第49卷 第4期 2025年2月25日
DOI：10.7500/AEPS20240508001
电力系统自动化
Automation of Electric Power Systems
Vol. 49 No. 4 Feb. 25, 2025

电网稳定智能判别模型可信度指标体系构建与综合评价方法

刘慧玉[1]，王渝红[2]，石 访[1]，周 旭[2]，李保罗[1,3]，姬凯旋[3]
（1. 电网智能化调度与控制教育部重点实验室（山东大学），山东省济南市 250061；
2. 四川大学电气工程学院，四川省成都市 610065；3. 中国电力科学研究院有限公司，北京市 100192）

摘要：人工智能算法的不可信问题阻碍了其在电网稳定性分析及控制等场景中的实际应用。目前，尚无特定适用于电网稳定智能判别模型的可信度评价指标。文中针对电力行业特点开展人工智能稳定判别模型的可信度评价，选取准确性、复杂度、鲁棒性、可迁移性、可解释性共5个分项指标构建了电网稳定智能判别可信度评价指标体系，并总结了各分项指标的具体计算方法。同时，引入模糊层次分析法，结合主客观评价，确定各分项指标的权重并计算综合可信度指标。最后，以电压支撑强度、宽频带振荡以及频率稳定判别模型为例进行了可信度评价和分析，结果证明了所提出方法的有效性。

关键词：人工智能；电网稳定性；可信度；可迁移性；可解释性；模糊层次分析法

0 引言

随着电网规模逐渐扩大和新能源的广泛接入，时域仿真等传统的稳定评估方法在实际应用中面临着建模不准、运行方式失配、计算耗时长等问题，电网稳定性判别也面临着新的挑战。基于人工智能（artificial intelligence，AI）的电网稳定性判别方法，因其强大的数据处理能力和学习能力得到广泛的关注[1-2]。随着AI技术的进步，其在稳定性判别中的应用不再局限于传统的支持向量机等机器学习模型，以神经网络为代表的深度学习模型已成为研究热点，并成为高效、准确评估新型电力系统稳定性的潜在可行途径。

目前，AI已被广泛应用于图像识别、医疗等领域以及电力系统的各个方面，包括负荷预测、故障诊断、设备缺陷识别、电价预测等[3]，但在实际应用中仍存在诸多风险和不足。例如，在图像识别领域，模型对训练过程中的输入数据非常敏感。文献[4]通过实验证明，增加输入干扰会导致深度神经模型在图像识别中的判断失误，即训练数据中的噪声会干扰模型的判断。在自动驾驶领域，曾因AI未正确识别道路清扫车而造成人员伤亡事故[5]。在电力领域，2015年乌克兰电力系统因遭受黑客的网络攻击引发了长时间的停电事故[6]，表明完全依赖于信息系统存在安全风险。此外，AI模型的黑箱特性使得基于数据驱动的电网故障诊断结果只能作为参考，辅助电网运行人员做出最终判断[3]。电网稳定性判别中，模型的漏判、误判可能导致严重的停电事故，且缺乏可解释性，调度人员难以信任模型的判别结果[7]。同时，深度神经网络等深度学习方法在电力系统的应用中存在鲁棒性差等缺陷[8]。综上，由于在可解释性、鲁棒性等方面存在局限性，AI模型在电网稳定性判别等对安全性要求较高的实际场景中并未得到广泛应用。评价AI模型的可信度并确保模型的安全可信，对模型的实际应用推广具有重要的现实意义。

目前，AI领域可信度方面的研究已取得了一定的成果。美国国家标准技术研究院发布的《人工智能风险管理框架》总结了可信赖AI系统的特点[9]。文献[10]梳理了现有研究工作以及可信属性，包括鲁棒性、公平性、可解释性、隐私性、安全性、防危性、可追责性、透明性、普惠性等，认为AI的可信属性应包含可靠性、鲁棒性、公平性、隐私保护等9个方面。2019年，欧盟发布了《可信AI伦理指南》，从可信AI的根基、实现以及评估3个层面介绍了可信AI的伦理框架[11]。上述研究提出了可信AI的特点或属性，但未能充分考虑AI在电网稳定性判别应用中的专业需求与特性。

在电力行业，AI算法在实际应用中的局限性同样备受关注。针对电网数据不可信问题，文献[12]

收稿日期：2024-05-08；修回日期：2024-08-29。
上网日期：2024-09-25。
国家重点研发计划资助项目（2021YFB2400800）；国家电网公司科技项目（SGSDDK00WJJS2200092）。

构建了层次化、动态化的分析模型，以评估电力数据的可信度。针对AI模型在电力系统应用中存在训练样本获取困难、鲁棒性差等问题，文献[13]提出一种电力知识经验与机器学习相结合的引导学习机制，以提升模型的泛化性、鲁棒性等指标。针对可解释性，文献[14]梳理了可解释AI（explainable artificial intelligence，XAI）在电力系统中的应用和局限性，并提出了电力系统XAI未来的研究方向。文献[15]分析了电力行业AI存在的数据风险、技术风险等，并提出了相应的防护措施。文献[16]展望了大语言模型在电力系统中的潜在应用，讨论了所面临的数据管理、可解释性、可靠性等问题，并提出了相应的解决方案。文献[17-18]指出了电力系统AI的可信伦理及进化迁移瓶颈，提出了数据与知识的融合驱动机制以及模型进化趋优机制等解决方法。

以上研究指出了AI在电力系统应用中面临的挑战和局限性，如可解释性、鲁棒性等，并提出了相应的解决措施以及未来的研究方向。例如，文献[12]提出了可信性度量模型，但侧重于评价电网数据而非AI模型；其他文献从方法层面提出了可解释性、可迁移性等特定评价指标的改进方法。目前，针对电网AI稳定判别模型可信度评价方面的研究较匮乏，尚未形成一套公认的、适应于电网稳定智能判别应用场景的可信度指标体系和综合评价方法。文献[10]从软件可信度量的角度，提出了基于属性的AI系统可信度量评估框架，但并非专门针对电力系统稳定性判别这一特定应用领域。

本文旨在提出适用于电网稳定智能判别的可信度评价指标体系和综合评价方法。针对电力系统稳定性评估AI模型，考虑AI模型特点和电力系统的实际应用需求，从准确性、复杂度、鲁棒性、可迁移性、可解释性5个方面构建可信度评价指标体系，给出各项指标的具体计算方法；提出基于模糊层次分析的综合可信度评价方法，通过问卷调查收集百余名行业专家关于各分项指标权重以及相对重要程度的反馈意见，基于调查结果确定各分项指标权重并计算综合可信度指标；对电压支撑强度、宽频带振荡以及频率稳定判别AI模型分别进行了可信度评价和对比分析，结果证明了所提出可信度评价方法的有效性。

1　可信度评价指标体系构建

1.1　电网稳定性判别可信度分项指标选取

电力系统的安全稳定运行与人民生产、生活密切相关，AI模型的安全可信是其在电力领域应用的前提。然而，AI在电网稳定性判别的应用中还存在以下问题。

1）鲁棒性风险

在电网数据采集和传输过程中，传感器精度不足、通信线路不稳定、电磁干扰等多种因素可能会造成量测误差、数据丢失等数据质量问题[19]。由于训练集数据的局限性以及模型自身的复杂性，AI模型根据训练集产生的决策边界与真实的决策边界无法完全重合，使得模型对特定输入数据异常敏感，与训练样本差异较大的样本或存在微小扰动的对抗样本可能导致模型发生误判，进而导致对电网失稳状态的错误判断，增加停电事故的风险。

2）迁移局限性

AI稳定判别模型的训练需要完备的样本集，而样本集的构建需要花费大量的时间和精力。实际运行中，电网通常处于稳定运行状态，发生失稳的概率极低，实际故障数据难以获取，且电力系统的时变性导致历史运行数据适用性下降[2]。因此，目前的AI稳定判别模型训练主要基于仿真数据，应用仿真软件根据研究需求设置各种故障场景和运行方式，以生成完备的样本集。

AI稳定判别模型的训练和测试用样本集，通常基于特定的运行条件和故障场景，难以完全覆盖所有场景。当系统的拓扑结构、运行方式等发生变化时，新场景下的数据特征和分布与原始样本之间可能存在较大差异，导致已有模型的准确性等评估能力大幅下降。而在新场景下，重新训练模型需再次获取样本集、数据处理、模型构建和优化等步骤，花费大量的时间和精力。

3）可解释性风险

AI稳定判别模型的目标是学习和建立输入特征和电力系统稳定性指标之间的映射关系。为保证信息的充分性，通常选取高维的输入特征作为模型的输入，且模型的训练涉及大量的参数计算，相较于传统的浅层神经网络，深度学习模型强大的非线性数据处理能力使其在稳定性判别方面具有显著优势。但是，深度学习模型的本质是黑箱模型，其输入特征与稳定性指标的映射关系缺乏可解释性，使得在实际应用过程中，业务人员无法得知模型的决策依据，故也无法完全信任模型的判断结果。例如，对于给定样本的输入，稳定判别模型做出“失稳”的判断，但其缺乏物理机理、逻辑推理以及因果解释，业务人员无法理解模型的判断结果。AI模型的不可解释性可能导致潜在的安全风险，阻碍了其在电网稳定性判别、电力调度等场景中的应用推广。

