DOI:10.19651/j.cnki.emt.2519259

基于 CSSA-CatBoost-LSTM 的风机状态预测*

孙家栋[1] 李子恒[2] 陈德基[1] 施 珮[1]
(1.无锡学院江苏高校优秀科技创新团队(实时工业物联网) 无锡 214105;
2.南京信息工程大学计算机学院 南京 210044)

摘 要:针对风机状态预测中特征提取不充分及单一模型预测精度不足的问题,提出一种融合 CatBoost 算法与长短期记忆网络(LSTM)的风机运行状态预测方法。首先,基于风机传感器特征和时序特征,使用 SVFE 和 MVFE 方法交叉融合生成全局复合特征,并结合熵权法改进的灰色关联分析实现特征降维。其次,通过引入混沌映射改进的麻雀搜索算法(CSSA)对 LSTM 模型超参数进行全局寻优,实现最优参数组合的自适应筛选与精准确定。最后,通过最优加权组合策略对 CatBoost 与优化后的 LSTM 进行深度融合,以提升预测精度与模型泛化能力。以中国宜昌某磷化工企业风机为例,对所提 CSSA-CatBoost-LSTM 风机状态预测方法进行了验证,验证结果表明该方法在准确性和可靠性方面有显著提升。

关键词:CatBoost;长短期记忆网络;熵权法;灰色关联分析方法;混沌映射;麻雀搜索算法
中图分类号:TH43;TN98 **文献标识码**:A **国家标准学科分类代码**:460.2020

Wind turbine health state prediction based on CSSA-CatBoost-LSTM

Sun Jiadong[1] Li Ziheng[2] Chen Deji[1] Shi Pei[1]
(1. Excellent Science and Technology Innovation Teams of Jiangsu Universities (Real-time Industrial Internet of Things), Wuxi University, Wuxi 214105, China; 2. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: To address the issues of insufficient feature extraction and inadequate prediction accuracy of single models in wind turbine condition forecasting, this study proposes a wind turbine operational condition prediction method that integrates the CatBoost algorithm with the Long Short-Term Memory network (LSTM). Firstly, based on the wind turbine sensor features and temporal features, the SVFE (a feature extraction method, assume its full name is known in the specific context) and MVFE (another feature extraction method, assume its full name is known in the specific context) methods are employed for cross-fusion to generate global composite features. Additionally, feature dimension reduction is achieved by incorporating grey relational analysis improved with the entropy weight method. Secondly, the Sparrow Search Algorithm (SSA) enhanced by chaotic mapping, termed CSSA, is introduced to conduct global optimization of the hyperparameters of the LSTM model, enabling adaptive screening and precise determination of the optimal parameter combination. Finally, the CatBoost model and the optimized LSTM model are deeply fused using an optimal weighted combination strategy to enhance prediction accuracy and model generalization capability. Taking the wind turbines of a phosphorus chemical enterprise in Yichang, China, as an example, the proposed CSSA-CatBoost-LSTM wind turbine condition prediction method was validated. The validation results demonstrate significant improvements in both accuracy and reliability of this method.

Keywords: CatBoost; long short-term memory network; entropy weight method; grey correlation analysis method; sparrow search algorithm; chaotic mapping

0 引 言

随着工业 4.0 和智能制造的快速发展,旋转机械设备的健康状态监测与故障预测在能源、化工、冶金等重工业领域的重要性日益凸显[1]。作为工业生产的核心动力设备,风机的稳定运行直接关系到生产安全、能效优化和设备寿

收稿日期:2025-07-02
*基金项目:国家自然科学基金(62072216)、江苏省高校自然科学研究面上项目(21KJB520020)、无锡学院引进人才科研启项经费(2023r005)项目资助

命。以磷化工行业为例，罗茨风机在原料输送、废气处理等关键环节中承担着不可替代的作用，其故障可能导致生产中断、环境污染甚至安全事故[2]。因此，准确预测风机运行状态以保障其正常运行，对确保工业生产的稳定性与高效性具有重要意义。

近年来，风机状态预测技术作为保障设备可靠运行的关键手段，已取得显著进展。现有方法主要分为 3 类：基于物理模型的方法[3-4]、基于传统数据驱动的方法[5-6]以及基于深度学习的方法[7-8]。基于物理模型法虽能深入剖析风机运行机理，但受限于风机结构复杂性与运行环境不确定性，存在精确建模难度大、计算复杂度高及实时应用困难等问题。传统数据驱动方法虽可提升预测精度并丰富数据，但存在解释性不足，且在处理风机数据时面临特征提取与泛化性能受限的挑战。鉴于上面两种方法的局限性，基于深度学习的预测方法逐渐成为研究热点，其中循环神经网络(recurrent neural network，RNN)及其变体的应用最为广泛。

长短时记忆网络(long short-term memory，LSTM)凭借其对风机数据时序性与非线性关系的建模能力[9-10]，及 CatBoost 算法具备出色特征识别能力，在风机状态检测领域得到广泛应用[11]。当前研究为提升模型的预测精度，主要聚焦于多模型融合策略与超参数智能调优技术两大方向。例如：文献[12]构建了卷积神经网络、LSTM 和注意力机制的组合模型，实现电站风机状态的多工况适应性预测。文献[13]提出了由自适应噪声完备集成经验模态分解与反向传播神经网络组成的组合模型，通过改进麻雀搜索算法(sparrow search algorithm，SSA)算法对反向传播神经网络的超参数进行寻优，以提升风机功率短期预测精度。现有研究虽可提升风机状态预测的精度，但仍存在因模型输入特征维度不足，特征关联性低、或优化算法改进不足导致的风机状态预测效果欠佳、时间成本较高等的问题。

在预测特征选择方面，风机异常时会伴随振动、发热和噪音等物理现象[14]。尽管三者均可用于风机预测和故障识别，但振动和噪音易受环境噪声干扰且信号处理流程复杂[15]，而温度特征凭借数据采集成本低、信号处理逻辑简单、工业现场适配性强等显著优势，成为当前风机状态预测领域的优选技术方案[16-17]。

本文结合 LSTM 和 CatBoost 的优势，建立了基于 CatBoost-LSTM 的风机工作温度预测模型，同时，为了降低特征冗余，减少网络计算负担，本文提出了一种基于距离相关系数的最小冗余最大相关(minimal-redundancy-maximal-relevance，mRMR)算法用于特征筛选。本文对残差信号使用贝叶斯检验，以达到在线监测和预警的效果。

综上所述，本文提出一种基于混沌映射改进麻雀搜索算法(chaotic sparrow search algorithm，CSSA)优化类别提升与 LSTM 耦合的风机状态预测方法。在特征工程阶段，从时间戳、原始特征中提取年、月、日、振动等新特征并将其交互扩展特征集，后运用信息熵加权的灰色关联分析方法筛选扩展的特征集，实现原始特征降维与关键特征提取，进一步提升预测精度。并且运用 Tent 混沌算法对麻雀搜索算法进行改进，可以优化算法的收敛速度，使算法更快地找到最优解。然后，利用最优加权组合法将 CatBoost 和 LSTM 模型进行加权组合，将两组特征作为模型的输入特征进行模型训练；最后，利用某磷化工企业实际风机运行数据对所提方法和模型的预测精度进行验证，为风机故障预警提供了兼具理论创新性与工程实用性的解决方案。

