第48卷 第4期 2024年2月25日
DOI：10.7500/AEPS20230509007
电力系统自动化
Automation of Electric Power Systems
Vol. 48 No. 4 Feb. 25，2024

可解释人工智能在电力系统中的应用综述与展望

王小君[1]，窦嘉铭[1]，刘　墨[1]，刘畅宇[1]，蒲天骄[2]，和敬涵[1]
（1. 北京交通大学电气工程学院，北京市 100044；2. 中国电力科学研究院有限公司，北京市 100192）

摘要： 可解释人工智能(XAI)作为新型人工智能(AI)技术，具有呈现AI过程逻辑、揭示AI黑箱知识、提高AI结果可信程度的能力。XAI与电力系统的深度耦合将加速AI技术在电力系统的落地应用，在人机交互的过程中为电力系统的安全、稳定提供助力。文中梳理了电力系统XAI的历史脉络、发展需求及热点技术，总结了XAI在源荷预测、运行控制、故障诊断、电力市场等方面的电力应用，并围绕解释含义、迭代框架、数模融合等方面展望了电力系统XAI的应用前景，可为推动电力系统智能化转型与人机交互迭代提供理论参考与实践思路。

关键词： 电力系统；人工智能；可解释性；机器学习

0　引言

近期，以Chat Generative Pre-trained Transformer（ChatGPT）[1]为首的大型语言模型因其能够模拟人类语言行为而震惊世界，引发人类社会关于人工智能（artificial intelligence，AI）应用的大讨论。电力系统作为人类社会重要构成，半个世纪以来，人工智能的发展同样为其提供了崭新道路：联结主义、知识主义、行为主义和群体智能等常用方法[2]不仅能够及时处理电力系统运行控制的大量随机高维动态数据，还能结合机理知识推动电力系统物理信息耦合，实现能源利用效率进一步提高。“AI+”已成为电力系统实现能源转型与“双碳”目标的关键举措与重要推手。

然而，不同形式可再生能源高渗透率的接入与高比例电力电子设备的广泛应用促使电力系统复杂性进一步加剧：对内，源荷随机性显著与不确定性增强，运行工况样式繁多，潮流问题复杂多变；对外，密切联系政策法规、社会价值取向，在与社会网络的交互过程中战略性与基础性更加突出[3]，对“经济-安全-环境”三角平衡提出更高要求。为保障具有高维随机特性并兼有数据知识耦合、内外利益纠缠的复杂巨系统安全可信，以深度神经网络（deep neural network，DNN）为首的人工智能黑箱模型越来越难以为继，表现出深度学习模型不透明、网络参数意义不确切、学习样本不均衡且数据不完备等问题，阻碍了人工智能在电力系统中的推广应用。

可解释人工智能（explainable artificial intelligence，XAI）进入电力系统为解决上述问题提供了重要思路。XAI的目的是使人类更易于理解人工智能的行为[4]，其对于电力系统的作用和意义主要表现在：1）在运行层面，XAI可使电力专家充分了解、验证人工智能黑箱模型的控制与操作过程，结合人工经验协助优化调整运行策略，有助于电力系统更安全稳定运行、更高效能源利用与更少量环境污染；2）在管理层面，XAI可用于识别电力系统外部风险因素，分析利益相关方的策略选择与影响程度，有助于协助人类阐明电力系统演化机理并制定优化引导策略；3）在知识层面，XAI能够挖掘和传递隐含在复杂电力系统数据中的高维知识和洞见，揭示人类领域知识未涉猎的新规律、新机制，有助于完善电力系统知识体系。同时，国家自然科学基金委员会近期也发布指南[5]，要求发展可解释、可通用的下一代人工智能方法，推动人工智能在科学领域的创新应用，以支撑中国在新一轮国际科技竞争中的主导地位。因此，适时开展电力系统XAI研究，探索黑箱人工智能模型结构与机理挖掘，对推动人工智能方法在电力科学领域的创新应用、构建新一代清洁低碳安全高效的智能电网、提高中国在能源与人工智能领域的国际竞争地位具有重要的现实意义与理论价值。

针对XAI的研究，国内外学者已提出诸多理论与实践方法。部分文章对现有XAI技术进行总结：文献[6]从社会科学角度对XAI进行了论述；文献

收稿日期：2023-05-09；修回日期：2023-11-07。
上网日期：2024-01-03。
国家自然科学基金资助项目（52377071）；国家自然科学基金青年基金资助项目（52107068）。

[7-8]对机器学习可解释性方法进行了全面的分析整理；文献[9]关注了医疗领域的XAI，文献[10]则聚焦于XAI在6G通信的应用。总体而言，尽管XAI的定义范围与评估方式未有普遍认同的标准与体系，但各领域学者正努力找寻XAI与本领域业务知识的关键结合点，以使领域内的智能水平与可信程度并驾齐驱。

在能源领域，已有一些初步关于XAI的综述工作[11-15]。文献[11]从XAI框架入手，探索了电力系统中机器学习可解释性方法的基本概念、研究框架与关键技术，其主要贡献在于形成了电力XAI的整体研究思路与框架。文献[12]基于文献计量学方法，对智慧能源系统中XAI与治理方法选择多个关键词进行分析，但其对关键词的选择和分析较为宽泛，未能深入挖掘具体应用场景的细节，导致分析结果有限，难以为实际工程应用提供有效的指导和建议。文献[14]对能源领域的部分XAI应用进行了简要介绍，包括电网应用、能源行业、建筑能源管理等，但由于其针对多个能源领域开展综述，未能对电力系统中的运行规划、优化调度、电力市场等特定应用进行专门的详细探讨与分析。

总的来看，现有研究对电力系统XAI应用划分与局限讨论较为宽泛笼统，场景的分类仍需统筹规划。合适的应用场景是XAI成功落地的关键，利用因地制宜的思想思考电力系统XAI应用可以从问题本身特点出发，得到适配问题本身的应用方式，从而将XAI优势最大化凸显。因此，为使XAI在新型电力系统中发挥更大价值，本文特别关注于分析XAI在电力系统的应用方式与本质特征异同，以期为电力能源结构转型与“双碳”目标实现提供助力。本文综合梳理了电力系统中XAI的5种应用，尤其补充优化调度、运行规划、电力市场等方面的XAI应用，并深入探讨每类应用的局限性和可能的创新方法；展望了电力系统XAI的研究方向，特别讨论了解释范围界定困境、解释能力与准确性矛盾、数据模型主导平衡、数据机理融合范式挖掘等共性难题，以期为相关研究提供有益参考。

1　电力系统XAI发展与现状

XAI是人工智能发展早期就存在的概念。20世纪八九十年代，电力系统专家系统中已引入“解释”概念，文献[16]中指出专家系统解释部分的主要功能是解释系统本身的推理结果，使用户接受。2004年，文献[17]首先运用英文explainable artificial intelligence表示具有可解释能力的人工智能，但此概念仅限于基于全谱命令的人工智能系统在军事仿真游戏中的解释能力。

XAI概念自提出以来，逐渐进入快速发展通道。电力系统中，文献[18]综述了系统中故障诊断的各种研究方法，认为基于人工神经网络的电力系统故障诊断方法缺乏解释自身行为和输出结果的能力，限制了其在大型电力系统中的应用。2006年，“深度学习”的概念[19]被提出，人工智能研究进入了新的阶段。然而，深度学习虽然加深了网络深度，提升了模型效果，但却大大减少了人们对模型内部的理解。面对该问题，2017年5月，美国国防部高级研究计划局启动了XAI计划[20]，从解释模型学习、解释界面设计与解释心理研究三方面入手，较为完整和明确地阐述了关于人工智能可解释性的概念，即一整套能够产生更多可解释模型，维持高水平学习性能，使用户理解、信任和有效管理人工智能的机器学习技术，正式拉开了XAI快速发展的帷幕[21]。

近年来，为促使人工智能应用快速落地，中国出台了各类XAI相关政策。2017年，国务院公开印发的《新一代人工智能发展规划》中强调：“实现具备高可解释性、强泛化能力的人工智能”[22]。2021年，新一代人工智能治理专业委员会发布了《新一代人工智能伦理规范》[23]，指出：“在算法设计、实现、应用等环节，提升透明性、可解释性、可理解性、可靠性、可控性”。在产业界，阿里[24]、京东[25]等公司都针对XAI展开前沿部署。值得注意的是，2022年12月，国网山东省电力公司联合阿里巴巴公司共同实现了基于XAI的负荷预测，有力地支撑电网安全稳定运行[26]。

在学术界，XAI与电力系统的联系日益紧密，已有一些相关综述与研究论文发表。已发表的电力系统XAI综述论文如表1所示，其分别从电力系统可解释框架、能源领域XAI文献计量、建筑能源XAI应用等方面开展综述，对XAI在电力系统落地应用具有较强的借鉴价值与意义。同时，本文基于知识图谱软件——visualization of similarities viewer（VOSviewer）[27]分析了电力系统XAI论文（基于中国知网），通过聚类分析了解XAI在电力系统的关键领域，如图1所示。在技术层面，XAI与机器学习、深度学习、决策树、注意力机制、随机森林、卷积神经网络（convolutional neural network，CNN）等人工智能技术关系密切；在应用层面，XAI与故障诊断、暂态稳定评估、电力负荷预测、风电功率预测等电力系统关系密切。

随着国家扶持与行业推动，XAI在电力系统的交互已向纵深发展，成为电力系统智能化升级的重要趋势。

表1　近期发表的电力/能源系统相关XAI综述总结
Table 1　Summary of XAI reviews in power/energy system published in recent years

文献	主要特点	主要内容
[11]	尝试解决电力系统智能分析中机器学习可解释性尚未形成整体研究思路与框架的问题	1)给出了机器学习可解释性基本概念、数学描述、主要理论与性能评价；2)构建了面向电力系统智能分析的机器学习可解释性基本框架；3)给出了电力系统智能分析的机器学习可解释性方法总结与展望
[12]	从文本挖掘的角度入手，分析能源人工智能的可解释性和可治理性，侧重于管理	1)将能源人工智能解释性和管理性分为四大类内容，即人工智能行为与治理、技术、设计与开发、运营，共发现了15个主题词；2)从文献计量学的角度给出了展望
[13]	从可再生能源和资源的解决方案、应用和挑战对XAI研究进行综述，侧重于可再生能源	1)介绍了XAI的研究背景；2)分析了人工智能与可再生能源相结合的挑战，例如缺乏理论知识、实践知识、基础架构过时、财政压力；3)给出了人工智能和XAI在未来可再生能源中的作用
[14]	对能源领域XAI进行了综述，涉及多个能源部门，内容较广	1)分析了在能源和电力系统领域XAI技术的主要挑战；2)介绍了XAI在能源领域的应用(稳定性评估、建筑能源等)；3)给出了与XAI和能源系统相关的潜在应用和未来研究方向
[15]	可解释机器学习在建筑能源管理中应用的全面综述，侧重于建筑侧能源的利用	1)综述了建筑能源管理中的可解释机器学习技术；2)给出了可解释性机器学习在建筑能源管理中的现状；3)讨论了可解释机器学习在建筑能源管理的挑战

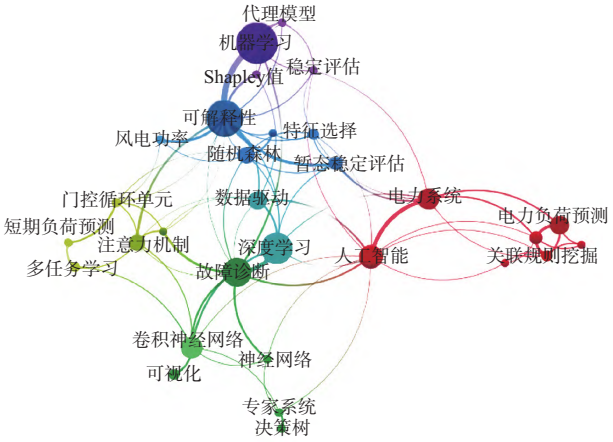

图1　电力系统XAI文献关键词聚类结果
Fig. 1　Clustering results of XAI literature keywords related to power system

2　经典XAI方法

考虑到在电力系统中引入XAI具有重要意义，本章将介绍几种经典的XAI技术，并阐述其在电力系统中的常见应用。经典XAI可根据解释方式分为两类，即具有自解释能力的人工智能模型与人工智能模型解释方法。前者指本身具备一定解释能力的模型，包括线性模型、决策树、随机森林、注意力机制等；后者则指一类用于解释人工智能模型的方法，该类方法往往与模型关系不大，常见的如沙普利值加性解释方法(Shapley additive explanations，SHAP)[28]、模型无关的局部可解释性方法(local interpretable model-agnostic explanations，LIME)[29]和梯度加权类激活映射方法(gradient-weighted class activation mapping，Grad-CAM)[30]等等。各类方法的对比如表2所示。

2.1　自解释人工智能模型

2.1.1　线性模型

线性模型是机器学习中最简单且具有自解释性的人工智能模型，其通过建立输入特征与输出特征之间的线性关系来进行预测。线性模型的预测结果通常可以建模为：

$$y=w_1x_1+w_2x_2+\cdots+w_nx_n+\varepsilon \tag{1}$$

式中：y为预测结果；x_i为输入特征；w_i为模型需要学习的特征权重；ε为预测结果与真实值的误差。线性模型的主要目标是找到一组最佳参数w_i，使得预测输出与实际输出之间的误差最小，常用方法包括最小二乘法和梯度下降法等。

线性模型的解释性体现在其可直观解释每个输入特征对输出结果影响的重要性，由于只涉及加法和乘法运算，线性模型的计算效率较高，适用于数据维度较高的实际问题。然而，线性模型假设输入与输出之间存在线性关系，故在处理非线性问题时，其拟合能力有限，且存在对异常值较为敏感的问题。

2.1.2　决策树和随机森林

决策树[31]是一种具有可解释性的学习方法，通过计算叶子节点的熵以寻找最优的分割特征，从而在每个节点上进行分裂。决策树可解释性体现在其能把数据转化为可解释的形式(如图2所示)，每个节点指代一种可判断的决策，沿着树的分支、经过一系列决策即可得到结论。基于决策树构成的随机森林[32]也同样具有一定程度的可解释性。作为一种集成学习算法，该方法通过构建多个决策树聚合所有树的预测结果，决定一个共同的输出结果。