AI模型的鲁棒性风险、迁移局限性以及可解释性风险阻碍了AI稳定判别模型的应用推广，迫切需

要构建AI稳定判别模型的可信度评价指标体系，对模型性能进行系统性评价并筛选出最具应用潜力的模型。结合现有的通用AI可信属性，即鲁棒性、公平性、可解释性、隐私性等[10]，以及AI在稳定性判别应用中存在的鲁棒性风险、迁移局限性以及可解释性风险，电网AI稳定判别可信度属性应包括准确性、复杂度、鲁棒性、可迁移性以及可解释性5个属性。

电网稳定性判别通常可归结为分类问题或回归问题，而AI模型的准确性对于评估模型性能、指导模型改进以及提升用户信任度具有重要意义。电力系统因其对稳定性有着极高的要求，必须对潜在的失稳场景进行精准判断，从而确保在必要时能够迅速采取控制措施，防止停电事故的发生。鉴于AI稳定判别模型在决策中的关键作用，任何误判都可能引发严重后果，首先需保证模型预测的准确性。

深度神经网络等AI模型参数众多、计算量大，考虑到模型的执行效率和在线应用需求，模型结构不应太复杂。对于稳定性判别而言，在系统失稳时，模型的判断时间不能过长，应及时反馈预测结果从而为安全稳定装置提供足够的反应时间。此外，适当的模型复杂度有助于提高模型性能，可在不过拟合的前提下深度挖掘数据的内在规律，并在新数据集上表现出较好的预测能力。因此，AI稳定判别模型的复杂度是重要的评价指标之一。

电网数据的采集和传输不可避免地会存在一定的误差和噪声[20-21]，部分AI模型对数据极为敏感，当输入数据中存在微小扰动时，可能会导致模型判断错误，并导致严重事故。此外，电网运行环境复杂，存在恶意攻击等不确定性因素，模型必须保证在数据异常条件下仍能维持输出的稳定性。鲁棒性要求当数据中存在较小的偏差时，不会对模型的性能产生影响，或者只能产生较小的影响；而当数据中出现较大的偏差时，不能对模型的性能产生灾难性的影响，即能够保持稳定的性能和准确的预测结果。因此，鲁棒性也是AI稳定判别模型需要考虑的重要指标之一。

当电网的拓扑结构、运行方式以及故障场景等发生变化时，已有模型的性能在新的场景下可能会大幅下降[22]，造成分析结果与实际的稳定指标相差较大。因此，必须保证模型在不同场景、不同任务或不同数据集上的适应能力。可迁移性是指能够在经过少量调整或无须调整的情况下，将已训练好的模型应用于新的场景时可解决相似问题的能力，从而节省重新训练模型所需的时间和资源，这也是衡量模型性能和应用范围的重要指标之一。

AI模型通常为黑箱模型，模型缺乏可解释性[23]，决策者往往无法得知模型判断的依据和原因，因而不能完全信任模型的判断结果，在一定程度上阻碍了AI在稳定性判别等高可靠性场景中的推广应用。可解释模型可以帮助调度员理解其决策依据和推理过程，并根据模型提供的结果做出更准确的决策，从而有助于及时发现潜在问题，提高模型的鲁棒性和可靠性，确保电网的稳定运行。因此，可解释性也是AI稳定判别模型的重要评价指标之一。

综上，结合电网实际应用存在的风险，考虑AI模型的准确性要求、模型在实际应用中的计算效率、模型对数据变化和复杂场景的容忍程度和应用到新场景的能力，以及模型训练和决策过程能够被人类所理解的程度，选取准确性、复杂度、鲁棒性、可迁移性和可解释性5个指标构建可信度评价指标体系，从这5个方面评价模型的可信度。

1.2 准确性

基于AI的电网稳定性判别主要过程如下：首先，利用海量的训练数据进行离线训练，构建满足性能要求的预训练模型，包括训练样本的生成和预处理、特征提取以及特征优化、算法的选取、模型的训练以及模型的评价；然后，预训练模型根据实时的电网数据快速评估系统的稳定状态，同时根据实际的评估结果实时反馈，及时对模型进行修正[7]，即在线应用阶段。AI稳定判别模型可归结为分类问题或者回归问题：若输出为稳定或不稳定此类二值标签，则为分类问题[24-26]；若输出为暂态稳定裕度、极限切除时间等连续性指标，则为回归问题[27-29]。通常而言，针对功角稳定、电压稳定、频率稳定以及宽频带振荡等不同稳定形态或问题，AI稳定判别模型的输入、输出及模型结构均有所不同，但在评估结果准确性评价方面的要求基本一致。

对于分类问题，评价模型准确性的指标有：准确率(accuracy)、精确率(precision)、召回率(recall)、特异性指标(false positive rate，FPR)、误判率等。

1)准确率λ_{Acc}：表示分类正确的样本在所有样本中所占的比重，即

$$\lambda_{Acc}=\frac{f_{TP}+f_{TN}}{f_{TP}+f_{FN}+f_{FP}+f_{TN}} \tag{1}$$

式中：f_{TP}为真正为正类且被预测为正类的样本；f_{TN}为真正为负类且被预测为负类的样本；f_{FN}为真正为负类但被预测为正类的样本；f_{FP}为真正为正类但被预测为负类的样本。

2)精确率λ_{Pre}：表示在被识别为正类的样本中，实际为正类的样本所占的比重，即

$$\lambda_{\mathrm{Pre}}=\frac{f_{\mathrm{TP}}}{f_{\mathrm{TP}}+f_{\mathrm{FP}}} \tag{2}$$

3）召回率λ_{Rec}：表示在所有真正为正类的样本中，被模型识别出来的正类样本所占的比重，即

$$\lambda_{\mathrm{Rec}}=\frac{f_{\mathrm{TP}}}{f_{\mathrm{TP}}+f_{\mathrm{FN}}} \tag{3}$$

4）特异性指标λ_{FPR}：在稳定判别场景中也可称之为漏判率，表示在所有真正为负类的样本中被错误预测为正类的样本所占的比重，即

$$\lambda_{\mathrm{FPR}}=\frac{f_{\mathrm{FP}}}{f_{\mathrm{FP}}+f_{\mathrm{TN}}} \tag{4}$$

5）误判率λ_{error}：表示所有被预测为负类的样本中实际为正类的样本所占的比例，即

$$\lambda_{\mathrm{error}}=\frac{f_{\mathrm{FN}}}{f_{\mathrm{FN}}+f_{\mathrm{TN}}} \tag{5}$$

对于回归问题，常见的评估指标有：平均绝对误差（mean absolute error，MAE）、平均绝对百分比误差（mean absolute percentage error，MAPE）、均方根误差（root mean square error，RMSE）。

1）平均绝对误差λ_{MAE}表示预测值与真实值之间的平均绝对差异值，即

$$\lambda_{\mathrm{MAE}}=\frac{1}{N}\sum_{q=1}^{N}|y_{\mathrm{T}q}-y_{\mathrm{P}q}| \tag{6}$$

式中：$y_{\mathrm{T}q}$为第q个样本的真实值；$y_{\mathrm{P}q}$为第q个样本的预测值；N为样本数量，且$\forall q=1,2,\cdots,N$。

2）平均绝对百分比误差λ_{MAPE}表示预测值与真实值之间的平均百分比误差，可以反映输出值与实际值的偏差度，即

$$\lambda_{\mathrm{MAPE}}=\frac{1}{N}\sum_{q=1}^{N}\left|\frac{y_{\mathrm{T}q}-y_{\mathrm{P}q}}{y_{\mathrm{T}q}}\right|\times 100\% \tag{7}$$

3）均方根误差λ_{RMSE}表示预测值与真实值之间误差平方和的均值的算数平方根，即

$$\lambda_{\mathrm{RMSE}}=\sqrt{\frac{1}{N}\sum_{q=1}^{N}(y_{\mathrm{T}q}-y_{\mathrm{P}q})^2} \tag{8}$$

对于频率偏移极值等频率安全指标预测的回归问题，可以采用上述误差类指标评估模型的准确性[30]；对于暂态稳定裕度等指标，可通过设定合适的阈值将回归性问题转化为二分类问题。例如，在暂态稳定分析中，暂态稳定裕度γ可以表示为系统不同故障位置的极限清除时间T_{CCT}与故障清除时间T_{CT}的差值[27]，即

$$\gamma=T_{\mathrm{CCT}}-T_{\mathrm{CT}} \tag{9}$$

将稳定裕度阈值设置为0，当暂态稳定裕度指标大于0时代表系统稳定、小于0时代表系统失稳。因此，可将暂态稳定裕度指标转化为稳定或不稳定的分类问题，从而在判断系统是否失稳的同时，计算系统的稳定程度[31]。

在实际评估中，可以选取准确率评估分类问题的准确性，也可以选取准确率、误判率、漏判率等指标的均值形成综合准确性指标来综合反映模型的准确性；还可以视具体问题而定，如AI稳定判别模型更看重对失稳样本的漏判率，故可以对漏判率赋予更高的权重。为使综合指标更具说服力，也可以采用模糊层次分析法确定各指标的权重，形成综合准确性指标。回归问题的评估方法与分类问题类似。