1　基本理论

1.1　CatBoost 模型

CatBoost 是一种基于梯度提升决策树(gradient boosting decision tree，GBDT)框架构建的集成学习算法。通过引入了排序提升法机制，优化梯度估计过程，能够有效缓解传统 GBDT 算法梯度偏差和预测偏移现象，从而在抑制模型过拟合的同时，显著提升预测精度和泛化性能。CatBoost 核心思想在于通过构建有序的数据子集序列，打破传统 GBDT 中梯度估计的依赖关系。

1.2　LSTM 网络

LSTM 是一种主要用于序列数据处理的循环神经网络，具有记忆效应，可处理任意长时序数据。相比 RNN，LSTM 引入一个新的内部状态 c^t 专门进行线性循环信息传递，LSTM 网络的核心单元结构如图 1 所示，其通过动态门控机制实现时间序列信息的选择性记忆与传递。

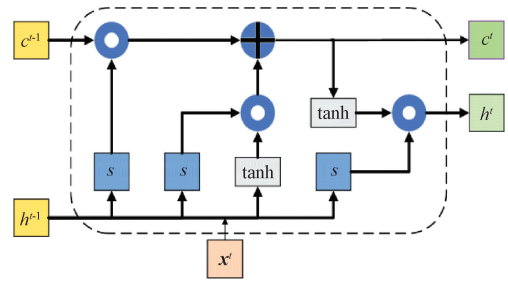


图 1　LSTM 单元结构

Fig. 1　LSTM unit structure

具体而言，x^t 代表当前时刻的输入向量，c^{t-1} 和 c^t 分别表示上一时刻与当前时刻的内部状态(即记忆单元)，用于存储长期历史信息；h^{t-1} 和 h^t 则为对应时刻的隐藏层状态，负责向后续层传递特征信息。在计算过程中，符号 ⊙ 表示逐元素乘积(Hadamard 积)，用于门控信号与状态向量的交互；⊕ 表示逐元素加法，用于状态更新。激活函数方面，采用 Sigmoid 函数(输出范围[0，1])生成门控权重，图 1 中为 s，tanh 则用于状态值的归一化处理。LSTM 通过 3 个关键门控信号调控信息流：遗忘门 z^f 动态决定上一时刻内部状态 c^{t-1} 中需丢弃的信息比例；输入门 z^i 控制当前候选状态 z 中需写入新状态 c^t 的信息量；输出门 z^o 则调节当前内部状态 c^t 中需输出至隐藏层 h^t 的信息比例。最终，内部状态与隐藏状态通过公式 c^t 和 h^t 协同更新。

1.3 麻雀搜索算法

麻雀搜索算法是由薛建凯[18]于 2020 年提出的一种新型智能优化算法,其设计灵感源自麻雀群体在觅食过程中规避捕食者的自然行为模式。该算法通过模拟麻雀种群中不同角色个体的协作机制,构建了一种兼具局部搜索能力与全局探索性能的优化框架。在 SSA 中,麻雀种群被划分为发现者(Scout)、加入者(Joiner)和侦察者(Sentinel)3 类角色[19],3 类个体通过动态交互与角色转换,在搜索空间中协同完成寻优任务。这种基于生物群体智能的机制赋予了 SSA 算法较高的收敛精度与较快的收敛速度,使其在复杂优化问题中展现出良好的性能。

1)发现者负责在搜索空间探索潜在最优位置,其"能源储备"(由适应度函数值衡量)较高,通过更新位置引导种群全局搜索。

2)加入者依据发现者位置调整搜索路径,发现者与加入者数量动态平衡且可角色互换,增强搜索多样性。

3)种群感知威胁时触发警报,发现者带领种群向安全区域迁移,体现算法对动态环境的适应性。部分加入者因适应度低,主动迁移至其他区域,促进算法在局部区域的深度探索。

4)加入者因竞争资源发生位置冲突并更新位置,增强局部搜索能力。

5)警报触发时,边缘麻雀向中心聚集,中心麻雀随机分散,平衡全局探索与局部开发能力。

1.4 混沌映射算法

在多数启发式智能算法中,种群初始位置常通过伪随机数生成(服从[0,1]均匀分布),但此类伪随机数可能引发变量在搜索空间分布不均,影响初始化质量[20]。为优化初始种群分布,本文引入混沌映射生成随机数。混沌映射作为一种具有初始条件敏感依赖性和长期不可预测性的数学模型,可确保变量在解空间中均匀分布,从而拓宽算法搜索范围、提升寻优精度与收敛速度。尽管 Circle、Logistic 及 Chebyshev 映射等经典混沌映射被广泛应用,但易导致变量分布不均及寻优效率低下[21]。相较之下,Tent 混沌映射因具有更优的遍历性和收敛速率,能够有效克服上述缺陷。基于此,本文采用 Tent 混沌映射对算法初始种群生成过程进行优化,以提升算法性能。

Tent 混沌映射的具体公式如式(1)所示,其中,$\alpha \in (0,1)$,根据实际情况选择适合的 α 值。

$$x_n+1=f(x_n)=\begin{cases}\dfrac{x_n}{\alpha}, x_n \in [0,\alpha) \\ \dfrac{(1-x_n)}{(1-\alpha)}, x_n \in [\alpha,1)\end{cases} \tag{1}$$

2 特征构造与特征挖掘

2.1 基于温度的风机状态特征构造

本方法选择温度作为关键特征来表征风机的运行状态。在正常工作状态下,风机系统内机械部件摩擦生热、电机损耗产热与强制对流散热等过程达成热力学动态平衡,使得设备温度维持在相对稳定的区间。当风机出现异常工况时,诸如轴承磨损加剧、润滑失效或电机过载等故障诱因,将显著提升系统产热功率,打破原有的热平衡机制,进而导致温度呈现超阈值攀升现象。而在停机过程中,随着机械运转动能与空气动力学作用的消失,设备产热机制停止,温度遵循牛顿冷却定律逐渐衰减,并在与环境达成热平衡后趋于稳定。

以中国宜昌某磷化工企业风机为例,异常工况、正常状态及停机状态下的温度时序数据。如图 2 所示,该图截取并呈现了风机在 3 种典型运行状态下的温度参数变化特征。实验观测数据显示,当风机处于异常运行状态时,其温度参数普遍高于 35℃,呈现出显著的过热特征;在突发停电工况下,系统温度出现明显下降,整体低于 25℃,处于低温区间;而当风机维持正常运行状态时,温度参数能够稳定控制在 25℃～35℃ 的合理区间内。上述结果表明,风机不同运行状态与温度参数之间存在显著的对应关系,温度变化可作为表征风机运行状态的重要指标。

