

基于去噪扩散概率模型的混凝土配合比数据增强方法

梁 涛¹, 李 悦¹, 杨 斌¹, 刘江林¹, 李亚强², 梁美洁¹

(1. 北京工业大学城市与工程安全减灾教育部重点实验室, 北京 100124; 2. 北京林业大学水土保持学院, 北京 100083)

摘 要: 机器学习模型被广泛应用于混凝土配合比设计和性能预测, 数据集的质量与数量对模型的拟合精度有着直接影响。然而, 由于试验周期长、测试成本高、人为影响因素大等原因, 构建大规模、高质量的混凝土配合比数据集仍面临显著挑战。本工作基于去噪扩散概率模型, 构建了一种生成混凝土配合比虚拟数据的深度学习模型——MixDiff。定量分析结果显示, MixDiff 生成的虚拟数据集的特征分布、特征相关性与实际数据集具有较高相似性, 两者的瓦瑟斯坦距离仅为 0.019 5, K-S 统计量仅为 0.095 2, J-S 散度仅为 0.120 7, 皮尔逊相关矩阵差异的弗罗贝尼乌斯距离仅为 0.332 7。使用多种机器学习模型分别基于真实小样本数据集和 MixDiff 生成的虚拟数据集进行训练, 试验表明: 通过引入虚拟数据集, RF、SVR、KNN、XGBoost、LightGBM、CatBoost 和 DNN 共 7 种模型的平均 R^2 相对上升 16.72%, 平均 RMSE 相对下降 47.96%, 平均 MAE 相对下降 46.86%。MixDiff 生成的配合比数据在 7 d 强度、28 d 强度、流动度、坍落度、初凝时间与终凝时间上的平均 MAPE 仅为 3.8%。本工作为混凝土配合比的数据扩充提供了一种新方法, 可提升机器学习模型在小样本条件下的预测精度。

关键词: 混凝土; 配合比; 去噪扩散概率模型; 数据增强; 深度学习

中图分类号: TU528 文献标志码: A 文章编号: 0454-5648(2026)10-3499-19

网络出版时间: 2026-09-02



随着人工智能技术的发展, 机器学习已被广泛应用于混凝土材料的宏观性能预测与配合比优化设计中^[1-3]。在此类任务中, 模型的预测精度高度依赖于训练数据的数量与质量。然而, 由于混凝土配合比试验周期长、测试成本高、人为影响因素大等客观原因, 构建大规模、高质量的混凝土配合比数据集面临挑战。

为应对小样本数据带来的建模挑战, 现有研究主要从模型优化与数据扩充 2 个方面展开。在模型优化方面, 响应面法^[4]和贝叶斯方法^[5]是混凝土材料配合比设计与性能预测中常用的建模方法。响应面法通常通过构建输入特征与目标变量之间的近似函数, 揭示关键配合比特征对宏观性能的影响规律, 并为配合比优化提供可解释的数学依据; 贝叶斯方法则能够在有限样本条件下引入先验信息并描述预测结果的不确定性, 适用于小样本场景下的性能预测与优化搜索。随着机器学习的发展, 研究者进一步通过决策树引导神经网络预训练^[6]、结合

高斯过程的深度贝叶斯网络^[7]以及迁移学习^[8]等方法, 提升模型在小样本数据集上的拟合能力。这类方法的共同特点是通过优化模型结构、引入先验知识等方式, 实现有限样本中有效信息的深度挖掘, 对降低小样本场景下的建模误差有着一定作用。然而, 这类方法并未从根本上增加可利用的样本数量, 其性能提升仍受到原始样本规模的限制。相比之下, 数据扩充方法可通过生成虚拟样本来缓解数据不足的问题, 为小样本混凝土性能预测任务提供了一种新思路。其中, 以生成对抗网络及其变体为代表的生成式模型已被初步应用于混凝土性能试验数据生成^[9, 10]。然而, 已有研究表明, 生成对抗网络在高维数据建模中常面临训练不稳定、样本多样性不足以及模式崩塌等问题^[11, 12], 限制了其在复杂混凝土配合比数据生成任务中的使用。因此, 亟须引入一种具有更强建模能力与训练稳定性的生成范式, 以实现小样本场景下混凝土配合比数据的有效扩充。

收稿日期: 2026-05-09。 修订日期: 2026-06-05。

基金项目: 国家自然科学基金重点项目(U2368207); 北京市自然科学基金(24JL003); 国家自然科学基金(52508244); 中国博士后科学基金(GZC20252135)。

第一作者: 梁 涛(1997—), 男, 博士研究生。

通信作者: 李 悦(1972—), 男, 博士, 教授。

Received date: 2026-05-09. Revised date: 2026-06-05.

First author: LIANG Tao (1997—), male, Doctoral candidate.

E-mail: li971102@163.com

Correspondent author: LI Yue (1972—), male, Ph.D., Professor.

E-mail: liyue@bjut.edu.cn

去噪扩散概率模型^[13]通过前向逐步加噪与反向逐步去噪过程学习真实数据的潜在分布,具有出色的复杂联合概率分布建模能力,已在生成式人工智能领域取得了巨大进展。去噪扩散概率模型^[13]通过固定的前向马尔可夫添加噪声过程和参数化的反向去噪过程学习数据分布,在无需对抗训练的情况下实现了高质量生成,为扩散生成模型的发展奠定了基础。随后,去噪扩散隐式模型^[14]通过构造非马尔可夫反向生成过程,显著减少采样步数,提高了扩散模型的生成效率。Score-SDE^[15]则从连续时间随机微分方程的角度对扩散过程进行建模。该方法将扩散模型与基于分数匹配的生成模型联系起来,为复杂数据分布建模提供了更一般的理论框架。此外,潜在扩散模型^[16]将扩散过程由原始数据空间转移至低维潜空间,在降低计算成本的同时保持了较高的生成质量。然而,上述研究主要面向图像、语音等高维非结构化数据,其网络结构多依赖卷积、注意力或潜空间编码机制,难以直接适配混凝土配合比这类低维、连续、强相关的小样本表格数据。

针对上述问题,本工作提出了一种面向混凝土配合比数据生成的残差全连接去噪扩散概率模型——MixDiff,旨在实现小样本条件下混凝土配合比数据的扩充,精确复现材料复杂组分与性能之间的潜在映射关系,生成具有统计一致性和物理合理性的虚拟配合比数据。针对配合比数据的二维结构化特点,MixDiff并未使用图像扩散模型中常用的卷积或U-Net结构,而是采用全连接层^[17]搭建整个网络,以更直接地表征各配合比变量及性能指标之间的映射关系;同时,MixDiff在去噪网络中引入残差全连接^[18],改善小样本训练过程中深层网络梯度衰减的问题,提高反向去噪过程的稳定性与表达能力。

为系统评估MixDiff生成数据的可靠性,本工作首先构建了小样本真实数据集,并利用MixDiff学习其联合分布特征,生成大规模虚拟配合比数据集。采用瓦瑟斯坦距离、K-S统计量、J-S散度等指标,从边缘分布与特征相关性两个角度定量评估虚拟数据质量。进一步使用多种机器学习算法在基准数据模式与数据增强模式下开展对比试验,验证了MixDiff在提升预测精度、降低预测方差方面的效果;最后通过试验证明了MixDiff生成数据的有效性。本工作的整体技术路线图如图1所示。

1 小样本数据库建立

1.1 实验原材料

实验用胶凝材料包含水泥、粉煤灰和矿渣粉。其中,水泥选用P·O 42.5级普通硅酸盐水泥,其主要粒径(D_{10} ~ D_{90})分布区间为5~35 μm 。选用II级粉煤灰,其主要粒径分布在5~85 μm 之间。矿渣粉则选用S95级矿渣粉,主要集中在4~31 μm 。胶凝材料的化学组成见表1,粒径分布曲线见图2。

为优化混凝土的堆积密度并模拟真实的工程应用场景,本工作采用了不同细度模数的细骨料与连续级配的粗骨料。其中,细骨料包含细度模数为2.2的天然河砂和细度模数为2.9的机制砂。粗骨料选用碎石,粒径分别为5~10 mm和5~25 mm。减水剂为固含量为31%的AN400型聚羧酸系高性能减水剂。

1.2 试件制备与性能测试

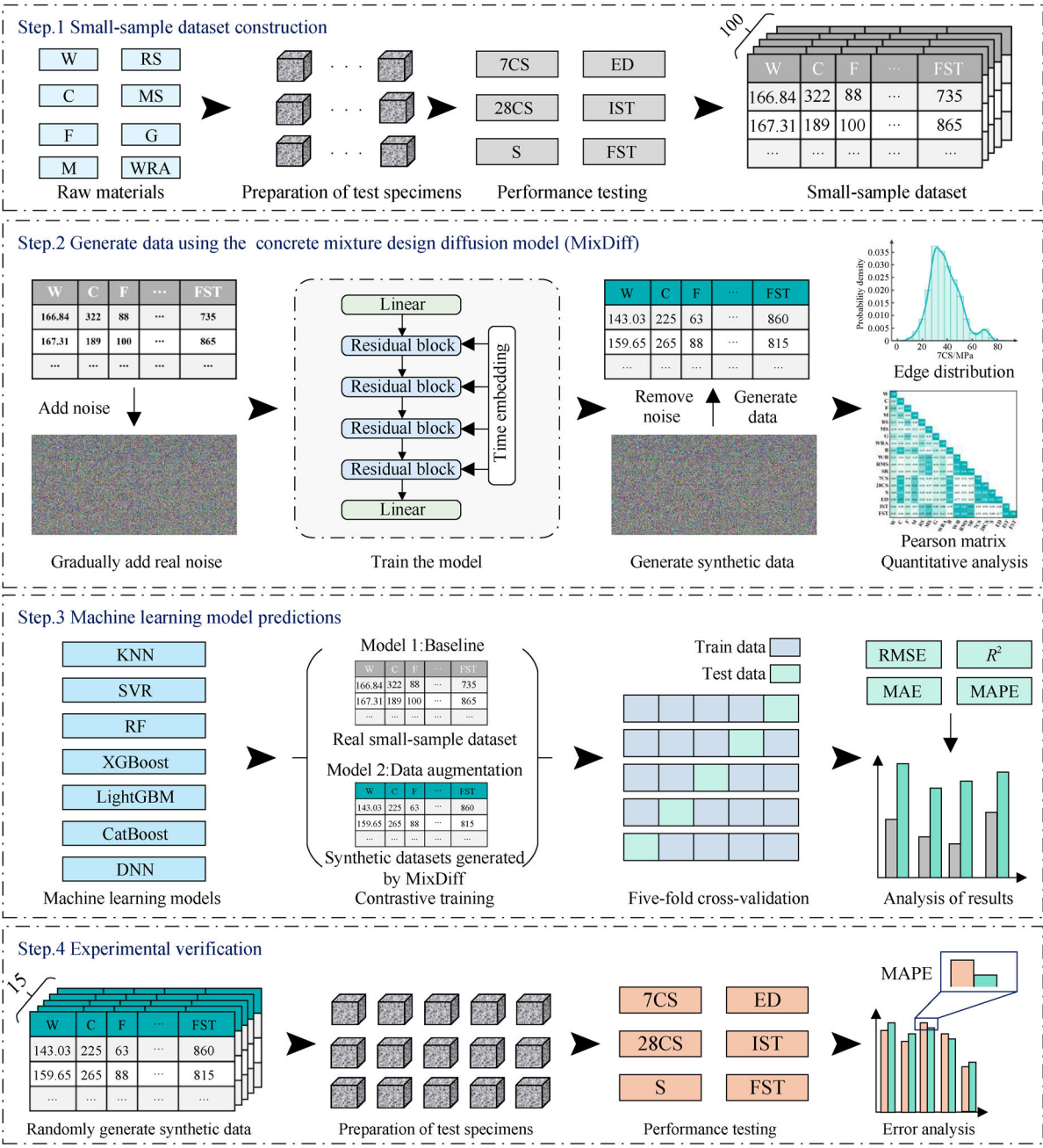
针对小样本场景,制备了100组C20~C80混凝土试件。测试了混凝土的7 d抗压强度(7CS)、28 d抗压强度(28CS)、坍落度(S)、扩展度(ED)、初凝时间(IST)及终凝时间(FST),测试方法参照国标GB/T 50080—2016《普通混凝土拌合物性能试验方法标准》和GB/T 50081—2019《混凝土物理力学性能试验方法标准》。试验及试件养护实况如附录1所示。