表2　经典XAI方法对比
Table 2　Comparison of classical XAI methods

类型	方法	对应章节	优点	缺点	适用范围	适用数据类型
自解释人工智能模型	线性模型	2.1.1节	1)简单易理解;2)计算效率高;3)可以处理高维数据	1)对非线性问题拟合能力有限;2)对异常值敏感	适用于回归和分类问题,特别是线性可分的问题	表格型数据
	决策树	2.1.2节	1)易于理解和实现;2)能够处理分类和回归问题;3)能够处理多输出问题	1)容易过拟合;2)对连续型数据处理能力较弱;3)对异常值敏感	适用于分类和回归问题,能够处理离散和连续属性的数据	表格型数据
	注意力机制	2.1.3节	1)提高模型对输入的关注度;2)改善长序列处理能力;3)提高模型解释性	1)计算复杂度较高;2)不易处理大规模数据	适用于序列到序列的任务	序列数据(文本、时间序列等)
人工智能模型解释方法	SHAP	2.2.1节	1)适用于多种模型;2)理论基础扎实;3)可以处理高维数据	1)计算复杂度较高;2)需要大量样本数据	适用于多种机器学习模型,如神经网络等	表格型数据、图像型数据、序列数据等
	LIME	2.2.2节	1)适用于多种模型;2)可解释局部预测结果;3)生成可视化解释	1)对全局解释能力有限;2)线性近似可能不准确;3)计算复杂度较高	适用于多种机器学习模型,如线性模型、决策树、神经网络等	表格型数据、图像型数据、序列数据等
	Grad-CAM	2.2.3节	1)适用于多种CNN;2)可生成可视化热力图;3)不需要修改原有模型	1)仅适用于CNN;2)依赖于梯度信息,可能受到梯度消失或爆炸影响	适用于CNN,特别是用于电力系统中图像分类、目标检测等任务	图像型数据

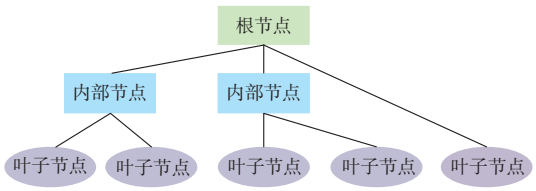

图2　决策树模型结构
Fig. 2　Structure of decision tree model

然而,决策树与随机森林存在着共同的缺陷,即当决策树层数较大或具有多棵决策树的复杂推理过程时,根据先验知识对推理过程进行解释也会变得非常困难,可解释性会受到一定限制。但层数较大的决策树与随机森林也有可用的解释分析方法。例如:决策树拥有weight、gain、cover等特征重要性指标[33],其中,weight表示在所有树中某一特征用于分裂节点的次数,gain表示在所有树中某一特征用于分裂节点的平均增益,cover则表示在所有树中某一特征用于分裂节点的平均覆盖率;另外,SHAP也可对决策树与随机森林模型进行局部解释,揭示每个样本点的预测是如何根据所有特征的联合贡献得出的。

在电力系统中,决策树和随机森林的解释性体现在其可了解不同特征如何在决策过程中影响人工智能模型的结果。例如,以历史数据为基础,将电力系统的状态和控制变量分解为若干特征(如系统负荷、风速、气象等),通过学习不同特征组合的控制策略,基于决策树开展调度优化。在该过程中,决策树可以描述系统状态与控制决策之间的关系,并直观地反映出不同特征组合对应的控制策略,具有一定解释效果。文献[34]针对变电站运维专家系统扩充困难与专家依赖的问题,提出基于决策树的专家规则提取方法。所提出的规则能够满足高解释度、高置信度、路径清晰的基本要求。

2.1.3　注意力机制

深度学习中的注意力机制[35]是一种仿效人类注意力机制的计算模型,最初是为了解决序列到序列模型在处理长距离依赖问题的困难而提出的。其能够在处理输入数据时,动态地将注意力集中在重要的特征上,从而在保持高精度的同时,减少不必要的计算[36]。

注意力机制模型的解释性主要体现在其能够清晰显示模型关注的重点(如图3所示[37]),更加灵活地处理输入数据,并且根据不同的任务和数据动态地调整注意力权重,从而提高了模型的性能和可解释性。

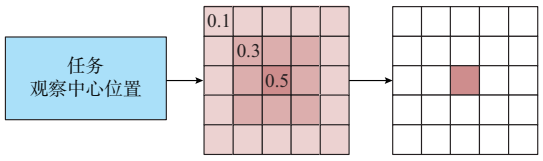

图3　注意力机制处理过程
Fig. 3　Processing process of attention mechanism

在电力领域,注意力机制及其构成的模型已被用来解释电力系统暂态稳定评估[38]、设备故障诊断[39]和可再生能源出力预测[40]中的深度学习模型,用于识别时间序列数据中对模型决策有最大贡献的特征。另外,注意力机制可以与基于强化学习

（reinforcement learning，RL）的调度方法相结合，例如利用注意力机制的智能体嵌入式向量表示方法揭示调度模型的内部工作原理。在这种情况下，注意力机制可帮助调度员理解模型在做出调度决策时所关注的关键因素，从而提高模型的可解释性与可靠性。

2.2　人工智能模型解释方法

人工智能模型解释方法主要是指能够解释现有人工智能黑箱模型的方法。一般来说，电力系统中常见的人工智能黑箱模型包括DNN、深度强化学习（deep reinforcement learning，DRL）和高斯过程等。

1）DNN的不可解释性主要反映在其复杂网络结构与多层次网络参数间的相互作用，使得最终的预测结果难以直观地解释。此外，对于同一输入，不同神经元的激活方式可能存在多种组合，亦增加了人类对于模型内部原理的理解难度。

2）DRL的不可解释性主要反映在其学习过程中，通过与环境的交互来优化策略。这种交互过程中的状态转换与奖励变化通常难以直观地解释。同时，DRL策略网络与价值网络的参数更新过程复杂且相互依赖，使得模型的决策过程难以解释。

3）高斯过程的不可解释性主要反映在其基于概率分布的推理过程。高斯过程通过对输入数据的协方差矩阵进行计算，从而得到一个预测分布。然而，这种基于概率的推理过程往往在处理高维数据和复杂关系时难以直观地解释。

针对人工智能黑箱有一些经典的解释方法，包括SHAP、LIME、Grad-CAM等等，这些方法通过不同的技术和原理，揭示了模型工作机制，提高人工智能黑箱模型的可解释性。下面分别进行分析。

1）SHAP

SHAP是一种用于解释机器学习模型的非参数方法，适用于预测问题。SHAP发展历史可追溯到上世纪50年代，经济学家L.Shapley提出了“Shapley值”[41]的概念，用于衡量每个人在集体中的贡献度。2017年，文献[28]提出解释预测的统一框架SHAP，其可帮助解释机器学习模型之外的因素如何影响模型的输出，以及不同属性对预测结果的贡献大小。

在电力系统中，SHAP可应用在大部分监督学习领域，表现出了显著的优势。在源荷预测应用中，文献[42]提出一种结合长短期记忆（long short-term memory，LSTM）递归神经网络与多任务学习的综合能源预测方法，并利用SHAP解释技术进行解释，提高了能源预测的准确性和可解释性。在暂态稳定研究中，文献[43]提出基于SHAP理论的暂态电压稳定评估归因分析框架，量化了暂态电压稳定特征对稳定评估结果的影响。

2）LIME

LIME是由Ribeiro等人于2016年提出的一种模型无关的可解释方法[29]。其基本原理是针对每个待解释样本，在其附近选取一定数量的邻近数据点，并为这些邻近数据点赋予与待解释样本的距离成反比的权重，形成了一个新的、局部有限的加权数据集。随后，LIME基于该数据集训练一个简单、可解释的模型（如线性回归模型），尽可能逼近原模型在该局部范围内的预测结果。

LIME的目标是解释原模型在目标样本上的预测行为，而不是重新拟合整个训练集。它通过在局部范围内拟合一个简单的代理模型，揭示了预测结果是如何受到各特征值和模型参数的影响。这样，LIME可为每个样本提供针对性的解释，有助于理解原模型在不同样本上的预测差异。

在电力系统中，LIME模型已被广泛应用到稳定性评估[44]、源荷预测[45]等领域。通过探究原模型在各个样本上的局部可解释性，LIME有助于提高电力系统预测与评估的准确性和可信性。

3）Grad-CAM

Grad-CAM是一种解释CNN预测过程的可视化方法，适用于任何CNN模型且无须改变网络结构。其通过生成热力图来揭示网络在识别图像时所关注的区域，以帮助人类更好地理解模型的决策依据。该方法的核心在于计算预测类别相对于某个中间层的梯度，以表示该层特征图的重要性。然后，将特征图权重与梯度相乘，得到类激活映射（class activation mapping，CAM）。最后，将CAM上采样至原始图像大小，生成直观的热力图。从输入输出关系来说，Grad-CAM显示的是输入特征对输出标签概率的重要性。

在电力系统中，Grad-CAM方法已在多个图像处理领域得到研究，例如电力线路安全检测[46]、螺栓多属性分类[47]、绝缘子故障检测[48]等。通过使用Grad-CAM生成的热图，研究人员和工程师可以直观地了解输入图像的哪些区域对模型的预测结果产生了重要贡献，从而提高模型预测的可解释性，辅助决策制定和诊断过程。

3　XAI在电力系统中的应用

为促使XAI在电力系统的进一步发展与应用，本章回顾部分文献并讨论经典研究，将XAI在电力系统中的应用划分为源荷预测、系统规划、运行控制、故障诊断、电力市场等5类应用（如图4所示），并

在每类应用最后讨论不足，以期为相关领域研究提供思路。

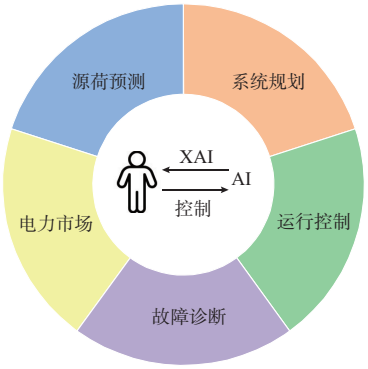

图4　电力系统XAI应用范围

Fig. 4　Scope of XAI applications in power system

3.1　XAI在源荷预测中的应用

XAI可帮助人类理解源荷预测模型的工作原理并提高预测结果的可信度，经由人机交互后可提高未来源荷预测的准确性。XAI在源荷预测中的应用可以划分为新能源出力预测和负荷预测。

3.1.1　新能源出力预测

准确的新能源出力预测可帮助相关部门提前做好规划和调度工作，合理统筹部署发输电设备，提升消纳能力。XAI通过分析出力预测模型内在机制，发现非预期因素，针对人工智能黑箱模型存在的问题进行校正和改进，提升预测结果的解释性能。本节综述了XAI在新能源出力预测中的两类应用：风电出力预测与光伏出力预测，如表3所示。

表3　XAI在新能源出力预测中的相关文献

Table 3　Related literatures on application of XAI in power output prediction of renewable energy

能源	文献	方向	特征	预测	解释	测试系统
风能相关	[40]	风电超短期功率预测	气象条件、运行状态、空气动力学特征	双重注意力+LSTM	注意力机制	中国东海大桥海上风电场数据
	[49]		气象条件、空气动力学特征	门控循环单元(GRU)	特征重要性、LIME	中国东北某地区风电场
	[50]		气象条件、空气动力学特征、风机知识	改进权重无关神经网络(WANN)	WANN	中国宁夏某风电场某风机数据
	[51]		气象条件、空气动力学特征	多元注意力机制+LSTM+轻量梯度提升机(LightGBM)	LightGBM、注意力机制、可视化	中国西北某风电场气象预测数据以及实测数据
	[52]	风电消纳能力预测	风电出力、系统状态	DNN+极限梯度提升机(XGBoost)	XGBoost	2018年北爱尔兰电网历史数据
	[53]	风电场景生成		变分自动编码器	T分布-随机近邻嵌入	2018—2021年韩国海上风速数据
太阳能相关	[54]	光伏出力预测	气象条件	随机森林	LIME、SHAP、ELI5(Explain Like I'm 5)	2014年全球能源预测竞赛数据
	[55]		气象条件	XGBoost	SHAP	光伏电池板与气象站数据
	[56]		气象条件	XGBoost	ELI5	欧道明大学某光伏电站实际数据
	[57]	太阳能辐射预测	气象条件	两阶段的聚类模糊规则学习框架	数据库和规则库特征	校内Vantage Pro 2气象站数据

1)风电出力预测

现有风电出力预测均具有较高的准确度，但其内部决策机制无法完全被使用者理解。为解决此问题，文献[49]提出一套基于可解读概念的风电功率超短期预测模型分析方法，并通过LIME对风电功率预测效果较好和较差的个体分别解读，分析影响风电功率预测模型实际表现因素(例如170 m高度风速、100 m高度风速和气压)。同时，根据解释结果更改风电超短期功率预测策略，提高了风电出力预测的准确率，形成了“建模—测试—解释—提升”的反馈闭环。文献[50]改进了WANN[58]并应用到风机功率超短期预测中。WANN是一种独特的神经网络架构，其强调网络结构而非权重的重要性。相比于DNN中数百万甚至数十亿的参数，WANN的权重简化使得神经网络架构更易于分析和理解。在该文中，WANN的解释性主要体现在网络激活函数依据特征本身相关的属性进行选择，例如，风速相关特征以线性函数或绝对值函数为主，变桨角度、风向角等角度特征以正弦函数或余弦函数为主，这与风机物理特性相近。为挖掘海上风电场出力预测的时空关联并实现模型的可解释，文献[59]提出了基于多重时空注意力图神经网络(graph neural network，GNN)的风电出力预测模型。该模型在空间、特征和时间3个维度引入多头注意力机制，借由各维度注意力权重的热力图解释模型工作机理，并结合专家知识验证了解释结果的合理性。