1.3　复杂度

模型复杂度可以从时间复杂度、空间复杂度两方面进行评价：时间复杂度体现了模型训练和预测所花费的时间；空间复杂度体现了算法执行过程中需要的内存资源[32]。

时间复杂度通常表示为$\mathrm{O}(f(n))$，其中，$f(n)$是与输入规模n相关的函数，随着输入规模n的增加，时间复杂度描述了算法执行时间的增长趋势。常见的时间复杂度有：常数时间$\mathrm{O}(1)$、对数时间$\mathrm{O}(\log n)$（常数因子底数在O记法中被丢弃）、线性时间$\mathrm{O}(n)$、线性对数时间$\mathrm{O}(n\log n)$、平方时间$\mathrm{O}(n^2)$。其相互间的大小关系为：

$$\mathrm{O}(1)<\mathrm{O}(\log n)<\mathrm{O}(n)<\mathrm{O}(n\log n)<\mathrm{O}(n^2) \tag{10}$$

为保证计算的准确性和可靠性，空间复杂度通常采用均匀加权平均法计算模型空间复杂度。假设模型有M个属性，如决策树的属性包括树的数量、树的深度、节点数目、分支数目，每个属性的值分别为$a_1,a_2,\cdots,a_M$，则模型空间复杂度可表示为：

$$c_{\mathrm{s}}=\frac{1}{M}\sum_{m=1}^{M}a_m \tag{11}$$

式中：c_{s}为空间复杂度的值；a_m为第m个属性的值。

考虑到时间复杂度和空间复杂度的特征尺度差异较大，采用最大绝对值缩放（max absolute scaling，MAS）计算模型综合复杂度。MAS以数据集中的绝对最大值为基准，将时间复杂度和空间复杂度映射到相对一致的尺度上，确保在综合复杂度计算中不会被某一特征的尺度差异所主导，则综合复杂度的表达式为：

$$c=\frac{1}{2}\left(\frac{c_{\mathrm{t}}}{c_{\mathrm{t,max,abs}}}+\frac{c_{\mathrm{s}}}{c_{\mathrm{s,max,abs}}}\right) \tag{12}$$

式中：c为模型综合复杂度的值；c_{t}为时间复杂度的值；$c_{\mathrm{t,max,abs}}$为时间复杂度的最大绝对值；$c_{\mathrm{s,max,abs}}$为空间复杂度的最大绝对值。

1.4 鲁棒性

通常通过对比在训练样本中增加噪声前后模型性能指标的变化情况评估模型的鲁棒性，体现了模型对于不同程度噪声和变化的敏感性。双样本KS（Kolmogorov-Smirnov）检验是一种用于检验两个数据集是否来自同一分布的非参数检验方法。为评估模型的鲁棒性[33]，KS检验可用于比较基于正常样本和添加噪声后样本的模型预测分布。添加噪声前后模型预测分布越相似，则模型的鲁棒性越好。

KS检验的计算方法如下：计算添加噪声前后模型的预测分布F_1和F_2；计算KS检验统计量D，D越小则模型的鲁棒性越好，其表达式为：

$$D=\max|F_1(\varepsilon)-F_2(\varepsilon)| \tag{13}$$

式中：F_1为添加噪声前的预测分布；F_2为添加噪声后的预测分布；ε为待检测的样本。

1.5 可迁移性

传统的评估模型可迁移性的方法是通过源域数据完成模型训练，计算该模型的准确率等反映模型性能的指标，再利用目标数据集对源模型进行微调，并计算微调后的模型准确率等指标，通过对比准确率等指标评估模型的迁移效果[34-37]。

传统的评估模型可迁移性的方法既要在源域数据训练模型，又要对源模型进行微调。目前，计算机视觉领域的部分学者提出了条件熵（conditional entropy，CE）、H分数（H-score）、对数预期经验预测（log expected empirical prediction，LEEP）、最大证据对数（logarithm of maximum evidence，LogME）、基于最优传输的条件熵（optimal transport based conditional entropy，OTCE）等评价指标。CE表示一个随机变量的值对另一个随机变量的值所提供的信息量的量度，从源任务迁移到目标任务时，通过计算每个任务训练集的标签序列的条件熵来衡量可迁移性的大小[38]。H-score是一种基于统计学和信息论原理的评估方法，通过类间方差和特征冗余度来近似估计目标标签的最优对数损失，并将可迁移性定义为最优源任务特征相对于目标任务的归一化H-score[39]。LEEP不需要对源模型进行微调，只需将目标数据在源模型上进行一次前向传递即可。首先，利用源任务训练好的模型计算目标数据集中输入的伪标签分布；然后，计算目标标签在伪标签上的经验条件分布；最后，计算LEEP值即可[40]。OTCE根据源任务和目标任务的域差异和任务差异来表征源任务和目标任务之间的可迁移性[41]。但是，以上方法只能用于评估分类问题。LogME是一种同时适用于分类和回归任务的方法，其将预训练模型视为一个固定的特征提取器，根据提取的特征和标签之间的对数最大证据评估可迁移性[42]。该方法避免了最大似然法的过拟合问题，覆盖了视觉、自然语言处理（natural language processing，NLP）、分类、回归、有监督预训练模型、无监督预训练模型等各类应用，是目前通用性最强的指标。

大多数文献评估可迁移性指标好坏的整体思路如下：首先，通过源模型和目标数据集计算待评判模型的可迁移性指标；然后，利用目标训练集对源模型进行微调，将目标测试集作为微调后模型的输入，并计算模型评估准确率，其反映了模型真实的可迁移性；进而，评估可迁移性指标对准确率的预测程度，即该指标能否真实地反映模型的可迁移性[43]。目前，该思路在计算机视觉领域应用较多，可推广至其他行业领域。

1.6 可解释性

AI模型的黑箱特性导致其在实际应用中受到诸多限制，为提高模型的可解释性，模型可解释性方法得到了研究发展。目前，常用的方法有排序重要性（permutation importance）、模型无关的局部可解释性（local interpretable model-agnostic explanations，LIME）方法、沙普利值可加性解释（Shapley additive explanation，SHAP）方法等。

排序重要性方法[44]的基本原理是在训练好模型之后，随机打乱某个特征，而其余特征保持不变，对比打乱该特征前后模型预测精度的变化，得到特征的重要性排序。

LIME方法[45]是一种与模型无关、适用于分类以及回归问题的局部可解释性方法，其本质是在局部使用简单的可解释性模型对原模型的预测结果进行拟合，如线性模型、决策树等。如图1所示，分类的决策边界如紫色区域的边界所示，使用线性模型无法很好地逼近，选取其中的一个待测样本，如图中虚线上紫色五角星所示，在该样本的邻域进行采样，距离待测样本越近的采样点权重越高，用简单模型拟合原模型对采样样本的预测结果，得到的简单线性模型（图中虚线所示）称为局部可解释模型，该线性模型的权重可以反映出每个特征的重要程度以及贡献程度。

SHAP方法[46]是一种基于合作博弈论的局部以及全局可解释方法，可以解释各种机器学习模型以及深度学习模型，其原理是计算每个样本的每个特征对模型预测的贡献度，即Shapley值，进而可以得知每个特征对模型预测结果的影响，并且可以分析某个样本的各个特征对预测结果的影响，既可以进

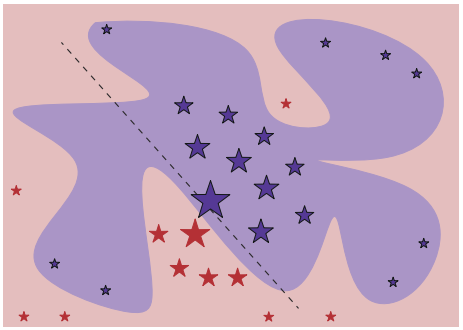

图1　LIME方法原理图
Fig. 1　Principle diagram of LIME method

行全局解释，也可以进行局部解释。Shapley值满足下式：

$$y_i = y_{\text{base}} + \sum_{j=1}^{l} \phi_j \tag{14}$$

式中：y_i为第i个样本的预测值；y_{base}为整个模型的基准值（通常是所有样本的目标变量均值）；l为一个样本包含的特征总数；ϕ_j为特征j的Shapley值，$\forall j = 1, 2, \cdots, l$。

特征j的Shapley值的含义是遍历所有可能的特征子集，对去除特征j前后预测输出的差值加权平均得到特征j对预测结果的贡献，即Shapley值。其计算公式如下：

$$\phi_j = \sum_{S \subseteq Z \setminus \{j\}} \frac{|S|!(|Z| - |S| - 1)!}{|Z|!} \left[g_x(S \cup \{j\}) - g_x(S) \right] \tag{15}$$

式中：Z为所有特征的集合；S为不包含特征j的特征子集；$|S|$、$|Z|$分别为集合S和Z中非零元素的个数；$g_x(S)$为模型在给定特征子集S时的预测输出；x表示输入数据的一个样本。

2　综合可信度评价指标计算

在AI稳定判别模型的实用化过程中，模型的准确性、可解释性等多重属性往往会被考虑在内，且上述5个分项指标间相互独立，从不同的角度和维度反映了模型的性能，而单一的评价指标并不足以全面地反映模型的可信度。本文提出的可信度评价体系更侧重于针对特定的问题，评价模型间的相对表现而非单一模型的绝对优劣，并从多个候选模型中优选出最具可信度的模型，为模型在实际部署前提供必要的验证和筛选流程。将各分项指标同时考虑在内，借助综合评价方法形成综合可信度评价指标，为决策者提供了一个更全面的评价视角，并且可以更简洁、直观地反映模型的综合性能。

