基于 LightGBM-XGBoost 融合模型的风电场中短期功率预测方法

陈　硕[1*]，徐凯宏[2]，余子恒[2]

（1. 国网浙江杭州余杭区供电公司，浙江 杭州 311100；
2. 国网浙江杭州市供电公司，浙江 杭州 310000）

摘要：为提高风电场功率预测精度，解决传统方法对复杂时空特征及风机工况差异建模不足的问题，文章提出了一种基于分组策略的双模型融合预测方法。首先，采用随机森林算法对原始数据进行清洗，并基于场站周边气象站点及风机经纬度信息构建气象数据空间模型；其次，引入多种时空特征以增强输入数据的表征能力；在此基础上，针对风机工况差异进行分组，分别采用LightGBM和XGBoost模型进行训练，并采用误差加权的方式进行模型融合。实验结果表明，所提方法相比传统单一模型预测精度显著提升，验证了其有效性和实用性。本研究通过多模态数据协同建模与分组融合机制，为高精度风电预测提供了可推广的技术路径。

关键词：风电功率预测；空间插值；分组建模；LightGBM；XGBoost；模型融合　　**中图分类号：**TM76

DOI：10.13882/j.cnki.ncdqh.2501A035　　**CSTR：**32400.14.ncdqh.2501A035

Hybrid LightGBM-XGBoost Model for Medium and Short Term Wind Farm Power Forecasting

CHEN Shuo[1*], XU Kaihong[2], YU Ziheng[2]

(1. State Grid Zhejiang Hangzhou Yuhang District Power Supply Company, Zhejiang Hangzhou 311100, China;
2. State Grid Zhejiang Hangzhou Power Supply Company, Zhejiang Hangzhou 310000, China)

Abstract: To improve the accuracy of wind farm power forecasting and address the limitations of traditional methods in modeling complex spatiotemporal features and turbine operating condition variations, this paper proposes a dual-model fusion forecasting method based on a grouping strategy. First, the Random Forest algorithm is employed to preprocess the raw data, and a spatial meteorological model is constructed using data from surrounding weather stations and turbine location coordinates. Second, diverse spatiotemporal features are incorporated to enhance the input data representation. The turbines are then clustered according to their operating condition differences, with each group trained separately using LightGBM and XGBoost models, followed by an error-weighted fusion strategy. Experimental results demonstrate that the proposed method achieves significantly higher accuracy compared to conventional single-model approaches, validating its effectiveness and practicality. Through multi-source data integration and a grouped fusion mechanism, this study provides a generalizable framework for high-precision wind power forecasting.

Keywords: wind power forecasting; spatial interpolation; grouping strategy; LightGBM; XGBoost; model fusion

0　引言

近年来，伴随着经济的迅猛增长，社会的电力需求持续上升，导致发电量逐年增加。为了实现《巴黎协定》中设定的将全球气温上升控制在 2 ℃以内的目标，必须加大在新能源、低碳技术和清洁能源研发上的投资，以优化能源结构。风力发电由于其高效、成本低廉且对环境友好等优势，受到了广泛的关注和青睐[1]。然而，在实际工程应用中，由于气象条件的波动性和不确定性，风力发电在时间和空间上表现出复杂的耦合与聚合特性，这给风电并网工程带来了极大的挑战[2-3]。

为确保电力系统的稳定与安全运行，风电功率预测的及时性和准确性需达到高标准和严要求。这不仅有助于调度机构科学合理地安排各种电源的启

收稿日期：2025-01-16

停，还能有效保障发电侧与消费侧之间的功率平衡。在这一背景下，提高风电功率预测精度，优化调度策略，已成为实现电力系统高效、可靠运行的关键所在[4-5]。

风电原始数据具有高维度和非线性的特点，针对数据特征的复杂性，学者们提出了多种解决方法。文献 [6] 提出了一种基于时间依赖特征挖掘的风电场单机功率预测方法，通过引入注意力机制权重来提升预测精度。文献 [7] 构建了 CNN-LSTM 混合分析算法，可有效捕捉输入数据之间的关键信息。文献 [8] 结合交叉局部异常因子和注意力机制，提高了异常数据环境下的预测鲁棒性。文献 [9] 考虑到风电功率的非平稳特性，提出了基于改进经验模态分解和支持向量机的组合预测方法。文献 [10] 则采用谱聚类算法对风电机组历史运行数据进行聚类分析，以捕捉机组间的运行差异性。

