引用格式：陶冠宏, 张婉渝, 许文波, 等. 基于 Shapley 值的可解释 AI 在风机齿轮箱健康监测与故障定位中的应用 [J]. 电子科技大学学报, 2025, 54(6): 924-934.
TAO G H, ZHANG W Y, XU W B, et al. Application of shapley-value based explainable AI in health monitoring and fault localization for wind turbine gearboxes[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(6): 924-934.

基于 Shapley 值的可解释 AI 在风机齿轮箱健康监测与故障定位中的应用

陶冠宏[1,2]，张婉渝[2]，许文波[1*]，范振军[2]

（1. 电子科技大学 资源与环境学院，成都 611731；2. 成都天奥集团有限公司，成都 611731）

摘要：在风力发电领域，风机齿轮箱健康状态直接影响风电机组的发电量，而当前基于领域知识和数据驱动的齿轮箱故障诊断与定位技术受限于领域知识的完备性、数据量不足及算法透明度不足。为解决此问题，提出了一种既具有学习能力又能提供可解释输出的可解释人工智能（explainable AI, XAI）框架。通过将 Shapley 值分析法引入无监督和监督学习算法中实现算法改进，缓解模型对数据量的过度依赖，同时增强模型的可解释性。实验通过两个典型风机齿轮箱案例验证了该框架的有效性：案例 1 结果表明，相较于无监督和监督学习算法，所提出的框架在数据标签稀缺情况下显著提升了聚类效果；案例 2 结果表明，所提出的框架通过模型可解释性分析，实现风机齿轮箱故障成因定位，能够为齿轮箱故障预防与维护提供指导性建议。实验结果展示了“知识+数据”结合方式在工程应用中的显著效果，为可解释人工智能的落地应用提供有价值的参考。

关键词：风力发电；齿轮箱故障；可解释 AI（XAI）；Shapley 值；聚类
中图分类号：TP391 **文献标志码：**A DOI: 10.12178/1001-0548.2024200

Application of shapley-value based explainable AI in health monitoring and fault localization for wind turbine gearboxes

TAO Guanhong[1,2], ZHANG Wanyu[2], XU Wenbo[1*], and FAN Zhenjun[2]

(1. School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. Chengdu Spaceon Group Co., Ltd., Chengdu 611731, China)

Abstract: In the field of wind power generation, the health status of wind turbine gearboxes directly impacts the power output of wind turbine units. Current gearbox fault diagnosis and localization techniques, which are based on domain knowledge and data-driven approaches, are constrained by the completeness of domain knowledge, insufficient data volume, and lack of algorithm transparency. To address this issue, we propose an explainable AI framework that possesses both learning capabilities and provides interpretable outputs. By incorporating the Shapley value analysis method into unsupervised and supervised learning algorithms, the framework achieves improvements, alleviating the model's excessive dependence on data volume and enhancing the model's interpretability. The effectiveness of the proposed framework was validated through experiments on two typical wind turbine gearbox cases. The results of case 1 indicate that, compared to unsupervised and supervised learning algorithms, the proposed framework significantly improves clustering performance in situations with scarce data labels. The results of case 2 demonstrate that the framework, through model interpretability analysis, achieves the localization of wind turbine gearbox fault causes, providing guiding suggestions for gearbox fault prevention and maintenance. The experimental results showcase the significant effectiveness of the 'knowledge + data' integration approach in engineering applications, offering valuable references for the practical implementation of explainable artificial intelligence.

Key words: wind power generation; gearbox fault; explainable AI (XAI); Shapley value; clustering

风力发电作为一种清洁可再生能源，在全球能源供应体系中的地位日益显著。根据国际能源署的报告，风电是增长最为迅猛的能源之一，2023 年总装机量为 117 GW，较 2022 年增长了 50%，并

收稿日期：2024−07−31
基金项目：中电天奥产业发展基金（202201090404）
作者简介：陶冠宏，博士，高级工程师，主要从事模式识别、计算机视觉等方面的研究。
*通信作者 E-mail: xuwenbo@uestc.edu.cn

预计到 2030 年，全球风电容量将至少扩展至 320 GW[1]。风力发电的效率与可靠性在很大程度上取决于其核心组件——齿轮箱的性能。作为风力发电机的关键部件，齿轮箱的健康状况直接影响整个系统的稳定运行。

齿轮箱故障是导致风电机组停机的常见原因之一，尽管其发生频率相对较低，但一旦发生，往往造成长时间停机，并伴随高昂的维修与更换成本，对发电效率和运营经济性构成显著威胁[2]。因此，研究有效的齿轮箱故障诊断技术，不仅能够提前识别潜在问题、减少停机时间、降低维护成本，还能显著提升风力发电系统的整体效率与可靠性。

这一领域近年来已受到学术界和企业界的广泛关注。传统齿轮箱故障诊断方法主要依赖领域知识，通过专家经验和预设规则构建模型。这类方法具有明确的逻辑结构和高可解释性，如基于油液分析[3]、振动信号分析[4]和声发射分析[5]的诊断技术，能够清晰揭示故障机理。然而，这些方法的局限性在于处理复杂多变的实际场景能力不足。由于高度依赖专家知识，其适应性较差，且在面对高维、噪声干扰严重的运行数据时，诊断效果往往受限。近年来，人工智能技术凭借其强大的数据处理能力，在齿轮箱故障诊断中展现出显著优势。特别是基于数据迭代学习的方法，如高斯混合模型[6]、支持向量机[7]和神经网络[8]，通过聚类、分类和回归等手段，为齿轮箱故障诊断和预测提供了更精确、自动化的解决方案。

具体而言，聚类和分类方法利用无监督和有监督学习技术，对特征数据进行分组或分类，以识别齿轮箱故障相关的模式。如文献 [9] 提出了一种改进的狮群优化密度聚类算法，用于检测风电齿轮箱异常状态，相较于 K-means 算法表现出更高的准确性。文献 [10] 构建了基于深度残差网络（ResNet）的故障诊断模型，通过建立故障特征与类别间的非线性映射，实现齿轮箱故障的分类识别。文献 [11] 则提出了结合长短时记忆网络（long short-term memory, LSTM）与支持向量机的混合模型，进一步提升了诊断精度。

另一方面，回归方法通过构建监测变量（如传感器读数、运行条件）与输出指标（如齿轮箱健康状态或温度）之间的关系模型，实现故障预警。如文献 [12] 结合自适应共振理论提出神经网络和高斯混合模型，有效检测风力涡轮机的故障及预故障状态。文献 [13] 则融合 Xgboost、LightGBM 和 LSTM 模型的优势，预测齿轮箱轴承温度，从而实现故障预警。