2.1　综合可信度评价方法

由于各分项指标的含义及量纲不同，并不能直接利用各分项指标的属性值计算综合可信度，在计算之前应对各个指标进行规范化处理。

2.1.1　指标属性值的规范化处理

根据指标取值的优劣，可将指标分为效益型指标、成本型指标、偏离型指标等[47]。其中，效益型指标越大越好（正向指标），成本型指标越小越好（负向指标），偏离型指标距离标准值越近越好。通常，将负向指标转化成正向指标以实现指标的同向化，转换方式为：

$$k_1 = \frac{1}{p + k_{\max} + k} \tag{16}$$

式中：k为同向化前的指标属性值；k_1为同向化后的指标属性值；$k_{\max}$为指标的最大值；p为协调系数，一般取为0.1。

同向化处理之后，需对指标的单位和数量级进行无量纲化处理：

$$k_2 = \frac{k_1 - k_{1,\min}}{k_{1,\max} - k_{1,\min}} \tag{17}$$

式中：k_2为无量纲化处理后指标的属性值；$k_{1,\min}$为处理前指标的最小值；$k_{1,\max}$为处理前指标的最大值。无量纲化处理后，各指标值的范围在0～1之间。

可信度评价指标体系中的复杂度属于负向指标，即复杂度越低越好，其余4个指标为正向指标。因此，在综合可信度的计算时，需要对指标进行同向化处理，将复杂度转换成正向指标，各分项指标的评分值越大代表模型在该指标上的表现越好，并对各指标进行无量纲化处理，将指标范围归一到0～1之间。

2.1.2　综合指标的计算

规范化后的指标为归一化的正向值，对准确性、模型复杂度、模型鲁棒性、可解释性、可迁移性5个指标加权形成可信度综合评价指标。其计算公式如下：

$$H = \sum_{b=1}^{5} h_b H_b \tag{18}$$

式中：H为综合可信度；H_b为第b个指标规范化后的属性值；h_b为第b个指标的权重，且5个指标的权重之和为1。

综合可信度计算的整体流程如下：首先，评估模型的准确性、复杂度等5个方面，得到5个指标的属性值；其次，对指标属性值进行规范化处理；然后，根据专家打分表确定各指标的权重；最后，将各指标的属性值以及权重代入式（18）得到模型的综合可

信度。

各分项指标的范围为[0,1],则综合可信度的范围为[0,100%],且模型的准确性、鲁棒性、可解释性等指标的属性值越大,综合可信度的值也越大,代表模型更加可信。

2.2 指标权重的确定

在综合指标的构建过程中,各指标权重的分配至关重要。目前,常用的权重分配方法有主观赋权法、客观赋权法和组合赋权法3类,而主观赋权法又包括专家咨询法、层次分析法等。其中,层次分析法是美国运筹学家T. L. Saaty提出的一种定性和定量相结合的系统分析方法,其主要依赖于专家的专业知识和个人经验,主观性较强,且没有考虑到专家判断的模糊性,难以做到客观准确。为此,模糊层次分析法在层次分析法的基础上引入模糊理论,改进了层次分析法主观性强等缺点[48]。

2.2.1 模糊层次分析法原理

模糊层次分析法是一种基于层次分析法的定性和定量相结合的系统分析方法,对具有模糊性的因素进行定量计算[49],其计算方法如下。

1)比较任意两个指标的相对重要程度,按照表1所示的方法对指标相对重要程度进行定量描述,若包含U个指标,r_{uv}为模糊判断矩阵第u行第v列的元素,可得模糊判断矩阵为$\boldsymbol{R}=(r_{uv})_{U\times U}(u,v=1,2,\cdots,U)$。若指标$u$比指标$v$重要得多,则标度$r_{uv}$为0.8,其他情况以此类推。

表1 重要程度量化

Table 1 Quantification of importance

标度	定义	说明
0.5	同等重要	两因素相比较,同等重要
0.6	稍微重要	两因素相比较,一因素比另一因素稍微重要
0.7	明显重要	两因素相比较,一因素比另一因素明显重要
0.8	重要得多	两因素相比较,一因素比另一因素重要得多
0.9	极端重要	两因素相比较,一因素比另一因素极端重要
0.1、0.2、0.3、0.4	反比较	若因素r_u与因素r_v相比较得到的判断为r_{uv},则因素r_v与因素r_u相比较得到的判断为$r_{vu}=1-r_{uv}$

2)通过模糊互补判断矩阵计算各指标的权重,计算公式为:

$$W_u=\frac{\sum_{u,v=1}^{U}r_{uv}+\frac{U}{2}-1}{U(U-1)} \tag{19}$$

式中:W_u为指标u的权重。

3)进行一致性检验,判断计算得到的权重是否合理。计算各个模糊矩阵之间的相容性指标,以及模糊互补判断矩阵与特征矩阵的相容性指标,判断模糊判断矩阵是否具有一致性。特征矩阵的计算如下:

$$\boldsymbol{W}=(W_{uv})_{U\times U}=\left(\frac{W_u}{W_u+W_v}\right)_{U\times U} \tag{20}$$

式中:$\boldsymbol{W}$为特征矩阵,且$\forall u,v=1,2,\cdots,U$;W_{uv}为特征矩阵第u行第v列的元素。

相容性指标的计算公式为:

$$I(\boldsymbol{A},\boldsymbol{B})=\frac{1}{U^2}\sum_{u,v=1}^{U}|a_{uv}+b_{uv}-1| \tag{21}$$

式中:I为矩阵$\boldsymbol{A}$和矩阵$\boldsymbol{B}$之间的相容性指标;a_{uv}和b_{uv}分别为矩阵$\boldsymbol{A}$和矩阵$\boldsymbol{B}$中第u行第v列的元素。

在多个专家参与评判的情况下,如果每个判断矩阵与其对应的特征矩阵满足一致性,且每个判断矩阵之间满足一致性,则可以将各专家权重的均值作为最终的权重[49]。因此,权重的计算步骤如下:建立各个专家的模糊互补判断矩阵;通过模糊互补判断矩阵求解对应的权重,得到权重集;进行一致性检验;将权重集进行算术平均,求得模糊权重。

2.2.2 指标权重的计算

为确定各指标的权重,设计了专家打分表调研业内专家关于指标权重分配的意见。考虑到部分专家判断的模糊性,采用了确切权重和相对重要程度评价两种方式。专家可以直接对各个指标赋予确切的权重,也可以仅给出指标之间的相对重要程度。为使调研结果更具说服力,收集了来自国家电网有限公司、中国南方电网有限责任公司各级调度以及科研高校等的百余位业内专家的反馈意见。

对于确切权重相关问卷,在将所有确切权重的问卷统计后,分别计算5个指标权重的均值,称之为平均确切权重。对于相对重要程度评价相关问卷,利用模糊层次分析法进一步计算权重,称之为模糊权重。将平均确切权重和模糊权重按照各自的专家人数进行加权平均,得到最终的指标权重。开发Python程序统计各专家的反馈意见并计算最终的指标权重,得到各指标的平均确切权重为[0.416,0.120,0.171,0.171,0.122]。

按照模糊层次分析法计算模糊权重,以一位专家为例,具体的计算步骤如下。

首先,根据相对重要程度生成对应的模糊判断矩阵:

$$\boldsymbol{R}=\begin{bmatrix}0.5&0.6&0.7&0.7&0.5\\0.4&0.5&0.6&0.7&0.5\\0.3&0.4&0.5&0.4&0.4\\0.3&0.3&0.6&0.5&0.4\\0.5&0.5&0.6&0.6&0.5\end{bmatrix} \tag{22}$$

按照式(19)计算各指标权重为[0.225, 0.210, 0.175, 0.180, 0.210]，按照式(20)计算特征矩阵为：

$$\boldsymbol{W}=\begin{bmatrix}0.500 & 0.517 & 0.563 & 0.556 & 0.517\\ 0.483 & 0.500 & 0.546 & 0.539 & 0.500\\ 0.438 & 0.455 & 0.500 & 0.493 & 0.455\\ 0.444 & 0.462 & 0.507 & 0.500 & 0.462\\ 0.483 & 0.500 & 0.546 & 0.539 & 0.500\end{bmatrix} \tag{23}$$

其他专家权重的计算与之类似。一致性校验通过后，对所有进行模糊评价的专家权重取均值，得到模糊权重为[0.230, 0.186, 0.204, 0.195, 0.185]。

图2所示为有效调研表所收集的各指标精确权重以及模糊权重的分布情况。从图中可以看出，准确性指标的权重分布具有一定的均匀性，复杂度、鲁棒性、可迁移性以及可解释性指标的分布具有一定的对称性。因此，将各指标权重的均值作为最终的权重具有合理性。将平均确切权重和模糊权重取均值，最终得到准确性、复杂度、鲁棒性、可迁移性、可解释性等指标的权重计算结果分别为0.33、0.15、0.19、0.18、0.15。

