人工智能可解释性的研究现况及在医学领域应用效果评测

何晓曦，蔡云鹏*

（中国科学院深圳先进技术研究院，深圳，518000）

摘 要 人工智能可解释性是指人们能够理解和解释机器学习模型的决策过程。这一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。解释性在人工智能系统中至关重要，尤其是在对于决策敏感和关键的领域，如医疗、金融和法律。通过提供可解释性，人们可以更好地理解模型的推理基础，确保其决策是公正、健壮且符合伦理标准。在不断发展的人工智能领域中，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。文章梳理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，特别是在医疗领域的可解释性方面进行了更深入的探讨。对当前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尝试方向。

关键词 人工智能；可解释性；医学影像；医患交互；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 the Medical Field

He Xiaoxi, Cai Yunpeng*

（Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China)

Corresponding Author: Cai Yunpeng. ShenZhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518000, China. Email: yp.cai@siat.ac.cn

Abstarct Artificial intelligence interpretability refers to the ability of people to

来稿日期：2024-3-12 修回日期：2024-04-03
基金项目：国家自然科学基金（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，Email: yp.cai@siat.ac.cn

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 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. The 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

Funding National Natural Science Foundation of China

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愈发广泛，从医疗保健到保险、金融、法律等，都展现了巨大的潜力。然而，这些强大而复杂的系统所做出的决策往往被视为黑盒，让人难以理解其内在的决策过程，这在一些特定的使用场景中是难以接受的。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够被广泛接受和适应的必要条件。以上所言引发了人工智能可解释性的重要议题，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。这不仅为我们提供了对模型的深入了解，还为我们提供了一种监督、调整和干预的手段，以确保模型的决策符合道德、法律和社会的期望。在探索 AI 可解释性的道路上，我们面临着诸多挑战。随着算力的发展，越来越深的网络层数和海量的数据使得解释模型决策变得愈发复杂。此外，我们还需解决模

型的不公平型、偏见以及对抗攻击性等问题，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。

本综述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，最后评测了部分可解释方法在医学影像数据集上的表现。旨在为构建更加透明、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

可解释的人工智能来自英文 Explainable Artificial Intelligence[1]，缩写为 XAI。这一概念最先被提出是在 2016 年，由美国国防高级研究计划局（DARPA）提出，为了避免混淆，可解释人工智能的缩写应该被写作 XAI，其中的 X 代表 Explainable。此后 XAI 作为可解释人工智能的缩写被广泛接受。与可解释的人工智能相关的研究则可最早追溯与 1982 年，当时研究人员提出了一种名为 Neocognitron 的人工神经网络[2]，这个网络采用了一种分层设计，通过多层学习的方式使计算机能够逐渐“学会”识别视觉模式。通过多次重复激活的强化策略训练，网络的性能得到了逐步增强。由于层数相对较少，学习内容相对固定，这使得网络在初步阶段具有一定的可解释性能。可以将其看作深度学习可视化的起点。这篇文章提出了基于统计结果的敏感性分析方法[3]，从机器学习模型的结果对模型进行分析，试图得到模型的可解释性。早期的这些系统的规则集合提供了直观且易于理解的解释，然而，这些系统在处理复杂问题和大规模数据上表现有一定局限性，因此并不能满足后来越来越复杂的需求。但是，这些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研究道路。

随着深度学习等复杂模型的广泛应用，对于 AI 系统可解释性的需求愈发迫切。研究者开始寻找各种方法来揭示这些模型的决策过程。从 2010 年代中期至今，逐渐出现了一系列解释性技术，并发展出了若干分支方向。这些方法按照解释的范围不同，可以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来进行解释。全局可解释性模型提供了对整个模型决策的洞察，从而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，来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 是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、解释的能力。在实际应用中，在一些领域，比如涉及到个体权利、隐私、伦理等重要问题的决策时，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愈加广泛，可解释的重要性也越发突显。虽然基于深度学习的算法系统在诸如图像分类、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，所以其在上述许多领域的实际生产部署中都受到了严重的限制。由此可见，建立一套可解释的 AI 系统意义重大。XAI 的发展对于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，同时也能够降低人们对于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。本文结合文献[4]和其他文献中的说法，将可解释性目标的多样性进行了如下总结：