通过分析上述研究趋势，可以清晰梳理齿轮箱故障诊断技术的发展脉络：从依赖领域知识的传统方法，逐步转向基于数据迭代学习的现代方法（如深度学习）。传统方法在可解释性方面具有优势，但因主观性强、量化能力不足，常导致诊断结果不一致，且其数据处理能力有限，难以应对高维、复杂或噪声数据，同时对专家知识的过度依赖使其难以适应快速变化的运行条件，限制了大规模应用。相比之下，数据驱动的迭代学习方法通过大规模数据训练，能够捕捉复杂模式和细微特征，显著提升诊断精度。然而，这些方法通常表现为“黑箱”特性，缺乏可解释性，导致模型决策过程的不透明性和复杂性增加，降低了其在实际应用中的可信度。

近年来，可解释人工智能（explainable AI, XAI）在提升模型性能和揭示决策过程等方面发挥了至关重要的作用。通过“知识+数据”的结合策略，XAI 不仅增强了模型的透明度和实用性，还为破解深度学习“黑箱”问题提供了有效途径。其核心目标是通过技术手段使 AI 模型的决策过程被人类可理解。具体而言，XAI 借助基于博弈论的 Shapley 值和局部可解释模型无关方法（local interpretable model-agnostic explanations, LIME）等技术，系统分析特征对预测结果的影响机制。如 Shapley 值通过量化振动信号、温度参数等特征对故障诊断的边际贡献，融合领域知识的逻辑性与数据驱动的统计规律，保留关键物理参数、剔除冗余干扰，从而提升诊断的可解释性。文献 [14] 展示了 Shapley 值在监督聚类分析中的应用，通过揭示 COVID-19 症状学研究中的子群体信息，验证了其在解析复杂数据结构中的价值。同样，文献 [15] 研究了 XAI 在机器故障诊断中的应用，利用 Shapley 值进行特征选择，有效减少特征数量，提升了模型效率和性能。而 LIME 方法则通过对特征扰动采样构建局部代理模型，提取针对特定预测样本的解释规则，既保留了数据驱动的优势，又增强了规则的可解释性。如文献 [16] 提出了一种基于改进 LIME 的时间–频率图像质量评估方法，用于评估旋转机械振动信号的质量，从模型输入端优化了故障诊断效果。文献 [17] 则通过 LIME 量化 1DCNN 模型中故障预测的特征影响权重，进一步提升了模型的可信度和透明度。

然而，在故障诊断领域，现有研究往往聚焦于诊断精度或可解释性的单方面优化，缺乏对故障定位的深入探索。如文献 [15] 通过 Shapley 值实现特征简化并提升性能，文献 [16] 借助 LIME 改进诊断效果，但两者均未充分解决故障定位问题，且适用范围局限于监督学习场景。针对这些不足，本文提出了一种融合知识与数据驱动的可解释人工智能框架。该框架通过 Shapley 值量化特征贡献，指导无监督聚类的标签生成，在标注数据不足时显著减少对人工标注的依赖；在标注数据充足时，进一步整合 Shapley 值技术，深入分析特征对模型决策的影响，精准定位故障成因，为故障预防与维护提供更具指导性的建议，从而提升模型决策的透明度和可信度。此外，结合蒙特卡洛采样策略，该框架将 Shapley 值计算复杂度从传统的 $O(2^N)$ 降至 $O(KN)$，有效缓解了高维数据场景下的实时性瓶颈。

为验证所提框架的有效性，本文通过两个典型风机齿轮箱案例进行实验分析。在标注数据不足的场景下，结合无监督聚类与有监督分类识别齿轮箱健康状态，并引入 Shapley 值优化聚类效果；在标注数据充分的场景下，通过 Shapley 值分析影响轴承温度的特征贡献度，定位故障类型与原因，从而验证框架的实用性与优越性。

1 研究方法

本文提出的可解释人工智能分析框架，利用 Shapley 分析实现了既能够迭代学习又能提供可解释输出的功能，兼顾不同规模的标注数据量，适配不同情境下的需求。框架如图 1 所示，首先数据经过统一的预处理后，在标签稀缺的情况下，一方面利用聚类模型挖掘数据内在结构性信息，从而获取相应的类别划分；另一方面通过构建基分类模型，计算大量无标签数据的 Shapley 值，进而结合聚类模型，实现对无标签数据的标签预测。当数据标签充足时，则构建完全监督学习模型，充分利用标签信息的完整性，为实现高效的故障定位与诊断提供科学依据。

本框架的核心优势在于可解释 AI 方法在风机齿轮箱状态监测与故障定位中的创新应用，该技术允许模型在不同的学习场景中有效地捕捉和解释特征与故障之间的关系。在数据标签稀缺或充裕条件下，该框架均能提供稳定且可靠的预测性能。目前经典的解释方法主要包括 SHAP、LIME 以及 DeepLIFT 等，其中，SHAP 和 LIME 属于模型无关的解释方法，而 DeepLIFT 则是一种与模型相关的解释方法。鉴于实际应用中模型的类型和结构具有多样性，选择模型无关的方法（如 SHAP 和 LIME）能够确保解释过程不受特定模型架构的限制。具体而言，SHAP 方法具有坚实的数学理论基础，并且能够从全局和局部两个层面对模型进行解释，为深入理解模型的决策机制提供了有力的工具。相比之下，LIME 主要专注于局部解释，对全局特征重要性的分析能力较弱。因此，本文选择 SHAP 作为主要的解释方法，并以 LIME 为对比，深入探讨 SHAP 解释方法在故障定位任务中的有效性和稳定性。通过这种对比分析，旨在为风机齿轮箱故障诊断提供更具参考价值的解释性分析策略。

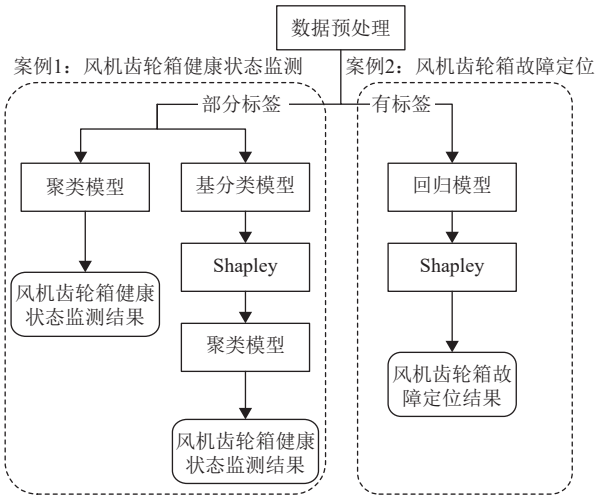

图 1 框架流程图

SHAP 是基于 Shapley 值的一种模型解释方法，能够同时捕捉局部和全局特征的可解释人工智能技术，最初由美国加州大学教授 Lloyd Shapley 提出[18]，用于解决合作博弈中公平分配收益的问题。在一场包含多个玩家共同参与的博弈中，不同玩家所获得的奖励会有所不同，这是因为部分玩家对博弈的贡献较大，而部分玩家的贡献相对较小。Shapley 值是一个能够根据玩家的不同贡献，将总奖励公平地分配给每个玩家的度量指标。