2）光伏出力预测

除风电外，光伏发电自身所具有的不确定性与光伏装机能力的加强也需重新设计预测方法，通过考虑多种气象因素（如云量、辐照度等因素），提高了光伏预测的准确性。针对随机森林回归实现的光伏发电预测模型，文献[54]应用LIME、SHAP和ELI5这3个XAI工具进行了解释。其中，ELI5全称Explain Like I'm 5（即像“我五岁”一样解释），即解释非常容易理解。作为一种可解释机器学习模型的Python库[60]，ELI5可以解释线性分类器和回归器的权重和预测，显示决策树特征重要性并解释决策树和基于树的组合的预测。在该文中，ELI5被用来检查单个光伏出力样本的预测情况，LIME利用光伏原始特征值解释特定数据，而SHAP则用于分析光伏预测的全局特征重要性，以及局部特征间的交互影响（如表面太阳辐射率下降值与每日时刻间的关系），在其算例中得出了“表面太阳能辐射是光伏出力预测中最重要的特征”等知识。另外，文献[55]利用树结构自组织映射与XGBoost实现了光伏功率预测，并借由SHAP开展解释，在算例中发现“总辐射度平均值对大多数样本预测结果影响最大”等相关知识。针对可再生能源发电不确定性与预测负荷之间存在的误差问题（鸭子曲线），文献[61]提出一种结合注意力机制和LSTM模型的XAI方法，用于理解气象因素对光伏出力的影响。注意力机制在预测输出时聚焦于特征的某些部分，并利用上下文向量、注意力权重和输入变量来显示对该预测值的解释。文献[62]针对数据驱动难以综合考虑模型结构和数据特征给出合理解释的问题，提出一种融合多注意力DNN的可解释光伏功率区间预测模型，利用神经元电导梯度法解释模型结构，结合注意力权重分析光伏功率预测特征，发现：随着模型训练迭代次数的增加，特征注意力权重变化可以很好地揭示与光伏出力预测最相关的因素。

3.1.2　负荷预测

除了出力预测外，XAI在负荷预测中的应用也已引起广泛关注（表4总结了相关文献），特别是在电力负荷和多元负荷的预测中，以下将详细介绍这两个方向的文献。

表4　XAI在负荷预测中的相关文献

Table 4　Related literatures on application of XAI in load forecasting

文献	年份	研究方向	输入特征	预测方法	解释方法	测试系统
[45]	2020	电力负荷预测	气象因素、室内气候因素	LSTM构成seq2seq模型	LIME	韩国能源研究所实际数据、EnergyPlus生成的数据集
[63]	2021		气象因素、耦合特征、历史负荷	XGBoost	SHAP	大西洋中部地区2014—2016年电力负荷数据
[64]	2021		气象因素、室内气候因素、历史负荷	注意力机制＋LSTM	注意力机制	中国山东省青岛市某办公楼数据
[65]	2022		气象因素、日类型因素、能源价格（图像数据）	循环扩张机制＋卷积GRU＋Transformer	注意力机制	2014年全球能源预测大赛数据
[42]	2021	多元负荷预测	气象因素、日类型因素等	LSTM＋多任务学习	SHAP	美国国家可再生能源实验室官网中某实际楼宇综合能源系统数据
[66]	2022		气象因素、耦合特征、历史负荷	耦合特征挖掘算法＋LSTM＋多任务学习	SHAP	美国亚利桑那州立大学坦佩校区综合能源系统数据
[67]	2022		气象因素	LSTM	标准化回归系数法（SRC）、LIME和SHAP	EnergyPlus生成2 000个建筑信息数据
[68]	2022	冷负荷预测	气象因素、日类型因素等	XGBoost	SHAP	中国深圳某办公建筑数据
[69]	2021		气象因素	XGBoost	SHAP	美国国家海洋和大气管理局和国家太阳辐射数据库气候数据、EnergyPlus模拟仿真数据
[70]	2021		气象因素、耦合特征、历史负荷	注意力机制＋循环神经网络	注意力机制	某商业建筑数据

在电力负荷预测领域，文献[71]利用SHAP解释了基于LightGBM的建筑电力负荷预测模型，在该模型中发现“在电力负荷预测中，温热指数和风寒温度表现出比气象信息的温度、湿度和风速更大的影响”等知识。文献[72]提出基于LSTM、多任务学习和人口流动性优化的新冠肺炎期间短期负荷预测模型，利用Shapley可视化模型进行解释，认为：相关性高的指标与其对预测结果的贡献并不同步。同样，文献[73]利用SHAP解释了用于日前负荷预测多层感知机模型，发现：在模型中添加节日变量会

被赋予一个相对较大的负Shapley值。

在多元负荷预测领域，文献[67]为了寻找建筑负荷预测模型影响最大的输入变量，采用SRC和LIME、SHAP确定建筑负荷预测模型的输入特征。其中，SRC是一种用于线性模型的全局敏感性分析方法[74]，其核心思想是将自变量与因变量进行标准化处理，从而消除量纲对回归系数的影响，并通过比较标准化后的回归系数来评估各自变量对因变量的相对重要性。SRC与LIME、SHAP共同用于比较对负荷预测中输入变量的简化能力，SHAP比LIME在建筑负荷预测模型中更有优势，因为在大多数情况下SHAP可以比LIME用更少的基本输入变量保持准确性。在交通能耗方面，文献[75]利用LIME解释家庭交通能源消耗的预测结果，分析家庭出行特征（如家庭旅行、人口统计和邻里）对特定交通分析区交通能源消耗的重要性和影响，认为XAI不仅能够提供透明的预测结果，而且能提供不同特征对预测家庭交通能源使用的影响。

3.1.3 XAI在源荷预测应用的进一步思考

在源荷预测应用中，XAI可提供相应输入参数对预测结果的影响，从而更好地了解预测模型的行为。尽管有多种XAI方法可以应用，但它们仍有一定局限性。下面分别从XAI方法与XAI在源荷预测中应用的问题两方面进行分析。

1）源荷预测中的XAI方法

第一，源荷预测场景下的特征选取。用于源荷预测的XAI方法所解释的特征一般采用预先定义特征，并非自动提取特征。一方面，预先定义特征的优势在于更易于包含电气领域的专业知识，相对于自动提取特征的可解释性更强；另一方面，预先定义的特征会造成XAI方法解释范围受限，只限于解释预先定义好的特征，存在泛化能力差的问题。如何平衡好二者的关系需要结合工程实际开展研究。

第二，源荷预测场景下的时空关系。源荷预测所使用的XAI方法难以有效提取特定时刻前后的特征，导致模型难以考虑源荷预测样本所在空间的全局景象。同时，源荷预测可能涉及从几分钟到几天甚至几个月的时间尺度。不同的时间尺度可能需要不同的模型结构和特性。因此，可进一步开发具备多时间尺度全局解释能力XAI方法，如概率图模型[76]、循环神经网络的代理模型[77]等，以在源荷预测中实现不同时刻间可解释性特征的比较和分析。

2）XAI在源荷预测的应用

第一，源荷预测场景下的XAI算例构建。目前，尚未存在针对源荷预测与解释的标准算例系统。源荷预测在电力系统运行、调度及规划中具有重要地位，关系到系统稳定性与经济效益。XAI能够揭示源荷预测内部运作机制，为决策者提供透明、可靠的预测理由，从而确保更加科学、合理的电力调度和规划决策。因此，构建一个针对电力系统源荷预测的标准XAI算例系统，不仅可以为研究者提供统一的研究平台和基准，还有助于推动电力系统预测及决策支持的智能化和精确化。

第二，源荷预测场景下的人机交互。需进一步提升源荷预测场景下XAI的交互性优化与人类的协同作用。例如，利用可视化手段，将生成的可解释特征直观地展示给工业部门或物业负责人等用户，以帮助其更好地理解负荷预测结果，实现进一步能源管理。另外，还可与ChatGPT等大型预训练语言模型合作，通过问答形式与用户交互，提供有针对性的解释及建议。例如，当工业用户询问为什么明天的负荷预测比今天高时，智能解释系统可根据模型分析结果回答："明天的气温预计会降低，导致暖气使用增加，同时晚上将进行一次安保演练，需要照明设备加班，这也是负荷预测增加的原因。"这种高度互动的方式有助于管理者深入洞悉模型预测背后的驱动因素，并根据实际情况调整能源使用策略。此过程不仅实现了技术与人的紧密融合，更是将能源管理知识与策略传递给了人类。

第三，源荷预测场景下的XAI应用限制性。现有源荷预测研究中，XAI方法多为直接套用SHAP或LIME的官方案例。这种统计学方法在某种意义上忽视了电力领域的物理原理与特性，无法从物理意义上解释预测结果的依据。同时，现有研究未能针对不同源荷预测的特点设定独特的人工智能预测方法与解释方法，这限制了将特定领域知识深入、准确地传递给人类。因此，亟须开发面向特定应用场景的解释工具与理解方式，以定量分析预测结果的特定影响特征和交互特征之间的相互关系。

3.2 XAI在电力系统规划中的应用

"双碳"背景下，高比例可再生能源与高比例电力电子设备接入使得新型电力系统规划问题愈发困难。系统规划的目标是设计和管理复杂的电力系统，以提供高质量、可信、安全和持续的电力服务，XAI可以为此提供重要的支持。近期，文献[78]开发了一种机器学习流程，用于预测发电项目能否成功。该模型应用梯度提升树构建模型，利用工厂容量、燃料、网络连接、所有权和融资情况等特征进行项目成功预测，并利用SHAP对预测结果解释。值得注意的是，该文不仅研究了一阶SHAP值（某一特征对预测结果的影响），还研究了二阶SHAP值（某一特征与其他特征的相互影响），丰富了SHAP

在电力系统中的应用。针对风能和太阳能电站的优化选址问题，文献[79]提出一种基于XAI的风能太阳能发电厂选址适宜性评估方法，利用超过55 000个真实风光电站的特征因素训练了随机森林、支持向量机和多层感知机模型，基于SHAP解释了风光电站选址过程中技术、经济、环境和社会因素对风光电站选址空间决策的影响，解释结果展示："风电场最适合选址的地点是水位高、距离城市远的地方，而靠近城市和温度适中的地方最有可能投资太阳能光伏电站。"

总的来看，XAI在系统规划中的应用较为匮乏，未来可以在下面几个方面探索：

第一，系统规划场景下的XAI风险研究。电力系统中，人工智能技术可实现高效的规划方案生成，但在生成方案的过程中需保证规划方案的可解释性。项目管理人员对人工智能所生成的规划方案及其潜在风险深度理解是至关重要的，使管理者了解方案的优劣性与可能后果。此方向可重点研究人工智能规划模型与模型解释工具的交互式迭代，以实现更好的风险管理。

第二，系统规划场景下的XAI公平性研究。XAI技术应用带来更高效的电力系统规划，同时可挖掘各个方面的优势和劣势。可防止由于数据样本偏差造成的规划问题，平衡不同区域、企事业单位和消费者的利益，推进电力系统规划的公平性。

3.3　XAI在运行优化与稳定控制中的应用

电力系统的运行优化与稳定控制是指在保障电力系统安全稳定的前提下，提高其运行效率与经济性。而XAI技术的引入有望提高其决策透明度，从而增强决策者的信任。本节将分别介绍运行优化、稳定控制两方面应用。

3.3.1　运行优化

运行优化场景中，XAI可有效地提高系统运行优化有效性，并且帮助用户理解系统的优化结果。例如，文献[80]将SHAP用于光伏系统无功控制的DNN，量化了负荷和光伏有功功率对相应的最佳光伏无功功率的影响，明确了哪些电网状态信息对每个光伏系统的最佳无功功率调度有重大影响，并分离出重要的特征子集用于建立分散的DNN控制器。针对电能质量扰动分类问题，文献[81]提出一种基于Grad-CAM和遮挡敏感性CNN的分类器解释方法。该方法通过提高分类器的透明度，使电力专家能够在发生电能质量扰动事件的情况下做出明智和可信的决策。同时，该论文还给出了电能质量扰动应用中关于正确解释的精确定义。

除上述基于监督学习的运行优化场景外，XAI目前比较受到关注的是解释性用于优化调度的RL模型。尽管该问题研究尚处于起步阶段，但仍有一些研究可供参考借鉴。下面分别从能源领域RL可解释性、能源领域RL物理知识嵌入，以及人工智能领域RL可解释性三方面进行介绍。

1）能源领域RL可解释性研究

由于多数RL模型由各类神经网络组件构成，针对神经网络应用的人工智能黑箱解释方法均可在此类RL中套用，以提高整体可解释性。例如，文献[82]提出利用SHAP反向传播深度解释器为基于DRL的电力系统应急控制提供合理的解释，其中，SHAP的作用为解释DRL所有输入特征对所有输出动作的重要性。在建筑能源领域，文献[83]为提高RL策略的透明度，在利用SHAP揭示每个输入对最终决策影响的同时，利用树方法提取RL中的关键控制规则，以实现最终控制决策的可解释性。另外，文献[84]针对电力系统调度中的图深度Q网络模型，提出了一种改进样本平衡深度沙普利加性解释（sample-balanced deep Shapley additive explanation，SE-DSHAP）与子图解释器相结合的多层级解释方法。一方面，SE-DSHAP能够对图深度Q网络模型的特征贡献进行排序，并通过样本均衡提高解释效率；另一方面，子图解释器用于选择与SE-DSHAP高亮节点相关的关键区域。总的来看，利用经典的深度学习模型解释方法（如SHAP）确实能够提供一定DRL内部可解释性，但现有研究大多针对其中的神经网络部分，而针对RL的其他要素（如环境、目标等）[85]及其相互关系的解释研究则关注较少。针对电力系统环境中RL各类要素的解释，还需要结合电力领域的专业知识进一步分析。

2）能源领域RL物理知识嵌入的RL研究

物理知识嵌入是一种将领域知识直接嵌入RL模型的方法。电力系统优化调度中，其主要表现为将电力系统的物理规律、操作约束和业务规则等领域知识嵌入RL模型中。通过物理知识嵌入，提高了整体RL优化调度框架的性能、鲁棒性和可解释性（但RL内部神经网络的可解释性不受影响）。文献[86]提出一种大型电网潮流计算收敛的自适应调节策略，将知识和经验整合到学习过程中，以减小搜索范围。通过模拟人工调整与设置多重奖励机制，使得搜索算法具有明确的方向，提高了算法动作的可解释性。文献[87]对电力调度中应用的数据与知识联合驱动的人工智能方法做了总结，认为：知识数据联合驱动方法在理论上可借助因果思想与机理知识来提高数据驱动方法的解释性，但在增强解释性的过程中，也可能由于知识融合方式不恰当而导致

