重点选题“人工智能与医学”·专题专栏

医学人工智能可解释性的研究与应用进展▲

张会勇[1,2] 王富博[2,3*]

（1 桂林电子科技大学计算机与信息安全学院，广西桂林市 541004；2 广西医科大学基因组与个体化医学研究中心，广西南宁市 530021；3 数字医学与健康广西高校工程研究中心，广西南宁市 530021）

王富博，教授，博士研究生导师，美国加州大学洛杉矶分校联合培养博士，美国西达赛奈医学中心博士后，现任数字医学与健康广西高校工程研究中心主任。目前主要从事医学人工智能研究，包括恶性肿瘤及心脑血管重大疾病跨时空多模态数据融合策略和人工智能感知关键技术，基于临床大数据和人工智能的重大疾病临床决策支持系统的研发与应用，泛肿瘤标志物与肿瘤特异性标志物的筛选、鉴定和试剂盒的研发，恶性肿瘤“医防融合”与全生命周期主动健康研究及应用等。担任中国生物医学工程学会器官与类器官芯片分会委员、中国老年保健协会泌尿系统增龄性疾病防治分会常务委员、广西医师学会泌尿外科专业委员会委员、中国研究型医院学会细胞外囊泡专业委员会青年委员、广西生物信息学学会理事等。主持国家重点研发计划课题1项、国家自然科学基金面上和青年项目各1项、中央引导地方科技发展资金专项1项、广西杰出青年科学基金1项等。以第一作者或通信作者在*Nature Cell Biology*、*Molecular Cancer*、*BMC Medicine*、*Cell Death & Disease*、*International Journal of Cancer*等SCI期刊发表论文36篇。先后在国际泌尿外科年会（SIU）、美国泌尿外科年会（AUA）、韩国泌尿外科年会（KUA）等国际会议上作会议报告。获得第十届中国国际大学生创新大赛全国总决赛铜奖2项，第八届、第九届、第十届中国国际大学生创新大赛广西赛区选拔赛金奖，2022第七届南宁市创新创业（人才）项目投融资路演大赛二等奖，2021“力合星空杯”生物医药大健康成果转化创业大赛一等奖。

【提要】 人工智能（AI）在医疗领域的深度渗透催生了精准诊断、个性化治疗、主动健康等创新治疗模式，但其决策过程的不透明性（也称为“黑箱”效应）可引发临床信任危机与监管合规挑战。可解释性人工智能（XAI）通过揭示模型决策逻辑，成为破解AI技术潜力与应用瓶颈矛盾的核心路径。XAI正从辅助解释工具发展为系统性解决方案，其将推动医学AI从“结果输出”转向“过程透明”，为构建可信赖的智能医疗生态奠定基础。本文系统解析医学AI可解释性的多维内涵，梳理特征重要性分析、因果推断、多模态融合等关键技术的演进脉络，结合影像诊断、药物研发等前沿场景的实证研究，深入探讨医学AI数据异质性、责任界定等核心挑战。

【关键词】 人工智能；可解释性；医学；因果推断；临床决策支持

【中图分类号】 R 319 **【文献标识码】** A **【文章编号】** 0253-4304（2025）08-1088-11

DOI：10.11675/j.issn.0253-4304.2025.08.04

Research and application progresses on interpretability of medical artificial intelligence

ZHANG Huiyong[1,2], WANG Fubo[2,3]

(1 School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China; 2 Research Center for Genomic and Personalized Medicine, Guangxi Medical University, Nanning 530021, Guangxi, China; 3 Guangxi University Engineering Research Center of Digital Medicine and Healthcare, Nanning 530021, Guangxi, China)

【Abstract】 The deep integration of artificial intelligence (AI) in healthcare has given rise to innovative paradigms such as precision diagnosis, personalized treatment, and proactive health. However, the opacity of AI decision-making (often

▲基金项目：广西壮族自治区自然科学基金面上项目（桂科发〔2025〕69号）；广西壮族自治区杰出青年科学基金（2023GXNSFFA026003）

第一作者简介：张会勇，在读博士研究生，中级工程师，研究方向为医学人工智能。

*王富博为通信作者。

termed the "black box" effect) can trigger clinical trust crises and monitor regulatory compliance challenges. Explainable artificial intelligence (XAI), by unveiling the logical framework behind model decisions, has emerged as a pivotal approach to reconciling the tension between AI's potential and its application bottlenecks. XAI is evolving from an auxiliary explanatory tool into a systematic solution. It is poised to shift medical AI from "outcome delivery" to "process transparency", laying the foundation for a trustworthy intelligent healthcare ecosystem. This paper systematically deconstructs the multidimensional implications of interpretability in medical AI, traces the evolution of key technologies, including feature importance analysis, causal inference, and multi-modal fusion, and delves into core challenges such as data heterogeneity and accountability demarcation, supported by empirical studies in cutting-edge applications like medical imaging diagnosis and drug discovery.

【Key words】 Artificial intelligence, Interpretability, Medicine, Causal inference, Clinical decision-supporting

近年来，人工智能（artificial intelligence，AI）与机器学习在医疗领域的应用呈爆发式增长，改变了疾病诊断与医疗决策的模式。例如，在医学影像领域，Hendrix等[1]开发了一项基于深度学习的AI系统，该系统在鉴别肺部良性结节、原发性肺癌和肺转移瘤中的敏感性分别为94.3%、96.9%和92.0%，均高于大多数放射科医生的主观鉴别敏感性。在临床疾病预测方面，Li等[2]建立了分时段的AI集成学习模型，能够预测ICU患者发生脓毒症的风险，预测精度达到85%，比基线模型提高了23%；Zheng等[3]所构建的机器学习模型ShockSurv，能够准确预测脓毒性休克患者的28 d死亡率，曲线下面积（area under the curve，AUC）达0.903。在药物研发领域，AI通过提高效率和识别新型治疗靶点，改变了阿尔茨海默病的药物发现方式，基于AI的分析方法（从药物设计、虚拟筛选到药物-靶点相互作用预测）结合多组学数据分析，不仅推动了精准医学的发展，也有助于更深入地理解疾病的病理生理学机制[4-5]。

然而，复杂神经网络的"黑箱"特性引发多重挑战。以深度学习为例，其多层非线性变换形成的决策边界难以被人类理解，因此，即使AI建议更准确，人们在医疗决策中对AI建议的接受度仍较低，普遍更愿意接受来自临床医生的医疗建议[6]。2023年被*Nature*杂志撤稿的一项AI制药研究[7]暴露了数据偏差未被解释的风险——模型依赖不平衡的生物数据导致靶点误判，造成数百万研发资金被浪费。与此同时，欧洲联盟于2024年出台的《人工智能法案》将医疗AI列为高风险系统，要求提供"可追溯的决策证据链"，而美国食品药品监督管理局（Food and Drug Administration，FDA）的审批中仅14.7%的AI设备提交了透明性验证报告[8]，这反映AI的可解释性已从技术需求升级为伦理与法律的刚性约束。

