ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 3,May 2025,pp. 562－584
DOI:10. 16337/j. 1004-9037. 2025. 03. 002

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

医疗大模型发展现状与展望

钱　波[1,2]，李富江[1,2]，郑常乐[1,2]，张道强[1,2]

(1. 南京航空航天大学人工智能学院,南京 211106;2. 南京航空航天大学脑机智能技术教育部重点实验室,南京 211106)

摘　要：医疗大模型是大规模预训练模型技术在医疗领域的重要应用成果,已成为智能辅助医疗的重要研究方向。通过在海量医学数据上进行预训练,这类模型展现出跨任务迁移、多模态理解和复杂推理等关键能力,突破了传统神经网络在医学应用中的多项限制。借助这些能力,医疗大模型正在重塑辅助诊断、病例报告生成和医学影像分析等核心任务的实现路径,对实现医疗“通用智能”具有深远意义。基于此,本文对医疗大模型的发展现状与未来趋势进行综述。首先,回顾了医疗人工智能模型在人工智能快速演进背景下的发展历程;其次,重点介绍了大模型在病理学、眼科和脑疾病等医学子领域的研究进展;最后探讨了当前医疗大模型面临的挑战,并展望其未来的发展方向。

关键词：医疗大模型;人工智能;预训练模型;辅助医疗

中图分类号：TP183　　**文献标志码**：A

A Review of Development and Future Directions of Medical Foundation Models

QIAN Bo[1,2], LI Fujiang[1,2], ZHENG Changle[1,2], ZHANG Daoqiang[1,2]

(1. School of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China;2. Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Medical foundation models represent a significant application of large-scale pre-trained model technology in the healthcare domain and have become a key research focus in intelligent medical assistance. By leveraging pretraining on vast amounts of medical data, these models exhibit critical capabilities such as cross-task transfer, multimodal understanding, and complex reasoning, overcoming several limitations of traditional neural networks in medical applications. With these capabilities, medical foundation models are reshaping the implementation of core tasks such as assisted diagnosis, clinical report generation, and medical image analysis. They hold profound implications for achieving general intelligence in healthcare. Based on this, this paper provides a comprehensive review of the current state and future trends of medical foundation models. First, it reviews the development of medical AI models in the context of rapid advancements in artificial intelligence. Then, it highlights research progress of large models in medical subfields such as pathology, ophthalmology, and neurological disorders. Finally, it discusses the challenges currently faced by medical foundation models and explores their future development directions.

Key words: medical foundation model; artificial intelligence; pre-trained model; auxiliary medical care

---

基金项目:国家自然科学基金(62136004);江苏省重点研发计划(BE2022842)。
收稿日期:2025-03-30;**修订日期**:2025-05-11

引　言

Transformer架构[1]的出现标志着人工智能进入新的技术范式,为大模型的出现奠定了技术基础。该架构通过引入多头自注意力机制,在突破序列长度对模型建模能力限制的同时,实现了高效的并行处理,显著提升了模型的上下文理解能力和训练效率。基于这一设计理念,随后涌现出一系列参数规模不断扩大的模型,在视觉、语言及多模态等任务中表现出卓越性能,为构建跨模态、跨任务的通用模型提供了技术支撑,推动了大模型技术的发展。

另一方面,人工智能进入以大规模预训练模型(Large-scale pre-trained model,LPM)为代表的大模型时代,不仅得益于算法结构的突破,更深层地源于数据规模和算力这两大基础要素的同步突破。数据被视为训练大模型的"燃料",其规模和质量直接决定了模型学习能力的上限[2]。随着互联网、物联网以及数字化技术的普及,全球数据量呈指数级增加。据预测,全球数据总量在2025年将达到175 ZB[3]。与此同时,英伟达A100/H100等GPU集群的浮点运算能力提升至每秒千万亿次,使训练千亿参数模型成为可能。以BERT[4]和GPT[5]系列为代表的预训练大模型相继涌现,在多项自然语言处理(Natural language processing,NLP)任务上达到甚至超越人类水平,推动人工智能进入"大模型时代"。

这一技术发展迅速从自然语言处理领域扩展至各个垂直领域,其中医疗健康被视为最具潜力的落地场景之一。由于医疗任务具有高度复杂性,其往往涉及到不同模态数据的整合、个体化诊疗方案的制定以及医学知识的不断更新。传统的人工智能方法在应对这些挑战时常常面临数据孤岛和模型泛化能力不足等问题。而医疗大模型则在多任务处理、医疗知识泛化、上下文推理和跨模态数据融合等方面表现出显著优势。例如,在中文医学场景中,于2024年发布的DeepSeek-R1[6]因具备广泛的医学知识、任务适配性和多轮推理能力已逐步进入实际临床应用阶段。首都医科大学附属北京中医医院本地化部署DeepSeek-R1,推出了"临床智能助手Copilot"和"医问DeepSeek",分别服务于病历质控与辅助诊断环节。东软集团则与多家医院合作,将DeepSeek嵌入其智能医疗解决方案,实现了从接诊、问询到辅助决策的端到端流程重构。这些实例表明,大模型正在从实验室走向临床实践,在提升医疗效率和增强诊疗准确性等方面展现出巨大潜力。

医疗大模型不仅提升了模型的表征能力与泛化能力,也为医疗人工智能从"窄任务专用"向"通用智能"转变提供了技术路径。当前,医疗大模型正加速实现从文本理解、图像分析到多模态协同的统一建模框架,赋能疾病识别、临床决策支持和健康管理等核心场景。基于此背景,本文系统梳理了大模型技术与医疗人工智能融合的发展路径,并对当前一些具有代表性的医疗大模型进行介绍,最后探讨了医疗大模型所面临的挑战和未来发展方向。

1　医疗AI模型发展历史

医疗人工智能(Medical artificial intelligence, Medical AI)自20世纪中叶萌芽以来,经历了多个重要发展阶段,不断推动临床诊断的智能化和自动化。图1展示了不同阶段医疗AI的发展历程。在早期的萌芽阶段,医疗AI以专家系统为代表,通过规则推理和知识库技术初步实现了疾病诊断和治疗建议的智能辅助。随后进入机器学习阶段,医学研究者开始采用基于数据驱动的方法,如支持向量机(Support vector machine, SVM)[7]等算法,在影像识别、生物医学信号处理和电子病历分析等领域取得显著进展。2012年起,随着卷积神经网络(Convolutional neural networks, CNNs)[8]的流行,医疗AI迈入深度学习阶段,模型性能大幅提升,尤其在医学影像分类和病理识别等任务中展现了超越传统机器学习方法的性能。而进入大模型阶段后,预训练大模型成为医疗AI新的核心动力。模型的跨模态理解、统一推理和领域知识融合等能力不断增强,为医学影像分析、医学问答系统和临床决策支持等应用开辟了新的可能性。

