文章类型：论文|刊号（ISSN）：2630-4651 /（中图刊号）：860GL005

人工智能在电力系统故障诊断中的可解释性研究

蒋高飞
南水北调中线信息科技有限公司
DOI:10.32629/btr.v8i8.4935

[摘　要] 随着人工智能(AI)技术的迅猛发展,其在电力系统故障诊断领域的应用日益广泛。深度学习、强化学习等先进AI模型凭借强大的非线性建模能力和高精度预测性能,在提升故障识别效率与准确性方面展现出巨大潜力。然而,这些“黑箱”模型普遍缺乏可解释性(Explainability),导致其决策过程难以被人类专家理解与信任,严重制约了其在安全关键型电力系统中的实际部署。本文系统梳理了当前主流AI故障诊断方法的技术特点,深入剖析了可解释性缺失对电力系统运行带来的潜在风险,并从模型内在可解释性、事后解释方法以及人机协同机制三个维度,全面探讨了提升AI故障诊断可解释性的关键技术路径。在此基础上,结合典型应用场景,提出了一种融合图神经网络与注意力机制的可解释故障诊断框架。最后,本文对AI可解释性在电力系统中的未来发展方向进行了展望,强调构建“可信AI”体系对于实现智能电网安全、可靠、高效运行的重要意义。
[关键词] 人工智能；电力系统；故障诊断；可解释性；图神经网络；注意力机制；可信AI
中图分类号：TM76　文献标识码：A

Research on Explainability of Artificial Intelligence in Power System Fault Diagnosis

Gaofei Jiang
South-to-North Water Diversion Middle Route Information Technology Co., Ltd.

[Abstract] With the rapid development of artificial intelligence (AI) technology, its application in the field of power system fault diagnosis has become increasingly widespread. Advanced AI models such as deep learning and reinforcement learning, leveraging their powerful nonlinear modeling capabilities and high-precision prediction performance, have shown great potential in improving the efficiency and accuracy of fault identification. However, these "black box" models generally lack explainability, making their decision-making processes difficult for human experts to understand and trust, which severely restricts their practical deployment in safety-critical power systems. This paper systematically reviews the technical characteristics of current mainstream AI fault diagnosis methods, deeply analyzes the potential risks posed by the lack of explainability to power system operation, and comprehensively explores key technical paths for improving the explainability of AI fault diagnosis from three dimensions: intrinsic model explainability, post-hoc explanation methods, and human-machine collaboration mechanisms. On this basis, combined with typical application scenarios, an explainable fault diagnosis framework integrating graph neural networks and attention mechanisms is proposed. Finally, this paper discusses the future development directions of AI explainability in power systems, emphasizing the importance of building a "trustworthy AI" system for achieving safe, reliable, and efficient operation of smart grids.
[Key words] artificial intelligence; power system; fault diagnosis; explainability; graph neural network; attention mechanism; trustworthy AI

引言

电力系统关乎国民经济与社会民生,其安全稳定运行至关重要。当下,新能源大规模并网、负荷结构复杂化、电网规模扩大,电力系统呈现高度非线性等特征,传统故障诊断方法在海量数据与复杂故障前,暴露出响应慢、泛化弱、适应性差等局限。近年来,以多种神经网络为代表的人工智能技术在电力系统故障诊断领域进展显著,能自动学习复杂映射关系,快速准确判断故障,提升电网智能化水平。但多数高性能AI模型是“黑箱”系

文章类型：论文|刊号（ISSN）：2630-4651 /（中图刊号）：860GL005

统,决策逻辑抽象难追溯,在电力系统高要求场景下,调度员等难以理解诊断结论,无法建立信任。AI误判或引发连锁故障,造成重大损失。故而,在保持AI模型高性能的同时提升其可解释性,是智能电网研究的前沿热点与核心难题,可解释人工智能是连接AI技术与工程实践的关键桥梁。

1 人工智能在电力系统故障诊断中的应用现状

1.1主流AI故障诊断方法

目前,应用于电力系统故障诊断的AI方法主要包括：(1)基于深度学习的方法：利用DNN、CNN等模型处理电压、电流等时序信号。例如,将故障录波数据作为输入,通过CNN提取局部特征,实现故障类型的分类。(2)基于循环神经网络的方法：RNN及其变体(如LSTM、GRU)擅长处理长序列依赖关系,适用于分析故障暂态过程中的动态特征。(3)基于图神经网络的方法：电力网络天然具有图结构(节点为母线/发电机,边为输电线路)。GNN能够直接在拓扑图上进行信息传播与聚合,有效捕捉元件间的电气耦合关系,对定位故障区域具有独特优势[1]。(4)基于集成学习与迁移学习的方法:通过集成多个基模型提升鲁棒性,或利用源域知识辅助目标域(如不同区域电网)的故障诊断,解决数据稀缺问题。