在此背景下，可解释性人工智能（explainable artificial intelligence，XAI）通过技术手段将AI决策转化为人类可理解的形式，成为破局关键。在技术层面，特征重要性分析，如SHAP[9]、LIME[10]与可视化技术（Grad-CAM[11-12]、LRP[13]等）能揭示模型决策逻辑；在临床层面，XAI有助于优化医生的决策过程，例如，一项针对黑色素瘤与痣的鉴别诊断研究表明，采用XAI系统后，皮肤科医生的诊断信心及对AI建议的信任度均显著提升（分别提升12.25%、17.52%），其机制源于XAI系统引入了融合文本和局部区域的多模态解释性[14]；在监管层面，XAI可满足WHO发布的《医疗卫生中人工智能的伦理和治理指南》中提出的"可追溯、可解释、可审计"要求[15]，为AI医疗设备合规上市提供技术支撑。这种多维价值体系正推动医学AI从"结果输出"向"过程透明"转型，构建可信赖的智能医疗生态。

当前，XAI研究已形成从基础理论到应用创新的完整体系：在方法学上，从早期的特征可视化分析发展到多模态因果推理[16]；在技术路径上，从事后解释（代理模型[17]等）转向内生可解释架构设计（融合决策树与注意力机制[18]等）；在应用场景上，从影像诊断拓展到个性化治疗、药物研发等复杂领域[19]。然而，数据异质性导致的解释泛化不足、隐私保护与解释精度的悖论、责任界定模糊等问题，仍需要进一步探讨。本文从医学AI可解释性的核心维度出发，梳理关键技术、前沿进展及转化挑战，以期为跨学科研究提供理论框架，为临床转化指明实践路径。

1 医学AI可解释性的核心维度解析

1.1 技术透明度——从数据输入到决策逻辑的全路径揭示 技术透明度要求医学AI系统清晰展示数据输入、特征处理及决策生成的全过程。在数据层面，需要公开训练集的人口统计学分布特征（种族、年龄比例等）、数据标注流程（标注指南、审核机制等）。例如，哈佛医学院眼科AI实验室在构建青光眼检测数据集Harvard-GF时，不仅公开了数据集中亚裔、非裔和白种人样本的分布情况，还发布了一种新的公平性缩放度量方法[20]。在特征层面，SHAP值可量化每个临床指标对预测结果的贡献度。例如，在前列腺癌生化复发风险的机器学习预测模型中，前列腺特异性抗原（prostate specific antigen，PSA）最低值在整体样本中的综合SHAP值最高，表明其为该模型最具影响力的特征；具体到某患者时，初诊PSA值对其预测结果的贡献尤为显著，而相较于训练集的平均预测值（17.06），该患者初诊PSA值对应的SHAP值为9.18，使得其最终预测值达到28.46，成为影响其生化复发概率提升的关键变量[21]。在决策逻辑层面，注意力机制通过热力图展示AI模型在影像诊断中关注的解剖区域。例如，Wollek等[22]在利用ViT模型对胸部X线片进行气胸分类时，采用TMME法生成基于注意力机制的显著性热力图，精准标注模型关注的解剖区域（气胸病变），不仅提升了分类性能（AUC达0.95），还增强了模型的可解释性。

1.2 临床可靠性——解释内容与医疗实践的深度契合 临床可靠性包含公平性、可信赖性与临床相关性。公平性要求AI模型在不同人群中的表现一致。当前，许多工具可以检测模型这一特性，例如，AIF360工具包[23]可以检测AI在性别、种族维度的偏差。此外，梯度协调框架FairGrad也可以平衡预测性能与多个敏感属性（种族、性别等）之间的公平性。在精神活性物质使用障碍的治疗效果预测任务中，FairGrad法有效降低了种族维度上的平等机会差异（equalized odds difference，EOD），相较于基线模型降低了28.8%，在脓毒症诊断任务中，其将种族EOD降低了47.8%[24]。可信赖性依赖于严格的验证流程。欧洲癌症未来信息技术联盟提出构建AI技术临床验证框架用于预测癌症的治疗反应，即基于真实世界数据，通过多中心验证、标准化数据模型OMOP、量化性能指标（AUC等）及算法可解释性流程，确保AI技术预测结果的可信赖性，强调严格验证是临床可靠性的核心支撑[25]。临床相关性则要求AI模型解释符合循证医学指南。根据《临床XAI指南》提出的五项准则，其中第二项准则（G2）——临床相关性，明确规定解释形式须与循证医学指南或临床路径高度一致，以确保XAI生成的推荐对临床决策具有直接临床指导意义[26]。因此，XAI的解释框架应与临床指南的目标保持一致。例如，在放射学中，XAI的热图可视化应与影像学诊断标准相匹配，以提升临床接受度和安全性[27]；Lucia心房颤动应用程序（Lucia APP）通过分析心电图实现节律检测，并计算CHA2DS2-VASc和HAS-BLED评分来提供基于指南的抗凝治疗建议，该应用程序利用循证指南（美国心脏病学会/美国心脏协会相关指南），能详细解释为何特定风险因素（如年龄≥75岁或有心力衰竭病史）会成为推荐使用口服抗凝药物的原因，从而增强了临床相关性和可依赖性[28]。可见，通过以上手段，XAI可与现有临床指南紧密结合，提升了其在实际应用中的有效性和可信度。

1.3 解释实用性——面向多元用户的分层设计 针对医生、患者、管理者的不同需求，医学AI的可解释性须具备分层适配性。对于医生，医学AI应提供深入的技术性解释，详细阐明病理生理学机制并直接映射至循证指南。例如，可通过SHAP方法量化各基因对分类结果的贡献度，并结合生物实验的验证结果，来解释关键基因如何通过代谢紊乱、细胞分化异常或免疫逃逸等途径驱动癌症发生。以甲状腺癌为例，当*TG*基因表达水平最高（SHAP值最大）时，甲状腺激素平衡被打破，进而引发基因组不稳定，显著提高甲状腺癌的发生风险[29]。对于患者，医学AI应将复杂的模型预测转化为易于理解的生活化语言，以具体的行为建议（饮食调整或增加运动等）帮助其掌握健康状况并推动行为改变[30]。而对于管理者，医学AI则需要以摘要形式呈现关键性能指标、合规要点与完整的审计轨迹，包括模型训练日志、参数调整记录、临床验证数据等，以满足《人工智能法案》的“可追溯性”要求[15]。这种分层设计可确保医学AI可解释性在专业性与可及性之间获得平衡，避免“技术术语过剩”或“关键信息缺失”。

2 医学AI可解释性的关键技术体系

2.1 特征重要性分析——量化决策驱动因素的核心技术 特征重要性分析通过数学建模识别对预测结果影响最大的输入特征，是最基础的解释技术。SHAP值基于博弈论的边际贡献理论，为每个特征分配全局