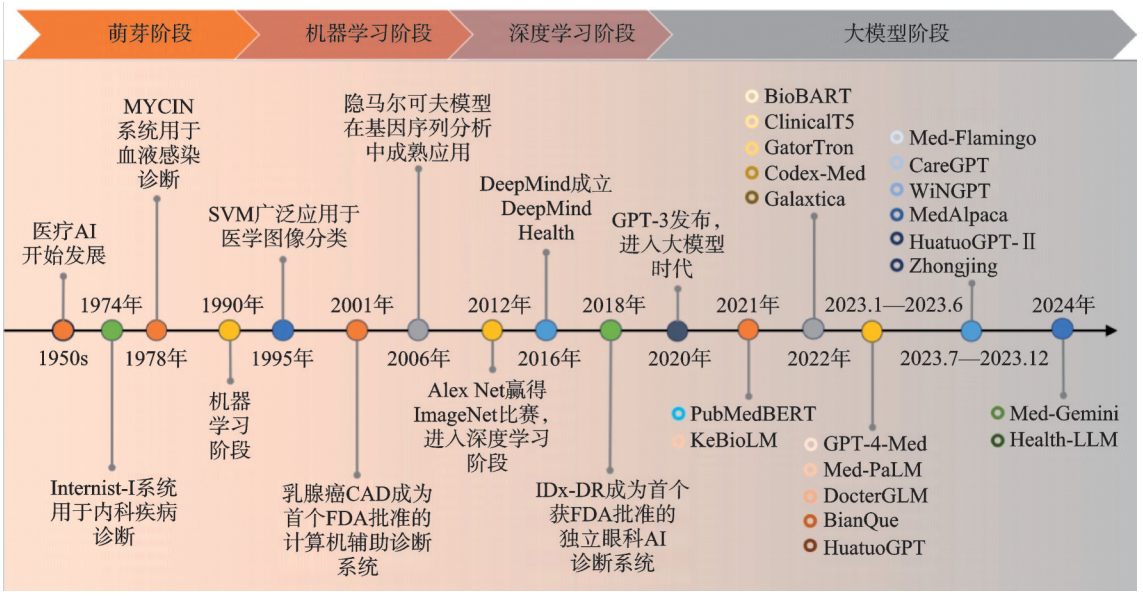

图1 医疗AI发展历史

Fig.1 Development timeline of medical AI

1.1 萌芽阶段

在人工智能发展的早期阶段，计算机的应用主要集中在数值计算和数据处理方面。对于医学领域而言，解决问题的方法不是数学模型或数据处理程序，而是定性推理技术，即通过判断规则或启发式方法将理论知识与实际问题相连接。医生在诊断过程中，常常基于经验规则、生理学知识和临床联想进行推理。这一时期的医疗人工智能的核心算法思想为知识显示表示和符号主义推理，通过人工构建的规则库和推理机实现诊断，其中最具代表性的形式是专家系统[9]。专家系统通过由领域专家编写的大量“如果-那么”规则，模拟临床诊断中的逻辑推理过程，实现对症状、检查结果与潜在疾病之间关系的形式化表达。例如，斯坦福大学于20世纪70年代开发了MYCIN系统[10]，将其用于血液感染和细菌性疾病的诊断。该系统使用了基于规则的知识库和基于不确定性的模糊推理策略，在模拟医生诊断流程方面取得了一定成功[11]。尽管在某些特定任务上表现出一定能力，但专家系统的构建验证依赖于人工编写和维护的成千上万条规则。而一旦知识库规模扩大，规则之间的冲突、冗余和一致性问题将显著增加[12]。其次，专家系统普遍缺乏对不确定性、异质性和高维度复杂医疗数据模式的建模能力，难以处理现实临床环境中的模糊症状和个体差异[13]。

1.2 统计学习阶段

随着数字化技术的广泛应用，医院信息系统(Hospital information system, HIS)和电子健康记录(Electronic health record, EHR)在医疗机构中逐渐普及，医学数据逐步从传统纸质文档转向数字化存储。得益于数字革命带来的低成本、高效率数据采集与存储手段，现代医院配备了完善的监控与数据收集设备，使得医疗数据得以持续积累与共享[14]。大量医学数据的涌现，为统计学习方法在医疗人工智能中的应用奠定了坚实基础。

在统计学习阶段，医疗AI系统主要依赖传统机器学习算法，基于监督学习范式[15-16]实现疾病诊断与风险预测等任务。其基本原理是在大量带标签的结构化医疗数据上学习一个从输入特征到目标变量的映射关系，并通过最小化损失函数来优化模型参数，使得预测结果尽可能接近真实标签。广泛使

用的机器学习算法包括支持向量机、逻辑回归、决策树和随机森林等方法[17]。这些算法具备良好的泛化能力和可解释性,适用于小样本和结构化特征明确的医疗场景[18]。例如,在糖尿病诊断任务中,Delen等[19]使用SVM模型对乳腺癌的良恶性进行了分类,显著提升了分类准确率。Williams等[20]使用J48决策树和朴素贝叶斯算法对乳腺癌发病风险进行建模,分别取得了94.2%和82.6%的预测准确率。

尽管统计学习方法在结构化数据处理方面表现优异,但也存在明显局限。一方面,这类方法高度依赖人工特征选择与构建,特征工程的质量直接影响模型效果;另一方面,统计学习方法在处理非结构化数据(如影像和文本数据)时表现不佳,难以捕捉深层语义和潜在关联,影响了模型的泛化能力和广泛适用性[21]。

1.3　深度学习阶段

2012年AlexNet[22]在ImageNet挑战赛中取得突破性成果,标志着深度学习时代的到来。该阶段的人工智能算法通过构建多层非线性的神经网络结构,使模型能够自动从原始输入数据中学习特征表示,逐层提取从底层纹理特征到高层语义信息,有效摆脱了传统机器学习算法对人工特征设计的依赖。同时,深度神经网络具备端到端的特征学习能力,相比于传统"特征提取+分类器"的建模流程,端到端模型通过统一优化大幅简化了建模流程。

在医学图像分析领域,早期研究主要集中在利用CNN处理二维图像。Setio等[23]提出多视角CNN架构,有效提升了肺结节检测的准确率。Esteva等[24]利用Inception-v3构建眼底图像分类模型,实现糖尿病视网膜病变的自动检测,其性能在多个指标上接近人类眼科专家水平。Ronneberger等[25]提出的U-Net则广泛应用于医学图像分割任务,在肝脏、脑肿瘤和心脏等器官的像素级分割中取得显著成果。郝小可等[26]提出了多尺度残差融合图卷积网络框架MSRF-GCN,实现了自闭症谱系障碍(Autism spectrum disorder, ASD)等脑疾病的辅助诊断。该模型设计了一种功能连接生成模块,能提取有远程依赖关系的时间序列相关特征,辅助定位潜在异常脑区。同时,融合人口图中的多尺度信息,提升了图结构建模的表征能力。

尽管深度学习技术推动了医学AI的快速发展,但主流模型仍然普遍具有"任务专一性强、迁移能力弱"的局限。例如,上述模型依赖大规模高质量标注数据进行监督学习,难以泛化到数据分布差异较大的其他任务场景。因此,近年来研究者逐步转向探索统一的多任务医学基础模型,以解决跨任务迁移和数据不足等核心难题。

1.4　大模型阶段

Transformer架构通过引入自注意力机制,取代了循环神经网络(Recurrent neural network, RNN)[27]和CNN,被广泛应用于不同的深度学习任务。与RNN按时间步递归处理序列、CNN依赖固定感受野不同,Transformer能够在每一层并行地建模输入序列中所有位置间的依赖关系,从根本上突破了序列长度对模型建模能力的限制,缓解了长序列条件下反向传播中常见的梯度消失和梯度爆炸问题[28]。此外,该架构中引入的多头自注意力机制不仅增强了模型的上下文建模能力,使其能够捕捉语义空间中不同维度的特征表征,还通过投影矩阵将查询($\boldsymbol{Q}$)、键($\boldsymbol{K}$)、值($\boldsymbol{V}$)分别投影至多个子空间,从而在GPU(Graphics processing unit)的SIMD(Single instruction, multiple data)架构下实现并行处理,显著提升了训练效率。而且,由于其采用由多个相同模块堆叠而成的编码器——解码器结构,使得模型能够简单地通过增加层数和扩展宽度来增加参数的数量,为模型参数的大规模扩展提供了结构基础。在此架构的支持下,大模型技术得以快速发展,并在实践中表现出强大的扩展潜力和通用能力。例如,OpenAI所提出的GPT-3模型[29]在参数量扩展至1 750亿个后出现上下文学习和复杂推理能力,从实证角度验证了语言模型的"缩放定律"[30]。这表明当模型参数和数据量达到一定规模时,性能可以持续随