联合驱动模型的拟合性能和精确性下降。因此，如何优化领域知识在数据模型中的嵌入方法及控制嵌入尺度，是需要根据电力领域的实际问题开展研究的。

3)人工智能领域RL可解释性研究

在人工智能领域，已出现一些关于RL可解释性的论文，可为电力系统优化调度应用提供启发。文献[88]对RL可解释方法进行综述，将其按照关键要素的构成划分为环境解释、任务解释与策略解释。文献[89]将RL可解释问题分为策略与目标模型解释问题、响应与局部目标解释问题和模型检查问题。文献[90]认为RL的解释性应该体现在智能体不仅能够执行所请求的任务，而且可以在自己不确定的时候将任务交还给人类。总的来看，上述文献均认为现有RL可解释性需要研究更全面的解释方法(尤其是在应对安全敏感任务时)，以解释RL智能体在各个方面的行为，从而为人类(例如调度员)提供易于理解的决策解释。

基于上述分析，XAI在运行优化场景未来可探索的方向包括：

第一，运行优化场景下的RL环境感知解释。在智能体环境中，可研究电力系统调度环境重要因素的定量评估方法，利用注意力机制分析大电网环境状态变化特征与智能体(机组、园区等)动作结果相关关系，量化分析电力系统环境状态因素对动作结果的冲击影响；同时，研究基于多园区(例如多微网、综合能源系统群等)的多智能体调度问题时，可考虑多智能体群体间交互特征的解释方法，结合博弈论、电网物理规律、群体动力学等解释掌握环境规律。

第二，运行优化场景下的RL动作决策解释。可利用因果科学、反事实推理，考虑已观测的智能体动作经验链条，对动作过程的已知变量构造不同取值，进行调度过程的因果推断与反事实推理以及规则提取。现有研究有一些可供参考，例如，文献[91]提出一种用于RL的软注意力机制，能够通过软注意力机制迫使智能体顺序查询环境视图来关注任务相关信息，可以直接观察到智能体在选择行动时所使用的信息。

第三，运行优化场景下的RL奖励反馈解释。奖励是环境对智能体动作的反馈，研究基于敏感性分析的智能体奖励解释方法，例如，通过在大电网调度过程中对智能体动作添加不同程度的微小扰动，分析智能体受到环境给予奖励反馈的变化程度，建立智能体与环境之间的“策略-奖励”融合表达形式，提炼电力系统的调度奖励规律与奖励函数分解设计方法，为后续同类RL问题的奖励设计提供思路。

第四，运行优化场景下的人机交互。研究RL智能体与人类调度员的交互机制对于调度过程的实时风险评估处理与知识双向传递至关重要。例如：文献[92]开发一种基于实时人工指导的DRL方法，用于端到端自动驾驶策略训练。通过所设计的控制转移机制，人类能够在模型训练过程中实时干预和纠正智能体的不合理行为。这种“人在回路”的指导机制可尝试在电力系统调度过程使用。通过设计直观的可视化界面展示智能体状态、动作及奖励机制，开展高效的知识传递，有助于调度员快速捕捉关键信息和趋势。在关键决策点引入调度员参与，利用人类专家经验提升决策质量。通过实时学习调度员动作反馈，使智能体不断优化决策策略，实现智能体自主进化。

3.3.2 稳定控制

稳定控制场景中，人工智能模型的任务是判断系统在受到一定程度扰动后能否正常运行。然而，当人工智能黑箱模型做出稳定性评估时，往往难以给出判断的原因[93]，阻碍了系统稳定性的进一步分析。针对该问题，本节分别针对暂态稳定、静态稳定和新能源并网三方面XAI应用进行了综述。XAI在稳定控制领域应用的相关文献如表5所示。

1)暂态稳定

针对暂态稳定问题，研究者们主要从人工智能模型解释方法和具有自解释能力的人工智能模型两个方面进行深入探索。首先，在人工智能模型解释方法方面。文献[43]提出基于SHAP的暂态电压稳定评估归因分析框架，通过计算Shapley值的平均绝对值大小得到暂态电压稳定特征重要性排序，并根据每个特征的边际贡献，进一步量化不同输入特征对模型输出结果的影响。针对如何实现快速、准确的电力系统暂态稳定在线评估的问题，文献[104]提出基于改进一维CNN的电力系统暂态稳定评估方法，通过Grad-CAM对暂态评估模型的类激活图，并结合系统拓扑结构进行可视化分析。文献[105-106]利用电力系统故障时发电机的运行特征构建基于XGBoost的暂态稳定预测模型，并对具体故障利用LIME进行解释，包括特征对单个样本分类贡献率以及它们如何影响预测结果。例如，针对某支路预测结果(三相短路故障)的情况，给出了具体特征对该情况的作用。文献[107]采用SHAP从全局和局部两个维度对XGBoost模型和样本开展解释，找出了稳定关键特征，使模型更加透明。其次，一些学者采用具有自解释能力的人工智能模型，提升了暂态稳定评估的可解释性。针对暂态稳定问

表5　XAI在稳定控制领域应用的相关文献
Table 5　Related literatures on application of XAI in stability control

参考文献	评估方法	解释方法	测试系统
[43]	支持类别特征的梯度提升树	SHAP	1)IEEE 39节点系统；2)美国南卡罗来纳州500节点系统
[94]	LightGBM		1)美国西部电力协调委员会3机9节点系统；2)IEEE 39节点系统；3)IEEE 300节点系统
[95]	梯度提升树		欧洲互联电网平台的公开数据集
[96]	DNN		IEEE 39节点系统
[97]	DNN	加权线性回归＋正则化的局部代理	IEEE 39节点系统
[98]	改进深度信念网络	局部线性解释模型	IEEE 39节点系统
[44]	CNN	LIME	1)IEEE 39节点系统；2)IEEE 118节点系统
[99]	XGBoost		IEEE 39节点系统
[35]	自注意力机制＋Transformer	注意力机制	IEEE 39节点系统
[100]	自注意力机制		IEEE 39节点系统
[101]	斜回归树＋Boosting集成	斜回归树Lasso/Ridge正则化	1)改进IEEE 30节点系统；2)美国国家可再生能源实验室73节点稳定测试系统
[102]	决策树	决策树	IEEE 68节点系统
[103]	GRU＋决策树		IEEE 39节点系统
[104]	改进一维CNN	Guided Grad-CAM	IEEE 39节点系统

题中常规DNN可信性与时间信息提取能力同时不足的问题，文献[36]在DNN内部工作过程，利用自注意力引导模型在训练迭代中自适应聚焦于重要特征，使模型快速捕获电力系统前后时刻间的状态依赖，展示出良好的解释性。针对深度学习预测器在基于相量测量单元的故障后暂态稳定评估预测结果无法解释的问题，文献[100]提出一种嵌入自注意力层的深度学习模型和迁移学习策略。这种嵌入的注意力层能够识别受干扰最大的生成器，以增加模型的可解释性。

2)静态稳定

针对静态电压稳定裕度估计中关键特征量筛选问题，文献[94]基于累积贡献率和SHAP的关键特征量开展特征筛选，利用SHAP模型依据贡献值大小对特征降序排列，采用累积贡献率增量的循环优化过程剔除冗余特征，体现了SHAP模型在事前进行特征选择和优化的潜力。文献[108]在电压稳定裕度预测中利用自然梯度提升树与SHAP解释理论，分析了电压稳定裕度预测的影响因素。

3)新能源并网

针对新能源发电并网问题，XAI同样可判断稳定程度。其中，孤岛问题是指在电力系统中，当主电网侧出现计划或非计划的断电情况时，分布式发电系统（如光伏、风电等）仍然继续为本地负载供电的情况。这种情况会形成一个“孤岛”，即分布式发电系统和本地负载构成一个独立于主电网运行的微电网。针对如何检测孤岛的问题，文献[109]提出一种基于LightGBM的孤岛检测可解释模型。分别利用SHAP、LIME和累积局部效应（accumulated local effect，ALE）对训练好的孤岛识别模型输出决策进行成因分析。ALE[110]作为一种可解释技术，其首先计算每个特征在相邻特征值之间的局部效应，然后将这些局部效应累加，得到该特征对模型预测的总贡献。通过这种方式，可将预测结果分解为各个特征的贡献，从而实现模型的可解释性。ALE被应用于计算局部效应来消除电气特征量强相关性的干扰，实现了对并网检测系统孤岛模型的可解释分析。针对风电系统与电网互动的次同步振荡问题，文献[111]提出一种风电并网系统次同步振荡在线预测和优化控制方法，基于LIME分析了各风电场开机台数对次同步振荡的影响，识别出对系统次同步振荡影响最大的风电场和风机，并将该知识传递给人类，人类基于此知识制定切除风机或施加阻尼等策略缓解了次同步振荡。

综上所述，与同样基于监督学习的源荷预测应用相似，稳定评估场景下大量工作采用LightGBM与LIME、SHAP等后验解释技术相结合，通过对已有模型的分析和解释来对评估结果进行解释。此类事后解释方法一般遵循“离线训练＋在线部署”形式，需在模型预测或决策完成后对结果进行分析与解释。因此，人工智能模型解释方法总体计算时间通常在秒级范围，具有较好的实时性，能够满足绝大多数实时稳定性评估的应用场景。但稳定控制应用中仍有如下问题：

第一，稳定控制场景下的人机交互。稳定控制需要考虑系统受到扰动后能否正常运行。而经由XAI获得的部分评估结果仍需借助人类电力专家的领域知识、行业经验和判断力来解释模型输出结果的需求，XAI方法在很多情况下仍无法完全透明和直观地解释评估结果，可能存在偏差、局限性或不完整性。同时，可解释性这一性质并不直接等同于决策者对算法和其输出结果的完整信任度。未来，还需将可解释性与可信性结合研究。

第二，稳定控制场景下的拓扑知识及其解释性。传统监督学习方法在稳态评估上主要依赖系统状态数据与稳定类别的映射关系，往往忽视对网络拓扑及其高维属性的解释分析。因此，需研究如何将包括拓扑结构、运行方式在内的高维信息嵌入稳定评估智能模型中，旨在实现不依赖人类专家输入而具备高度可解释性的稳态评估机制。值得注意的是，人工智能领域已出现GNN解释相关研究[112]，具有一定潜力，为解读图结构黑箱模型行为提供了直接的理论支撑。未来研究可侧重于GNN解释性[113]、异质信息网络[114-115]、表征学习[116-117]等前沿技术的综合应用。

3.4 XAI在电力系统故障诊断中的应用

高效且精确地实现故障诊断对电力系统的安全可靠运行至关重要。现有人工智能黑箱的故障诊断方法由于缺乏可解释性，难以提供诊断过程的详细解释，使得运维人员对其信任度下降。因此，XAI可提供诊断辅助信息等相关知识，帮助运维人员理解人工智能模型诊断结果，同时指导人工智能模型的改进和优化，进一步提升诊断结果准确性和可信性。本节分别从系统级故障诊断、设备级故障诊断以及储能系统状态分析三方面给出XAI的应用。

3.4.1 系统级故障诊断

在系统级故障诊断场景下，人工智能模型可用于检测、诊断系统可能存在的故障，同时，XAI应提供针对故障清晰、透明的解释。在此基础上，运维人员应基于XAI所传递的知识，根据系统情况进行及时的决策。针对高压输电线路的发展性故障难以辨识问题，文献[118]利用带有解释能力的注意力机制模块（convolutional block attention module，CBAM）[119]嵌入CNN实现一种发展性故障识别方法。通过在卷积层间添加CBAM，使网络自动聚焦于电流波形的幅值、突变等重要特征，可视化异常波形的整体概貌，提高了模型内部的可解释性。文献[120]使用SHAP解释了基于LSTM的电力系统事件识别结果，提出可将SHAP值纳入神经网络损失函数中，作为约束条件引入知识领域。实际上，该方法可以与物理信息神经网络（physics-informed neural network，PINN）[121]进行对比区分。PINN可将带有物理知识的微分方程作为约束加入神经网络的训练过程中，并通过调整网络参数匹配微分方程的物理约束条件。尽管PINN能在一定程度上实现知识模型与数据模型的耦合，并通过近似解析解来提高模型的精度与鲁棒性，但其主要目的仍在于提高模型训练效果，而非直接增强模型的可解释性，其内部的决策过程仍难以解释和理解。相比于PINN，文献[120]在引入知识领域约束条件的同时，SHAP值嵌入损失函数的方式提高了人工智能模型的可解释性。

3.4.2 设备级故障诊断

在设备级故障诊断场景下，XAI通过分析设备故障数据等相关信息，可向人类传递设备故障关键因素、形成原因等决策信息。针对光伏阵列故障，文献[122]设计了一种实际的可解释性故障检测系统，由二极管设备、XGBoost、LIME和相关边缘架构组成，使用LIME构建XAI。在该系统的实际应用过程中，带有可解释属性的模型帮助现场工程师理解了故障原因，显著减少故障排除和故障检测时间，同时通过XAI减少了维护操作，节省了资源。同样针对光伏阵列故障，文献[123]利用SHAP、Anchors[124]和反事实解释（diverse counterfactual explanations，DiCE）[125]3种事后XAI解释了基于神经网络的光伏故障检测结果。其中，Anchors是一种基于规则的解释方法，其通过产生一组规则以在观察值附近稳定地锚定预测值，这些规则揭示了模型在给定输入的情况下所依据的关键特征。而DiCE则专注于生成多样化的反事实解释，提供了在不同分类结果下的替代实例，从而阐明模型的决策边界。相对于Anchors和DiCE，SHAP在光伏故障检测中表现出更高的稳定性和一致性。针对海上风电机组齿轮箱运行状态的监测问题，文献[126]提出一种GRU和注意力机制相结合的监测方法，利用注意力权重表征输入特征对目标建模特征建模的贡献率，发现了“齿轮箱油温与电机转速、叶轮转速等风机运行状态紧密相关”等知识，并结合训练过程的权重变化与领域知识验证了所提方法的合理性。在制冷系统，为了优化不规则露点冷却器的运行，文献[127]开发了一个可解释的DNN模型，基于SHAP解释了设备操作与参数设计对设备性能的影响。为了增加用户对模型的信任，文献[128]开发了一个移动应用程序，利用深度学习图像识别模型，从捕获的图片中自动识别能源设备品牌名称和型号，利用Score-CAM[129]识别由模型捕获的无关特征进行