重要性分数。例如，在一项纳入52 645例健康成年人的前瞻性队列研究中，研究者利用1 463种血浆蛋白质组学数据构建预测模型，并借助SHAP值进行量化评估与可视化，结果显示，胶质纤维酸性蛋白（glial fibrillary acidic protein，GFAP）在全因痴呆预测模型中贡献度最高，其次是神经纤维丝轻链蛋白（neurofilament light chain，NEFL）与生长分化因子15（growth differentiation factor 15，GDF15），该模型可在发病前10~15年识别痴呆高风险个体；针对该模型及SHAP分析筛选出的蛋白进行富集分析，结果表明GFAP可反映星形胶质细胞活化参与神经炎症进程，NEFL与GDF15则分别关联轴突损伤和血管病变通路[31]。该分析框架不仅实现了临床特征贡献度的定量排序，更通过生物学机制映射为早期筛查标志物的临床验证提供了可解释的决策依据。LIME则聚焦局部解释，通过在单个样本附近生成线性代理模型，解释特定患者的决策逻辑。例如，应用LIME解释XGBoost模型对高血压风险预测的结果时发现，在无高血压人群中，未使用高血压药、无高脂血症及年龄是模型预测结果的主要影响因素；而在高血压人群中，使用高血压药和年龄为模型预测结果的关键影响因素[32]。可视化技术如Grad-CAM，可通过热力图将抽象的模型决策转化为视觉证据。例如，在鉴别结节病与淋巴瘤的研究中，Grad-CAM热图聚焦氟代脱氧葡萄糖异常积累部位，同时结合正侧位最大密度投影图像，获得优异的鉴别性能，平均准确率为0.890，AUC为0.963[33]，其有助于识别疾病特征，为临床鉴别提供更多的依据。这些方法的优势在于模型无关性，可普遍适用于神经网络、随机森林等算法，但仍面临SHAP计算复杂度高、LIME局部解释稳定性不足等挑战。

2.2 *代理模型与基于案例的推理——连接技术逻辑与临床经验的桥梁* 代理模型通过将复杂黑箱模型训练为简单可解释模型，实现全局解释的可视化[17]。例如，通过将基于深度学习的肺癌分类模型映射到决策树结构，可以使医生清晰看到“无发现（判定分数≤0.43）→结节（判定分数≤0.16）→肿块（判定分数≤0.06）→判定为恶性病灶”的规则路径，显著降低专业理解门槛[34]。基于案例的推理则利用临床类比思维，通过检索相似病例来解释当前的预测结果。该机制通过构建“案例检索-特征匹配-结果对比”的逻辑框架，将模型决策与医生经验性推理相结合，提升解释的临床可理解性与实用性。例如，在糖尿病自我管理场景中，当AI系统因当前温度为7 ℃而建议使用低剂量胰岛素时，同步呈现历史相似案例（7 ℃情况下低剂量有效、31 ℃情况下正常剂量有效），通过基于温度等特征的剂量决策对比分析，辅助用户理解建议依据，具象化临床类比思维的解释效能[35]。此类方法的核心优势在于解释形式贴近临床思维，但代理模型存在“保真度”不足的问题，即简单模型可能无法完全复现复杂模型的决策边界，从而导致解释偏差，因此需要通过模型融合技术（多代理模型集成等）提升准确性。

2.3 *内生可解释模型——从架构设计实现透明决策的革新* 内生可解释模型在算法设计阶段融入透明性，从根本上解决黑箱问题，其核心优势在于解释性与预测性能的内生一致性。早期研究聚焦于天然可解释的基础模型，如决策树与基于规则的系统，其分层决策逻辑与临床指南高度契合，成为最初的XAI[36]。例如，在葡萄球菌菌血症治疗中，基于决策树的临床决策支持系统通过显式规则分层实现透明决策：针对金黄色葡萄球菌感染，依据病例的复杂程度，直接输出“立即咨询感染病专家”（复杂病例）或“推荐β-内酰胺类抗生素”（非复杂病例）；针对凝固酶阴性葡萄球菌感染，基于血培养次数生成“重复培养”或“排查血管内装置”的建议，其规则路径可直接追溯至临床诊疗规范[37]。而注意力机制的引入则赋予了神经网络显式解释能力，其通过量化模型对输入数据的关注程度，生成可解释信号。例如，在数字病理学中，基于注意力的多示例学习模型对CAMELYON16数据集的肿瘤组织图像进行分析时，可通过注意力图显式标注对癌灶区域的关注程度，无论是合成图块扰动实验还是特征采样验证，注意力权重均能准确映射病理医生的诊断关注点，实现“像素级决策归因”[38]。此类模型的优势在于解释性与预测性的一致，但常面临准确性与透明性的权衡——过于简化的结构可能牺牲复杂模式识别能力，因此需要通过混合架构（神经符号系统等）平衡性能与解释性。例如，癌症神经符号AI系统是将神经符号方法与命名实体识别、实体链接相结合的一项混合AI系统，其可将非结构化的临床笔记转化为结构化术语，实体识别准确率较传统方法提升58%[39]。

3 医学AI可解释性的前沿研究进展——从方法创新到范式变革

3.1 多模态XAI——融合异质数据生成深度解释的突破 医疗数据的多模态特性(图像、文本、基因组、生理信号等)推动XAI向融合方向演进。多模态XAI通过图神经网络构建跨模态知识图谱,以节点特征表征不同模态实体(基因、影像区域、临床指标等),以边权重进行实体间因果关系建模,从而实现异质数据的语义关联与协同解释。多模态XAI的典型代表模型为TMO-Net,其可解释预训练多组学模型,通过自监督学习生成跨组学联合嵌入,促进跨组学相互作用,实现联合表征学习及不完整多组学数据的推断,在乳腺癌亚型预测、癌症转移预测、药物反应预测和患者预后预测等多任务中表现优异,可有效提取关键特征并助力解析肿瘤发病机制,为转移风险评估提供跨尺度解释依据[40]。一项临床研究结果显示,在转移性非小细胞肺癌一线免疫治疗效果预测中,通过整合临床、PET/CT影像、病理和转录组数据而构建的多模态机器学习模型,其预测患者治疗后1年死亡风险的AUC及预测患者生存情况的一致性指数显著优于单模态模型;93%的多模态特征组合可有效区分高低风险人群,其中74%的多模态特征组合的预测效能优于单一指标细胞程序性死亡配体1,这表明跨模态融合分析能捕捉传统单模态分析忽略的复杂生物标志物间的关联,并进一步证实了多模态融合的解释优势[41]。针对多模态数据的语义鸿沟,如图像像素与基因序列的异质表示,预训练模型CLIP[42]可通过跨模态对比学习对齐异质数据空间。而领域知识注入技术则可将统一医学语言系统(unified medical language system,UMLS)嵌入模型训练,例如,某研究基于UMLS构建增强型大型语言模型(large language model,LLM),经多名医生评估,该增强型LLM在医学问题回答的事实准确性、内容相关性和完整性方面显著优于原LLM,解释内容符合临床标准[43]。

