第 13 卷 第 6 期
2024 年 11 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 13 No. 6
Nov. 2024

引文格式：
何晓曦, 蔡云鹏. 人工智能可解释性的研究现况及在医学领域的应用效果评测 [J]. 集成技术, 2024, 13(6): 76-89.
He XX, Cai YP. Current research status of explainability in artificial intelligence and evaluation of its application effects in medical fields [J]. Journal of Integration Technology, 2024, 13(6): 76-89.

人工智能可解释性的研究现况及在医学领域的应用效果评测

何晓曦 蔡云鹏*

（中国科学院深圳先进技术研究院 深圳 518055）

摘 要 人工智能可解释性指人们理解和解释机器学习模型决策过程的能力。该领域的研究旨在提高机器学习算法的透明度，使其决策更加可信和可解释。可解释性在人工智能系统中至关重要，尤其是在医疗、金融和法律等敏感和关键的决策领域。提供可解释性有助于人们更好地理解模型决策背后的逻辑推理，从而确保其决策过程的公正性和稳健性，并符合伦理标准。在不断发展的人工智能领域，提高模型可解释性是实现可信、可持续发展人工智能的关键一步。该文概述了人工智能可解释的发展历史和各种可解释方法的技术特点，在医疗领域的可解释性方面进行了更深入的探讨。此外，该文还对当前各种人工智能可解释方法在医学影像数据集上的局限性进行了分析，并提出了未来可能的探索方向。

关键词 人工智能；可解释性；医学影像；医患交互；可解释方法评估
中图分类号 TP30 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20240312001

Current Research Status of Explainability in Artificial Intelligence and Evaluation of Its Application Effects in Medical Fields

HE Xiaoxi CAI Yunpeng*

(*Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)

*Corresponding Author: *yp.cai@siat.ac.cn*

Abstract Artificial intelligence interpretability refers to the ability of people to understand and interpret the decision-making process of machine learning models. Research in this field aims to improve the transparency of machine learning algorithms, making their decisions more trustworthy and explainable. Interpretability is crucial in artificial intelligence systems, especially in sensitive and critical decision-making domains such as

收稿日期：2023-03-12 **修回日期**：2024-04-03
基金项目：国家自然科学基金项目(U22A2041)
作者简介：何晓曦，硕士研究生，研究方向为人工智能可解释；蔡云鹏(通讯作者)，研究员，研究方向为人工智能可解释、生物信息学和医学健康大数据挖掘，E-mail：yp.cai@siat.ac.cn。

healthcare, finance, and law. By providing interpretability, people can better understand the reasoning behind the model's decisions, ensuring that they are fair, robust, and ethical. In the continuously evolving field of artificial intelligence, enhancing the interpretability of models is a key step towards achieving trustworthy and sustainable AI. This article outlines the development history of interpretable artificial intelligence and the technical characteristics of various interpretability methods, with a particular focus on interpretability in the medical field. It provides a more in-depth discussion of the limitations of current methods on medical imaging datasets and proposes possible future directions for exploration.

Keywords　artificial intelligence; explainability; medical imaging; doctor-patient interaction; evaluation of interpretable methods

Funding　This work is supported by National Natural Science Foundation of China (U22A2041)

1　引　言

随着人工智能(artificial intelligence，AI)技术的飞速发展，机器学习模型在各个领域的应用越来越广泛，从医疗保健到保险、金融和法律等，都展现出巨大潜力。然而，这些强大而复杂的机器学习系统做出的决策往往被视为黑盒，让人难以理解其内在决策过程，这在一些特定的使用场景中难以被接受。在当前人工智能爆炸性发展的时代，理解和信任 AI 系统的决策不仅仅是一种好奇心的满足，更是确保其在现实应用中能被广泛接受和适应的必要条件。可解释性方法旨在揭示 AI 系统内在的工作机制，使决策过程对用户、开发者和决策制定者透明可见。可解释性方法不仅为研究者提供了深入了解模型的机会，还为监督、调整和干预模型提供了重要手段，确保其决策符合道德、法律和社会的期望。然而，在探索 AI 可解释性的过程中，仍面临诸多挑战。随着计算能力的提升、网络层数的加深和数据规模的扩大，模型决策的解释变得越来越复杂，增加了理解和透明化的难度。此外，解决模型的不公平性、偏见和对抗攻击性等问题，以确保 AI 系统的可解释性，不仅仅是一种理论上的构想，更是能在实际应用中取得实质性成果的必要条件。

本文将深入探讨人工智能可解释性的现状、挑战和未来发展方向。首先，介绍人工智能可解释的概念，回顾可解释性方法的发展历史；其次，总结当前主流可解释方法的技术特点；最后，评测部分可解释方法在医学影像数据集上的表现。本文旨在为构建更加透明和可信赖的 AI 系统提供启示，并推动人工智能技术在更广泛的领域，尤其是医学领域取得更为显著的成功。

2　人工智能可解释的发展历史和目标

2.1　人工智能可解释的发展历史

可解释的人工智能(explainable artificial intelligence，XAI)[1]最先于 2016 年由美国国防高级研究计划局(DARPA)提出。与 XAI 相关的研究最早可追溯至 1982 年，当时，研究人员提出一种名为 Neocognitron 的人工神经网络[2]，该网络利用一种分层设计，通过多层学习的方式使计算机能逐渐“学会”识别视觉模式。通过多次重复激活的强化策略训练，该网络的性能得到逐步增强。由于该网络的层数较少，学习内容较

固定，因此在初步阶段具有一定的可解释性，可以将其看作深度学习可视化的起点。Garson[3]提出了基于统计结果的敏感性分析方法，从机器学习模型的结果进行分析，试图得到模型的可解释性。上述早期系统机器学习的规则集合提供了直观和易于理解的解释，然而，它们在处理复杂问题和大规模数据上表现出一定局限性，不能满足越来越复杂的需求。但是，上述类型的方法提出了人工智能可解释性的关键概念，为后来的研究者开辟了一条道路。

随着深度学习等复杂模型的广泛应用，对 AI 系统可解释性的需求越来越迫切。研究者开始寻找各种方法揭示深度学习模型的决策过程。从 21 世纪 10 年代中期至今，逐渐出现了一系列解释性方法，并发展出了若干分支。按照解释范围，解释性方法可分为局部解释和全局解释。局部可解释性方法通过在数据总体中表示单个输入数据的个体特征归因进行解释。全局可解释性方法提供了对整个模型决策的洞察，可理解一系列输入数据的归因。本文主要根据可解释模型背后核心算法的实现方法进行分类，并对各种可解释方法进行讨论。