之提升。

大模型在NLP领域取得巨大成功标志着大模型时代的到来，随后通用预训练模型开始在医疗等各个领域得到广泛应用。以Transformer架构为基础的预训练模型因其具有庞大的参数量而具备在大规模医疗数据中学习通用表示的能力，然后利用下游任务数据进行微调以实现对下游任务的适配。与传统人工智能“训练-测试”的方案相比，“预训练-微调”范式显著提升各类任务性能，模型学习到的通用表征能迁移到分类、分割、推理和问答等多种任务，具有多任务泛化能力。如图2所示，大模型通过在大规模训练数据集上进行预训练学习通用表征，然后在下游专业领域数据集上进行微调Adapter模块或者直接应用于下游任务。

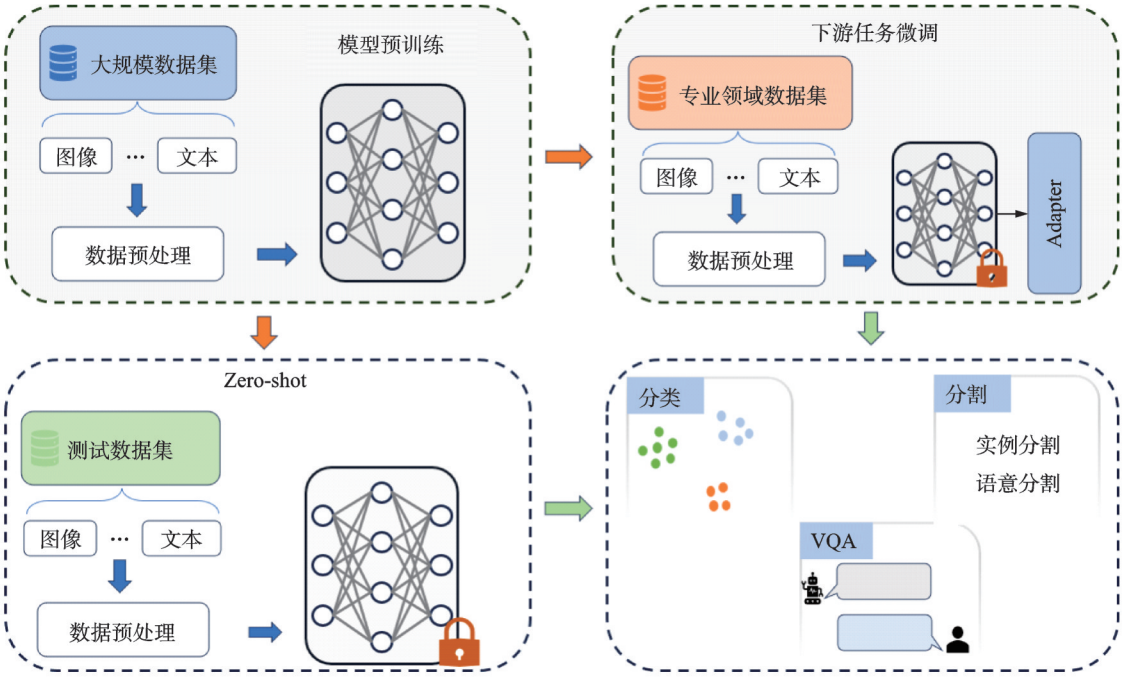

图2 大规模预训练模型工作流程

Fig.2 Workflow of large-scale pre-trained model

例如，Google DeepMind于2023年发布了Med-PaLM模型[31]，首次系统地将大语言模型用于医学知识理解与问答。其在美国医学执照考试(United States Medical Licensing Examination, USMLE)题目中的准确率高达67%以上，表现接近专业临床医生，而且在多个现实临床任务中展现出潜在实用性。在国内，以DeepSeek为代表的大模型也不断涌现。DeepSeek-R1模型拥有千亿级训练参数，采用多阶段预训练与指令微调策略，训练数据覆盖大量中文资料。更重要的是，DeepSeek-R1拥有多轮推理能力，将其应用于医疗任务时，能够在上下文推理中整合症状、检验指标和病史信息等异构数据，执行接近临床医师的推理式诊断流程。在CMExam[32]和C-Eval[33]等中文医学基准测试中，DeepSeek-R1在涉及因果解释、跨段落信息整合等高复杂度问答任务上显著超越同类模型[6]，表明模型内部已隐式学习到“分析—权衡—判断”的临床推理逻辑链条。

2 医疗大模型分类

随着大模型在自然语言处理和计算机视觉等通用领域取得突破性进展，医疗人工智能也正加速迈

向以“基础模型”为核心的新范式。传统医疗AI模型多以特定任务和特定疾病为目标,通常采用“任务驱动”的监督学习方式。但是这种学习方式需要精细标注数据来进行训练,而且模型的泛化能力十分有限。相比之下,大模型通过在大规模通用数据上进行预训练,获得具备迁移能力的通用表征,使其在仅依赖少量标注数据甚至无监督的条件下也能完成多种下游任务。

2.1　数据集和评测指标

2.1.1　数据集

为了系统评估医疗大模型在不同任务中的泛化能力与性能表现,研究者们广泛采用多个具有代表性的公共医疗数据集作为评测基准。以下是常见的医疗大模型评测数据集。

(1)MIMIC-CXR数据集[34]是由MIT发布的大规模胸部X射线影像数据集,包含约37万张图像及其文本报告,广泛应用于图文对齐、疾病分类和报告生成等任务。

(2)CheXpert数据集[35]由斯坦福大学发布,包含来自65 240名患者的224 316张胸部X光图像,涵盖14种常见胸部疾病的弱标签或不确定标签,广泛应用于多标签分类。

(3)TCGA-BRCA数据集[36]来自美国癌症基因组计划,包含来自多个中心的高分辨率乳腺癌病理切片图像、临床数据和分子信息,广泛用于乳腺癌分类、分级和区域分析任务。

(4)MedBench数据集[37]由上海人工智能实验室联合多家机构发布,包含约30万个覆盖50多个临床专科的医学问题。其数据来源于真实考试题目、临床病例和医学教材,广泛用于评估中文医学大语言模型的知识掌握、推理能力及伦理安全性等多维度表现。

(5)PAIP 2019数据集[38]是韩国主办的病理图像挑战赛数据集,其中包括病理全切片图像、病理学家标注的注释以及根据注释生成的肿瘤区域和存活肿瘤区域的真值二进制像素掩码,常用于肿瘤区域分割。

(6)EyePACS数据集[39]是Kaggle 2019糖尿病视网膜检测比赛使用的数据集。该数据集由印度的眼科医院技术人员拍摄,经验丰富的眼科医生分类标注,包含不同像素大小的视网膜眼底图像,常用于眼底图像分类。

(7)REFUGE数据集[40]是MICCAI 2018举办的视网膜青光眼挑战赛使用的数据集,其中包括视盘/视杯分割标注和青光眼分类标签,常用于眼科分割和分类任务。

(8)BraTS数据集[41]由医学影像计算与计算机辅助干预协会于2012年发起,旨在推动脑肿瘤影像分析研究。该数据集汇集了多模态核磁共振数据,每例都包含T1、T1Gd、T2和FLAIR四种序列,涵盖多种类型脑肿瘤,广泛应用于脑肿瘤分割研究。