3.2 因果推理驱动的解释深化——从相关到因果的范式跃迁 传统AI依赖相关性分析,而因果推理技术[结构因果模型(structural causal model,SCM)和反事实解释等]通过揭示数据因果机制,显著提升了解释深度。SCM通过有向无环图(directed acyclic graph,DAG)建模变量间因果关系,支持干预性推理,能够有效解答“若采取特定措施将如何影响结果”这一临床关键问题。例如,在关于阿尔茨海默病诊断的研究中,以因果变量载脂蛋白ε4等位基因构建的模型,相较于以脑脊液tau蛋白浓度这类反因果变量构建的模型,在跨人群应用时展现出更强的迁移能力;基于DAG的马尔可夫毯理论筛选预测变量并构建模型,不仅校准效果更优,且输入变量更为精简,为临床风险预测模型的优化提供了理论依据[44]。此外,SCM在解析基因调控网络关键路径,探究脑区功能连接方向性,以及分析流行病学中共病与疾病进展之间因果关联等医学场景中,推动了从“数据相关”到“因果归因”的模式转变[45]。

反事实解释则通过构建“最小差异”假设场景,帮助用户理解模型决策边界。例如,在医学图像分类领域,利用基于CycleGAN技术[46]的GANterfactual方法,对胸部X线图像的肺野透明度等纹理特征进行定向修改,生成反事实图像,直观呈现特征变化对分类结果的影响。该方法使AI决策预测准确率达94.68%,显著优于传统显著性图方法(LIME、LRP等),同时显著提升了用户对解释的满意度与信任度[47]。因果推理技术与机器学习的结合(因果增强神经网络等)还可以识别数据中的混杂因素,减少虚假关联[48],例如,在心脏疾病的心电图分类中,可通过后门调整消除患者的年龄、性别或身体成分等混杂因素对分类结果的间接影响,结合反事实推理构建因果图,切断“混杂因素→信号特征→疾病标签”的虚假关联路径,使模型聚焦于感染诱导的心脏异常波形与病理生理机制之间的真实因果关系,提升对心肌炎、败血症相关心肌损伤的诊断准确性与解释可靠性[49]。这些技术的突破标志着医学领域XAI从“模式识别”向“机制理解”的模式转变。

3.3 大模型解释——破解复杂神经网络黑箱的关键路径 LLM与深度学习模型在医学文本处理、影像分析中展现出强大的能力,但其千亿级参数导致“黑箱”特性,使得模型的决策过程难以解释,进而催生了一系列专用可解释性增强技术。当前主要有以下技术路径:一是通过结构化知识图谱提升推理的规范性与临床一致性。例如,DR.KNOWS模型融合医学知识图谱(UMLS等),利用图神经网络和多头注意力机制检索并输入症状相关的临床路径,再将输出映射到标准诊断术语,有效提升了模型在真实EHR数据中的预测准确性及与人工评估结果的契合度[50]。二是通过思维微调来校正人类偏好。例如,MedFound的优化版MedFound-DX-PA可通过自引导链式思维微调来校正人类偏好,从而提升临床推理能力,其在多场景诊断表现优异,还能辅助医生提升诊断准确率[51]。例如,在对1名呼吸系统疾病患者进

行诊断时，经模型提示医生将“急性支气管炎”修正为“慢性支气管炎急性加重”；在甲状腺疾病案例中，模型依据抗体指标协助医生确诊“自身免疫性甲状腺炎”[51]。三是通过模型蒸馏提升决策链条的透明度。学生模型通过模仿教师模型的推理过程，学习其生成的解释性标注数据，并结合特征蒸馏与反馈机制，在推理路径、领域知识迁移、逻辑表征等方面实现决策逻辑的多维校正，特别适用于医疗、法律等高合规性需求场景[52]。例如，LEADER框架通过“教师-学生”架构，将复杂LLM LEADER(T)的药物推荐知识压缩至轻量模型LEADER(S)，在初次就诊患者场景中有效保留核心能力，在MIMIC-Ⅲ和MIMIC-Ⅳ数据库的相关数据集中的精准率-召回率曲线AUC分别达0.7631和0.7033[53]。此外，神经符号系统也为提升模型解释性提供了新的思路。通过将医学逻辑规则引入模型结构，可对模型输出进行符号层面的约束，从而提升推理的逻辑一致性。例如，逻辑神经网络模型在糖尿病预测中将临床规则转化为可微分逻辑操作，并通过可学习的参数建模多路径风险因素，模型中可视化的特征权重与阈值与临床认知高度一致，使决策过程更加合理透明[54]。

3.4 *自适应交互解释系统——构建动态人机协同的新生态* 面向临床决策的实时性与个性化需求，XAI正从“静态解释”转向“动态交互”，通过用户反馈机制优化解释策略。例如，Clinician-in-the-Loop框架基于强化学习生成“集值策略”，在MIMIC数据库的脓毒症临床数据集中，将由不同静脉输液量与血管加压药剂量组合的25个治疗选项按预后相似性划分为近等效行动集合，将次优边际参数ζ设定为0.05时，50%的状态可映射到多个近等效行动集合，同时可视化热图可显示不同治疗选项之间的等效关联性，为医生提供直观的决策参考[55]。Huang等[56]利用SurgicalVLM-Agent构建动态交互框架以高效校正Llama 3.2模型，使模型能够适应手术动态环境，并在任务规划与提示生成上获得优异的性能，从而推动XAI从静态模型向动态智能协同工具演进。可视化交互工具如3D Grad-CAM[57]、动态决策树[58]等，允许医生交互式探索模型决策逻辑。PATHFx Version 3.0基于贝叶斯信念网络对发生骨骼转移的癌症患者的生存情况进行预测，医生可依据治疗后不同时间点的预测结果及多源数据动态调整治疗方案，实现临床决策的交互式支持[59]。在机器人辅助前列腺切除术的“目标活检”阶段，Tanzi等[60]构建的一套两步自动对齐系统，利用U-Net与MobileNet进行语义分割，提取导管锚点坐标与旋转角度，将术前3D模型精准叠加至2D内镜视频流，辅助医生进行定位活检。而XAI机器人辅助软性膀胱镜检查系统则可通过探头进行半自主导航与可视化分析，从而实现膀胱肿瘤的标准化扫描与病灶智能标注，生成包含预处理图像的结构化报告供医生审核，提升检测可重复性并减少漏诊与过度诊断[61]。此类系统的核心价值在于支持“解释-验证-修正”的闭环流程，使AI从“决策输出者”转变为“协作探索伙伴”，但对交互界面的易用性、解释生成的实时性提出了更高要求，需要结合自然语言处理与可视化技术提升用户体验。

4 XAI的临床应用——从辅助工具到智能伙伴

4.1 *影像诊断——透明化决策提升临床信任的实证研究* 在基于CT影像的肺癌筛查中，结合Grad-CAM与SHAP的XAI工具可精确标注结节的恶性特征，并量化其对诊断准确率的贡献度。在CT影像肺癌检测中，有研究者融合ResNet50与Grad-CAM构建XAI模型，通过可视化技术精准定位并量化标注结节恶性特征（分叶征、胸膜牵拉征等），结果显示，在肺腺癌预测中，热图所集中的区域与此类肺癌的典型分布区域高度吻合，而在肺大细胞癌预测中，模型的置信度分数达98.02%，该模型不但错误分类率低，而且为诊断提供可追溯的视觉证据[62]。在针对病理全切片的分析中，视觉注意力机制通过多轮编码器聚焦肿瘤细胞密集的关键区域，实现细粒度特征推理，基于WBCD数据库进行训练后，该方法对肺结节分类的准确率达99.86%，对乳腺癌分类的准确率达99.89%，均高于传统方法[63]。基于注意力机制的PathOrchestra[64]、人类可解释PathAI[65-66]等病理大模型已进入商业实践阶段，透明化决策逻辑将提升临床医生对AI分级结果的信任。更重要的是，XAI技术促进了“AI初筛-人工复核”的协同工作模式。例如，在基于CT影像的肺癌检测中，ResNet50和Inception V3模型经XAI技术（LIME、Grad-CAM）解释，ResNet50测试准确率达100%，Inception V3达99.92%，但模型存在聚焦错误区域问题，如LIME显示模型关注非癌症区域，而放射科医生可以根据提示进行可疑区域复核，弥补模型不足[67]。显然，这种“AI初筛-人工复核”协同模式是必要的，XAI或成为分级诊疗体系中的关键技术支撑。