1.3 小样本数据集构建结果

1.3.1 基本信息 基于试验构建了一个包含100条数据的小样本数据集。输入特征的维度为12,包括8个组分特征与4个衍生特征。其中组分特征包含水(W)、水泥(C)、粉煤灰(F)、矿渣粉(M)、河砂(RS)、机制砂(MS)、骨料(G)及减水剂(WRA)。衍生特征包含胶凝材料总量(B)、水胶比(W/B)、细骨料总量(RMS)及砂率(SR)。输出目标变量包括7CS、28CS、S、ED、IST及FST。

输入特征和目标变量的数理统计结果分别见表2和表3。尽管样本总量有限,但该数据集在特征空间中仍表现出良好的覆盖性。例如,W/B的范围为0.274~0.686,28CS的范围为23.50~87.70,实现了从普通混凝土到高强混凝土的覆盖。这种多维度的参数分布显著增强了输入特征与目标变量之间的非线性映射难度。

1.3.2 特征分布 图3展示了小样本数据集的皮尔逊相关系数矩阵。由图可见,输入特征和目标变量之间的皮尔逊相关系数^[19]存在明显的相关性。例如,组分特征中的C与7CS和28CS表现出正相关,其相关系数均为0.92;衍生特征中的W/B与IST及



W is water, C is cement, F is fly ash, M is mineral fines, RS is river sand, MS is manufactured sand, G is coarse aggregate, WRA is water-reducing agent, 7CS is 7 d compressive strength, 28CS is 28 d compressive strength, S is slump, ED is expansion degree, IST is initial setting time, and FST is final setting time.

图 1 整体技术路线图

Fig. 1 Overall technology roadmap

表 1 胶凝材料的化学组成

Table 1 Chemical composition of cementitious materials							w/%
Materials	CaO	SiO ₂	Al ₂ O ₃	SO ₃	MgO	Fe ₂ O ₃	Other
Cement	51.70	23.41	8.61	4.23	6.73	2.60	2.72
Fly ash	5.45	44.19	38.48	1.13	2.28	3.96	4.51
Mineral fines	33.32	29.96	18.50	3.22	11.69	0.43	2.88

w is mass fraction.

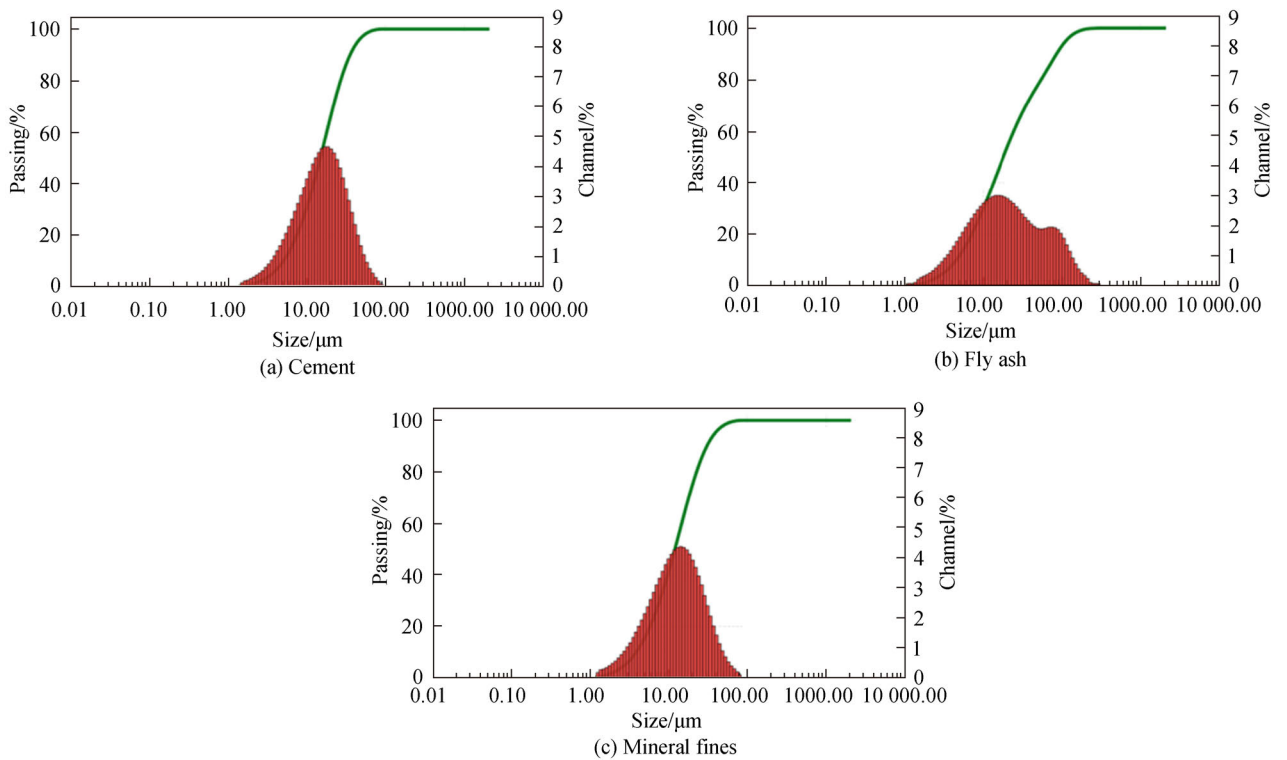


图 2 粒径分布曲线图
Fig. 2 Particle size distribution curve

表 2 输入特征的数理统计结果

Table 2 Mathematical statistics results for input features

Feature	Component features							Derived features				
	$\rho(W)/$ ($\text{kg}\cdot\text{m}^{-3}$)	$\rho(C)/$ ($\text{kg}\cdot\text{m}^{-3}$)	$\rho(F)/$ ($\text{kg}\cdot\text{m}^{-3}$)	$\rho(M)/$ ($\text{kg}\cdot\text{m}^{-3}$)	$\rho(RS)/$ ($\text{kg}\cdot\text{m}^{-3}$)	$\rho(MS)/$ ($\text{kg}\cdot\text{m}^{-3}$)	$\rho(G)/$ ($\text{kg}\cdot\text{m}^{-3}$)	WRA	$\rho(B)/$ ($\text{kg}\cdot\text{m}^{-3}$)	W/B	$\rho(RMS)/$ ($\text{kg}\cdot\text{m}^{-3}$)	SR
Mean	159.3	263.5	93.2	40.7	203.0	562.3	1040.4	0.016 3	397.4	0.414	765.1	0.4232
Std	8.5	60.5	32.5	26.5	45.3	65.0	27.0	0.001 4	73.8	0.077	62.9	0.0223
Min	141.4	85.0	34.0	0	155.0	342.0	966.0	0.014 0	242.0	0.274	621.0	0.3699
Max	181.2	448.0	192.0	95.0	388.0	646.0	1108.0	0.020 0	600.0	0.686	893.0	0.4601

ρ is density, Std is standard deviation, B is total binder content, W/B is the water-to-binder ratio, RMS is total fine aggregate content (RS + MS), and SR is the sand ratio.

表 3 目标变量的数理统计结果

Table 3 Mathematical statistics results for target variable

Feature	7CS/MPa	28CS/MPa	S/mm	ED/mm	IST/min	FST/min
Mean	38.62	54.47	225.40	563.6	647.80	793.95
Std	12.23	13.22	12.26	37.68	53.25	60.95
Min	13.60	23.50	180.00	470.00	535.00	660.00
Max	74.70	87.70	255.00	675.00	780.00	925.00

FST 呈正相关，其相关系数分别为 0.94 和 0.91。这种特征之间的强相关性表明，输入特征包含了目标变量所需的关键特征信息，为机器学习模型构建高精度的非线性映射提供了基础。

图 4 与图 5 为输入特征与目标变量的概率密度分布。由图可知，多数特征呈现出近似单峰分布，

表明样本主要集中于相对稳定的取值区间，有利于模型捕捉主要变化规律。部分特征(如 W/B 与 WRA)呈现出明显的双峰结构，说明该小样本数据集包含多样化的配合比数据，该类分布能够反映更丰富的物理信息，但在小样本条件下也会提高模型拟合难度并降低预测稳定性。

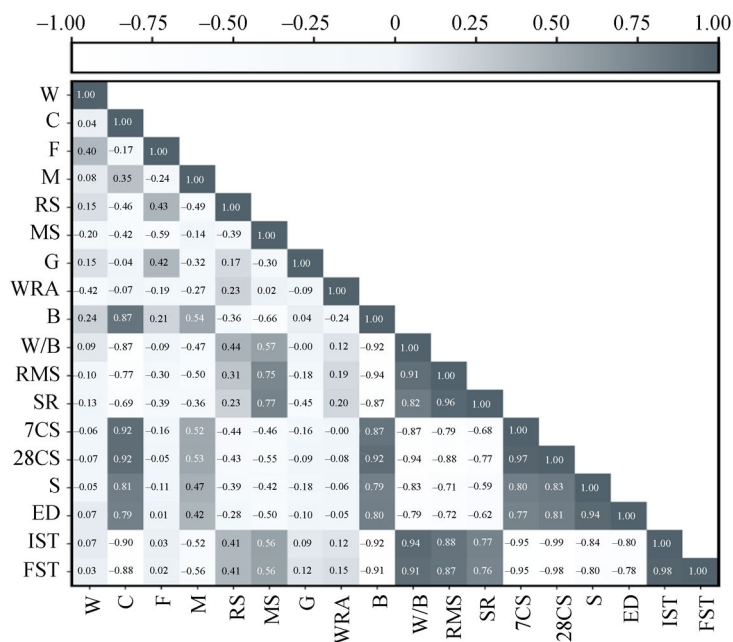
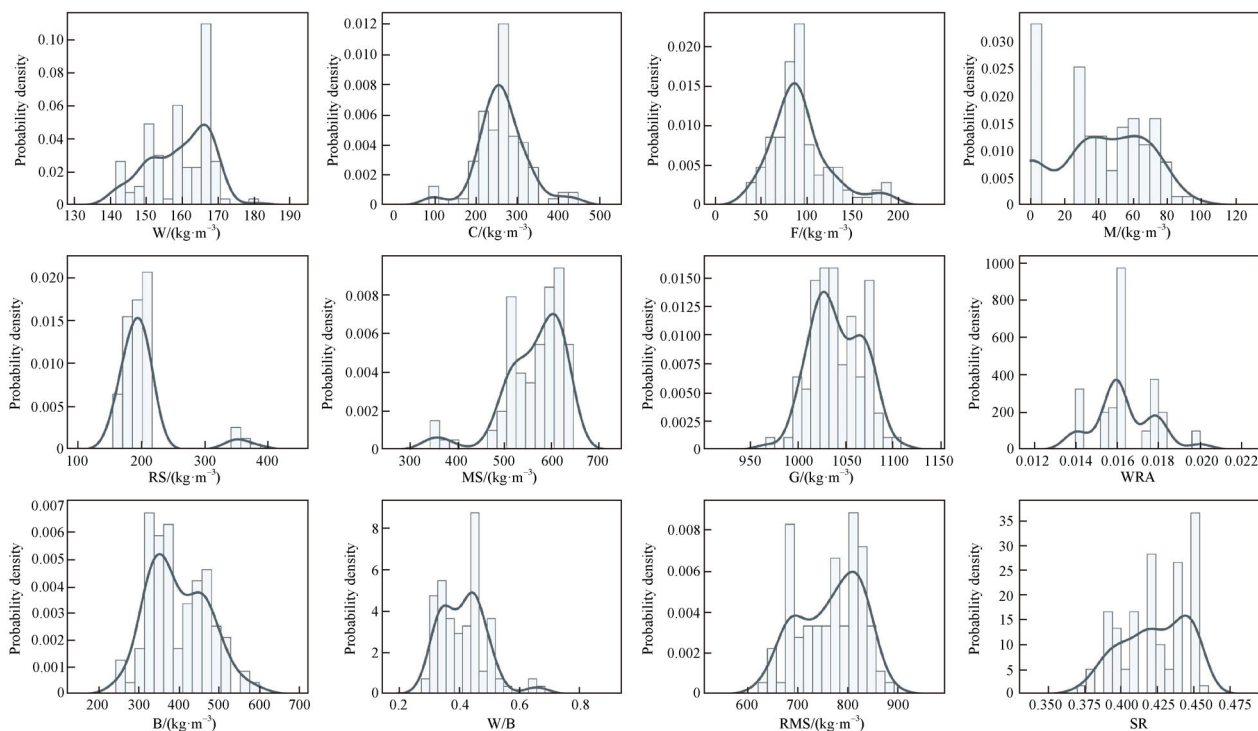


图 3 小样本数据集的皮尔逊相关系数矩阵

Fig. 3 Pearson correlation matrix for a small dataset



Panels in the same row share the same y-axis label and unit, which are displayed on the leftmost panel.