(9)ADNI数据集[42]是美国国家卫生研究院于2004年发起的一个多模态脑部影像研究项目,旨在支持阿尔茨海默病的早期诊断与进展监测。该数据集涵盖了轻度认知障碍、阿尔茨海默病患者及健康对照者的结构MRI(Magnetic resonance imaging)、PET(Positron emission tomography)扫描、脑脊液生物标志物、基因信息和认知行为评估等多种数据类型,广泛用于多模态建模和疾病分期预测。

(10)OASIS数据集[43]由华盛顿大学医学院的研究团队发布,旨在为神经科学研究提供一个公开、高质量的脑成像数据资源。该数据集包含了来自不同年龄段的健康和认知障碍个体的脑部MRI扫描图像,广泛用于认知退化建模、阿尔茨海默病早期检测和多模态医学大模型训练任务中。

2.1.2　评价指标

在医疗AI的研究过程中,模型评估指标的选择直接影响着模型性能的可信度与可比较性。不同任务类型,如疾病分类、图像分割和文本生成等适用不同的评价标准,科学合理地选用评估指标是进行横向模型比较与结果解释的前提。不同任务的评价指标主要有以下几种。

准确率(Accuracy)是分类任务中最直观的指标,定义为分类正确的样本与总样本数的比值,数值范

围为[0,1]。公式如下

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{1}$$

式中TP(True positive)、TN(True negative)、FP(False positive)和FN(False negative)分别表示真正例、真负例、假正例和假负例。尽管准确率简单直观,能够简洁地反映模型整体预测的正确性,但在类别极度不均衡的任务中,该指标往往被多数类主导,无法真实反映模型在少数类上的识别效果。

为了更全面衡量模型的表现,还会使用精确率(Precision)、召回率(Recall)和F_1-score等指标来综合评估模型的分类性能,公式如下

$$\text{Precison}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{2}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

$$F_1\text{-score}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{4}$$

精确率强调预测为正类中真正为正的比例,而召回率强调所有正类中被正确识别的比例。F_1分数作为二者的调和均值,适用于对精确率和召回率均有较高要求的场景,能在不平衡任务中提供更公平的评价。

AUC(Area under the curve)是分类任务中广泛采用的性能评估指标,其基于ROC(Receiver operating characteristic)曲线计算得到。ROC曲线以假阳性率(False positive rate, FPR)为横轴,真正例率(True positive rate, TPR)为纵轴,刻画了模型在不同判别阈值下的分类能力。AUC表示ROC曲线下的面积,其取值范围为[0,1],数值越接近1表示模型对正负样本的区分能力越强;相反,数值接近0.5则表明模型几乎不具备有效判别能力。公式如下

$$\text{TPR}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{5}$$

$$\text{FPR}=\frac{\text{FP}}{\text{FP}+\text{TN}} \tag{6}$$

$$\text{AUC}\approx\sum_{i=1}^{n-1}\frac{\text{FPR}_{i+1}-\text{FPR}_i}{2}\times(\text{TPR}_{i+1}+\text{TPR}_i) \tag{7}$$

式中:n表示ROC曲线上离散点数量;FPR_i、TPR_i分别对应第i个离散点的横纵坐标轴数值。AUC的理论定义是对ROC曲线下区域进行积分,但由于实际应用中ROC曲线通常由有限数量的离散点构成,因此常采用梯形法则进行近似计算。AUC的显著优势在于其阈值无关性,不依赖于特定的分类阈值,能够稳定地衡量模型在整体判别能力上的表现。

Dice系数是医学图像分割任务中最常用的重叠度评估指标,主要用于量化模型预测区域与真实标注区域之间的相似性,广泛应用于病灶检测及器官分割等任务。其定义为

$$\text{Dice}=\frac{2\times|X\cap Y|}{|X|+|Y|} \tag{8}$$

式中:X表示模型预测的分割区域;Y表示真实标注区域;$|X\cap Y|$表示预测区域与真实区域的交集大小,即两个区域重叠像素点数量;$|X|$和$|Y|$分别表示预测区域和真实区域的像素点总数。Dice的数值范围为[0,1],值越接近1表示模型预测结果与真实标注越接近,模型的分割效果越好。Dice系数的优势在于其对“前景”类别有较强的关注,能较好地处理目标区域远小于背景的情况。然而它对类别不平衡数据的敏感性也可能导致在类别分布严重不均时出现评价偏差。

BLEU指标在医学报告生成、图文对齐和多模态问答等任务中被广泛用于自动化文本生成质量评估。其衡量的是生成文本与参考文本之间的n-gram匹配程度,常见的为BLEU-1到BLEU-4,分别对应1-gram到4-gram的精确度。公式如下

$$\mathrm{BLEU}=\mathrm{BP}\times\exp\left(\sum_{n=1}^{N}\omega_n\times\log p_n\right) \tag{9}$$

$$\mathrm{BP}=f(x)=\begin{cases}1 & c>r\\ \mathrm{e}^{\left(1-\frac{r}{c}\right)} & c\leqslant r\end{cases} \tag{10}$$

式中:N为使用n-gram的最大阶数;p_n为生成文本中与参考文本匹配的n-gram比例;ω_n为各阶n-gram的权重;BP为文本长度惩罚因子,其中,c表示生成文本的长度,r为参考文本的长度。BLUE数值范围在[0,1],数值越大表示生成文本与参考文本越接近。

2.2　病理大模型

作为疾病诊断的"金标准",病理学在癌症诊断与亚型分类中扮演着不可或缺的角色[44]。然而,传统病理分析高度依赖病理医生的专业经验。样本制备、切片染色、显微镜观察及报告撰写等流程繁琐,且易受主观因素干扰,难以满足大规模临床应用的需求。近年来,随着数字病理图像的积累和深度学习的发展,具备大规模参数与跨任务迁移能力的病理大模型应运而生,正逐步成为推动病理智能化转型的关键力量。

当前的病理大模型主要聚焦于肿瘤识别与分型、图像与文本的问答任务和自动化病理报告生成等任务。在实际应用中,病理全切片(Whole slide image, WSI)数据在实际分析中由于尺寸巨大、组织分布不均,传统的随机切图方式常导致关键区域被遗漏。为解决这一问题,研究者从多尺度建模与注意力机制两个维度进行优化。例如,Chen等[45]提出了病理大模型HIPT(Hierarchical image pyramid Transformer),采用coarse-to-fine的层次化Transformer架构。如图3所示,该模型首先将WSI切分为Cell-Level级别的图像块,通过Transformer构建局部区域的Patch-Level表征。随后,该模型将局部表征与Patch-Level token融合后输入第二级Transformer聚合更高层次的Region-Level表征。最后将Region-Level表征与Region-Level级别token融合输入Tranformer,生成Slide-Level表征,实现高分辨率条件下的细节捕捉。研究人员在模型预训练中使用了10 678张千兆像素WSI、408 218张4 096像素×4 096像素图像和1.04亿张256像素×256像素图像进行模型预训练,训练过程采用MAE(Masked autoencoder)[46]自监督学习策略。这种coarse-to-fine架构有效捕捉了病理图像中的组织结构与细胞层级特征,模型在癌症亚型划分和生存预测任务上的性能均优于目前最先进的方法。