4.2 *个性化治疗——因果解释赋能精准干预的创新*

实践　反事实解释技术在临床决策中发挥重要价值，如在针对肺炎患者分析中，规则模型通过反事实解释发现哮喘病史患者因接受强化治疗而死亡率更低，避免了黑箱模型因忽略治疗策略差异导致的干预误导[68]。在国际卒中实验中，研究人员运用反事实框架构建临床预测模型，以推断急性缺血性卒中患者的阿司匹林个体化治疗效果，结果显示该模型的区分度良好，为精准医学中依据个体特征制订个性化治疗策略提供了参考，优化临床决策[69]。Li等[70]在针对癫痫诊疗的研究中，将CR-VAE模型结合Granger因果理论，从脑电图数据中发现脑内因果关系集中于深层结构，与临床前脑叶切除术靶点区域一致，这可为手术定位提供因果证据。这些研究构建了“因果发现-精准干预-效果评估”的闭环体系，通过可解释性分析增强临床决策合理性，为疾病机制解析与个性化治疗方案设计提供创新模式。

4.3　药物研发与安全性评估——从数据驱动到知识融合的跨越　在药物研发与安全性评估领域，AI正推动该领域从依赖统计关联的“数据驱动”模式向整合领域知识、因果推理与多模态数据的“知识融合”模式转型。在靶点识别研究领域，Ren等[71]利用PandaOmics平台整合多组学数据与生物网络分析，精准筛选出TRAF2和NCK相互作用激酶作为纤维化治疗靶点，并开发出小分子抑制剂INS018_055，该小分子抑制剂在临床试验中展现出良好的安全性和耐受性。这一研究结果验证了AI驱动靶点发现的临床价值。针对药物-靶点相互作用预测中知识图谱嵌入（knowledge graph embedding，KGE）方法的置信度问题，基于因果干预的方法通过扰动实体嵌入向量、计算分数序列一致性，6个KEG模型在Hetionet和BioKG数据集中显著降低预期校准误差，其中CrossE模型在Hetionet数据集中的高置信度预测准确率提升至52.1%，为精准筛选候选药物-靶点提供可靠依据[72]。在药物设计阶段，AI技术通过多技术融合提升研发效率：AlphaFold[73-74]和RosettaFold[75]基于蛋白序列直接预测三维结构及配体结合构象，突破传统结构依赖瓶颈；RetroExplainer框架将逆合成任务转化为分子组装过程，通过图神经网络提升合成路径预测的可解释性，性能优于传统方法[76]。此外，因果推理框架在机制解析中发挥关键作用。例如，基于DAG的六步迭代法整合先验知识与数据，解决观察性数据混杂问题：沙利度胺是两种对映异构体的集合体，具有致畸与抗肿瘤双重作用，利用该方法针对沙利度胺的双重作用机制进行解析，发现其（S）-对映异构体具有通过E3泛素连接酶降解SALL4蛋白及靶向IKZF1/3蛋白的双重作用，这一发现推动靶向蛋白降解嵌合体类新药和多特异性药物的研发[77]。在临床应用阶段，基于核形态学的机器学习模型在细胞衰老检测中的准确率为76.32%，并成功验证天然化合物表儿茶素对环氧酶2的抑制作用，为毒理学机制解析提供新的路径[78]。

上述研究结果充分体现了XAI在整合领域知识、因果推理与多模态数据方面的突破性价值，推动药物研发从依赖数据统计关联的“效率工具”，升级为融合因果机制解析的“知识融合决策伙伴”，为精准靶点识别、可靠药物设计及安全性评估提供了跨模式解决方案。

5　挑战与未来方向

5.1　技术挑战——从方法局限到系统适配的突破路径　（1）数据异质性难题：不同医疗机构的影像设备参数（CT层厚、MRI场强等）、电子病历结构化程度差异显著，导致解释模型的跨中心泛化性不足，而设备参数的细微差异即可对模型性能产生显著影响[79]。针对这一挑战，基于联邦学习的跨域解释技术展现出潜在的解决方案，即通过设计域不变特征对齐机制，在保护数据隐私的前提下共享特征重要性分布而非原始数据，实现解释逻辑的跨中心迁移与适配[80]。该路径通过系统性整合数据治理、模型架构优化与跨域知识迁移，为破解异质性数据引发的解释可靠性难题提供了参考。（2）隐私-解释性悖论：详细解释可能泄露患者的基因、病史等敏感信息，而严格匿名化会导致特征失真，降低解释准确性。差分隐私技术通过向解释数据添加可控噪声，在保护隐私的同时维持解释质量，成为平衡二者的核心方案，但噪声参数的优化仍需要结合具体医疗场景[81]，例如在基因检测解释中，需要确保噪声不掩盖致病性突变的关键信息。

5.2　伦理法律困境——责任界定与监管滞后的破解策略　在医学领域，AI在广泛应用的同时，也引发了知情同意、安全与透明度、算法公平性和偏见及数据隐私等一系列核心伦理挑战，加之全球在AI监管方面存在显著差异，构建系统完善的法律框架已成为当务之急[82]。当前，医学AI主要面临两大困境：（1）医疗事故责任划分。当AI诊断错误导致患者损害时，现行法律难以明确开发者（算法错误）、医疗机构（部署审核）、临床医生（决策采纳）的责任边界。欧洲联

盟出台的《人工智能法案》要求医疗AI提供“决策证据链”，推动企业采用区块链技术记录模型训练数据来源、算法迭代历史及临床应用反馈，为责任追溯提供技术支撑[83]。(2)监管标准缺失。当前各国对医学AI可解释性的评估缺乏统一指标，例如FDA中仅12%的AI设备提交解释质量报告[84]。为此，2023年9月，科技部、教育部等多部门联合印发《科技伦理审查办法(试行)》[85]，明确从事生命科学、医学、AI等科技活动的单位设立科技伦理(审查)委员会的必要性。今后，还需要建立包含“临床相关性”“可理解性”“忠实度”的量化评估体系(“系统因果性量表”[84]等)，并制订分级认证标准(基础解释级、深度因果级等)，从而规范市场准入，推动医学AI行业稳健发展。