尽管上述研究在风电功率预测精度提升方面取得了显著进展，但仍存在若干亟待解决的问题：首先，现有方法普遍忽视了风电场空间特征与功率输出的高相关性，时空特征的协同建模不足；其次，单一模型的泛化能力有限，难以适应复杂多变的风电场景；再者，实际运行数据常因采集设备故障等因素包含大量异常值和缺失值，现有研究缺乏应对方案；此外，单一气象站点的观测数据难以准确反映全场风机的运行环境，而不同型号风机的功率特性差异也尚未得到充分考虑。

针对上述问题，本文采用随机森林对风电场的原始数据进行清洗，并基于场站周边多个气象站点及风机的经纬度信息，对风机的气象数据进行空间建模，同时引入多种时空特征以提高预测精度。此外，本文根据风机的工况差异进行分组，每组分别采用 LightGBM 和 XGBoost 模型进行预测，并通过双模型融合的方法有效减少单一模型的误差，从而显著提高了整体预测精度。

1　数据集及模型介绍

风电功率预测模型的构建需要大量实测场站数据和历史气象预报数据来完成，通过对数据进行合理的清洗和特征构造，将有利于机器学习模型的拟合。同时，选择合适的机器学习算法对提高预测精度和效率至关重要。风电数据的高维度、复杂的非线性关系以及大规模数据集对算法提出了更高的要求。

1.1　数据集

场站数据来源于广东省南部沿海某风电场站 2022 年 3 月—2024 年 8 月的实测数据，包括风机轮廓风速、实时温湿度、单风机功率以及场站实发功率等，采样间隔为 15 min。实际数据中包括部分缺失值和异常值。

Open-Meteo 是一个由德国气象学家和软件开发者共同创建的开放数据平台，面向公众提供免费的气象数值模型短中期预报和 API 服务。为了弱化单一数值模型计算过程中可能存在的奇异点对预测结果的影响，本文在 Open-Meteo 平台调用 ECMWF、NOAA 和 CMA 数值天气预报，并结合 3 个模型的预报数据结果来做出更加准确可靠的气象数据集。针对大于等于 1 h 时间间隔的气象源数据，采用插值方法得到 15 min 间隔数据。气象数据特征见表 1。

表1　气象特征

特征	含义	特征	含义
Wind Speed (10 m)	10 m风速	Temperature (2 m)	2 m高度气温
Wind Speed (80 m)	80 m风速	Relative Humidity (2 m)	2 m高度相对湿度
Wind Speed (120 m)	120 m风速	weather_code	天气代码
Wind Speed (180 m)	180 m风速	precipitation	降水
Wind Direction (10 m)	10 m风向	pressure_msl	海平面气压
Wind Direction (80 m)	80 m风向	cape	对流有效位能
Wind Direction (120 m)	120 m风向	wind_speed_1000 hPa	0.1 MPa高度处风速
Wind Direction (180 m)	180 m风向	wind_direction_1000 hPa	0.1 MPa高度处风向
Wind Gusts (10 m)	10 m阵风		

1.2　模型介绍

在机器学习中，梯度提升决策树能够有效捕捉风速、风向及其他气象因素之间复杂的非线性关系。其中，LightGBM 和 XGBoost 作为梯度提升决策树中的代表性算法，都具备良好的可扩展性，支持分

布式计算，能够高效处理大规模数据集，并且通过并行化处理和防止过拟合的机制，都能提高模型的训练速度和泛化能力。

2 数据预处理

2.1 数据清洗

风电功率预测的准确性高度依赖于输入数据的质量，包括数值天气预报数据和相应的功率观测数据。然而，这些数据在实际测量过程中可能会受到环境噪声、通信稳定性的影响，导致数据异常或数据缺失。为了提高数据质量，通常应对采集到的原始数据进行一系列预处理步骤，包括缺失值处理、异常值检测、不变值处理及数据平衡等。

