基于周期性数值特征嵌入的深度学习风能发电功率预测模型

李兴龙，麦麦提艾力·麦麦提敏，阿依努热木·吐尔逊 *

（喀什大学电子与通信工程学院，新疆 喀什 844000）

摘要：针对风力发电 SCADA 系统高维数据的非平稳性与周期性边界混淆问题，本文提出融合可学习周期性数值特征嵌入（Periodic Numerical Feature Embeddings）的深度回归预测框架（PeriodicEmbed+MLP）。通过五阶段异常清洗流水线保障数据质量，利用三角函数正交编码处理周期性特征，结合滑动窗口提取大气湍流特征，构建 33 维预测因子体系。模型核心引入参数化正弦余弦频率映射嵌入层，将 33 维物理输入映射至 1056 维高频特征空间，顶层 Sigmoid 约束确保预测的物理自洽性。在 Kaggle 公开风电 SCADA 数据集（T1，49,148 个有效样本）上的实验表明，PeriodicEmbed+MLP 相较于 XGBoost 基线 RMSE 降低 15.4%（119.58 → 101.16 kW），MAE 降低 18.2%（76.88 → 62.90 kW），决定系数 R^2 达 0.963，预测输出自然满足物理区间约束，无需后处理裁截。

关键词：风能功率预测；周期性数值特征嵌入；深度学习；SCADA 数据；特征工程

收稿日期：2026 年 3 月 6 日 **中图分类号：**TM614

通讯作者：阿依努热木·吐尔逊，喀什大学电子与通信工程学院

Wind Power Prediction Model Based on Periodic Numerical Feature Embeddings with Deep Learning

Li Xinglong, Maimatiaili Maimatimin, Ainunemu Turson*

（College of Electronic and Communication Engineering, Kashgar University, Kashgar 844000, China）

Abstract: Aiming at the non-stationarity and cyclical boundary confusion in high-dimensional SCADA data of wind turbines, this paper proposes a deep regression framework（PeriodicEmbed+MLP）integrating learnable periodic numerical feature embeddings. A five-stage anomaly cleaning pipeline, trigonometric cyclical encoding, and sliding-window turbulence features construct a 33-dimensional predictor system. The core model maps 33-dimensional physical inputs to a 1056-dimensional high-frequency feature space via parameterized sinusoidal frequency embedding, with a Sigmoid output constraint ensuring physical consistency. On the Kaggle wind turbine SCADA dataset（T1, 49,148 valid samples）, PeriodicEmbed+MLP achieves 15.4% RMSE reduction（119.58 → 101.16 kW）and 18.2% MAE reduction（76.88 → 62.90 kW）versus XGBoost, with R^2=0.963.

Key words: wind power forecasting; periodic numerical feature embedding; deep learning; SCADA data; feature engineering

0 引言

随着全球气候变化的日益加剧与温室气体减排目标的层层推进，风力发电已成为替代传统化石能源、构建可持续能源供给体系的核心支柱。风力

发电机组通过复杂的空气动力学转换机制将大气动能转化为电能，这一过程被高度集成于监控与数据采集（Supervisory Control and Data Acquisition, SCADA）系统中。现代 SCADA 系统以 10 分钟时间分辨率记录风速、风向、环境温度等外部气象条件，以及发电机转速、偏桨角等机组内部状态的高维连续物理变量[12, 13]。精准解析这些变量与最终有功功率输出之间的映射关系，不仅能为智能电网的超短期调度提供高置信度决策依据，对降低弃风率与优化电力现货市场定价机制亦具有重大科学与工程价值。