2.2 人工智能可解释的目标

XAI 指在人工智能系统中，使机器学习模型的决策和行为能够被理解和解释的能力。在实际应用中，当一些领域涉及个体权利、隐私、伦理等重要问题决策时，理解和信任机器学习模型的工作原理显得尤为重要。随着深度学习应用范围的更加广泛，可解释的重要性也越发突显。基于深度学习的算法系统虽然在图像分类、语音识别、情绪分析等领域已经达到，甚至超过了人类的水平，但是由于其产生的结果不可解释，甚至不可控，因此在上述涉及个体权利、隐私、伦理等问题领域的实际生产部署中都受到了严重限制。由此可见，建立一套可解释的 AI 系统意义重大。XAI 的发展对推动人工智能的应用和接受度具有重要意义。解释性的模型更容易被广泛应用于医疗、金融、司法等领域，同时也能降低人们对人工智能系统的担忧和疑虑。在研究和开发人工智能技术时，关注可解释性不仅是一种技术追求，更是对社会责任和伦理的体现。本文结合孔祥维等[4]的说法，将可解释性目标的多样性进行了如下总结：

(1) 可信任度。可信任度指当面对一个决策时，人类认为模型性能的置信度高，具有鲁棒性和稳定性，同时还能产生可靠的、可信的解释。

(2) 伦理和社会考虑。从社会角度出发，模型要能确保提供的解释具有决策公平的能力。人工智能可解释性涉及一系列伦理和社会考虑，包括确保模型的决策不带有歧视、不侵犯隐私，以及在重要领域的决策中，能提供足够的解释和证据。这对模型应用与实际生产和避免社会问题至关重要。

(3) 反馈性。反馈性指用户能向模型提供反馈，且模型能根据反馈进行调整的能力。这种反馈循环不仅有助于改善模型的性能和减小误差，还有助于提高用户对模型的满意度。

(4) 透明性。可解释性要求机器学习模型的内部逻辑和决策过程是透明的。透明性使得用户和相关利益方能理解模型是如何得出某一决策的，这对建立信任和对模型进行调整至关重要。

(5) 可理解性。与透明性相关，强调模型的输出结果能以人类可理解的方式进行解释，包括解释模型中的特征重要性、影响决策的因素和模型对不同输入的响应。

2.3 医学领域对可解释人工智能的应用需求

在当今世界各地，预期寿命的延长、慢性病发病率的飙升和昂贵的新疗法的不断开发[5]促使人们在医疗保健领域投入越来越多的研究。人工智能有望通过改善医疗保健条件并使其更具成本效益来缓解这些趋势所带来的挑战与压力[6]。然而，人工智能虽然潜力较大，但它并不是一个通用的解决方案。医学领域往往要求相关从业人员

的态度认真、严肃，因为它与人的健康和生命息息相关，与其他领域相比，具有其特殊性，因此，其对可解释提出了较高的要求。

在人工智能可解释性的背景下，病灶定位指解释模型在诊断或预测中是如何对患者的病灶进行定位的过程。该过程对医学影像分析领域较为重要，因为医生和临床专业人员需了解 AI 模型是如何做出诊断或定位病灶的，以便对患者的状况做出正确的评估和决策。例如，在前列腺癌的诊断中，磁共振成像（magnetic resonance imaging，MRI）已被证明可以极大地改善前列腺癌的检测[7]，而且越来越多地用于指导医生的治疗方案，被认为是一种最敏感的非侵入性成像模式，能可视化检测和定位前列腺癌。与单独的超声引导系统活检相比，MRI-超声融合活检可改善临床意义的前列腺癌的检测[8-9]。对通过活检获得的前列腺组织进行病理学分析，可确定前列腺癌的存在和分级[8-11]。在非侵入性成像中，侵袭性癌症的位置和范围可帮助指导治疗决策，即是否进行根治性前列腺切除术或局部治疗或主动监测[8]。然而，MRI 上良性与癌性组织之间微妙的视觉差异经常导致放射科医生难以对诊断结果进行解释，此时，一个精确的、可解释的病灶定位方法就很重要。它不但能帮助医生更准确地完成疾病诊断，减少医疗事故的发生，还能提高医护人员的效率，释放医疗资源。

在医学领域，医生无时无刻都在面对与患者沟通的场景，从患者的角度来看，可解释性要关注一个问题：使用人工智能驱动的决策辅助工具是否符合以患者为中心的护理的内在价值观。以患者为中心的护理旨在响应并尊重患者个体的价值观和需求[12]，它将患者视为护理过程中的积极合作伙伴，强调患者的选择权和对医疗决策的控制权。以患者为中心的护理的一个关键组成部分是共同决策，旨在确定最适合患者个体情况的治疗方法[13-14]。以患者为中心的护理涉及患者和临床医生之间的公开对话，临床医生告知患者可用行动方案的潜在风险和益处，患者讨论他们的价值观和优先事项[15-16]。如果临床医生不能完全理解决策辅助的内部运作和计算，则将无法向患者解释某些结果或建议是如何得出的[17]。可解释性可通过为临床医生和患者提供基于患者个人特征和风险因素的个性化对话帮助解决上述问题。通过模拟不同治疗或生活方式干预措施的影响，可解释的人工智能决策辅助可帮助提高患者的选择意识，并支持临床医生了解患者的价值观和偏好[18]。

3　人工智能可解释方法研究现状

可解释的模型在许多决策场景中都很受青睐，因为它们不仅能提高模型产生结果的可靠性和用户识别能力，而且能帮助人们挖掘数据集中蕴藏的知识和内在规律。因此，近年来，人们对从机器学习模型中获得的解释或开发内在可解释模型[19]越来越感兴趣。从解释的范围来看，可解释方法可以是局部的或全局的，有些方法可以扩展到局部和全局。局部可解释方法着重于理解模型在特定输入附近的行为。与全局可解释性相比，局部可解释性关注模型在某个具体实例或一个小组实例上的解释，而不是在整个数据集上的解释。全局可解释的方法提供对模型作为一个整体决策的洞察，可加深人们对输入数据的属性的理解。除此之外，根据所使用的算法原理，还可将可解释方法归类为基于扰动的解释方法、基于梯度的解释方法、基于反事实生成的解释方法和基于知识蒸馏的解释方法。本章将会对可解释方法进行介绍和讨论。