2.1.1 缺失值处理

为确保数据集的完整性和连续性，以便更准确地进行预测，须对缺失值做相应的处理。

本文采用三次样条插值法来填充数据中的缺失值，三次样条插值不但可以提供准确的插值结果，而且可以保证生成的插值函数变化光滑，能真实反映数据关系。此外，该方法在处理大量数据点时表现稳定，避免了震荡，且计算效率高，是应用研究和工程中的常用选择。

2.1.2 异常值检测

风机输出功率与风速的三次方成线性关系。但在实际工程中，由于风机受控制算法、测量误差、停机检修和弃风限电等因素影响，使得部分情况下功率曲线大大偏离理论值。

随机森林（random forest，RF）通过集成多个决策树的预测结果，能够减少单棵树对异常值的敏感性，在捕获数据中的复杂非线性关系的同时提高整体模型的鲁棒性和稳定性。随机森林模型在训练时，能够更好地保留数据的内在模式，避免了因异常值导致的模型失真。对于一些包含复杂关系的数据集，随机森林能够更准确地识别出那些偏离正常模式的异常值。

本文使用 RF 模型对数据进行训练，建立正常数据的预测模型。接着，通过将模型的预测结果与真实标签进行对比，可以识别出预测误差显著高的实例，并将其界定为异常值。

图 1 是采用四轮随机森林法清洗的结果。从图中不难看出，随着清洗轮次的上升，RF 模型筛选出的有效数据越发拟合风机的理论功率曲线。

2.1.3 不变值处理

在风电历史功率数据集中，部分风机在某些时间段的气象和功率数据丧失波动性，在较长时间内保持恒定，与邻近风机对比呈现明显异常，这是因为传感器出现信号丢失或设备卡顿等故障情况，称其为“不变值”。

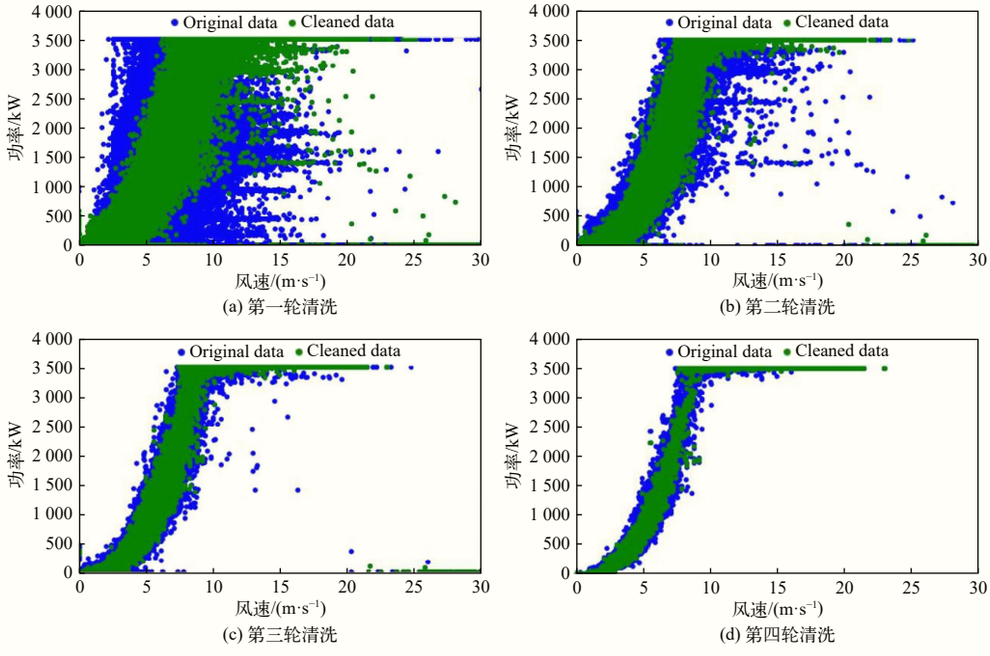

图1 随机森林法清洗效果

例如：图 2 为 34 号风机在 2024 年 4 月 21 日的风速和功率数据，场站的数据同样存在不刷新现象。“不变值”的风速和功率虽然满足功率曲线，但是却与实际气象不对应，这将影响模型的训练过程。本文选择根据时移特征和差值衍生特征对疑似“不变点”的异常重复数据进行甄别和剔除，仅保留时序中的第一个点。