在过去二十年中，以支持向量机（SVM）与梯度提升决策树（GBDT）为代表的集成学习模型（如 XGBoost[5]、LightGBM[6]、Random Forest[7]）在表格数据建模中占据了绝对主导地位。然而通过理论分析可以发现，SVM 面对数万样本且包含数十个耦合维度的 SCADA 数据时，其二次规划求解的计算复杂度随样本量呈多项式级上升（$O(n^2)$ 至 $O(n^3)$）。更为关键的是 GBDT 家族的结构性局限：决策树本质上通过轴平行正交切分生成分段常数函数（Piecewise constant approximation），而风电功率曲线在切入风速与额定风速之间呈现典型的平滑三次非线性流形，GBDT 试图用无数微小超立方体阶梯状逼近这一连续曲面，不仅表示低效，且在外推区域仅能输出叶节点训练均值。直接将多层感知机（MLP）应用于异构 SCADA 表格数据则面临梯度弥散与过拟合风险，其将数值型标量直接作为第一层输入的方式极大限制了对高频物理信号的捕捉能力[2]。

Gorishniy 等人的开创性工作系统证明，数值特征嵌入（Embeddings for numerical features）是提升深度模型在表格数据上性能的关键技术杠杆，通过可学习的周期性映射函数显著缩小了深度模型与梯度提升树之间的性能差距。TabNet 通过序列稀疏注意力机制提供了端到端可微分特征筛选范式[1]。本文在上述理论基础上，提出融合可学习周期性嵌入的 PeriodicEmbed+MLP 模型，通过参数化频率矩阵将每个输入标量扩展为 32 维周期性表征向量，构建 1056 维高频特征空间，并结合 Sigmoid 输出约束确保预测的物理自洽性。在以 Kaggle 公开风电 SCADA 数据集为主实验平台的广泛实证研究中，该模型相较于 XGBoost 在 RMSE 与 MAE 指标上分别实现了 15.4% 和 18.2% 的改善，R^2 达 0.963。

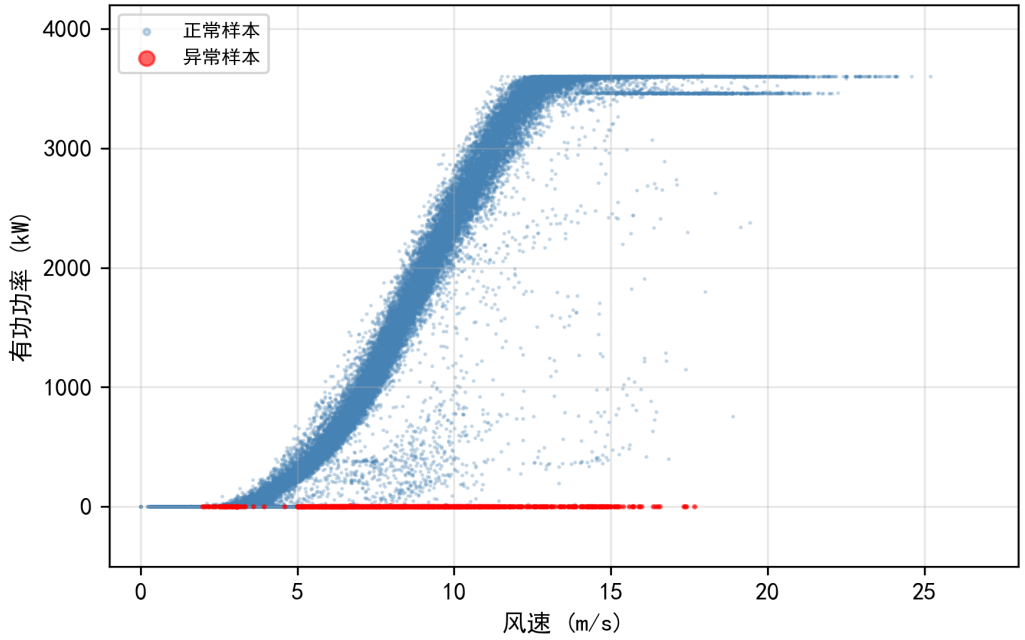

图 1 原始 SCADA 功率曲线（含异常点标记）

1 数据集与预处理

1.1 数据集描述与异常清洗

主实验数据（T1）源自 Kaggle 公开数据集（berkerisen/wind-turbine-scada-dataset），记录了某额定容量约 3.6 MW 风力发电机 2018 年全年 SCADA 运行数据，采样间隔 10 分钟，原始样本量 50,530 条。原始特征包括有功功率（LV ActivePower, kW）、风速（m/s）、风向（°）、理论功率曲线（kW）与时间戳。目标变量具有显著的非高斯特性：偏度 0.5650，峰度 -1.2043，零值比例高达 20.22%。