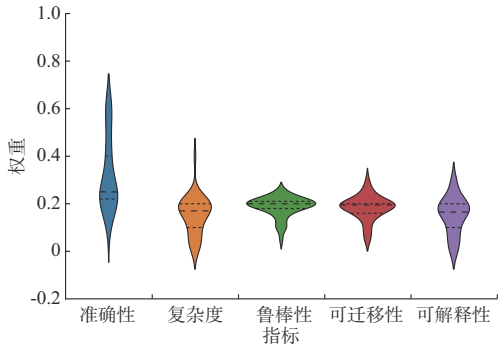

图2　各指标权重分布情况

Fig. 2　Distribution of weights for various indicators

上述指标权重的计算结果显示，准确性所占的比重最大，其次是鲁棒性、可迁移性、可解释性与复杂度。模型的准确性直接决定了其在实际应用中的效果，足够高的准确性是模型应用于电力场景的前提。鲁棒性衡量了模型对数据噪声的容忍程度，电力AI模型的在线应用依赖于实时采集的电力数据，而此类数据中往往存在一定的噪声，模型具有一定的鲁棒性十分必要。考虑到电力系统始终处于动态变化之中，当系统运行方式和拓扑结构发生变化时，原有模型的性能可能会大幅下降，模型具有一定的可迁移性可以节省重新训练模型的时间和精力。可解释性反映了模型的决策过程被人类所理解的程度，而复杂度反映了模型决策过程中的内存需求和时间需求。权重计算结果反映了在综合可信度评价中各专家对各指标的重视程度，符合电力系统的实际应用需求。

此外，随着AI算法在电网稳定性判别中的应用不断深化，相关领域专家学者关注的具体指标及其权重也会随之动态调整。本文给出的指标及权重在电网AI稳定判别模型应用初期，可为模型可信度评价提供一定的参考。在实际应用中，为确保各指标权重的时效性以及评价结果的合理性，可根据所关注的稳定形态或具体问题，采取滚动式的数据收集机制，定期收集专家打分表，并据此动态调整具体指标和权重。

3　实例验证

3.1　电压支撑强度增强指标的综合可信度评价

选取新能源多场站短路比(multiple renewable energy station short-circuit ratio，MRSCR)临界阈值智能增强模型作为测试模型[50]，评价增强前后模型的可信度。该测试模型在中国电力科学研究院万节点机电仿真测试系统(包含581台同步电源、894个新能源场站、6回直流、10 577个三相节点)上，构建了10 000个训练样本集以及3 000个测试样本集，覆盖14套典型方式和6套非典型方式(考虑送端负荷水平、新能源出力占发电比例以及输送功率的不同，受端负荷水平、直流受端占负荷比例、新能源出力占发电比例的不同)，涵盖500类确定性常规故障以及继电保护拒动等1 200类非预想故障。MRSCR是电压支撑强度的量化指标。目前，该指标的计算参数存在阻抗矩阵等非直接量测量，无法实时计算，且其临界阈值通过典型参数折算和仿真设置，即临界短路比(critical short-circuit ratio，CSCR)为经验值。然而，不同工况下计算参数和CSCR是动态变化的，该类指标及其临界阈值仅能粗略表征系统的电压支撑强度。针对MRSCR和CSCR计算公式中的非直接量测量和经验性参数，利用智能算法进行增强，计算等值电势的精确值。智能增强模型是以母线电压幅值、相角、节点注入功率作为输入，等值电势作为输出的回归模型。

选取平均绝对误差作为准确性指标，选取传统的可迁移性评估方法计算可迁移性指标，利用SHAP方法对模型进行可解释性分析，最终将各指标进行规范化处理。评价智能增强前后模型的可信度，得到各分项指标及综合可信度指标如表2所示，各分项指标对比如图3所示。可见，智能增强后模型的准确性、鲁棒性、可迁移性指标均有所提升，而复杂度和可解释性有所下降，综合可信度提高。

表2　电压支撑强度模型各指标对比
Table 2　Indicator comparison of voltage support strength model

增强前后	准确性	复杂度	鲁棒性	可迁移性	可解释性	综合可信度
增强前	0.89	0.98	0.51	0.85	1.00	0.84
增强后	0.95	0.96	0.95	0.91	0.88	0.93
提升比例/%	6.74	−2.04	86.27	7.06	−12.00	10.71

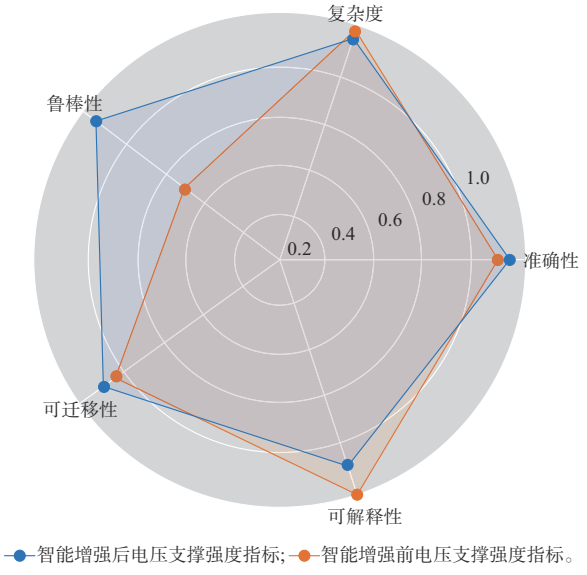

图3　电压支撑强度模型智能增强前后的各指标雷达图
Fig. 3　Radar chart of indicators before and after intelligent enhancement of voltage support strength model

在增强前，将等值电势取经验值代入短路比指标中进行分析，无需AI模型的预测过程，而且增强前的方法本身是具有可解释性的。在增强后，将电网实时量测量输入智能增强模型中获取等值电势的精确值，故智能增强模型的准确性更高。智能增强模型在预测过程中需要花费额外的时间和内存资源，故模型更加复杂。智能增强模型的训练依赖于海量的训练数据，增强后的模型对噪声具有更高的容忍度，鲁棒性提高。图注意力网络具有较好的表示能力和适应性，不但能处理异构图数据，而且具有适应不同拓扑结构变化的能力，提高了模型的可迁移性。由于图注意力网络本身在等值电势的预测过程中缺乏一定的可解释性，故可解释性降低。总体而言，智能增强后的模型可信度提高，与综合可信度指标的计算结果一致，证明了本文所提方法的有效性。

3.2　宽频带振荡增强指标的综合可信度评价

选取宽频带振荡阻尼系数智能增强模型作为另一测试模型[51-52]。该测试模型在中国电力科学研究院万节点电磁仿真测试系统（包含357台同步电源、385个新能源场站、6回直流、13 139个三相节点）上，构建了10 000个训练样本集以及3 000个测试样本集，覆盖14套典型方式和6套非典型方式。宽频带振荡通常存在全局主导模式，即所有量测点得到的振荡频率和阻尼基本一致。支路阻尼系数（支路电流瞬时值响应）仅能量化局部阻尼情况，通过全网支路加权可反映系统阻尼情况，但权重系数难以确定，振荡风险评估的准确性受限。智能增强后的模型是一种融合PageRank算法和自编码器的宽频带振荡风险辨识框架，将阻尼因子计算嵌入神经网络以实现系统阻尼系数的智能增强，显著提高了振荡风险评估的准确率及计算效率。

计算准确率、精确率、召回率等作为准确性指标，其他计算与3.1节类似，此处不再赘述。评价智能增强前后模型的可信度，得到各分项指标及综合可信度指标如表3所示，各分项指标对比情况如图4所示。可见，智能增强后模型的准确性、鲁棒性、可迁移性以及可解释性均有提高，而复杂度下降，综合可信度提高。

表3　宽频带振荡模型各指标对比
Table 3　Indicator comparison of broadband oscillation model

增强前后	准确性	复杂度	鲁棒性	可迁移性	可解释性	综合可信度
增强前	0.91	0.99	0.44	0.85	0.75	0.80
增强后	1.00	0.63	0.84	0.96	0.89	0.90
提升比例/%	9.89	−36.36	90.91	12.94	18.67	12.50

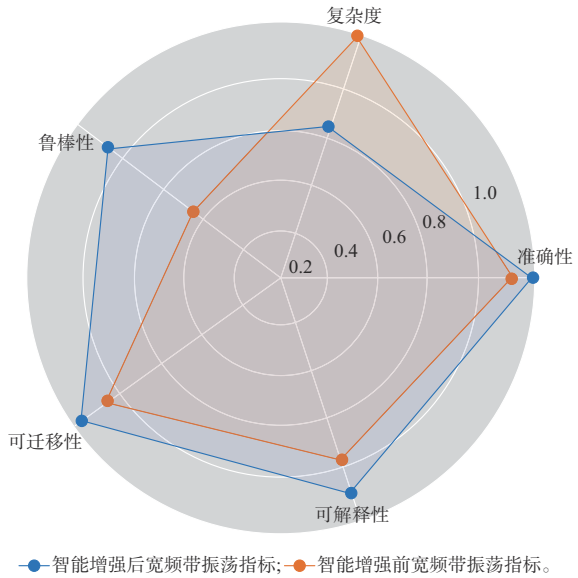

图4　宽频带振荡模型智能增强前后各指标雷达图
Fig. 4　Radar chart of indicators before and after intelligent enhancement of broadband oscillation model

增强前模型的系统阻尼采用线路阻尼加权平均的计算方式，而智能增强模型采用智能算法计算智能增强线路权重，相比于增强前的加权平均更加合理，提高了模型的准确性。智能增强算法计算线路权重需要花费额外的时间和内存资源，故复杂度变差。智能增强模型将系统分为多局部区域并采用平均共识机制，在数据缺失与可变拓扑情况下具有更好的鲁棒性与评估稳定性。智能增强模型结合了系统拓扑特性，可迁移性提高。智能增强线路权重可以反映线路对系统振荡情况的影响，可解释性提高。总体来看，智能增强后的模型可信度提高，与综合可信度指标的计算结果趋于一致。