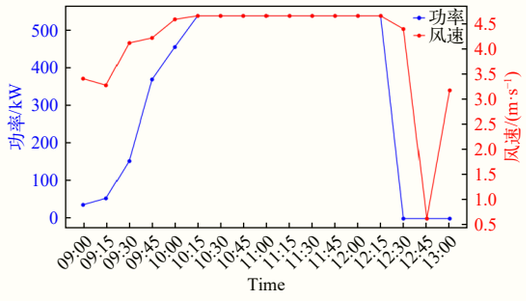

图2　34 号风机风速、功率曲线

2.1.4　数据平衡

原始数据集中存在大量风速小于切入风速的工况，图 3 为 34 号风机经数据清洗之后的风速占比，小于切入风速的风机理论输出功率为 0，可见 0 功率的占比高达 40% 以上。

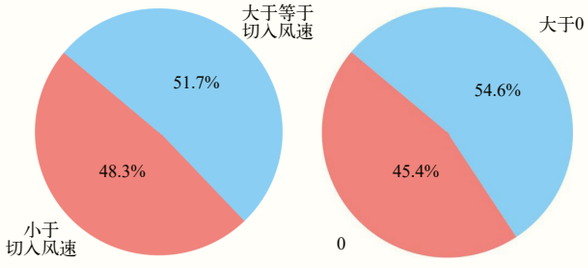

图3　风速及功率分布情况（切入风速 = 3 m/s）

在回归模型中，数据集的分布不均衡可能会导致模型在优化损失函数时受到数据统计分布情况的强烈影响，从而在罕见气象下表现出较弱的拟合能力。这意味着模型可能会将更多的关注点放在低风速区间，这些点可能具有较大的权值和较高的更新频率。实际中更希望模型重视中、高风速区间的数据，力求在大风天拥有更准确地预测结果。

因此，本文采用下采样（downsampling）的方式，随机丢弃 90% 小于切入风速的数据，来平衡数据在不同功率大小的数值分布。

2.2　特征构造

当处理时间序列数据时，将时间信息转化为模型可理解的特征尤为重要。时间特征提取能够帮助模型更好地理解和利用数据中的时间依赖性和周期性，从而提高预测的准确性和模型的整体性能。

2.2.1　天数特征

计算日期在整个月中的相对位置来构造特征 days_in_month，可以帮助模型更好地理解每日气象的连续性变化情况。

2.2.2　月特征

月是一个周期性的变量（12 个月为一周期），通过这种方式可以有效地捕捉周期性模式。海风的月变化主要体现在季节性上，这与地-日轨道变化和太阳直射点的变化相关。所涉及的风电场位于亚热带季风区，海陆气温的季节性变化影响季风活动特征的变化。

$$\begin{cases} m_{\sin} = \sin\left(\dfrac{2\pi m}{12.0}\right) \\ m_{\cos} = \cos\left(\dfrac{2\pi m}{12.0}\right) \end{cases} \tag{1}$$

式中：m 为月份；$m_{\sin}$ 为月份的正弦分量；$m_{\cos}$ 为月份的余弦分量。

式中：m 为月。

2.2.3　小时特征

与处理月的方法类似，通过正弦和余弦函数将小时信息转换为周期性特征，小时也是一个周期性变量（24 h 为一周期），这种转换有助于模型更好地捕捉一天中的周期性模式，例如（近海岸地区）陆海风的交替转换。

$$\begin{cases} h_{\sin} = \sin\left(\dfrac{2\pi h}{24.0}\right) \\ h_{\cos} = \cos\left(\dfrac{2\pi h}{24.0}\right) \end{cases} \tag{2}$$