1）可信任度。可信任度是指，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，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到一系列伦理和社会考虑。这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，能够提供足够的解释和证据。这对于模型应用与实际生产以及避免社会问题是至关重要的。

3）反馈性。反馈性是指用户能够向模型提供反馈，并且模型能够根据反馈进行调整的能力。这种反馈循环有助于改善模型的性能和减小误差，同时也提高了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够理解模型是如何得出某一决策的，这对于建立信任和对模型进行调整都至关重要。

5）可理解性。与透明性相关，可理解性强调的是模型的输出结果能够以人类可理解的方式进行解释。这包括了解释模型中的特征重要性、影响决策的因素以及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以及昂贵的新疗法的不断开发[5]促成人们在医疗保健领域投入越来越多的研究目光。人工智能有望通过改善医疗保健并使其更具成本效益来减轻这些发展的影响[6]。然而，尽管人工智能具有不可否认的潜力，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有认真而且严肃的态度，因为它与人的健康和生命息息相关，在与其他领域相比时有其特殊性，所以这些特点使其对可解释提出了更高的要求。

在人工智能可解释性的背景下，病灶定位是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。这一过程对于医学影像分析等领域特别重要，因为医生和临床专业人员需要了解 AI 模型是如何做出诊断或病灶定位的，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（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。它被认为是最敏感的非侵入性成像模式，能够可视化、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组织病理学分析，以确定前列腺癌的存在和分级[8-11]。非侵入性成像中侵袭性癌症的位置和范围也可以帮助指导治疗决策，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 上良性和癌性组织之间微妙的视觉差异经常使得放射科医生对图像的解释具有挑战性，这个时候一个精确地的，可解释的病灶定位方法就有很重要的价值。它不但可以帮助医生更准确的完成疾病诊断，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着与患者的沟通，从患者角度看待可解释性就要关注一个问题：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，强调他们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]。它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]。如果临床医生不能够完全理解决策辅助的内部运作和计算，他们无法向患者解释某些结果或建议是如何得出的[17]。可解释性可以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来解决这个问题。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以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，因为它们不仅提高了可靠性和用户识别能力，而且还可以帮助知识的发现。因此，近年来人们对从机器学习模型中获得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，这些方法可以是局部的或全局的。有些方法可以扩展到这两者。局部可解释方法着重于理解模型在特定输入附近的行为。相比于全局可解释性，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了对模型作为一个整体的决策的洞察，从而加深了我们对输入数据数组的属性的理解。除此之外，根据可解释方法背后使用的算法原理的不同，还可以将这些方法归于基于扰动的方法、基于梯度的方法、基于反事实生成的方法和基于知识蒸馏的方法。在本节中，将会对这些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探测具有不同变化的输入观察输出变化来进行解释。这些扰动可以在特征级别进行，通过将某些特征替换为零或随机的对抗实例，选择一个或一组像素（超像素），进行模糊、平移或屏蔽操作等。

1）LIME[20]（*Local Interpretable Model-Agnostic Explanations*）这种方法为了得到一个对人类可理解的表示，通过评估在输入附近的局部区域内的模型预测的变化来确定哪些特征对于模型的输出具有重要性。它使用超像素（一组相邻的像素）来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于特定的类别输出最重要。