校正。

3.4.3　储能系统状态分析

在储能系统状态分析场景下，XAI也有部分应用。文献[130]采用随机森林与可解释方差评价指标对储能系统运行单元进行重要性分析。可解释方差是一种机器学习的评价指标，可解释方差数值越大意味着模型的自变量越能解释因变量。针对新能源汽车动力电池荷电状态预测，文献[131]考虑了电池非线性和模型可解释性，采用较容易解释的集成学习方法构建荷电状态预测模型，取得了较好效果。

3.4.4　XAI在故障诊断应用的进一步思考

未来，XAI在故障诊断场景下可能的研究方向包括：

第一，故障诊断场景下的多模态诊断解释。现有故障诊断方案依赖于一维时间序列以及特征图的二维展示，可利用图像相关XAI方法（例如Grad-CAM）开展解释。未来可研究不同细粒度、不同数据结构（包括文本数据、图像数据等）的特征，可通过多模态实现特征融合，从而达到不同层次的解释能力。在这方面，文献[132]综述了多模态DNN的可解释性，可为电力系统多模态故障诊断解释提供参考。

第二，故障诊断场景下的特征选取。故障诊断领域，未来须开展复杂特征的解释，如故障情况下的声音特征、电场磁场变化及有关系统运行的其他参数，或者地理信息、故障地点等图结构特征。故障的子系统可能存在多重耦合和高度相关的特征集合，以及复杂的网络结构和行为变化。此背景下，XAI可当作一种特征选择方法，帮助人类进一步理解特征重要性及特征间相互关系，确保复杂耦合的特征能够被有效地表示和转化为维数可控、算法可用的机器学习特征。

第三，故障诊断场景下的解释体系构建。特定应用场景需建立对应的解释体系，以便对故障类型和原因进行精确地分类和解读。为增强解释的可信度，可考虑将不同的解释方法进行耦合。例如，结合全局与局部的解释策略，或者在模型的前期与后期进行解释耦合。

第四，故障诊断场景下的XAI迁移能力。针对不同的电力系统和设备，其故障产生的原因与规律可能存在差异。因此，尽管XAI在电力系统故障场景下具有良好的预测能力，但其模型仍然具有一定的局限性。为提高XAI模型的可迁移性，并使其适用于不同的系统和设备，需要在XAI模型的开发中考虑这些差异，并进行相关的适配和优化。

第五，故障诊断场景下，故障解释结果与故障知识库联合开发利用。通过将XAI生成的解释结果与已有的故障知识库结合，可进一步提高故障诊断的准确性与便利性。一方面，解释结果可为故障知识库提供更深入的洞察力，揭示故障现象与特征间潜在的关联，从而丰富知识库的内容并提高其实用性；另一方面，故障知识库可为解释结果提供实际案例与经验，有助于验证和修正模型生成的解释，使其更符合实际情况。上述过程中，未来可开展系统故障的知识表示（因果图分析[133]、反事实表示[134]等）、知识融合（贝叶斯网络推理[135]等）等方面的研究。综合来看，可解释结果与故障知识库的结合，既能提升知识库的质量，又能增强可解释结果的实用性与准确性，从而为故障诊断带来有效的技术支持。

3.5　XAI在电力及能源市场中的应用

电力市场作为复杂的多方参与市场，需要不断精细运作、长期规划。在此过程中，人工智能黑箱模型难以向用户阐明其计算规则和预测过程，这种情况不仅阻碍用户对算法结果的理解和信任，也可能导致市场决策的不稳定性和市场风险的增加。XAI通过提供透明的计算过程与可解释的预测结果，有助于提高电力市场参与者对人工智能方法的信任度，降低决策不确定性与潜在风险。为此，本节探讨了XAI在能源价格预测、市场出清方面的应用。

3.5.1　能源价格预测

能源价格预测能够帮助社会各界对能源价格变化做好准备，以更好地应对能源价格风险。同时，能源价格预测也可为政府制定政策与投资决策提供重要参考依据。文献[136]提出了一种考虑人机协作的批发市场日前电价预测的新框架，由5个阶段组成：数据理解模块（提高用户对数据集的洞察力）、模型性能模块（用于评估模型质量）、模型审核模块（对残差进行诊断性评估）、特征灵敏度与特征影响模块（用于分析影响电价变化的输入数据来源）和模型简化模块（采用决策树近似表示预测模型）。这5个模块通过量化技术与可视化技术增强了模型可解释性。针对单个市场传统特征集难以支撑高精度预测需求的问题，文献[137]提出一种考虑校准窗口集成与耦合市场特征的可解释双层日前电价预测框架，基于SHAP分析了耦合市场特征与传统特征对电价预测的影响，发现对北欧电价影响最大的特征依次为总发电量、风力发电量、历史电价及水力发电量。针对家庭能源价格问题，文献[138]搭建了一个面向非专家用户的XAI系统，该系统由可视化和文本组件组成，使用游戏化的方式加深用户对家庭能源价格预测模型（随机森林）的理解。同时，分别使用置换特征重要性（permutation feature importance，

PFI)[32]和SHAP进行解释。其中,PFI是一种传统的无偏见全局解释方法,具有直观、通用、简单的优势。PFI通过随机打乱各样本某一特征取值,对比打乱前后的结果,来判断各特征的重要程度:如果打乱该值增大了模型误差,则该特征较为“重要”;如果打乱该值对模型误差影响不大,则该特征为“不重要”。由PFI构成的程序让用户体验到不同特征如何影响家庭能源价格的预测结果。

3.5.2 市场出清

市场出清是电力市场的一个重要概念,指在某个时段内,按照市场规则和机制,通过市场交易将所有发电方案与需求方案进行匹配,确定最终的市场价格和电力供需。市场出清是电力市场的核心任务之一,也是市场竞争和资源配置的重要手段。文献[139]利用孤立森林算法对电力市场非典型价格进行识别,孤立森林作为一种基于决策树的算法,本身就具有可解释性,该文结合机组受限状态解释分析了机组启停原因、机组受限状态、断面阻塞原因以及非典型价格原因等因素影响。

3.5.3 XAI在能源市场应用的进一步思考

XAI在电力及能源市场中的应用可能会存在以下几个潜在的研究方向:

第一,能源市场应用下的多智能体RL的解释方法。目前,能源市场中的多方博弈问题(例如用能用户与能源供应商间的策略互动)可通过多智能体RL进行有效求解。然而,考虑到能源市场的高度时空动态性,其模式和关系随着时间的推移可能发生变化,能源价格、供需情况同样易受到各种复杂内部因素和外部环境因素的综合作用。因此,在研究多智能体可解释RL基础算法(如文献[140-142])的同时,还需进一步探究针对能源市场复杂动态的多智能体决策解释方法。

第二,能源市场应用下的XAI隐私保护和安全保障能力。在电力市场中,如何确保市场参与者的关键数据和隐私在使用XAI过程中不被泄露或恶意攻击是一个研究焦点。需要研究针对XAI的数据泄露和恶意攻击防御方法,防止电力市场XAI的解释结果被恶意篡改,造成经济损失。未来可能的研究方向包括能源市场中的联邦学习的XAI[143]和XAI攻防技术[144]等。

第三,能源市场应用下的XAI监管与合规问题。在电力及能源市场中,监管机构对于模型的可解释性和透明度有着严格的要求。在应用XAI时,需要针对电力及能源市场的特点和监管规定进行模型设计和优化。通过与监管机构保持紧密沟通,确保模型能够符合标准法规所给定的可解释性与透明度要求。

4 XAI在电力系统中的挑战与展望

虽然XAI在电力系统日益得到重视,但由于电力系统人工智能的特殊性较为显著,目前对于XAI在电力系统中的运用仍存在较多悬而未决的挑战。因此,本章依照从小到大、由浅入深的逻辑尝试论述XAI在电力系统应用的挑战与展望(如图5所示),以期为XAI技术在电力系统的应用发展提供有价值的启示。

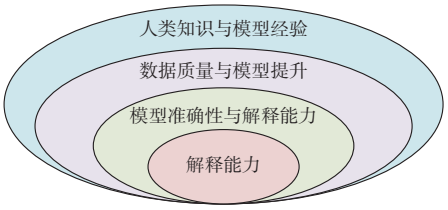

图5 XAI在电力系统中的挑战
Fig. 5 Challenges of XAI in power systems

4.1 电力XAI含义解释问题

目前,对于电力系统各领域XAI解释的具体含义还没有达成共识,主要受限于两方面:1)虽然现有解释形式对于电力领域中需具备可解释性的某些应用场景已足够适用,但XAI在理论和技术上仍然存在许多不足之处,尚未形成具有广泛适用性的、成熟的XAI电力系统解释性能评价框架[11];2)电力系统不同局部领域、不同应用场景、不同面向人群所需的解释能力也有所不同。可根据以下两方面开展电力领域人工智能解释的含义挖掘。

第一,在解释层次上,根据不同应用场景与目标受众区分解释层次。例如,可将解释目标受众划分为人工智能研究人员、电力专家和普通用户等。对于人工智能研究人员,可解释性需要在模型的内部机制、算法原理、参数调整等方面进行解释;对于电力专家,需要解释人工智能模型是否适合电力系统,以及电力人工智能应用的影响参数;对于普通用户,可解释性应主要体现在结果可视化,其可能更关注模型如何得出某个结果,以及是否符合他们的预期。根据不同的目标受众,可解释性需要在不同解释层次进行选择和调整,以确保其最大限度地满足不同层次的用户需求。

第二,在解释方式上,提高不同解释方式的适用性。“解释”存在着文本、图表、交互式界面等不同表达方式,但“解释”本身就是一种相对而言的概念,其效果的好坏与具体应用场景、目标受众、解释方式等因素都有关系。需要开发面向能源领域的多种解释形式,以根据具体应用场景和目标受众,结合实际需

求和预期，进行灵活的选择和调整，以最大限度地满足用户预期。

在高比例可再生能源渗透与高比例电力电子设备接入过程中，超大规模电网这一实际物理系统将进一步呈现高维复杂动态特性，如何在多时间尺度、多运行场景等不同层面下，准确提取系统外部环境数据与历史数据的事前解释指标，精确分析刻画高维非线性人工智能黑箱内部的事中解释指标，深入挖掘高置信、低偏差、可验证的人工智能黑箱模型结果的事后解释指标，形成电力领域人工智能黑箱可信度量证据与“事前-事中-事后”度量评估方法以助力电力系统人工智能结果可信评估与知识发现，是未来需要解决的关键问题。

4.2　模型准确性与解释能力矛盾

算法准确性代表着算法的性能和效果，而可解释性指的是算法的行为和决策过程的可理解性。一方面，可解释方法因专注于理解算法的决策过程，而忽视了算法的性能和效果；另一方面，准确性可能会损害可解释性，因为人工智能算法（尤其是深度学习）可能会通过一系列复杂的变换来提升算法的性能，可解释性就会受到影响。因此，大量模型在构建过程中需要在准确性和解释性之间进行权衡。例如，电力系统海量设备元件的高度耦合与非线性关系逐渐增加了其运行复杂性，相关人工智能模型可能会针对某些目标的实现设计复杂神经网络结构。虽然多层结构与节点互联能增加网络的表达能力、拟合高维复杂的非线性关系，提高准确性，但其网络隐层特征并未展示出明确的电力领域物理意义，网络连接较多时单个神经元的贡献也难以分离，不可避免地带来了可解释性低的问题。因此，在未来电力系统人工智能大规模应用的情况下，模型准确性与解释能力的矛盾亟须得到讨论。

模型准确性与解释能力的矛盾对抗可通过一定形式转化。机器学习算法之间在性能上的微小差异可能会被解释结果和在下一轮迭代中对参数的调整所消除[145]。因此，作为超大规模的复杂巨系统，电力系统中人工智能模型与XAI模型的交互迭代必将随着未来新能源出力、负荷水平、网架拓扑、运行方式等多因素的变化而变化，体现出高维动态的复杂特性。如何基于工程思维平衡人工智能模型准确性与XAI模型解释性之间的矛盾，形成面向特定电力场景的人机交互解释框架，构建“数据准备-模型调整-结果解释”的人工智能/XAI算法迭代循环，并推动电力系统“经济-安全-环境”矛盾三角向平衡演进，是未来需要解决的关键问题。

4.3　数据质量与模型提升矛盾

机器学习在电力系统中的应用研究通常关注于特定的模型设计，忽略了对电力系统数据质量提升。数据质量对可解释性的效果产生一定影响，这种影响在以模型为中心和以数据为中心的方法中存在差异。

以人工智能模型为中心的方法重点关注模型算法设计，可能会忽视电力系统所用数据的广度和深度，以及对潜在问题或规律的准确认识，且由于输入数据已经固定，通过复杂操作提升的模型能力可能会进一步限制模型的自解释性。同时，由于输入数据已经确定，输出结果的好坏较大程度上取决于人工智能模型的优劣，通过“输入-输出”映射实现的事后解释效果也已确定。因此，对于大部分自解释方法以及模型无关的解释方法来说，人工智能黑箱模型的好坏会直接决定解释性能。

以数据为中心的方法[146]主要关注数据质量提升与推理数据的针对性设计，其是特征工程与人工智能模型的核心关键。这种方法强调了数据预处理、特征构建和特征选择等环节的重要性，以便更好地捕捉和反映电力系统中的关键信息。然而，数据质量的提升并不一定会带来整体性能的提升，这是因为模型的复杂性和泛化能力也对性能产生较大影响。此外，以数据为中心的方法可能会遇到数据不足、噪声干扰或不平衡等问题，这些问题会影响模型的准确性与解释性。因此，对于模型无关的解释方法来说，数据好坏无法直接决定解释性能。

综上，基于更高维度、更大范围、更多模态的电力系统数据开展人工智能算法的设计、优化和解释的过程中，如何开展电力系统复杂数据的标注与预处理，如何提高人工智能模型清晰、透明与设计性进而提升可解释性，最终平衡以模型为中心和以数据为中心的矛盾主次双方，实现更高维度、更细粒度、更深层次的电力系统知识发现，是未来需要解决的关键问题。

