & gas corrosion research: A comprehensive analysis of ANN and RF models[J]. *Ocean Eng*, 2024, 295: 116796.
- [23] LIU J, FU S M, LIU F, et al. Intelligent fault diagnosis of belt conveyor rollers using a polar KNN algorithm with audio features[J]. *Eng Fail Anal*, 2025, 168: 109101.
- [24] YANG P B, CHAN Y J, YAZDI S K, et al. Optimisation and economic analysis of industrial-scale anaerobic co-digestion (ACoD) of palm oil mill effluent (POME) and decanter cake (DC) using machine learning models: A comparative study of Gradient Boosting Machines (GBM), K-nearest neighbours (KNN), and random forest (RF)[J]. *J Water Process Eng*, 2024, 58: 104752.
- [25] ZHANG C Y, PAN Y, CHEN J J. Image-based prediction for enclosure structure deformation in pipe-roof tunnel construction using a physical-guided and generative deep learning method[J]. *Autom Constr*, 2025, 171: 106002.
- [26] ZHANG Z L, PAN Q J, YANG Z H, et al. Physics-informed deep learning method for predicting tunnelling-induced ground deformations[J]. *Acta Geotech*, 2023, 18(9): 4957–4972.
- [27] MANDAL B. Leveraging machine learning for analyzing the nexus between land use and land cover change, land surface temperature and biophysical indices in an eco-sensitive region of Brahmani-Dwarka interfluvium[J]. *Results Eng*, 2024, 24: 102854.
- [28] CANDELIERI A, SORMANI R, AROSIO G, et al. A hyper-resolution SVM classification framework: Application to on-line aircraft structural health monitoring[J]. *Procedia Soc Behav Sci*, 2014, 108: 57–68.
- [29] MARIANO C, MÓNICA B. A random forest-based algorithm for data-intensive spatial interpolation in crop yield mapping[J]. *Comput Electron Agric*, 2021, 184: 106094.
- [30] 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, California, USA. ACM, 2017: 4768–4777.
- [31] RIBEIRO M T, SINGH S, GUESTRIN C. “Why should I trust you?”: Explaining the predictions of any classifier[C]//*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA. ACM, 2016: 1135–1144.
- [32] Vedishcheva N M, Shakhmatkin B A, Wright A C. Thermodynamic modelling of the structure and properties of glasses in the systems $\text{Na}_2\text{O}-\text{B}_2\text{O}_3-\text{SiO}_2$ and $\text{Na}_2\text{O}-\text{CaO}-\text{SiO}_2$ [J]. *Phys Chem Glasses*, 2005, 46(2): 99–107.
- [33] REN M G, LU X N, DENG L, et al. $\text{B}_2\text{O}_3/\text{SiO}_2$ substitution effect on structure and properties of $\text{Na}_2\text{O}-\text{CaO}-\text{SrO}-\text{P}_2\text{O}_5-\text{SiO}_2$ bioactive glasses from molecular dynamics simulations[J]. *Phys Chem Chem Phys*, 2018, 20(20): 14090–14104.
- [34] WAINER J, FONSECA P. How to tune the RBF SVM hyperparameters? An empirical evaluation of 18 search algorithms[J]. *Artif Intell Rev*, 2021, 54(6): 4771–4797.
- [35] WANG Z N, WU F X, HAO N, et al. The combined machine learning model SMOTER-GA-RF for methane yield prediction during anaerobic digestion of straw lignocellulose based on random forest regression[J]. *J Clean Prod*, 2024, 466: 142909.

Impacts of Preprocessing Methods and Machine Learning Algorithms on Glass Material Property Prediction: Comparative Study and Optimization Strategies

WANG Siyu^{1,2}, SUN Yuancheng¹, YU Rui², DU Xiurong¹, ZHOU Lisheng¹, LIU Xing¹, CHEN Zhiwei^{1,3}

(1. China Building Materials Academy, Beijing 100024, China; 2. School of Materials and Microelectronics, Wuhan University of Technology, Wuhan 430070, China; 3. Materials Science and Engineering, Wuhan University of Technology, Wuhan 430070, China)

Extended Abstract

Introduction The existing research on modeling glass properties has some limitations in critical modeling stages, thus hindering predictive accuracy and interpretability. Primarily, systematic data pre-processing is often neglected, *i.e.*, most literature relies solely on basic data cleaning (*e.g.*, removing outliers and duplicates), lacking thorough investigation into optimal missing value imputation methods and the crucial sequence of imputation versus feature standardization. Secondly, hyperparameter optimization strategies remain simplistic. As some advanced techniques like random search or Bayesian optimization are introduced, studies generally overlook a critical adaptation of tuning methods to specific algorithm characteristics and a lack comparative analysis of hyperparameter optimization approaches across different models. This potentially restricts the performance of hyperparameter-sensitive models like Support Vector Machines (SVM). To address these challenges, this study was to propose a comprehensive, optimized framework for glass property prediction.