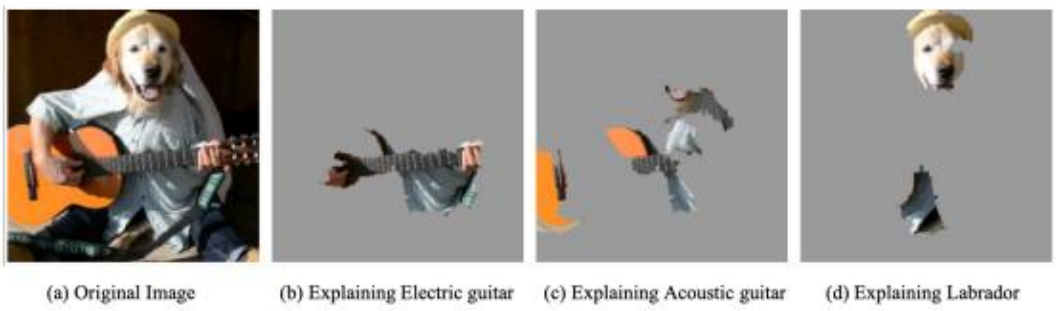

图 1[20]：LIME 效果图

Figure 1：LIME Explanation Visualization

2）SHAP[21]由 Lundberg 等人提出 。在这个方法中，Shapley 值的计算基于博弈论中的 Shapley 值理论，这个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，以确定它们在共同博弈中的价值。这种方法能够更全面地解释模型对于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，从而增强对模型决策的可解释性。SHAP 这一方法同样得到了研究者们的广泛关注，在这些文献中[22-25]，对这一方法做了进一步的研究工作。

3）Zeiler 等人[26]通过遮挡输入图像的不同部分并使用反卷积网络（DeConvNet）来可视化深度卷积网络各层的神经激活。DeConvNets 是使用滤波器和反池化操作设计的 CNN，以呈现与传统 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。个别激活图有助于理解感兴趣的深度模型的内部层学习的内容和方式，允许对深度神经网络进行精细研究。

4）RISE(Randomized Input Sampling for Explanation)[27]这种方法通过将输入图像与随机掩码相乘来扰动输入图像。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。根据置信度分数加权平均掩码，用于找到具有正值热图的个别预测的最终显著性地图。

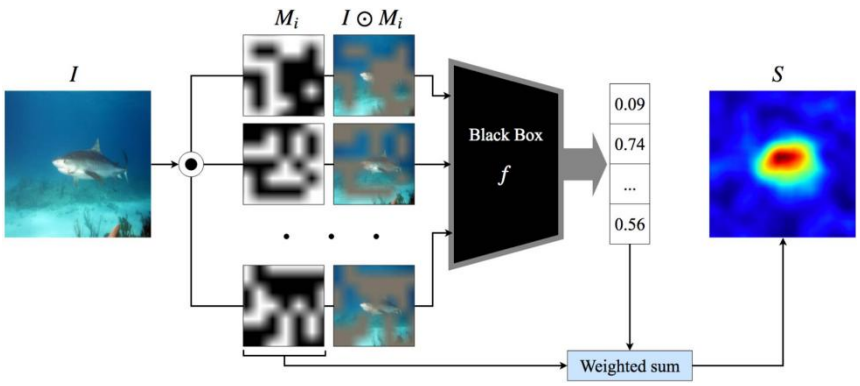

图 2[27]：对输入图像使用随机掩模进行扰动，找到单个掩蔽输入的置信度分数，使用加权函数生成显著图

Figure 2: Apply random masks to disturb the input image, find the confidence score of the masked input, and generate a saliency map using a weighted function

5）由 Burns 等人[28]引入的可解释性随机化测试（IRT）和一次性特征测试（OSFT）专注于通过用无信息的对立事实替换特征来发现重要特征。通过使用假设检验框架建模特征替换，作者展示了一种检查上下文重要性的有趣方式。不幸的是，对于深度学习算法，由于预

训练深度模型的严格输入尺寸，无法从输入中删除一个或多个特征。将值归零或填充对立事实值可能会导致不令人满意的性能，因为特征之间存在相关性。

基于扰动的方法往往简单有效，适用性广泛，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合使用的。而且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，通过反向传播算法计算损失函数相对于模型参数的梯度。这些梯度表示了在参数空间中移动时损失函数的变化情况。这些方法通过计算输入相对于输出的梯度，然后将梯度映射回输入空间，产生一个热图或显著性图。这些图表明输入的哪些部分对于模型的决策贡献较大。

1）Gradient 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 等人[29]通过使用类激活映射（CAM）的线性加权组合生成显著性图，以此强调图像空间中的重要部分。在 CAM 的基础上，后面又有研究者对此做了进一步的工作。Grad-CAM[30]和 Grad-CAM++[31]对于更深的 CNNs 进行了 CAM 操作的改进，相比于 CAM 需要改变网络结构，后面的这两种方法更加灵活，但是也具有噪声多，解释的细粒度不够等问题。