式中：h 为时刻小时数；$h_{\sin}$ 为小时数的正弦分量；$h_{\cos}$ 为小时数的余弦分量。

2.2.4　分钟特征

利用式 3 将分钟数分为每 15 min 一个区间（0～14 min 为 0，15～29 min 为 1，以此类推）。

$$m_{\mathrm{c}} = \left\lfloor \frac{m_{\mathrm{i}}}{15} \right\rfloor + h \times 4 \tag{3}$$

式中：m_{i} 为时刻的分钟数；m_{c} 为分钟数的类别变量。

这种离散化相比小时特征做了进一步细化，并转换为类别变量以减小模型复杂度，不仅可以帮助

模型更好地理解和利用日周期性特征，同时也可以15 min间隔预测数据。

2.2.5 季节特征

季节性变化是风资源最显著的特征之一。不同的季节，风速和风能产出往往表现出明显的周期性变化。例如，夏季风期间受到来自太平洋的暖湿气流影响，通常伴有强降水和强对流天气；冬季风期间受到西伯利亚冷空气影响。通过构造季节特征，可以准确反映这种周期性变化。表2为季节特征。

表2 月季节对应表

月	季节特征
12、1、2	0
3、4、5	0.25
6、7、8	0.75
9、10、11	1

2.2.6 经纬度特征

由于在同一批气象数据中，各风机相对于气象观测点的地理位置存在显著差异，将风机的经纬度特征纳入数据集一定程度上可以反映风机的空间分布，还能帮助模型更精准地识别不同区域的风力资源变化。此外，地理空间的差异往往伴随着微气候的变化，通过对经纬度的考量，可以进一步揭示这些局部环境对风机性能的影响。

2.2.7 年平均风速

考虑到不同风机所处的地理环境和自然条件存在差异，每年感受到的风速也会有所不同。为了更准确地反映这一动态变化，本文将每个风机的年平均风速特征加入特征数据中，这一特征将有助于模型捕捉风速的长期趋势。

3 空间建模及风机分类

3.1 空间建模

在风电功率预测中，由于气象数据通常以网格点位置（如气象站或数值天气预报网格）的形式存在，而风机分布在不同的地理位置，风机的地理分布可能与气象数据的网格点位置不完全匹配，合理的空间建模对于提升预测准确性至关重要。通过匹配风机位置与气象数据、捕捉局部气象特征以及分析时空动态变化，空间建模能够显著提高风电功率预测的准确性，对风电场的有效管理和运营至关重要。

通过空间建模，可以将不同网格点的气象数据有效插值到具体的风机位置。这一过程能够使模型获取更贴近风机实际运营环境的气象信息，从而避免因位置不匹配导致的预测误差。

griddata插值是一种常用的数据插值技术，主要用于在空间或时间上对散点数据进行插值，以生成平滑的网格数据。这种技术广泛应用于各种需要处理空间数据的领域，如地理信息系统、气象学、风电功率预测等。通过griddata可以将离散的数据点转化为连续的曲面，从而更全面地描述空间变量的分布情况。

选择风机周围（111.5，21.25）、（111.5，21.38）、（111.62，21.25）、（111.62，21.38）、（111.75，21.25）、（111.75，21.38）6个气象点对风机气象数据进行插值。图4是气象站点10 m风速数据和1、27、97、136号风机的风速插值结果。

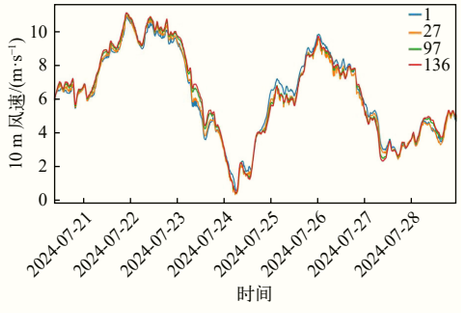

图4 风速插值曲线

3.2 风机分组

考虑到场站内相同型号、相同额定功率风机的风速功率曲线相似，因此考虑对风电场的风机进行分组。每组将构建单独的模型去预测。按额定功率、切入切出风速等可以分为4类风机，见表3。