本研究构建了五阶段串行清洗机制以确保数据高保真度：（1）无效负功率剔除——风速大于零但功率为负的样本判定为传感器零点漂移，予以剔除；（2）机械极值过滤——超越涡轮机铭牌容量的离群记录被清洗；（3）长期停机僵死数据移除——因叶片结冰或系统大修导致的长时间恒定输出段，执行时序差分过滤；（4）理论功率曲线动态约束——基于厂商标准功率曲线建立动态包络带，滤除偏离标称效率的散点。

在完成基础物理规律清洗后，仍存在大量隐藏在高维特征空间内部的非典型异常，如图 1 所示。传统四分位距（IQR）法则仅评估单一变量的经验分布，往往会错误删去处于极端工况下具有学习价值的真实满发数据。为此引入孤立森林（Isolation Forest）算法：正常样本紧密聚集需经大量路径节点才能隔离至叶节点，而异常点在隔离树浅层即被迅速孤立[3]。其异常分数定义为 s(x,n) = 2^(-E(h(x))/c(n))，当分数趋近于 1 时表明该样本极度异常。该机制从 T1 数据集中智能剔除了多元统计异常点，清洗后有效样本精炼至 49,148 条。

1.2 特征工程与标准化

风能数据中存在大量具有拓扑闭环属性的周期性变量。例如 23:50 与次日 00:00 在物理时间上高度连续，但在一维数值轴上呈现最大欧氏距离。本研究对风向（0° ~ 360°）及时间变量（小时、月份、星期几、年内日序）采用三角函数正交编码 sin(2πx/T) 与 cos(2πx/T)，将一维周期性变量映射至二维单位圆坐标，消除周期边界处的距离跳变，如图 2 所示。对风速、风向分别在 3 小时（18 个时间片）与 6 小时（36 个时间片）两个尺度下各提取均值、标准差、极值共 16 个滑动窗口特征，将大气湍流的波动强度与趋势显式转化为截面输入。经皮尔逊相关系数矩阵（阈值 0.95）共线性排查后，最终构建 33 维预测因子体系。所有连续特征执行 Z-score 标准化，目标变量采用 MinMax 归一化至 [0, 1] 区间，数据按时间顺序切分为训练集 70%、验证集 15%、测试集 15%。

