

基于机器学习的小样本超高性能混凝土性能预测及可解释性分析

张增起^{1,2}, 杜伟杰², 刘 宇³, 马善亮², 刘晓明^{1,2}, 荆 锐⁴

(1. 北京科技大学绿色低碳钢铁冶金全国重点实验室, 北京 100083; 2. 北京科技大学冶金与生态工程学院, 北京 100083; 3. 中建一局集团建设发展有限公司, 北京 100102; 4. 唐山工业职业技术大学, 河北 唐山 063299)

摘 要: 为在小样本条件下实现超高性能混凝土(UHPC)性能的准确预测, 并揭示多组分胶凝材料对 UHPC 力学性能的影响规律, 融合了条件生成对抗网络(CTGAN)、贝叶斯优化的极限梯度提升树(BO-XGBoost)模型以及可解释性机器学习方法, 建立了一套用于多响应预测与分析的框架。以 160 条实验数据为原始数据集, 采用 CTGAN 生成合成样本进行数据增强, 建立以矿物掺合料类型和掺量、减水剂掺量和养护龄期为输入, 水膜层厚度与抗压强度为输出的 BO-XGBoost 预测模型, 并结合沙普利可加性解释值(SHAP)和部分依赖图(PDPs)分析各变量对输出性能的影响规律。结果表明, 利用 CTGAN 生成的数据在分布形态与变量相关性上与真实数据保持高度一致, 有效缓解了小样本带来的数据不足的问题。BO-XGBoost 模型在测试集上对水膜层厚度和抗压强度的预测决定系数(R^2)分别达到 0.986 和 0.976, 误差分布集中且对称, 表现出优异的预测精度与泛化能力。SHAP 与 PDPs 分析表明了不同矿物掺合料对水膜层厚度与抗压强度的影响具有显著差异。粉煤灰微珠和硅灰在中低掺量下对水膜层厚度具有正向作用, 而多种掺合料在高掺量条件下因颗粒堆积致密化表现为负向影响。对于抗压强度, 超细矿渣粉、硅灰和粉煤灰微珠整体呈正向贡献, 矿渣粉与粉煤灰在高掺量条件下因活性受限表现为负向作用。该框架为小样本条件下 UHPC 的配合比设计与性能优化提供了有效的数据驱动手段。

关键词: 条件生成对抗网络; 超高性能混凝土; 数据增强; 水膜层厚度; 抗压强度

中图分类号: TU526 文献标志码: A 文章编号: 0454-5648(2026)10-3436-17

网络出版时间: 2026-09-01



超高性能混凝土(UHPC)因具有优异的力学性能与耐久性能, 已在桥梁工程、装配式结构以及 3D 打印建筑等领域得到广泛应用^[1-3]。在材料组成方面, UHPC 通常通过掺入硅灰、矿渣、粉煤灰等矿物掺合料, 并配合高效减水剂调控颗粒级配与界面结构, 以获得高强度和高致密性的材料体系^[4, 5]。大量研究表明, 矿物掺合料的种类、粒度、掺量等对浆体流变行为、结构建立以及力学性能具有显著影响, 其作用机制复杂且差异明显^[6]。同时, 由于 UHPC 体系中固相体积分数高、颗粒间相互作用强, 其性能往往表现出明显的非线性及时间依赖性, 使得传统经验方法难以准确预测其性能^[7]。

水膜层厚度(WFT)是表征颗粒体系中自由水在固体颗粒表面分布状态的重要微观参数, 本质上反映了单位颗粒表面积上可用于润滑的水量^[8]。对于 UHPC 这类高固相体积分数体系而言, 拌合水在填充颗粒间孔隙后所形成的剩余水膜直接决定了颗粒

间的润滑作用与分散状态, 从而显著影响浆体的流动性与结构建立过程^[9]。研究表明, 当水膜层厚度不足时, 颗粒间摩擦与絮凝作用增强, 体系表现出较高屈服应力与黏度; 而随着水膜层厚度增加, 颗粒分散性改善, 流动性显著提升, 但过厚的水膜又可能导致结构稳定性下降^[8]。此外, 不同矿物掺合料由于其粒径分布、比表面积及表面化学特性方面存在差异, 可通过改变颗粒堆积密实度和体系需水量间接调控水膜层厚度, 使水膜层厚度受多因素耦合作用影响, 呈现出较强的非线性变化特征^[10]。

近年来, 机器学习方法在水泥基材料性能预测领域展现出应用前景。Fan 等^[11]基于遗传算法优化的人工神经网络(GA-ANN)实现了密实颗粒堆积的 UHPC 性能的高精度预测。Aylas-Paredes 等^[12]通过构建大规模数据集并开发先进的机器学习模型, 系统验证了机器学习在 UHPC 性能预测中的可行性与优越性。此外, 部分研究还结合了特征选择与超参

收稿日期: 2026-05-01。 修订日期: 2026-06-13。

基金项目: 国家自然科学基金面上项目(52478232)。

第一作者: 张增起(1991—), 男, 博士, 教授。

Received date: 2026-05-01. Revised date: 2026-06-13.

First author: ZHANG zengqi (1991—), male, Ph.D. Professor.

E-mail: zhangzq@ustb.edu.cn

数优化的方法,如遗传算法、粒子群优化等,进一步提升模型预测性能^[13]。与此同时,沙普利可加性解释值(SHAP)、特征重要性分析等可解释性机器学习方法也逐渐被用于揭示各输入变量对模型输出的影响规律,从而增强模型结果的可解释性和工程应用价值^[14]。然而,现有研究多依赖较大规模的实验数据集,而 UHPC 的试验成本高昂、数据获取难度大,小样本问题尤为突出,严重限制了模型预测精度和泛化能力的进一步提升^[15]。

针对小样本问题,生成式模型为数据增强提供了新的解决思路。条件生成对抗网络(Conditional Tabular GAN, CTGAN)能够学习表格数据中的复杂联合分布,并生成符合原始数据统计特征的合成样本,从而扩展数据规模、提高模型训练效果^[16]。在此基础上,本工作以不同矿物掺合料掺量、减水剂掺量和养护龄期为输入变量,以水膜层厚度和抗压强度为并行输出变量,构建融合 CTGAN、贝叶斯

优化的极限梯度提升树(BO-XGBoost)和可解释性机器学习的智能化设计框架。通过 CTGAN 扩充 UHPC 小样本数据集,采用贝叶斯优化提升 XGBoost 模型预测性能,并结合 SHAP 和部分依赖图(PDPs)分析不同输入变量对水膜层厚度及抗压强度的影响规律,以期 UHPC 配合比设计和性能调控提供参考。

1 实验与算法

1.1 原材料

制备 UHPC 所用水泥为 P·I 42.5 基准水泥(PC),所用骨料为粒径范围在 0.84~2.00 mm 的石英砂。综合选取了硅灰(SF)、S95 矿渣粉(GGBS)、超细矿渣粉(sGGBS)、二级粉煤灰(FA)、粉煤灰微珠(sFA) 5 种矿物掺合料。此外,使用减水剂控制胶凝材料的流动度。使用的各类材料的化学成分见表 1。

表 1 各原材料的化学组成
Table 1 Chemical composition of raw materials

Material	SiO ₂	Al ₂ O ₃	Fe ₂ O ₃	CaO	MgO	SO ₃	Na ₂ O	LOI
PC	22.11	4.43	3.13	62.38	2.28	2.62	0.53	2.04
SF	96.36	0.33	0.05	0.58	0.34	1.05	0.86	2.11
GGBS	26.76	15.26	0.77	45.92	6.30	1.16	0.79	1.30
sGGBS	24.11	13.92	0.90	51.16	5.40	1.89	0.54	1.66
FA	45.52	34.39	7.34	6.45	0.70	0.60	1.22	5.08
sFA	47.62	30.92	8.61	5.27	0.73	0.88	1.59	2.32

w is mass fraction; LOI is loss on ignition; PC is portland cement; SF is silica fume; GGBS is ground granulated blast-furnace slag; sGGBS is superfine ground granulated blast-furnace slag; FA is fly ash; sFA is fly ash microspheres.

1.2 实验配比与数据库

本工作设置了二元、三元以及四元 3 种胶凝材料体系,通过实际实验获得水膜层厚度和抗压强度相关原始数据。二元体系中,分别使用 1 种矿物掺合料等体积替代水泥,替代率设置为 0、5%、10%、15%和 20%。三元体系在二元体系的基础上,固定硅灰掺量为 10%。四元体系是在三元体系的基础上将矿渣粉的掺量设定为 20%,但不添加粉煤灰和粉煤灰微珠。为计算水膜层厚度,本工作采用湿堆积法测定胶凝材料的堆积密度,该方法在新拌浆体条件下进行测试。通过测试胶凝材料浆体的湿堆积密度和孔隙率,通过下列公式计算水膜层厚度。

$$u_s = u_w - g \quad (1)$$

$$A_M = A_a \times R_a + A_b \times R_b \quad (2)$$

$$\text{WFT} = \frac{u_s}{A_M} \quad (3)$$

其中: u_s 是剩余水体积与固相体积的比值; u_w 是水胶体积比; g 为孔隙率; A_M 是固相的比表面积; A_a

和 R_a 分别是水泥的比表面积与体积分数; A_b 和 R_b 分别是矿物掺合料的比表面积和体积分数;WFT是水膜层的平均厚度。

具体胶凝材料净浆的配合比如表 2 所示。按照配比制得 UHPC 后养护 3、7、28 d 和 90 d 并测试其抗压强度,从而获得原始数据集,共包含 160 条数据。

1.3 数据增强模型

CTGAN 是在传统 GAN 基础上发展而来的一种专门用于表格数据建模的生成模型,其核心是通过生成器(Generator)与判别器(Discriminator)之间的对抗训练,使得生成数据逐步逼近真实数据分布^[17]。CTGAN 引入了条件向量与模式特定归一化等机制,对复杂分布的拟合能力较强。在本工作中,针对本工作所构建的 UHPC 小样本数据库,采用 CTGAN 模型进行数据增强。首先,对原始数据中的连续变量与类别变量进行识别与预处理,并构建条件输入向量;随后,通过对抗训练学习各输入变量之间的

表 2 不同胶凝材料体系的配合比设计

Table 2 Mix proportions of different cementitious material systems

 $\varphi/\%$

Samples	Cementitious materials					
	Cement	SF	GGBS	sGGBS	FA	sFA
CS5-20	95, 90, 85, 80	5, 10, 15, 20				
CG5-20	95, 90, 85, 80		5, 10, 15, 20			
CsG5-20	95, 90, 85, 80			5, 10, 15, 20		
CF5-20	95, 90, 85, 80				5, 10, 15, 20	
CsF5-20	95, 90, 85, 80					5, 10, 15, 20
CSG5-20	85, 80, 75, 70	10	5, 10, 15, 20			
CSsG5-20	85, 80, 75, 70	10		5, 10, 15, 20		
CSF5-20	85, 80, 75, 70	10			5, 10, 15, 20	
CSsF5-20	85, 80, 75, 70	10				5, 10, 15, 20
CSGsG5-20	65, 60, 55, 50	10	20	5, 10, 15, 20		

 φ is volume fraction.