2）Simonyan 等人[32]引入了一种基于梯度的方法来生成卷积网络的显著性图。Zeiler 等人之前提到的 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，是因为在反向传播过程中相对重要性给予了梯度值。在使用修正线性单元（ReLU）激活时，对传统 CNN 进行反向传播会导致负梯度的零值。然而，在 DeConvNets 中，梯度值不会被截断为零。这使得能够进行准确的可视化。GuidedBP 方法[33, 34]也是基于梯度的解释方法的一类，这是对[32]工作的改进。

3）Salient Relevance (SR) Maps[35]这一方法是由 Li 等人提出的，这是一种基于输入图像的 L 的上下文感知显著性图。因此，第一步是使用相同的输入尺寸找到感兴趣的输入图像的相关性图。上下文感知显著性关联图算法获取相关性图并为各个像素找到显著性值。在这里，如果一组相邻像素在相同和多个尺度上与其他像素块不同，则像素是显著的。这是为了区分图像的背景和前景层。

4）Ancona 等人[36]表明在梯度方法中，其中对应于输入的输出梯度乘以输入，对生成可解释的模型结果解释是有用的。然而，在这篇文章中[37]，作者提出了集成梯度（IG）并认为大多数基于梯度的方法在某些“axioms”上存在不足，这些’axioms’是任何基于梯度的技术所期望的特征。作者认为诸如 DeepLift[38]、逐层相关传播（LRP）[39]、反卷积网络（DeConvNets）[26]和引导反向传播[33]等方法具有违反某些 axioms’’的特定反向传播逻辑。

基于梯度的方法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的展示模型对输入的关注程度；这类方法的计算相对简单，通常只需要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；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，因此具有广泛的适用性。但是，基于梯度的解释方法通常只能提供对输入特征的局部解释，而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；此外，对模型的复杂非线性关系进行线性近似的过程中，可能导致无法准确捕捉模型的真实行为；最后，它们可能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

反事实生成[40, 41]的方法从某种角度来说也是一种基于扰动的方法，它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，并使用生成对抗网络来生成尊重数据分布的合成样本，以这种方式超过启发式填充样本。最近的一种方法[42, 43]通过对抗式自动编码器[44]生成范例和反范例，该方法通过编码器将数据映射到潜在空间，并对潜在向量应用随机扰动来解码作为范例和反范例的真实合成样本。这些工作提出了一种方法来解释机器学习模型，通过观察预测结果如何在语义修改的合成样本之间转换。

1）Xie[45, 46]等人提出了一种可以实现全局解释的方法，该方法通过在学习的低维流形中创建传输路径来实现解释，允许将样本转换为模拟样本，查看不同的类，但保留其背景特征，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，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和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，可视化分类决策模式。然后可以通过连续修改样本的方式，在一个方向上移动其类关联代码沿着引导路径，直到其分类结果发生变化，以实现对每个个体样本的解释。

2）这篇文章[47]利用反事实解释来改善预先训练的深度神经网络（DNN）的不确定性量化。在现有的反事实解释工作基础上进行了改进，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，有助于改善决策边界。经过微调的模型，可以成功地捕捉模糊样本、未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 等人[48]使用了反事实解释作为审核给定黑盒分类器并评估该分类器使用的放射图像特征是否具有任何临床相关性的方法。主要思路是开发了一个基于 cGAN 的框架来生成查询图像的逐渐变化的扰动，使得分类决策对于给定的目标类别从负面变为正面。此外作者通过添加专门的重建损失来保留生成图像中的解剖形状和异物（例如支撑装置），该损失结合了语义分割和异物检测网络的上下文。特别值得注意的是，所提出的模型很好地解释了心脏扩大和胸腔积液的决定，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，并且在实践中很难，但是与完整注释相比，图像级别标签的获取速度和难度要快得多。这篇文章[49]仅使用图像级别的标签构建了一个病变分割模型。具体来说，首先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。它结合了监督学习（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。

反事实生成方法同样具有很强的灵活性，通常不依赖于特定类型的模型，能够生成与实际预测相反的情况，并通过比较模型在两种情况下的行为来解释模型的决策过程，从而提供了较强的解释性；还可以通过构建反事实情景，探索潜在模式；有些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。但是也会存在着计算复杂度高，依赖于特征表示，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，来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，从而使得原始模型的预测结果更容易理解和解释。