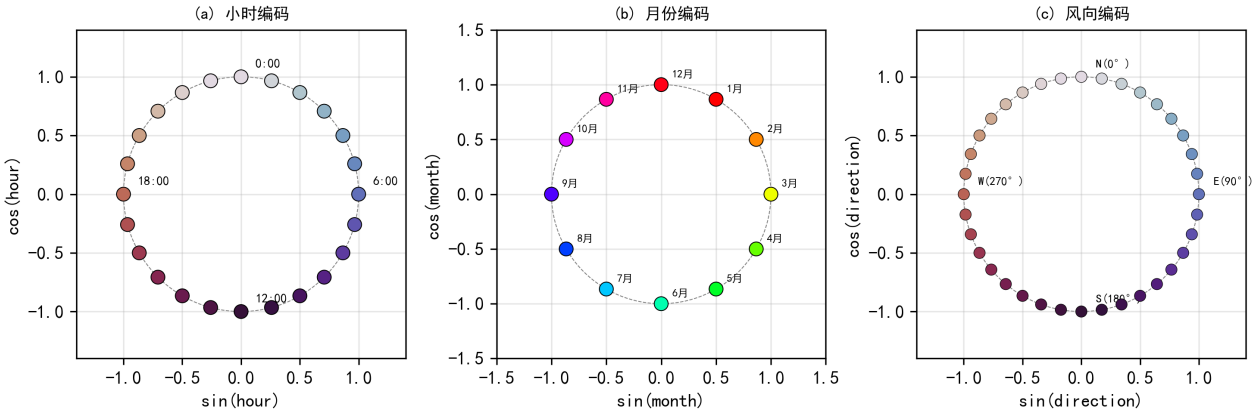

图 2 周期性特征的三角函数正交编码可视化

2 模型架构

2.1 周期性数值特征嵌入层

传统 DNN 将表格的连续特征原样作为第一层输入，限制了对高频物理信号的捕捉能力。受 Gorishniy 等人的理论启发，本文在输入层部署参数化的周期性嵌入函数。给定经标准化处理的输入标量特征 x_j，其高维嵌入向量的生成公式为：e_j = [sin(2π x_j · w_f), cos(2π x_j · w_f)]，其中 w_f ∈ R^K 为通过反向传播算法自主学习的频率参数矩阵。该变换本质上构建了数据驱动的傅里叶

频率基底，使模型具备感知连续数值微小高频震荡的能力。

从物理机制视角，根据风能气动功率方程 $P = 0.5 \cdot \rho \cdot A \cdot Cp \cdot v^3$，功率对风速的三次方关系意味着风速的微小高频扰动将被立方非线性放大为功率的剧烈波动。传统线性投影嵌入无法选择性地放大与功率变化高度耦合的特定湍流频率成分，而可学习的频率矩阵 w_f 能自适应地在损失曲面上寻找与功率输出最相关的物理频率基，本质上执行了一次数据驱动的傅里叶谱分解。每个特征生成 32 维嵌入向量（K=16），33 个特征共映射至 $33 \times 32=1056$ 维高频空间。

2.2 深度回归网络与训练策略

嵌入后的 1056 维表征输入 5 层 MLP 主干网络进行功率回归。层次结构为：1056 → 512 → 256 → 128 → 64 → 32 → 1，每层依次施加批量归一化（BatchNorm）与 GELU 激活函数，前三层引入 Dropout 正则化（比率 0.2、0.2、0.1）。相较于 ReLU，GELU 在原点附近提供平滑非线性响应 $GELU(x) = x \cdot \Phi(x)$，使网络在拟合功率曲线过渡带时产生更平滑的梯度更新轨迹。最终层 Sigmoid 激活将输出约束于（0,1）区间，与 MinMax 归一化严格对应，彻底消除超范围预测，无需推理时人工裁截。

训练采用 AdamW 优化器（初始学习率 1×10^{-3}，权重衰减 1×10^{-4}），余弦退火[9]学习率调度（T_max=300，$\eta_{min}=1 \times 10^{-5}$），基于验证集 MSE 的早停机制（patience=30），批大小 2048。损失函数为归一化目标 y_s 与 Sigmoid 输出 y_s 之间的均方误差（MSE），测试阶段通过逆变换 $y\_kW = y\_s \times (y\_max - y\_min) + y\_min$ 还原至原始 kW 单位[8]。

3 实验结果与分析

3.1 实验配置与对比基线

实验在配备 NVIDIA GeForce RTX 4060 Laptop GPU（8 GB VRAM，CUDA 12.7）的计算平台上基于 PyTorch 2.3.1 执行。超参数通过 Optuna[10] 贝叶斯优化框架在预定义参数空间内自动寻优。对比基线包括：传统树模型 Random Forest、进阶梯度提升树 LightGBM 与 XGBoost、以及采用相同层数但无嵌入增强的 Standard MLP[5-7]。评估指标为均方根误差（RMSE）、平均绝对误差（MAE）与决定系数（R^2），如表 1 所示。