联合分布关系，从而生成新的合成样本^[18]。通过多次调整实验参数确定了数据合成规模：以水膜层厚度为研究对象时合成 1000 条样本数据，以抗压强度为研究对象时合成 900 条样本数据。其主要考虑是在小样本条件下充分扩展数据分布空间，同时避免生成数据过多可能带来的分布偏移与冗余信息问题。在生成数据过程中，CTGAN 能够保持原始数据中变量之间的相关性及分布特征，有效缓解小样本带来的数据稀疏问题。

1.4 机器学习模型

XGBoost 是一种基于梯度提升决策树的集成学习方法，其基本原理是通过加性模型，逐步向模型中构建新的回归树，每一棵回归数都对前一轮模型的残差进行迭代拟合，从而不断提升机器学习模型的预测精度^[19]。对于本工作所构建的 UHPC 数据库而言，其输入变量涵盖二元、三元及四元胶凝材料体系。这些多元体系之间的组合方式复杂多变，输入特征呈现出较强的非线性耦合关系，对模型的解析能力提出了较高要求。XGBoost 具有较强的抗过拟合能力，并支持多种正则化机制和特征子采样策略，适用于本工作这种变量维度较多但样本量有限的场景，其能够有效捕捉上述多组分体系中各变量之间复杂的非线性关系，并建立输入特征与 UHPC 的水膜层厚度、抗压强度之间的映射关系。本工作采用 CTGAN 方法扩增原始数据集，得到的增强数据集被用于对 XGBoost 模型进行训练，从而提高 XGBoost 模型对不同配合比组合的学习能力，改善小样本条件下的预测性能，提升模型的泛化性能(同分布条件下的预测性能)。此外，为了进一步提高模型的预测精准度，本工作引入贝叶斯优化(BO)方法

对 XGBoost 模型的超参数进行自动寻优。贝叶斯优化是一种基于概率模型的全局优化方法，其可通过构建目标函数的代理模型在有限的搜索次数内高效寻优最优超参数组合^[20]。引入贝叶斯优化方法，不仅可以提高 XGBoost 模型在小样本数据条件下的预测精度与稳定性，同时还可以增强模型对不同配合比条件的泛化能力。

1.5 解释性方法

为了提高模型输出结果的可解释性，本工作引入了 SHAP 与 PDPs 两种解释性方法，用于分析 UHPC 配比参数对水膜层厚度和抗压强度的影响规律。

SHAP 方法来源于博弈论中的 Shapley 值理论，其核心思想是量化每个输入特征对模型预测结果的边际贡献，从而实现对复杂机器学习模型的全局与局部解释。具体而言，对于每个预测样本，SHAP 值能够量化各输入变量在预测过程中的正向或负向作用，从而揭示不同矿物掺合料对 UHPC 的水膜层厚度和抗压强度的影响程度及其作用方向^[21]。对任意一个样本 x ，假设模型输出为 $f(x)$ ，其 SHAP 解释可表示为式(4)：

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i \quad (4)$$

其中： ϕ_0 为模型在训练集上的基准输出值(Expectation value)； ϕ_i 为特征 i 对预测结果的 Shapley 贡献值； M 为输入特征数量。

第 i 个特征的 Shapley 值计算公式为：

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (5)$$

其中： N 表示全部特征的集合； S 为不包含特征

i 的特征子集； f_S 表示仅使用特征子集 S 时模型对样本的预测。

PDPs 方法可以用于描述单一输入变量或变量组合对模型输出的平均影响趋势^[22]。该方法通过在给定目标变量不同取值的条件下，对其他变量进行边际平均，并计算模型输出的变化情况。从而揭示输入变量与模型预测结果之间的响应关系^[22, 23]。

本工作将 SHAP 和 PDPs 方法结合使用，一方面，能够利用 SHAP 识别影响水膜层厚度和抗压强度的关键影响因素及其贡献方向；另一方面，通过 PDPs 分析关键变量在不同取值区间内对模型输出的影响趋势，可以进一步揭示多组分胶凝材料体系中配比参数与性能响应之间的非线性关系。

1.6 性能评估

为了评价模型的输出和拟合性能，本工作采用了决定系数(R^2)、平均绝对百分比误差(MAPE)、均方根误差(RMSE)以及平均绝对误差(MAE) 4 种常用的回归性能评价指标作为评价标准^[24]。具体的指标计算公式如下：

$$MAE = \frac{1}{n} \sum_{x=1}^n |y_x - \hat{y}_x| \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{x=1}^n (y_x - \hat{y}_x)^2} \quad (7)$$

$$MAPE = \frac{1}{n} \sum_{x=1}^n \left| \frac{y_x - \hat{y}_x}{y_x} \right| \quad (8)$$

$$R^2 = 1 - \frac{\sum_{x=1}^n (\hat{y}_x - y_x)^2}{\sum_{x=1}^n (\bar{y}_x - y_x)^2} \quad (9)$$

其中： n 是样本量； y_x 是真实值； $\hat{y}_x$ 是预测值； $\bar{y}_x$ 是真实值的平均值。

2 结果与讨论

2.1 真实与合成数据质量分析

本工作以硅灰、S95 矿渣粉、超细矿渣粉、二级粉煤灰和粉煤灰微珠掺量，以及养护龄期和聚羧酸减水剂(PCE)掺量为输入变量，以 UHPC 的水膜层厚度(WFT)和不同龄期的抗压强度作为输出变量，基于标准化实验获得了 160 条不同配合比的 UHPC 抗压强度数据，并以此构建了 UHPC 真实数据集。表 3 详细列出了原始真实数据集中输入变量与输出变量的描述性统计信息。图 1 和图 2 分别展示了真实数据集和 CTGAN 合成数据集中各变量的直方分布，图 3 和图 4 展示的是各变量的累积分布图。可以看出，CTGAN 生成样本能够较好地保留原始数据的主要分布特征，合成数据在变量取值范围和分布形态上与真实数据基本一致。部分原始数据中的离散值经过增强后变为连续值，数据集变得更加丰富，有助于提高模型预测的准确性。不同输入变量的数据分布均匀，有效避免了数据堆积对模型训练的影响。图 5 和图 6 给出了真实数据集与合成数据集中各变量之间的相关系数矩阵。对比可知，合成数据在关键变量相关关系上与真实数据保持较好一致性，说明 CTGAN 不仅能够学习单变量的边缘分布特征，也能够一定程度上捕捉变量之间的相关结构。总体来看，真实数据与合成数据在变量分布和相关性方面具有较高相似性，表明所建立的 CTGAN 模型能够较好地模拟原始数据分布，为后续 BO-XGBoost 模型训练提供可靠的数据基础。

表 3 真实数据集中输入变量与输出变量的描述性统计特征

Table 3 Descriptive statistics of the input and output variables in real dataset

Variable	Unit	Count	Mean	Std	Min	25%	50%	75%	Max	Type
Cem		160	0.81	0.11	0.50	0.75	0.80	0.86	0.95	Input
PCE		160	0.64	0.17	0.35	0.52	0.63	0.76	1.10	Input
SF		160	0.06	0.06	0.00	0.00	0.10	0.10	0.20	Input
GGBS		160	0.05	0.08	0.00	0.00	0.00	0.06	0.20	Input
sGGBS		160	0.04	0.07	0.00	0.00	0.00	0.05	0.20	Input
FA		160	0.03	0.06	0.00	0.00	0.00	0.00	0.20	Input
sFA		160	0.03	0.06	0.00	0.00	0.00	0.00	0.20	Input
CA	d	160	32.00	34.92	3.00	6.00	17.50	43.50	90.00	Input
WFT	nm	160	11.84	3.34	5.17	10.56	11.89	12.62	20.19	Output
CS	MPa	160	154.33	25.03	114.00	131.00	157.00	176.00	203.00	Output

Cem is cement; PCE is polycarboxylate superplasticizer; SF is silica fume; CA is curing age; WFT is water film thickness; CS is compressive strength.

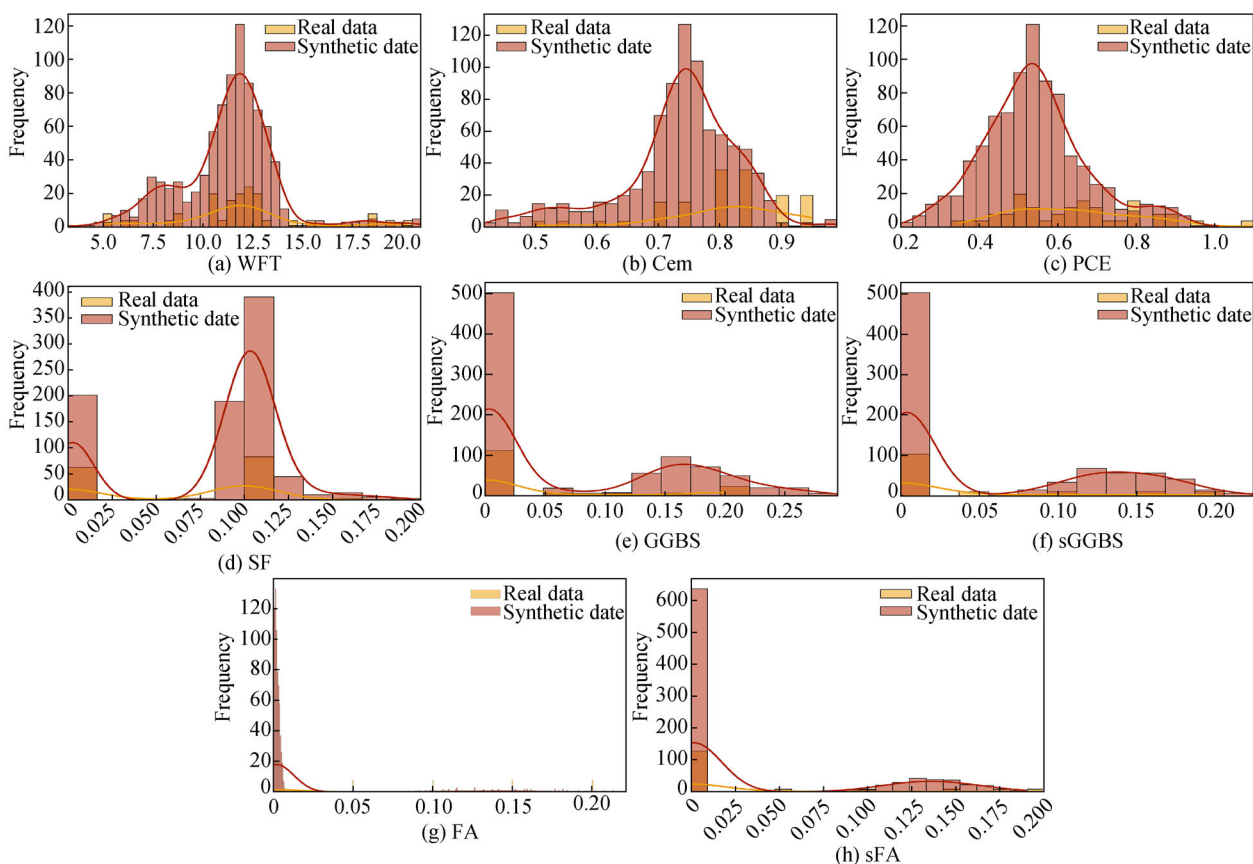


图 1 以水膜层厚度为输出时真实数据与合成数据中各变量的分布对比

Fig. 1 Distribution comparison of variables in real and synthetic datasets for WFT as output