1）黑盒风险评分模型在我们的生活中通常是专有或不透明的。Tan 等人[50]提出了一种模型蒸馏和比较的方法，用于审计这类模型。为了深入了解黑盒模型，作者将其视为教师，训练透明的学生模型来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来洞察黑盒模型。使用这种方法的一个关键优势是无需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 等人[51]描述了一种使用已训练的神经网络创建更可解释模型的方法，该模型以软决策树的形式呈现，并通过随机梯度下降进行训练，使用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。这种软决策树比直接在数据上训练的模型更具泛化性能，但性能低于用于训练它的神经网络。

3）Che 等人[52]引入了一种称为可解释模仿学习（Interpretable Mimic Learning）的新型知识蒸馏方法，以学习可解释的表型特征，实现强大的预测，并模仿深度学习模型的性能。这种框架使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时记忆网络）中学习可解释的特征。作者提出，对于一个真实世界的临床时间序列数据集的详尽实验证明，此方法在性能上达到了与深度学习模型相似或更好的水平，并为临床决策提供了可解释的表型。

4）这篇文章[53]是第一个在多类别数据集上将深度神经网络蒸馏成普通决策树的研究。为了解决深度神经网络的复杂模型使其难以理解和推理预测结果的问题，作者应用知识蒸馏技术将深度神经网络蒸馏成决策树，以同时实现良好的性能和解释性。将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，学生模型的准确性性能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。这有助于用户更好地理解模型的行为和决策过程；在知识蒸馏的过程中，即使降低了模型的复杂度，也可以保持较高的预测准确率；相比于原始的复杂模型，减少了模型的训练和推理所需的计算资源。但是在知识蒸馏的过程中，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，使得可解释模型无法完全捕捉原始模型的复杂性；模型过度拟合训练数据，会降低泛化能力；依赖原始模型，如果原始模型的性能较差或者不稳定，可能会影响到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

在上面的一节中，我们介绍了基于扰动，基于梯度等四种可解释方法，接下来提出的一个挑战是如何设计合理的算法对各种可解释方法进行评估。虽然研究者在可解释方法评估这一块已经取得了一定的进展，但是目前还处于一个初步的探索阶段。这篇文章[54]从算法维度讨论了深度学习的可解释问题与挑战，也提出对解释需要保证的性质和定量描述缺乏统一的标准，是一个高度开放的问题，还需研究者们继续努力。

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，比如评估解释的准确性，即解释是否真实反映了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，即是否包含了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，可以通过用户调查、实验等方式收集用户对解释的信任程度和满意度来评估可信度；评估解释生成的时间和资源消耗等等。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了效果展示和评测。我们使用了 BraTS2020[55, 56]数据集，BraTS2020 是一个医学影像数据集，包含了来自多个医疗中心的多模态脑部磁共振成像（MRI）数据。数据集包含了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。为了生成训练数据，我们在 z 轴上选择了特定坐标 80、82、84、86、88 和 90 处的特定切片，从而获得了每个患者训练集中的六个 2D 脑图像和六个相应的地面真实掩膜。对于 BraTS2020 数据集，我们在实验中使用了共计 1,298 个脑图像。其中，1,005 个图像包含肿瘤，而 293 个图像为正常图像。在实验过程中所有数据分为有病和无病两种类别。通过我们的调研发现，并不是所有可解释方法都适用于生成显著图且根据显著图得到分割后的二值图。所以我们挑选了适合本次实验的一共 10 种可解释方法进行了评测，为了证明我们的结论，我们还在动物类别预测的任务上进行了对比实验。我们首先在这个数据集上微调了 ResNet50[57]的模型，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，为了对分割结果进行定量评估，我们计算了各个分割和人工标注的交并比（IOU）和重叠度量（DICE）。IOU 和 DICE 各自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加敏感，而 DICE 对于目标大小、形状的变化更加鲁棒。因此，结合使用这两个指标可以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