表3 风机分组

风机ID	额定功率/kW	切入风速/(m·s⁻¹)	额定风速/(m·s⁻¹)	切出风速/(m·s⁻¹)
63—109、153—167	6.45	3	10.3	25
110—152	7.00	3	10.5	25
168	5.50	3	10.3	25
169—199	6.45	3	11.0	25

4　模型设计

在风电功率预测中，模型设计的核心在于构建一个能够精准捕捉数据特征、高效进行预测且具备良好泛化能力的模型。这不仅涉及算法的选择和优化，还包括对数据的深入理解、特征的有效提取以及模型性能的全面评估。

选定 LightGBM 和 XGBoost 作为核心算法后，本文根据风电功率预测的具体需求和数据特性进行模型的定制化处理，具体实现流程见图 5。

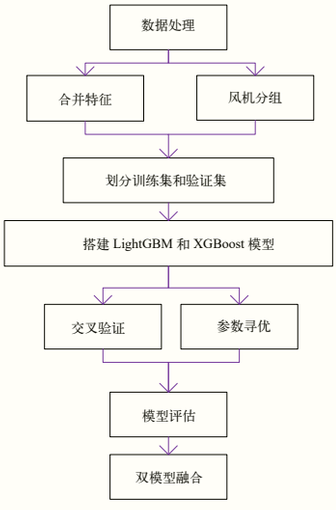

图5　方案流程

首先，将每个风机气象特征、时间特征以及空间特征进行合并，同时对风机进行分组处理，合并组内风机数据，每组构建单独的模型。其次，对每组数据进行训练集、验证集、测试集划分，测试集取最后 10 天的数据（960 点），训练集和验证集按 9∶1 进行分割。

每组数据分别搭建 LightGBM 和 XGBoost 进行训练，模型寻优的最大评估次数为 50，即寻优过程中尝试的超参数组合的最大数量为 50，交叉验证的损失函数采用 mean_squared_error。

利用每组模型预测组内风机的发电功率，再将所有风机的发出功率求和，减去厂用电及损耗，最后得到全站发电功率的预测。以测试集对模型预测结果进行评估，评估标准采用均方误差（MSE）和平均绝对误差（MAE）以及 r2_score。最后，对 LightGBM 和 XGBoost 进行融合，评估融合后的精度。

5　模型评估

表 4 展示了 LightGBM 和 XGBoost 模型在验证集上的测试结果。从结果可以看出，2 个模型的 MSE 和 MAE 均不超过 0.5，表明模型的预测误差在合理范围内。此外，r2_score 的测试结果表明，预测曲线与真实曲线具有较高的吻合度，进一步验证了模型的有效性和准确性。

表4　验证集测试结果

算法名称	风机组别	MSE/kW^2	MAE/kW	r2_score
XGBoost	63—109、153—167	0.216	0.249	0.958
	110—152	0.294	0.333	0.941
	168	0.391	0.355	0.798
	169—199	0.141	0.224	0.904
LightGBM	63—109、153—167	0.213	0.234	0.963
	110—152	0.279	0.328	0.948
	168	0.467	0.409	0.727
	169—199	0.132	0.220	0.908

图 6 展示了风电功率影响因素中排在前 10 位的特征重要性。由此可见，发电功率受时空特征和风速特征的影响较大。这一结果表明，这些特征在预测模型中起着关键作用，对提高模型的预测精度具有重要影响。

图 7、图 8 展示了 4 组类别中 50 号和 105 号风机的模型预测结果。可以看出，预测结果与实际结果之间存在很高的相关性，表明模型具有良好的预测能力。这一发现进一步验证了所提出的模型在风电功率预测中的有效性和准确性。