3.3　频率稳定判别模型的综合可信度评价

基于万节点机电仿真数据（同3.1节）分别训练卷积神经网络（convolutional neural network，CNN）、长短期记忆（long short-term memory，LSTM）网络、CNN-LSTM网络等3种纯AI模型进行频率稳定判别，并开展可信度评价，评价结果如表4和图5所示。从可信度的评价结果可知，3种模型的准确性相差无几；复杂度方面，CNN模型明显优于其他两种模型；3种模型的鲁棒性指标均在80%左右，其中，CNN-LSTM网络的鲁棒性较高；3种模型的可迁移性与可解释性都较差；由于CNN模型的复杂度表现更好，使得其综合可信度略高于其他两个模型。

表4　不同模型的评价结果
Table 4　Evaluation results of various models

模型	准确性	复杂度	鲁棒性	可迁移性	可解释性	综合可信度
CNN	0.98	0.90	0.79	0.61	0.57	0.80
LSTM	0.97	0.45	0.81	0.55	0.56	0.72
CNN-LSTM	0.98	0.89	0.88	0.42	0.59	0.79

通过机理层面的分析可知，LSTM网络内部每个单元都需通过多个门（如输入门、遗忘门和输出门）进行复杂的计算，故LSTM和CNN-LSTM网络进行预测时需要花费更多的时间，复杂度更差。同时，CNN和LSTM网络具有良好的特征提取能力，能够从输入数据中提取关键的特征信息，一定程度上能够忽略输入数据中的噪声，具有一定的抗噪性能和鲁棒性。CNN-LSTM网络结合了CNN和LSTM网络的优点，能够同时提取数据关键局部特征和时间依赖关系，鲁棒性更高。但由于3种模型并未使用对抗样本进行对抗训练，或采取其他提高模型鲁棒性的措施，整体上模型的鲁棒性并不高。

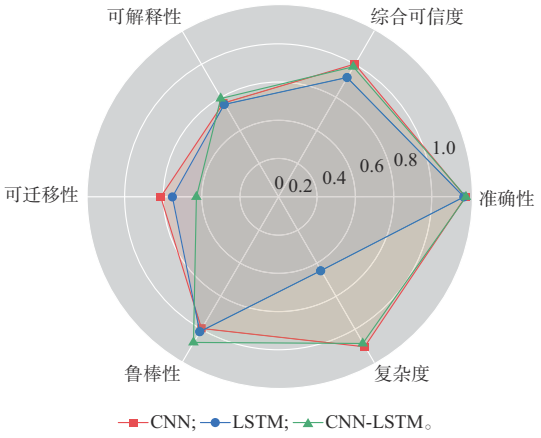

图5　不同模型评价结果雷达图
Fig. 5　Radar chart of evaluation results of various models

CNN和LSTM网络作为深度学习中的两种通用模型，相较于CNN-LSTM网络这一组合模型在不同的任务上具有更好的适应性，故可迁移性高于CNN-LSTM网络，但3种模型在构建过程中都没有考虑到系统网络的拓扑特征，所以模型的可迁移性都较低。LSTM网络作为一种循环神经网络，通过门控机制来控制信息的流动和存储，保证其内部状态在时间上具有连续性，其内部决策过程和推理机制相对复杂，可解释性更差，且3种模型均为纯AI模型，所以整体上可解释性都较低。各分项指标的评价结果与理论分析结果基本一致，证明了可信度评价体系的有效性。

重复两次划分训练集、验证集以及测试集，改变数据的分布规律，重新训练CNN、LSTM网络、CNN-LSTM网络3种模型，并测试模型的可信度，评价结果如表5所示。CNN1、LSTM1、C-L1代表重复一次划分的模型，CNN2、LSTM2、C-L2代表重复二次划分的模型。分析评价结果可知，3种模型的准确性相近，LSTM网络的复杂度效果最差，CNN-LSTM网络的鲁棒性更好；3种模型的可迁移性和可解释性都较差，LSTM网络的综合可信度得分最低。

表5　各指标评价结果
Table 5　Evaluation results of various indicators

模型	准确性	复杂度	鲁棒性	可迁移性	可解释性	综合可信度
CNN1	0.98	0.91	0.76	0.50	0.59	0.78
LSTM1	0.98	0.47	0.81	0.50	0.56	0.71
C-L1	0.97	0.90	0.85	0.48	0.56	0.72
CNN2	0.99	0.95	0.75	0.56	0.58	0.80
LSTM2	0.98	0.46	0.80	0.49	0.54	0.71
C-L2	0.97	0.87	0.90	0.41	0.60	0.78

由表5可见,改变数据分布时各评价指标的具体数值会呈现出一定的波动,但其相对排名与整体评估结果保持不变。而本文提出的可信度评价体系侧重于评价模型间的相对表现而非单一模型的绝对优劣。因此,数据分布的变化不会对可信度的评价结果造成实质性影响。

4 模型可信度提升策略的探讨及展望

为切实推进AI稳定判别模型在电力系统中的实际应用,本文构建了可信度评价体系和综合评价方法以评估模型的可信度。获取评价结果后的重点在于如何采用有效可行的思路和策略,对模型进行优化和完善。

针对AI稳定判别模型的鲁棒性风险,可考虑对数据进行降噪处理或利用对抗样本进行对抗训练以提升模型的鲁棒性。AI模型的拟合边界难以与真实的决策边界重合,在训练集中加入对抗样本进行对抗训练使拟合边界逼近真实边界,保证模型能抵御对抗样本的攻击,可有效提升模型的鲁棒性。对抗训练的过程如下:首先,利用投影梯度下降或快速梯度符号攻击等方法生成对抗样本;然后,将对抗样本加入训练集中对模型进行训练和优化。对抗训练可使模型学习到更加丰富的特征表示,同时显著提升其对抗攻击的防御能力,增强模型的鲁棒性和稳定性。

针对AI稳定判别模型的迁移局限性问题,可在模型的选择和设计阶段采用通用的模型架构和算法、在构建特征集时考虑系统的拓扑结构特征或者采用迁移学习方法等手段进行解决。通用的模型架构和算法如CNN、循环神经网络等,可适用于不同的问题,在多种任务上都具有良好的表现。图深度学习模型是一种专门用于处理图形数据的深度学习架构[53],能够捕获和利用图形数据中的复杂结构和关系,将其应用于稳定判别中,可有效利用描述电网拓扑结构的离散矩阵参量与连续型运行变量,在输入特征中融合系统的拓扑信息,并将各节点的多类连续特征进行聚合,提高稳定判别模型的特征提取能力,从而有效改善拓扑结构变化时模型的泛化性,提高模型的可迁移性。迁移学习是一种将模型从源域中学到的知识迁移到目标域中的方法,当电力系统的运行场景发生变化时,根据原有模型的网络参数等知识辅助模型在新场景下的学习,从而减少重新训练模型所需的时间和样本数量。

针对AI模型的可解释性问题,可通过以下方法进行改善:

1)在模型的构建中,可选取本身具有可解释性的模型,如线性模型、决策树等。线性模型为输入特征的加权组合,特征权重可直观体现各个特征对输出特征的影响效果,决策树的每一个决策都可对应于具体的逻辑规则,具有一定的可解释性。

2)构建数据与机理融合的模型。在AI模型中融入一定的电力系统相关物理机理和专业经验,如利用深度神经网络学习电网故障后动态过程的微分-代数方程,将物理方程与数据驱动模型进行深度融合,以提升AI稳定判别模型的可解释性[54]。

3)使用可解释性方法对模型的决策过程进行解释。例如,建模前利用主成分分析等方法对数据进行处理,挖掘或展示数据主特征,帮助研究人员了解数据分布特性;或在训练模型后,利用类激活映射、LIME、SHAP等可解释性方法对模型进行解释。

5 结语

为应对电网AI稳定判别模型缺乏相适应的可信度评价方法问题,本文提出了可信度评价指标体系和综合可信度评价方法。结合电网的特点和需求,从准确性、复杂度、鲁棒性、可迁移性以及可解释性5个方面构建可信度评价体系。引入模糊层次分析法,构建基于5个分项指标的综合可信度评价指标。针对回归模型和分类模型两个典型案例开展可信度测试,评价结果与理论分析相一致,表明了所提方法的实用性和有效性。此外,对模型可信度的提升策略进行了初步探讨。本文提出的可信度评价方法以及指标权重计算可对行业以及工程应用提供参考。

目前,本文的工作仅限于可信度评价方法,而模型可信度的提升是一个复杂且多维度的过程,涉及模型的鲁棒性、可解释性、可迁移性等多个方面,具体的改进策略和技术手段需进行更为深入和系统的研究。如何根据不同稳定形态和AI模型的特点,确定模型的薄弱环节并有针对性地提出模型优化和改进的方向,从而提升模型的综合可信度和工程实用水平,将是后续工作的重点。

参 考 文 献

[1] 刘俊,孙惠文,吴柳,等.电力系统暂态稳定性评估综述[J].智慧电力,2019,47(12):44-53.
LIU Jun, SUN Huiwen, WU Liu, et al. Overview of transient stability assessment of power systems[J]. Smart Power, 2019, 47(12): 44-53.

[2] 汤奕,崔晗,李峰,等.人工智能在电力系统暂态问题中的应用综述[J].中国电机工程学报,2019,39(1):2-13.