5.3 *临床转化障碍——认知鸿沟与系统整合的解决路径* (1)医工协作壁垒：AI开发者与临床医生的知识体系差异导致解释需求错位。例如，医生更关注“该特征是否符合病理生理学机制”，而现有技术多提供“特征重要性排名”，因此需要建立“临床需求驱动”的解释框架，通过自然语言生成技术将技术解释转化为病理生理学机制分析[86-87]。同时，加强医工交叉培训，建立“临床-算法”双导师培养体系，从源头促进解释技术的针对性研发。(2)系统整合挑战：医院信息系统与AI解释模块的接口标准化不足，导致解释结果难以嵌入电子病历。斯坦福大学DEPLOYR框架通过HL7 FHIR标准实现AI解释的实时传输与结构化存储，为跨系统整合提供参考[88]。未来，需要推动建立全国统一的医疗AI解释接口规范，促进不同厂商系统的相互操作性。

5.4 *未来技术演进方向——从单一技术到生态构建的前瞻布局* (1)内生可解释架构创新：已有学者研发了“自解释AI”，将解释生成模块作为模型训练的内生组件，例如在神经网络中引入“解释损失函数”，强制模型学习可视化的决策模式，从源头避免黑箱问题[89]。(2)因果不变性学习：融合因果推理与表示学习，旨在确保模型解释在数据分布偏移下仍保持稳定与有效，使模型解释不受数据分布偏移影响，例如通过干预不变特征提取，确保AI在不同地域、不同人群中的解释逻辑一致，提升跨场景可靠性[48]。(3)动态自适应解释系统：开发基于用户画像的智能解释引擎，根据医生的亚专科背景、患者的教育程度动态调整解释深度，实现“千人千面”的精准解释，例如为基层医生提供基于指南的简化规则，为专科医生提供分子机制级别的因果分析[90]。

5.5 *生态构建展望——技术、伦理、监管的协同进化* 未来医学AI可解释性的发展需要构建“技术研发-临床验证-政策规制”的协同生态。在技术层面上，推动多学科融合，将生物医学知识、临床指南嵌入AI模型，提升解释的生物学合理性；在应用层面上，建立“临床场景定义-解释工具开发-实证效果反馈”的闭环，例如，通过真实世界研究验证XAI对医生决策效率的提升；在政策层面上，加快制订可解释性技术标准与伦理指南，明确不同风险等级AI的解释要求，例如对高风险的手术机器人AI实施“解释实时审计”制度。通过系统性推进，医学AI将逐步实现从“结果输出”到“过程透明”的模式转变，最终构建“可解释、可信赖、可协同”的智能医疗生态，为精准医疗的落地提供坚实支撑。

6 小 结

医学AI的可解释性研究不仅是技术问题，更是融合临床需求、伦理规范、监管要求的系统性工程。当前，XAI技术已从单一的特征可视化发展到多模态因果推理，从辅助解释工具进化为决策过程的有机组成部分。尽管面临数据异质性、责任界定等挑战，但其在提升临床信任、保障患者安全、促进技术合规等方面的价值已得到充分验证。随着技术创新与生态构建的协同推进，可解释性将成为医学AI的必备属性，推动这一技术从“实验室创新”转化为“临床刚需”，最终实现“AI赋能医疗，解释构建信任”的终极目标。

参 考 文 献

[1] Hendrix W, Hendrix N, Scholten ET, et al. Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest CT scans[J]. Commun Med(Lond), 2023, 3(1): 156.

[2] Li X, Xu X, Xie F, et al. A time-phased machine learning model for real-time prediction of sepsis in critical care[J]. Crit Care Med, 2020, 48(10): e884-e888.

[3] Zheng F, Wang L, Pang Y, et al. ShockSurv: a machine learning model to accurately predict 28-day mortality for septic shock patients in the intensive care unit[J]. Biomed Signal Process Control, 2023, 86(Part A): 105146.

[4] Qiu YG, Cheng FX. Artificial intelligence for drug discovery and development in Alzheimer's disease[J]. Curr Opin Struct Biol, 2024, 85: 102776.

[5] Cheng F, Wang F, Tang J, et al. Artificial intelligence and open science in discovery of disease-modifying medicines for Alzheimer's disease[J]. Cell Rep Med, 2024, 5(2): 101379.

[6] Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence[J]. J Consum Res, 2019, 46(4): 629-650.

[7] Nikolaev A, McLaughlin T, O'Leary DD, et al. RETRACTED ARTICLE: APP binds Dr6 to trigger axon pruning and neuron death *via* distinct caspases[J]. Nature, 2009, 457(7232): 981-989.

[8] Muehlematter UJ, Bluethgen C, Vokinger KN. FDA-cleared artificial intelligence and machine learning-based medical devices and their 510(k) predicate networks[J]. Lancet Digit Health, 2023, 5(9): e618-e626.

[9] Lundberg SM, Lee SI. A unified approach to interpreting model predictions[C]//von Luxburg U. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 4768-4777.

[10] Ribeiro MT, Singh S, Guestrin C. Why should I trust you? ": explaining the predictions of any classifier[DB/OL]. (2016-08-09)[2025-06-30]. http://arxiv.org/abs/1602.04938.

[11] Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks *via* Gradient-based localization[J]. Int J Comput Vis, 2020, 128(2): 336-359.

[12] Chattopadhyay A, Sarkar A, Howlader P, et al. Grad-CAM++: Improved visual explanations for deep convolutional networks[C]// IEEE. Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe: IEEE, 2018: 839-847.

[13] Binder A, Montavon G, Lapuschkin S, et al. Layer-wise relevance propagation for neural networks with local renormalization layers[C]//Villa AEP, Masulli P, Rivero AJP. Artificial Neural Networks and Machine Learning-ICANN 2016. Cham: Springer, 2016: 63-71.

[14] Chanda T, Hauser K, Hobelsberger S, et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma[J]. Nat Commun, 2024, 15(1): 524.

[15] 隗冰芮, 薛 鹏, 江 宇, 等. 世界卫生组织《医疗卫生中人工智能的伦理治理》指南及对中国的启示[J]. 中华医学杂志, 2022, 102(12): 833-837.

[16] Holzinger A, Malle B, Saranti AN, et al. Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI[J]. Inf Fusion, 2021, 71: 28-37.

[17] Muhammad D, Bendechache M. Unveiling the black box: a systematic review of explainable artificial intelligence in medical image analysis[J]. Comput Struct Biotechnol J, 2024, 24: 542-560.

[18] Okada Y, Ning Y, Ong MEH. Explainable artificial intelligence in emergency medicine: an overview[J]. Clin Exp Emerg Med, 2023, 10(4): 354-362.

[19] Li JJ, Guan ZY, Wang J, et al. Integrated image-based deep learning and language models for primary diabetes care[J]. Nat Med, 2024, 30(10): 2886-2896.

[20] Luo Y, Tian Y, Shi M, et al. Harvard glaucoma fairness: aretinal nerve disease dataset for fairness learning and fair identity normalization[J]. IEEE Trans Med Imaging, 2024, 43(7): 2623-2633.

[21] Lu W, Zhao L, Wang S, et al. Explainable and visualizable machine learning models to predict biochemical recurrence of prostate cancer[J]. Clin Transl Oncol, 2024, 26(9): 2369-2379.