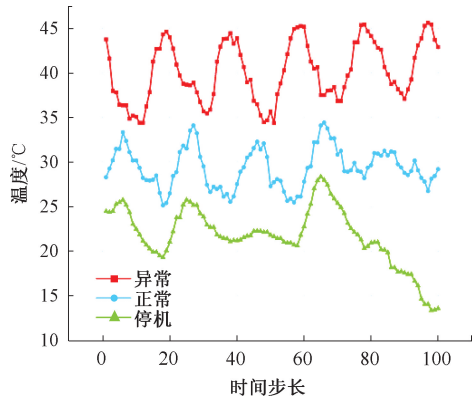


图 2 温度信号时序图

Fig. 2 Timing diagram of temperature signal

2.2 基于 SVFE 和 MVFE 的时序特征挖掘

针对数据集中各特征均呈现时间序列特性的场景,特征提取策略可分为两大类:其一为单变量特征提取(single variable feature extraction,SVFE),该方法聚焦于单一时间序列变量,通过时域统计、频域分析或非线性变换等手段,从原始序列中衍生出新的特征,以揭示数据内部的潜在模式与规律;其二为多变量特征提取(multiple variable feature extraction,MVFE),该方法则侧重于两个或多个时间序列变量之间的交互作用,通过相关性分析、协整检验或动态建模等方法,构建能够反映变量间协同效应的复合特征,从而捕捉数据中的复杂依赖关系。

针对罗茨风机数据集,其原始数据不仅涵盖了各类传感器采集的物理量时间序列(如温度、压力、振动等),还包含了以"%Y-%m-%d%H:%M:%S"格式记录的时间戳特征序列。为充分挖掘时间戳中蕴含的周期性、趋势性

信息，并增强模型对风机温度与特征间复杂非线性关系的建模能力，本研究采用 SVFE 方法对时间戳特征进行深度解析。具体而言，通过正则表达式匹配与时间格式解析技术，从时间戳字符串中提取出年(%Y)、月(%m)、日(%d)、时(%H)、分(%M)、秒(%S)等基础时间特征，这些特征作为新的时间序列变量，能够反映数据的时间维度变化。此外，为进一步捕捉数据的季节性规律，本研究还基于月特征构建了春、夏、秋、冬 4 个季节性时间序列特征，这些特征将共同用于模型训练[22]。

另外，如图 3 所示，将时间戳与振动速度特征进行融合构造新特征，采用 MVFE 进行特征提取，利用直接运算的方式构造复合特征，通过组合，扩充特征维度，加强特征隐藏关联的学习能力，从而提高模型预测精度。

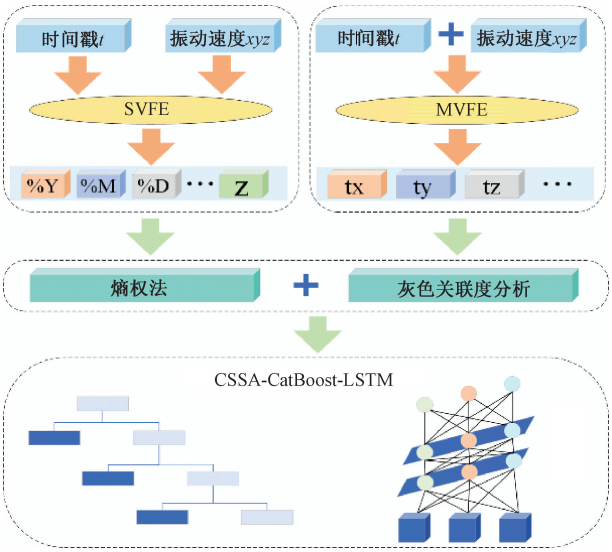


图 3　特征提取示意图

Fig. 3　Schematic diagram of feature extraction

2.3　基于熵权法和灰色关联的高关联特征筛选

灰色关联分析作为一种多因素统计分析方法，其核心思想是基于系统行为序列的几何形状相似性来量化因素间的关联程度。该方法通过构建参考序列与比较序列的关联度矩阵，能够有效揭示数据集中各因素之间的内在联系。

在风机运行状态监测领域，温度参数作为关键性能指标，其变化受多源传感器特征(如振动、转速等)的协同影响针对包含 $m\times n$ 维结构的数据集(n 为特征维度，m 为每个特征的时间序列长度)，灰色关联分析可系统化地量化风机温度序列 $x=\{x(k)\mid k=1,2,\cdots,m\}$ 与各特征序列 $x_i=\{x_i(k)\mid k=1,2,\cdots,m\}(i=1,2,\cdots,n-1)$ 之间的关联强度，其核心计算步骤以灰色关联系数的求解为核心，通过度量各特征序列与温度目标序列的几何相似性，揭示特征对温度变化的贡献权重。式(2)是灰色关联系数值计算公式：

$$c_i(k)=\frac{a+\rho b}{|x(k)-x_i(k)|+\rho b} \tag{2}$$

其中，$i=1,2,3,\cdots,n-1$；ρ 为分辨系数，取 0.5；a，b 为对绝对差值，具体计算公式如式(3)～(4)所示。

$$a=\min_i\{\min_k[|x(k)-x_i(k)|]\} \tag{3}$$

$$b=\max_i\{\max_k[|x(k)-x_i(k)|]\} \tag{4}$$

r_i 为灰色关联度：

$$r_i=\frac{1}{m}\sum_{k=1}^{m}c_i(k) \tag{5}$$

在基于灰色关联分析的评估体系中，式(5)采用对关联系数取平均值的方式来确定关联度大小。然而，该方式存在局限性，容易将各个特征的个性特征掩盖，无法充分体现除风机温度特征之外的其他特征相对于风机温度特征的重要程度差异。由此所得的关联度，难以满足对特征重要性进行精准评估的需求[23]。

信息熵作为一种重要的信息度量指标，能够有效地描述各事件发生的不确定性程度。在特征权重确定问题中，通过信息熵计算得到的熵值大小与特征所蕴含的信息量呈正相关关系。基于此原理，采用信息熵加权方法对特征进行处理，能充分利用实际数据中所包含的丰富信息来确定各特征的权重，避免了主观因素对评估结果的干扰，使得权重分配更加符合数据本身的内在规律。信息熵大小的计算过程具体如下：

根据式(6)计算熵值 S_i：

$$S_i=-[\ln(m)]^{-1}\sum_{k=1}^{m}p_i(k)\ln[p_i(k)] \tag{6}$$

其中，$p_i(k)$ 为特征 i 中第 k 个归一化值占该特征所有归一化值之和的比例，如式(7)所示。

$$p_i(k)=\frac{x'_i(k)}{\sum_{k=1}^{m}x'_i(k)} \tag{7}$$

将所求熵值 S_i 作为权重代入式(4)中，得出改进后的灰色关联度大小 $\widetilde{r_i}$，如式(8)所示。

$$\widetilde{r_i}=\frac{1}{m}\sum_{k=1}^{m}S_ic_i(k) \tag{8}$$

3　基于 CSSA-CatBoost-LSTM 模型的风机状态预测方法

3.1　CSSA-CatBoost-LSTM 模型

如图 4 所示，呈现了 CSSA-CatBoost-LSTM 模型结构化框架。具体来看，输入的特征是历数据清洗、深度特征挖掘以及筛选后获取的高关联度时序变量，在子模型层，CatBoost 模块通过构建多棵决策树，对输入特征实施非线性映射，进而输出预测分量；而 LSTM 模块则依托其特有的长短期记忆机制，敏锐捕获特征序列的时序依赖关系，输出另一路预测分量。