Shapley 值表示玩家a对总收益的贡献：

$$\phi(a)=\sum_{S\subseteq N\setminus\{a\}}\frac{|N|!(|N|-|S|-1)!}{|N|!}[v(S\cup\{a\})-v(S)] \quad (1)$$

式中，$N=\{a,b,c,\cdots\}\in\mathbb{R}^N$，玩家组合的子集$S\in N$是不包含玩家$a$的集合；$v(S)$表示该子集 S 中玩家参与博弈时所获得的奖励值。

Shapley 值具有以下重要性质。

1）公平性：每个玩家的 Shapley 值反映了其对总收益的公平贡献。

2）对称性：如果两个玩家在所有子集中的贡献相同，则他们的 Shapley 值相等。

3）线性性：Shapley 值对收益函数是线性的，即$\phi(a+b)=\phi(a)+\phi(b)$。

在机器学习和数据科学领域，Shapley 值可应用于量化数据特征在预测模型中的贡献度，如在监督学习中，Shapley 值可用于解释模型的预测结果，对于每个特征，Shapley 值表示其对模型预测的贡献；在无监督学习中，Shapley 值可以用于提升聚类或降维效果；在强化学习中，Shapley 值可以用于解释智能体的决策过程[18]。

在机器学习领域，对于特征i和样本x的 Shapley 值定义为[19]：

$$\phi_i(f,x)=\sum_{S\subseteq F\setminus\{i\}}\frac{|s|!(|m|-|s|-1)!}{|m|!}[f_x(S\cup\{i\})-f_x(S)] \quad (2)$$

式中，F表示包含所有特征的集合；m是特征的总数；而S是一个排除了特征i的子集；s是集合S中特征的数量；$f_x(S\cup\{i\})$是在包含特征i的条件下，模型f对样本x的预测值；$f_x(S)$是在不包含特征i的条件下，模型f对样本x的预测值。这个公式表示了特征i对样本x预测的边际贡献。

但是，精准计算 Shapley 值的成本很高，时间复杂度随特征维度的增加呈指数级增长，这使得在处理高维数据时，精确计算变得异常困难。为了解决这一问题，研究者们提出了多种近似计算方法，旨在降低计算复杂度并提高计算效率。蒙特卡洛采样通过随机特征排列估计边际贡献均值[20]，通用性强但收敛相对较慢；KernelSHAP 将问题转化为加权线性回归[19]，理论完备但高维效率低；TreeSHAP 利用树结构特性实现线性复杂度[21]，但仅限树模型；分层抽样通过重要性预计算减少采样量[22]，需额外开销；特征分组合并相关特征降低维度[23]，但依赖先验知识；动态规划通过缓存中间结果加速[24]，实现复杂且内存占用高。

若需在高维、模型不可知的场景下平衡通用性与实现成本，蒙特卡洛采样是较优选择。尽管其计算复杂度为 O（Km），其中 K 代表采样次数，m 代表特征数量，但它不依赖模型结构假设（如 TreeSHAP 对树模型的限制），无须预计算特征重要性（如分层抽样）或分组（如特征分组），可直接应用于神经网络、集成模型等黑箱系统。尤其在解释性需求优先于实时性的场景（如医疗、金融风险评估、设备健康监测等），蒙特卡洛的模型无关性和理论可解释性使其成为最优解决方案。

因此，本文选择文献 [20] 提出的无偏随机顺序蒙特卡洛估计器 IME（incremental meta-knowledge estimator）作为计算 Shapley 值近似方法，该算法通过随机排列特征，确保每个特征的贡献被公平地估计，从而保证了估计值的无偏性。并结合伪随机采样和自适应采样，进一步提高计算效率和精度。IME 算法具体公式如下：

$$\hat{\phi}_i(f,x)=\frac{1}{K}\sum_{k=1}^{K}\Delta f_i(\sigma_k) \quad (3)$$

式中，$\hat{\phi}_i(f,x)$表示样本x中特征i的平均边际贡献，f是预测模型；$\Delta f_j(\sigma_k)$是特征i在特征序列σ_k上的边际贡献；K是采样次数。

上式中得到了单个样本x的特征i的平均边际贡献，那么对于一个样本原始预测的解释，则存在以下近似：

$$f(x)\approx\hat{f}(x)=\hat{\phi}_0+\sum_{i=1}^{m}\hat{\phi}_i(f,x) \quad (4)$$

式中，$f(x)$是真实预测值；$\hat{f}(x)$是蒙特卡洛采样近似预测值；$\hat{\phi}_0$是基准预测值；m是特征总数。

相比于传统 Shapley 值须计算所有 2^m 个子集的边际贡献，该方法单次排列的计算复杂度为 O（m），总复杂度降为 O（Km）。一般情况下（$K\geqslant10^4$）可满足多数高维场景（如 m=20 时误差<5%），能够在不牺牲太多准确性的情况下，有效地处理高维数据集[20]，为特征重要性提供一个近似的评估，同时保持计算的可行性。由于每个排列的计算相互独立，可通过多线程并行处理提高计算效率。如 1 万次采样在 100 核集群上仅需约 100 次串行计算时间。权衡计算效率和估计精度，本文采用 IME 算法通过式 (3) 和式 (4) 重复K次近似计算 Shapley 值，从而在高维数据场景下实现高效且准确的特征重要性评估。

2　案例 1：风机齿轮箱健康状态监测

本案例聚焦于风力发电机齿轮箱的健康状态监测问题，基于有限标注样本，利用无监督聚类和有监督分类实现对齿轮箱健康状态识别，然后通过引入 Shapley 值进行聚类分析实现效果提升，以此验

证 Shapley 值在聚类中的效用。

2.1 实验数据集构建

本案例数据来源于东南大学 2023 年发布的风机行星齿轮箱的数据集[25]。该数据集涵盖了齿轮箱的 5 种状态，包括健康、断齿、齿轮磨损、齿根裂纹和缺齿，具体状态如图 2 所示。