4.4　人类知识与模型经验矛盾

电力系统内部已累积大量物理规律、人类经验、安全规则和因果逻辑等先验知识，其可通过知识-数据融合的方式嵌入数据驱动模型中，以提高整体模型的内在解释性与知识发现。2016年，薛禹胜院士提出因果方法与机器学习融合分析的思考[147]之后，知识-数据融合驱动的思想在能源系统开始得到初步研究，已在系统参数辨识[148]、暂态稳定分析[149]等领域取得了较好效果，展示出良好的潜力。知识-数据融合驱动技术的研究主要目的是将人类先验知识与数据驱动的模型相结合，以提高模型的准确性、鲁

棒性、泛化能力等。其与XAI的关联和区别主要表现在以下两方面：

1)在知识-数据融合的模型构建过程中，首先需要人类能够理解人工智能黑箱算法是如何工作的，才能设计出合适的知识模型与之交互。此时，XAI就发挥了“人类-人工智能”的桥梁作用。

2)虽然知识与数据的融合往往不会直接提高数据驱动模型的可解释性，但知识的加入能够增强融合模型内部的可解释性，并在一定程度上为模型输出提供更有价值的解释。

然而，现有模型内在解释方法仍难以统一电网不同业务下机理知识与数据驱动融合的差异，缺乏或难以形成对数模混合驱动模型构成规律的统一认识，如何将基于人类智能所形成的物理知识、实践经验，与基于人工智能所构建的高维表达、抽象行为相融合，形成人-机“你中有我，我中有你”混合驱动的物理信息融合新范式，是未来需要解决的关键问题。

5 结语

XAI已受到电力系统学术界、产业界及政府部门的高度重视，对现有XAI的研究、总结和归纳对促进人工智能在电力系统中快速发展与落地应用至关重要。本文从XAI兴起入手，对其背景意义、基本方法和应用范围3个方面进行分析与梳理，以探讨其对电力系统发展的影响与价值：从背景意义看，XAI应突破现有人工智能公平、信任、操作、安全瓶颈；从基本方法看，注意力机制、SHAP、LIME等已成为电力系统XAI常用方法；从应用范围看，XAI已渗透源荷预测、系统规划、运行控制、故障诊断、电力市场等领域。未来，XAI在电力系统的应用还面临解释范围界定困境、解释能力与准确性矛盾、数据模型主导平衡、数据机理融合范式挖掘等诸多问题，需把握不同矛盾的不同特点，结合电力系统实际开展具体研究。

参考文献

[1] Introducing ChatGPT [EB/OL]. [2023-03-29]. https://openai.com/blog/chatgpt.

[2] 韩笑，郭剑波，蒲天骄，等.电力人工智能技术理论基础与发展展望(一)：假设分析与应用范式[J].中国电机工程学报，2023，43(8)：2877-2891.
HAN Xiao, GUO Jianbo, PU Tianjiao, et al. Theoretical foundation and directions of electric power artificial intelligence (Ⅰ): hypothesis analysis and application paradigm [J]. Proceedings of the CSEE, 2023, 43(8): 2877-2891.

[3] 郭剑波.对新型电力系统演进趋势和关系的认识[C]//2022年中国电机工程学会年会，2023年2月9日，武汉，中国.
GUO Jianbo. Insights into the evolutionary trends and relationships of new power systems[C]// 2022 Annual Meeting of the Chinese Society for Electrical Engineering, February 9, 2023, Wuhan, China.

[4] GUNNING D, STEFIK M, CHOI J, et al. XAI—explainable artificial intelligence [J]. Science Robotics, 2019, 4 (37): eaay7120.

[5] 国家自然科学基金委.关于发布可解释、可通用的下一代人工智能方法重大研究计划2023年度项目指南的通告[EB/OL]. [2023-11-06]. https://www.nsfc.gov.cn/publish/portal0/tab434/info89087.htm.
National Natural Science Foundation of China. Notice on the publication of the 2023 project guidelines of the major research program on interpretable and generalizable next-generation artificial intelligence methods [EB/OL]. [2023-11-06]. https://www.nsfc.gov.cn/publish/portal0/tab434/info89087.htm.

[6] MILLER T. Explanation in artificial intelligence: insights from the social sciences[J]. Artificial Intelligence, 2019, 267: 1-38.

[7] MONTAVON G, SAMEK W, MÜLLER K R. Methods for interpreting and understanding deep neural networks [J]. Digital Signal Processing, 2018, 73: 1-15.

[8] CARVALHO D V, PEREIRA E M, CARDOSO J S. Machine learning interpretability: a survey on methods and metrics [J]. Electronics, 2019, 8(8): 832.

[9] TJOA E, GUAN C T. A survey on explainable artificial intelligence (XAI): toward medical XAI[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(11): 4793-4813.

[10] GUO W S. Explainable artificial intelligence for 6G: improving trust between human and machine[J]. IEEE Communications Magazine, 2020, 58(6): 39-45.

[11] 蒲天骄，乔骥，赵紫璇，等.面向电力系统智能分析的机器学习可解释性方法研究(一)：基本概念与框架[J].中国电机工程学报，2023，43(18)：7010-7030.
PU Tianjiao, QIAO Ji, ZHAO Zixuan, et al. Research on interpretable methods of machine learning applied in intelligent analysis of power system (part Ⅰ): basic concept and framework[J]. Proceedings of the CSEE, 2023, 43(18): 7010-7030.

[12] ALSAIGH R, MEHMOOD R, KATIB I. AI explainability and governance in smart energy systems: a review [J]. Frontiers in Energy Research, 2023, 11: 1071291.

[13] ERSOZ B, SAGIROGLU S, BULBUL H I. A short review on explainable artificial intelligence in renewable energy and resources [C]// 2022 11th International Conference on Renewable Energy Research and Application (ICRERA), September 18-21, 2022, Istanbul, Turkey: 247-252.

[14] MACHLEV R, HEISTRENE L, PERL M, et al. Explainable artificial intelligence (XAI) techniques for energy and power systems: review, challenges and opportunities [J]. Energy and AI, 2022, 9: 100169.

[15] CHEN Z, XIAO F, GUO F Z, et al. Interpretable machine learning for building energy management: a state-of-the-art review[J]. Advances in Applied Energy, 2023, 9: 100123.

[16] 刘觉，谢汉中.人工智能，专家系统(三)[J].电力系统自动化，1988，12(2)：58-61.

LIU Jue, XIE Hanzhong. Introduction to artificial intelligence & expert system (part Ⅲ)[J]. Automation of Electric Power Systems, 1988, 12(2): 58-61.

[17] VAN LENT M, FISHER W, MANCUSO M. An explainable artificial intelligence system for small-unit tactical behavior [C]// Proceedings of the 16th Conference on Innovative Applications of Artifical Intelligence, July 25-29, 2004, San Jose, USA: 900-907.

[18] 郭创新，朱传柏，曹一家，等.电力系统故障诊断的研究现状与发展趋势[J].电力系统自动化，2006，30(8)：98-103.
GUO Chuangxin, ZHU Chuanbai, CAO Yijia, et al. State of arts of fault diagnosis of power systems [J]. Automation of Electric Power Systems, 2006, 30(8): 98-103.

[19] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[20] GUNNING D, AHA D W. DARPA's explainable artificial intelligence program[J]. AI Magazine, 2019, 40(2): 44-58.

[21] 张成洪，陈刚，陆天，等.可解释人工智能及其对管理的影响：研究现状和展望[J].管理科学，2021，34(3)：63-79.
ZHANG Chenghong, CHEN Gang, LU Tian, et al. Explainable artificial intelligence and its impact on management: research status and prospects [J]. Journal of Management Science, 2021, 34(3): 63-79.

[22] 国务院.关于印发新一代人工智能发展规划的通知[EB/OL].[2023-01-02].https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.
The State Council of the People's Republic of China. Circular on the issuance of the development plan for a new generation of artificial intelligence [EB/OL]. [2023-01-02]. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

[23] 国家新一代人工智能治理专业委员会.《新一代人工智能伦理规范》发布[EB/OL].[2023-01-02].https://www.safea.gov.cn/kjbgz/202109/t20210926_177063.html.
National Professional Committee on the Governance of Next Generation Artificial Intelligence. Code of ethics for the new generation of artificial intelligence released [EB/OL].[2023-01-02]. https://www.safea.gov.cn/kjbgz/202109/t20210926_177063.html.

[24] 阿里巴巴集团，中国信息通信研究院.人工智能治理与可持续发展实践白皮书[EB/OL].[2023-11-06].https://s.alibaba.com/cn/aaigWhitePaperDetails.
Alibaba Group, China Academy of Information and Communication Research. White paper on artificial intelligence governance and sustainable development practices [EB/OL].[2023-11-06]. https://s.alibaba.com/cn/aaigWhitePaperDetails.

[25] 中国信息通信研究院，京东探索研究院.可信人工智能白皮书[EB/OL].[2023-01-02].http://www.caict.ac.cn/kxyj/qwfb/bps/202107/P020210709319866413974.pdf.
China Academy of Information and Communication Research, Jingdong Discovery Institute. Trusted artificial intelligence white paper [EB/OL].[2023-01-02].http://www.caict.ac.cn/kxyj/qwfb/bps/202107/P020210709319866413974.pdf.

[26] 助力"新型电力系统"落地！国网山东电力打造可信AI负荷预测[EB/OL].[2023-03-28].http://www.xinhuanet.com/tech/20221215/9046d1e1fb3945a88164a3ec1676597b/c.html.
Helping the "new type of power system" come to fruition! State Grid Shandong Power builds credible AI load forecasting [EB/OL]. [2023-03-28]. http://www.xinhuanet.com/tech/20221215/9046d1e1fb3945a88164a3ec1676597b/c.html.

[27] VOSviewer. Visualizing scientific landscapes[EB/OL]. [2023-03-28]. https://www.vosviewer.com.

[28] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, USA: 4768-4777.

[29] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?": explaining the predictions of any classifier[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, 2016, San Francisco, USA: 1135-1144.

[30] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy: 618-626.

[31] QUINLAN J R. C4.5: programs for machine learning [M]. San Mateo, USA: Morgan Kaufmann Publishers, 1993.

[32] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[33] Xgboost developers. Python API reference—xgboost 2.0.1 documentation [EB/OL]. [2023-11-06]. https://xgboost.readthedocs.io/en/stable/python/python_api.html.

[34] 刘雁文，胡炎，邰能灵.基于决策树的智能变电站运维专家系统规则提取方法[J].电力科学与技术学报，2019，34(1)：123-128.
LIU Yanwen, HU Yan, TAI Nengling. Rule extraction method of operation and maintenance expert system for an intelligent substation based on the decision tree [J]. Journal of Electric Power Science and Technology, 2019, 34 (1): 123-128.

[35] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL].[2014-09-01]. https://arxiv.org/abs/1409.0473.pdf.

[36] 刘畅宇，王小君，尚博阳，等.基于域自适应迁移学习的有源配电网故障选线方法[J/OL].高电压技术：1-11[2023-11-05].https://doi.org/10.13336/j.1003-6520.hve.20230521.
LIU Changyu, WANG Xiaojun, SHANG Boyang, et al. Fault line selection for active distribution network based on domain adaptive transfer learning[J/OL]. High Voltage Engineering: 1-11 [2023-11-05]. https://doi.org/10.13336/j.1003-6520.hve.20230521.

[37] 李佳玮.基于图神经网络的配电网故障定位方法[D].北京：北京交通大学，2023.
LI Jiawei. Distribution network fault location based on graph neural network[D]. Beijing: Beijing Jiaotong University, 2023.

[38] 房佳姝，刘崇茹，苏晨博，等.基于自注意力Transformer编码器的多阶段电力系统暂态稳定评估方法[J].中国电机工程学报，2023，43(15)：5745-5759.

FANG Jiashu, LIU Chongru, SU Chenbo, et al. Multi-stage transient stability assessment of power system based on self-attention Transformer encoder[J]. Proceedings of the CSEE, 2023, 43(15): 5745-5759.

[39] 山衍浩.基于数据驱动的海上风机状态监测与故障诊断研究[D].上海:上海电力大学,2021.
SHAN Yanhao. Data-driven research on condition monitoring and fault diagnosis of offshore wind turbine[D]. Shanghai: Shanghai Electric Power University, 2021.

[40] 苏向敬,周汶鑫,李超杰,等.基于双重注意力LSTM神经网络的可解释海上风电出力预测[J].电力系统自动化,2022,46(7):141-151.
SU Xiangjing, ZHOU Wenxin, LI Chaojie, et al. Interpretable offshore wind power output forecasting based on long short-term memory neural network with dual-stage attention [J]. Automation of Electric Power Systems, 2022, 46(7): 141-151.

[41] KUHN H W. Classics in game theory[M]. Princeton, USA: Princeton University Press, 1997.

[42] 孙庆凯,王小君,张义志,等.基于LSTM和多任务学习的综合能源系统多元负荷预测[J].电力系统自动化,2021,45(5):63-70.
SUN Qingkai, WANG Xiaojun, ZHANG Yizhi, et al. Multiple load prediction of integrated energy system based on long short-term memory and multi-task learning[J]. Automation of Electric Power Systems, 2021, 45(5): 63-70.

[43] 周挺,杨军,詹祥澎,等.一种数据驱动的暂态电压稳定评估方法及其可解释性研究[J].电网技术,2021,45(11):4416-4425.
ZHOU Ting, YANG Jun, ZHAN Xiangpeng, et al. Data-driven method and interpretability analysis for transient voltage stability assessment[J]. Power System Technology, 2021, 45(11): 4416-4425.

[44] AN J, YU J C, LI Z H, et al. A data-driven method for transient stability margin prediction based on security region[J]. Journal of Modern Power Systems and Clean Energy, 2020, 8(6): 1060-1069.

[45] KIM M, JUN J A, SONG Y J, et al. Explanation for building energy prediction [C]// 2020 International Conference on Information and Communication Technology Convergence (ICTC), October 21-23, 2020, Jeju, South Korea.

[46] 付磊.基于级联式深度学习的电力线路安全性检测算法研究[D].成都:四川大学,2021.
FU Lei. Research on power line safety detection algorithm based on cascade deep learning [D]. Chengdu: Sichuan University, 2021.

[47] 何颖宣.基于多标签学习的螺栓多属性分类方法研究[D].北京:华北电力大学,2021.
HE Yingxuan. Research on multi-attribute classification method of bolts based on multi-label learning[D]. Beijing: North China Electric Power University, 2021.