CSSA 算法进行参数寻优：一方面，针对 CatBoost 和 LSTM 模型，对学习率、神经元数量、学习次数等核心参数进行优化；另一方面，针对组合过程，优化各子模型预测分量的权重，从而实现最优加权融合。最终，经过 CSSA 优化

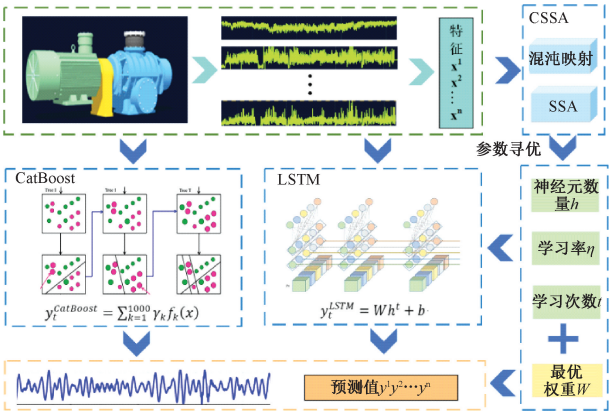


图 4　CSSA-CatBoost-LSTM 模型方法框图

Fig. 4　CSSA-CatBoost-LSTM model method flowchart

的加权组合模块,对 CatBoost 与 LSTM 的预测分量进行融合,输出风机状态的预测结果。

3.2　基于 CSSA-CatBoost-LSTM 模型的预测流程

本文提出一种基于 CSSA-LSTM-CatBoost 的组合预测模型用于磷化工罗茨风机预测。LSTM 网络凭借其独特的门控机制和记忆单元,能够有效捕捉振动信号、温度变化等时序数据的长期依赖关系和非线性动态特征;而 CatBoost 算法则通过有序提升策略和对称决策树结构,挖掘时间、温度、振动等多维特征间的复杂交互作用和隐含模式。两种方法的结合形成了时空特征的全方位捕捉能力,但简单的等权融合无法充分发挥其协同效应。为此,本研究引入改进的麻雀搜索算法进行 LSTM 超参数优化和最终权重优化。该算法通过立方映射混沌初始化确保种群多样性,采用自适应搜索策略平衡开发与探索过程,并设计复合目标函数综合考量预测误差的多个维度。预测流程图如图 5 所示。

1)多源数据采集与结构化:从化工厂数据库提取风机振动、电气及温度数据,通过 Python 存储,确保数据完整性与可访问性。

2)数据清洗与标准化:对连续型变量采用双重窗口策略进行缺失值插补,包括滑动短窗口(5 个时间步,对应 5 h)均值插补与周期性长窗口(24 个时间步,对应 24 h)均值插补;对类别型变量则采用众数进行缺失值填充;基于 3σ 原则识别连续型变量中的异常观测值,对检测到的异常值采用局部中位数(窗口大小为 7 个时间步)进行替换;最后应用 Min-Max 缩放方法,将所有连续型特征线性映射至[0,1]区间,实现特征量纲的统一化处理。该流程兼顾了时序数据的局部波动性与周期性特征,通过统计性方法确保数据处理的科学性与一致性,为后续建模分析奠定标准化的数据基础。

3)改进灰色关联分析(IGRA):结合信息熵与互信息计算特征权重,按式(7)计算温度与振动特征的加权灰色关联度,突出高贡献特征。

4)特征选择与时序特征构造:选取关联度排名前 10 的特征。后提取时间戳的年、月、日等维度,并构造时间戳-特征联合特征,最后取两组特征中关联度前 6 的特征作为模型输入。

5)最优加权组合模型构建:采用最优加权组合构建。

6)预测结果输出并可视化。

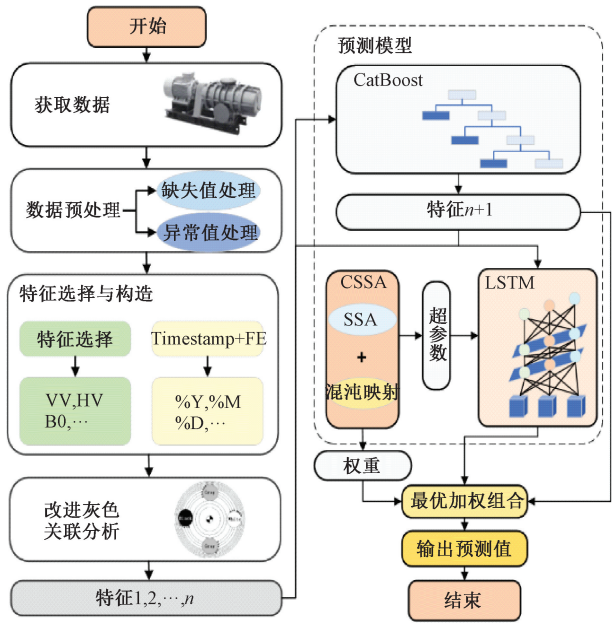


图 5　预测流程

Fig. 5　Prediction process

4　案例分析

算例数据选取的是某磷化工企业一罗茨风机 2023 年 1 月 1 日 00:00:00～2024 年 1 月 1 日 00:00:00 的一年的实测数据,主要包含罗茨风机前端的振动数据和温度数据,时间分辨率为 1 h,共 8 640 组数据,将数据集按照 8∶2 划分为训练集和测试集进行预测。并且选用温度低谷、高峰、平稳上升及下降 4 种情况进行预测效果展示。利用 3.3 节建立的磷化工风机预测模型进行仿真分析,验证所提方法的有效性和准确性。

4.1　数据清洗

表 1 给出了所用数据集的相关特征信息,包括特征名称、英文缩写以及单位。

表 1　风机原始特征

Table 1　Original characteristics of wind turbine

特征	单位	特征	单位
垂直振动速度	mm/s	电池电压	V
轴向振动速度	mm/s	震动加速度	g′s
水平振动速度	mm/s	润滑指数	%

考虑到风机工作异常、定期维护以及数据采集装置故障带来的数据缺失或异常,使得原始数据出现缺失或异常

等不良数据，为避免不良数据对模型训练学习过程带来不利影响，对原始数据集进行缺失值检测，随后采用前后数据求平均值的方式对缺失值进行填充，并且剔除异常值。

鉴于风机运行过程中可能因设备故障、定期停机维护以及数据采集装置的传感器漂移或通信中断等因素，导致原始数据集中存在缺失值或异常值，此类不良数据会显著干扰模型训练的收敛性与泛化能力。为此，首先对原始数据集进行缺失值检测，针对连续型数据采用基于滑动窗口（窗口大小设为 5）的均值插补策略，利用前后相邻时刻的有效数据均值填充缺失值，以保留局部时序特征；同时，借助 2024 年 1 月～4 月期间所采集的时间序列数据，运用 3σ 准则对数据中的异常值进行精准识别。针对识别出的超出阈值范围的数据点，直接予以剔除处理。