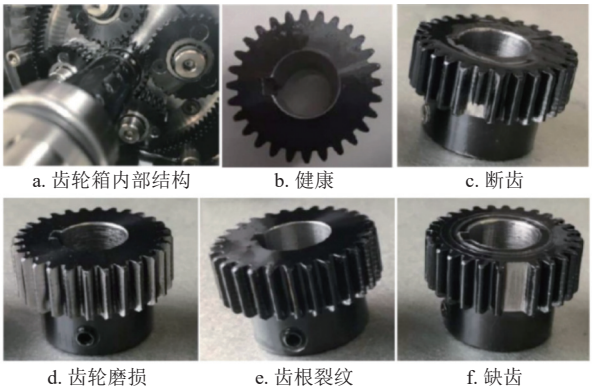

图 2 齿轮箱及齿轮状态示意图[25]

数据采集是通过安装在齿轮箱外壳的加速度计和输入轴的编码器完成的。每个数据样本包含 4 个字段，但由于第 4 个字段的含义不明确，本案例仅使用了前 3 个字段：传感器在 X 轴和 Y 轴上的加速度信号以及齿轮的转速。实验从 5 种状态的数据中，每类随机等比例地选取了 4 万条数据，总计 20 万条数据作为实验数据集。

2.2 实验步骤

为了验证 Shapley 值分析在有限标注样本条件下提升齿轮箱健康状态的聚类效果，将实验数据进行分层抽样形成 20 万、10 万、5 万这 3 种不同规模数据集，每个数据集的实验遵循以下步骤。

1）数据集划分：将每个数据集以 2∶8 的比例划分为小样本训练集 S 和大样本测试集 T。训练集 S 用于有监督模型的训练和 Shapley 值的计算，而测试集 T 则用于评估和对比不同方法的性能。

2）无监督聚类：对测试集 T 中的无标签样本应用多种算法进行无监督聚类，得到初步的聚类结果 $R1$。

3）有监督模型训练：利用有标签的小样本训练集 S，训练一个 Xgboost 基分类模型 M，该模型将用于后续的预测和 Shapley 值计算。

4）模型预测与 Shapley 值计算：使用训练好的模型 M 对测试集 T 进行预测得到结果 $R2$，并计算每个样本的 Shapley 值。

5）Shapley 值聚类：基于计算得到的 Shapley 值，在测试集 T 上再次应用聚类算法进行聚类，得到基于 Shapley 值的聚类结果 $R3$。

6）结果对比与评估：分别计算 K-means 聚类结果 $R1$、模型 M 的预测结果 $R2$ 以及基于 Shapley 值的 K-means 聚类结果 $R3$ 与真实测试集 T 标签之间的归一化互信息（normalized mutual information, NMI）值。通过比较 NMI 值，评估不同方法在聚类和预测准确性方面的表现。

2.3 实验结果与分析

1）无监督聚类结果 $R1$

在无监督聚类中，针对不同规模数据集，以测试集 T 中的样本作为实验数据集，分别应用 K-means、GMM 和 Spectral 聚类算法进行聚类，聚类算法超参数如表 1 所示。

表 1 聚类模型参数表

模型	参数名称	参数值
K-means	簇数K	5
	初始化方法	K-means++
	初始化次数	10
	最大迭代次数	1 000
GMM	簇数K	5
	初始化方法	K-means
	初始化次数	10
	最大迭代次数	1 000
Spectral	簇数K	5
	初始化方法	K-means
	初始化次数	10

K-means 作为一种广泛应用的聚类算法，具有较高的知名度和理解度。为了提升对聚类结果的理解并实现更为直观的数据可视化，本研究采用 3D 散点图的形式对基于 20 万数据集的 K-means 聚类分析的结果进行了展示。这种方法不仅简化了复杂数据集的解析过程，同时也为观察不同簇之间的关系提供了清晰且直观的视角。作为对比，图 3 展示了测试集 T 真实标签分布情况，图 4 展示了基于原始特征的 K-means 聚类结果，其中 X、Y、Z 轴分别代表 X、Y 方向的加速度值和齿轮转速值。

对比图 3 和图 4 可以观察到，原始特征经过 K-means 聚类处理后陷入局部最优解，而非全局最优，使得聚类结果未能充分反映数据的真实结构，与真实的标签分布存在显著偏差。原因是 K-means 聚类算法在划分簇时，主要依据样本点之间的距离，而忽略了样本的其他潜在特性。因此，聚类结

果的 NMI 值仅为 20.57%，这表明聚类产生的分布与真实标签的分布之间的一致性相当低。

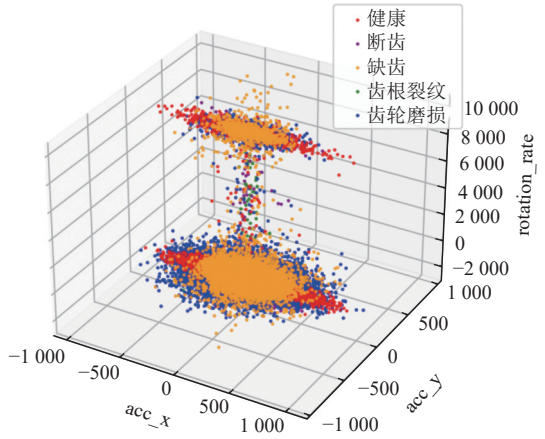

图 3　测试集 T 数据标签三维分布示意图

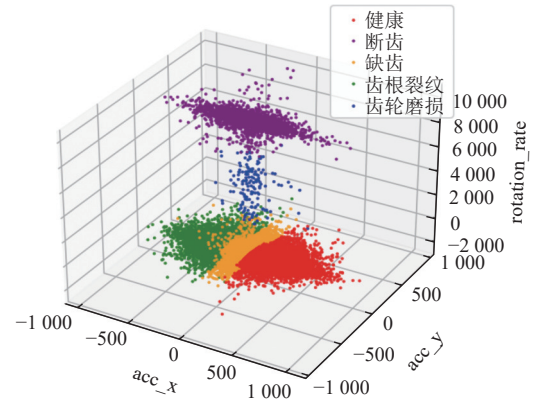

图 4　基于原始特征的 K-means 聚类结果三维分布示意图

2）有监督分类预测结果 $R2$

通过构建 Xgboost 模型对训练集 S 进行有监督训练，得到风机齿轮箱健康状态基分类模型 M，该模型的超参数如表 2 所示。

表 2　Xgboost 分类模型参数表

参数名称	参数值
基学习器	gbtree
基学习器个数	1 000
学习率	0.1
树最大深度	5
子节点权重	1

通过对分类结果进行统计，图 5 展示了 Xgboost 模型在测试集 T 上的有监督分类结果。

从图 4 和图 5 的分布对比来看，分类结果的分布和实际标签的分布具有更强的相似性，说明 Xgboost 模型不仅成功地捕捉到了样本数据中的潜在特征，而且准确地反映了真实的健康状态分布。预测结果与真实标签之间的 NMI 值为 75.07%，表明模型预测性能良好，证实了基分类模型能够较好地区分和识别齿轮箱的不同健康状态。