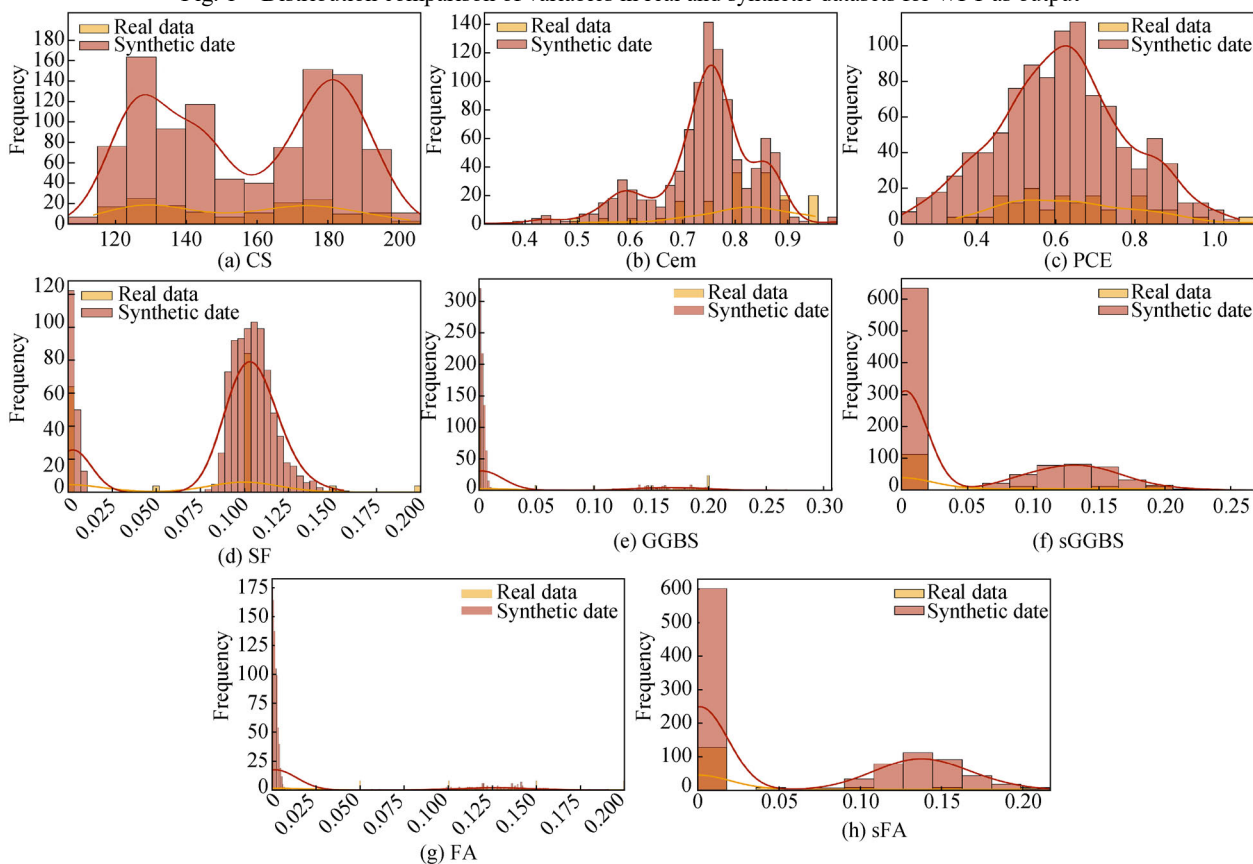


图 2 以抗压强度为输出时真实与合成数据中各变量的分布对比

Fig. 2 Distribution comparison of variables in real and synthetic datasets for compressive strength as output

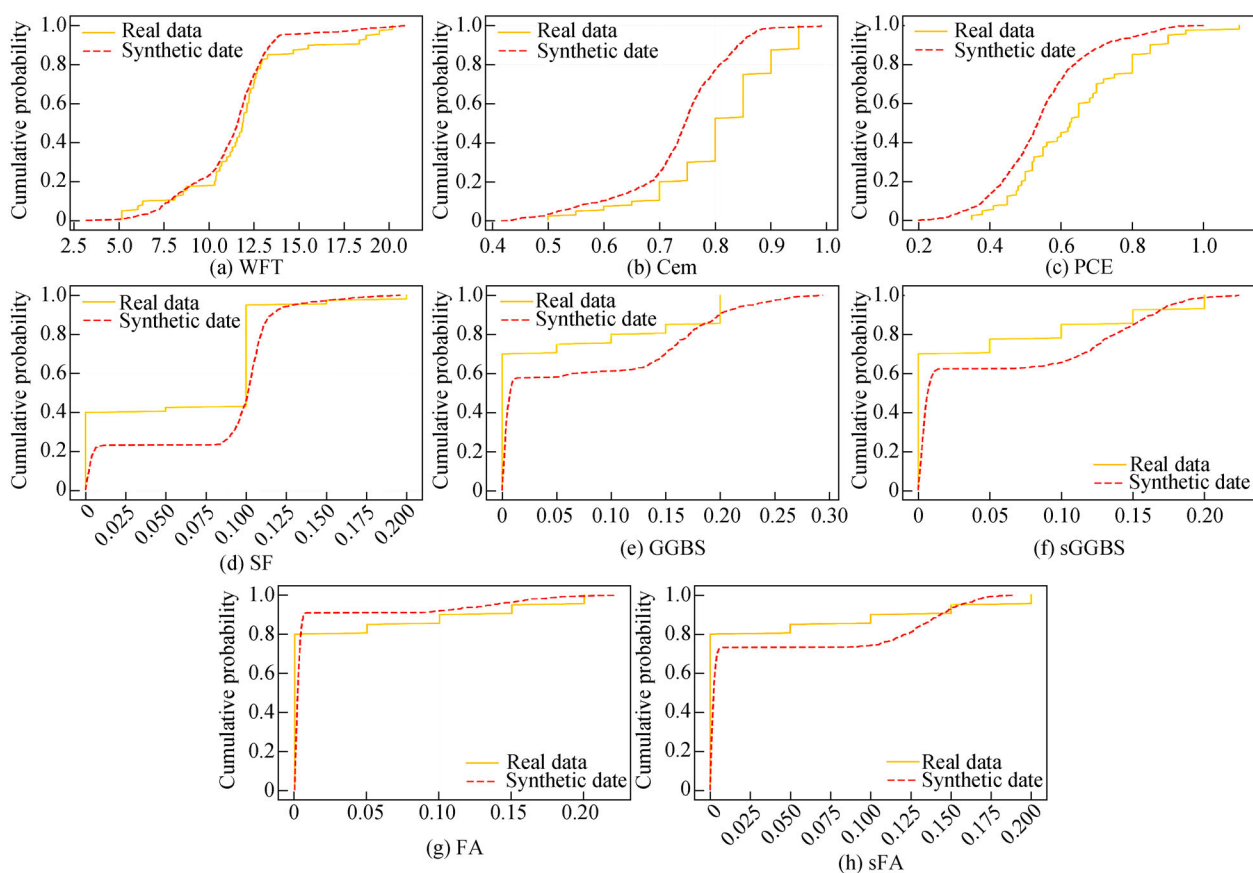


图 3 数据增强后以水膜层厚度为输出的特征变量累积分布

Fig. 3 Cumulative distribution of feature variables in real and synthetic datasets for WFT as output after data augmentation

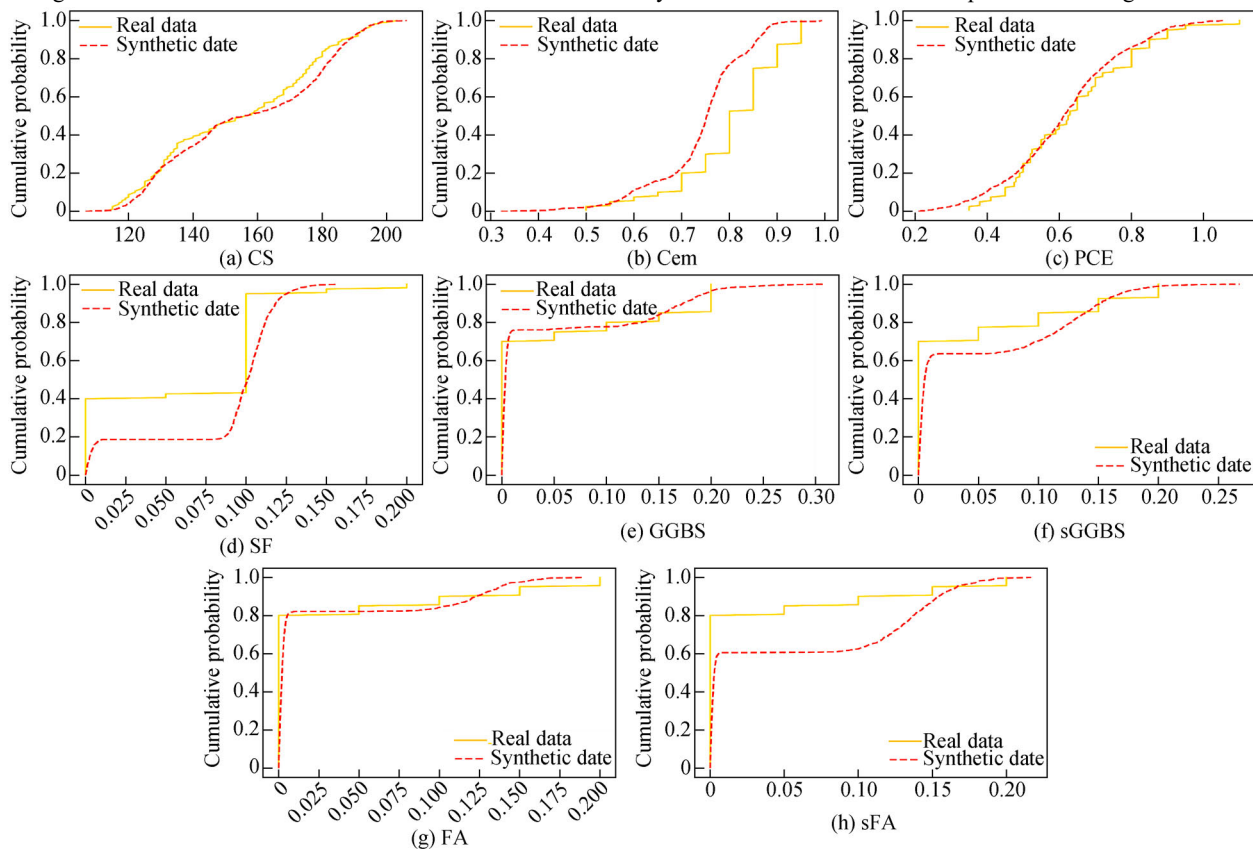


图 4 数据增强后以抗压强度为输出的特征变量累积分布

Fig. 4 Cumulative distribution of feature variables after data augmentation for compressive strength as output after data augmentation

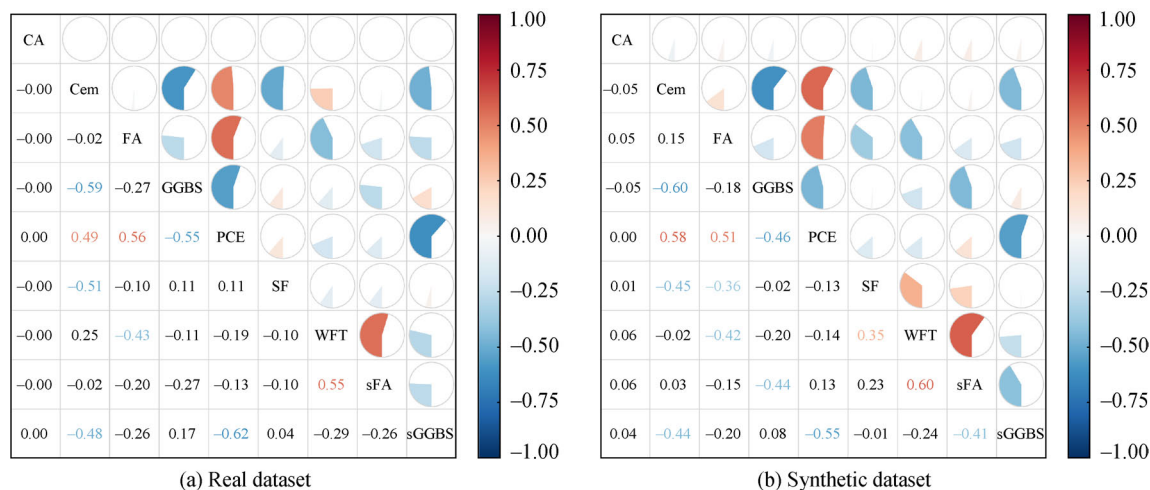


图5 以水膜层厚度为输出的特征变量相关性矩阵

Fig. 5 Correlation matrices of feature variables for WFT as output

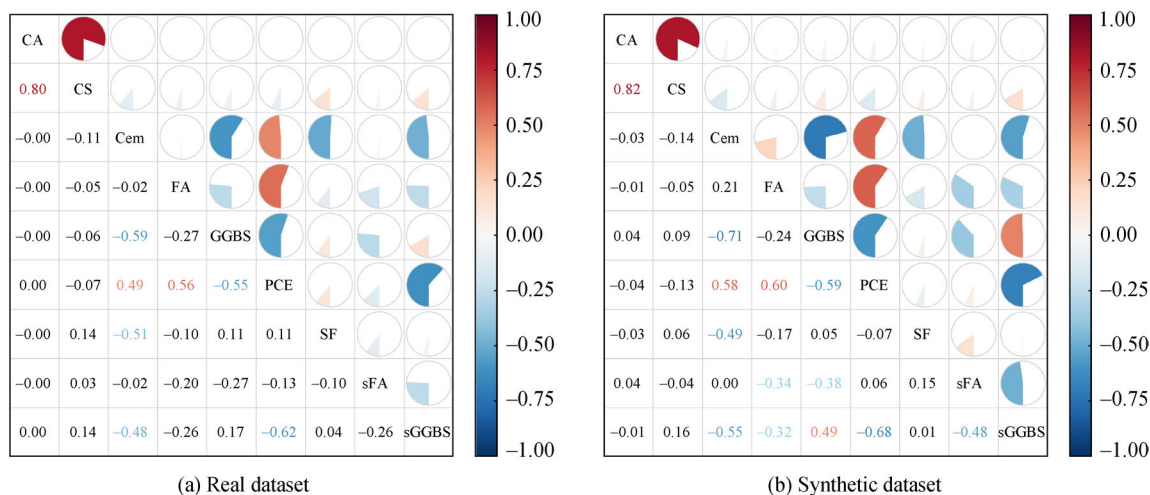


图6 以抗压强度为输出的特征变量相关性矩阵

Fig. 6 Correlation matrices of feature variables for compressive strength as output