4.2　特征挖掘和筛选

首先，由式(2)～(8)对表 1 中所有候选特征进行灰色关联分析，量化各特征与目标温度特征间的非线性关联程度，并按灰色关联度降序排列，随后，基于信息熵理论计算各特征的熵值，以评估其信息冗余度，其中熵值越小表明特征蕴含的信息量越集中；计算结果及排序如表 2 所示，排序结果与改进前基本一致，其中垂直振动速度与振动加速度互换了排序位置，究其原因，是因为垂直振动速度与振动加速度相比，特征熵值较小，即蕴含的信息量比更集中；尽管特征 VV 与 HV 的信息熵值在候选特征集中排名靠前，表明其数据分布的离散程度较低、冗余信息较少，但二者在改进前后的灰色关联度及加权灰色关联度排序中均位列末端，表明其与目标温度特征的非线性关联强度较弱，信息贡献度不足以支撑模型预测性能。因此，将 VV 与 HV 从预测模型的输入特征集中剔除，以避免引入冗余噪声并优化模型复杂度。

表 2　风机原始特征改进前、后关联度大小计算结果

Table 2　Calculation results of the correlation degree before and after improving the original features of the wind turbine

特征	加权前		熵值	加权后	
	大小	排序		大小	排序
垂直振动速度(VV)	0.538 7	3	5.261 5	0.532 9	4
轴向振动速度(AV)	0.543 3	2	5.554 8	0.537 1	2
水平振动速度(HV)	0.537 5	5	5.115 3	0.531 9	5
电池电压(BO)	0.534 2	6	2.229 2	0.531 7	6
振动加速度(VA)	0.538 4	4	4.641 5	0.533 3	3
润滑严重性指数(LS)	0.592 1	1	5.760 0	0.585 1	1

由于特征选择较少，选择通过时间戳与特征相互交互扩展特征集，最终通过将原始特征集与新构建的特征集相互比较，选出前 10 名做出表 3，通过表 3 和上述分析，特征 BO-year、LS-hour、LS-day、LS-month、VA-day、VA-hour 灰色度大小均在 0.72 以上，但 BO-year 的熵值过低，是由于其变化模式单一，不利于进行模型预测的训练，所以添加特征 AV-day 作为第 6 个特征。

表 3　时间戳-风机特征改进前、后关联度大小计算结果

Table 3　Calculation results of correlation degree before and after timestamp fan feature improvement

特征	加权前		熵值	加权后	
	大小	排序		大小	排序
BO-year	0.983 9	1	2.229 2	0.979 4	1
LS-hour	0.788 0	2	7.510 6	0.775 9	2
LS-day	0.769 3	3	7.764 0	0.757 0	3
LS-month	0.757 7	4	7.150 5	0.744 6	5
VA-day	0.755 9	5	7.297 5	0.746 7	4
VA-hour	0.752 3	6	7.030 2	0.741 5	6
AV-day	0.739 7	7	7.994 1	0.727 6	7
AV-hour	0.727 9	8	7.715 6	0.716 4	8
VV-day	0.723 1	9	7.847 1	0.711 5	9
VV-hour	0.717 9	10	7.566 0	0.706 8	10

4.3　模型测试

为系统评估新构造时序特征对模型预测性能的贡献，本研究分别构建特征提取前后的 LSTM 与 CatBoost 模型，并对比其预测效果。具体而言，针对原始特征集与融合新时序特征的特征集，分别训练 LSTM 与 CatBoost 模型，将特征提取后的模型命名为 FE-LSTM（feature-extracted LSTM）与 FE-CatBoost，并通过均方误差（mean squared error，MSE）、平均绝对误差（mean absolute error，MAE）等指标量化特征提取对模型预测精度的提升作用如式(9)～(11)所示。

$$E_{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|x(i)-x^{*}(i)\right| \tag{9}$$

$$E_{MSE}=\frac{1}{N}\sum_{i=1}^{N}(x(i)-x^{*}(i))^{2} \tag{10}$$

$$R^{2}=1-\frac{\sum_{k=1}^{m}[y(k)-Y(k)]^{2}}{\sum_{k=1}^{m}[y(k)-\bar{Y}(k)]^{2}} \tag{11}$$

其中，$y(k)$ 为预测值；$Y(k)$ 为真实值；$\bar{Y}(k)$ 为真实值的平均值；m 为数据样本数。

此外，为验证组合模型 LSTM-CatBoost 在特征增强后的性能优势，将特征提取处理后的组合模型 FE-LSTM-CatBoost 与单一模型 FE-LSTM、FE-CatBoost 的预测结果进行对比，分析多模型融合策略对复杂时序依赖关系的建模能力。

通过时间戳特征与特征交互得到新的输入特征，并利用 CSSA 对模型超参数进行优化，将处理后的特征数据输入到 FE-LSTM-CatBoost-CSSA 组合模型中进行训练，并与单一模型以及未优化 SSA 的组合模型的预测效果进行对比。如表 4 所示，在测试集中，FE-CSSA-CatBoost-

LSTM 模、型比其他 4 种模型的 RMSE 和 MAE 均有较大程度的下降。而 R^2 相比与其他 4 种模型,精准度也更高。在训练时间方面,虽然 FE-CatBoost 模型训练耗 FE-CSSA-LSTM-CatBoost 模型 46 min 的训练时间处于一个相对合理的范围,既不会因为训练时间过长而影响模型的快速部署和应用,又能够保证模型具有较高的预测精度和拟合优度。

将测试集数据分别输入本文提出的 FE-CSSA-CatBoost-LSTM 模型与 BP-SVM、GBDT、LSTM 和 LSTM-Transformer 相对比,包括训练集与验证集大小也控制在相同范围内。预测结果如表 4 和图 6 所示。FE-CSSA-CatBoost-LSTM 模型表现出显著优势,其 MAE 为 7.05、MSE 为 11.83、R^2 达 99.49%。其在训练时间相对合理的情况下,实现了性能的显著优化,在训练效率和模型性能之间提供了平衡,训练时间短的模型往往在预测精度和拟合优度上有所欠缺。

表 4　五种不同模型预测结果对比

Table 4　Comparison of prediction results of five different models

模型	E_{MSE} /%	E_{MAE} /%	R^2 /%	训练时间 /min
BP-SVM	44.24	34.21	88.72	15
GBDT	39.32	30.31	91.08	12
LSTM	22.04	48.58	97.15	43
LSTM-Transformer	21.83	17.09	97.32	52
FE-CSSA-LSTM-CatBoost	11.83	7.05	99.49	46