图 4 小样本数据集输入特征的概率密度分布

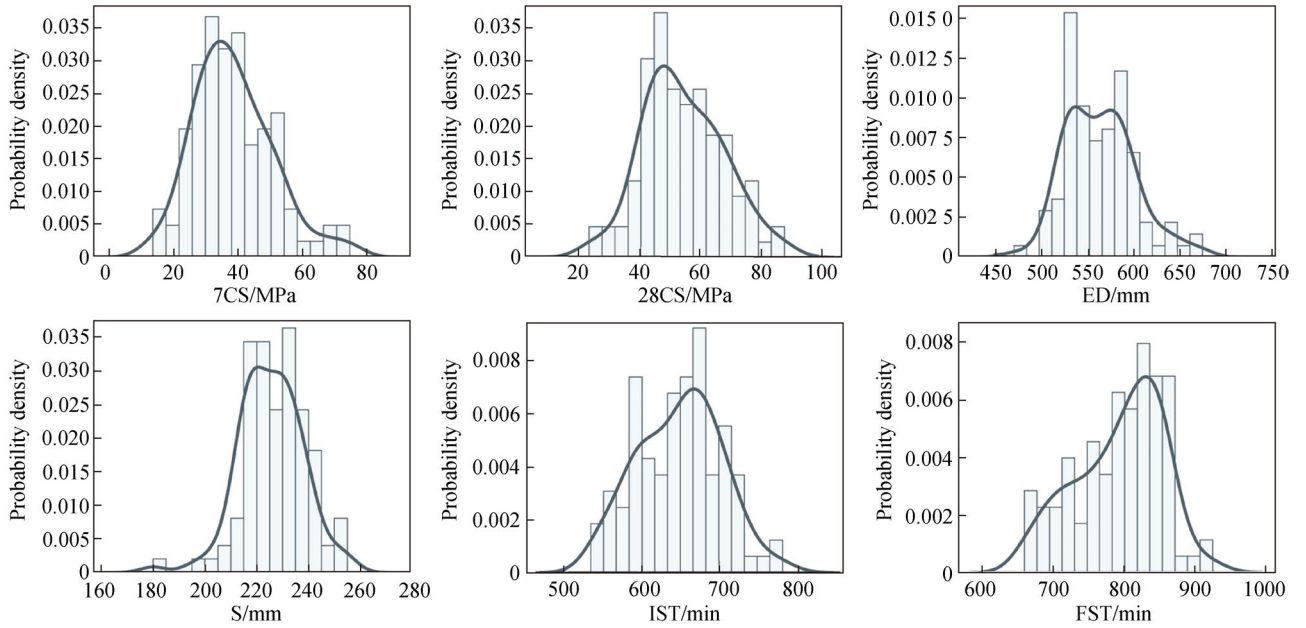
Fig. 4 Probability density distribution of input features in small-sample datasets

2 基于去噪扩散概率模型的小样本数据扩充方法

2.1 去噪扩散概率模型的原理

2.1.1 前向扩散过程 前向扩散过程是一个向真

实数据逐步注入噪声的过程^[20]。给定从实验数据真实分布中采样的配合比数据 $x_0 \in \mathbb{R}^D$ (D 是特征维度), 通过预设的余弦方差调度^[21]逐步向数据添加高斯噪声。对于任意的离散时间步 $t \in (0, T)$, T 为总扩散步数, 其状态转移分布定义为:



Panels in the same row share the same y-axis label and unit, which are displayed on the leftmost panel.

图5 小样本数据集目标变量的概率密度分布
Fig. 5 Probability density distribution of the target variable in a small-sample dataset

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

其中: t 为扩散时间步; x_{t-1} 表示第 $t-1$ 步的含噪数据; x_t 表示第 t 步的含噪数据; β_t 为第 t 步注入的噪声方差系数, 由余弦方差调度策略动态决定, 用于控制真实数据信息随扩散过程衰减的速率; $\mathbf{I}$ 为 D 维单位矩阵; $\mathcal{N}(\cdot)$ 表示高斯分布。

利用高斯分布的可加性与重参数化技巧, 任意时刻 t 的边缘分布 $q(x_t | x_0)$ 可由初始真实数据 x_0 直接得到, 其闭式解为

$$q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) \mathbf{I}) \quad (2)$$

其中: $\alpha_t = 1 - \beta_t$, 表示第 t 步中原始信号的保留比例; $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, 表示从第 1 步到第 t 步累积信号保留系数; 随着 t 的增加, $\bar{\alpha}_t$ 逐渐趋近于 0, 意味着原始数据中的有效信息不断减少, 最终分布逐渐接近标准高斯噪声。该性质使得模型在训练阶段可以高效采样任意时间步的含噪样本, 从而覆盖从真实数据分布到纯噪声分布的整个样本空间。

2.1.2 反向生成过程 数据生成的核心在于反向生成过程, 其目标是从无信息的标准高斯分布 $x_T \sim \mathcal{N}(0, \mathbf{I})$ 中逐步去除噪声, 最终恢复出具有物理意义的配合比数据 x_0 ^[20]。由于真实的后验分布 $q(x_{t-1} | x_t)$ 需要对整个数据分布进行积分, 难以直接

解析求解, 因此采用深度神经网络 $\epsilon_\theta(x_t, t)$ 对其进行拟合。本工作采用噪声预测策略, 即训练神经网络预测当前样本 x_t 中包含的噪声分量。基于贝叶斯定理推导, 从状态 x_t 重构上一时刻 x_{t-1} 的采样公式可表示为

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left[x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right] + \sigma_t z \quad (3)$$

其中: x_{t-1} 为反向生成过程中第 $t-1$ 步的数据状态; x_t 为第 t 步的含噪声数据; $\alpha_t = 1 - \beta_t$ 表示第 t 步的信号保留系数; $\bar{\alpha}_t$ 为累积信号保留系数; $\epsilon_\theta(x_t, t)$ 为神经网络预测的噪声分量; θ 表示神经网络中的可训练参数; σ_t 为反向采样过程中的噪声标准差; $z \sim \mathcal{N}(0, \mathbf{I})$ 为随机高斯扰动项, 用于维持生成样本的随机性与多样性。

整个反向过程是从纯噪声到真实数据的逐步重构过程。模型从标准高斯分布出发, 经过 T 次迭代逐步剥离预测噪声, 最终将随机噪声映射为符合真实数据分布特征的混凝土配合比样本。

2.2 残差全连接去噪扩散概率模型

基于 2.1 小节的理论, 本工作构建了用于生成混凝土配合比数据的扩散模型(MixDiff), 其结构图如图 6 所示。

由于混凝土配合比数据属于表格型结构化数据, MixDiff 并未使用计算机视觉领域常用的卷积神

神经网络架构,而是基于残差全连接网络。MixDiff 主要包含以下 3 个核心模块:

1) 时间嵌入模块:为了使模型感知扩散过程中的时间步信息,引入正弦位置嵌入机制^[22],将离散的时间步 t 映射为高维连续向量 e_t ,以保留不同时间

间步之间的相对位置信息。其数学表达式如下:

$$PE(t, 2i) = \sin(t / 10\,000^{2i/d}) \quad (4)$$

$$PE(t, 2i+1) = \cos(t / 10\,000^{2i/d}) \quad (5)$$

其中: t 表示当前扩散时间步; i 表示嵌入向量中的维度索引; d 表示时间嵌入向量的维度。

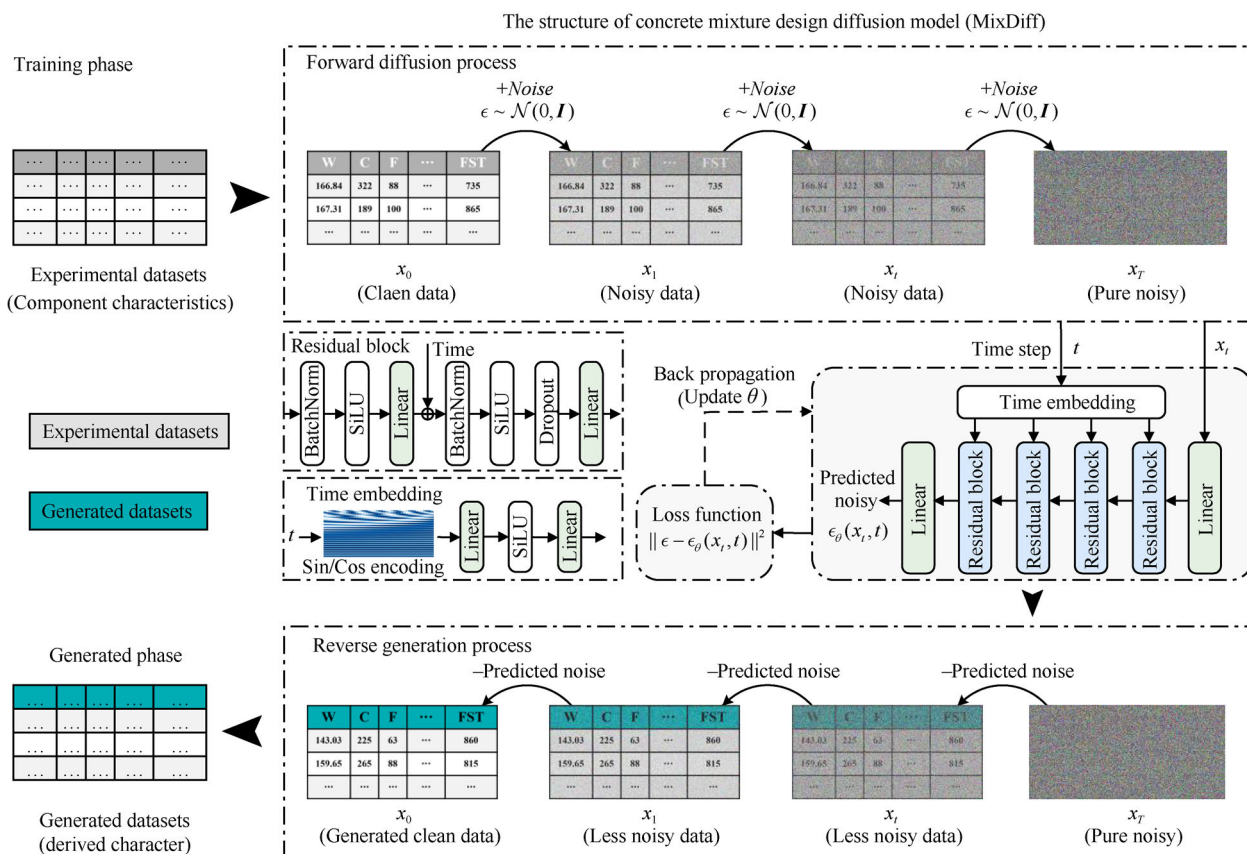


图 6 MixDiff 的结构图
Fig. 6 Structure diagram of MixDiff

随后, e_t 经过 2 个全连接层与 SiLU 激活函数^[23]进行非线性映射,并被投影到与特征空间相同的维度。投影后的时间信息通过加法操作注入到每个残差块中,使网络能够在反向去噪的不同阶段,根据当前噪声强度动态调整特征提取策略。