3.1　基于扰动的解释方法研究进展

基于扰动的解释方法的基本思路是通过迭代改变输入数据，观察其对模型输出的影响，从而分析模型对不同输入特征的敏感性。这种方法通

过探测输入与输出之间的关系，帮助解释模型的决策过程和每个特征的重要性。这种方法可以在特征级别进行，通过将某些特征替换为 0 或随机的对抗实例，选择一个或一组像素(超像素)进行模糊、平移或屏蔽等操作。

(1) LIME[20]为了得到一个人类可理解的表示，通过评估在输入附近的局部区域内的模型预测的变化，确定哪些特征对模型的输出具有重要性，如图 1。该方法使用超像素(一组相邻的像素)构建可解释的局部模型，通过计算每个超像素的权重，生成一个二进制向量，表示在解释模型中，哪些超像素对特定的类别输出最重要。

(2) 在 Lundberg 等[21]提出的 SHAP 方法中，Shapley 值的计算基于博弈论中的 Shapley 值理论，该理论提供了一种公平分配合作价值的方法。在解释模型的预测结果时，SHAP 将数据特征视为博弈中的玩家，通过计算各个特征对最终输出预测的贡献，确定它们在共同博弈中的价值。SHAP 方法能更全面地解释模型对输入特征的预测过程，并以公平的方式为每个特征分配贡献值。它提供了一种基于博弈论的理论最优解，用于解释模型预测的特征贡献，可增强模型决策的可解释性。SHAP 方法同样得到了研究者的广泛关注，文献[22-25]对 SHAP 方法做了进一步的研究工作。

(3) Zeiler 等[26]通过遮挡输入图像的不同部分和使用反卷积网络(deconvolutional network，DeConvNet)，可视化深度卷积网络各层的神经激活。DeConvNet 用滤波器和反池化操作设计的卷积神经网络(convolutional neural network，CNN)，呈现与传统 CNN 相反的结果。因此，与减小特征维度相反，DeConvNet 用于创建一个激活图，该图映射回输入像素空间，从而创建神经(特征)活动的可视化。个别激活图能帮助人们理解深度模型的内部层是如何学习特征的，以及它们关注的具体内容，这允许人们对深度神经网络进行更精细的研究。

(4) RISE[27]方法通过将输入图像与随机掩码相乘来扰动输入图像，如图 2 所示。具体而言，对掩盖的图像进行输入，并捕获与个别图像相对应的显著性地图。然后，根据置信度分数加权平均掩码，从而得到最终的显著性地图。该图能突出显示个别预测中具有正值热图的区域。这些区域表明了模型在做出该预测时最为关注的图像部分，有助于解释模型的决策依据。

(5) Burns 等[28]引入的可解释性随机化测试(IRT)和一次性特征测试(OSFT)通过将特征替换为无信息的对立事实，来识别模型中的重要特征。Burns 等[28]提出了一种有趣的方法，通过使用假设检验框架建模特征替换，以检查上下文的重要性。然而，对于深度学习算法，由于预训练深度模型对输入尺寸有要求严格，无法简单地从输入中删除一个或多个特征。因此将特征值设为 0 或填充对立事实值可能会导致不理想的性能表现，因为特征之间可能存在相关性。

基于扰动的方法往往简单有效、适用性较

(a) Original Image (b) Explaining Electric guitar (c) Explaining Acoustic guitar (d) Explaining Labrador

图 1 LIME 解释结果可视化[20]

Fig. 1 LIME explanation visualization[20]

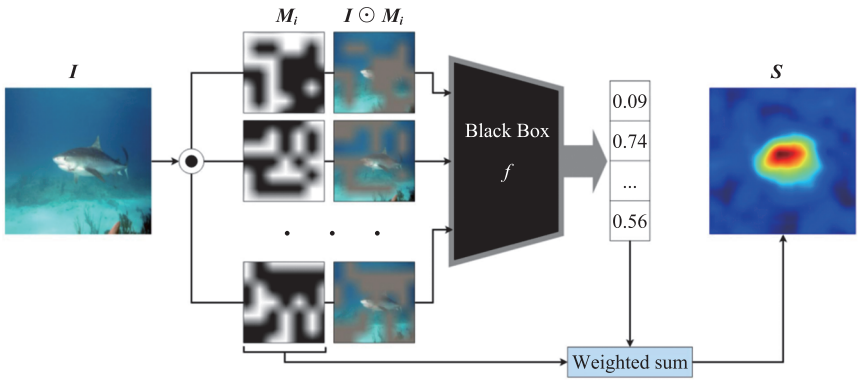

图 2　RISE 原理图[27]

Fig. 2　RISE schematic image[27]

广，可以与各种类型的模型结合使用，包括神经网络、决策树等，但是需要对输入进行多次扰动和评估，时间成本较高，这在一些对时间消耗有严格控制的场景下是不适用的。此外，基于扰动的方法通常假设模型在附近是光滑的，可能无法捕捉到对抗性扰动等非光滑情况下的模型行为。

3.2　基于反向传播或梯度的解释方法研究进展

在训练深度学习模型时，研究人员往往通过反向传播算法计算损失函数相对于模型参数的梯度。这些梯度表征了它们在参数空间中移动时损失函数的变化情况。反向传播算法通过计算输入相对于输出的梯度，将梯度映射回输入空间，产生一个热图或显著性图。这些图可表明输入的哪些部分对模型的决策贡献较大。

(1) 类激活映射 (class activation mapping，CAM)：大多数显著性方法使用全局平均池化层进行所有池化操作，而不是最大池化。Zhou 等[29]利用 CAM 的线性加权组合生成显著性图，强调图像空间中的重要部分。在 CAM 的基础上，又有研究者对 Grad-CAM[30]和 Grad-CAM++[31]做了进一步工作。Grad-CAM[30]和 Grad-CAM++[31]对更深的 CNN 进行了 CAM 操作的改进，与需要改变网络结构的 CAM 相比，后面的两种方法更灵活，但存在噪声较多、解释的细粒度不够等问题。