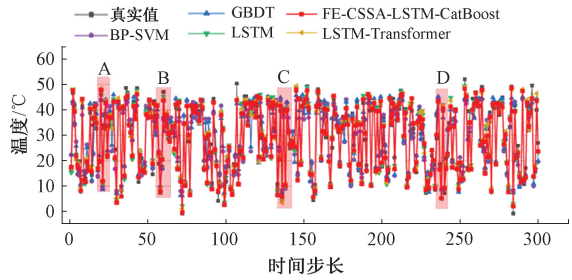
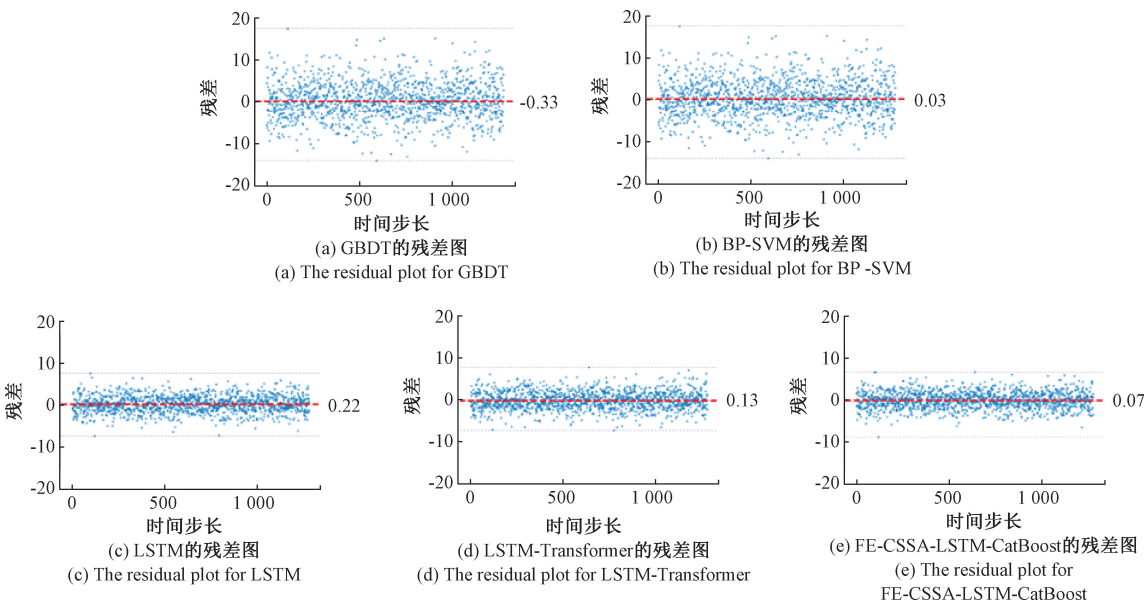


图 6　五种模型预测值与实际值的对比

Fig. 6　Comparison between actual values and predicted values of five models

为了展示预测值与实际值之间的差异,随时间或观测序号的变化情况。图 7 中分别呈现了 5 种模型的残差分布。GBDT 模型的残差在零值附近随机分布,虽然整体上没有明显的趋势,但残差的分布范围较广,表明在某些时间步长上,预测误差较大。BP-SVM 模型相较于 GBDT,残差的波动范围更大,且在某些区域残差较为密集,表明模型在某些特定时间步长上的预测能力较弱。LSTM 模型残差分布相对较为均匀,围绕零值上下波动,显示出 LSTM 模型在捕捉时间序列特征方面具有一定优势,但仍有部分时间步长上存在较大误差。LSTM-Transformer 模型的残差在零值附近呈现出较为紧密的分布,波动范围相对较小,表明该模型在整合 LSTM 和 Transformer 的优点后,能够更有效地捕捉时间序列中的复杂模式,预测精度有所提高。本文使用的模型,中位值 0.07 为最小,残差分布最为集中,且围绕零值紧密波动,具有出色的预测性能和稳定性,能够有效减少预测误差。



(a) GBDT的残差图
(a) The residual plot for GBDT
(b) BP-SVM的残差图
(b) The residual plot for BP -SVM
(c) LSTM的残差图
(c) The residual plot for LSTM
(d) LSTM-Transformer的残差图
(d) The residual plot for LSTM-Transformer
(e) FE-CSSA-LSTM-CatBoost的残差图
(e) The residual plot for FE-CSSA-LSTM-CatBoost

图 7　五种不同模型的残差对比图

Fig. 7　Residual comparison plot for five different model

如图 8 所示，展示了 5 种预测模型在风机温度预测任务中的性能表现。从图 6 中选择上升拐点 A、下降拐点 B、波峰 C 和波谷 D 四个典型区域展开详细分析。图 8(a)～(d)分别为对应区域的局部放大图。从图 8 中可清晰观察到：在温度变化平稳阶段，FE-CSSA-LSTM-CatBoost 模型的预测误差较 GBDT 降低约 29%，较 BP-SVM 降低 33%；在温度突变阶段，改进模型的误差峰值仍低于传统模型 20%～30%。