1.2可解释性缺失带来的问题

尽管性能优越,但上述AI模型普遍存在以下可解释性问题：(1)决策依据不透明：模型输出“线路L12发生单相接地故障”,但无法说明是依据哪些传感器数据、哪些时间点的特征做出的判断。(2)错误归因困难：当模型误判时,运维人员难以定位是数据质量问题、模型结构缺陷还是训练不足所致。(3)缺乏物理一致性:AI模型可能学习到与物理规律相悖的虚假相关性(如将天气与特定故障错误关联),导致在新场景下失效。(4)合规与审计障碍:在电力行业监管严格的背景下,无法解释的AI决策难以通过安全审查,也难以满足事故回溯与责任认定的要求。这些问题使得AI诊断系统往往停留在“辅助参考”层面,难以获得调度员的完全信任,限制了其在闭环控制等关键环节的应用。

2 可解释人工智能(XAI)的核心内涵与技术路径

2.1可解释性的定义与维度

在电力系统语境下,AI可解释性应包含以下维度:一是透明性(Transparency):模型结构本身易于理解(如决策树),或其内部状态可被可视化。二是因果性(Causality)：能揭示输入特征与输出决策之间的因果或强相关关系。三是一致性(Consistency)：解释结果应与电力系统物理规律、运行经验相符。四是交互性(Interactivity)：支持用户通过提问、反事实推理等方式与模型进行对话式解释。

2.2提升可解释性的主要技术路径

2.2.1内在可解释模型

内在可解释模型通过在算法设计阶段嵌入可理解的结构,从根本上避免“黑箱”问题。例如,广义加性模型(GAMs)将整体预测分解为若干单变量函数的加和,每个函数对应一个输入特征,使得用户可以清晰地看到每个量测值对最终诊断结果的独立贡献。虽然这类模型在表达复杂非线性关系时可能不如深度神经网络强大,但其结构的简洁性使其在对安全性要求极高的子系统(如主保护后备逻辑)中具有独特优势[2]。在图神经网络领域,研究者开始探索可解释的GNN变体。典型做法是在消息传递机制中引入注意力权重,使模型在聚合邻居信息时能够动态分配不同的重要性。这种注意力机制不仅提升了模型性能,其权重本身即构成了一种天然的解释:高注意力值的边或节点,正是模型在诊断过程中重点关注的对象。通过可视化这些权重,调度员可以直观地看到AI“目光”所及之处,从而建立起对模型判断的信任。

2.2.2事后解释方法

对于已经训练完成的高性能黑箱模型,事后解释方法提供了一种灵活的补救策略。其中,SHAP(SHapleyAdditiveexPlanations)方法基于博弈论中的Shapley值,为每个输入特征分配一个公平的贡献度评分。在电力系统中,这意味着可以量化SCADA系统中每一个遥测点对故障诊断结果的影响大小,生成按重要性排序的特征列表。LIME(LocalInterpretableModel-agnosticExplanations)则通过在局部邻域内拟合一个简单的可解释模型(如线性回归),来近似黑箱模型的行为,适用于解释单个故障案例的决策逻辑。梯度类方法如Grad-CAM,则利用输出对输入的梯度信息生成热力图,特别适合处理图像化表示的电网数据(如将拓扑图栅格化)。而在电力领域更具前景的是反事实解释,它通过构造“如果……那么……”的假设场景,帮助用户理解模型的敏感边界。例如,系统可以回答:“若故障线路对端的电流未超过阈值,则不会触发跳闸指令。”这种解释方式贴近工程师的思维习惯,有助于快速评估模型在异常工况下的可靠性。

2.2.3基于知识融合的混合建模

将电力系统深厚的领域知识融入AI模型,是提升可解释性与鲁棒性的根本途径。物理信息神经网络(PINNs)在损失函数中显式加入由微分方程描述的物理约束(如潮流方程),迫使神经网络的输出不仅拟合数据,还要满足基本物理规律。这种方法有效抑制了模型学习虚假相关性的倾向,使其在训练数据覆盖不足的区域仍能保持合理推断。另一种思路是构建符号-神经混合系统。该系统将感知层(由神经网络处理原始量测数据)与认知层(由符号推理引擎执行基于规则的逻辑判断)分离[3]。神经网络负责从噪声数据中提取高层语义特征(如“检测到零序电流突增”),而符号引擎则根据预设的保护逻辑(如“零序电流>定值且持续时间>时限→判定为接地故障”)进行最终决策。这种分层架构不仅使整个推理链条清晰可溯,还便于专家对规则库进行维护与更新,实现了AI智能与人类智慧的有机融合。

3 面向电力系统故障诊断的可解释 AI 框架设计

为兼顾性能与可解释性,本文提出一种融合图神经网络与注意力机制的可解释故障诊断框架(ExplainableGNN-basedFaultDiagnosis,XGFD)。