(2) Simonyan 等[32]引入了一种基于梯度方法生成的卷积网络显著性图。Zeiler 等[26]提到的 DeConvNet 作为一种扰动方法，使用反向传播进行激活可视化。DeConvNet 的工作之所以引人注目，是因为它在反向传播过程中根据梯度值分配了相对重要性。使用修正线性单元 (ReLU) 这种激活函数时，传统 CNN 的反向传播会将负梯度置为 0，表明当梯度为负时，ReLU 层将其完全忽略，从而阻止某些信息的反向传播。此时，DeConvNet 通过不同的处理方式，提供了更好的模型内部工作机制的可视化和解释。GuidedBP 方法[33-34]也是一种基于梯度的解释方法，是对 Simonyan 等[32]工作的改进。

(3) Li 等[35]提出的 Salient Relevance Maps 方法基于输入图像上下文感知显著性图。该方法通过逐像素计算模型的相关性得分，生成初步的相关性图。在此基础上，SRM 进一步通过多尺度分析和上下文感知的显著性检测来分离前景和背景区域，增强显著区域与非显著区域的对比度，使解释更直观。该方法旨在有效区分图像中的前景和背景层。

(4) Ancona 等[36]的研究表明，在梯度方法中，通过将输出的梯度与对应输入相乘，可生成用于解释模型结果的可解释性图。这种方法

有助于识别输入中对模型预测贡献较大的特征。然而，Sundararajan 等[37]提出了集成梯度(IG)，认为大多数基于梯度的方法在某些“axioms”上存在不足，而这些“axioms”是任何基于梯度的技术所期望的特征。Sundararajan 等[37]认为，DeepLift[38]、逐层相关传播(LRP)[39]、DeConvNet[26]和引导反向传播[33]等方法具有违反某些“axioms”的特定的反向传播逻辑。

基于梯度的解释方法的优点如下：首先，提供了视觉上直观的解释，可以通过梯度值或热力图显而易见地展示模型对输入的关注程度；其次，计算相对简单，通常仅需计算模型输出相对于输入的梯度，因此计算效率较高，适用于大规模数据和复杂模型；最后，适用于各种类型的模型，包括深度学习模型、线性模型和非参数模型等，因此适用性较广。基于梯度的解释方法的缺点如下：首先，通常仅能提供对输入特征的局部解释，不能全面解释模型的整体决策过程，这可能导致对模型行为的全面理解不足；其次，在对模型的复杂非线性关系进行线性近似的过程中，可能无法准确捕捉模型的真实行为；最后，它们对对抗性攻击具有敏感性，从而影响可解释性的稳定性和可信度。

3.3 基于反事实生成的方法

从某种角度来说，基于反事实生成[40-41]的方法也是一种基于扰动的方法，其工作于上下文感知的语义级特征(如超像素或词项)，而不是原始输入。此外，该方法利用对抗网络生成尊重数据分布的合成样本，超越了传统的启发式样本填充方法。Guidotti 等[42-43]通过对抗式自动编码器[44]生成范例和反范例，并通过编码器将数据映射到潜在空间，然后对潜在向量应用随机扰动，以解码作为范例和反范例的真实合成样本。文献[40-43]提出一类基于反事实生成的方法，通过观察预测结果如何在语义修改的合成样本之间转换，来解释机器学习模型产生的结果。

(1) Xie 等[45-46]提出一种可实现全局解释的方法，通过在学习的低维流形中创建传输路径实现解释。通过这些路径，可以探索和可视化区分规则背后的领域知识。将图像样本编码为一个向量空间，并将该向量空间分为两个子空间：一个样式子空间编码类相关特征，另一个背景子空间编码有条件独立于类的样本特征。将一个样本的背景代码与另一个样本的样式代码相结合，就会产生一个真实的合成样本。在提取的低维流形特征空间中，首先可视化分类决策模式，然后可通过连续修改样本的方式，在一条引导路径上移动样本的类关联编码，直到样本的分类结果发生变化，以实现对每个个体样本的解释。

(2) Singla 等[47]利用反事实解释，改善了预先训练的深度神经网络(DNN)的不确定性量化这一缺陷，在现有的反事实解释工作的基础上，提出了一个基于 StyleGANv2 的骨干网络。通过对增强数据进行微调，并结合软标签，可达到改善决策边界的目的。经过微调的模型可成功捕捉模糊样本、不可见的近分布外样本，以及标签转移情况下的不确定性，改善了模型高度准确但是过于自信的问题。

(3) Singla 等[48]的主要思路是开发一个基于 cGAN 的框架，通过生成查询图像逐渐变化的扰动，使得分类决策从对给定目标类别的负面转变为正面。同时，Singla 等[48]通过添加专门的重建损失(该损失结合了语义分割和异物检测网络的上下文)，保留生成图像中的解剖形状和异物(如支撑装置)。此外，这个模型很好地解释了心脏扩大和胸腔积液的影像结果，并得到了经验丰富的放射科住院医师的证实。

(4) 医学图像需要领域专家进行注释，这在实践中很难，但是与完整注释相比，图像级别标签的获取速度要快得多，而且难度要低很多。Tao 等[49]仅使用图像级别的标签构建了一个病变

分割模型。首先，病变分割模型使用图像级别标签训练图像分类器；其次，利用模型可视化工具根据训练的分类器为每个训练样本生成一个对象热图；最后，基于生成的热图（作为伪注释）和对抗性学习框架，构建和训练用于水肿区域分割（EAS）的图像生成器。病变分割模型结合了监督学习（具有病变感知）和对抗性训练（用于图像生成）的优点，在实现可解释的同时，解决了医学影像完整注释获取困难的问题。

反事实生成方法具有很强的灵活性，通常不依赖于特定类型的模型，能生成与实际预测结果相反的情境，可帮助理解模型决策背后的机制。此外，通过比较模型在预测为正样本和负样本下的行为，解释模型的决策过程，提供了较强的解释性。反事实生成方法还可通过构建反事实情景，探索数据中的潜在规律。有些反事实生成方法不会局限在二分类问题上，在一些多分类场景中也有不错的表现。但是也会存在计算复杂度高、依赖于特征表示和反事实生成的过程中可能引入偏差等问题。

3.4　基于知识蒸馏的方法

基于知识蒸馏的可解释方法是一种利用深度学习模型的预测结果和其对应的可解释模型之间的关系，提高深度学习模型的可解释性的方法。这种方法的基本思想是通过将一个复杂的黑盒深度学习模型的知识传递给一个可解释模型，使得原始模型的预测结果更容易理解和解释。