基础模型代表了医学AI研发的新前沿[47]。在病理学领域,研究人员正积极探索构建具备泛化能力的病理基础模型。这些模型通过在大规模、多样化的数据集上进行预训练,并通过Zero-shot或Fine-tuning方式迁移至各类下游任务。例如,Wang等[48]提出了通用病理基础模型CHIEF,解决了不同数字化流程与人群差异性导致的模型泛化能力受限问题。该模型采用两种互补的预训练策略,通过无监督学习提取图块级显微特征,弱监督学习捕捉全切片图像的全局结构信息,实现不同层级特征的融合。研究团队在全球24家医疗机构的病理图像数据上对CHIEF进行了系统评估,实验结果表明,其在多个下游任务中比现有最优模型性能提升36.1%,表现出卓越的泛化能力。Vorontsov等[49]提出基础模型Virchow,是当时规模最大的计算病理学基础模型。该模型采用ViT-H/14架构,并采用基于学生-教师范式的DINOV2[50]自监督学习算法。学生模型通过匹配教师模型提供的类别标签学习全局表征,同时通过匹配教师模型的图像区块标记来学习局部表征,训练学生模型以匹配教师模型的表示。该模型在来自MSKCC的约10万名患者的150万张H&E染色WSI图像上进行了预训练,实验表明其在泛癌症检测与生物标志物预测方面已接近临床应用级性能。

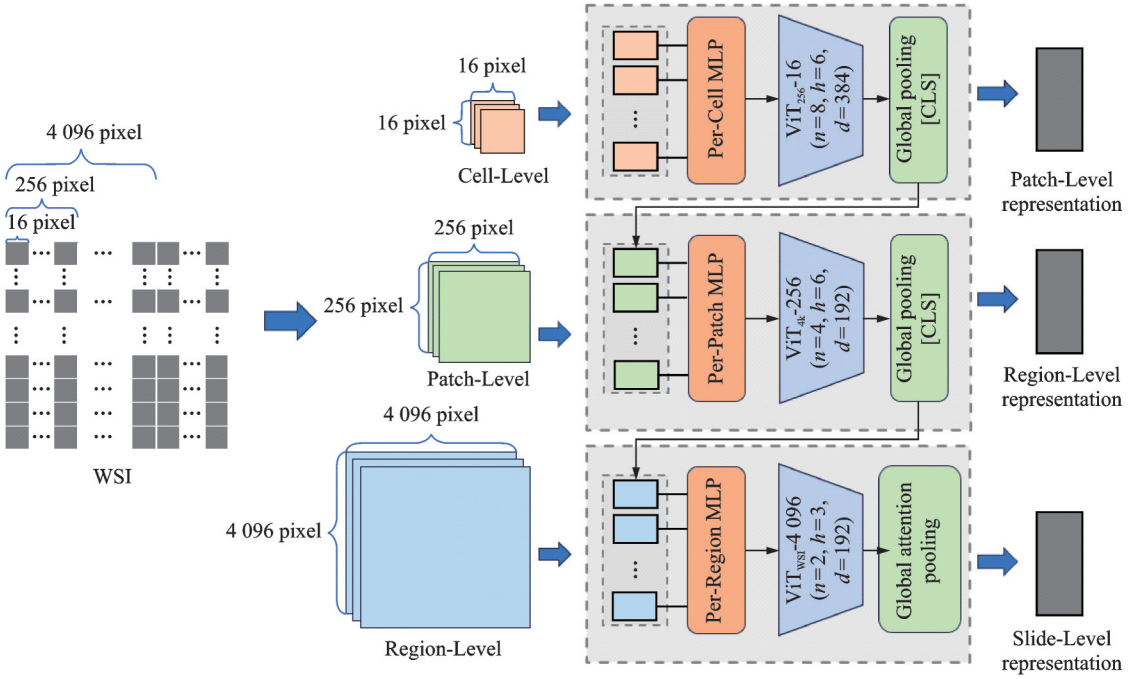

图3 HIPT模型框架图[45]

Fig.3 HIPT model architecture[45]

为克服传统病理模型在标签稀缺和单一模态方面的局限，越来越多研究人员开始研究多模态融合的病理基础模型。Huang等[51]提出的PLIP(Pathnology language-image pretraining)模型成功将视觉语言模型引入病理学。该模型联合优化WSI图像编码器与文本编码器，采用InfoNCE损失实现图像与文本特征在语义空间中的对齐。通过在包含208 414张病理图像及其自然语言描述的OpenPath数据集上进行预训练，模型展现出良好的图像与文本理解能力。Lu等[52]提出了视觉语言模型CONCH，采用对比学习策略，在117万对病理图像-文本对上进行与任务无关的预训练。在14个不同的基准测试上的结果表明，模型可迁移至涉及组织病理图像与文本的多种下游任务，在图像分类、分割以及文本转图像和图像转文本检索等任务中均取得较好性能。为进一步利用大量未标注、未配对的图像和文本数据，Xiang等[53]研究了统一掩码建模的多模态大模型MUSK。该模型首先在来自11 577名患者的5 000万张病理图像和10亿个病理相关文本标记上进行自监督预训练，随后在100万对配对的图像-文本对上进行进一步预训练，实现视觉与语言特征的高效对齐。模型无需进一步微调，即可在图像-文本检索、视觉问答、图像分类与分子生物标志物预测等23项病理图像块级和像素级基准测试中表现出卓越性能。此外，MUSK还在黑色素瘤复发预测、全癌种预后预测以及肺癌与胃食管癌免疫治疗反应预测等任务中表现优异。

2.3 眼科大模型

随着全球人口老龄化加剧以及影响眼健康的风险因素不断变化，眼科医疗服务的需求日益增长，已超过高素质眼科医生和专业医疗团队的供给能力。人工智能被视为应对这一挑战的关键技术路径。然而，现阶段的大多数眼科AI模型依赖大量人工注释数据，成本高、效率低，且多数模型仅针对单一疾病或图像模态，难以满足实际临床需求。相比之下，基础模型因其更高的效率、强适应性与良好的可扩展性，正在成为应对全球眼科挑战的新机遇。

Zhou等[54]提出了一个视网膜图像基础模型RETFound。该模型基于自监督学习方法,利用大量无标签图像提取具有泛化能力的特征,用于后续的有监督任务。在160万张未标注的视网膜图像上进行掩码自监督训练后,RETFound在眼部疾病和系统性疾病的诊断方面优于基线模型。Silva-Rodrigue等[55]提出了一种引入专家知识的通用预训练视觉语言模型FLAIR。如图4所示,该模型首先将彩色眼底图像的标签通过专家知识模块转化为细粒度、结构化的病理特征描述,作为文本监督。随后,将彩色眼底图像和专家知识分别输入图像编码器和文本编码器,利用对比学习将彩色眼底图像的层次结构、类别之间的关系以及目标疾病感兴趣区域的特征信息与图像特征实现对齐。实验结果表明,模型表现优于在具体数据集上训练的模型。而且模型的整体性能也超越了规模更大的通用视觉语言基础模型,展现出医学领域专家知识嵌入的巨大潜力。Shi等[56]提出了一种多模态视觉语言基础模型EyeCLIP。该模型通过联合眼底彩照、OCT图像及临床文本报告进行预训练,结合自监督重建、多模态图像对比学习和图文对比学习等策略,捕捉不同模态间的共享表征。在眼病分类、跨模态检索和视觉问答等任务中均取得较好的表现。Du等[57]提出RET-CLIP模型,借鉴CLIP框架,设计了“单眼级-双眼级-患者级”三重对比优化策略,用于模拟真实临床检查中左右眼以及患者整体信息的贡献。实验结果表明,该模型的诊断性能优于现有基准模型。Luo等[58]提出了一种基于最优传输的FairCLIP方法,通过减小总体样本分布与每个人口统计群体对应分布之间的辛克霍恩距离,以此来提升视觉语言模型的公平性。实验结果显示,FairCLIP在多个群体上的AUC和公平性指标上均优于原始模型。Jiang等[59]提出了一种用于早期糖尿病视网膜病变检测的实时微动脉瘤病变分割框架GlanceSeg。该方法结合了SAM(Segment anything model)基础模型和眼科医生的注释图,通过生成显著图和提示点,引导模型高效定位并分割微小病灶,实现了人机协同的无标签分割。但上述模型仅支持眼底彩照或光学相干断层扫描(Optical coherence tomography, OCT)两类图像模态,临床应用范围仍有限。为此,Qiu等[60]提出了基础模型VisionFM。该模型使用来自560 457名个体的340万张眼科图像进行自监督预训练。实验结果表明,VisionFM的表现不仅优于多个强基线深度网络,还具有诊断可解释性与跨模态泛化能力。即使面对在训练阶段未出现的新型成像设备或眼科疾病,模型仍能保持高精度预测。