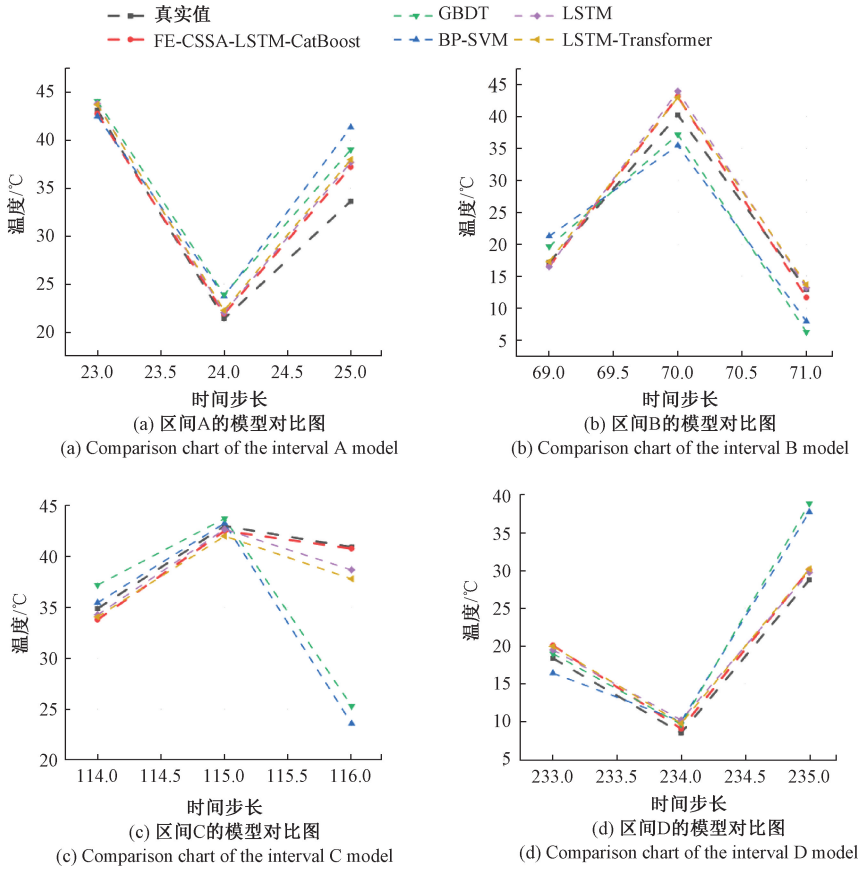


图 8　五种模型预测值与实际值的局部对比图

Fig. 8　Local comparison of predicted and measured values among five models

定量评估结果也印证了这一结论：在 4 组测试集上，CSSA-LSTM-CatBoost 模型的平均 RMSE 较 GBDT 降低 43.1%，较 BP-SVM 降低 39.7%；R^2 值从 GBDT 的 0.75～0.80 和 BP-SVM 的 0.72～0.78 提升至 0.91～0.94，充分证明了改进方法在复杂工业场景中的有效性。

如表 5 所示，消融实验数据得出基于标准麻雀搜索算法(SSA)优化的 LSTM 模型 LSTM-SSA 虽通过全局搜索改善了 LSTM 的时序建模能力，但受限于 SSA 算法易陷入局部最优的缺陷，仍存在优化空间。该模型对风机温度序列中非平稳工况的适应能力较弱，例如在功率快速波动阶段，预测值与真实值的滞后偏差导致误差指标偏高。

在 LSTM-SSA 基础上引入 CatBoost 梯度提升模块从原始温度序列中提取了具有高区分度的时序特征，并将其与 LSTM-CatBoost 模型结合为 TFE-LSTM-CatBoost。实验数据显示，该模型 E_{MAE} 降至 1.095，R^2 提升至 98.83%，验证了 TFE 模块在特征维度上的降维增效能力。具体而言，熵值法通过量化特征信息量，动态调整灰色关联分析的权重分配，使得模型更聚焦于与温度变化强相关的特征子集，从而增强了特征与目标变量的非线性映射能力。

表 5　单一模型和组合模型的预测误差结果

Table 5　Prediction error results of single model and combined model

模型	E_{MSE} /%	E_{MAE} /%	R^2 /%	训练时间 /min
FE-LSTM	22.04	14.86	97.36	43
FE-CatBoost	19.17	27.31	97.77	20
LSTM-CatBoost	19.94	12.18	98.58	65
FE-LSTM-CatBoost	18.04	10.95	98.83	58
FE-CSSA-LSTM-CatBoost	11.83	7.05	99.49	46

通过熵值法改进的灰色关联分析(TFE)模型 LSTM-CatBoost 后，尽管 E_{MSE} 略有上升(199.35)，但 E_{MAE} 显著降低至 1.218 0，且 R^2 提升至 98.58%。这一现象表明，CatBoost 通过迭代拟合 LSTM 预测残差，有效缓解了单一

LSTM 模型对极端值的敏感性。然而,由于未针对时序特征进行针对性优化,模型在处理高维动态特征时仍存在信息冗余,导致 RMSE 指标未进一步改善。

最终提出的 TFE-LSTM-CatBoost-CSSA 模型通过混沌映射初始化 SSA 种群,显著提升了算法的全局搜索能力。相较于基线 LSTM-SSA 模型,其 E_{MSE} 降低 39.0%,E_{MAE} 降低 52.6%,R^2 提升 2.13%。通过 Logistic 混沌映射生成初始种群,打破了标准 SSA 算法的均匀分布假设,增强了种群多样性,避免了算法早熟收敛;CSSA 算法同时调优 LSTM 的隐藏层神经元数量、学习率与 CatBoost 的树深度、学习率等超参数,实现了时序建模与梯度提升模块的参数联动优化。

5 结 论

本文针对风机健康状态预测中特征提取不充分及单一模型预测精度受限的问题,提出了一种基于混沌优化策略的混合深度学习架构 CSSA-CatBoost-LSTM 模型。首先,通过对风机传感器特征与时间戳特征进行多维度解析与交互融合,构建了更具表征能力的复合特征集,提升了特征维度与信息丰富度。其次,引入熵权修正的灰色关联分析方法,动态评估特征重要性,实现了特征降维与关键特征提取,进一步优化了模型输入。随后,设计了 CatBoost 梯度提升树与长短期记忆网络的混合架构,发挥了 CatBoost 在特征交互挖掘与 LSTM 在时序数据处理方面的优势。为优化 LSTM 模型超参数,本文引入了混沌映射改进的麻雀搜索算法(CSSA),提升了算法的全局搜索能力与收敛速度。实验结果表明,无论是在风机正常运行状态或者异常运行状态,CSSA-CatBoost-LSTM 组合模型均能保持较高的预测精度。

本研究为风机故障预警提供了兼具理论创新性与工程实用性的解决方案,其技术路径可推广至其他旋转机械状态监测领域。受限于单一企业数据集,模型在多工况跨领域应用时的适应性仍需验证,未来将结合迁移学习技术拓展应用场景。

参考文献

[1] 王慧慧. 先进机械传动系统在工业 4.0 背景下的优化策略——以电动汽车驱动系统为例[J]. 汽车维修技师,2025(4):100-101.
WANG H H. Optimization strategies for advanced mechanical transmission systems in the context of industry 4.0—taking electric vehicle drive systems as an example [J]. Auto Maintenance Technician,2025(4):100-101.

[2] 李乐. 基于机器学习的火电厂风机故障预警方法[J]. 自动化应用,2024,65(14):183-185.
LI L. A fault early warning method for fans in thermal power plants based on machine learning [J]. Automation Application,2024,65(14):183-185.

[3] 王洋,迟耀丹,杨浩远,等. 基于灰色预测模型的风机机侧控制研究[J]. 电器工业,2025(6):24-28.
WANG Y,CHI Y D,YANG H Y,et al. Research on wind turbine generator-side control based on grey prediction model[J]. Electrical Equipment Industry,2025(6):24-28.

[4] 陈冬玲,刘有冠. 基于灰色预测模型的焦炉煤气风机动态寿命预测研究[J]. 轻工科技,2017,33(1):48-49.
CHEN D L,LIU Y G. Research on dynamic life prediction of coke oven gas fan based on grey prediction model [J]. Light Industry Science and Technology,2017,33(1):48-49.

[5] 郝炎军,李庆华. 基于 PCA-ALOCO-SVM 模型的引风机状态预测及预警研究[J]. 机电信息,2024(7):1-5.
HAO Y J,LI Q H. Research on induced draft fan condition prediction and early warning based on PCA-ALOCO-SVM model[J]. Mechanical and Electrical Information,2024(7):1-5.

[6] 陈维刚,张会林. 基于 RF-LightGBM 算法在风机叶片开裂故障预测中的应用[J]. 电子测量技术,2020,43(1):162-168.
CHEN W G,ZHANG H L. Application of RF-LightGBM algorithm in fault prediction of wind turbine blade cracking[J]. Electronic Measurement Technology,2020,43(1):162-168.