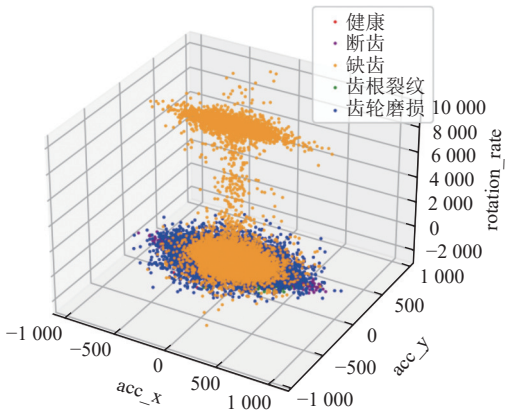

图 5　Xgboost 模型的有监督分类结果三维分布示意图

3）基于 Shapley 值的聚类结果 $R3$

利用已经训练完成的 Xgboost 模型，依据式 (3) 和式 (4) 对测试集 T 进行了 Shapley 值计算，以识别和量化每个特征对模型预测的贡献度。为了确保与原始特征聚类结果的可比性，采用与之前聚类算法相同的参数设置（如表 1 所示），将计算得到的 Shapley 值作为输入并进行聚类，其中 K-means 聚类的结果如图 6 所示。

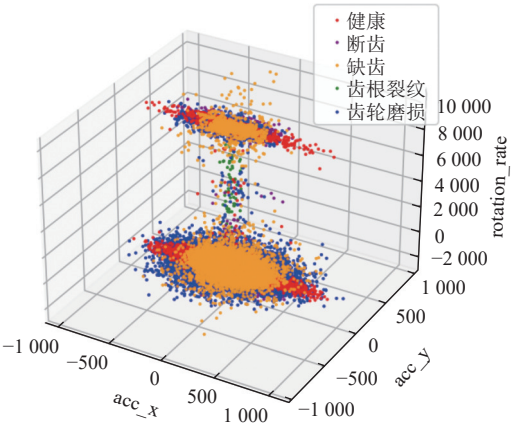

图 6　基于 Shapley 值的 K-means 聚类结果三维分布示意图

表 3 对 3 种不同的聚类方法（K-means、GMM 和 Spectral）在不同规模的数据集上的表现进行了比较。实验在 Intel(R) Core(TM) i7-10700 CPU @ 2.90 GHz 处理器上进行。对于 K-means 和 GMM 算法，均设定 1 000 次迭代，Spectral 算法由于没有迭代次数参数，则采用其默认设置。从结果可以看出，K-means 和 GMM 算法的计算效率相对较高，更适合处理大规模数据集。Spectral 算法涉及求解拉普拉斯矩阵和获取特征向量等计算密集型过

程，效率相对较低。因此，Spectral 算法更适合于小规模数据集。

通过对比 *R*1（原始特征 K-means 聚类）、*R*2（有监督分类）和 *R*3（基于 Shapley 值的 K-means 聚类）3 项实验结果，可以得出结论：1）无论是在 20 万、10 万还是 5 万的数据规模下，Shapley 值的聚类效果均优于原始特征和预测值，表明 Shapley 值能够更好地捕捉数据的内在结构，从而提供更准确的聚类结果；2）Shapley 值的聚类效果不受具体模型的影响，无论是 K-means、GMM 还是 Spectral 方法，Shapley 值都能提供稳定的聚类效果，体现了其模型无关性的特点。最后，不同模型计算效率差异显著，在实际应用中需进行权衡。

表 3　聚类模型 NMI 指标

模型	数据规模/万	原始特征	预测值	Shapley值	计算耗时/s
K-means	20	0.205 7	0.750 7	0.773 5	0.113
	10	0.212 4	0.762 1	0.769 7	0.063
	5	0.214 5	0.751 6	0.770 3	0.054
GMM	20	0.409 4	0.750 7	0.751 5	0.219
	10	0.412 4	0.762 1	0.779 1	0.182
	5	0.404 9	0.751 6	0.755 7	0.071
Spectral	20	0.574 3	0.750 7	0.764 0	7.935
	10	0.577 8	0.762 1	0.771 9	6.970
	5	0.491 4	0.751 6	0.758 9	0.239

3　案例 2：风机齿轮箱故障定位

本案例将以齿轮箱轴承温度为研究对象，构建齿轮箱轴承温度回归模型，并利用 Shapley 值对该模型的预测结果进行可解释性分析，锁定对温度影响的关键部件，进一步定位导致温度异常的具体因素，以实现齿轮箱故障事后定位。

3.1　实验数据集构建

该案例数据来自山东德州某风场采集的 SCADA（supervisory control and data acquisition）数据，数据涵盖 25 台风机，数据采集的时间是 2023 年 1 月至 2023 年 7 月，采集间隔为 10 min。根据风机运行原理从 SCADA 数据中选择以下数据项，如表 4 所示，作为本案例齿轮箱轴承温度的解释特征，并分别命名为 F1, F2,⋯, F12。

风力发电机齿轮箱作为能量传递的核心部件，其物理特性表现为多级行星齿轮的非线性动力学耦合、滚动轴承接触摩擦热累积以及润滑油承载的时变特性。故障发生本质上是机械能−热能−电磁能转换失衡的物理过程：叶轮转速（F1）与齿轮箱转速（F5）传动比异常反映齿轮啮合刚度退化；低速轴前后轴温度梯度（F3、F4）表征轴承滚道磨损引发的摩擦热分布畸变，齿轮箱入口油温（F6）与压力（F7）的变化揭示润滑失效过程，发电机转速（F8）/扭矩（F9）的谐波分量能暴露轴系中电磁转矩脉动，而偏航误差（F11）则能反映齿轮接触疲劳，变流器冷却介质温度（F12）异常则间接反映齿轮箱热边界条件的改变。这 12 个物理特征能够构建齿轮箱健康状态的跨尺度（机械−热−电）耦合表征体系，为数据驱动的故障诊断提供了物理可解释的特征空间。