2.2 CTGAN-BO-XGB 模型建立与训练

基于实验获得的 UHPC 原始数据集, 本工作采用 CTGAN 引入条件生成机制进行数据增强。通过同时引入随机噪声和条件信息, 生成器能够学习原始数据中不同变量的分布特征及其关联关系, 从而生成与真实样本统计特征相近的合成数据。该过程有助于扩展小样本数据规模, 丰富不同配比组合下的数据分布, 提高后续机器学习模型的训练充分性和泛化能力。

在完成数据集质量分析的基础上, 本工作首先对真实数据与增强数据进行了统一的划分与预处理。全部 160 条真实实验数据构成独立测试集, 用于最终评价模型的泛化性能; 基于 CTGAN 生成的合成数据集则用于模型的训练。连续型变量在输入模型前统一进行归一化或标准化处理, 分类变量的编码方式与 CTGAN 训练阶段保持一致, 从而避免

因特征表示差异导致的数据分布偏移。

在数据增强阶段, 本工作利用 CTGAN 对原始 160 条数据进行建模, 通过生成器与判别器的对抗训练学习输入变量与输出变量(包括水膜层厚度和抗压强度)的联合分布。训练过程中采用 Adam 优化器, 并对学习率、潜在空间维度、隐藏层结构、条件向量维度以及批量大小等关键超参数进行网格搜索与经验调优^[25]。不同超参数组合下生成数据的质量通过概率分布对比、相关系数矩阵等多维度进行评估, 优先选择在分布相似度和相关性保持方面表现更优且训练过程稳定的参数配置^[18]。最终确定的 CTGAN 超参数组合如表 4 所示, 对应的结果表明合成数据在捕捉原始数据复杂分布关系方面具有较高可靠性, 为后续模型训练提供了充足且质量可控的数据基础。

在性能预测阶段, 本工作以 XGBoost 作为核心

回归模型, 对 CTGAN 生成的训练集进行拟合, 并利用验证集进行模型选择和调优。XGBoost 模型训练中综合考虑了最大树深、基学习器数量、学习率、子采样比例、列采样比例以及最小叶节点权重等关键参数, 以验证集上 RMSE 最小为主要优化目标, 并结合 R^2 、MAE 和 MAPE 等指标对模型性能进行综合评价。为在有限计算成本下高效搜索高维超参数空间, 在此基础上引入贝叶斯优化构建 BO-XGBoost 模型。该方法通过高斯过程回归对超参数与模型性能之间的映射关系进行建模, 并基于期望改进(Expected improvement)等采集函数迭代选择新的候选参数组合。随着迭代次数的增加, 验证集 RMSE 逐步降低并趋于稳定, 表明贝叶斯优化能够有效锁定性能更优的参数区域。最终得到的最优超参数组合作为 BO-XGBoost 模型的最终配置, 其搜索范围与最终取值如表 5 所示, 超参数调优过程如图 7 所示。

表 4 CTGAN 的超参数
Table 4 Hyperparameters of the CTGAN model

Hyperparameters	Optimization ranges	Optimal value
Embedding_dim	[64, 128]	128
Generator_dim	[(128, 128), (256, 256), (512, 512)]	(256, 256)
Discriminator_dim	[(128, 128), (256, 256), (512, 512)]	(512, 512)
Generator_lr	$[2 \times 10^{-5}, 2 \times 10^{-4}, 2 \times 10^{-3}]$	2×10^{-4}
Generator_decay	$[1 \times 10^{-6}, 1 \times 10^{-5}]$	1×10^{-5}
Discriminator_lr	$[2 \times 10^{-5}, 2 \times 10^{-4}, 2 \times 10^{-3}]$	2×10^{-4}
Discriminator_decay	$[1 \times 10^{-6}, 1 \times 10^{-5}]$	1×10^{-5}
Batch size	[100, 800]	500
Epochs	[1000, 15 000]	12 000

表 5 BO-XGBoost 模型的超参数
Table 5 Hyperparameters of the BO-XGBoost model

Hyperparameters	Optimization ranges	Optimal value
Max_depth	3–10	6
Learning_rate	0.01–0.20	0.10
N_estimators	50–1000	500
Min_child_weight	1–10	7
Subsample	0.5–1.0	0.8
Colsample_bytree	0.1–1.0	0.8
Reg_lambda	0.01–1.00	0.40

2.3 建模结果分析

图 8 和图 9 展示了在采用 CTGAN 数据增强条

件下, BO-XGBoost 模型对水膜层厚度和抗压强度的预测性能及其泛化能力。结果显示, 本工作所构建的模型在小样本条件下依然可以保持良好的拟合精度与预测稳定性。从散点拟合结果来看, 无论是水膜层厚度(图 8a 和图 8d)还是抗压强度(图 9a 和图 9d), 预测值整体分布于 45°理想线附近, 大多数样本位于 $\pm 10\%$ 的误差线范围内, 其测试集 R^2 分别为 0.986 和 0.976, 均优于训练集的 0.882 和 0.957。从误差指标上看, 水膜层厚度在训练集与测试集上的 RMSE 分别为 0.909 3 nm 和 0.387 6 nm, 抗压强度的 RMSE 分别为 5.244 8 MPa 和 3.837 9 MPa, 整体误差水平较低, 进一步表明模型具有较好的预测精度。误差分布图(图 8b、图 8e、图 9b、图 9e)进一步表明, 预测误差呈近似对称分布, 且主要集中在较小范围内, 说明模型不存在明显的系统性偏差。此外, 箱线图结果(图 8c、图 9c)表明训练集与测试集的残差分布具有较好的一致性, 反映出模型具有稳定的泛化能力。损失函数随迭代次数的变化曲线(图 8f、图 9f)显示模型在训练过程中逐步收敛, 未出现明显过拟合现象, 说明模型训练过程相对稳定。上述分析表明, CTGAN 数据增强在一定程度上改善了小样本条件下数据分布不足的问题, 使模型能够更充分地学习不同配比条件下的非线性关系, 从而提升预测精度与稳定性。总体而言, BO-XGBoost 模型能够较好地描述配比参数与水膜层厚度及抗压强度之间的复杂关系, 并在有限样本条件下表现出良好的泛化性能。

2.4 特征可解释性分析

为进一步揭示 BO-XGBoost 模型在 UHPC 的性能预测中的决策机理, 并从材料组成角度解释模型输出结果的物理意义, 本工作结合 SHAP 和 PDPs 方法, 对关键特征变量的影响机制进行分析, 结果如图 10 和图 11 所示。

从图 10a 中的 SHAP 全局特征重要性排序及摘要图可以看出, 掺合料的种类对水膜层厚度的影响存在明显差异。其中, 粉煤灰微珠的贡献度相对较高, 整体呈现显著的正向驱动作用, 表明在一定掺量范围内, 这类细颗粒材料能够通过填充效应及颗粒分散作用增加体系中水膜层厚度^[26]。相比之下, 粉煤灰与矿渣粉的 SHAP 值在正负区间均有分布, 表现出较强的非线性和条件依赖特征。单特征 SHAP 分析再次表明了各掺合料对水膜层厚度的影响规律具有显著差异。对于水泥(图 10b), 其 SHAP 值随着水泥掺量增加而整体呈上升趋势, 尤其在在

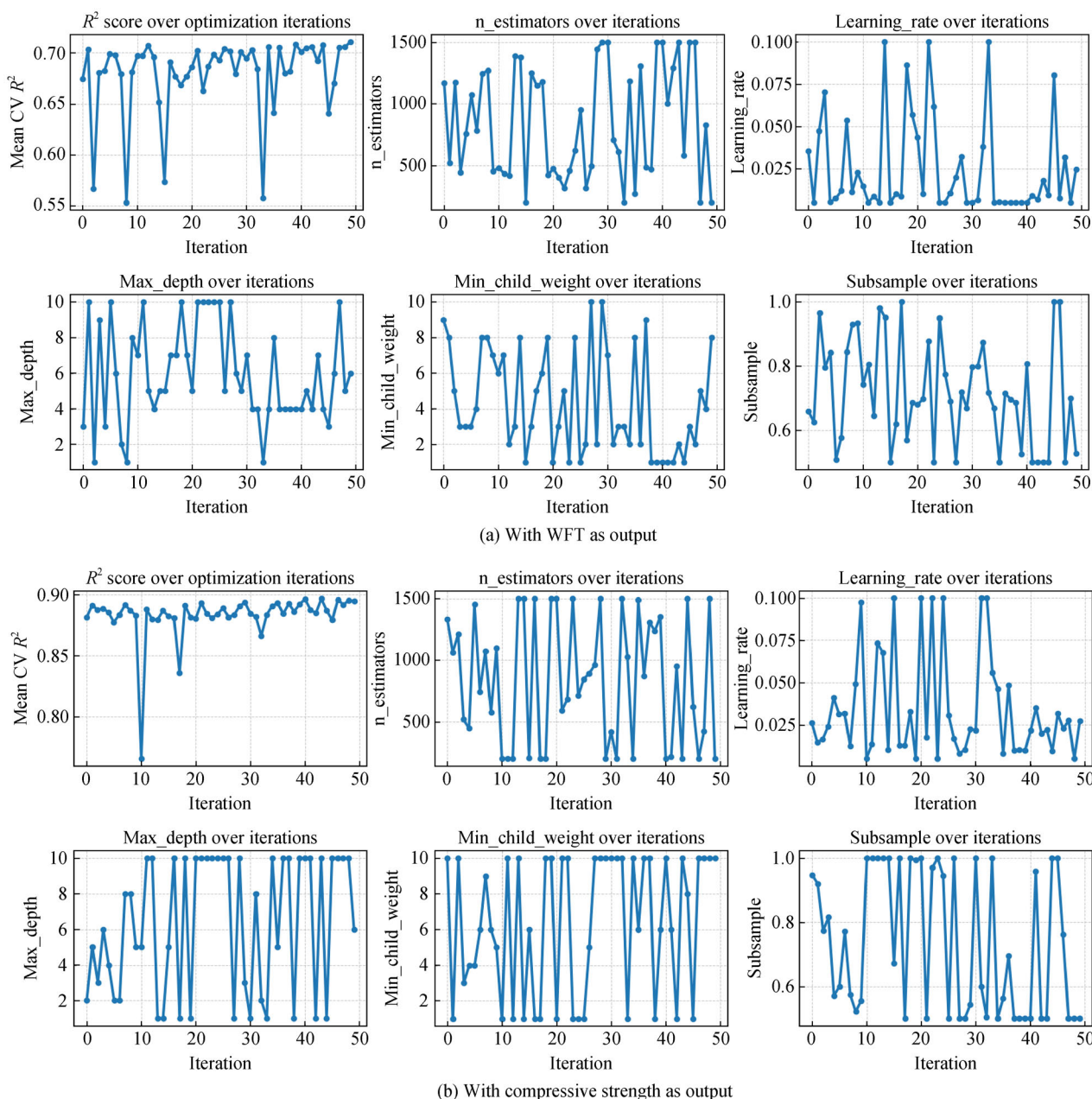


图7 BO-XGBoost 模型超参数调优过程

Fig. 7 Hyperparameter optimization process of the BO-XGBoost model

取值区间,水膜层厚度表现出明显的正向响应。这可能是因为,水泥替代部分高比表面积掺合料后,体系总比表面积降低、颗粒堆积状态发生改变有关^[27]。图 10c 显示,随着 PCE 掺量的增加,其在水膜层厚度的影响表现为先增后减的变化趋势。在中等掺量范围内,PCE 对水膜层厚度具有较好的正向促进作用,但过高掺量导致性能下降。这可能是因为过多的 PCE 导致体系分散过度,水膜连续性被削弱^[28]。对于矿渣粉与超细矿渣粉(图 10e 和图 10f),其 SHAP 值随掺量增加整体呈下降趋势,在高掺量