（1）生活中，黑盒风险评分模型通常是专有的或不透明的。Tan 等[50]提出了一种模型蒸馏和比较的方法，用于评估黑盒风险评分模型。为深入了解黑盒模型，Tan 等[50]将其视为教师，通过训练透明的学生模型，模仿黑盒模型分配的风险分数。将经过蒸馏训练的学生模型与在有标注的结果上进行训练的第二个未蒸馏的透明模型进行比较，并利用两个模型之间的差异洞察黑盒模型。黑盒风险评分模型的一个关键优势是无须事先知道要查找什么，而且对具有未知偏见来源的复杂实际数据最为有用。

（2）Frosst 等[51]描述了一种使用已训练的神经网络创建更加可解释的模型的方法，该方法以软决策树的形式呈现，通过随机梯度下降进行训练，将神经网络的预测结果作为更具信息性的目标。软决策树使用学到的过滤器基于输入示例做出分层决策，最终选择特定的静态概率分布作为其输出。与直接在数据上训练的模型相比，软决策树的泛化性能更好，但性能低于用于训练它的神经网络。

（3）Che 等[52]引入了一种称为可解释模仿学习的新型知识蒸馏方法，不仅能学习可解释的表型特征，实现强大的预测，还能模仿深度学习模型的性能。Che 等[52]引入的新型知识蒸馏方法利用梯度提升树，从深度学习模型（如堆叠去噪自动编码器和长短期记忆网络）中学习可解释的特征。Che 等[52]提出了一种方法，并通过对真实世界的临床时间序列数据集进行详尽实验，证明该方法在性能上达到甚至超过了深度学习模型的水平，为临床决策提供了可解释的结果。

（4）Liu 等[53]首次在多类别数据集上将深度神经网络蒸馏成普通决策树。为解决深度神经网络的复杂模型难以理解和推理预测结果的问题，Liu 等[53]应用知识蒸馏技术将深度神经网络蒸馏成决策树，以同时实现良好的解释性能。将模型要解决的问题表述为一个多输出回归问题后，实验证明，在相同深度的树水平上，学生模型的准确性显著优于普通决策树。

基于知识蒸馏的方法通过将复杂模型的知识转移到可解释模型中，使得原始模型的预测结果更易于理解和解释，有助于用户更好地理解模型的行为和决策过程。在知识蒸馏过程中，即使降低模型的复杂度，也可以保持较高的预测准确率。与原始的复杂模型相比，基于知识蒸馏的方法减少了模型训练和推理所需的计算资源。但在

知识蒸馏过程中，由于将复杂模型的知识压缩到可解释模型中可能会导致一定程度上的信息损失，因此导致如下问题：可解释模型无法完全捕捉原始模型的复杂性；可解释模型过度拟合训练数据，会降低模型本身的泛化能力；可解释模型依赖原始模型，如果原始模型的性能较差或者不稳定，则可能影响最终可解释模型的质量。

4 医学影像数据集上当前可解释方法的效果评测

本文第 3 节介绍了基于扰动、梯度、反事实生成和知识蒸馏的可解释方法，而如何设计合理的算法对各种可解释方法进行评估存在挑战性。虽然某些研究者在可解释方法评估方面已取得了一些进展，但目前还处于初步探索阶段。吴俊杰等[54]从算法维度讨论了深度学习中的可解释性问题与挑战，指出对解释所需保证的性质和定量描述仍然缺乏统一标准。这是一个高度开放的问题，现阶段仍需研究者在理论和实践中进一步探索，以寻求更为规范和一致的解决方案。

4.1 可解释方法评测

一般来讲，对解释的结果进行评估通常需要综合考虑多个因素，例如：评估解释的准确性，即解释是否真实反映人工智能系统的决策依据或行为原因；评估解释在不同情境下是否保持一致性，即相同的输入或任务下，解释是否一致；评估解释是否全面，即是否包含影响决策或行为的所有关键因素；评估解释的可信度，即用户对解释是否有信任感，可通过用户调查、实验等方式收集用户对解释的信任程度和满意度，并根据其结果评估可信度；评估解释生成的时间和资源消耗等。本文从各种可解释方法中选择了部分适用于影像分类，且结果直观的可解释方法进行效果展示和评测。本文使用的 BraTS2020[55-56]数据集是一个医学影像数据集，不仅包含多个医疗中心的多模态脑部 MRI 数据，还包含由专业认证的神经放射科医生提供的 3D 脑部 MRI 扫描和相应的地面真实标签。为生成训练数据，本文在三维坐标系的 z 轴上选择了特定坐标 80、82、84、86、88 和 90 处的特定切片，获得了每个患者训练集中的 6 个 2D 脑图像和 6 个相应的地面真实掩膜。本文使用了 BraTS2020 数据集中的 1 298 个脑图像。其中，1 005 个图像包含肿瘤，293 个图像为正常图像。实验过程中的所有数据可分为有病和无病两种。本文调研发现，并不是所有可解释方法均适用于生成显著图，因此，本文挑选了适合本次实验的 10 种可解释方法进行评测。首先在 BraTS2020 数据集上微调 ResNet50[57]模型，然后使用各种可解释方法定位病灶区域和生成红鹳的分割。此外，为对分割结果进行定量评估，本文还计算了各个分割及人工标注的交并比(intersection over union，IOU)和 DICE 系数。IOU 和 DICE 强调不同方面的性能。前者对边界框或区域之间的准确性更敏感，而 DICE 对目标大小和形状的变化更鲁棒。因此，结合使用 IOU 和 DICE 可提供更全面的性能评估，确保算法在各个方面都表现良好。

4.2 评测结果展示及现有方法在医学影像的局限性

本文分别在脑影像和鸟类影像上实验了各种可解释方法，效果如图 3 和图 4 所示，对应方法的评价指标结果如表 1 所示。基于 CAM 及其变种的可解释方法在医学影像上的解释精确度普遍较高，无论是在直观视觉上，还是在 DICE 和 IOU 上，均支持这一结论；基于梯度的方法处于中间水平；而基于 RISE 的方法表现最差，甚至没有将病灶区域标注出来(IOU 和 DICE 也是最低)。但是，当本文将这些方法用在自然影像时(如识别鸟类)，则会得到一些违反直觉的结论。例如：在脑肿瘤上表现最差的 RISE，当换到鸟影像上时，评价指标处于中间水平；基于 CAM