表 4　SCADA 数据项与特征对应表

部件	SCADA数据项	特征名
叶轮	叶轮转速	F1
轮毂	轮毂温度	F2
低速轴	低速轴前轴温度	F3
	低速轴后轴温度	F4
齿轮箱	齿轮箱转速	F5
	轮箱入口油温	F6
	齿轮箱入口压力	F7
发电机	发电机转速	F8
	发电机扭矩	F9
变桨系统	桨叶角度偏差	F10
偏航系统	偏航误差	F11
散热系统	变流器冷却介质温度	F12

为确保数据质量和分析有效性，本研究基于选定特征对数据进行了预处理，包括异常处理、字段处理和标准化处理。异常处理针对停机数据，当有功功率平均值≤0 或风速小于 0 时，删除相关记录。字段处理中，定义桨叶角度偏差（Dev）为特征 F10，用于衡量桨叶 1、桨叶 2 和桨叶 3 角度的不一致性，其计算公式为：

$$\mathrm{Dev}=\frac{1}{3}\sum_{i=1}^{3}(|\theta_i-\mu|) \tag{5}$$

式中，θ_i代表第 i 个桨叶的角度；μ代表 3 个桨叶角度的均值。

最后，通过最大最小标准化处理，让数值归一到 [0,1] 区间，消除量纲影响。

标准化后的数据被划分为训练集和测试集，其中训练集包含 20 台正常工作的风机，共 51.84 万条运行记录；测试集则包括 4 台正常风机和 1 台故障风机，共计 2.16 万条记录。其中，在年度例行检修中，A3 号风机润滑油位过低导致润滑不足，风机无法正常发电而被判定为故障风机。

3.2　实验步骤

为了识别齿轮箱轴承的温度异常，并提出排查策略，定位异常的具体部件。实验将遵循以下步骤。

1）数据集构建：将风机 2023 年 1 月至 6 月的历史数据作为训练集，2023 年 7 月的数据作为测试集。

2）温度预测模型构建：对清洗后的训练集基于监督学习方法建立齿轮箱轴承温度预测模型。

3）异常发现：根据 3 倍标准差原则，定位异常风机及高温预警时间。

4）Shapley 值计算：输入轴承预测模型和测试集样本进行 Shapley 值计算。

5）排查策略及故障定位：基于 Shapley 值和 LIME 方法进行定位分析，提出排查策略并定位到导致异常发生的因素。

3.3　实验结果与分析

3.3.1　温度预测模型训练结果

基于表 4 中列出的 12 个可解释特征，本研究分别采用 Xgboost、CatBoost 和 Lightgbm 算法构建了齿轮箱轴承温度预测模型。该模型选取平均绝对误差（mean absolute error, MAE）作为损失函数以优化预测性能。模型参数和模型评估指标如表 5 和表 6 所示，最终选择 Xgboost 作为预测模型。

表 5　Xgboost、CatBoost 和 Lightgbm 回归模型参数表

模型	参数名称	参数值
Xgboost	基学习器	gbdt
	基学习器个数	1 000
	学习率	0.1
	树最大深度	4
	子节点权重	100
CatBoost	基学习器	Oblivious Trees
	基学习器个数	1 000
	学习率	0.1
	树最大深度	4
	子节点正则化	3
Lightgbm	基学习器	gbdt
	基学习器个数	1 000
	学习率	0.1
	树最大深度	4
	子节点权重	100

表 6　Xgboost、CatBoost 和 Lightgbm 回归模型评价指标

模型	指标名称	指标值
Xgboost	MAE	0.019
	MAPE	3 945 585 133
	MSE	7×10^{-4}
	RMSE	0.027
CatBoost	MAE	0.026
	MAPE	6 137 840 279
	MSE	0.001
	RMSE	0.033
Lightgbm	MAE	0.026
	MAPE	7 684 653 852
	MSE	0.001
	RMSE	0.034

3.3.2　异常发现

通过对齿轮箱轴承温度分析，其分布符合正态分布的特性。基于这一观察，采用了 3 倍标准差法以识别轴承温度的异常波动，该方法能有效捕捉到大部分极端情况。具体异常发现的步骤如下：1）计算温度数据的平均值 μ 和标准差 σ；2）设定温度异常阈值的上限为 μ+3σ，即平均温度加上 3 倍标准差；3）遍历测试集中的每个数据点，并将其与设定的异常值阈值上限进行比较；4）若某个温度数据点的值超出了该上限，即将其判定为异常值，并触发高温预警系统。

通过上述异常发现方法，发现 A3 号风机在 2023 年 7 月 6 日，齿轮箱轴承出现了持续不断的高温预警。将 2023 年 7 月每天温度平均得到图 7，从图中可以初步明确该风机存在故障，但无法定位故障原因，需要采用 Shapley 值对该风机进行进一步分析。

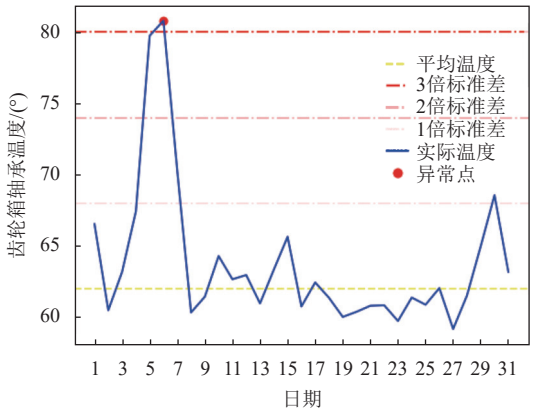

图 7　异常样本示例图

3.3.3　重要特征分析

部件损坏并非一蹴而就，而是经过长时间的累积效应逐渐显现。因此，对于故障定位的分析分为两步策略：非预警时间段的特征分析和预警时间段的特征分析。

在未产生预警时，对 A3 号风机的样本进行整体的特征分析。旨在从日常维护中捕捉可能预示部件损坏的征兆，进而对潜在风险部件实施重点维护，以预防故障的发生。

如图 8 所示，齿轮箱轴承温度的主要影响因素依次为：F9 发电机扭矩、F6 齿轮箱入口油温、F5 齿轮箱转速、F1 叶轮转速。这些因素分别关联于发电机、齿轮箱、叶轮等关键部件。在常规的检查和维护中，F9 发电机扭矩应作为重点检查对象。其中，F6 齿轮箱入口油温的影响呈现出较为分散的特点，即它对轴承的高温区和低温区均产生