Methods This research established a fully optimized framework for glass property prediction. The methodology comprised four integrated stages, *i.e.*, 1) Systematic Data Preprocessing: For dataset A, the framework rigorously evaluated the combinatorial effects

of three missing value imputation strategies (*i.e.*, Median Imputation, Polynomial Imputation, Random Forest Imputation) and two standardization sequences (*i.e.*, Imputation followed by Standardization; Standardization followed by Imputation); 2) Predictive Model Construction and Evaluation: Six distinct machine learning algorithms, which were chosen for their varied inductive biases (*i.e.*, Random Forest-RF, XGBoost-XGB, Classification and Regression Trees-CART, K-Nearest Neighbors-KNN, Support Vector Machines-SVM, Artificial Neural Networks-ANN), were employed. The model predictive efficacy was rigorously quantified using a triple evaluation metric (*i.e.*, Mean Squared Error-MSE, Mean Absolute Error-MAE, Coefficient of Determination- R^2). The model generalization was validated by a novel dataset (A); 3) Interpretable Analysis: The SHAP (SHapley Additive exPlanations) framework was applied to elucidate the contribution paths of key glass components towards the predicted properties; and 4) Hyperparameter Optimization Strategy Analysis: For representative models SVM and RF, a comparative experiment was conducted by three hyperparameter tuning methods, *i.e.*, Random Search, Grid Search, and Bayesian Optimization. This stage was to establish adaptable criteria matching hyperparameter tuning strategies to model architectures.

Results and Discussion

The results show that the sequence of standardization and imputation significantly affects model performance metrics, with a substantially larger impact observed on MSE and MAE (up to 49.64% variation), compared to R^2 . Conversely, the choice of imputation method exerts a greater influence on R^2 (up to 63.68%) under conditions of high missing data rates (*i.e.*, >10%). The optimal preprocessing is standardization-first followed by Random Forest Imputation. The SHAP analysis reveals that preprocessing sequence does not alter the fundamental trends in feature contributions, but significantly changes the numerical distribution morphology of SHAP values, thus introducing systematic bias. Imputation method, particularly at high missing rates (*i.e.*, >40%), directly affects the ranking of feature importance derived from SHAP.

The comparative performance of the six algorithms, evaluated across metrics, follows an order of $RF \approx XGB > KNN \approx ANN > SVM \geq CART$. The RF and XGB demonstrate a superior performance, leveraging their ensemble learning advantages. The predictive accuracy of model (as measured by MSE, MAE, and R^2) is robust to varying missing data rates. However, the interpretability of model predictions *via* SHAP becomes less reliable at high missing rates.

The SHAP analysis indicates distinct governing components for different properties, sometimes diverging from conventional understanding. While B_2O_3 dominates the Glass Transition Temperature (T_g), M_g content is primarily governed by SiO_2 , and followed by B_2O_3 . Na_2O is a dominant contributor to both Density and Refractive Index. The primary contributors to the thermal expansion coefficient (TEC) are temperature-range dependent. Na_2O followed by K_2O dominates at 20–300 °C, whereas SiO_2 followed by Na_2O dominates at 20–400 °C.

The use of hyperparameter tuning method proves critical, especially for hyperparameter-sensitive models. The Bayesian optimization can obtain the most substantial improvements for the lower-accuracy SVM model, achieving a maximum R^2 improvement of 76.56%. In contrast, its impact on the high-performing ensemble algorithm RF is marginal, indicating that alternative, potentially more explorative methods like Genetic Algorithms (GA) can be more suitable for such models. This highlights a necessity for model-specific hyperparameter optimization strategies.

Conclusions This study proposed and validated a comprehensive optimization framework for glass property prediction, addressing key limitations in data preprocessing and hyperparameter tuning prevalent in the existing literature. The systematic evaluation demonstrated that preprocessing choices, particularly the sequence of operations and imputation method under high missing data scenarios, had profound and quantifiable impacts on both predictive accuracy (MSE, MAE, R^2) and the reliability of interpretability frameworks like SHAP. Among the evaluated algorithms, RF and XGB consistently outperformed others. Critically, the SHAP analysis provided insights into composition–property relationships, thus revealing temperature–dependent dominance for TEC and challenging assumptions regarding M_g 's primary drivers. Finally, the hyperparameter optimization experiments established that tuning strategy efficacy could be highly model-dependent. The Bayesian optimization offered significant gains for sensitive models like SVM but was less effective for robust ensembles like RF, necessitating tailored approaches. This framework could provide a validated, end-to-end methodology for enhancing the accuracy, reliability, and interpretability of predictive models in glass science and potentially other material domains.

Keywords machine learning algorithm comparison; property prediction; hyperparameter tuning; data preprocessing