及其变种的方法在评价指标上也不再呈压倒性优势。这表明一些方法虽然在自然影像数据集上表现很好，但在医学影像上，性能会出现较大下降，表明医学影像数据集的特殊性。未来，在进行医学影像方面的工作时，可以留意上述问题。此外，本文还发现，与自然影像相比，绝大多数方法用在医学影像数据集上时(如 BraTS2020)，定位出来的病灶和专家标注的金标准之间，IOU 和 DICE 都出现了显著下降，表明这些可解释方法并不能很好地将病灶部位完整地区分开来，现有的方法在医学影像任务上存在语义解释性差等问题，也就是说，其在解释模型决策时，在语义方面的指向性不够强或不够准确。这可能涉及模型输出的解释与实际语义或领域知识之间的差距。在可解释性方面，了解假阳性或者假阴性的原因和模型是如何产生这种错误解释的，对医生和临床专业人员理解模型决策的不确定性和局限性至关重要。与此同时，假阳性和假阴性的解释还能为改进模型性能和提高医学应用中的信任度提供指导。为解决以上问题，可以考虑在如下方面进行改进：

(1) 多模态信息融合。对于涉及多模态输入(如图像、文本、数值等)的任务，综合考虑不同模态的信息，以提高解释的语义指向性。

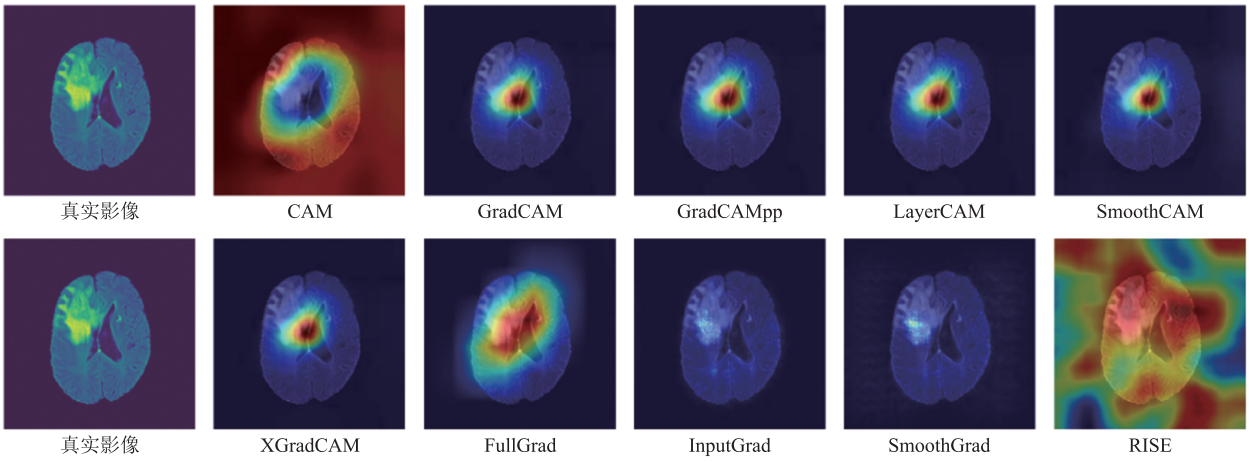

图 3　脑影像可解释方法效果图

Fig. 3　Effect diagram of brain imaging explainable methodology

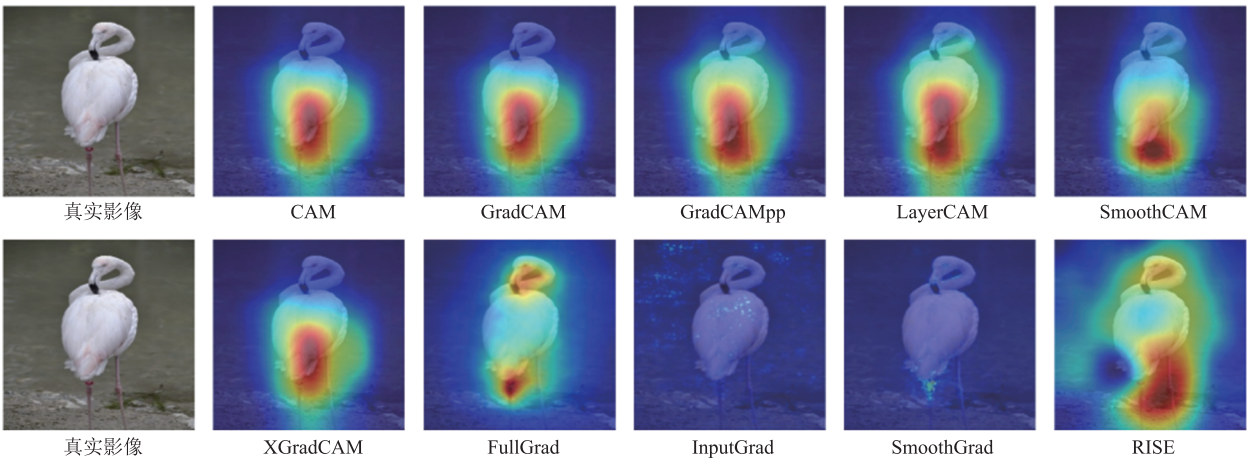

图 4　自然影像可解释方法效果图

Fig. 4　Effect diagram of natural image explainable method

表 1 各种可解释算法的 IOU 和 DICE

Table 1 IOU and DICE for various explainable algorithms

方法	IOU		DICE	
	大脑	火烈鸟	大脑	火烈鸟
CAM	0.292 1	0.402 3	0.452 2	0.573 7
GradCAM	0.304 8	0.402 3	0.467 1	0.573 7
GradCaMpp	0.303 5	0.445 1	0.465 7	0.616 1
LayerCAM	0.302 9	0.449 0	0.465 0	0.619 8
SmoothCAM	0.272 6	0.457 1	0.428 4	0.627 4
XGradCAM	0.305 2	0.402 3	0.457 7	0.573 7
FullGrad	0.281 8	0.693 8	0.439 7	0.812 3
InputGrad	0.179 6	0.147 6	0.304 5	0.257 3
SmoothGrad	0.185 2	0.116 8	0.312 5	0.209 2
RISE	0.174 7	0.406 3	0.297 5	0.577 9