实验结果如图 3、图 4 和表 1 所示，其中图 3 和图 4 我们分别在脑影像和鸟类影像上实验了各种可解释方法，表 1 是对应方法的评价指标结果。我们观察到，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表现普遍较高，无论是在直观视觉上，还是 DICE 和 IOU 的量化指标上都支持了这一结论，基于梯度的方法处于中间水平，而 RISE 这种方法表现最差，它甚至没有将病灶区域标注出来（IOU 和 DICE 也是最低）。但是当我们把这些方法用在自然影像时（如识别鸟类），会得到一些违反直觉的结论。如在脑肿瘤上表现最差的 RISE，换到鸟影像上时，评价指标处于中间水平。此外，基于 CAM 及其变种的方法这时候也不再在评价指标上处于压倒性优势。这表明一些方法虽然在自然影像数据集上表现很好，但是在医学影像上性能会出现较大下降，这反映了医学影像数据集的特殊之处，后面在进行这方面工作时，可以留意这一问题。此外，我们还发现，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 都出现了显著下滑，这反映了这些可解释方法并不能很好的将病灶部位完整的区分开来，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。这可能涉及到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，对于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。为了解决以上问题，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。

2）领域专业知识的整合： 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。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。

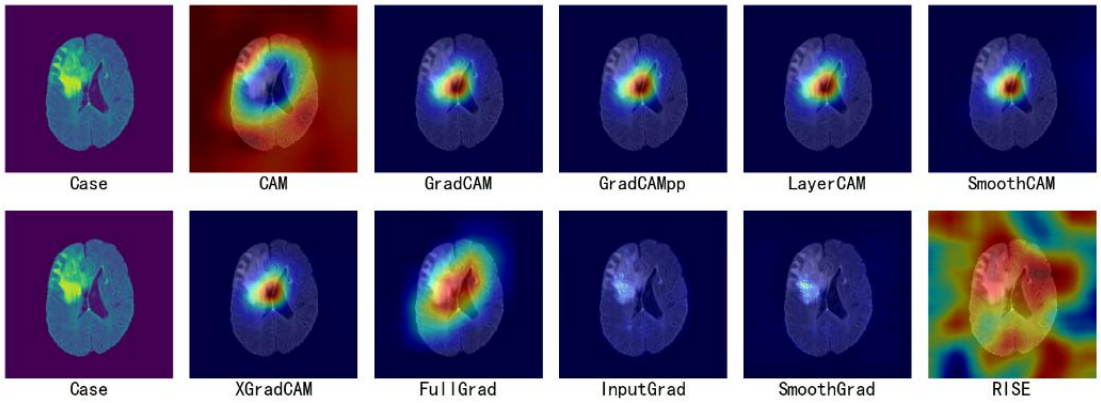

图 3：脑影像可解释方法效果图

Figure 3: Effect Diagram of Brain Imaging Explainable Methodology

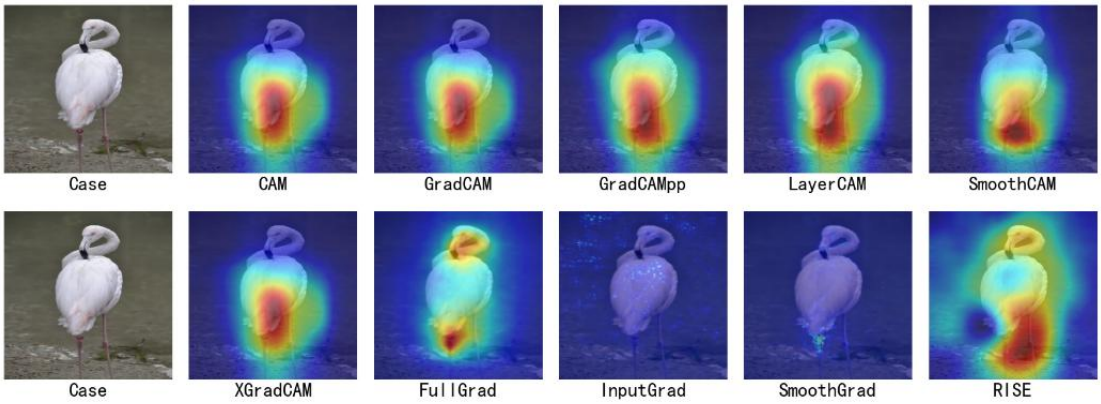

图 4：自然影像可解释方法效果图

Figure 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	CAM	GradCAM	GradCaMpp	LayerCAM	SmoothCAM
IOU(brain)	0.2921	0.3048	0.3035	0.3029	0.2726
TOU(flamingo)	0.4023	0.4023	0.4451	0.4490	0.4571
DICE(brain)	0.4522	0.4671	0.4657	0.4650	0.4284
DICE(flamingo)	0.5737	0.5737	0.6161	0.6198	0.6274
方法	XGradCAM	FullGrad	InputGrad	SmoothGrad	RISE
IOU(brain)	0.3052	0.2818	0.1796	0.1852	0.1747
IOU(flamingo)	0.4023	0.6938	0.1476	0.1168	0.4063
DICE(brain)	0.4577	0.4397	0.3045	0.3125	0.2975
DICE(flamingo)	0.5737	0.8123	0.2573	0.2092	0.5779

5. 总结与展望

一个优秀的人工智能系统应该具备强大的智能性能，同时保持透明、可解释、可靠、公平、注重隐私保护、可持续，并且能够与人类友好互动，同时具备社会责任感，确保其发展和应用符合伦理和社会价值观。根据当今的标准，盲目地相信预测性分类器的结果是不可取的，因为在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。在本文中，我们探讨了可解释人工智能（XAI）至关重要的原因，涵盖了 XAI 的几个方面，并根据它们背后的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，我们的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着精度低、语义指向不明确等问题。针对这些局限性，我们提出了一些改进建议，以供后来的研究者继续探究。

总体而言，人工智能可解释性是一个备受关注的领域，研究人员正在努力改进模型解释

技术，使深度学习模型的决策过程更容易理解。随着可解释性工具的不断发展，预计这些工具将在医疗、金融等领域得到广泛应用，我们认为未来人工智能的可解释研究将可从以下几个方面着手探究：