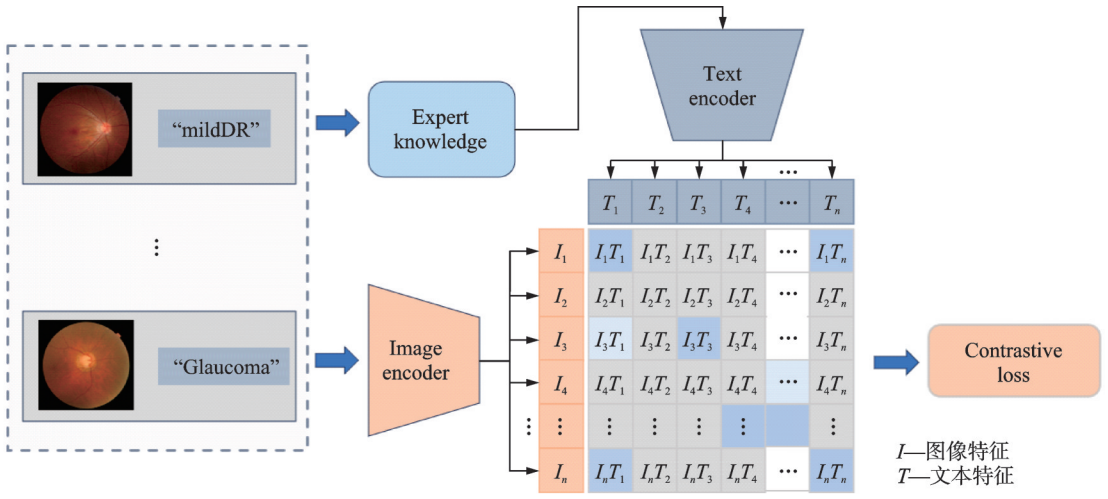

图4　FLAIR模型框架图[55]

Fig.4　FLAIR model architecture[55]

2.4　脑疾病大模型

脑疾病,如阿尔茨海默病、帕金森病、脑肿瘤和脑卒中等,其诊断通常依赖于多模态数据的综合判

断，包括MRI、CT（Computed tomography）、PET影像和EEG（Electroencephalogram）等。然而，这些任务普遍存在数据模态异构、标注成本高和样本不平衡等问题，导致传统机器学习模型的泛化能力有限。尤其在早期诊断、亚型分型和疾病进展预测等复杂任务上，传统模型的表现非常不稳定。为应对这些问题，研究人员开始探索将基础模型引入脑疾病分析，利用预训练机制提升模型的通用性与下游任务适应能力。

在MRI影像数据方面，Chen等[61]利用掩码自监督学习方法提出了一个针对增强脑部MRI扫描数据的基础模型，利用大量无标注数据提升了模型在脑肿瘤监测、鉴别和分子状态预测方面的性能。Barbano等[62]提出了一种脑部MRIs的解剖基础模型AnatCL，旨在充分利用核磁共振成像数据中的信息。该模型采用对比学习框架，对之前研究采用患者年龄作为元数据的方法进行改进，将MRI中的解剖信息和年龄一起作为元数据来指导模型学习，并在损失函数中加入额外的解剖学度量。在12种不同的下游任务中的测试结果表明，得益于MRIs中丰富的解剖学信息，模型在预训练过程中可以获得更稳健和更泛化的表征。Sun等[63]提出了一种基础模型，通过组织分类网络与组织感知增强网络协同提升MRI图像质量，从而优化了分割、配准和诊断任务。Cox等[64]提出了一种基于ViT的适用于大脑结构的大模型BrainSegFounder。该模型采用两阶段的自监督预训练策略，专注于MRI影像的空间特征分布，其在医学影像分割任务上的性能显著优于传统全监督模型。Caro等[65]基于自监督掩码训练策略和Transformer架构，提出了BrainLM基础模型。该模型在77 298名受试者的6 700小时fMRI数据上进行预训练，其方法是通过随机掩盖部分时间序列数据来训练模型预测被掩盖的部分。实验结果表明，其在对未来脑活动状态的预测任务中优于其他基线模型的性能。

在EEG信号建模方面，Jiang等[66]提出了LaBraM基础模型，以解决传统深度学习方法在泛化能力和适应性上的不足。如图5所示，模型引入EEG通道切片和向量来量化神经频谱预测，将连续的原始EEG信号编码为离散的神经代码，突破任务间EEG采集设置不一致的问题。该模型结合MAE框架与

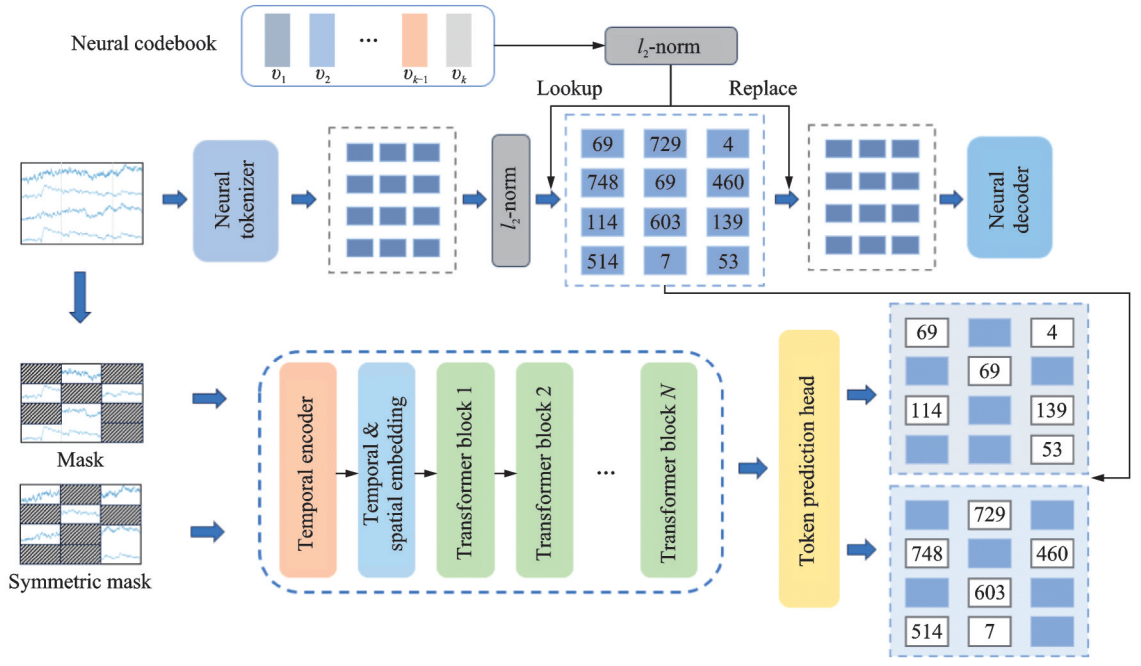

图5 LaBraM模型架构图[66]

Fig.5 LaBraM model architecture[66]