(2) 领域专业知识的整合。不仅需要确保解释技术不仅仅是基于模型权重或梯度的数学解释，还需要考虑领域专业知识，即将领域专家的知识整合到解释中，以确保解释在语义上更为准确。

(3) 用户参与和反馈。引入用户参与和反馈机制，以便用户可以与解释进行交互，并提供关于解释准确性的反馈。这有助于不断改进解释技术，使其更符合用户的语义期望。

(4) 可解释性评估指标。制定并使用适当的可解释性评估指标，以量化解释技术在语义指向性方面的准确性，并与实际任务的语义一致性进行比较。

5 总结与展望

一个优秀的人工智能系统不仅应具备出色的智能表现，还应保持透明、可解释、可靠、公平、注重隐私保护、可持续等特性，并能与人类友好互动，同时具备社会责任感，确保其发展和应用符合伦理与社会价值观。盲目地相信预测性分类器的结果是不可取的，因为机器学习中存在数据偏差、可信性和对抗性示例的强烈影响。本文探讨了 XAI 至关重要的原因，涵盖了 XAI 的几个方面，并根据算法原理对它们进行了分类，以解释深度神经网络算法。此外，考虑到医学影像数据的特殊性，本文评测实验观察到：现有的可解释方法应用到医学影像数据集上时存在精度低、语义指向不明确等问题。针对这些局限性，本文提出一些改进建议，以为未来的研究提供参考。

总体而言，人工智能可解释性是一个备受关注的领域，研究人员正努力改进模型解释技术，使深度学习模型的决策过程更容易理解。可解释性工具正在不断发展，预计将在医疗、金融等领域得到广泛应用。本文认为，未来的人工智能可解释研究可从以下几个方面进行更深入的研究：

(1) 全局方法与局部方法结合。全局方法和局部方法各有优缺点，若将两种方法结合起来，则可以更全面地理解和解释人工智能系统的工作原理。先利用全局方法分析整个模型的行为，识别模型在不同情况下的工作模式和决策规律；再利用全局分析结果指导局部解释的生成，会使得局部解释的准确性和连贯性更好。

(2) 有导向的反事实生成。现有的可解释方法在进行反事实生成时，大多不能对生成的样本进行连续和有意识的修改。未来，研究人员可以考虑把数据集投影到一个有特征信息或者语义信息的流行空间中，通过在流行空间中操控，达到有导向的反事实生成解释。

(3) 交互式可解释。当前的 XAI 系统通常强调对“模型如何产生决策过程”的解释，不管用户有多少主动的输入或者互动，均仅影响机器“生成解释”的过程，不影响机器“做出决策”的过程，这是一种单向的价值目标对齐。未来，研究人员可以考虑在“双向价值对齐”方面进行更深入的研究，即通过与人交互，逐渐更新系统的价值函数，与人类的价值保持一致。

(4) 嵌入更多的外部知识。在当前的深度学习研究中，大多模型通过数据驱动的方式进行训

练和学习，较少关注知识驱动的观点。然而，将人类知识嵌入深度学习模型中，尤其是以知识图谱的形式，与深度学习技术相结合，构建具有解释性的深度学习模型，可以作为一个重要的研究方向。通过这种方式，可以利用人类积累的知识和规则，指导模型的学习过程，使模型更具有解释性和可理解性。

参 考 文 献

[1] Gunning D, Aha DW. DARPA's explainable artificial intelligence (XAI) program [J]. AI Magazine, 2019, 40(2): 44-58.

[2] Fukushima K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position [J]. Biological Cybernetics, 1980, 36(4): 193-202.

[3] Garson D G. Interpreting neural-network connection weights [J]. AI Expert, 1991, 6(4): 46-51.

[4] 孔祥维, 唐鑫泽, 王子明. 人工智能决策可解释性的研究综述 [J]. 系统工程理论与实践, 2021, 41(2): 524-536.
Kong XW, Tang XZ, Wang ZM. A survey of explainable artificial intelligence decision [J]. Systems Engineering-Theory & Practice, 2021, 41(2): 524-536.

[5] Amann J, Blasimme A, Vayena E, et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective [J]. BMC Medical Informatics and Decision Making, 2020, 20: 1-9.

[6] Higgins D, Madai VI. From bit to bedside: a practical framework for artificial intelligence product development in healthcare [J]. Advanced Intelligent Systems, 2020, 2(10): 2000052.

[7] Bhattacharya I, Khandwala YS, Vesal S, et al. A review of artificial intelligence in prostate cancer detection on imaging [J]. Therapeutic Advances in Urology, 2022, 14: 17562872221128791.

[8] Ahmed HU, Bosaily AES, Brown LC, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study [J]. The Lancet, 2017, 389(10071): 815-822.

[9] Bosaily AES, Parker C, Brown LC, et al. PROMIS—prostate MR imaging study: a paired validating cohort study evaluating the role of multi-parametric MRI in men with clinical suspicion of prostate cancer [J]. Contemporary Clinical Trials, 2015, 42: 26-40.

[10] Johnson DC, Raman SS, Mirak SA, et al. Detection of individual prostate cancer foci via multiparametric magnetic resonance imaging [J]. European Urology, 2019, 75(5): 712-720.

[11] Van Der Leest M, Cornel E, Israel B, et al. Head-to-head comparison of transrectal ultrasound-guided prostate biopsy versus multiparametric prostate resonance imaging with subsequent magnetic resonance-guided biopsy in biopsy-naïve men with elevated prostate-specific antigen: a large prospective multicenter clinical study [J]. European Urology, 2019, 75(4): 570-578.

[12] Committee on Quality of Health Care in America. Crossing the quality chasm: a new health system for the 21st century [M]. Washington (DC): National Academies Press, 2001.

[13] Barry MJ, Edgman-Levitan S. Shared decision making: the pinnacle of patient-centered care [J]. New England Journal of Medicine, 2012, 366(9): 780-781.

[14] Kunneman M, Montori VM, Castaneda-Guarderas A, et al. What is shared decision making? (and what it is not) [J]. Academic Emergency Medicine, 2016, 23(12): 1320-1324.

[15] O'neill ES, Grande SW, Sherman A, et al. Availability of patient decision aids for stroke prevention in atrial fibrillation: a systematic review [J]. American Heart Journal, 2017, 191: 1-11.