1）全局方法与局部方法结合。全局方法和局部方法各自有其优势和局限，如果将这两种方法结合起来可以更全面地理解和解释人工智能系统的工作原理。使用全局方法来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。然后，利用这些全局分析结果来指导局部解释的生成，会使得局部解释更具有准确性和连贯性。

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，大多不能做到对生成的样本进行连续和有意识的修改，研究人员后面可以考虑把数据集投影到一个有特征信息或者语义信息流行空间中，通过在流行空间中操控，来达到有导向的反事实生成解释。

3）交互式可解释。当前的 XAI 系统通常强调的是对 “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，都只能影响机器 “生成解释” 的过程，而不影响机器 “做出决策” 的过程，这是一种单向的价值目标对齐。后续工作可以考虑在“双向价值对齐”方面进行更深入的研究。即通过系统和人的交互逐渐更新其价值函数来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，大多数模型都是通过数据驱动的方式进行训练和学习的，而较少关注到知识驱动的观点。然而，将人类知识嵌入到深度学习模型中，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来指导模型的学习过程，使其更具有解释性和可理解性。

参考文献

[1] Gunning D, Aha D. 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]. 1991.

[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): 524-36.
Kong XW, Tang XZ, Wang ZM. A survey of explainable artificial intelligence decision [J]. Syst Eng Theory Pract, 2021, 41: 524-36.

[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 V I. 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 Y S, 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 H U, Bosaily A E-S, Brown L C, 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): 815-22.

[9] Bosaily A E-S, Parker C, Brown L, 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 D C, Raman S S, Mirak S A, et al. Detection of individual prostate cancer foci via multiparametric magnetic resonance imaging [J]. European urology, 2019, 75(5): 712-20.

[11] Van Der Leest M, Cornel E, Isrnel 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): 570-8.

[12] America C O Q O H C I. Crossing the quality chasm: a new health system for the 21st century [M]. National Academies Press, 2001.

[13] Barry M J, Edgman-Levitan S. Shared decision making—The pinnacle patient-centered care [J]. 2012.

[14] Kunneman M, Montori V M, Castaneda-Guarderas A, et al. What is shared decision making?(and what it is not) [J]. Acad Emerg Med, 2016, 23(12): 1320-4.

[15] O'neill E S, Grande S W, 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 P A, Brito J P, 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: 159-63.

[17] Bjerring J C, Busch J. Artificial intelligence and patient-centered decision-making [J]. Philosophy & technology, 2021, 34: 349-71.