了显著的影响。因此，在出现高温预警时，应特别关注 F6 齿轮箱入口油温的监测和排查。这些分析结果为齿轮箱轴承温度的预警和日常维护提供了重要的指导依据。

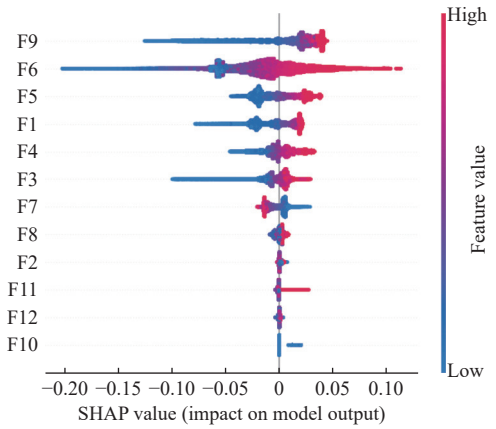

图 8 特征重要性排序图

3.3.4 异常定位

产生预警时，则聚焦于重点特征对预警时间段进行深入分析，此时的特征分析更具有及时性和精准性，从而更准确地理解该时刻高温预警发生的原因，不仅对于即时故障的诊断至关重要，也为事后故障的排查提供了指导依据。针对 A3 号风机在特定时刻（2023 年 7 月 6 日）出现的高温预警现象，分别采用 Shapley 值分析法和 LIME 分析法对该异常时刻进行特征定位。图 9 展示了 Shapley 值分析法对该异常时刻的可解释结果，而图 10 则呈现了 LIME 分析法多次运行后对该时刻异常行为的可解释结果。

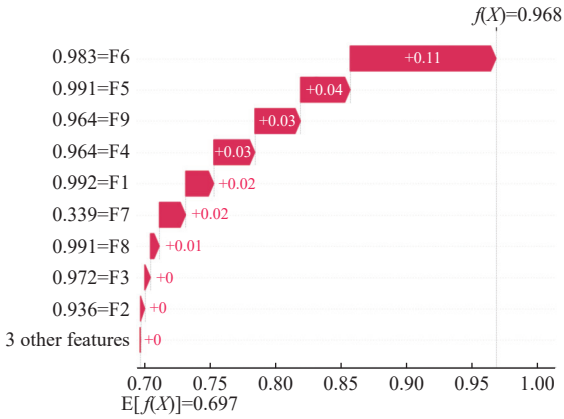

图 9 A3 号风机高温预警点 Shapley 值分析图

Shapley 值和 LIME 分析结果共同显示，A3 号风机在 2023 年 7 月 6 日齿轮箱轴承温度显著上升的主要原因为 F6 齿轮箱入口油温的异常升高。这一异常产生的主要原因是过度负载或润滑不足问题。考虑到风场风机负载基本一致，可以排除过度负载的原因，因此可以定位轴承温度升高是由润滑系统故障导致。以上结论与 A3 号风机实际工况吻合，A3 号风机在事后排查过程中发现润滑油泄漏，导致润滑油位过低无法正常工作，事故现场照片如图 11 所示。

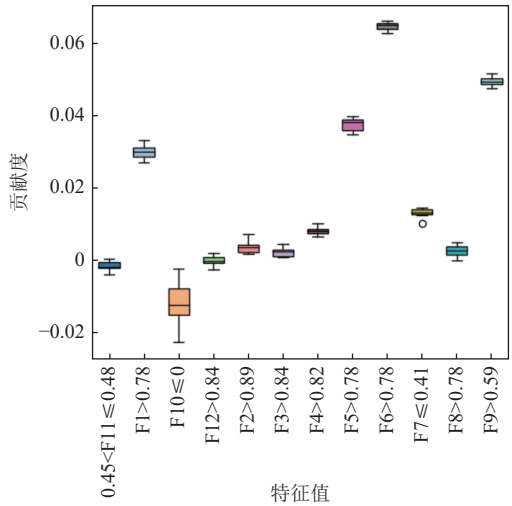

图 10 A3 号风机高温预警点 LIME 箱线图

图 11 A3 号风机润滑油泄露现场图

本案例通过齿轮箱轴承温度异常的 A3 号风机，构建了温度回归模型，再对 A3 号风机在预警时间段与非预警时间段分别进行 Shapley 值分析和 LIME 分析，有效地追溯到 A3 号风机齿轮箱故障的根本源头是润滑系统故障导致齿轮箱入口油温的突发性显著上升，最终导致了风机的停机。在 LIME 分析过程中，由于受到随机扰动的影响，每次的分析结果都会存在细微的差异。如图 10 所示，F10 桨叶角度偏差的贡献值分布较为分散，这使得其对预测结果的可解释稳定性相对较弱，不如 Shapley 值分析结果稳定可靠。该案例展示了 Shapley 值在提供稳定可解释性分析方面的优势，直接且准确地锁定了故障产生的关键要素，从而为后续的维护决

策提供了可靠的理论基础。

4　结束语

无监督聚类的优势在于能够揭示数据内在结构，无须依赖先验标签，但却受限于算法的敏感性和鲁棒性问题。相对地，有监督分类虽然准确度高，却高度依赖高质量大规模标签数据。本文所提出的 Shapley 值聚类方法，通过量化特征贡献度，增强了聚类一致性并揭示了数据深层结构。该方法在 NMI 值指标上表现优于传统无监督聚类，并在一些情况下接近或超越有监督模型。其优势在于融合了特征重要性信息，同时保留了无监督学习的优势，为数据理解和解释提供了新视角。

在样本充分的情况下，Shapley 值在模型可解释方面展现出显著优势，尤其在高维特征空间与有限数据量的复杂场景中。通过识别并量化可解释特征对模型预测的具体贡献，Shapley 值方法能显著透明化模型内部运作机制，进而为特征选择与模型优化策略的制定提供了坚实的基础，显著增强模型的可解释性并提升用户信任度。

本文所提出的可解释人工智能框架，能够解决在不同样本量场景下的模型表现问题，尤其在有限样本标注的情况下提升聚类效果，同时在模型可解释性问题中发挥了关键作用，Shapley 值的引入将领域知识与数据驱动的方法相结合，显著提升复杂模型的可解释性。

本文通过两个案例研究，建立了一种可解释人工智能框架，同时深入探索了 Shapley 值在不同场景中的显著效用，实验结果展示了“知识+数据”结合方式在工程应用中的显著效果。本文提出的框架在齿轮箱故障诊断中实现了三重突破：1）通过 Shapley 值量化指导无监督聚类中的标签生成，减少了对标注数据的依赖；2）利用蒙特卡洛采样提升了高维数据处理的效率；3）通过 Shapley 值定位故障成因，相较于传统方法和黑盒模型，显著提高了诊断的透明度和实用性，为风电行业提供了更可靠的维护策略。但是，如何进一步提升框架扩展性，以覆盖更多类型的数据集与学习任务，需要进一步应用验证。其次，探索 Shapley 值在不同领域和更多样化的数据分布下的适用性和稳健性，也是未来研究的重要方向。

参 考 文 献

[1] Global Wind Energy Council. Global wind report 2024 [EB/OL]. [2024-07-15]. https://gwec.net/global-wind-report-2024/, 16 April 2024.