时空Transformer进行训练,构建了可迁移的EEG通用表征模型,在情绪识别和异常检测等任务中取得显著提升,推动了EEG研究从“小样本监督学习”向“大模型预训练+微调”的范式转型。Zhang等[67]则提出了Brant大模型,以Transformer作为编码器,设计了时间编码器和空间编码器,用于捕捉脑信号中的长期依赖性和空间相关性。Brant在多种下游任务,包括神经信号预测、频率-相位预测、数据填充和癫痫检测等任务中取得了最先进的性能,展现了其通用特征表征能力。

2.5　其他疾病相关大模型

除了在病理影像、眼科影像和脑疾病方向的广泛应用外,大模型技术在肺部疾病、心血管疾病、皮肤病及全身性疾病等多个医疗场景中也展现出显著的潜力。

在肺部疾病方面,Lai等[68]提出了一种视觉语言跨注意力对齐框架CARZero,以实现医学影像和诊断文本之间的高效配准。如图6所示,该框架首先采用基于大语言模型的提示对齐策略,将多样化的诊断表达转换为结构化的标准格式。随后,胸部X光图像和对应的标准化诊断描述文本分别输入图像编码器和文本编码器,提取图像和文本特征表示,并利用双向交叉注意力机制对图像特征和文本特征进行交互建模。最后,分别对融合后的特征进行线性投影,形成图像-文本相似性矩阵,实现跨模态对齐。Phan等[69]提出了一种多视角视觉语言匹配框架MAVL,旨在解决医学图像与非结构化报告中关键病理结果对齐不准确的问题。该框架将疾病描述分解为多个细粒度属性,如形状、不透明度和位置等,从而将任何新疾病的视觉外观与基础视觉知识联系起来,以提高对训练阶段未见过疾病的识别能力。Zhang等[70]提出了一种知识增强自动诊断(Knowledge-enhanced auto diagnosis, KAD)方法来应对医疗任务对细粒度和领域知识的高要求。该方法借助医学知识图谱训练知识编码器,将放射报告转化为结构化领域知识,引导图像与文本的联合表示学习,从而有效地将领域知识注入编码器,实现对任意疾病或放射学检查结果的诊断。Wu等[71]设计了一种三元组提取模块,将原始放射报告转化为包含医学实

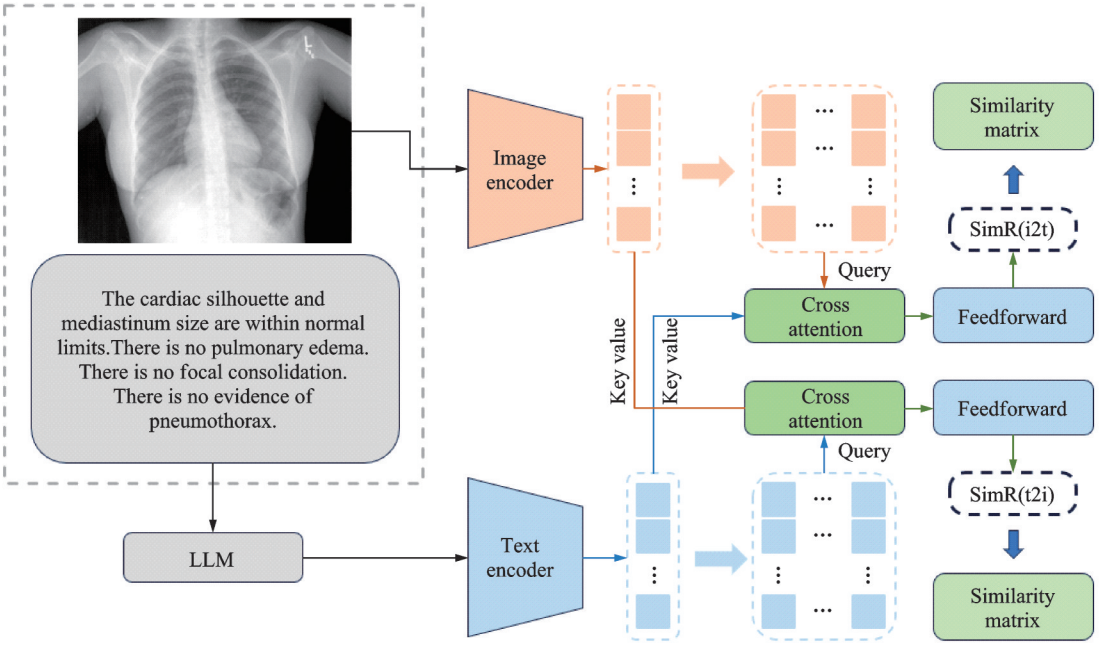

图6　CARZero模型框架图[68]

Fig.6　CARZero model architecture[68]

体的三元组,并通过查询知识库把医学实体翻译为丰富的医学领域知识,从而实现对图像patch的医学实体级别语言监督。Bannur等[72]提出的BioViL-T框架结合CNN-Transformer混合图像编码器与文本模型进行联合预训练,并利用时间相关性将之前的图像和给定的报告进行对比学习,充分利用了自监督学习来有效提升模型性能。Liu等[73]提出了一种基于多视角纵向对比学习方法。该方法通过融合多张纵向图像的空间与时间信息,并利用放射报告中固有的时空信息来监督视觉和文本表征的预训练。随后,该方法引入了标记化缺失编码技术来处理先验数据缺失的情况,使得模型能够根据可用的先验知识生成准确的放射学报告。

3 医疗AI面临的挑战

尽管医疗AI在疾病预测、辅助诊断和治疗决策等任务上展现出广阔前景,但其在实际部署和临床转化过程中仍面临一系列关键性挑战。这些挑战不仅制约了模型性能的进一步提升,而且也关系到其安全性和可靠性[74]。首先,模型可解释性的缺乏阻碍了其在高风险医疗场景中的可信应用与监管合规,使得医疗AI模型难以在临床实践中推广和应用;其次,对大规模、高质量医疗数据的依赖使得模型性能在跨机构和跨人群泛化时存在显著瓶颈;最后,构建和部署大型模型所需的高昂计算基础设施成本,进一步限制了其在资源有限地区和中小型医疗机构中的可及性。

3.1 模型可解释性差

深度学习在医学领域展现出卓越的性能,其在某些疾病的预测准确率堪比甚至超越人类专家[75]。然而,无论是CNN还是Transformer,其本质都是一种“黑箱”结构,模型缺乏可解释性和透明度[76]。但是,在医疗领域,模型诊断的可解释性尤为重要。医生在采纳AI辅助建议时,往往需要理解模型的推理过程与依据。“黑箱”特性使得模型预测结果难以追溯,在高风险医疗场景中可能造成严重后果[77],从而限制了先进的深度学习模型在临床实践中的推广与应用。

为了克服这一问题,人们提出了一些关于可解释性的研究。针对可解释性的研究主要集中在内在可解释和事后解释[78]。前者聚焦于模型结构本身具有良好的可解释性,通常适用于简单且便于理解的模型,如传统机器学习算法。后者则是先训练一个神经网络,再尝试解释其行为,包括学习特征分析、特征重要性评估、特征交互建模[79-81]以及基于显著性图的可视化解释方法[82-84]。例如,Zhou等[85]提出了类别激活映射(Class activation mapping, CAM),在最后卷积层特征图上应用全局平均池化,取代CNN末端的全连接层,再将其输出与分类结果之间的线性权重关系反向投影至特征图上,从而生成热力图以显示各区域对当前分类结果的贡献。在此基础上,Selvaraju等[86]提出了梯度加权类激活映射(Grad-CAM),该方法可用于任意CNN架构,通过梯度反向传播计算任意层特征图对目标类别的导数,进而生成与类别相关的激活图。Chattopadhay等[87]提出了Grad-CAM++方法。该方法在Grad-CAM的基础上引入导数加权机制,将特征图每个像素梯度的二阶和三阶导数进行加权,更精确地建模类别特征对空间位置的响应强度,从而生成更加平滑且准确的可视化特征图。Windisch等[88]使用Grad-CAM显示了脑部MRI中用于肿瘤分类的关键区域。Böhle等[89]使用LRP(Layer-wise relevance propagation)从脑磁共振图像中识别出与阿尔茨海默病相关的区域。通过将LRP提供的显著图与引导反向传播提供的显著图进行了比较,发现LRP在识别已知阿尔茨海默病区域方面更加具体有效。