[48] 黄杰.基于深度学习的绝缘子憎水性识别与故障检测方法研究[D].长沙:湖南大学,2021.
HUANG Jie. Research on insulator hydrophobic identification and fault detection method based on deep learning [D]. Changsha: Hunan University, 2021.

[49] 白玉莹.计及风资源时空特征的风电功率超短期预测研究[D].吉林:东北电力大学,2021.
BAI Yuying. Research on ultra-short-term wind power forecasting considering the temporal and spatial characteristics of wind resources[D]. Jilin: Northeast Electric Power University, 2021.

[50] 荆凯.基于权重无关神经网络的风机功率超短期预测算法研究[D].银川:宁夏大学,2020.
JING Kai. Research on ultra-short-term prediction algorithm of fan power based on weight-independent neural network [D]. Yinchuan: Ningxia University, 2020.

[51] 崔杨,王议坚,黄彦浩,等.基于多元注意力框架与引导式监督学习的闭环风电功率超短期预测策略[J].中国电机工程学报,2023,43(4):1334-1347.
CUI Yang, WANG Yijian, HUANG Yanhao, et al. Ultra-short-term forecasting strategy of closed-loop wind power based on multi-attention framework and guided supervised learning[J]. Proceedings of the CSEE, 2023, 43(4): 1334-1347.

[52] 余长青.基于人工智能方法的风电消纳能力预测及消纳措施优化[D].重庆:重庆大学,2020.
YU Changqing. Prediction of wind power absorptive capacity and optimization of absorptive measures based on artificial intelligence method [D]. Chongqing: Chongqing University, 2020.

[53] HEO S, KO J, KIM S, et al. Explainable AI-driven net-zero carbon roadmap for petrochemical industry considering stochastic scenarios of remotely sensed offshore wind energy [J]. Journal of Cleaner Production, 2022, 379: 134793.

[54] KUZLU M, CALI U, SHARMA V, et al. Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools[J]. IEEE Access, 2020, 8: 187814-187823.

[55] CHANG X M, LI W, MA J, et al. Interpretable machine learning in sustainable edge computing: a case study of short-term photovoltaic power output prediction[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 4-8, 2020, Barcelona, Spain.

[56] SARP S, KUZLU M, CALI U, et al. An interpretable solar photovoltaic power generation forecasting approach using an explainable artificial intelligence tool[C]// 2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), February 16-18, 2021, Washington, USA.

[57] BAHANI K, ALI-OU-SALAH H, MOUJABBIR M, et al. A novel interpretable model for solar radiation prediction based on adaptive fuzzy clustering and linguistic hedges [C]// Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications, September 23-24, 2020, Rabat, Morocco: 1-6.

[58] GAIER A, HA D. Weight agnostic neural networks[EB/OL]. [2019-06-11]. https://arxiv.org/abs/1906.04358.pdf.

[59] 苏向敬,聂良钊,李超杰,等.基于MSTAGNN模型的可解释海上风电场多风机出力预测[J].电力系统自动化,2023,47(9):88-98.
SU Xiangjing, NIE Liangzhao, LI Chaojie, et al. Interpretable power output prediction of multiple wind turbines for offshore

wind farm based on multiple spatio-temporal attention graph neural network model [J]. Automation of Electric Power Systems, 2023, 47(9): 88-98.

[60] ELI5. A library for debugging_inspecting machine learning classifiers and explaining their predictions [EB/OL]. [2023-11-06]. https://github.com/eli5-org/eli5.

[61] AZEMENA H J, AYADI A, SAMET A. Explainable artificial intelligent as a solution approach to the duck curve problem [J]. Procedia Computer Science, 2022, 207: 2747-2756.

[62] 武宇翔，韩肖清，牛哲文，等.融合多注意力深度神经网络的可解释光伏功率区间预测[J/OL].电网技术：1-19[2023-10-09]. https://doi.org/10.13335/j.1000-3673.pst.2023.0978.
WU Yuxiang, HAN Xiaoqing, NIU Zhewen, et al. Interpretable photovoltaic power interval prediction using multi-attention deep neural networks [J/OL]. Power System Technology: 1-19 [2023-10-09]. https://doi. org/10.13335/j.1000-3673.pst.2023.0978.

[63] 谷云东，刘浩.基于最优特征组合改进极限梯度提升的负荷预测[J].计算机应用研究，2021，38(9)：2767-2772.
GU Yundong, LIU Hao. Load forecasting based on optimal feature combination improved XGBoost [J]. Application Research of Computers, 2021, 38(9): 2767-2772.

[64] GAO Y, RUAN Y J. Interpretable deep learning model for building energy consumption prediction based on attention mechanism[J]. Energy and Buildings, 2021, 252: 111379.

[65] 逄宝中，李庚银，武昭原，等.基于循环扩张机制的ConvGRU-Transformer短期电力负荷预测方法[J].华北电力大学学报(自然科学版)，2022，49(3)：34-43.
TI Baozhong, LI Gengyin, WU Zhaoyuan, et al. A short-term load forecasting method based on recurrent and dilated mechanism of ConvGRU-Transformer [J]. Journal of North China Electric Power University (Natural Science Edition), 2022, 49(3): 34-43.

[66] 吕忠麟，顾洁，孟璐.基于耦合特征与多任务学习的综合能源系统短期负荷预测[J].电力系统自动化，2022，46(11)：58-66.
LYU Zhonglin, GU Jie, MENG Lu. Short-term load forecasting for integrated energy system based on coupling features and multi-task learning [J]. Automation of Electric Power Systems, 2022, 46(11): 58-66.

[67] CHUNG W J, LIU C D. Analysis of input parameters for deep learning-based load prediction for office buildings in different climate zones using explainable artificial intelligence[J]. Energy and Buildings, 2022, 276: 112521.

[68] 章超波，刘永政，李宏波，等.基于加权残差聚类的建筑负荷预测区间估计[J].浙江大学学报(工学版)，2022，56(5)：930-937.
ZHANG Chaobo, LIU Yongzheng, LI Hongbo, et al. Weighted residual clustering-based building load prediction interval estimation [J]. Journal of Zhejiang University (Engineering Science), 2022, 56(5): 930-937.

[69] CHAKRABORTY D, ALAM A, CHAUDHURI S, et al. Scenario-based prediction of climate change impacts on building cooling energy consumption with explainable artificial intelligence[J]. Applied Energy, 2021, 291: 116807.

[70] LI A, XIAO F, ZHANG C, et al. Attention-based interpretable neural network for building cooling load prediction [J]. Applied Energy, 2021, 299: 117238.

[71] MOON J, RHO S, BAIK S W. Toward explainable electrical load forecasting of buildings: a comparative study of tree-based ensemble methods with Shapley values[J]. Sustainable Energy Technologies and Assessments, 2022, 54: 102888.

[72] 张桢豪.新冠肺炎疫情下电力系统短期负荷预测模型及方法研究[D].南宁：广西大学，2022.
ZHANG Zhenhao. Study on short-term load forecasting model and method of power system under COVID-19 epidemic situation[D]. Nanning: Guangxi University, 2022.

[73] BOLSTAD D A, CALI U, KUZLU M, et al. Day-ahead load forecasting using explainable artificial intelligence [C]// 2022 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), April 24-28, 2022, New Orleans, USA.

[74] TIAN W. A review of sensitivity analysis methods in building energy analysis [J]. Renewable and Sustainable Energy Reviews, 2013, 20: 411-419.

[75] SHAMS AMIRI S, MOTTAHEDI S, LEE E R, et al. Peeking inside the black-box: explainable machine learning applied to household transportation energy consumption [J]. Computers, Environment and Urban Systems, 2021, 88: 101647.

[76] GHAHRAMANI Z. An introduction to hidden Markov models and Bayesian networks [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2001, 15(1): 9-42.

[77] WOOD-DOUGHTY Z, CACHOLA I, DREDZE M. Proxy model explanations for time series RNNs [C]// 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), December 13-16, 2021, Pasadena, USA.

[78] ALOVA G, TROTTER P A, MONEY A. A machine-learning approach to predicting Africa's electricity mix based on planned power plants and their chances of success [J]. Nature Energy, 2021, 6(2): 158-166.

[79] SACHIT M S, SHAFRI H Z M, ABDULLAH A F, et al. Global spatial suitability mapping of wind and solar systems using an explainable AI-based approach [J]. ISPRS International Journal of Geo-Information, 2022, 11(8): 422.

[80] UTAMA C, MESKE C, SCHNEIDER J, et al. Reactive power control in photovoltaic systems through (explainable) artificial intelligence[J]. Applied Energy, 2022, 328: 120004.

[81] MACHLEV R, PERL M, BELIKOV J, et al. Measuring explainability and trustworthiness of power quality disturbances classifiers using XAI—explainable artificial intelligence [J]. IEEE Transactions on Industrial Informatics, 2022, 18(8): 5127-5137.

[82] ZHANG K, ZHANG J, XU P D, et al. Explainable AI in deep reinforcement learning models for power system emergency control [J]. IEEE Transactions on Computational Social Systems, 2022, 9(2): 419-427.

[83] DAI X L, CHENG S Y, CHONG A. Deciphering optimal mixed-mode ventilation in the tropics using reinforcement learning with explainable artificial intelligence [J]. Energy and

Buildings, 2023, 278: 112629.

[84] ZHANG K, ZHANG J, XU P D, et al. A multi-hierarchical interpretable method for DRL-based dispatching control in power systems[J]. International Journal of Electrical Power & Energy Systems, 2023, 152: 109240.

[85] MILANI S, TOPIN N, VELOSO M, et al. A survey of explainable reinforcement learning [EB/OL]. [2022-02-17]. https://arxiv.org/abs/2202.08434.pdf.

[86] 王甜婧,汤涌,郭强,等.基于知识经验和深度强化学习的大电网潮流计算收敛自动调整方法[J].中国电机工程学报,2020,40(8):2396-2406.
WANG Tianjing, TANG Yong, GUO Qiang, et al. Automatic adjustment method of power flow calculation convergence for large-scale power grid based on knowledge experience and deep reinforcement learning[J]. Proceedings of the CSEE, 2020, 40(8): 2396-2406.

[87] 李鹏,黄文琦,王鑫,等.数据与知识联合驱动的人工智能方法在电力调度中的应用综述[J/OL].电力系统自动化:1-16[2023-06-30]. http://kns. cnki. net/kcms/detail/32.1180. TP. 20230608.1043.002.html.
LI Peng, HUANG Wenqi, WANG Xin, et al. Review on application of combined data-knowledge-driven artificial intelligence methods in power dispatching[J/OL]. Automation of Electric Power Systems: 1-16[2023-06-30]. http://kns.cnki.net/kcms/detail/32.1180.TP.20230608.1043.002.html.

[88] 刘潇,刘书洋,庄韫恺,等.强化学习可解释性基础问题探索和方法综述[J].软件学报,2023,34(5):2300-2316.
LIU Xiao, LIU Shuyang, ZHUANG Yunkai, et al. Explainable reinforcement learning: basic problems exploration and method survey[J]. Journal of Software, 2023, 34(5): 2300-2316.

[89] VOUROS G A. Explainable deep reinforcement learning: state of the art and challenges[J]. ACM Computing Surveys, 2022, 55(5): 92.

[90] HEUILLET A, COUTHOUIS F, DÍAZ-RODRÍGUEZ N. Explainability in deep reinforcement learning[J]. Knowledge-Based Systems, 2021, 214: 106685.

[91] MOTT A, ZORAN D, CHRZANOWSKI M, et al. Towards interpretable reinforcement learning using attention augmented agents [C]// Thirty-third Annual Conference on Neural Information Processing Systems, December 8-14, 2019, Vancouver, Canada.

[92] WU J D, HUANG Z Y, HU Z X, et al. Toward human-in-the-loop AI: enhancing deep reinforcement learning via real-time human guidance for autonomous driving[J]. Engineering, 2023, 21: 75-91.

[93] 杨博,陈义军,姚伟,等.基于新一代人工智能技术的电力系统稳定评估与决策综述[J].电力系统自动化,2022,46(22):200-223.
YANG Bo, CHEN Yijun, YAO Wei, et al. Review on stability assessment and decision for power systems based on new-generation artificial intelligence technology[J]. Automation of Electric Power Systems, 2022, 46(22): 200-223.

[94] 高晗,蔡国伟,杨德友,等.基于累积贡献率和可解释人工智能的静态电压稳定裕度估计特征量筛选方法[J].电力自动化设备,2023,43(4):168-176.
GAO Han, CAI Guowei, YANG Deyou, et al. Feature selection approach based on FCC-eAI in static voltage stability margin estimation[J]. Electric Power Automation Equipment, 2023, 43(4): 168-176.

[95] KRUSE J, SCHÄFER B, WITTHAUT D. Revealing drivers and risks for power grid frequency stability with explainable AI [J]. Patterns, 2021, 2(11): 100365.

[96] HAN T S, CHEN J F, WANG L, et al. Interpretation of stability assessment machine learning models based on shapley value[C]// 2019 IEEE 3rd Conference on Energy Internet and Energy System Integration (EI^2), November 8-10, 2019, Changsha, China.

[97] 韩天森,陈金富,李银红,等.电力系统稳定评估机器学习可解释代理模型研究[J].中国电机工程学报,2020,40(13):4122-4131.
HAN Tiansen, CHEN Jinfu, LI Yinhong, et al. Study on interpretable surrogate model for power system stability evaluation machine learning[J]. Proceedings of the CSEE, 2020, 40(13): 4122-4131.

[98] WU S A, ZHENG L, HU W, et al. Improved deep belief network and model interpretation method for power system transient stability assessment[J]. Journal of Modern Power Systems and Clean Energy, 2020, 8(1): 27-37.

[99] CHEN M H, LIU Q Y, CHEN S H, et al. XGBoost-based algorithm interpretation and application on post-fault transient stability status prediction of power system[J]. IEEE Access, 2019, 7: 13149-13158.

[100] HAN X X, JIN Y L, WU G, et al. A self-attention-embedded deep learning model for phasor measurement unit-based post-fault transient stability prediction[C]// 2022 Asian Conference on Frontiers of Power and Energy (ACFPE), October 21-23, 2022, Chengdu, China.