[7] 宋威,丁一,赵凯,等. 基于 EMD-LSTM 的风机故障停机发生时间预测[J]. 计算机仿真,2023,40(12):113-118.
SONG W,DING Y,ZHAO K,et al. Time prediction of wind turbine fault shutdown based on EMD-LSTM[J]. Computer Simulation,2023,40(12):113-118.

[8] 马明骏,赵海心,姜孝谟,等. 基于 LSTM-WPHM 模型的风机轴承故障报警与寿命预测方法[J]. 风机技术,2022,64(3):63-71.
MA M J,ZHAO H X,JIANG X M,et al. Fault alarm and life prediction method for wind turbine bearings based on LSTM-WPHM model[J]. Wind Turbine Technology,2022,64(3):63-71.

[9] 刘伟霞,田勋,肖家勇,等. 基于混合模型及 LSTM 的锂电池 SOH 与剩余寿命预测[J]. 储能科学与技术,2021,10(2):689-694.
LIU W X,TIAN X,XIAO J Y,et al. Prediction of lithium battery SOH and remaining useful life based on hybrid model and LSTM[J]. Energy Storage Science and Technology,2021,10(2):689-694.

[10] 卢鹏,年圣全,邹国良,等. 基于深度学习和 CatBoost 的海浪波高预测方法研究[J]. 海洋湖沼通报(中英文),2024,46(5):28-34.
LU P,NIAN SH Q,ZOU G L,et al. Research on wave

height prediction method based on deep learning and CatBoost[J]. Transactions of Oceanology and Limnology (in Chinese and English),2024,46(5):28-34.

[11] 袁建华,邵星,王翠香,等. 基于 AMCNN-BiLSTM-CatBoost 的滚动轴承故障诊断模型研究[J]. 噪声与振动控制,2025,45(2):82-89.
YUAN J H,SHAO X,WANG C X,et al. Research on rolling bearing fault diagnosis model based on AMCNN-BiLSTM-CatBoost[J]. Noise and Vibration Control,2025,45(2):82-89.

[12] 魏玮,吕游,齐欣宇,等. 基于 CNN-LSTM-AM 动态集成模型的电站风机状态预测方法[J]. 仪器仪表学报,2023,44(4):19-27.
WEI W,LYU Y,QI X Y,et al. Condition prediction method for power station fans based on CNN-LSTM-AM dynamic ensemble model[J]. Chinese Journal of Scientific Instrument,2023,44(4):19-27.

[13] 崔新苗,孙渊. 基于 CEEMDAN-ISSA-BPNN 组合模型的风机功率短期预测[J]. 上海电机学院学报,2023,26(3):141-146.
CUI X M, SUN Y. Short-term prediction ofwind turbine power based on CEEMDAN-ISSA-BPNN combined model [J]. Journal of Shanghai Dianji University,2023,26(3):141-146.

[14] 尚凤玲. 循环风机轴封机构改造[J]. 化工设计通讯,2017,43(10):93.
SHANG F L. Renovation of shaft seal mechanism for circulating fans [J]. Chemical Engineering Design Communications,2017,43(10):93.

[15] 郭壮丽,白琳,程英辉,等. 风机常见故障及智能诊断分析[J]. 现代制造技术与装备,2023,59(5):128-131.
GUO ZH L,BAI L,CHENG Y H,et al. Common faults and intelligent diagnosis analysis of fans[J]. Modern Manufacturing Technology and Equipment,2023,59(5):128-131.

[16] 赵金平,周胜伟. 基于 LSTM 的风机高速轴温度预测及预警方法研究[J]. 大电机技术,2023(S2):16-22.
ZHAO J P,ZHOU SH W. Research on temperature prediction and early warning method for high-speed shaft of fans based on LSTM[J]. Large Electric Machine and Hydraulic Turbine,2023(S2):16-22.

[17] 万安平,龚志鹏,张银龙,等. 基于 XGBoost-KDE 的风机主轴承温度预测与故障预警方法[J]. 热力发电,2022,51(12):164-171.
WAN AN P, GONG ZH P, ZHANG Y L, et al. Temperature prediction and fault early warning method for main bearings of fans based on XGBoost-KDE[J]. Thermal Power Generation, 2022, 51(12): 164-171.

[18] 薛建凯. 一种新型的群智能优化技术的研究与应用[D]. 上海:东华大学,2020.
XUE J K. Research and application of a novel swarm intelligence optimization technique [D]. Shanghai: Donghua University,2020.

[19] 郭雯博. 基于相似日和 Tent-SSA-BP 光伏发电功率预测研究[J]. 信息技术与信息化,2023(5):173-177.
GUO W B. Research on photovoltaic power generation forecasting based on similar days and Tent-SSA-BP[J]. Information Technology & Informatization, 2023 (5): 173-177.

[20] 陈浚锃,刘桂雄,谢方静. 既有办公建筑光伏发电预测的 SSA-LSTM 方法研究[J]. 电子测量技术,2025,48(6):171-178.
CHEN J K, LIU G X, XIE F J. Research on SSA-LSTM method for photovoltaic power generation forecasting in existing office buildings[J]. Electronic Measurement Technology,2025,48(6):171-178.

[21] 回立川,陈雪莲,孟嗣博. 多策略混合的改进麻雀搜索算法[J]. 计算机工程与应用,2022,58(16):71-83.
HUI L CH, CHEN X L, MENG S B. An improved sparrow search algorithm with multi-strategy hybridization [J]. Computer Engineering and Applications,2022,58(16):71-83.

[22] 徐恒山,莫汝乔,薛飞,等. 基于时间戳特征提取和 CatBoost-LSTM 模型的光伏短期发电功率预测[J]. 太阳能学报,2024,45(5):565-575.
XU H SH, MO R Q, XUE F, et al. Short-term photovoltaic power generation forecasting based on timestamp feature extraction and catboost-lstm model[J]. Acta Energiae Solaris Sinica,2024,45(5):565-575.

[23] 王雁凌,吴梦凯,周子青,等. 基于改进灰色关联度的电力负荷影响因素量化分析模型[J]. 电网技术,2017,41(6):1772-1778.
WANG Y L, WU M K, ZHOU Z Q, et al. Quantitative analysis model of power load influencing factors based on improved grey relational degree[J]. Power System Technology,2017,41(6):1772-1778.

作者简介

孙家栋,硕士,讲师,主要研究方向为工业物联网、模式识别与智能系统等。
E-mail:sunjiadong@cwxu.edu.cn

李子恒,硕士研究生,主要研究方向为工业互联网。
E-mail:202312490339@nuist.edu.cn

施珮,博士,副教授,主要研究方向为智慧交通、物联网。
E-mail:shipei@cwxu.edu.cn

陈德基(通信作者),博士,教授,主要研究方向为实时系统、工业物联网。
E-mail:dejichen@cwxu.edu.cn