[18] Politi M C, Dizon D S, Frosch D L, 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.

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

[20] Ribeiro M T, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier; proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, F, 2016 [C].

[21] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions [J]. Advances in neural information processing systems, 2017, 30.

[22] Antwarg L, Miller R M, Shapira B, et al. Explaining anomalies detected by autoencoders using SHAP [J]. arXiv preprint arXiv:190302407, 2019.

[23] Sundararajan M, Najmi A. The many Shapley values for model explanation; proceedings of the International conference on machine learning, F, 2020 [C]. PMLR.

[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 S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees [J]. Nature machine intelligence, 2020, 2(1): 56-67.

[26] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks; proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, F, 2014 [C]. Springer.

[27] Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models [J]. arXiv preprint arXiv:180607421, 2018.

[28] Burns C, Thomason J, Tansey W. Interpreting black box models via hypothesis testing; proceedings of the Proceedings of the 2020 ACM-IMS on foundations of data science conference, F, 2020 [C].

[29] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016 [C].

[30] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization; proceedings of the Proceedings of the IEEE international conference on computer vision, F, 2017 [C].

[31] ChattOpadhay A, Sarkar A, Howlader P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks; proceedings of the 2018 IEEE winter conference on applications of computer vision (WACV), F, 2018 [C]. IEEE.

[32] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps [J]. arXiv preprint arXiv:13126034, 2013.

[33] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: The all convolutional net [J]. arXiv preprint arXiv:14126806, 2014.

[34] Mahendran A, Vedaldi A. Salient deconvolutional networks; proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14, F, 2016 [C]. Springer.

[35] Li H, Tian Y, 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: 70-86.

[36] Ancona M, Ceolini E, Öztireli C, et al. Towards better understanding of gradient-based attribution methods for deep neural networks [J]. arXiv preprint arXiv:171106104, 2017.

[37] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks; proceedings of the International conference on machine learning, F, 2017 [C]. PMLR.

[38] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences; proceedings of the International conference on machine learning, F, 2017 [C]. PMLR.

[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 J, Lin Z, Yang J, et al. Generative image inpainting with contextual attention; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2018 [C].

[41] Chang C-H, Creager E, Goldenberg A, et al. Explaining image classifiers by counterfactual generation [J]. arXiv preprint arXiv:180708024, 2018.

[42] Guidotti R, Monreale A, Matwin S, et al. Explaining image classifiers generating exemplars and counter-exemplars from latent representations; proceedings of the Proceedings of the AAAI conference on artificial intelligence, F, 2020 [C].

[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 [J]. arXiv preprint arXiv:151105644, 2015.

[45] Xie R, Chen J, Jiang L, et al. Active Globally Explainable Learning for Medical Images via Class Association Embedding and Cyclic Adversarial Generation [J]. arXiv preprint arXiv:230607306, 2023.

[46] Xie R, Chen J, Cai Y, et al. Accurate Explanation Model for Image Classifiers using Class Association Embedding[C]. accepted.

[47] Singla S, Murali N, Arabshahi F, et al. Augmentation by Counterfactual Explanation-Fixing an Overconfident Classifier; proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, F, 2023 [C].

[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 Y, Ma X, Zhang Y, 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.

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

[51] Frosst N, Hinton G. Distilling a neural network into a soft decision tree [J]. arXiv preprint arXiv:171109784, 2017.

[52] Che Z, Purushotham S, Khemani R, et al. Distilling knowledge from deep networks with applications to healthcare domain [J]. arXiv preprint arXiv:151203542, 2015.

[53] Liu X, Wang X, Matwin S. Improving the interpretability of deep neural networks with knowledge distillation; proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW), F, 2018 [C]. IEEE.

[54] 吴俊杰, 刘冠男, 王静远, et al. 数据智能: 趋势与挑战 [J]. 系统工程理论与实践, 2020, 40(8): 2116-49.

Wu JJ, Liu GN, Wang JY, et al. Data intelligence: Trends and challenges [J]. Syst Eng-Theory Pract, 2020, 40: 2116-49.

[55] Menze B H, 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(1): 1-13.

[57] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016 [C].