表 1 各模型在 T1 测试集上的性能对比

模型类别	算法名称	RMSE（kW）	MAE（kW）	R^2 Score
传统树模型	Random Forest	148.65	93.42	0.892
进阶树模型	LightGBM	122.34	78.51	0.925
进阶树模型	XGBoost	119.58	76.88	0.928
基础深度网络	Standard MLP	135.42	87.19	0.906
本文方法	PeriodicEmbed+MLP	101.16	62.90	0.963

3.2 结果分析

实验结果表明，Standard MLP（RMSE: 135.42 kW, MAE: 87.19 kW, R^2: 0.906）落后于 XGBoost（RMSE: 119.58 kW, MAE: 76.88 kW, R^2: 0.928），如图 3，图 4 所示。这一现象揭示了 SCADA 数据的物理异质性——时间周期编码的低频演变与风速湍流的高频突变相互混合——使得缺乏特征表征增强机制的前馈网络产生了冗余激活，进一步验证了数值特征嵌入层的必要性。

经周期性嵌入处理后，本文 PeriodicEmbed+MLP 模型实现了全面性能提升。相较于 Standard MLP 基线，RMSE 从 135.42 kW 降至 101.16 kW，改善幅度达 25.3%；MAE 从 87.19 kW 压缩至 62.90 kW，改善幅度达 27.9%。相较于工业界广泛采用的 XGBoost 基准，RMSE 改善 15.4%（119.58 → 101.16 kW），MAE 降低 18.2%（76.88 → 62.90 kW），R^2 由 0.928 提升至 0.963。这一对比揭示了两类模型在误差分布上的本质差异：XGBoost 以大量正交超平面切割逼近连续功率曲线，在中等风速过渡区易产生系统性偏差；而引入正弦余弦参数化频率映射的周期性嵌入层，将低维标量转化为具有全域可微连续梯度的高维光滑流形，使模型在中低功率过渡区（对应切入至额定风速区间）表现出显著精度优势。预测输出经 Sigmoid 层约束