2) 残差全连接网络: MixDiff 的核心特征提取器由 4 个堆叠的残差块组成。每个残差块包含全连接层、批归一化层^[24]与 SiLU 激活函数。相比于传统的 ReLU 激活函数, SiLU 具有平滑、非单调的特点,有助于改善深层网络训练过程中的梯度传播,同时能够增强模型对混凝土组分非线性耦合关系的拟合能力。残差全连接可以使特征信息在深层网络中传递,减少了模型过拟合风险^[18]。

3) 损失函数: MixDiff 的训练目标是 최소화模

型预测噪声 $\epsilon_\theta(x_t, t)$ 与真实注入噪声 ϵ 之间的均方误差:

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{t, x_0, \epsilon} [\|\epsilon - \epsilon_\theta(x_t, t)\|^2] \quad (6)$$

其中: $\mathcal{L}_{\text{simple}}$ 表示预测噪声的损失函数; x_0 表示真实数据; x_t 为第 t 步的含噪声数据; t 表示随机采样的扩散时间步; $\|\cdot\|^2$ 表示二范数平方; $\mathbb{E}[\cdot]$ 表示对真实样本、时间步和噪声分布求期望。

此外, MixDiff 仅负责生成 C、W、F 等组分特征,而 W/B 等衍生特征则在后处理阶段计算得出。

2.3 模型实现细节

MixDiff 基于 PyTorch 框架构建并在单张 RTX 4060(8 GB)显卡上训练。训练过程采用 AdamW 优化器,并结合余弦退火学习率调度算法,以平

衡小样本条件下的模型收敛速度与泛化能力。为消除不同特征量纲差异对模型训练的影响,所有特征均进行归一化处理。超参数配置如表 4 所示。

表 4 MixDiff 的超参数设置
Table 4 MixDiff hyperparameter settings

Parameter	Value	Description
Diffusion steps	2000	Number of discrete steps in the forward diffusion process, controlling the gradual transformation of data into Gaussian noise
Hidden dimension	512	Hidden layer dimension, used to enhance the model's ability to capture complex nonlinear relationships
Batch size	16	Batch size, balancing GPU memory usage and the stability of gradient estimation
Learning rate	1×10^{-4}	Initial learning rate of the optimizer, dynamically adjusted using a cosine annealing schedule
Epochs	5000	Number of training epochs, ensuring sufficient convergence of the model under small-sample conditions
Optimizer	AdamW	Optimization algorithm

2.4 评价指标

为定量评估 MixDiff 生成数据与真实数据之间的边缘分布一致性,本工作使用瓦瑟斯坦距离(WD)、K-S 统计量(D_{KS})、J-S 散度(JSD)等 3 个评价指标进行评估。其中 WD 用于反映分布位置与形状的综合差异;KS 用于反映两个累积分布函数之间的最大偏差;JSD 用于衡量生成数据与真实数据在概率密度层面的相似性。公式如下所示:

$$WD = \inf_{\gamma \in \Pi(p, q)} \int |x - y| d\gamma(x, y) \quad (7)$$

$$KS = \sup_x |F_1(x) - F_2(x)| \quad (8)$$

$$JSD = \frac{1}{2}KL(p \| m) + \frac{1}{2}KL(q \| m) \quad (9)$$

其中: p 和 q 分别表示真实数据分布与生成数据分布; $\Pi(p, q)$ 表示所有以 p 和 q 为边缘分布的联合分布集合; $\gamma(x, y)$ 表示分布 p 中位置 x 与分布 q 中位置 y 之间的概率质量耦合关系; $F_1(x)$ 与 $F_2(x)$ 分别表示真实数据与生成数据的累积分布函数; $\sup$ 表示取最大值; $KL(\cdot)$ 表示 Kullback-Leibler 散度; $m = \frac{(p+q)}{2}$ 表示 p 和 q 的平均分布。

此外,为评估特征之间的相关结构一致性,引入基于皮尔逊相关矩阵的弗罗贝尼乌斯距离(FD)。该指标用于量化真实数据与生成数据在特征相关性

矩阵上的整体差异,其公式如下:

$$FD = \|R_{\text{real}} - R_{\text{gen}}\|_F = \sqrt{\sum_i \sum_j (r_{i,j}^{\text{real}} - r_{i,j}^{\text{gen}})^2} \quad (10)$$

其中: R_{real} 和 R_{gen} 分别表示真实数据与生成数据的皮尔逊相关系数矩阵; r_{ij} 表示矩阵中第 i 行第 j 列的元素; d 表示特征维度; FD 值越小,说明生成数据与真实数据在特征相关性上越接近。

2.5 实验结果

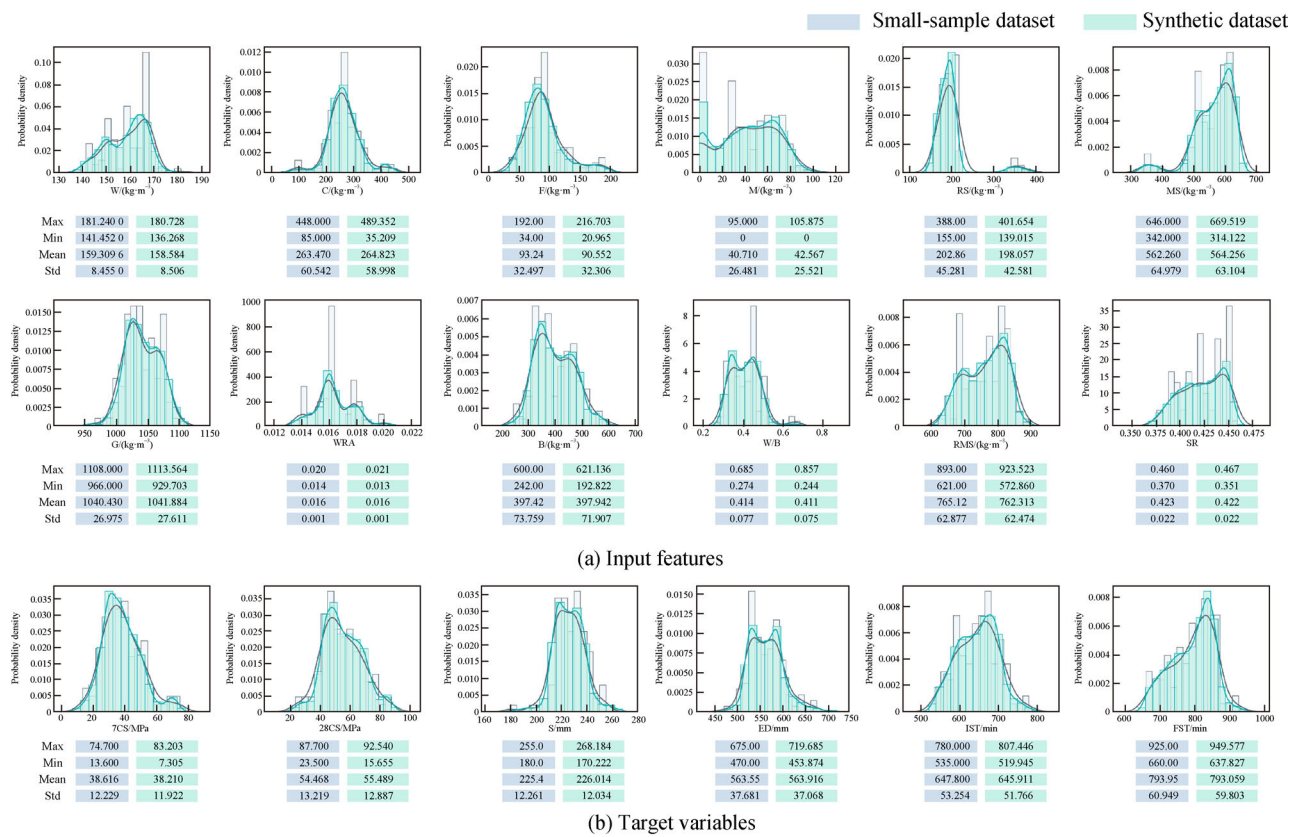
为评估 MixDiff 生成虚拟数据集与真实小样本数据集之间的分布,本小节以 3000 条生成的样本为例,从边缘分布、特征相关性两个维度对虚拟数据集进行评估。

2.5.1 边缘分布 图 7a 与图 7b 分别展示了虚拟数据集与真实小样本数据集中输入特征与目标变量的概率密度函数对比,其中灰色曲线为真实小样本数据集,绿色曲线代表了 MixDiff 生成的虚拟数据集。可以看出,生成数据与真实数据在峰值位置、分布宽度及整体形态上均表现出较高的一致性。对于 W/B 和 RS 等呈现显著双峰的变量, MixDiff 仍能够较好地捕捉其非正态分布特征,使生成数据在概率密度分布上与真实数据保持较高相似性。上述结果表明, MixDiff 能够有效学习真实混凝土配合比数据的分布规律,并具备较好的数据生成能力。

使用 2.4 小节所述的评价指标量化了虚拟数据集与真实小样本数据集之间分布差异,量化结果展示在表 5 中。量化结果表明, WD 的平均值仅为 0.019 5,表明生成数据在特征空间中与真实数据分布具有较高相似性,整体分布偏差较小。KS 统计量与 JSD 的平均仅为 0.095 2 和 0.120 7,进一步说明生成数据与真实数据在累积分布特征和概率密度重叠程度上具有良好一致性。表 5 中 SR、WRA 和 S 等特征的 JSD 相对较高,主要与其真实样本分布较集中以及 JSD 对局部概率密度差异较敏感有关。如图 7 所示,真实数据与生成数据的 SR 均值分别为 0.423 和 0.422,标准差均为 0.022,说明二者整体统计特征较为接近。综合来看, MixDiff 能够较好地学习真实小样本数据的分布规律,并生成与真实数据具有较高统计相似性的虚拟配合比数据。

2.5.2 特征相关性 混凝土的宏观性能取决于各组分之间复杂的耦合作用,因此,生成数据能否保持真实数据的特征相关性是评价其可靠性的重要依据。皮尔逊相关系数矩阵能够表征不同特征之间的线性相关关系,可用于对比真实小样本数据与虚拟数据在相关结构上的一致性。图 8a 和图 8b 分别展

示了真实小样本数据集与虚拟数据集的皮尔逊相关系数矩阵热力图，颜色越深代表了 2 个特征之间具有较强的相关性。观察发现，两者在整体色彩分布上表现出较高一致性，说明 MixDiff 能够较好地捕捉特征之间的相关性。例如，W/B 与 28CS 之间呈现负相关，与 IST 和 FST 均呈现正相关。



Panels in the same row share the same y-axis label and unit, which are displayed on the leftmost panel.