区间表现出明显的负向贡献。这可能是由于矿渣类材料掺量过高会提高颗粒堆积密实度,使得有效水膜空间受到压缩^[29]。从图 10g 和图 10h 可以看出,粉煤灰与粉煤灰微珠呈现出显著的掺量依赖特征。随着掺量增加,粉煤灰的 SHAP 快速转为负值,粉煤灰微珠的 SHAP 则快速转为正值,表明更细小的球形颗粒可以显著增加水膜层厚度^[30]。总体来看,各掺合料对水膜层厚度的影响不仅与掺量相关,还受到粒径分布、比表面积及颗粒形貌等因素的共同影响,体现出多因素耦合作用下的非线性调控特征。

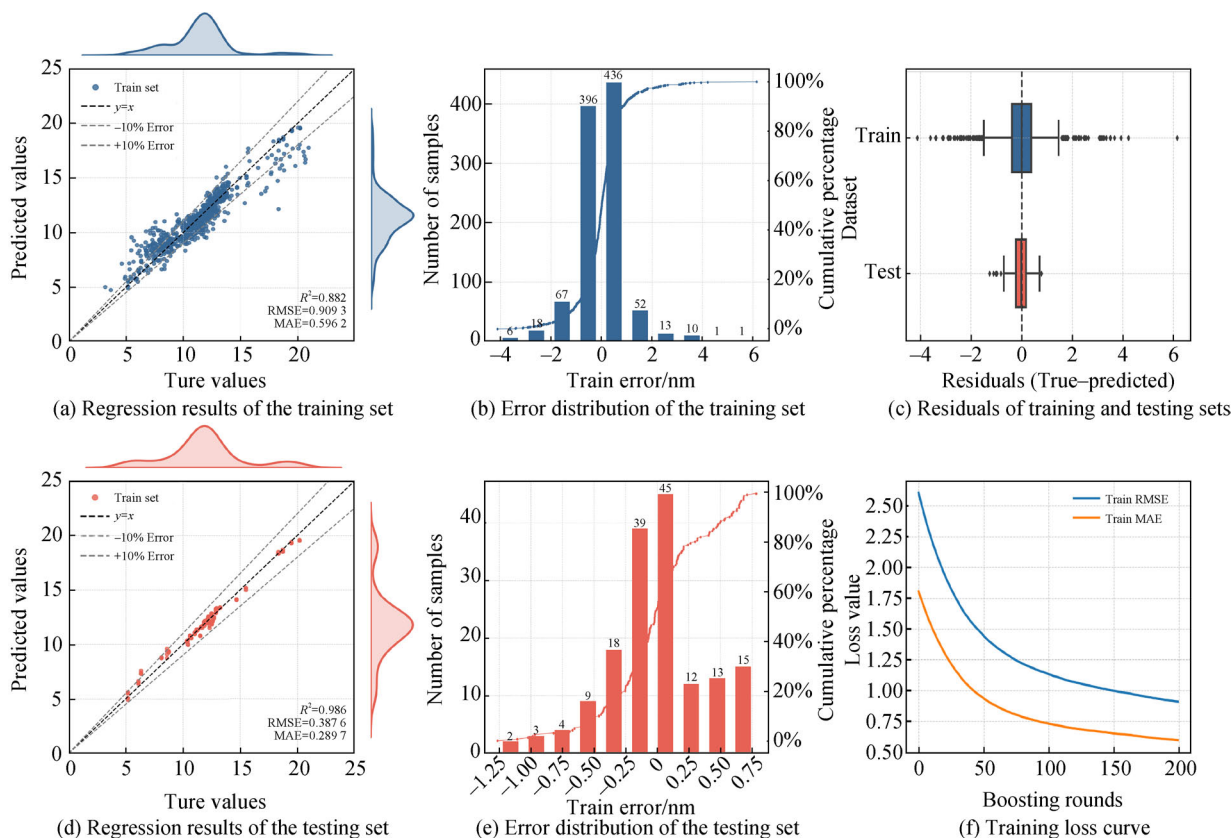


图 8 以水膜层厚度为输出的模型预测结果及评价指标

Fig. 8 Model prediction results and evaluation metrics for WFT as output

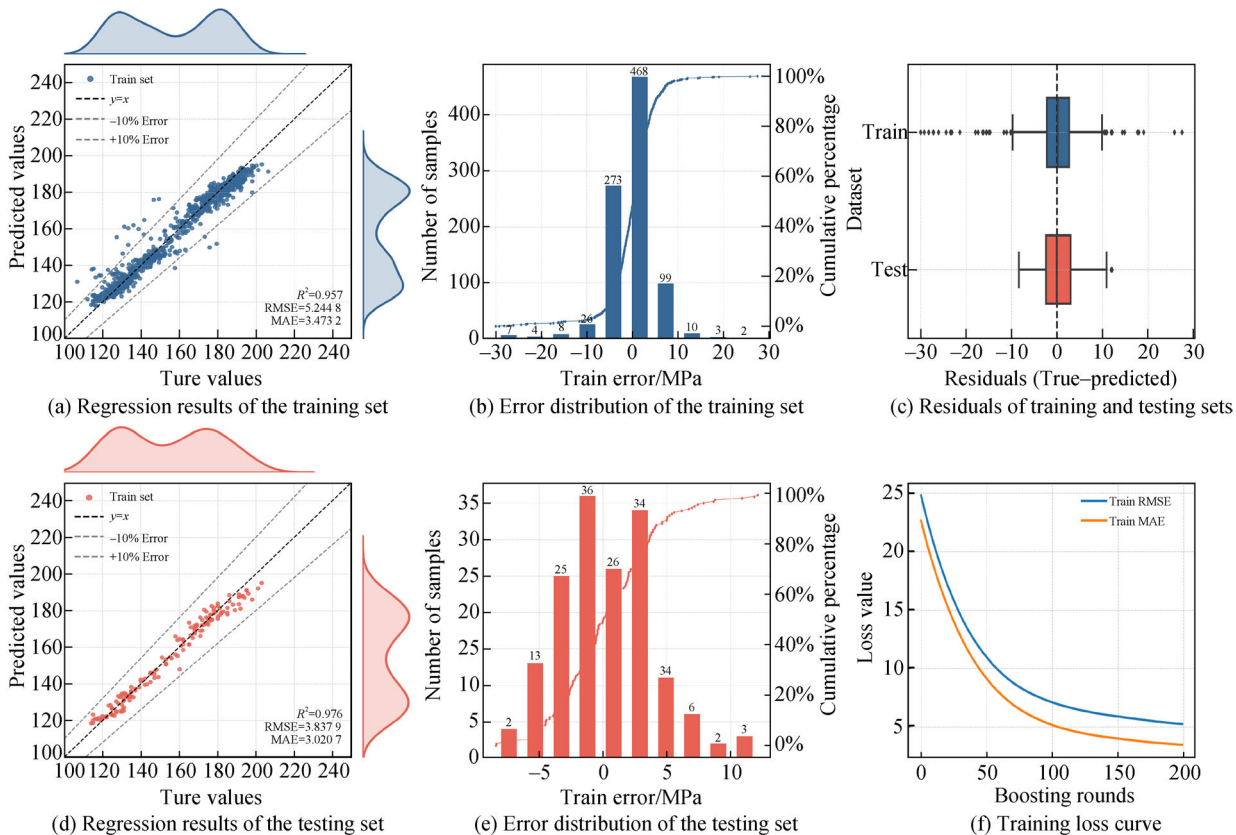


图 9 以抗压强度为输出的模型预测结果及评价指标

Fig. 9 Model prediction results and evaluation metrics for compressive strength as output