[2] MENG D B, NIE P, YANG S Y, et al. Reliability analysis of wind turbine gearboxes: Past, progress and future prospects[J]. International Journal of Structural Integrity, 2025, 16(1): 4-38.

[3] 耿珊. 油液监测在风机故障诊断中的应用[J]. 甘肃科技, 2020, 36(11): 50-54.
GENG S. Application of oil monitoring in fan fault diagnosis[J]. Gansu Science and Technology, 2020, 36(11): 50-54.

[4] 付松. 关于风电机齿轮箱传动系统振动特性的分析研究[D]. 乌鲁木齐: 新疆大学, 2011.
FU S. Analysis and research on vibration characteristics of gearbox transmission system of wind turbine[D]. Urumqi: Xinjiang University, 2011.

[5] LIU Z P, WANG X F, ZHANG L. Fault diagnosis of industrial wind turbine blade bearing using acoustic emission analysis[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(9): 6630-6639.

[6] 何群, 李晔阳, 江国乾, 等. 基于条件卷积自编码高斯混合模型的风电齿轮箱健康评估[J]. 太阳能学报, 2023, 44(12): 214-220.
HE Q, LI Y Y, JIANG G Q, et al. Health assessment of wind turbine gearbox based on conditional convolution autoencoding Gaussian mixture model[J]. Acta Energiae Solaris Sinica, 2023, 44(12): 214-220.

[7] LI Q, LI M, FU C, et al. Fault diagnosis of wind turbine component based on an improved dung beetle optimization algorithm to optimize support vector machine[J]. Electronics, 2024, 13(18): 3621.

[8] AMIN A, BIBO A, PANYAM M, et al. A Bayesian deep learning framework for reliable fault diagnosis in wind turbine gearboxes under various operating conditions[J]. Wind Engineering, 2024, 48(2): 297-309.

[9] 刘河生, 徐浩, 李宁, 等. 风电机组齿轮箱故障预警算法研究及应用[J]. 热力发电, 2024(4): 36-42.
LIU H S, XU H, LI N, et al. Research and application of fault early warning algorithm for gearbox of wind turbine[J]. China Industrial Economics, 2024(4): 36-42.

[10] 蔡昌春, 何捷, 承敏钢, 等. 基于改进 VMD-MCKD 和深度残差网络的风机齿轮箱故障诊断[J]. 山东电力技术, 2024, 51(2): 67-78.
CAI C C, HE J, CHENG M G, et al. Fault diagnosis of fan gearbox based on improved VMD-MCKD and deep residual network[J]. Shandong Electric Power, 2024, 51(2): 67-78.

[11] WANG H F, ZHAO X Y, WANG W J. Fault diagnosis and prediction of wind turbine gearbox based on a new hybrid model[J]. Environmental Science and Pollution Research International, 2023, 30(9): 24506-24520.

[12] BIELECKI A, WÓJCIK M. Hybrid AI system based on ART neural network and mixture of Gaussians modules with application to intelligent monitoring of the wind turbine[J]. Applied Soft Computing, 2021, 108: 107400.

[13] 俞国燕, 李少伟, 董晔弘. 基于 XGBoost-LightGBM-LSTM 的风机齿轮箱轴承故障预警[J]. 轴承, 2023(6): 140-145.
YU G Y, LI S W, DONG Y H. Fault warning of wind turbine gearbox bearings based on XGBoost-LightGBM-LSTM[J]. Bearing, 2023(6): 140-145.

[14] COOPER A, DOYLE O, BOURKE A. Supervised clustering for subgroup discovery: An application to COVID-19 symptomatology[C]//Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Cham: Springer International Publishing, 2021: 408-422.

[15] BRUSA E, CIBRARIO L, DELPRETE C, et al. Explainable AI for machine fault diagnosis: Understanding features' contribution in machine learning models for industrial condition monitoring[J]. Applied Sciences, 2023, 13(4): 2038.

[16] BAI Y H, CHENG W D, WEN W G, et al. A time–frequency image quality evaluation method based on improved LIME[J]. Applied Sciences, 2024, 14(7): 2917.

[17] LU F Y, TONG Q B, FENG Z W, et al. Explainable 1DCNN with demodulated frequency features method for fault diagnosis of rolling bearing under time-varying speed conditions[J]. Measurement Science and Technology, 2022, 33(9): 095022.

[18] SUNDARARAJAN M, NAJMI A. The many Shapley values for model explanation[EB/OL]. [2024-07-15]. https://doi.org/10.48550/arXiv.1908.08474.

[19] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[EB/OL]. [2024-07-15]. https://doi.org/10.48550/arXiv.1705.07874.

[20] ŠTRUMBELJ E, KONONENKO I. Explaining prediction models and individual predictions with feature contributions[J]. Knowledge and Information Systems, 2014, 41(3): 647-665.

[21] LUNDBERG S M, ERION G, CHEN H, et al. From local explanations to global understanding with explainable AI for trees[J]. Nature Machine Intelligence, 2020, 2(1): 56-67.

[22] CASTRO J, GÓMEZ D, MOLINA E, et al. Improving polynomial estimation of the Shapley value by stratified random sampling with optimum allocation[J]. Computers & Operations Research, 2017, 82: 180-188.

[23] FRYE C, ROWAT C, FEIGE I. Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability[EB/OL]. [2024-07-15]. https://doi.org/10.48550/arXiv.1910.06358.

[24] ANCONA M, CEOLINI E, ÖZTIRELI C, et al. Explainable AI: Interpreting, explaining and visualizing deep learning[M]. Berlin, Heidelberg: Springer-Verlag, 2019: 169-191.

[25] LIU D, CUI L, CHENG W. A review on deep learning in planetary gearbox health state recognition: Methods, applications, and dataset publication[J]. Measurement Science and Technology, 2023, 35(1): 1-23.

编 辑 张 莉