TANG Yi, CUI Han, LI Feng, et al. Review on artificial intelligence in power system transient stability analysis [J]. Proceedings of the CSEE, 2019, 39(1): 2-13.

[3] 赵晋泉，夏雪，徐春雷，等．新一代人工智能技术在电力系统调度运行中的应用评述[J]．电力系统自动化，2020，44(24)：1-10.
ZHAO Jinquan, XIA Xue, XU Chunlei, et al. Review on application of new generation artificial intelligence technology in power system dispatching and operation [J]. Automation of Electric Power Systems, 2020, 44(24): 1-10.

[4] DUAN R J, MAO X F, QIN A K, et al. Adversarial laser beam: effective physical-world attack to DNNs in a blink[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, USA: 16057-16066.

[5] 何积丰．安全可信人工智能[J]．信息安全与通信保密，2019，17(10)：5-8.
HE Jifeng. Secure and trusted artificial intelligence [J]. Information Security and Communications Privacy, 2019, 17(10): 5-8.

[6] 门天宇．国外电力系统网络安全事件对我国的启示[J]．电器工业，2022(10)：80-82.
MEN Tianyu. Enlightenment of foreign power system network security incidents to China [J]. China Electrical Equipment Industry, 2022(10): 80-82.

[7] 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.

[8] 陈羽飞，赵琦，何永君，等．人工智能在电力系统中的应用综述[J]．分布式能源，2023，8(6)：49-57.
CHEN Yufei, ZHAO Qi, HE Yongjun, et al. An overview of the application of artificial intelligence in power systems [J]. Distributed Energy, 2023, 8(6): 49-57.

[9] 孔勇，李美桃，王伟，等．美国《人工智能风险管理框架》解读[J]．中国信息化，2023(3)：39-44.
KONG Yong, LI Meitao, WANG Wei, et al. Interpretation of American "Artificial Intelligence Risk Management Framework" [J]. iCHINA, 2023(3): 39-44.

[10] 刘晗，李凯旋，陈仪香．人工智能系统可信性度量评估研究综述[J]．软件学报，2023，34(8)：3774-3792.
LIU Han, LI Kaixuan, CHEN Yixiang. Survey on trustworthiness measurement for artificial intelligence systems [J]. Journal of Software, 2023, 34(8): 3774-3792.

[11] 曹建峰，方龄曼．欧盟人工智能伦理与治理的路径及启示[J]．人工智能，2019，6(4)：39-47.
CAO Jianfeng, FANG Lingman. The path and enlightenment of EU artificial intelligence ethics and governance[J]. AI-View, 2019, 6(4): 39-47.

[12] 程晓荣，李天琦．电网数据可信性度量模型研究[J]．华北电力大学学报(自然科学版)，2017，44(2)：83-90.
CHENG Xiaorong, LI Tianqi. Research on credibility measurement model of power grid data [J]. Journal of North China Electric Power University (Natural Science Edition), 2017, 44(2): 83-90.

[13] 尚宇炜，马钊，彭晨阳，等．内嵌专业知识和经验的机器学习方法探索(一)：引导学习的提出与理论基础[J]．中国电机工程学报，2017，37(19)：5560-5571.
SHANG Yuwei, MA Zhao, PENG Chenyang, et al. Study of a novel machine learning method embedding expertise (Ⅰ): proposals and fundamentals of guiding learning[J]. Proceedings of the CSEE, 2017, 37(19): 5560-5571.

[14] 王小君，窦嘉铭，刘嬰，等．可解释人工智能在电力系统中的应用综述与展望[J]．电力系统自动化，2024，48(4)：169-191.
WANG Xiaojun, DOU Jiaming, LIU Zhao, et al. Review and prospect of explainable artificial intelligence and its application in power systems [J]. Automation of Electric Power Systems, 2024, 48(4): 169-191.

[15] 向英，韩玄．电力行业人工智能技术应用的网络安全风险分析[J]．信息安全与通信保密，2023，21(10)：67-74.
XIANG Ying, HAN Xuan. Cyber security risk analysis of artificial intelligence technology application in the power industry [J]. Information Security and Communications Privacy, 2023, 21(10): 67-74.

[16] 赵俊华，文福拴，黄建伟，等．基于大语言模型的电力系统通用人工智能展望：理论与应用[J]．电力系统自动化，2024，48(6)：13-28.
ZHAO Junhua, WEN Fushuan, HUANG Jianwei, et al. Prospect of artificial general intelligence for power systems based on large language model: theory and applications [J]. Automation of Electric Power Systems, 2024, 48(6): 13-28.

[17] 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.

[18] 蒲天骄，张中浩，谈元鹏，等．电力人工智能技术理论基础与发展展望(二)：自主学习与应用初探[J]．中国电机工程学报，2023，43(10)：3705-3718.
PU Tianjiao, ZHANG Zhonghao, TAN Yuanpeng, et al. Theoretical foundation and directions of electric power artificial intelligence (Ⅱ): self-directed learning and preliminary application[J]. Proceedings of the CSEE, 2023, 43(10): 3705-3718.

[19] 王钰．电网数据采集中的噪声分析与抑制[D]．哈尔滨：哈尔滨工业大学，2011.
WANG Yu. Noise analysis and suppression in power grid data acquisition[D]. Harbin: Harbin Institute of Technology, 2011.

[20] 蒲天骄，乔骥，韩笑，等．人工智能技术在电力设备运维检修中的研究及应用[J]．高电压技术，2020，46(2)：369-383.
PU Tianjiao, QIAO Ji, HAN Xiao, et al. Research and application of artificial intelligence in operation and maintenance

for power equipment[J]. High Voltage Engineering, 2020, 46(2): 369-383.

[21] 杨挺,耿毅男,郭经红,等.人工智能在新型电力系统智能传感、通信与数据处理领域应用[J].高电压技术,2024,50(1):19-29.
YANG Ting, GENG Yinan, GUO Jinghong, et al. Applications of artificial intelligence in sensing, communication, and data processing in the new power system[J]. High Voltage Engineering, 2024, 50(1): 19-29.

[22] 覃柳芸.基于迁移学习的电力系统暂态稳定自适应评估[D].北京:北京交通大学,2022.
QIN Liuyun. Adaptive assessment of power system transient stability based on transfer learning[D]. Beijing: Beijing Jiaotong University, 2022.

[23] 周念成,廖建权,王强钢,等.深度学习在智能电网中的应用现状分析与展望[J].电力系统自动化,2019,43(4):180-191.
ZHOU Niancheng, LIAO Jianquan, WANG Qianggang, et al. Analysis and prospect of deep learning application in smart grid [J]. Automation of Electric Power Systems, 2019, 43(4): 180-191.

[24] 卫志农,李超凡,丁爱飞,等.基于Tri-training-SSAE半监督学习算法的电力系统暂态稳定评估[J].电力自动化设备,2023,43(7):110-116.
WEI Zhinong, LI Chaofan, DING Aifei, et al. Power system transient stability assessment based on tri-training-SSAE semi supervised learning algorithm [J]. Electric Power Automation Equipment, 2023, 43(7): 110-116.

[25] 李欣,柳圣池,李新宇,等.基于CBAM-CNN的电力系统暂态电压稳定评估[J].电力系统及其自动化学报,2024,36(4):59-67.
LI Xin, LIU Shengchi, LI Xinyu, et al. CBAM-CNN-based short-term voltage stability assessment of power systems [J]. Proceedings of the CSU-EPSA, 2024, 36(4): 59-67.

[26] 武宇翔,韩肖清,牛哲文,等.基于变权重随机森林的暂态稳定评估方法及其可解释性分析[J].电力系统自动化,2023,47(14):93-104.
WU Yuxiang, HAN Xiaoqing, NIU Zhewen, et al. Transient stability assessment method based on variable weight random forest and its interpretability analysis[J]. Automation of Electric Power Systems, 2023, 47(14): 93-104.

[27] 王彤,刘九良,朱劭璇,等.基于随机森林的电力系统暂态稳定评估与紧急控制策略[J].电网技术,2020,44(12):4694-4701.
WANG Tong, LIU Jiuliang, ZHU Shaoxuan, et al. Transient stability assessment and emergency control strategy based on random forest in power system[J]. Power System Technology, 2020, 44(12): 4694-4701.

[28] XIE J, SUN W. A transfer and deep learning-based method for online frequency stability assessment and control [J]. IEEE Access, 2001, 9: 75712-75721.

[29] 钱倍奇,陈谦,张政伟,等.基于异构数据特征级融合的多任务暂态稳定评估[J].电力系统自动化,2023,47(9):118-128.
QIAN Beiqi, CHEN Qian, ZHANG Zhengwei, et al. Multi-task transient stability assessment based on feature-level fusion of heterogeneous data [J]. Automation of Electric Power Systems, 2023, 47(9): 118-128.

[30] 杨雨昕,姚伟,邓贤哲,等.基于Koopman时序延拓和CNN-Transformer模型的频率稳定指标预测[J/OL].电网技术:1-15[2024-06-14]. https://www.cnki.com.cn/Article/CJFDTotal-DWJS20240614004.htm.
YANG Yuxin, YAO Wei, DENG Xianzhe, et al. Frequency stability index prediction based on Koopman time series extension and CNN-Transformer model [J]. Power System Technology: 1-15 [2024-06-14]. https://www.cnki.com.cn/Article/CJFDTotal-DWJS20240614004.htm.