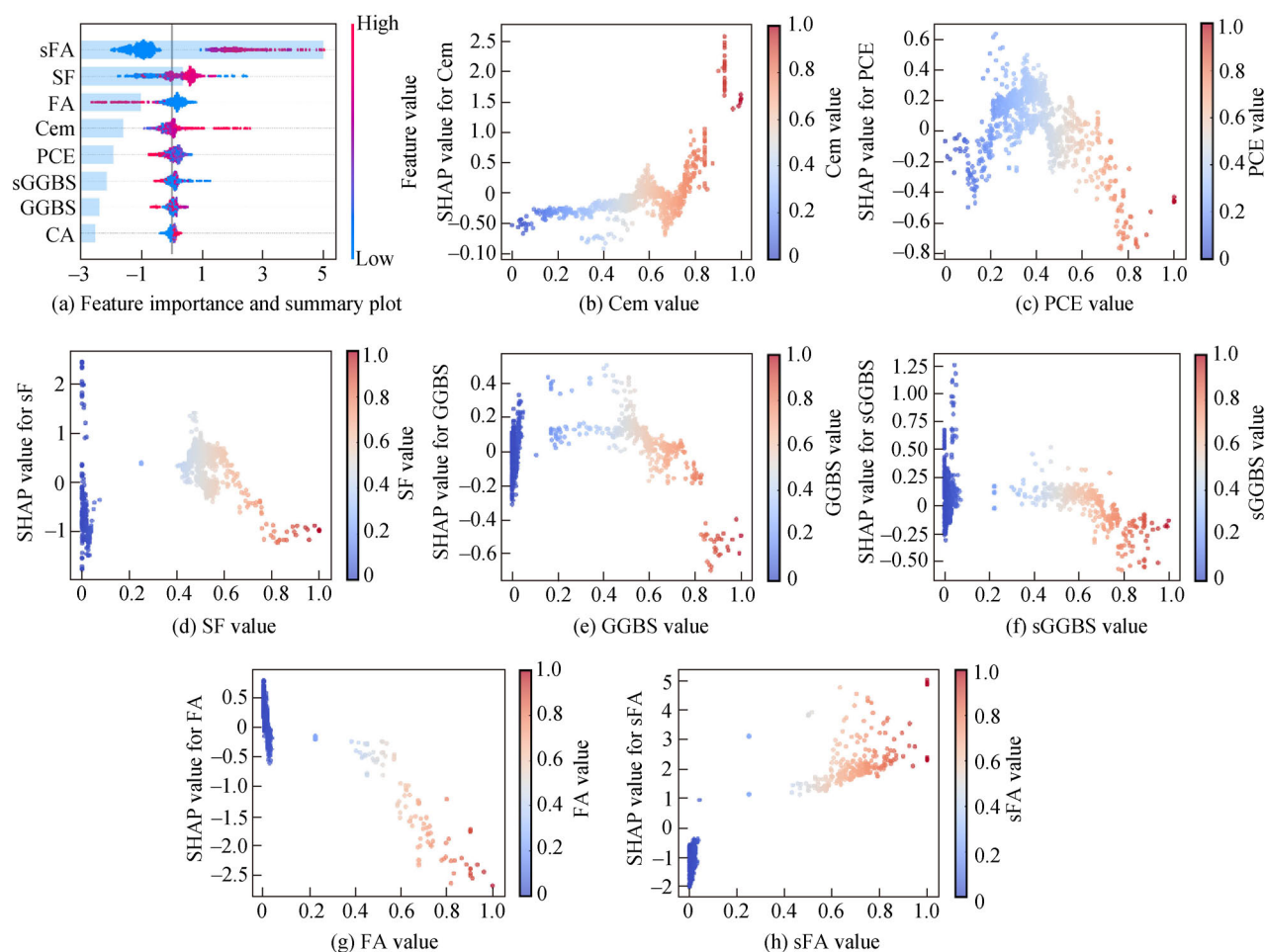


图 10 以 WFT 为输出的特征变量 SHAP 分析

Fig. 10 SHAP analysis of features variables for WFT as output

从图 11a 的 SHAP 全局特征重要性排序及摘要图可以看出, 养护龄期(CA)对抗压强度的贡献度最高, 整体表现为显著的正向作用。这与 UHPC 水化反应随时间推进、水化产物持续生成并填充孔隙的基本规律一致。除龄期外, 超细矿渣粉、硅灰和粉煤灰微珠均表现出正向贡献, 而矿渣粉和粉煤灰则呈现出负向影响。进一步的单特征 SHAP 分析表明, 各变量对抗压强度的影响规律具有显著差异。从图 11b 可以看出, 水泥的大比例掺入会对抗压强度产生不利影响, 这是由于 UHPC 需要考虑体系中的紧密堆积, 水泥用量过多会导致体系中缺乏细颗粒填充空隙^[31]。图 11c 显示, PCE 总体上表现为对强度发展有促进作用。而图 11d 则表明, 硅灰的对强度的发展有明显的掺量依赖性, 在较低的掺量区间, 硅灰对强度的贡献较弱甚至略呈负值, 而随着掺量的增加, 其迅速转为正向作用, 表明其高比表面积与微填充效应在较高掺量条件下更有利于强度提升^[32]。相比之下, 在低掺量区间, 矿渣粉的 SHAP 值(图 11e)偏正, 表明适量矿渣粉能够在一定程度上参与

或激发胶凝体系反应, 对强度发展起到补偿或促进作用^[33]。然而, 随着矿渣粉掺量的进一步提高, 其 SHAP 值迅速转为明显负值, 说明高掺量矿渣粉会对抗压强度产生显著不利影响。这一趋势反映了矿渣粉潜在活性依赖于水泥水化产物的激发, 在高掺量条件下, 由于熟料含量降低, 体系中 $\text{Ca}(\text{OH})_2$ 等激发源不足, 限制了矿渣粉的二次反应, 从而抑制强度发展^[34]。图 11f 可以看出, 超细矿渣粉的 SHAP 值随掺量增加由负逐步转为正, 在中高掺量区间, 大多数样本的 SHAP 值集中于正值区域, 说明在本工作范围内, 超细矿渣粉的掺入对体系强度的发展是有益的。这主要归因于其同时具有微填充效应及火山灰反应, 促进抗压强度的发展。图 11g 表现为图 11e 相似的变化趋势, 即在中高掺量条件下对强度产生负向影响, 但其作用程度相对较缓, 这可能与粉煤灰具有一定的填充效应有关。与之不同的是, 粉煤灰微珠的 SHAP 值显示(图 11h), 其对抗压强度整体呈现正向贡献, 说明粒径更细、分布更均匀的粉煤灰颗粒更有利于体系结构优化, 从而促进抗压强度的发展。

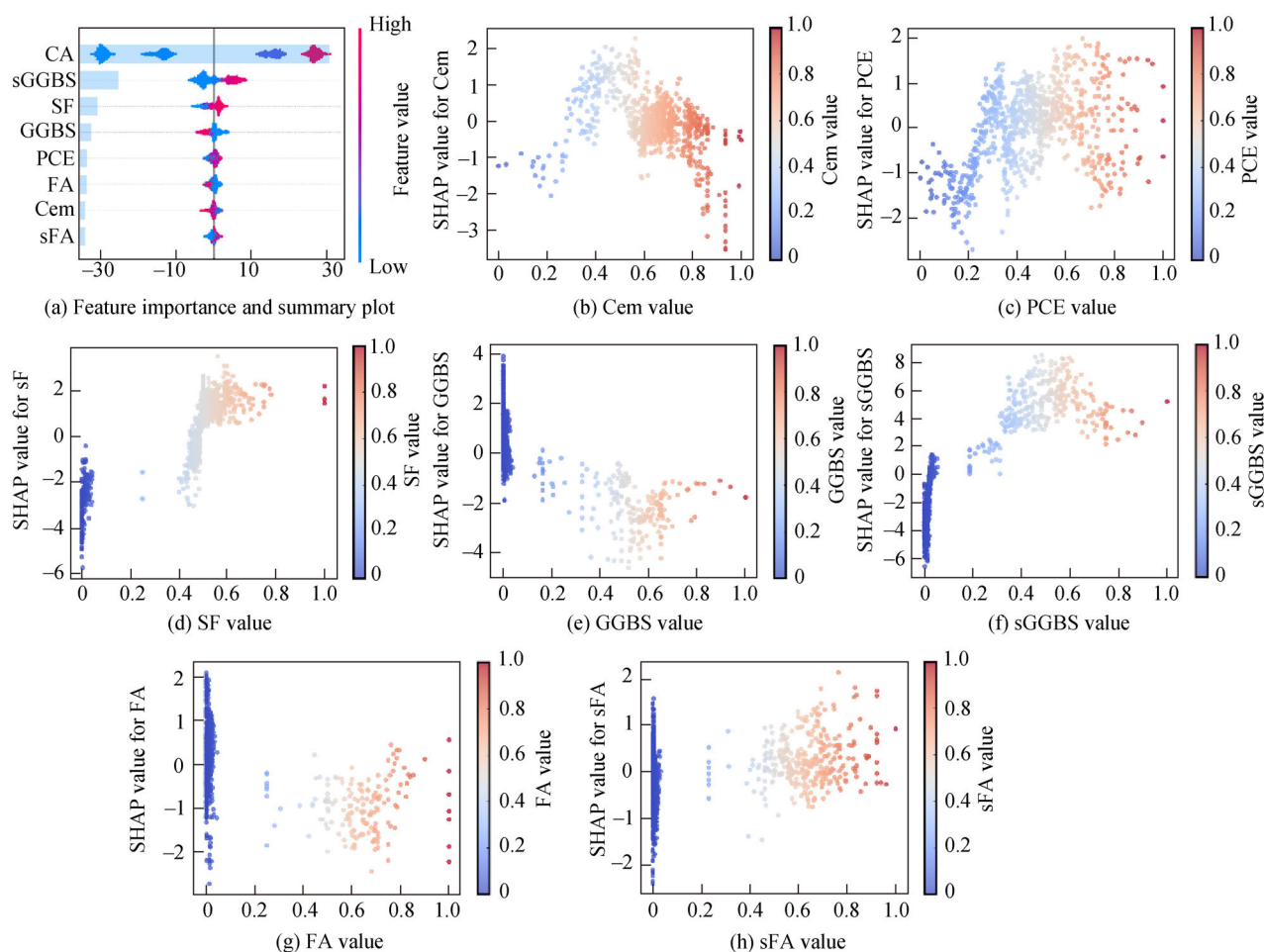


图 11 以抗压强度为输出的特征变量 SHAP 分析

Fig. 11 SHAP analysis of features variables for compressive strength as output

图 12 和图 13 展示了关键输入变量的部分依赖图(PDPs)分析结果。图 12 结果显示,随着水泥掺量的增加,水膜层厚度整体呈上升趋势,在低掺量区间增长较为平缓,而在中高掺量区间上升幅度明显增大。该趋势与图 10 中水泥的 SHAP 分析结果一致,即随着水泥掺量增加,其对水膜层厚度的正向作用逐渐增强,在较高取值区间,水膜层厚度表现出明显的正向响应。相比之下,PCE、硅灰在低到中等掺量范围内对水膜层厚度表现出逐渐增强的正向作用,而在高掺量条件下,其促进作用有所减弱甚至出现负向影响,表明过量掺入可能导致体系分散过度,从而不利于水膜结构的稳定形成^[35]。矿渣粉和超细矿渣粉在极低掺量范围内对水膜层厚度的发展具有明显的促进作用,在低掺量到中等掺量的大范围其增益逐渐减弱并趋于饱和,在高掺量条件下转为明显负向作用,这与其提高颗粒堆积密实度、压缩有效水膜空间的作用机制有关^[10]。粉煤灰与粉煤灰微珠则呈现出显著差异。随着掺量增加,粉煤灰对水膜层厚度的负向影响逐渐增强,而粉煤灰微

珠则表现出持续增强的正向贡献。这是因为粉煤灰微珠颗粒粒径更细、球形度更高,能够在胶凝材料颗粒间发挥更明显的填充和滚珠效应,有利于改善颗粒分散状态并提高有效水膜厚度;而粉煤灰颗粒级配及活性相对较弱,会削弱颗粒堆积协调性,从而表现出负向影响。

图 13 结果显示,随着水泥掺量的增加,抗压强度整体呈下降趋势,这可能与高水泥用量下体系颗粒堆积结构及水化反应平衡的变化有关。在本实验范围内,PCE 掺量的增加对强度具有促进作用,表明适当的分散效应有利于改善体系结构,从而提高力学性能。硅灰的影响表现出明显的掺量依赖性,在中低掺量区间,其对强度的提升作用较为有限,而在较高掺量下,其正向作用显著增强,说明较高掺量的硅灰能够通过微填充效应和结构致密化作用有效提升抗压强度^[36]。矿渣粉与超细矿渣粉则呈现出相反的变化趋势,随着掺量增加,矿渣粉对强度表现为逐渐抑制作用,而超细矿渣粉则持续表现出促进作用,这反映出超细颗粒在填充效应和潜在活