[22] Wollek A, Graf R, Čečatka S, et al. Attention-based saliency maps improve interpretability of pneumothorax classification[J]. Radiol Artif Intell, 2023, 5(2): e220187.

[23] Bellamy RKE, Dey K, Hind M, et al. AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias[DB/OL]. (2018-10-03)[2025-06-30]. http://arxiv.org/abs/1810.01943.

[24] Wang X, Yang CC. Balancing fairness and performance in healthcare AI: a gradient reconciliation approach[DB/OL]. (2025-04-19)[2025-06-30]. http://arxiv.org/abs/2504.14388.

[25] Tsopra R, Fernandez X, Luchinat C, et al. A framework for validating AI in precision medicine: considerations from the European ITFoC Consortium[J]. BMC Med Inform Decis Mak, 2021, 21(1): 274.

[26] Jin W, Li X, Fatehi M, et al. Guidelines and evaluation of clinical explainable AI in medical image analysis[J]. Med Image Anal, 2023, 84: 102684.

[27] Reddy S. Explainability and artificial intelligence in medicine[J]. Lancet Digit Health, 2022, 4(4): e214-e215.

[28] Schwab K, Nguyen D, Ungab GA, et al. Artificial intelligence MacHIne learning for the detection and treatment of atrial fibrillation guidelines in the emergency department setting (AIM HIGHER): assessing a machine learning clinical decision support tool to detect and treat non-valvular atrial fibrillation in the emergency department[J]. J Am Coll Emerg Physicians Open, 2021, 2(4): e12534.

[29] Dalmolin M, Azevedo KS, De Souza LC, et al. Feature selection in cancer classification: utilizing explainable artificial intelligence to uncover influential genes in machine learning models[J]. AI, 2025, 6(1): 2.

[30] Bhattacharya A, Ooge J, Stiglic G, et al. Directive explanations for monitoring the risk of diabetes onset: introducing directive data-centric explanations and combinations to support what-if explorations[C]//IUI. Proceedings of the 28th International Conference on Intelligent User Interfaces. New York: Association for Computing Machinery, 2023: 204-219.

[31] Guo Y, You J, Zhang Y, et al. Plasma proteomic profiles predict future dementia in healthy adults[J]. Nature Aging, 2024, 4(2): 247-260.

[32] Donmez TB, Kutlu M, Mansour M, et al. Explainable AI in action: a comparative analysis of hypertension risk factors using SHAP and LIME[J]. Neural Comput Appl, 2025, 37(5): 4053-4074.

[33] Aoki H, Miyazaki Y, Anzai T, et al. Deep convolutional neural network for differentiating between sarcoidosis and lymphoma based on ^{18}FFDG maximum-intensity projection images[J]. Eur Radiol, 2024, 34(1): 374-383.

[34] Horry M, Chakraborty S, Pradhan B, et al. Deep mining Generation of lung cancer malignancy models from chest x-ray images[J]. Sensors (Basel), 2021, 21(19): 6655.

[35] van der Waa J, Nieuwburg E, Cremers A, et al. Evaluating XAI: a comparison of rule-based and example-based explanations[J]. Artif Intell, 2021, 291: 103404.
[36] Mahbooba B, Timilsina M, Sahal R, et al. Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model[J]. Complexity, 2021, 2021(1): 6634811.
[37] Palm J, Alaid S, Ammon D, et al. Leveraging electronic medical records to evaluate a computerized decision support system for *Staphylococcus* bacteremia[J]. NPJ Digit Med, 2025, 8(1): 180.
[38] Albuquerque T, Yüce A, Herrmann MD, et al. Characterizing the interpretability of attention maps in digital pathology[DB/OL]. (2024-07-02)[2025-05-03]. https://arxiv.org/abs/2407.02484.
[39] García-Barragán Á, Sakor A, Vidal ME, et al. NSSC: a neuro-symbolic AI system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes[J]. Med Biol Eng Comput, 2025, 63(3): 749-772.
[40] Wang FA, Zhuang ZF, Gao F, et al. TMO-Net: an explainable pretrained multi-omics model for multi-task learning in oncology[J]. Genome Biol, 2024, 25(1): 149.
[41] Captier N, Lerousseau M, Orlhac F, et al. Integration of clinical, pathological, radiological, and transcriptomic data improves prediction for first-line immunotherapy outcome in metastatic non-small cell lung cancer[J]. Nat Commun, 2025, 16(1): 614.
[42] Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision[DB/OL]. (2021-02-26)[2025-06-30].http://arxiv.org/abs/2103.00020.
[43] Yang R, Marrese Te, Ke Y, et al. Integrating UMLS knowledge into large language models for medical question answering[DB/OL]. (2023-10-13)[2025-06-30]. http://arxiv.org/abs/2310.02778.
[44] Piccininni M, Konigorski S, Rohmann JL, et al. Directed acyclic graphs and causal thinking in clinical risk prediction modeling[J]. BMC Med Res Methodol, 2020, 20(1): 179.
[45] Upadhyaya P, Zhang K, Li C, et al. Scalable causal structure learning: scoping review of traditional and deep learning algorithms and new opportunities in biomedicine[J]. JMIR MedInform, 2023, 11: e38266.
[46] Zhu JY, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// IEEE. 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2242-2251.
[47] Mertes S, Huber T, Weitz K, et al. GANterfactual-counterfactual explanations for medical non-experts using generative adversarial learning[J]. Front Artif Intell, 2022, 5: 825565.
[48] Cheng Y, Song X, Wang Z, et al. Causally-informed deep learning towards explainable and generalizable outcomes prediction in critical care[DB/OL]. (2025-02-04)[2025-06-30]. https://arxiv.org/abs/2502.02109.
[49] Wang M, You C, Zhang W, et al. Causal ECGNet: leveraging causal inference for robust ECG classification in cardiac disorders[J]. Front Physiol, 2025, 16: 1543417.
[50] Gao Y, Li R, Croxford E, et al. Leveraging medical knowledge graphs into large language models for diagnosis prediction: design and application study[J]. JMIR AI, 2025, 4: e58670.
[51] Liu XH, Liu H, Yang GX, et al. A generalist medical language model for disease diagnosis assistance[J]. Nat Med, 2025, 31(3): 932-942.
[52] Xu X, Li M, Tao C, et al. A survey on knowledge distillation of large language models[DB/OL]. (2024-02-20)[2025-06-30]. http://arxiv.org/abs/2402.13116.
[53] Liu Q, Wu X, Zhao X, et al. Large language model distilling medication recommendation model[DB/OL]. (2024-02-05)[2025-06-30].http://arxiv.org/abs/2402.02803.
[54] Lu Q, Li R, Sagheb E, et al. Explainable diagnosis prediction through neuro-symbolic integration[J]. AMIA Jt Summits Transl Sci Proc, 2025, 2025: 332-341.
[55] Tang S, Modi A, Sjoding MW, et al. Clinician-in-the-loop decision making: reinforcement learning with near-optimal set-valued policies[C]//Proceedings of the 37th International Conference on Machine Learning. Vienna, 2020: 9387-9396.
[56] Huang J, He R, Khan DZ, et al. SurgicalVLM-Agent: towards an interactive AI co-pilot for pituitary surgery[DB/OL]. (2025-03-12)[2025-06-30]. http://arxiv.org/abs/2503.09474.
[57] Bhavani P, Chithra PL. 3D Grad-CAM in lung cancer images using deep learning techniques[C/OL]// Vijayalakshmi S, Jacob L, Savithri M, et al. Proceedings of the 1st International Conference on Artificial Intelligence, Communication, IoT, Data Engineering and Security. Lavasa: EAI, 2024. https://eudl.eu/doi/10.4108/eai.23-11-2023.2343226.
[58] Jammal M, Saab A, Abi Khalil C, et al. Impact on clinical guideline adherence of orient-COVID, a clinical decision support system based on dynamic decision trees for COVID19 management: a randomized simulation trial with medical trainees[J]. Int J Med Inform, 2025, 19: 105772.
[59] Anderson AB, Wedin R, Fabbri N, et al. External validation of PATHFx version 3.0 in patients treated surgically and nonsurgically for symptomatic skeletal metastases[J]. Clin Orthop Relat Res, 2020, 478(4): 808-818.
[60] Tanzi L, Piazzolla P, Porpiglia F, et al. Real-time deep learning semantic segmentation during intra-operative surgery for 3D augmented reality assistance[J]. Int J Comput Assist Radiol Surg, 2021, 16(9): 1435-1445.
[61] O'Sullivan S, Janssen M, Holzinger A, et al. Explainable artificial intelligence (XAI): closing the gap between image analysis and navigation in complex invasive diagnostic procedures[J]. World J Urol, 2022, 40(5): 1125-1134.