场站功率通过将所有风机的预测功率进行求和得出。表 5 展示了不同模型针对全站功率的 MAE

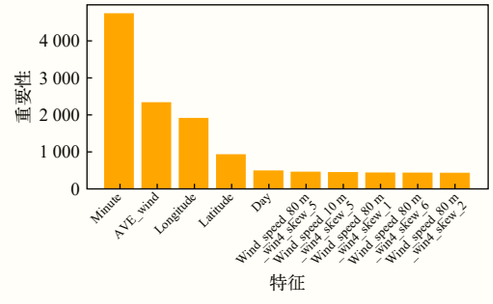

图6　特征重要性

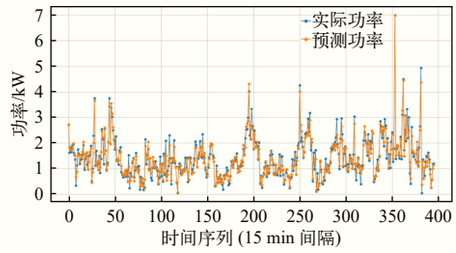

图7　50 号风机功率曲线

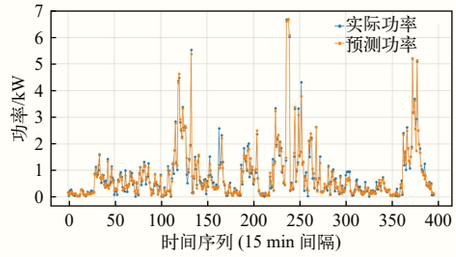

图8　105 号风机功率曲线

表5　不同模型下的 MAE

模型	MAE/kW	融合权重
LGB	36.449 056	0.315 149 79
XGB	64.452 798	0.165 533 8

值。在权重分配方面，采用了 MAE 倒数的占比来进行分配，以提高整体预测的准确性。

通过按照融合权重进行分配，将 XGBoost 和 LightGBM 2 种模型进行融合后，得到的 MAE 值为 33.102。与单个模型相比，这种融合方法显著提高了模型的性能，表明融合模型在预测全站功率方面更准确。

6　结束语

从模型的设计到验证结果表明，随机森林的清洗方法能够有效识别原始数据中的异常值，使输入数据更加符合实际的功率风速曲线。通过空间建模的技术手段，充分利用了时空信息来预测风机的功率，有效弥补了单一气象站点在适用性上的局限。此外，对风机进行分组并分别构建预测模型，使模型更加适用于不同风机的实际运行工况，增强了模型的针对性。最后，通过模型融合策略，进一步提升了预测模型的精度。

本研究为提升风电预测的精度和可靠性提供了坚实的基础，然而在模型训练过程中也观察到了计算效率的局限性。随着技术的持续进步和风电行业的快速发展，探索并应用更高效的人工智能模型将成为未来的研究方向。

参考文献

[1] 王立忠，王立林，洪义，等.海上风电技术发展趋势[J].能源工程，2024，44(6):3-12.
[2] 任建明.“碳中和”背景下的风电发展[J].农村电气化，2021(7):60-63.
[3] 安佰慧，袁博文，赵胜利，等.海上风功率预测精度的影响因素及提升方法研究[J].水电与新能源，2024，38(7):21-26，50.
[4] 时帅，张皓，黄冬梅，等.考虑爬坡特征量的海上风电短期分区功率预测[J].太阳能学报，2024，45(12):258-268.
[5] 路朋，叶林，汤涌，等.基于模型预测控制的风电集群多时间尺度有功功率优化调度策略研究[J].中国电机工程学报，2019，39(22):6572-6582.
[6] 苏向敬，宇海波，符杨，等.基于DALSTM和联合分位数损失的海上风电功率概率预测[J].中国电力，2023，56(11):10-19.
[7] 孙国强，项航，王新居，等.用于超短期风电预测的混合深度学习模型[J].电气开关，2022，60(4):50-53.
[8] 汪欣，蔡旭，李征.结合交叉局部异常因子和注意力机制的超短期风电功率预测方法[J].电力系统保护与控制，2020，48(23):92-99.
[9] 王涛，高靖，王优胤，等.基于改进经验模态分解和支持向量机的风电功率预测研究[J].电测与仪表，2021，58(6):49-54.
[10] 徐睿麟，郑建勇，梅飞，等.基于谱聚类和多元变分模态分解的风电机组功率预测[J].电网技术，2024，48(5):2043-2053.

作者简介

陈硕（1994—），男，硕士，从事电网自动化运维工作，E-mail：2019234251@tju.edu.cn。

（责任编辑：张峰亮）

资讯目次