[31] 李宝琴,吴俊勇,张若愚,等.融合多类型深度迁移学习的电力系统暂态稳定自适应评估[J].电力自动化设备,2023,43(1):184-192.
LI Baoqin, WU Junyong, ZHANG Ruoyu, et al. Adaptive assessment of transient stability for power system based on transfer multi-type of deep learning model[J]. Electric Power Automation Equipment, 2023, 43(1): 184-192.

[32] 周诚诚.第二类B-样条权函数神经网络的算法复杂度研究及应用[D].南京:南京邮电大学,2013.
ZHOU Chengcheng. Research and application of algorithm complexity of the second kind of B-spline weight function neural network [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2013.

[33] 罗进军.非高斯Alpha稳定分布噪声下的信号检测方法研究[D].长沙:国防科技大学,2018.
LUO Jinjun. Research on signal detection method under non-Gaussian Alpha stable distribution noise [D]. Changsha: National University of Defense Technology, 2018.

[34] 李楠,张帅,胡禹先,等.一种基于深度自适应网络迁移的暂稳评估模型更新框架[J].电力系统保护与控制,2024,52(14):25-35.
LI Nan, ZHANG Shuai, HU Yuxian, et al. An updating framework of a model for transient stability assessment based on a deep adaptive network transfer[J]. Power System Protection and Control, 2024, 52(14): 25-35.

[35] 欧阳福莲,王俊,周杭霞.基于改进迁移学习和多尺度CNN-BiLSTM-Attention的短期电力负荷预测方法[J].电力系统保护与控制,2023,51(2):132-140.
OUYANG Fulian, WANG Jun, ZHOU Hangxia. Short-term power load forecasting method based on improved hierarchical transfer learning and multi-scale CNN-BiLSTM-Attention[J]. Power System Protection and Control, 2023, 51(2): 132-140.

[36] 李宝琴,吴俊勇,李栌苏,等.基于主动迁移学习的电力系统暂态稳定自适应评估[J].电力系统自动化,2023,47(4):121-132.
LI Baoqin, WU Junyong, LI Lusu, et al. Adaptive assessment of power system transient stability based on active transfer learning[J]. Automation of Electric Power Systems, 2023, 47(4): 121-132.

[37] 李保罗,孙华东,张恒旭,等.基于两阶段迁移学习的电力系统暂态稳定评估框架[J].电力系统自动化,2022,46(17):

176-185.
LI Baoluo, SUN Huadong, ZHANG Hengxu, et al. Transient stability assessment framework of power system based on two-stage transfer learning [J]. Automation of Electric Power Systems, 2022, 46(17): 176-185.

[38] TRAN A, NGUYEN C, HASSNER T. Transferability and hardness of supervised classification tasks [C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, South Korea: 1395-1405.

[39] BAO Y J, LI Y, HUANG S L, et al. An information-theoretic approach to transferability in task transfer learning [C]// 2019 IEEE International Conference on Image Processing (ICIP), September 22-25, 2019, Taipei, China: 2309-2313.

[40] NGUYEN C, HASSNER T, SEEGER M, et al. LEEP: a new measure to evaluate transferability of learned representations [C]// Proceedings of the 37th International Conference on Machine Learning, April 26-30, 2020, Vienna, Austria: 7294-7305.

[41] TAN Y, LI Y, HUANG S L. OTCE: a transferability metric for cross-domain cross-task representations [C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, USA: 15774-15783.

[42] YOU K, LIU Y, WANG J, et al. LogME: practical assessment of pre-trained models for transfer learning [EB/OL]. [2023-12-20]. https://arxiv.org/abs/2102.11005v2.

[43] AGOSTINELLI A, PÁNDY M, UIJLINGS J, et al. How stable are transferability metrics evaluations? [M]// Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 303-321.

[44] ALTMANN A, TOLOŞI L, SANDER O, et al. Permutation importance: a corrected feature importance measure [J]. Bioinformatics, 2010, 26(10): 1340-1347.

[45] RIBEIRO M, SINGH S, GUESTRIN C. "Why should I trust you?": explaining the predictions of any classifier [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, August 13-17, 2016, San Diego, USA: 1135-1144.

[46] LUNDBERG S, LEE S. 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.

[47] 慈铁军. 基于决策者偏好的区间数多属性决策方法研究[D]. 天津：河北工业大学，2014.
CI Tiejun. Research on interval number multi-attribute decision-making method based on decision maker's preference [D]. Tianjin: Hebei University of Technology, 2014.

[48] 林爽. 城市污水处理厂MBR工艺综合评价研究[D]. 北京：清华大学，2015.
LIN Shuang. Study on comprehensive evaluation of MBR process in urban sewage treatment plant [D]. Beijing: Tsinghua University, 2015.

[49] 赖荣燊，林文广，吴永明. 面向绿色性能优化的产品族模块再设计优先次序识别[J]. 中国机械工程，2019，30(11)：1329-1335.
LAI Rongshen, LIN Wenguang, WU Yongming. Redesign priority identification of product family modules for green performance optimization [J]. China Mechanical Engineering, 2019, 30(11): 1329-1335.

[50] LI B L, XU S Y, SUN H D, et al. System strength assessment based on multi-task learning [J]. CSEE Journal of Power and Energy Systems, 2024, 10(1): 41-50.

[51] 王渝红，李晨鑫，周旭，等. 压缩感知和图卷积神经网络相结合的宽频振荡扰动源定位方法[J]. 高电压技术，2024，50(3)：1080-1089.
WANG Yuhong, LI Chenxin, ZHOU Xu, et al. Localization method of wide-band oscillation disturbance sources based on compressed sensing and graph convolutional neural networks [J]. High Voltage Engineering, 2024, 50(3): 1080-1089.

[52] 蒋奇良，王渝红，史云翔，等. 基于压缩感知与ISTA的宽频振荡扰动源分级定位方法[J]. 高电压技术，2024，50(8)：3725-3735.
JIANG Qiliang, WANG Yuhong, SHI Yunxiang, et al. Hierarchical localization method of broadband oscillation disturbance source based on compressive sensing and ISTA [J]. High Voltage Engineering, 2024, 50(8): 3725-3735.

[53] 黄济宇，管霖，郭梦轩，等. 图深度学习技术在智能暂态稳定评估中的应用及展望[J]. 电网技术，2023，47(4)：1500-1511.
HUANG Jiyu, GUAN Lin, GUO Mengxuan, et al. Application and prospect of graph deep learning technique in intelligent transient stability assessment [J]. Power System Technology, 2023, 47(4): 1500-1511.

[54] 乔骥，赵紫璇，王晓辉，等. 面向电力系统智能分析的机器学习可解释性方法研究(二)：电网稳定分析的物理内嵌式机器学习[J]. 中国电机工程学报，2023，43(23)：9046-9059.
QIAO Ji, ZHAO Zixuan, WANG Xiaohui, et al. Research on interpretable methods of machine learning applied in intelligent analysis of power system (Ⅱ): physics-embedded machine learning for power system stability analysis [J]. Proceedings of the CSEE, 2023, 43(23): 9046-9059.

刘慧玉(2001—)，女，硕士研究生，主要研究方向：电力系统智能判稳可信度。E-mail：liuhy0521@163.com

王渝红(1971—)，女，博士，教授，博士生导师，主要研究方向：高压直流输电、电力系统稳定与控制、新能源并网。E-mail：yuhongwang@scu.edu.cn

石　访(1982—)，男，通信作者，博士，副教授，主要研究方向：电力系统稳定分析与控制、新型电力系统同步测量技术与应用。E-mail：shifang@sdu.edu.cn

（编辑　章黎）

Construction and Comprehensive Evaluation Method for Trustworthiness Indicator System of Intelligent Assessment Model for Power Grid Stability

LIU Huiyu[1], *WANG Yuhong*[2], *SHI Fang*[1], *ZHOU Xu*[2], *LI Baoluo*[1,3], *JI Kaixuan*[3]

(1. Key Laboratory of Power System Intelligent Dispatch and Control of Ministry of Education (Shandong University), Jinan 250061, China;
2. College of Electrical Engineering, Sichuan University, Chengdu 610065, China;
3. China Electric Power Research Institute Co., Ltd., Beijing 100192, China)

Abstract: The untrustworthiness issues of artificial intelligence (AI) algorithm hinder its practical application in the scenarios such as power grid stability analysis and control. At present, there are no specific trustworthiness evaluation indicators applicable to the intelligent assessment model for power grid stability. Regarding the characteristics of the electric power industry, the trustworthiness evaluation for AI stability assessment models is carried out, and the five sub-indicators of correctness, complexity, robustness, transferability and interpretability are chosen to construct the trustworthiness evaluation indicator system of the intelligent assessment model for power grid stability, and the specific calculation methods for each sub-indicator are summarized. Meanwhile, the fuzzy analytic hierarchy process is introduced and combined with subjective and objective evaluation to determine the weights of each sub-indicator and calculate the comprehensive trustworthiness indicator. Finally, by taking the voltage support strength, broadband oscillation, and frequency stability assessment models as a case, the trustworthiness evaluation and analysis are carried out, and the results verify the effectiveness of the proposed method.

This work is supported by National Key R&D Program of China (No. 2021YFB2400800) and State Grid Corporation of China (No. SGSDDK00WJJS2200092).

Key words: artificial intelligence; power grid stability; trustworthiness; transferability; interpretability; fuzzy analytic hierarchy process