[101] 贾宏阳,侯庆春,刘羽霄,等.基于斜回归树及其集成算法的静态电压稳定规则提取[J].电力系统自动化,2022,46(1):51-59.
JIA Hongyang, HOU Qingchun, LIU Yuxiao, et al. Extraction of static voltage stability rule based on oblique regression tree and its ensemble algorithm[J]. Automation of Electric Power Systems, 2022, 46(1): 51-59.

[102] CREMER J L, KONSTANTELOS I, STRBAC G. From optimization-based machine learning to interpretable security rules for operation [J]. IEEE Transactions on Power Systems, 2019, 34(5): 3826-3836.

[103] REN C, XU Y, ZHANG R. An interpretable deep learning method for power system transient stability assessment via tree regularization[J]. IEEE Transactions on Power Systems, 2022, 37(5): 3359-3369.

[104] 赵恺,石立宝.基于改进一维卷积神经网络的电力系统暂态稳定评估[J].电网技术,2021,45(8):2945-2957.
ZHAO Kai, SHI Libao. Transient stability assessment of power system based on improved one-dimensional convolutional neural network[J]. Power System Technology, 2021, 45(8): 2945-2957.

[105] 陈明华,刘群英,张家枢,等.基于XGBoost的电力系统暂态

稳定预测方法[J].电网技术,2020,44(3):1026-1034.
CHEN Minghua, LIU Qunying, ZHANG Jiashu, et al. XGBoost-based algorithm for post-fault transient stability status prediction[J]. Power System Technology, 2020, 44(3): 1026-1034.

[106] 陈明华.电力系统暂态稳定性智能评估方法研究[D].成都:电子科技大学,2019.
CHEN Minghua. Research on intelligent evaluation method of power system transient stability[D]. Chengdu: University of Electronic Science and Technology of China, 2019.

[107] 胡润滋,马晓忱,孙博,等.基于特征选择的暂态安全评估方法及其可解释性研究[J].电网技术,2023,47(2):755-763.
HU Runzi, MA Xiaochen, SUN Bo, et al. Transient security assessment method based on feature selection and its interpretive approach[J]. Power System Technology, 2023, 47(2): 755-763.

[108] 王强,陈浩,刘炼.基于自然梯度提升的静态电压稳定裕度预测及其影响因素分析[J].电力系统及其自动化学报,2022,34(9):130-137.
WANG Qiang, CHEN Hao, LIU Lian. Prediction of static voltage stability margin based on natural gradient lifting and analysis of its influencing factors[J]. Proceedings of the CSU-EPSA, 2022, 34(9): 130-137.

[109] 朱春霖,余成波.基于LightGBM算法的光伏并网系统孤岛检测及其集成的可解释研究[J].电力自动化设备,2023,43(7):80-86.
ZHU Chunlin, YU Chengbo. Islanding detection of grid-connected photovoltaic system based on LightGBM algorithm and its integrated interpretability analysis[J]. Electric Power Automation Equipment, 2023, 43(7): 80-86.

[110] APLEY D W, ZHU J Y. Visualizing the effects of predictor variables in black box supervised learning models[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2020, 82(4): 1059-1086.

[111] 向玮华,班连庚,周佩朋.基于机器学习可解释代理模型的风电次同步振荡在线预测及优化控制方法[J].电力系统保护与控制,2021,49(16):67-75.
XIANG Weihua, BAN Liangeng, ZHOU Peipeng. Online prediction and optimal control method for subsynchronous oscillation of wind power based on an interpretable surrogate model for machine learning[J]. Power System Protection and Control, 2021, 49(16): 67-75.

[112] YUAN H, YU H Y, GUI S R, et al. Explainability in graph neural networks: a taxonomic survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022: 1-19.

[113] AGARWAL C, QUEEN O, LAKKARAJU H, et al. Evaluating explainability for graph neural networks [J]. Scientific Data, 2023, 10: 144.

[114] LI T, DENG J L, SHEN Y Y, et al. Towards fine-grained explainability for heterogeneous graph neural network [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(7): 8640-8647.

[115] MIKA G P, BOUZEGHOUB A, WEGRZYN-WOLSKA K, et al. HGExplainer: explainable heterogeneous graph neural network[EB/OL]. [2023-11-06]. https://hal.science/hal-04220962.

[116] CHANG S Y, HAN W, TANG J L, et al. Heterogeneous network embedding via deep architectures[C]// Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 10-13, 2015, Sydney, Australia: 119-128.

[117] ZHAO Y, LIU Z, SUN M. Representation learning for measuring entity relatedness with rich information [C]// Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, July 25-31, 2015, Buenos Aires, Argentina.

[118] 刘志远,于晓军,罗美玲,等.基于CBAM-FCN的高压输电线路发展性故障识别方法[J].电网与清洁能源,2022,38(9):25-33.
LIU Zhiyuan, YU Xiaojun, LUO Meiling, et al. Developmental fault identification method of high voltage transmission line based on CBAM-FCN[J]. Power System and Clean Energy, 2022, 38(9): 25-33.

[119] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[M]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[120] SANTOS O L D, DOTTA D, WANG M, et al. Performance analysis of a DNN classifier for power system events using an interpretability method [J]. International Journal of Electrical Power & Energy Systems, 2022, 136: 107594.

[121] RAISSI M, PERDIKARIS P, KARNIADAKIS G E. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations [J]. Journal of Computational Physics, 2019, 378: 686-707.

[122] SAIRAM S, SESHADHRI S, MARAFIOTI G, et al. Edge-based explainable fault detection systems for photovoltaic panels on edge nodes[J]. Renewable Energy, 2022, 185: 1425-1440.

[123] UTAMA C, MESKE C, SCHNEIDER J, et al. Explainable artificial intelligence for photovoltaic fault detection: a comparison of instruments [J]. Solar Energy, 2023, 249: 139-151.

[124] RIBEIRO M T, SINGH S, GUESTRIN C. Anchors: high-precision model-agnostic explanations[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 1527-1535.

[125] MOTHILAL R K, SHARMA A, TAN C H. Explaining machine learning classifiers through diverse counterfactual explanations[C]// Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 27-30, 2020, Barcelona, Spain: 607-617.

[126] 苏向敬,山衍浩,周汶鑫,等.基于GRU和注意力机制的海上风机齿轮箱状态监测[J].电力系统保护与控制,2021,49(24):141-149.
SU Xiangjing, SHAN Yanhao, ZHOU Wenxin, et al. GRU and attention mechanism-based condition monitoring of an offshore wind turbine gearbox[J]. Power System Protection and Control, 2021, 49(24): 141-149.

[127] GOLIZADEH AKHLAGHI Y, ASLANSEFAT K, ZHAO X D, et al. Hourly performance forecast of a dew point cooler using explainable artificial intelligence and evolutionary optimisations by 2050 [J]. Applied Energy, 2021, 281: 116062.

[128] ZAHHARI S, LE J P, GROSSIN B, et al. DataPoste: a mobile application based on Understandable Deep Learning to characterise electrical equipment in secondary distribution substation [C]// CIRED 2021-The 26th International Conference and Exhibition on Electricity Distribution, September 20-23, 2021.

[129] WANG H F, WANG Z F, DU M N, et al. Score-CAM: score-weighted visual explanations for convolutional neural networks [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 14-19, 2020, Seattle, USA.

[130] 许格健.电池储能系统典型单元运行状态参数预测方法研究[D].吉林:东北电力大学,2020.
XU Gejian. Study on prediction method of operating state parameters of typical cells in battery energy storage system[D]. Jilin: Northeast Electric Power University, 2020.

[131] 滕曦.基于Stacking融合模型的新能源汽车动力电池SOC预测研究[D].重庆:重庆工商大学,2022.
TENG Xi. Research on SOC prediction of power battery of new energy vehicle based on Stacking fusion model [D]. Chongqing: Chongqing Technology and Business University, 2022.

[132] JOSHI G, WALAMBE R, KOTECHA K. A review on explainability in multimodal deep neural nets [J]. IEEE Access, 2021, 9: 59800-59821.

[133] PEARL J. Causality [M]. Cambridge, UK: Cambridge University Press, 2009.

[134] PEARL J, MACKENZIE D. The book of why: the new science of cause and effect [M]. New York, USA: Basic Books, 2018.

[135] SCHWARZ G. Estimating the dimension of a model[J]. The Annals of Statistics, 1978, 6(2): 461-464.

[136] BELTRÁN S, CASTRO A, IRIZAR I, et al. Framework for collaborative intelligence in forecasting day-ahead electricity price[J]. Applied Energy, 2022, 306: 118049.

[137] 刘慧鑫,沈晓东,魏泽涛,等.基于校准窗口集成与耦合市场特征的可解释双层日前电价预测[J/OL].中国电机工程学报:1-14 [2023-04-18]. http://kns.cnki.net/kcms/detail/11.2107.TM.20221107.1537.010.html.
LIU Huixin, SHEN Xiaodong, WEI Zetao, et al. Interpretable two-layer day-ahead electricity price forecast based on calibration window combination and coupled market characteristics [J]. Proceedings of the CSEE: 1-14 [2023-04-18]. http://kns.cnki.net/kcms/detail/11.2107.TM.20221107.1537.010.html.

[138] MA H N, MCAREAVEY K, MCCONVILLE R, et al. Explainable AI for non-experts: energy tariff forecasting[C]// 2022 27th International Conference on Automation and Computing (ICAC), September 1-3, 2022, Bristol, UK.

[139] 吴洋,辛茹,邹文滔,等.提升电力现货市场出清结果可解释性的综合分析方法[J].南方电网技术,2022,16(6):113-123.
WU Yang, XIN Ru, ZOU Wentao, et al. Comprehensive analysis method for enhancing the explainability of electricity spot market clearing results [J]. Southern Power System Technology, 2022, 16(6): 113-123.

[140] LIU Z C, ZHU Y Y, WANG Z, et al. MIXRTs: toward interpretable multi-agent reinforcement learning via mixing recurrent soft decision trees[EB/OL]. [2022-09-15]. https://arxiv.org/abs/2209.07225.pdf.

[141] MILANI S, ZHANG Z C, TOPIN N, et al. MAVIPER: learning decision tree policies for Interpretable multi-agent reinforcement learning [M]// Machine Learning and Knowledge Discovery in Databases. Cham: Springer Nature Switzerland, 2023: 251-266.

[142] MOTOKAWA Y, SUGAWARA T. MAT-DQN: toward interpretable multi-agent deep reinforcement learning for coordinated activities [C]// Proceedings of Artificial Neural Networks and Machine Learning-ICANN 2021: 30th International Conference on Artificial Neural Networks, September 14-17, 2021, Bratislava, Slovakia: 556-567.

[143] WANG G. Interpret federated learning with Shapley values [EB/OL]. [2023-05-11]. https://arxiv.org/abs/1905.04519.pdf.

[144] 纪守领,李进锋,杜天宇,等.机器学习模型可解释性方法、应用与安全研究综述[J].计算机研究与发展,2019,56(10):2071-2096.
JI Shouling, LI Jinfeng, DU Tianyu, et al. Survey on techniques, applications and security of machine learning interpretability [J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096.

[145] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead [J]. Nature Machine Intelligence, 2019, 1(5): 206-215.

[146] ZHA D C, BHAT Z P, LAI K H, et al. Data-centric artificial intelligence: a survey [EB/OL]. [2023-03-17]. https://arxiv.org/abs/2303.10158.pdf.

[147] 薛禹胜,赖业宁.大能源思维与大数据思维的融合:(一)大数据与电力大数据[J].电力系统自动化,2016,40(1):1-8.
XUE Yusheng, LAI Yening. Integration of macro energy thinking and big data thinking: Part one big data and power big data[J]. Automation of Electric Power Systems, 2016, 40(1): 1-8.

[148] 李峰,王琦,胡健雄,等.数据与知识联合驱动方法研究进展及其在电力系统中应用展望[J].中国电机工程学报,2021,41(13):4377-4390.
LI Feng, WANG Qi, HU Jianxiong, et al. Combined data-driven and knowledge-driven methodology research advances and its applied prospect in power systems [J]. Proceedings of the CSEE, 2021, 41(13): 4377-4390.

[149] 王琦,李峰,汤奕,等.基于物理-数据融合模型的电网暂态频率特征在线预测方法[J].电力系统自动化,2018,42(19):1-9.
WANG Qi, LI Feng, TANG Yi, et al. On-line prediction method of transient frequency characteristics for power grid

based on physical-statistical model[J]. Automation of Electric Power Systems, 2018, 42(19): 1-9.

王小君(1978—),男,博士生导师,主要研究方向:电力系统分析与控制、综合能源系统优化运行。E-mail:xjwang1@bjtu.edu.cn

窦嘉铭(1996—),男,通信作者,博士研究生,主要研究方向:可解释人工智能、强化学习在能源优化调度中的应用。E-mail:djmmjddjmmjd@163.com

刘　嬰(1991—),男,博士,主要研究方向:电力系统稳定性分析与控制、综合能源系统优化。E-mail:liuzhao1@bjtu.edu.cn

(编辑　孔丽蓓)

Review and Prospect of Explainable Artificial Intelligence and Its Application in Power Systems

WANG Xiaojun[1], *DOU Jiaming*[1], *LIU Zhao*[1], *LIU Changyu*[1], *PU Tianjiao*[2], *HE Jinghan*[1]

(1. School of Electrical Engineering, Beijing Jiaotong University, Beijing 100044, China;
2. China Electric Power Research Institute, Beijing 100192, China)

Abstract: Explainable artificial intelligence (XAI), as a new type of artificial intelligence (AI) technologies, can present the logic of the AI process, reveal the AI black-box knowledge, and improve the credibility of the AI results. The deep coupling between XAI and power systems may accelerate the AI technology application in the power system and assist with the safety and stability of human-machine interaction. Therefore, this paper reviews the historical context, development needs, and hot technologies of XAI in the power system, summarizes its applications in source-load forecasting, operation control, fault diagnosis, and electricity market, and explores the application prospects of XAI in the power system around aspects such as interpretability, iterative framework, and number-matrix fusion. This paper aims to provide theoretical references and practical ideas for promoting the intelligent transformation and iterative human-machine interaction of the power system.

This work is supported by National Natural Science Foundation of China (No. 52377071, No. 52107068).

Key words: power system; artificial intelligence; explainability; machine learning