图 7 边缘分布对比
Fig. 7 Comparison of edge distributions

表 5 边缘分布量化结果
Table 5 Quantification results of the edge distribution

Feature	Metric (10^{-2})			Feature	Metric (10^{-2})		
	WD	KS	JSD		WD	KS	JSD
W	2.66	14.03	14.11	RMS	1.62	5.83	8.07
C	1.27	7.87	7.25	SR	3.08	13.47	37.25
F	2.13	11.43	8.32	7CS	0.97	4.03	3.65
M	2.97	13.20	12.53	28CS	1.73	6.80	5.82
RS	2.13	14.13	4.88	S	2.04	10.47	29.46
MS	1.19	5.40	7.54	ED	1.20	6.40	5.19
G	1.74	7.50	9.43	IST	1.37	5.37	8.71
WRA	4.79	25.60	28.63	FST	1.19	5.93	6.89
B	1.19	5.27	8.68	AVG	1.95	9.52	12.07
W/B	1.77	8.67	10.86				

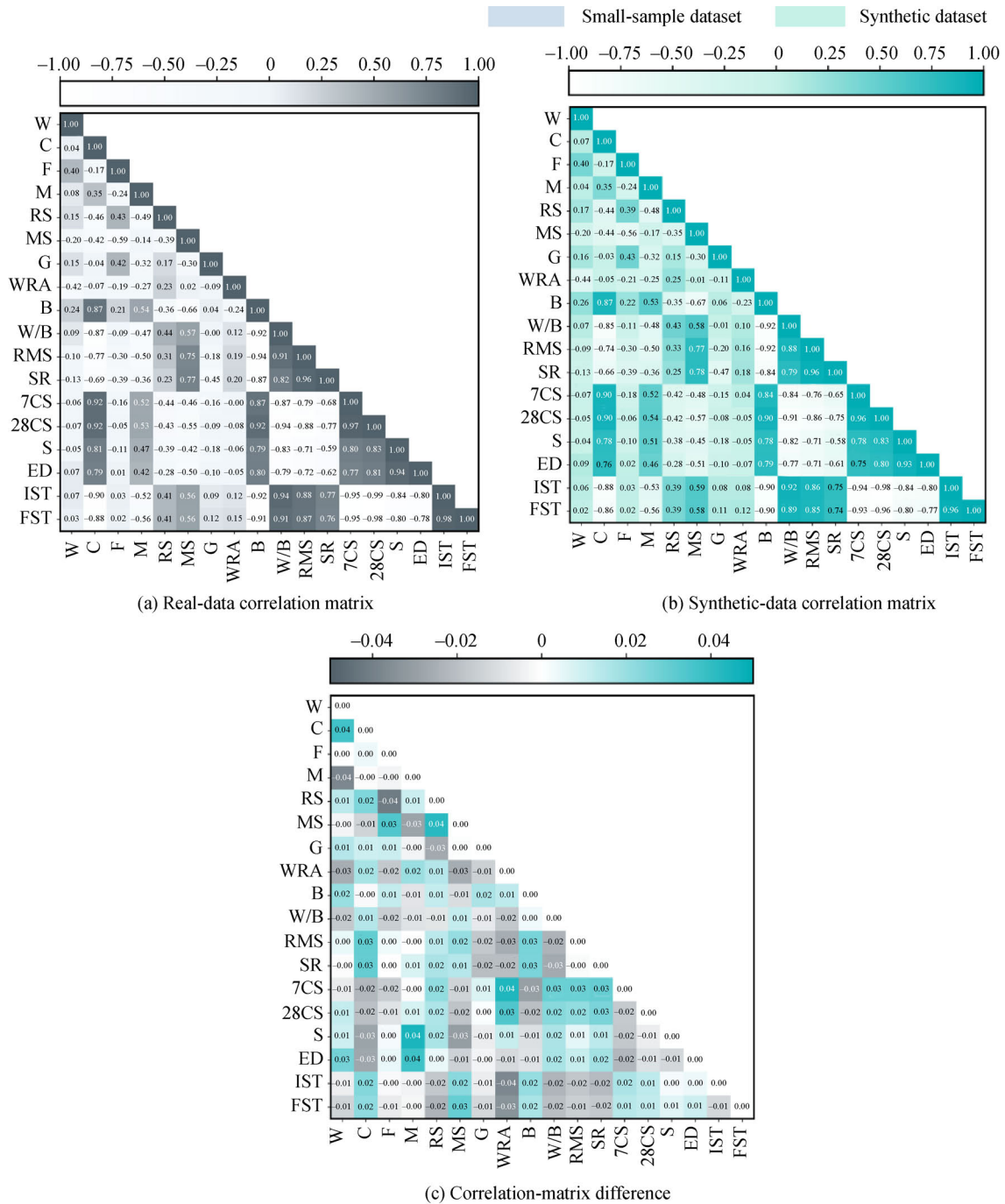


图8 特征相关性对比

Fig. 8 Comparison of feature correlations

为进一步分析2组相关矩阵之间的差异,图8c展示了真实数据与虚拟数据皮尔逊相关系数矩阵的差值。观察发现,特征之间的相关性差异整体较小,且主要集中在 $[-0.04, 0.04]$ 这一极小区间。进一步使用2.4小节定义的FD指标对2个矩阵的整体差异进行定量分析,计算结果仅为0.3327。这表明MixDiff生成的虚拟数据能够在特征相关性上较好地保持与真实小样本数据的一致性。

3 数据扩充对机器学习模型性能的影响

3.1 机器学习模型构建

大量研究表明,训练数据的数量与质量会直接影响机器学习模型的预测精度和泛化能力^[25–27]。为进一步验证MixDiff生成数据的有效性,并降低单一模型评价带来的偶然性,本工作选取多种典型机

器学习模型进行对比分析。所采用的模型如下:

1) K-最邻近回归(KNN)。

KNN 是一种非参数化的机器学习算法^[28], 是根據待预测样本在特征空间中的邻近样本信息估计目标值。该模型不依赖显式函数假设, 能够较好地反映局部样本分布特征, 适用于数据关系相对复杂但样本规模较小的回归任务。

2) 支持向量回归(SVR)。

SVR 是基于结构风险最小化原则的回归方法^[29]。其通过引入不敏感损失函数, 在允许一定预测误差的同时提高模型的泛化能力。对于非线性任务, SVR 采用核函数将低维输入投影至高维空间, 从而增强对复杂非线性关系的拟合能力。SVR 的泛化性能对超参数配置高度敏感, 实际建模中必须依赖交叉验证与参数寻优架构, 以规避过拟合或欠拟合风险。

3) 随机森林(RF)。

RF 是一种基于集成学习与 Bagging 范式的算法, 通过构建多棵决策树并对其预测结果进行平均来完成回归任务^[30]。该模型利用样本重采样和特征随机选择降低单棵树的过拟合风险, 具有较好的抗噪能力和稳定性, 适合处理非线性特征关系。

4) 极端梯度提升(XGBoost)。

XGBoost 是一种基于集成学习与 Boosting 范式的回归算法^[31]。该方法通过串行构建弱学习器, 逐步修正前一轮模型的残差, 以降低预测偏差。相比传统梯度提升方法, XGBoost 引入二阶梯度信息和正则化项, 有助于提高模型拟合能力并控制过拟合风险。

5) 轻量级梯度提升机(LightGBM)。

LightGBM 是一种高效的集成回归模型, 利用直方图算法和 Boosting 方法解决了大规模高维度数据的训练问题^[32]。该方法采用按叶节点生长的策略, 通过引入深度限制、单边梯度采样等技术, 减少了节点分裂时计算复杂度。对于高维特征和非线性回归问题, LightGBM 能够较快捕捉特征间的交互关系, 适合用于大规模数据增强后的建模任务。

6) 分类特征梯度提升(CatBoost)。

CatBoost 是一种基于 Boosting 范式的集成学习模型, 采用对称决策树结构和排序提升机制^[33]。该方法能够有效降低预测偏移, 并具有良好的稳定性和泛化能力。虽然本工作数据以连续特征为主, 但 CatBoost 在处理复杂非线性关系和特征交互方面仍

具有良好表现。

7) 深度神经网络(DNN)。

DNN 是一种基于多层非线性映射的表征学习模型^[34]。通过堆叠全连接层和非线性激活函数, DNN 能够自动提取变量之间的高阶交互特征, 具有较强的函数逼近能力。对于混凝土配合比这类多因素耦合问题, DNN 可作为评估数据扩充效果的重要模型之一。

3.2 训练设置

本工作将机器学习模型的训练过程划分为两种模式: 基准数据模式与数据增强模式。为客观评估 MixDiff 生成数据对模型预测性能的影响, 2 种模式均采用 5 折交叉验证策略。在基准数据模式中, 小样本数据集被划分为 5 个互斥的子集, 每次选取其中一个子集作为验证集, 其余 4 个子集作为训练集, 循环训练 5 次, 以保证每个子集均作为验证集参与一次模型评估。在数据增强模式中, 验证集仍采用真实小样本数据集中的对应子集, 而模型训练集由真实训练样本及 MixDiff 生成的虚拟样本共同构成, 从而评估生成数据对模型训练效果的提升作用。各项性能指标最终取 5 次验证结果的算术平均值。

为保证实验结果的可复现性, 本工作将随机种子固定为 42。模型训练前, 对所有输入特征进行归一化处理, 以降低不同特征量纲差异对模型训练的影响, 并提高模型收敛的稳定性。此外, 训练过程中采用网格搜索方法^[35]对关键超参数进行优化, 以减少人工设定参数带来的主观偏差。

3.3 机器学习模型的评价指标

为量化机器学习模型的预测精度与泛化能力, 本工作采用了 4 个评价指标: 决定系数(R^2)、均方根误差(RMSE)、平均绝对误差(MAE)与平均绝对百分比误差(MAPE)。其定义如下:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (11)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (12)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (13)$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (14)$$

其中: N 为样本数; y_i 为真实值; $\hat{y}_i$ 为对应的预测值; $\bar{y}$ 为真实值的平均值。 R^2 越接近 1, 说明模型

对目标变量方差的解释能力越强; RMSE、MAE 及 MAPE 越接近 0, 说明模型预测误差越小, 预测性能越优。

3.4 实验结果

3.4.1 扩充样本量对模型性能的影响 图 9 展示了 MixDiff 生成虚拟数据集样本量对机器学习模型性能的影响。图中实线表示不同机器学习模型在不同虚拟样本数量下的 R^2 变化趋势, 浅色阴影表示基于 5 折交叉验证结果计算得到的 95% 置信区间。结果表明, 当虚拟数据集样本量处于 100~500 时, 各模型的 R^2 随样本量增加而明显提升, 说明 MixDiff 生成数据能够有效补充小样本数据集中稀疏的特征空间; 当虚拟样本量由 500 进一步增加至 1000 时, 各模型性能仍持续提升, 但提升幅度较前一阶段有所减小, 表明数据扩充带来的性能提升效果开始下降; 在 1000~3000 样本量范围内, 各模型的 R^2 并未完全稳定, 而是仍存在缓慢上升趋势; 当虚拟样本

量进一步增加至 3000 以上时, 各模型的 R^2 基本不再明显提高, 曲线趋于稳定, 且置信区间内的波动较小, 表明继续增加虚拟样本量对模型预测性能的提升作用已较为有限。与此同时, MixDiff 在虚拟数据生成过程中表现出较高的生成效率, 单条虚拟数据的平均生成时间仅为 0.043 7 s, 表明该方法能够在较低时间成本下快速构建较大规模的虚拟数据集。因此, 综合考虑模型预测性能、结果稳定性以及计算成本, 本工作将后续实验中的虚拟数据集样本量设置为 3000。

3.4.2 数据扩充对模型性能的影响 在基准数据模式与数据增强模式下, 前述机器学习模型在网格搜索策略下的最优参数列于表 6 中。

在表 7 与图 10 中展示了 2 种训练模式下不同机器学习模型的预测结果。表 7 中的 R^2 、RMSE 和 MAE 为 6 个输出性能指标的综合评价结果, 统计的目标变量包括 7CS、28CS、S、ED、IST 和 FST。