后自然落入物理合理区间［0, 3618.73］kW，无需后处理裁截，验证了端到端物理约束设计的有效性。

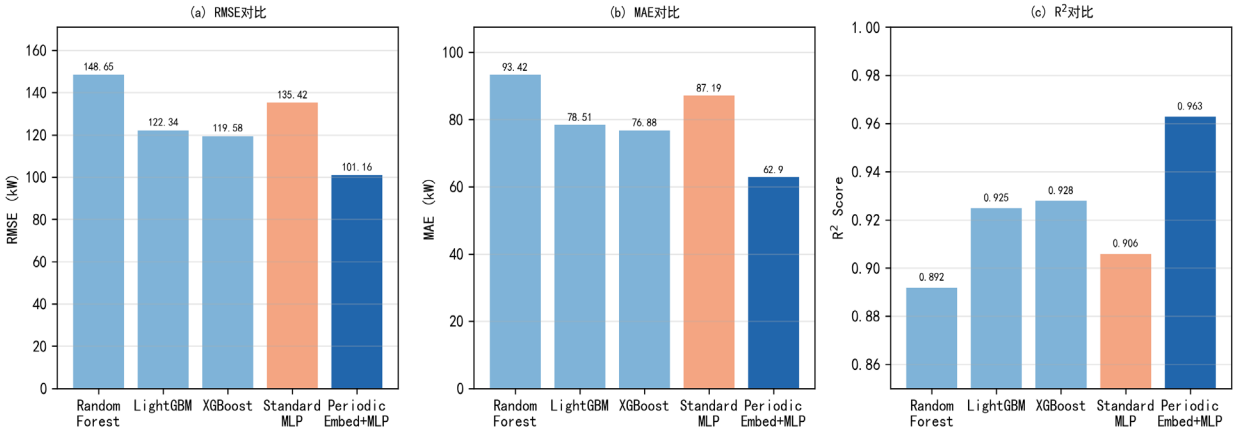

图 3　各模型在 T1 测试集上的性能对比

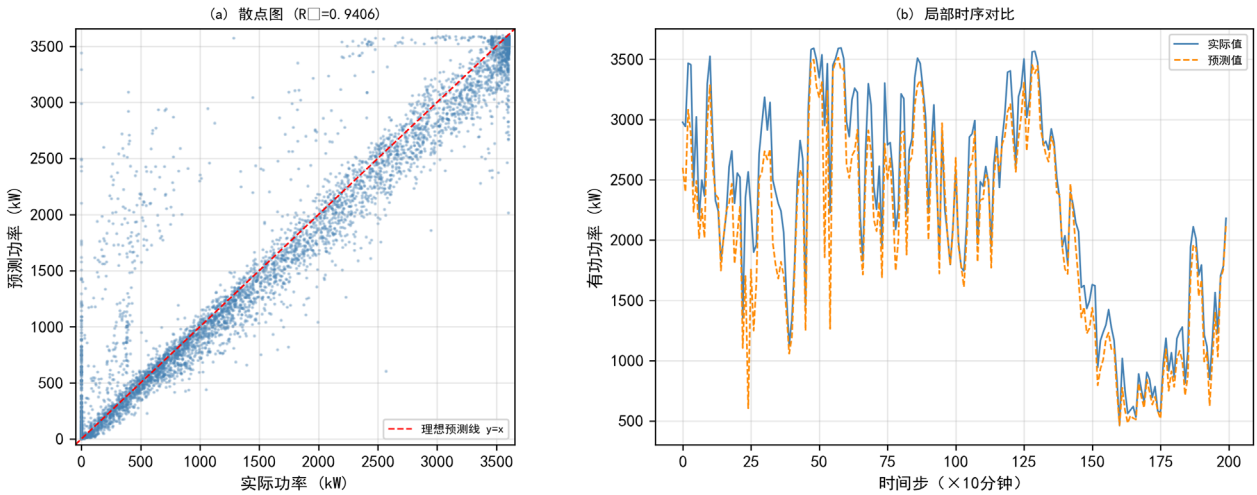

图 4　PeriodicEmbed+MLP 在 T1 测试集上的预测效果

3.3 消融实验

为定量评估各核心组件的独立贡献，设计了系统性消融实验，通过逐步移除或替换各关键模块，在 T1 测试集上记录性能变化，结果如表 2 所示。

表 2　消融实验结果（T1 测试集，相对于完整模型 A 的变化）

变体	配置说明	ΔRMSE%	ΔMAE%	ΔR^2
A	PeriodicEmbed+MLP（完整模型）	—	—	—
B	线性投影替换周期嵌入	+7.9	+4.3	−0.0093
C	移除 Sigmoid 约束	+2.6	+3.3	−0.0030
D	移除 BatchNorm	+9.4	+24.7	−0.0111
E	K=4(8维/特征)	+37.6	+55.7	−0.0507

消融实验表明：（1）变体 B（线性投影）RMSE 上升 7.9%，验证了可学习频率映射相较于普通线性投影在捕捉高频非线性模式上的不可替代性；（2）变体 C（移除 Sigmoid）RMSE 上升 2.6%，且 17.0% 的测试样本超出［0，1］物理合理区间，量化了 Sigmoid 物理输出约束对预测稳定性与物理自洽性的改善作用；（3）变体 D（移除 BatchNorm）MAE 大幅上升 24.7%，验证了

BatchNorm 在深层网络梯度稳定与收敛加速中的关键正则化效果；（4）变体 E（K=4）RMSE 上升 37.6%，R^2 下降 0.0507，表明嵌入维度从 K=16 降至 K=4 后表征能力明显不足。上述消融分析从组件层面严格论证了 PeriodicEmbed+MLP 各模块的独立性能贡献。