3.1框架架构

XGFD框架包含三个核心模块：

(1)图构建模块：将电网拓扑建模为有向图$\mathcal{G}=(\mathcal{V},\mathcal{E})$,其中节点$v_i \in V$表示母线或发电机,边 $e_{ij} \in \mathcal{E}$ 表示输电线路。节点特征包括电压幅值、相角、注入功率等；边特征包括线路阻抗、潮流方向等。

(2)可解释GNN编码器：采用带注意力机制的图卷积网络(GAT)。在消息传递过程中,计算节点i对邻居j的注意力系数α_{ij}:

$$\alpha_{ij}=\frac{\exp(\mathrm{LeakyReLU}(a^T[Wh_i\|Wh_j]))}{\sum_{k\in\mathcal{N}(i)}\exp(\mathrm{LeakyReLU}(a^T[Wh_i\|Wh_k]))}$$

其中hi为节点嵌入,W为可学习权重,a为注意力向量。注意力权重α_{ij}直观反映了在诊断过程中,节点j对节点i的影响程度,构成天然的解释依据。

(3)多任务解码器:同时输出故障类型(分类任务)与故障位置(节点/边回归或分类任务)。引入门控机制,使不同类型故障激活不同的子网络,提升决策逻辑的清晰度。

3.2可解释性实现机制

一是拓扑级解释:通过可视化注意力权重矩阵,展示在诊断某次故障时,哪些线路或母线被模型重点关注。例如,若某条线路的两端节点间注意力值显著高于其他边,则可推断该线路为可疑故障点。二是特征级解释：结合SHAP方法,对每个节点的输入特征(如电压、电流)计算贡献度,生成“特征重要性热力图”。三是反事实推理接口：允许用户修改部分量测值(如假设某PMU数据丢失),观察诊断结果的变化,评估模型的鲁棒性与敏感性。

4 讨论：挑战与未来方向

尽管本文提出的框架在可解释性方面取得了一定成效,但在迈向工程实用化的道路上仍面临诸多挑战。首要问题是实时性与解释复杂度的矛盾。电力系统故障诊断通常要求在毫秒级内完成,而诸如SHAP等高精度解释算法计算开销较大,难以满足在线应用需求。未来需研究轻量化的、与模型前向传播同步生成的解释机制。其次,现代电网数据来源日益多元,包括高速PMU、传统SCADA、行波测距装置乃至运维日志文本。如何构建统一的多模态融合解释框架,使AI能够综合各类异构信息并给出连贯一致的解释,是一个亟待解决的难题。再者,可解释性的最终服务对象是人,因此必须考虑人因工程因素。解释的形式(如拓扑图、数值表、自然语言)应匹配调度员的认知习惯与工作流程,这需要通过大量的人机交互实验来优化[4]。此外,详细的解释信息可能无意中泄露电网的敏感拓扑结构或运行状态,带来潜在的安全风险。如何在保障可解释性的同时,通过差分隐私、联邦学习等技术保护数据隐私,是另一个重要的研究方向。展望未来,发展在线可解释学习、构建电力专用XAI基准、探索大语言模型在自动生成故障分析报告中的应用,以及推动可解释性标准纳入行业规范,将是推动可信AI在电力系统中深度落地的关键路径。

5 结语

人工智能为电力系统故障诊断带来了革命性机遇,但其“黑箱”特性严重阻碍了在安全关键场景中的落地应用。本文系统论证了可解释性在提升AI诊断系统可信度、安全性与实用性方面的核心价值,综述了内在可解释模型、事后解释方法与知识融合等关键技术路径,并提出了一种基于注意力图神经网络的可解释故障诊断框架。未来,构建“性能-可解释性-鲁棒性”三位一体的可信AI系统,将成为智能电网发展的必然趋势。唯有让AI不仅“做得对”,而且“说得清”,才能真正赋能电力系统向更安全、更智能、更韧性的方向演进。

[参考文献]

[1]常欣远,王雯祺.基于人工智能的电力系统故障诊断与预测研究[N].市场信息报,2025-04-23(014).

[2]赵海萍,刘晓琴,邱昱.人工智能技术在电力系统故障诊断中应用[C]//中国电机工程学会电力信息化专业委员会,国家电网公司信息通信分公司.2022电力行业信息化年会论文集.国网甘肃省电力公司数字化事业部,2023:473-476.

[3]杨湛鸿.人工智能在电力系统故障诊断中的应用[C]//中国电力设备管理协会.中国电力设备管理协会第二届第一次会员代表大会论文集(2).广东电网有限责任公司湛江供电局,2022:49-53.

[4]方萌,史可敬.对于人工智能在电力系统故障诊断中的应用及研究[J].中国科技投资,2021,(22):63-64.