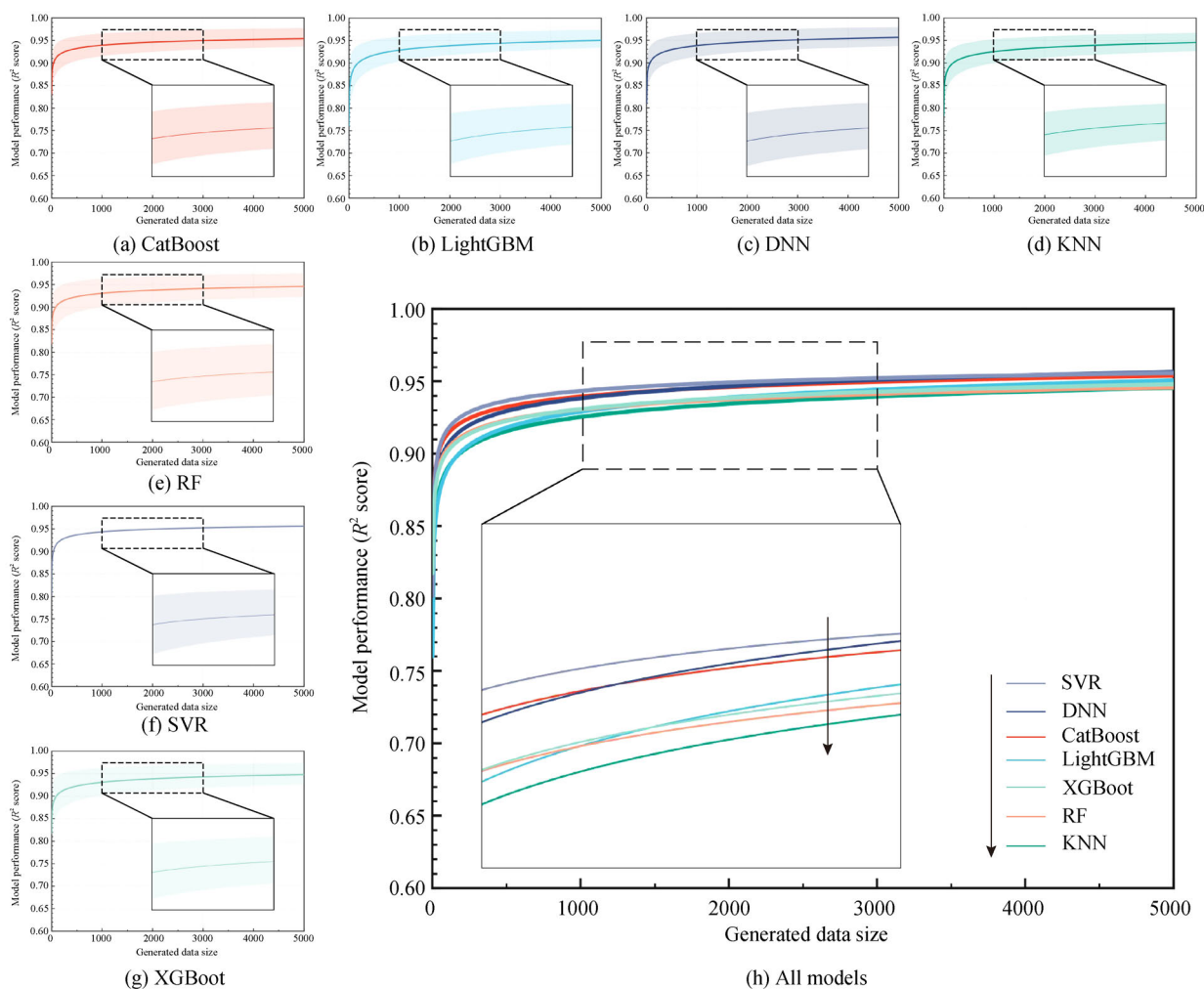


图 9 样本量对模型的影响

Fig. 9 Effect of synthetic sample size on model performance

表 6 机器学习模型最优参数
Table 6 Optimal parameters for machine learning models

Model	Hyperparameter	Search space	Optimal value	
			Baseline	Data augmentation
RF	n_estimators	[100,300,500]	100	500
	max_depth	[None,5,10,20]	None	20
	min_samples_leaf	[1,2,4]	1	2
	max_features	[1.0,"sqrt"]	1.0	1.0
SVR	C	[1,10,50,100]	10	10
	epsilon	[0.01,0.05,0.1]	0.05	0.1
	gamma	['scale','auto']	'scale'	'scale'
KNN	n_neighbors	[3,5,7,9]	3	9
	weights	['uniform','distance']	'distance'	'distance'
	p	[1,2]	1	2
XGBoost	n_estimators	[150,300,500]	300	150
	learning_rate	[0.01,0.05,0.1]	0.05	0.05
	max_depth	[3,5,7]	3	5
	subsample	[0.8,1.0]	0.8	0.8
LightGBM	n_estimators	[150,300,500]	300	300
	learning_rate	[0.01,0.05,0.1]	0.05	0.05
	num_leaves	[15,31,63]	15	15
	subsample	[0.8,1.0]	0.8	0.8
CatBoost	iterations	[150,300,500]	500	500
	learning_rate	[0.01,0.05,0.1]	0.05	0.05
	depth	[4,6,8]	4	6
DNN	hidden_layer_sizes	[(64,32),(128,64,32),(256,128,64)]	(256,128,64)	(256,128,64)
	learning_rate_init	[0.001,0.005,0.01]	0.005	0.001
	alpha	[0.000 1,0.001]	0.001	0.000 1

表 7 模型拟合结果
Table 7 Model fitting results

Model	Data augmentation	R^2	RMSE	MAE
RF	No	0.825 7±0.081 9	13.93±3.29	8.48±1.32
	Yes	0.949 3±0.022 8	7.58±0.96	4.80±0.48
SVR	No	0.805 0±0.053 9	14.38±2.32	8.65±0.89
	Yes	0.957 3±0.016 8	7.15±0.78	4.47±0.24
KNN	No	0.799 6±0.055 6	15.25±1.49	9.45±0.27
	Yes	0.947 6±0.014 0	7.96±0.92	5.01±0.33
XGBoost	No	0.831 7±0.058 1	13.77±2.54	8.49±0.86
	Yes	0.952 6±0.019 1	7.45±0.93	4.65±0.47
LightGBM	No	0.769 6±0.053 5	16.13±1.10	10.12±0.44
	Yes	0.953 6±0.018 2	7.41±1.04	4.71±0.53
CatBoost	No	0.836 8±0.052 7	13.51±2.37	8.13±0.65
	Yes	0.956 1±0.018 5	7.17±0.74	4.49±0.28
DNN	No	0.851 5±0.058 1	12.41±1.99	8.04±1.05
	Yes	0.959 4±0.016 6	7.00±0.71	4.52±0.29

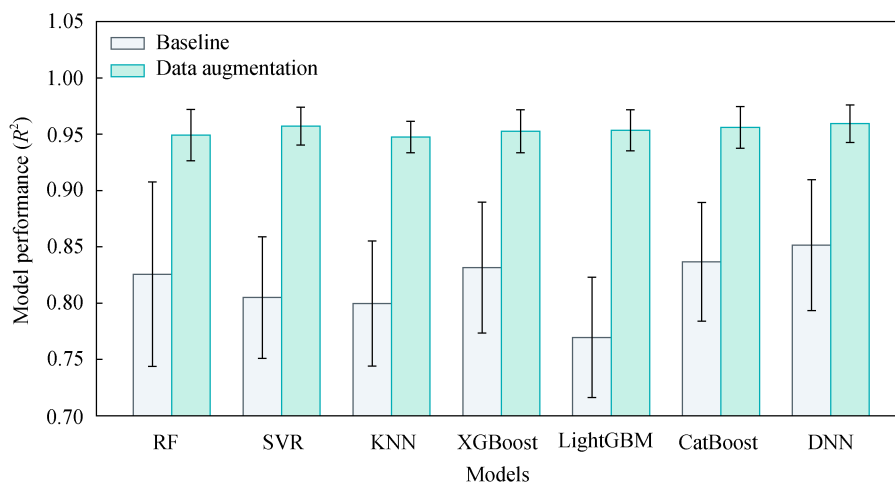


图 10 2 种模式下机器学习模型的拟合精度

Fig. 10 Fitting accuracy of machine learning models under two modes

总体来看, 使用 MixDiff 增强训练集样本数量显著提升了各类模型的预测性能。相比基准数据模式, RF、SVR、KNN、XGBoost、LightGBM、CatBoost、DNN 7 个机器学习模型在数据增强模式下的平均 R^2 由 0.817 1 提升至 0.953 7, 相对上升 16.72%; 平均 RMSE 由 14.20 降低至 7.39, 相对下降 47.96%; 平均 MAE 由 8.77 降低至 4.66, 相对下降 46.86%。这说明 MixDiff 生成的虚拟数据能够有效补充真实小样本数据中的特征空间信息, 从而提高模型对混凝土性能的预测能力。

具体而言, 在基准数据模式下, 由于真实样本数量有限, 各模型的 R^2 主要分布在 0.769 6 至 0.851 5 之间, RMSE 和 MAE 相对较高, 表明模型在小样本条件下难以充分学习复杂的非线性关系。引入 MixDiff 生成数据后, 各模型的 R^2 提升至 0.947 6~0.959 4, RMSE 降低到 7.00~7.96, MAE 降低到 4.47~5.01。该结果表明, 生成数据能够有效降低模型预测误差, 并增强模型对混凝土复杂配合比样本的拟合能力。从不同模型的表现来看, DNN 取得最高 R^2 和最低 RMSE, 其 R^2 和 RMSE 分别为 0.959 4 和 7.00。这说明在扩充数据的支持下, DNN 能够更充分地发挥其非线性表征能力。在图 11 中展示了 DNN 模型训练过程中训练集与验证集的损失变化情况。由图 11 可见, 在训练初期, 训练损失和验证损失均呈现快速下降趋势, 说明模型能够较快学习输入特征与目标变量之间的非线性映射关系。随着训练轮数的增加, 训练集损失与验证集损失的下降速度逐渐减缓, 并在 300 轮后趋于稳定, 表明模型参数更新逐步收敛, 训练过程已基本达到稳定状态。

同时, 验证损失整体变化趋势与训练损失保持一致, 二者之间未出现明显扩大差距。虽然验证损失在部分训练轮次存在轻微波动, 但其整体未表现出持续升高现象, 说明模型在验证集上的泛化性能较为稳定, 未发生过拟合现象。DNN 模型在训练过程中具有较好的收敛性和泛化能力。

与此同时, SVR、CatBoost、XGBoost 和 LightGBM 等模型也表现出较高的预测精度, 说明 MixDiff 生成数据的增益并不依赖于某一特定模型, 而是对多种机器学习方法均具有稳定的促进作用。

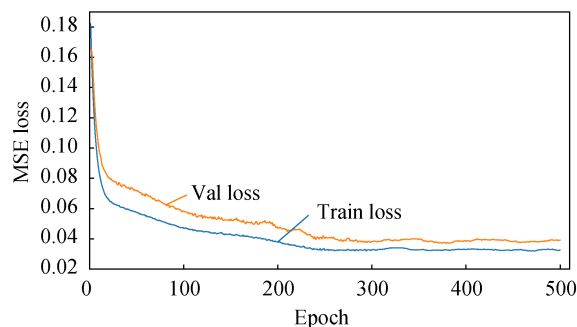


图 11 DNN 在训练过程中的损失变化曲线

Fig. 11 Loss curves during DNN training

此外, 从 5 折交叉验证的标准差可以看出, 数据增强模式下各模型的性能波动整体减小。例如, RF 的 R^2 标准差 0.081 9 降低至 0.022 8, KNN 由 0.055 6 降低至 0.014 0, DNN 由 0.058 1 降低至 0.016 6。这表明引入 MixDiff 生成的虚拟数据后, 模型在不同数据划分下的预测结果更加稳定, 泛化能力得到改善。

为进一步分析数据扩充对最优模型拟合效果的

影响,图 12 对比了 DNN 在两种训练模式下的预测结果。可以看出,在基准数据模式下,部分预测点与理想对角线存在一定偏离,说明小样本条件下模型对目标变量的拟合仍存在误差。而在数据增强模

式下,预测点整体更集中于理想对角线附近,尤其是在样本相对稀疏的区域,模型仍能保持较好的预测一致性。上述结果进一步说明,MixDiff 生成数据能够有效提升机器学习模型的预测精度与稳定性。