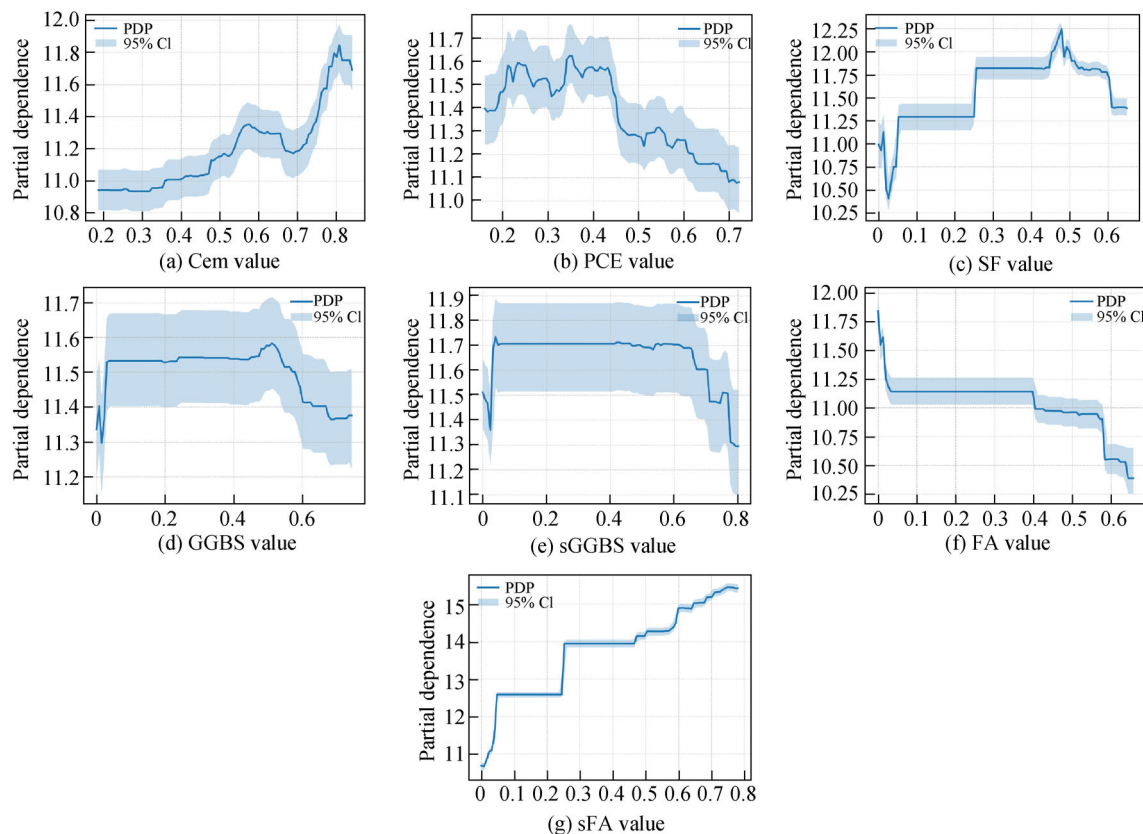


图 12 以水膜层厚度为输出的特征变量 PDPs 分析

Fig. 12 PDPs analysis of features variables for WFT as output

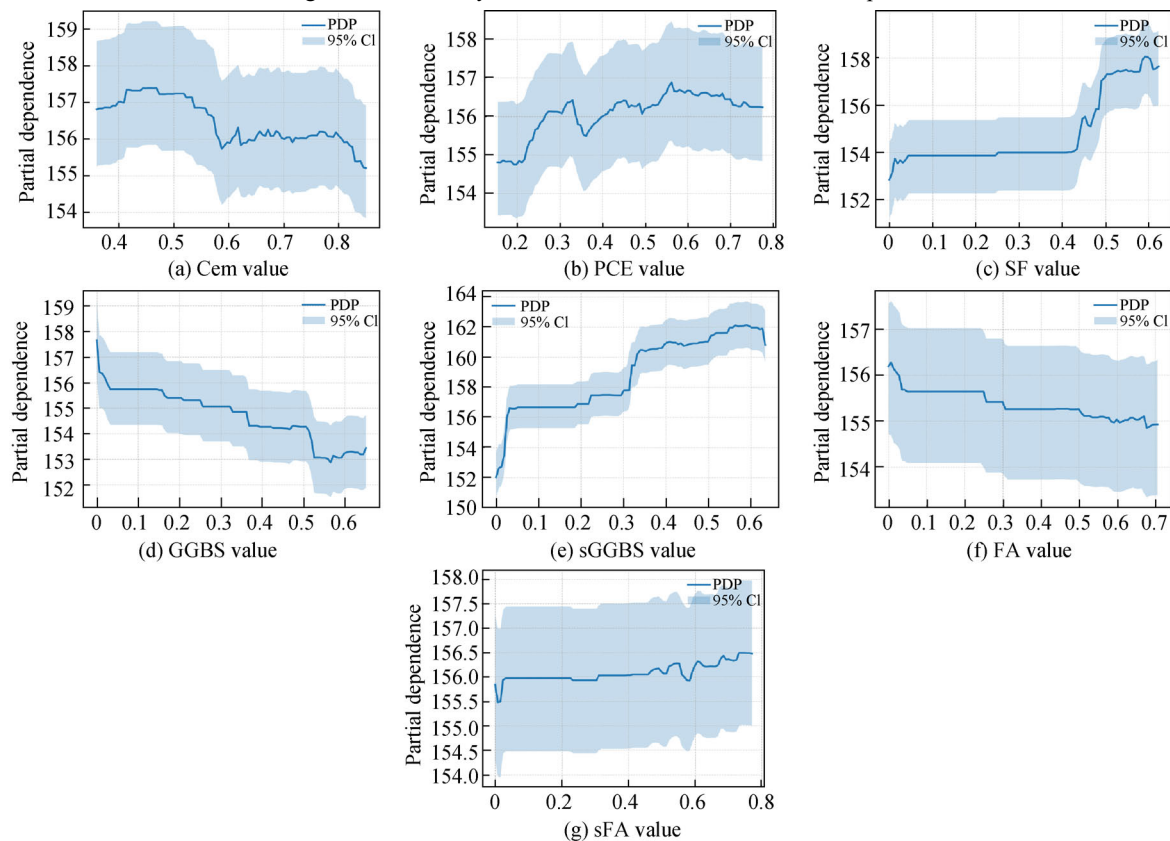


图 13 以抗压强度为输出的特征变量 PDPs 分析

Fig. 13 PDPs analysis of features variables for compressive strength as output

性方面的优势。粉煤灰的变化趋势与矿渣粉相似,即随着掺量增加对强度产生负向影响,但其下降幅度相对较缓,推测与其一定的填充作用有关。相比之下,粉煤灰微珠整体对抗压强度表现为正向贡献,但其随掺量增加的提升幅度较为平缓,说明其在改善体系结构方面具有一定效果,但增强作用相对有限。

3 结论

本工作基于 UHPC 小样本试验数据,构建了融合 CTGAN 数据增强、贝叶斯优化 XGBoost 及可解释性机器学习方法的多响应预测与分析框架,并对不同矿物掺合料对水膜层厚度及抗压强度的影响规律进行了系统分析,主要结论如下:

1) CTGAN 模型能够有效学习原始数据的分布特征及变量之间的相关关系,所生成的合成数据在分布形态和相关结构上与真实数据具有较好一致性,从而扩展了小样本数据规模,为后续模型训练提供了更为充分的数据支撑。

2) 基于 CTGAN 增强数据集构建的 BO-XGBoost 模型在水膜层厚度和抗压强度预测中均表现出较好的精度水平。测试集 R^2 分别达到 0.986 和 0.976,误差分布集中且对称。总体上看,该模型能够较好地刻画多组分 UHPC 体系中配比参数与性能之间的复杂非线性关系。

3) SHAP 分析表明,不同矿物掺合料对水膜层厚度与抗压强度的影响存在显著差异。对水膜层厚度而言,粉煤灰微珠整体表现为正向贡献,硅灰在中低掺量为正向贡献,而硅灰、超细矿渣粉、矿渣粉与粉煤灰在高掺量下呈现出负向影响。这主要与 UHPC 体系内的颗粒堆积状态有关。对抗压强度而言,超细矿渣粉、硅灰和粉煤灰因微填充效应及火山灰反应,整体表现为正向贡献,而矿渣粉与粉煤灰在高掺量条件下因反应活性受限表现出负向影响。

4) PDPs 分析表明,水膜层厚度主要受颗粒特征与堆积状态调控,而抗压强度则更多与材料活性及结构致密化过程密切相关,其分别从微观润滑状态与宏观力学性能两个层面反映了配比参数对 UHPC 体系结构演化的调控作用。

5) 总体而言,所构建的 CTGAN-BO-XGBoost 预测框架能够在小样本条件下实现对 UHPC 关键性能的高精度预测与机制解释,为多组分胶凝材料体系的配合比设计与性能优化提供了一种有效的数据驱动方法。

(第二十八届中国科协年会学术论文)

参考文献:

- [1] 刘一帆, 吴泽媚, 张轩翰, 等. 超高性能混凝土流变特性及调控研究进展[J]. 硅酸盐学报, 2023, 51(11): 3025–3038.
LIU Yifan, WU Zemei, ZHANG Xuanhan, et al. J Chin Ceram Soc, 2023, 51(11): 3025–3038.
- [2] 樊健生, 丁然. 超高性能混凝土在房屋建筑结构中的研究与应用进展[J]. 硅酸盐学报, 2023, 51(5): 1246–1258.
FAN Jiansheng, DING Ran. J Chin Ceram Soc, 2023, 51(5): 1246–1258.
- [3] FAN D Q, ZHU J Y, FAN M X, et al. Intelligent design and manufacturing of ultra-high performance concrete (UHPC)—A review[J]. Constr Build Mater, 2023, 385: 131495.
- [4] 张云升, 张文华, 陈振宇. 综论超高性能混凝土: 设计制备·微观结构·力学与耐久性·工程应用[J]. 材料导报, 2017, 31(23): 1–16.
ZHANG Yunsheng, ZHANG Wenhua, CHEN Zhenyu. Mater Rev, 2017, 31(23): 1–16.
- [5] SCHRÖFL C, GRUBER M, PLANK J. Preferential adsorption of polycarboxylate superplasticizers on cement and silica fume in ultra-high performance concrete (UHPC)[J]. Cem Concr Res, 2012, 42(11): 1401–1408.
- [6] KHAYAT K H, MENG W N, VALLURUPALLI K, et al. Rheological properties of ultra-high-performance concrete: An overview[J]. Cem Concr Res, 2019, 124: 105828.
- [7] 余睿, 陈飞翔, 范定强, 等. 基于机器学习技术的超高性能混凝土性能预测与精细化设计[J]. 硅酸盐学报, 2025, 53(5): 1165–1176.
YU Rui, CHEN Feixiang, FAN Dingqiang, et al. J Chin Ceram Soc, 2025, 53(5): 1165–1176.
- [8] LIU H R, SUN X, DU H, et al. Effects and threshold of water film thickness on multi-mineral cement paste[J]. Cem Concr Compos, 2020, 112: 103677.
- [9] 王圣杰, 李兵, 李传习, 等. 超高性能混凝土流动性与流变性关系[J]. 硅酸盐学报, 2023, 51(8): 1962–1970.
WANG Shengjie, LI Bing, LI Chuanxi, et al. J Chin Ceram Soc, 2023, 51(8): 1962–1970.
- [10] YE H, GAO X J, WANG R, et al. Relationship among particle characteristic, water film thickness and flowability of fresh paste containing different mineral admixtures[J]. Constr Build Mater, 2017, 153: 193–201.
- [11] FAN D Q, YU R, FU S Y, et al. Precise design and characteristics prediction of Ultra-High Performance Concrete (UHPC) based on artificial intelligence techniques[J]. Cem Concr Compos, 2021, 122: 104171.
- [12] AYLAS-PAREDES B K, HAN T H, NEITHALATH A, et al. Data driven design of ultra high performance concrete prospects and application[J]. Sci Rep, 2025, 15: 9248.
- [13] HE Y Z, GAO S J, LI Y, et al. Adaptive machine learning framework: Predicting UHPC performance from data to modelling[J]. Results Eng, 2025, 27: 106724.
- [14] 周祥胥, 段锋, 朱博. 基于机器学习的 SHAP 和 PDP 分析对 UHPC 流变性能的研究[J]. 材料导报, 2026, 40(2): 86–93.
ZHOU Xiangxu, DUAN Feng, ZHU Bo. Mater Rep, 2026, 40(2): 86–93.
- [15] 余睿, 李旺, 范定强. 基于条件生成对抗网络与马尔可夫链蒙特卡罗驱动的超高性能混凝土数据增强及智能预测方法[J]. 硅酸盐学报, 2026, 54(3): 842–856.