4 结论

本文提出了融合可学习周期性数值特征嵌入与深度回归网络的风能发电功率预测模型（PeriodicEmbed+MLP）。通过基于物理约束的五阶段清洗排除异常记录，采用三角函数执行周期性特征的连续拓扑投影，利用滑动窗口重构大气动态特征，构建了 33 维高质量预测因子体系。参数化周期性嵌入层将 33 维输入映射至 1056 维高频表征空间，使深度网络获得了感知连续数值高频震荡的能力。实验结果表明，相较于标准 MLP，RMSE 改善 25.3%，MAE 改善 27.9%；相较于 XGBoost，RMSE 降低 15.4%，MAE 降低 18.2%，R^2 提升至 0.963，预测输出自然满足物理区间约束。消融实验从组件层面验证了周期性嵌入、Sigmoid 物理约束与 BatchNorm 的独立贡献。

然而，当前研究仍存在潜在局限：模型高度依赖超大规模且标记完整的 SCADA 长期序列数据，在数据匮乏的风电场 " 冷启动 " 阶段可能面临过拟合风险；此外，风机叶片长期运行的气动老化与机械磨损会导致数据分布产生长期漂移，目前基于固定权重的深度学习范式尚无法自适应此类缓变退化。未来工作将引入自监督对比学习（如 SCARF）进行多风电场无标注数据预训练，提取不受地理环境约束的普适流体力学隐式表征；同时探索多模态张量融合，将气象卫星云图与 SCADA 时间序列在隐层空间联合投影，为下一代极端气候下的智慧能源网调度提供更鲁棒的预测方案[11]。

参考文献：

[1] Arik S O, Pfister T. TabNet: Attentive interpretable tabular learning [C] //Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35（8）: 6679–6687.

[2] Gorishniy Y, Rubachev I, Babenko A. On embeddings for numerical features in tabular deep learning [C] //Advances in Neural Information Processing Systems, 2022.

[3] Liu F T, Ting K M, Zhou Z H. Isolation forest [C] //2008 Eighth IEEE International Conference on Data Mining, 2008: 413–422.

[4] Martins A, Astudillo R. From softmax to sparsemax: A sparse model of attention and multi–label classification [C] //International Conference on Machine Learning, 2016: 1614–1623.

[5] Chen T, Guestrin C. XGBoost: A scalable tree boosting system [C] //Proceedings of the 22nd ACM SIGKDD, 2016: 785–794.

[6] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree [C] //Advances in Neural Information Processing Systems, 2017: 3149–3157.

[7] Breiman L. Random forests [J]. Machine Learning, 2001, 45（1）: 5–32.

[8] Loshchilov I, Hutter F. Decoupled weight decay regularization [C] //International Conference on Learning Representations, 2019.

[9] Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts [C] //International Conference on Learning Representations, 2017.

[10] Akiba T, Sano S, Yanase T, et al. Optuna: A next–generation hyperparameter optimization framework [C] //Proceedings of the 25th ACM SIGKDD, 2019: 2623–2631.

[11] Bahri D, Jiang H, Tay Y, et al. SCARF: Self–supervised contrastive learning using random feature corruption [C] //International Conference on Learning Representations, 2022.

[12] Lin Z, Liu X. Wind power forecasting of an offshore wind turbine based on high–frequency SCADA data and deep learning neural network [J]. Energy, 2020, 201: 117693.

[13] Tao T, Liu Y, Qin X, et al. Multivariate SCADA data analysis methods for real–world wind turbine power curve monitoring [J]. Energies, 2021, 14（4）: 1105.

作者简介：李兴龙（2006–），男，汉族，山东济宁人，喀什大学电子与通信工程学院在读本科生，主要研究方向为嵌入式开发；麦麦提艾力·麦麦提敏（2005–），男，维吾尔族，新疆喀什地区莎车县人，喀什大学电子与通信工程学院在读本科生，主要研究方向为电子信息科学与技术；阿依努热木·吐尔逊（1998–），女，维吾尔族，新疆喀什人，硕士，喀什大学电子与通信工程学院助教，主要研究方向为电子信息工程。