[62] Noviandy TR, Maulana A, Zulfikar T, et al. Explainable artificial intelligence in medical imaging: a case study on enhancing lung cancer detection through CT images [J]. Indonesian Journal of Case Reports, 2024, 2(1): 6-14.
[63] An F, Li X, Ma X. Medical image classification algorithm based on visual attention Mechanism-MCNN [J]. Oxid Med Cell Longev, 2021, 2021: 6280690.
[64] Yan F, Wu J, Li J, et al. PathOrchestra: a comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks [DB/OL]. (2025-03-31) [2025-05-04]. https://arxiv.org/abs/2503.24345v1.
[65] PathAI. Pathology Transformed [DB/OL]. [2025-06-30]. https://www.pathai.com/.
[66] Diao JA, Wang JK, Chui WF, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes [J]. Nat Commun, 2021, 12(1): 1613.
[67] Alomar A, Alazzam M, Mustafa H, et al. Lung cancer detection using deep learning and explainable methods [C]//2023 14th International Conference on Information and Communication Systems (ICICS). Irbid, Jordan: IEEE, 2023: 1-4.
[68] Tal E. Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare [C]//Rossi F. Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. New York: MachineryPublishing Co., 2023: 312-321.
[69] Nguyen TL, Collins GS, Landais P, et al. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials-an illustration with the International Stroke Trial [J]. J Clin Epidemiol, 2020, 125: 47-56.
[70] Li H, Yu S, Principe J. Causal recurrent variational autoencoder for medical time series generation [DB/OL]. (2023-01-16) [2025-06-30]. http://arxiv.org/abs/2301.06574.
[71] Ren F, Aliper A, Chen J, et al. A small-molecule TNIK inhibitor targets fibrosis in preclinical and clinical models [J]. Nat Biotechnol, 2025, 43(1): 63-75.
[72] Ye W, LiI C, Xie Y, et al. Causal intervention for measuring confidence in drug-target interaction prediction [DB/OL]. (2023-11-14) [2025-06-30]. http://arxiv.org/abs/2306.00041.
[73] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873): 583-589.
[74] Yang ZY, Zeng AA, Zhao Y, et al. AlphaFold2 and its applications in the fields of biology and medicine [J]. Signal Transduct Target Ther, 2023, 8(1): 115.
[75] Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network [J]. Science (1979), 2021, 373(6557): 871-876.
[76] Wang Y, Pang C, Wang YZ, et al. Retrosynthesis prediction with an interpretable deep-learning framework based on molecular assembly tasks [J]. Nat Commun, 2023, 14(1): 6155.
[77] Michoel T, Zhang JD. Causal inference in drug discovery and development [J]. Drug Discov Today, 2023, 28(10): 103737.
[78] Yu M, Li W, Yu Y, et al. Deep learning large-scale drug discovery and repurposing [J]. Nat Comput Sci, 2024, 4(8): 600-614.
[79] Glocker B, Robinson R, Castro DC, et al. Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects [DB/OL]. (2019-10-10) [2025-06-30]. http://arxiv, 2019: 04597.
[80] Gupta S, Sutar V, Singh V, et al. FedAlign: federated domain generalization with cross-client feature alignment [DB/OL]. (2025-01-26) [2025-06-30]. http://arxiv.org/abs/2501.15486.
[81] Bao W, Bauer L A, Bindschaedler V. On the importance of architecture and feature selection in differentially private machine learning [DB/OL]. (2022-05-13) [2025-06-30]. https://arxiv.org/abs/2205.06720.
[82] Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare [M]/Bohr A, Memarzadeh K. Artificial intelligence in healthcare. Pittsburgh: Academic Press, 2020: 295-336.
[83] Wu K, Wu E, Rodolfa K, et al. Regulating AI adaptation: an analysis of AI medical device updates [DB/OL]. (2024-06-24) [2025-04-12]. https://www.medrxiv.org/content/10.1101/2024.06.26.24309506v1.
[84] Holzinger A, Carrington A, Müller H. Measuring the quality of explanations: the system causability scale (SCS): comparing human and machine explanations [J]. Kunstliche Intell (Olderbourg), 2020, 34(2): 193-198.
[85] 科技部,教育部,工业和信息化部,等.科技伦理审查办法(试行)[EB/OL]. (2023-09-07) [2025-05-05]. https://www.gov.cn/zhengce/zhengceku/202310/content_6908045.htm.
[86] Young JB, Abraham WT, Albert NM, et al. Relation of low hemoglobin and anemia to morbidity and mortality in patients hospitalized with heart failure (insight from the OPTIMIZE-HF registry) [J]. Am J Cardiol, 2008, 101(2): 223-230.
[87] Felker GM, Adams KFJ, Gattis WA, et al. Anemia as a risk factor and therapeutic target in heart failure [J]. J Am Coll Cardiol, 2004, 44(5): 959-966.
[88] Corbin CK, Maclay R, Acharya A, et al. DEPLOYR: a technical framework for deploying custom real-time machine learning models into the electronic medical record [J]. J Am Med Inform Assoc, 2023, 30(9): 1532-1542.
[89] Hou J, Liu S, Bie Y, et al. Self-eXplainable AI for medical image analysis: a survey and new outlooks [DB/OL]. (2024-10-03) [2025-06-30]. http://arxiv.org/abs/2410.02331.
[90] Liu X. Adaptive explainable AI: designing user-centric explanation systems for enhanced interaction [D]. Austin: University of Texas at Austin, 2024.

（收稿日期：2025-05-20　修回日期：2025-07-24）