- YU Rui, LI Wang, FAN Dingqiang. J Chin Ceram Soc, 2026, 54(3): 842–856.
- [16] 孙浚博, 王雨飞, 赵宏宇, 等. 基于深度学习的 3D 打印混凝土蒸汽养护力学性能研究和抗压强度预测[J]. 硅酸盐通报, 2024, 43(5): 1723–1738.
- SUN Junbo, WANG Yufei, ZHAO Hongyu, et al. Bull Chin Ceram Soc, 2024, 43(5): 1723–1738.
- [17] WANG Y F, SUN J B, WANG X Y, et al. Multi-objective optimization of engineered cementitious composite based on machine learning and generative adversarial network[J]. J Build Eng, 2024, 96: 110471.
- [18] WANG S Q, XIA P, GONG F Y, et al. A Bayesian-physical informed conditional tabular generative adversarial network framework for low-carbon concrete data augmentation and hyperparameter optimization[J]. Eng Appl Artif Intell, 2025, 152: 110811.
- [19] HUU NGUYEN M, NGUYEN T A, LY H B. Ensemble XGBoost schemes for improved compressive strength prediction of UHPC[J]. Structures, 2023, 57: 105062.
- [20] HARUNA S I, IBRAHIM Y E, ABBA S I. Machine learning algorithm for predicting the compressive strength of sustainable ultra-high-performance concrete (UHPC) containing supplementary cementitious materials (SCMs)[J]. J Asian Archit Build Eng, 2025: 1–17.
- [21] KASHEM A, KARIM R, MALO S C, et al. Hybrid data-driven approaches to predicting the compressive strength of ultra-high-performance concrete using SHAP and PDP analyses[J]. Case Stud Constr Mater, 2024, 20: e02991.
- [22] 郭孟环, 谢咏菲, 周英武, 等. 基于多目标靶向性能的超高韧性地聚物复合材料智能化设计[J]. 硅酸盐学报, 2026, 54(3): 823–841.
- GUO Menghuan, XIE Yongfei, ZHOU Yingwu, et al. J Chin Ceram Soc, 2026, 54(3): 823–841.
- [23] NGUYEN N H, ABELLÁN-GARCÍA J, LEE S, et al. Efficient estimating compressive strength of ultra-high performance concrete using XGBoost model[J]. J Build Eng, 2022, 52: 104302.
- [24] JAISWAL R, BHATTA N, MUHIT I B. An interpretable machine learning approach for shear capacity prediction of prestressed UHPC bridge girders with GUI deployment[J]. Structures, 2026, 88: 111793.
- [25] XU W Y, ZHANG L Y, FAN D Q, et al. AI-infused characteristics prediction and multi-objective design of ultra-high performance concrete (UHPC): From pore structures to macro-performance[J]. J Build Eng, 2024, 98: 111170.
- [26] CHEN J J, KWAN A K H. Triple blending with fly ash microsphere and condensed silica fume to improve performance of cement paste[J]. J Mater Civ Eng, 2013, 25(5): 618–626.
- [27] MARCHETTI G, RAHHAL V F, IRASSAR E F. Influence of packing density and water film thickness on early-age properties of cement pastes with limestone filler and metakaolin[J]. Mater Struct, 2017, 50(2): 111.
- [28] LI L G, KWAN A K H. Effects of superplasticizer type on packing density, water film thickness and flowability of cementitious paste[J]. Constr Build Mater, 2015, 86: 113–119.
- [29] PHILIP S, MARAKKATH N. Compressive strength prediction and feature analysis for GGBS-Based geopolymers using optimized XGBoost and SHAP: A comparative study of optimization algorithms and experimental validation[J]. J Build Eng, 2025, 108: 112879.
- [30] KWAN A K H, LI Y. Effects of fly ash microsphere on rheology, adhesiveness and strength of mortar[J]. Constr Build Mater, 2013, 42: 137–145.
- [31] RANDL N, STEINER T, OFNER S, et al. Development of UHPC mixtures from an ecological point of view[J]. Constr Build Mater, 2014, 67: 373–378.
- [32] ZOU S, SUN Z P, LU Z C, et al. Fresh and hardened properties of cement paste under the synergistic influence of polycarboxylate superplasticizer and anionic polyacrylamide[J]. Case Stud Constr Mater, 2024, 21: e03895.
- [33] MOULA S, BEN FRAJ A, WATTEZ T, et al. The very early-age behaviour of Ultra-High Performance Concrete containing ground granulated blast furnace slag[J]. Constr Build Mater, 2023, 400: 132630.
- [34] MOULA S, BEN FRAJ A, WATTEZ T, et al. Mechanical properties, carbon footprint and cost of ultra-high performance concrete containing ground granulated blast furnace slag[J]. J Build Eng, 2023, 79: 107796.
- [35] HU Q H, ZHANG L, LUO Q, et al. Impact of polycarboxylate superplasticizer dosage on controlled low strength material flowability and bleeding: Insights from water film thickness[J]. Constr Build Mater, 2024, 447: 138145.
- [36] XI J Y, LIU J Z, YANG K, et al. Role of silica fume on hydration and strength development of ultra-high performance concrete[J]. Constr Build Mater, 2022, 338: 127600.

作者贡献声明:

张增起: 提出研究目标, 参与论文整体思路凝练与研究方案设计, 负责论文框架搭建与具体内容组织; 承担模型构建、参数设置与运行分析工作, 完成相关图表绘制与结果整理, 参与数据分析与结果讨论, 并负责论文初稿的撰写与修改

杜伟杰: 参与论文研究内容讨论与写作工作, 负责部分图表绘制与整理, 参与实验数据和模型结果的分析讨论, 对论文文本内容的完善与修改作出了贡献

刘 宇: 参与论文整体框架讨论与研究思路梳理, 对论文结构设置、内容表达和逻辑衔接提出修改意见; 参与论文修订与完善工作, 对稿件学术表达和整体质量提升作出了贡献

马善亮: 参与实验数据的整理与分析, 协助机器学习模型的调试与参数优化

刘晓明: 提出研究目标, 参与研究方案制定与论文框架设计, 对论文整体研究内容进行指导与把关; 参与论文重要内容的讨论与修改

荆 锐: 参与实验方案设计与数据分析工作, 对结果解释与论文内容完善提出建议

Performance Prediction and Interpretability Analysis of Ultra-High-Performance Concrete with Small Sample Size Based on Machine Learning

ZHANG Zengqi^{1,2}, DU Weijie², LIU Yu³, MA Shanliang², LIU Xiaoming^{1,2}, JING Rui⁴

(1. State Key Laboratory of Advanced Metallurgy, University of Science and Technology Beijing, Beijing 100083, China;

2. School of Metallurgical and Ecological Engineering, University of Science and Technology Beijing, Beijing 100083, China;

3. China Construction First Group Construction and Development Co., Ltd., Beijing 100102, China; 4. Tangshan Polytechnic University, Tangshan 063299, Hebei, China)

Extended Abstract

Introduction Ultra-high-performance concrete (UHPC) is widely applied in bridges, prefabricated structures, and advanced construction owing to its high strength, dense microstructure, and excellent durability performance. Its performance is strongly governed by the type and dosage of supplementary cementitious materials (SCMs), including silica fume (SF), ground granulated blast-furnace slag (GGBS), superfine GGBS (sGGBS), fly ash (FA), and fly ash microspheres (sFA), as well as polycarboxylate superplasticizer (PCE) dosage and curing age (CA). However, the relationships among mixture parameters, water distribution, hydration, and strength development are highly nonlinear, and conventional empirical models often fail to provide accurate predictions.

Water film thickness (WFT) is a key microscopic parameter reflecting the free water covering the surfaces of solid particles, directly influencing particle lubrication, dispersion, and fresh-state behaviour in high-solid-volume UHPC systems. Compressive strength (CS) is another essential index for mechanical evaluation. Accurate prediction of both WFT and CS is crucial for mixture design and performance optimization. Nevertheless, UHPC experiments are costly and time-consuming, leading to limited datasets. To overcome this small-sample challenge, this study proposes an integrated framework combining conditional tabular generative adversarial network (CTGAN) for data augmentation, Bayesian optimization (BO)-based eXtreme Gradient Boosting (XGBoost) for predictive modelling, and interpretable machine-learning methods (SHAP and PDPs) for mechanism analysis.

Methods A total of 160 experimental data points were obtained from binary, ternary, and quaternary cementitious systems. The input variables included cement content, PCE dosage, SF, GGBS, sGGBS, FA, sFA, and CA, while the output variables were WFT and CS. P-I 42.5 cement was used as the reference cement, and quartz sand with particle sizes of 0.84–2.00 mm was used as the aggregate. In the binary systems, one mineral admixture was used to replace cement by volume at replacement levels of 0%, 5%, 10%, 15%, and 20%. In the ternary systems, the SF content was fixed at 10%. In the quaternary systems, the GGBS content was fixed at 20% based on the ternary system. WFT was calculated using the wet packing method, and CS was measured after curing for 3, 7, 28 d, and 90 d.

CTGAN was used for data augmentation to expand the original small dataset. The model learned the joint distribution of the real experimental data through adversarial training between the generator and discriminator. After optimization, 1000 synthetic samples were generated for WFT prediction, and 900 synthetic samples were generated for CS prediction. The quality of the synthetic data was evaluated by comparing histograms, cumulative distribution curves, and correlation matrices between real and synthetic datasets.

The synthetic datasets were used to train XGBoost regression models. BO was applied to optimize hyperparameters, including tree depth, learning rate, number of estimators, subsample ratio, column-sampling ratio, minimum child weight, and regularization. Model performance was evaluated by R^2 , MAPE, RMSE, and MAE. SHAP and PDPs were employed to interpret feature contributions and nonlinear effects.

Results and discussion The comparison between real and synthetic datasets showed that CTGAN effectively preserved the main distributional characteristics of the original data. The generated samples had value ranges and distribution patterns similar to those of the real samples. The cumulative distribution curves also showed good consistency between real and synthetic datasets. In addition, the correlation matrices indicated that CTGAN captured not only the marginal distribution of individual variables but also the correlation structure among key variables. Therefore, the generated data could provide a reliable data basis for BO-XGBoost model training under small-sample conditions.

The BO-XGBoost model exhibited excellent prediction performance. For WFT, test R^2 reached 0.986, with training and test RMSE values of 0.909 3 nm and 0.387 6 nm, respectively. For CS, test R^2 was 0.976, with RMSE values of 5.244 8 MPa and 3.837 9 MPa. Predicted values clustered around the 45° ideal line, with most samples within $\pm 10\%$ error. Residuals were concentrated and symmetric, indicating no systematic bias, and loss curves showed stable convergence without overfitting. These results demonstrate that CTGAN augmentation effectively improved the ability of XGBoost to capture complex nonlinear relationships under small-sample conditions.

The SHAP and PDP analyses further revealed the influence mechanisms of different variables. For WFT, sFA showed a strong positive contribution, mainly because its fine and spherical particles improved particle dispersion and increased effective water-film formation. SF had a positive effect at low-to-medium dosages, but excessive SF could reduce WFT because of its high specific

surface area. GGBS and sGGBS tended to show negative effects at high dosages, which may be related to denser particle packing and compression of effective water-film space. PCE showed a nonmonotonic effect: moderate PCE improved WFT, while excessive PCE could weaken the stability of the water-film structure.

For CS, CA was the most important factor and showed a strong positive contribution, reflecting continuous hydration and microstructure densification with curing time. sGGBS, SF, and sFA generally promoted strength development. sGGBS contributed to strength through microfilling and potential pozzolanic activity, while SF improved strength through pore refinement and matrix densification. In contrast, GGBS and FA showed negative effects at high dosages because excessive cement replacement reduced clinker content and limited the availability of calcium hydroxide ($\text{Ca}(\text{OH})_2$) for secondary reactions. These results demonstrate that different mineral admixtures have dosage-dependent and material-specific effects on UHPC performance.

Conclusions This study establishes a multi-response prediction and interpretation framework integrating CTGAN, BO-XGBoost, SHAP, and PDPs for UHPC under small-sample conditions. CTGAN effectively generated high-quality synthetic data while preserving original distribution and correlation characteristics. The optimized BO-XGBoost model achieved outstanding predictive accuracy and generalization, with test R^2 of 0.986 for WFT and 0.976 for CS. The interpretability analysis demonstrated that the effects of mineral admixtures on WFT and CS are highly dependent on material type and dosage: sFA and SF are beneficial for WFT within suitable ranges, whereas sGGBS, SF, and sFA promote CS; excessive GGBS and FA impair CS due to limited reaction activity. The proposed framework offers an effective data-driven approach for UHPC mixture design, performance prediction, and mechanism interpretation when only limited experimental data are available.

Keywords conditional generative adversarial network; ultra-high-performance concrete; data augmentation; water film thickness; compressive strength