[16] Noseworthy PA, Brito JP, Kunneman M, et al. Shared decision-making in atrial fibrillation: navigating complex issues in partnership with the patient [J]. Journal of Interventional Cardiac Electrophysiology, 2019, 56(2): 159-163.

[17] Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making [J]. Philosophy & Technology, 2021, 34(2): 349-371.

[18] Politi MC, Dizon DS, Frosch DL, et al. Importance

of clarifying patients' desired role in shared decision making to match their level of engagement with their preferences [J]. BMJ, 2013, 347: f7066.

[19] Pintelas E, Livieris IE, Pintelas P. A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability [J]. Algorithms, 2020, 13(1): 17.

[20] Ribeiro MT, 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, 2016: 1135-1144.

[21] Lundberg SM, Lee SI. A unified approach to interpreting model predictions [C] // Proceedings of the 31st Conference on Neural Information Processing Systems, 2017: 1-10.

[22] Antwarg L, Miller RM, Shapira B, et al. Explaining anomalies detected by autoencoders using SHAP [Z/OL]. arXiv Preprint, arXiv: 1903.02407, 2019.

[23] Sundararajan M, Najmi A. The many Shapley values for model explanation [C] // Proceedings of the 37th International Conference on Machine Learning, 2020: 9269-9278.

[24] Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values [J]. Artificial Intelligence, 2021, 298: 103502.

[25] Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees [J]. Nature Machine Intelligence, 2020, 2(1): 56-67.

[26] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks [C] // Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, 2014: 818-833.

[27] Petsiuk V, Das A, Saenko K. RISE: randomized input sampling for explanation of black-box models [Z/OL]. arXiv Preprint, arXiv: 1806.07421, 2018.

[28] Burns C, Thomason J, Tansey W. Interpreting black box models via hypothesis testing [C] // Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference, 2020: 47-57.

[29] Zhou BL, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2921-2929.

[30] Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization [C] // Proceedings of the IEEE International Conference on Computer Vision, 2017: 618-626.

[31] ChattOpadhay A, Sarkar A, Howlader P, et al. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks [C] // Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018: 839-847.

[32] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps [Z/OL]. arXiv Preprint, arXiv: 1312.6034, 2013.

[33] Springenberg JT, Dosovitskiy A, Brox T, et al. Striving for simplicity: the all convolutional net [Z/OL]. arXiv Preprint, arXiv: 1412.6806, 2014.

[34] Mahendran A, Vedaldi A. Salient deconvolutional networks [C] // Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, 2016: 120-135.

[35] Li HY, Tian YK, Mueller K, et al. Beyond saliency: understanding convolutional neural networks from saliency prediction on layer-wise relevance propagation [J]. Image and Vision Computing, 2019, 83-84: 70-86.

[36] Ancona M, Ceolini E, Öztireli C, et al. Towards better understanding of gradient-based attribution methods for deep neural networks [Z/OL]. arXiv Preprint, arXiv: 1711.06104, 2017.

[37] Sundararajan M, Taly A, Yan QQ. Axiomatic attribution for deep networks [C] // Proceedings of the International Conference on Machine Learning, 2017: 3319-3328.

[38] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences [C] // Proceedings of the International Conference on Machine Learning, 2017: 3145-3153.

[39] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation [J]. PLoS One, 2015, 10(7): e0130140.

[40] Yu JH, Lin Z, Yang JM, et al. Generative image inpainting with contextual attention [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5505-5514.

[41] Chang CH, Creager E, Goldenberg A, et al. Explaining image classifiers by counterfactual generation [Z/OL]. arXiv Preprint, arXiv: 1807.08024, 2018.

[42] Guidotti R, Monreale A, Matwin S, et al. Explaining image classifiers generating exemplars and counter-exemplars from latent representations [C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 13665-13668.

[43] Guidotti R, Monreale A, Giannotti F, et al. Factual and counterfactual explanations for black box decision making [J]. IEEE Intelligent Systems, 2019, 34(6): 14-23.

[44] Makhzani A, Shlens J, Jaitly N, et al. Adversarial autoencoders [Z/OL]. arXiv Preprint, arXiv: 1511.05644, 2015.

[45] Xie RT, Chen JB, Jiang LM, et al. Active globally explainable learning for medical images via class association embedding and cyclic adversarial generation [Z/OL]. arXiv Preprint, arXiv: 2306.07306, 2023.

[46] Xie RT, Chen JB, Jiang LM, et al. Accurate explanation model for image classifiers using class association embedding [Z/OL]. arXiv Preprint, arXiv: 2406.07961, 2023.

[47] Singla S, Murali N, Arabshahi F, et al. Augmentation by counterfactual explanation-fixing an overconfident classifier [C] // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 4720-4730.

[48] Singla S, Eslami M, Pollack B, et al. Explaining the black-box smoothly: a counterfactual approach [J]. Medical Image Analysis, 2023, 84: 102721.

[49] Tao YH, Ma X, Zhang YZ, et al. LAGAN: lesion-aware generative adversarial networks for edema area segmentation in SD-OCT images [J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(5): 2432-2443.

[50] Tan S, Caruana R, Hooker G, et al. Distill-and-compare: auditing black-box models using transparent model distillation [C] // Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018: 303-310.

[51] Frosst N, Hinton G. Distilling a neural network into a soft decision tree [Z/OL]. arXiv Preprint, arXiv: 1711.09784, 2017.

[52] Che ZP, Purushotham S, Khemani R, et al. Distilling knowledge from deep networks with applications to healthcare domain [Z/OL]. arXiv Preprint, arXiv: 1512.03542, 2015.

[53] Liu X, Wang XG, Matwin S. Improving the interpretability of deep neural networks with knowledge distillation [C] // Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2018: 905-912.

[54] 吴俊杰, 刘冠男, 王静远, 等. 数据智能: 趋势与挑战 [J]. 系统工程理论与实践, 2020, 40(8): 2116-2149.
Wu JJ, Liu GN, Wang JY, et al. Data intelligence: trends and challenges [J]. Systems Engineering-Theory & Practice, 2020, 40(8): 2116-2149.

[55] Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS) [J]. IEEE Transactions on Medical Imaging, 2014, 34(10): 1993-2024.

[56] Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features [J]. Scientific Data, 2017, 4: 170117.

[57] He KM, Zhang XY, Ren SQ, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.