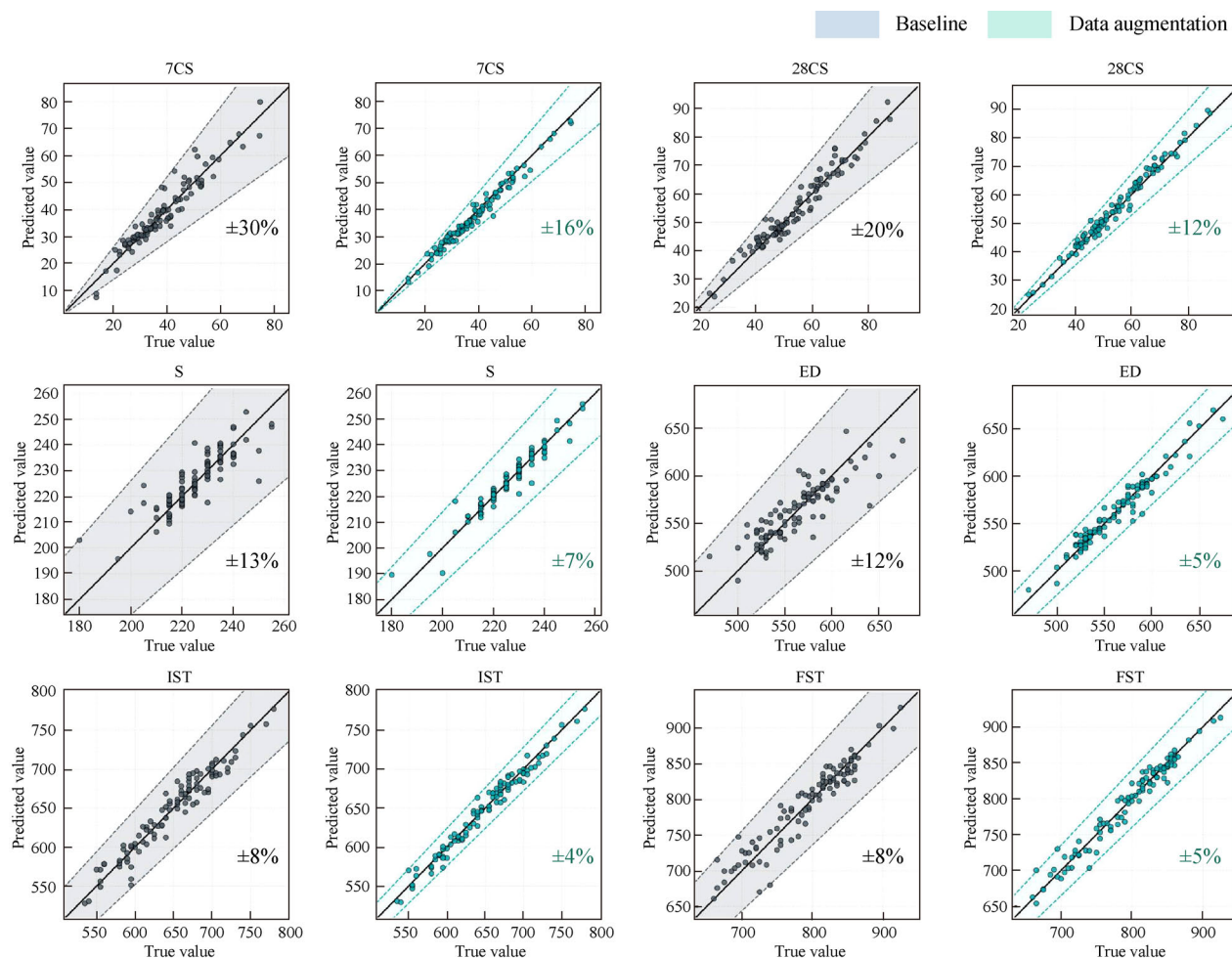


图 12 数据扩充对 DNN 模型预测结果的影响

Fig. 12 Effect of data augmentation on DNN prediction results

4 实验验证

为验证 MixDiff 生成数据的有效性,本工作从虚拟数据集中选取 15 组具有不同抗压强度的配合比,通过试验对其性能进行验证。所选取的配合比数据特征覆盖范围广泛,用于验证 MixDiff 在不同配合比条件下生成数据的合理性。试验过程中所用原材料与 1.1 小节保持一致。所选配合比详情见附录 2,试验结果见附录 3。

图 13 展示了 MixDiff 生成结果与真实试验结果在 7CS、28CS、S、ED、IST 和 FST 等指标上的对比。整体来看,生成结果与试验结果具有较好的一致性,各性能指标的平均 MAPE 均低于 5%,总体平均 MAPE 为 3.8%,说明 MixDiff 生成的配合比数

据在多项关键性能上能够保持较高的合理性。

需要说明的是,MixDiff 模型的适用性受原材料体系影响。当水泥、矿物掺合料、骨料类型等因素发生明显变化时,本工作所训练的模型权重不宜直接用于新材料体系,仍需开展少量代表性试验并重新训练模型。与重新构建大规模人工试验数据库相比,MixDiff 可利用少量试验数据生成虚拟样本并扩充训练集,从而降低试验成本和时间成本,提升小样本条件下机器学习模型的预测精度。对于类似的小样本、多特征、多输出混凝土材料性能预测任务,建议将样本量扩充至 500 组以上,并结合交叉验证结果、性能提升幅度及置信区间变化判断样本规模是否充足。当模型性能趋于稳定且置信区间波动较小时,可认为当前样本量基本满足训练需求。

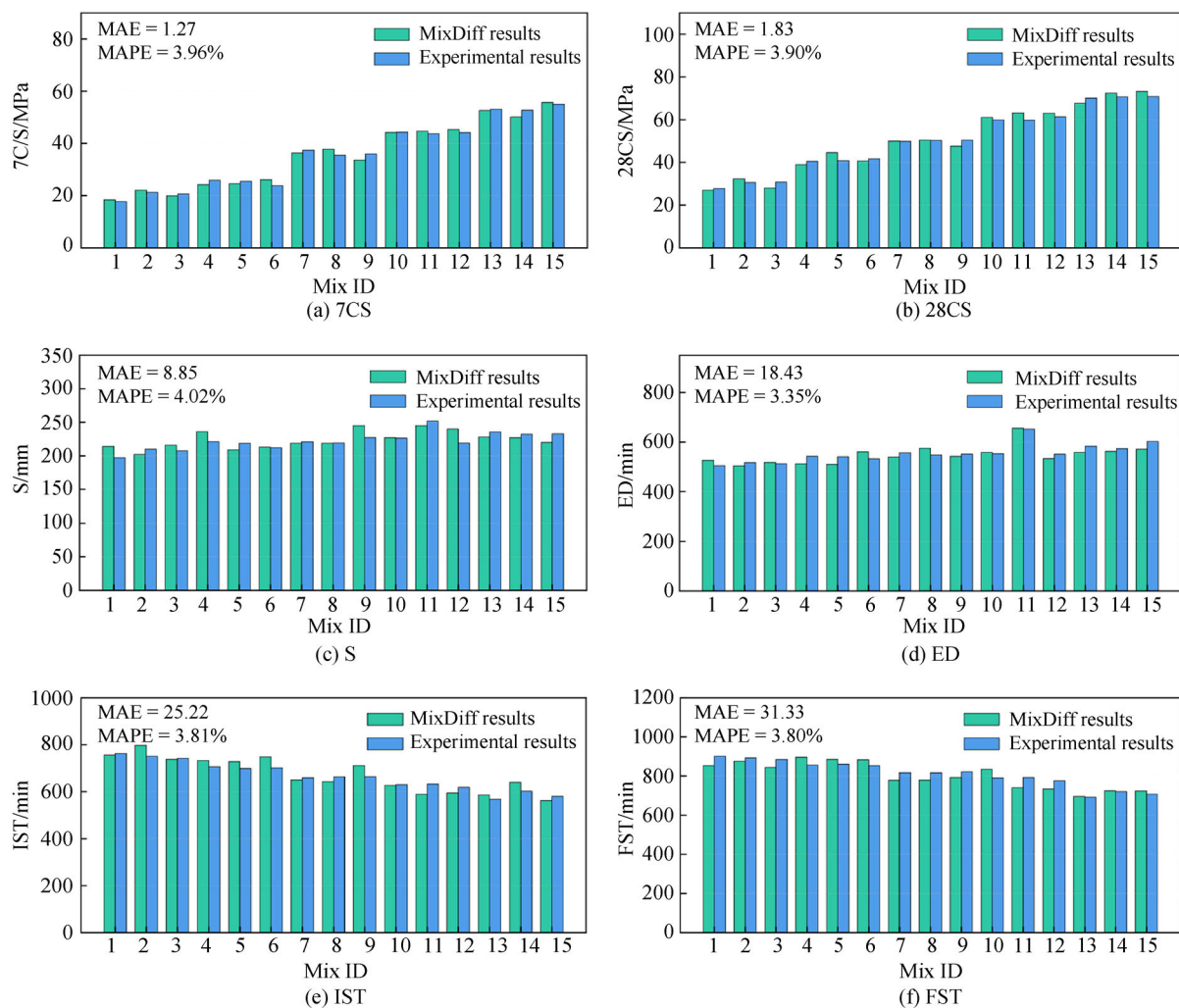


图 13 MixDiff 生成数据的实验结果与生成结果对比

Fig. 13 Comparison of Experimental Results with Generated Results from MixDiff

5 结论

本工作针对混凝土配合比数据获取周期长、成本高以及小样本条件下机器学习模型拟合能力不足的问题,提出了一种面向混凝土配合比数据生成的去噪扩散概率模型 MixDiff。通过分布一致性分析、机器学习建模对比以及试验验证,得到以下结论:

1) MixDiff 模型能够较好学习真实小样本数据的分布规律,生成数据在边缘分布上与真实数据具有较高一致性,WD、KS 与 JSD 指标均保持在较低水平。同时,生成数据能够有效保持特征间的相关性,其 FD 值仅为 0.3327,说明 MixDiff 能够较好反映混凝土配合比特征与性能之间内在关系。

2) MixDiff 生成数据能够有效提升机器学习模型的预测性能。数据规模敏感性实验表明,当虚拟样本量增加至约 3000 时,模型性能趋于稳定。引入

生成数据后,RF、SVR、KNN、XGBoost、LightGBM、CatBoost 及 DNN 7 种算法的预测性能均得到明显提升,所有模型的平均 R^2 均提升至 0.94 以上,相对提升 16.72%,平均 RMSE 相对下降 47.96%,平均 MAE 相对下降 46.86%,其中, DNN 取得最高 R^2 和最低 RMSE, SVR 取得最低 MAE。

3) MixDiff 生成的配合比数据具有较好的合理性。对 15 组生成配合比进行真实物理试验,生成结果与试验结果在 7CS、28CS、S、ED、IST 和 FST 等指标上保持较高一致性,各指标 MAPE 均低于 5%。这表明 MixDiff 能够学习混凝土配合比与宏观性能之间的物理规律,可作为小样本条件下混凝土配合比数据扩充的有效方法。当原材料体系发生变化时, MixDiff 模型虽然需结合少量新材料体系下的代表性试验数据进行微调或重新训练,但 MixDiff 仍可显著减少重新构建大规模试验数据库所需的试

验成本与时间成本。

附加材料

本文的补充材料见网络版(<https://link.cnki.net/urlid/11.2310.tq.20260901.1652.004>)。

(第二十八届中国科协年会学术论文)

参考文献:

- [1] 张剑. 面向多目标优化的超高性能混凝土抗压强度预测与配合比设计研究[J]. 土木工程与绿色建筑, 2026, 2(2): 84–89.
ZHANG Jian. Civ Eng Green Build, 2026, 2(2): 84–89.
- [2] 元强, 夏瑞, 刘易, 等. 基于多目标寻优的低碳混凝土智能设计[J]. 硅酸盐学报, 2026, 54(3): 878–893.
YUAN Qiang, XIA Rui, LIU Yi, et al. J Chin Ceram Soc, 2026, 54(3): 878–893.
- [3] 潘磊, 傅宇方, 王元丰. 粉煤灰混凝土冻融性能智能预测方法[J]. 硅酸盐学报, 2026, 54(2): 793–810.
PAN Lei, FU Yufang, WANG Yuanfeng. J Chin Ceram Soc, 2026, 54(2): 793–810.
- [4] 孟繁星, 许增光, 付涵, 等. 高海拔强紫外环境下花岗岩沥青混凝土老化性能研究[J]. 水力发电学报, 2026, 45(5): 95–108.
MENG Fanxing, XU Zengguang, FU Han, et al. J Hydroelectr Eng, 2026, 45(5): 95–108.
- [5] 马宗方, 戴溢阳, 贺静, 等. 基于机器学习的混凝土 3D 打印沉积线状态优化[J]. 工程设计学报, 2025, 32(4): 488–498.
MA Zongfang, DAI Yiyang, HE Jing, et al. Chin J Eng Des, 2025, 32(4): 488–498.
- [6] ZHANG Y, YUAN X H, ZHANG X P, et al. Prediction of high-performance concrete compressive strength using Decision Tree-Guided Artificial Neural Network Pretraining approach[J]. Eng Appl Artif Intell, 2025, 156: 110828.
- [7] WOO B H, CHOI D H, PARK C H, et al. Gaussian process-based data augmentable deep Bayesian neural network: Application to estimate the stress transfer length in pretensioned concrete beams[J]. Eng Struct, 2026, 348: 121809.
- [8] TIAN J C, WANG X T, LI Y L, et al. Small-sample modelling of early-stage operation of high concrete dams in cold regions based on deep transfer learning[J]. Appl Soft Comput, 2026, 188: 114492.
- [9] QU H L, HU T Y, QU J T, et al. A machine learning model based on GAN-ANN data augmentation for predicting the bond strength of FRP-reinforced concrete under high-temperature conditions[J]. Compos Struct, 2025, 369: 119321.
- [10] LI H R, YANG H F. Predictive model of bond strength between recycled aggregate concrete and rebar after high-temperature based on conditional table generation adversarial networks data augmentation and gene expression programming algorithm[J]. Eng Appl Artif Intell, 2025, 162: 112786.
- [11] ARJOSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, NSW, Australia. PMLR, 2017, 70: 214–223.
- [12] XU L, SKOULARIDOU M, CUESTA-INFANTE A, et al. Modeling tabular data using conditional GAN[C]//Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Red Hook, NY: Curran Associates, 2019: 6594–6604.
- [13] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Red Hook, NY: Curran Associates, 2020: 6840–6851.
- [14] SONG J, MENG C, ERMON S. Denoising diffusion implicit models[C]//International Conference on Learning Representations (ICLR 2021), Vienna, Austria (virtual), 2021.
- [15] SONG Y, SOHL-DICKSTEIN J, KINGMA D P, et al. Score-based generative modeling through stochastic differential equations[C]//Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Red Hook, NY: Curran Associates, 2020: 18483–18494.
- [16] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA. IEEE, 2022: 10674–10685.
- [17] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Netw, 1989, 2(5): 359–366.
- [18] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA. IEEE, 2016: 770–778.
- [19] PEARSON K. VII. Note on regression and inheritance in the case of two parents[J]. Proc R Soc Lond, 1895, 58(347–352): 240–242.
- [20] SOHL-DICKSTEIN J, WEISS E A, MAHESWARANATHAN N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//Proceedings of the 32nd International Conference on Machine Learning - Volume 37. ACM, 2015: 2256–2265.
- [21] NICHOL A Q, DHARIWAL P. Improved denoising diffusion probabilistic models[C]//Proceedings of the 38th International Conference on Machine Learning (ICML 2021). PMLR, 2021, 139: 8162–8171.
- [22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, California, USA. Red Hook, NY: Curran Associates Inc., 2017: 5998–6008.
- [23] ELFWING S, UCHIBE E, DOYA K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning[J]. Neural Netw, 2018, 107: 3–11.
- [24] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on Machine Learning (ICML 2015). PMLR, 2015, 37: 448–456.
- [25] 刘悦, 邹欣欣, 杨正伟, 等. 材料领域知识嵌入的机器学习[J]. 硅酸盐学报, 2022, 50(3): 863–876.
LIU Yue, ZOU Xinxin, YANG Zhengwei, et al. J Chin Ceram Soc, 2022, 50(3): 863–876.
- [26] 杨斌, 李悦, 林辉, 等. 荷载-离子-温度共同作用下地下混凝土损伤程度分析与预测[J]. 硅酸盐学报, 2026, 54(2): 653–665.
YANG Bin, LI Yue, LIN Hui, et al. J Chin Ceram Soc, 2026, 54(2): 653–665.
- [27] 张文生, 曹傅荔, 郅晓, 等. 机器学习方法用于水泥基材料的研究进展[J]. 硅酸盐学报, 2024, 52(11): 3617–3630.
ZHANG Wensheng, CAO Fuli, ZHI Xiao, et al. J Chin Ceram Soc, 2024, 52(11): 3617–3630.
- [28] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE

- Trans Inform Theory, 1967, 13(1): 21–27.
- [29] 李九阳, 陈立, 吴纪曙, 等. 基于可解释性机器学习算法的 FRP 筋 UHPC 粘结强度预测[J]. 辽宁工程技术大学学报(自然科学版), 2026, 45(2): 185–195.
- LI Jiuyang, CHEN Li, WU Jishu, et al. J Liaoning Tech Univ Nat Sci Ed, 2026, 45(2): 185–195.
- [30] BREIMAN L. Random forests[J]. Mach Learn, 2001, 45(1): 5–32.
- [31] 李华, 古金本, 陶毅. 基于机器学习预测纤维增强聚合物(FRP)筋-混凝土黏结强度的研究[J]. 工业建筑, 2025, 55(10): 102–114.
- LI Hua, GU Jinben, TAO Yi. Ind Constr, 2025, 55(10): 102–114.
- [32] 郭云杰, 刘俊乐, 孙虎, 等. 基于 LightGBM 算法的高性能混凝土抗压强度预测研究[J]. 建筑机械, 2025(3): 168–173.
- GUO Yunjie, LIU Junle, SUN Hu, et al. Constr Mach, 2025(3): 168–173.
- [33] PROKHORENKOVA L, GUSEV G, VOROBIEV A, et al. CatBoost: unbiased boosting with categorical features[C]//Advances in Neural Information Processing Systems 31 (NeurIPS 2018). Red Hook, NY: Curran Associates, 2018: 6638–6648.
- [34] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436–444.
- [35] YANG B, LI Y, LIN H, et al. Multi-objective design method of concrete mix ratio based on machine learning: Forward prediction and reverse design[J]. Arch Civ Mech Eng, 2026, 26(2): 57.

作者贡献声明:

梁 涛: 论文构思、模型算法、撰写初稿

李 悦: 提供资源、获取资助、调查研究、数据管理、项目管理、监督指导

杨 斌: 模型算法、审阅与修订

刘江林: 项目管理、审阅与修订、分析数据并推导结论

李亚强: 监督指导、审阅与修订、验证

梁美洁: 绘制图表、软件使用

A Concrete Mix Proportion Data Augmentation Method Based on Denoising Diffusion Probabilistic Models

LIANG Tao¹, LI Yue¹, YANG Bin¹, LIU Jianglin¹, LI Yaqiang², LIANG Meijie¹

(1. Key Laboratory of Urban and Engineering Safety and Disaster Mitigation, Ministry of Education, Beijing University of Technology, Beijing 100124, China; 2. College of Soil and Water Conservation, Beijing Forestry University, Beijing 100083, China)

Extended Abstract

Introduction Machine learning has been widely applied to various concrete materials tasks, such as macroscopic performance prediction and mix proportion optimization. The performance of machine learning models depends strongly on the quantity and quality of the available datasets. However, constructing large-scale and high-quality concrete mix proportion datasets remains challenging because concrete experiments are time-consuming, costly, and easily affected by material sources, curing conditions, and human operations. At present, small-sample modeling problems are mainly addressed through two strategies: model optimization and data augmentation. Model optimization methods mainly improve the extraction of information from existing small-sample datasets by adjusting model structures or training strategies. However, they do not fundamentally increase the amount of data, and their improvement in prediction accuracy is still limited by the original sample size. Conventional data augmentation methods mainly rely on generative adversarial networks, which learn the real data distribution through adversarial training between a generator and a discriminator, thereby generating virtual samples to expand the original dataset. Nevertheless, when dealing with complex tabular data, generative adversarial networks may encounter unstable training, insufficient sample diversity, and distribution shifts in generated data, resulting in limited consistency between virtual and real data. To address these issues, this study proposes MixDiff, a denoising diffusion probabilistic model (DDPM) for generating virtual concrete mix proportion data under small-sample conditions. This model aims to generate high-quality virtual concrete mix proportion data.

Methods A small-sample dataset containing 100 sets of concrete mix proportions was first established. The input features included water, cement, fly ash, mineral fines, river sand, manufactured sand, coarse aggregate, water-reducing agent, total binder content, water-to-binder ratio, total fine aggregate content, and sand ratio. The target variables included 7 d compressive strength, 28 d compressive strength, slump, expansion degree, initial setting time, and final setting time. MixDiff was developed using a fully connected neural network and a denoising diffusion probabilistic model. The model consisted of a sinusoidal timestep embedding module and four stacked residual fully connected blocks. Each residual block contained fully connected layers, batch normalization layers, and SiLU activation functions. In the forward diffusion process, Gaussian noise was progressively added to the normalized data according to a cosine variance schedule. In the reverse generation process, a neural network was used to predict the noise and progressively reconstruct concrete mix proportion samples with realistic distribution characteristics from standard Gaussian noise. Subsequently, the quality of virtual data generated by MixDiff was quantitatively evaluated from two perspectives: marginal distribution and feature correlation. Under five-fold cross-validation and grid search, several mainstream machine learning models were further used to compare the prediction performance of the baseline mode and the augmentation mode. Finally, 15 mix proportions were selected from the generated virtual dataset for experimental validation to evaluate the engineering plausibility of the generated data.

Results and discussion The quantitative evaluation results showed that the virtual dataset generated by MixDiff showed high similarity to the real dataset. The average Wasserstein distance, K-S statistic, and J-S divergence between the two datasets were 0.019 5, 0.095 2, and 0.120 7, respectively. The Frobenius distance between their Pearson correlation matrices was 0.332 7, indicating that MixDiff effectively preserved the marginal distribution characteristics of the real data and the correlation structure among variables. The machine learning comparison results showed that the prediction performance of all models improved significantly after the introduction of MixDiff-generated data. In the baseline mode, the average R^2 of the seven models was 0.817 1, and increased to 0.953 7 in the data augmentation mode. Meanwhile, the average RMSE decreased from 14.20 to 7.39, and the average MAE decreased from 8.77 to 4.66. Correspondingly, R^2 increased by 16.72%, while RMSE and MAE decreased by 47.96% and 46.86%, respectively. Among the machine learning models, the deep neural network achieved the best performance after data augmentation, with R^2 , RMSE, and MAE values of 0.959 4, 7.00, and 4.52, respectively. In addition, the standard deviations of the five-fold cross-validation results were significantly reduced in the data augmentation mode. This indicates that MixDiff-generated data not only improved the prediction accuracy of machine learning models but also enhanced model stability and generalization ability. Experimental validation of 15 virtual mix proportions showed that the data generated by MixDiff had high plausibility in six key performance indicators, including 7 d compressive strength, 28 d compressive strength, slump, expansion degree, initial setting time, and final setting time. The average MAPE of each performance indicator was lower than 5%, and the overall average MAPE was 3.8%.

Conclusions This study proposed MixDiff, a denoising diffusion probabilistic model for concrete mix proportion data generation. The results show that the virtual data generated by MixDiff were highly consistent with the real data in terms of marginal distribution and feature correlation. The Frobenius distance between the Pearson correlation matrices was only 0.332 7. After introducing MixDiff-generated data, the prediction performance of multiple machine learning models, including RF, SVR, KNN, XGBoost, LightGBM, CatBoost, and DNN, was significantly improved, with R^2 values of all models increasing above 0.94. Based on the average performance of the seven machine learning models for the six output performance indicators, R^2 increased by 16.72%, while RMSE and MAE decreased by 47.96% and 46.86%, respectively. Among these models, DNN achieved the highest R^2 and the lowest RMSE in the data augmentation mode. Experimental validation of 15 virtual mix proportions showed that the generated results were highly consistent with the measured results in terms of 7 d compressive strength, 28 d compressive strength, slump, expansion degree, initial setting time, and final setting time. The MAPE of each performance indicator was below 5%, and the overall average MAPE across the six performance indicators was 3.8%. These results indicate that MixDiff can learn the underlying relationships between concrete mix proportions and macroscopic properties, and it can serve as an effective method for concrete mix proportion data augmentation under small-sample conditions.

Keywords concrete; mixture proportion; denoising diffusion probabilistic model; data augmentation; deep learning