尽管已有众多可解释性方法被提出,但学术界尚未形成一个明确的概念来定义可解释性。不同可解释性方法关注不同的解释维度,这就导致了不同解释方法之间无法比较优劣[90]。此外,像CAM和Grad-CAM这些目前主流的可解释性方法,在实际应用中常常面临解释不稳定和易受到输入干扰的问题。而且,现有的可解释性方法大多针对单模态模型设计,对于像医学多模态大语言模型和图文联合诊断模型等,尚缺乏系统性可解释手段。例如,一个模型如何在“影像+电子病历+基因组数据”中做

出诊断决策,其跨模态信息融合与推理过程几乎无法溯源解释。

3.2　数据依赖性强

医疗大模型的发展极大地推动了医学人工智能的进步,但其训练过程中对高质量、大规模医疗数据的高度依赖也面临着一些挑战。医疗数据的获取与整合不仅关乎技术问题,更涉及法律、伦理和社会因素,已经成为制约医疗大模型进一步发展的关键瓶颈。

首先,医疗数据普遍存在隐私敏感问题。例如,电子健康记录不仅包含患者的疾病诊断记录、用药记录、检查检验数据和手术记录,还包含患者的年龄、性别和民族背景等数据,一旦泄露,可能对患者隐私造成严重侵害,甚至引发重大的社会和经济后果。此外,全球各地都对医疗数据的获取和使用设立了严格的法律法规,如欧盟的《通用数据保护条例》[91]等。这些法规在一定程度上限制了大规模开放医疗数据集的构建,使跨机构、跨国数据共享与整合变得较为困难。

其次,医疗数据结构复杂且异构性强,进一步增加了数据处理的难度。与自然图像和通用文本数据不同,医疗数据的数据类型极为丰富,不仅包含年龄、血压和各项检查数值等结构化数据,还包括医学检验报告这样的半结构化数据,以及医生的自由文本病程记录、病理报告和影像数据等非结构化数据[92-93]。而且,在实际临床记录中,检查项目的不同、患者依从性差异或医生记录习惯的不同都可能导致数据某些部分存在缺失值。这不仅提高了数据清洗和标准化的成本,还可能引入系统性偏差,进而影响模型的鲁棒性和可信度[94-95]。

最后,医疗数据存在明显的区域性、机构性和人群差异性。由于设备水平、诊疗流程、记录习惯以及人群基因背景和疾病谱的差异,不同地区、不同等级医疗机构之间采集到的数据在分布上往往存在显著差异。例如,大型医院的数据质量通常较高且记录规范详细,而资源有限地区的小型医疗机构数据可能存在缺失多、记录不规范等问题。这种由数据来源差异导致的分布偏移直接影响了医疗大模型的泛化能力,导致其可能仅在与训练数据分布相似的场景下表现良好,而在其他人群或机构中性能大幅下降[96-97]。这不仅在实际临床应用中带来性能风险,也可能放大已有的健康不平等,引发“人工智能公平性”[98]问题、伦理争议和社会问题。例如,有研究指出[99-100],部分基于历史医疗数据训练的AI系统在有色人种患者中预测准确率显著低于白人患者。

针对上述挑战,近年来学术界和产业界提出了多种应对策略,例如通过数据脱敏与匿名化处理降低隐私风险,采用联邦学习等隐私保护技术实现跨机构模型训练。与此同时,标准化医疗数据格式的制定与应用也在积极推动,以增加不同系统间的数据互操作性。然而,这些技术手段仍存在一定局限,例如脱敏过程可能损失有用信息,联邦学习在实际部署中面临通讯成本高、系统异构性处理难度大等问题。因此,如何在保障数据隐私和公平性的前提下,实现高效、可信和可持续的医疗大数据利用,仍然是医疗大模型领域亟需持续探索的重要方向。

3.3　算力需求大

近年来,LPM在自然语言处理和计算机视觉等领域都取得了革命性突破。其核心特征之一就是拥有巨大的参数规模,通常能达到百亿甚至万亿级别。这种超大规模模型在训练甚至微调过程中对计算资源的消耗极为庞大。例如,使用2 048块配备80 GB显存的A100 GPU对参数量为65亿的LLaMA模型进行训练,在处理1.4万亿个Token时耗时约为21 d[101]。即使直接微调GPT-3模型也需要处理约1 752亿个参数[102]。由于模型参数规模极为庞大,在大规模数据上进行预训练将消耗大量的电力,从而带来显著的环境成本。有研究估算,训练一个BERT模型的碳排放量约为630 kg[103]。而且,由于大模型的推理阶段同样需要大量计算资源,在大规模部署与持续运行过程中还将带来长期的环境负担[104]。

随着医疗大模型的兴起,类似的计算挑战同样日益突出。医疗大模型往往使用大量多模态的医疗

数据进行预训练,如EHR、医学影像和基因组信息等,并希望通过更大规模的模型参数来捕捉医学知识图谱中的复杂关系和潜在规律[105]。因此,这些模型通常也具有数十亿到千亿以上的可训练参数,使得其训练和推理过程中对算力、存储和能耗的需求急剧上升。更重要的是,超高的算力需求显著提升了医疗大模型开发的门槛,使得相关研究逐渐集中于拥有强大基础设施的科技企业和顶级科研机构。这不仅加剧了医疗人工智能领域的资源集中趋势,也直接限制了中小型医疗机构在真实场景中部署和应用大模型的能力。

4 医疗AI的未来展望

近年来,医疗大模型依托于Transformer架构和大规模预训练技术,在多种医学任务中展现了强大的感知、理解和生成能力。然而,面对现实世界中复杂、多变且高要求的医疗场景,现有的模型在推理能力和部署效率等方面仍然存在显著不足。为了推动医疗大模型迈向更高阶的智能化水平,不仅需要在性能上持续优化,更应从算法底层架构、学习范式到推理机制进行系统性革新。基于目前医疗大模型的发展现状与未来应用需求,本文提出以下3个关键发展方向:(1)实现模型的轻量化,以支持资源受限环境下的高效部署;(2)赋予模型真正的推理能力,使得模型能够突破传统关联学习的局限,增强其在因果推理、科学判断与知识问答等任务中的适应性与可信度;(3)从认知科学和脑启发的角度出发,探索新一代医疗大模型架构,突破现有的Transformer范式,推动医疗大模型向更具解释性、推理性和通用智能的方向演进。

4.1 轻量化

当前,医疗大模型的发展在性能指标上取得了显著突破,但随之而来的模型体积膨胀和推理开销加重等问题也成为阻碍其在真实医疗环境中广泛应用的重要瓶颈。特别是在医院终端、便携式医疗设备以及在资源受限地区,这些环境下的计算资源和能耗条件远无法支撑拥有数十亿参数规模的大模型高效运行。因此,面向医疗场景的大模型轻量化已成为亟需解决的关键问题。

现有医疗大模型普遍继承了通用大模型的架构,如Med-PaLM系列模型采用的架构拥有数百亿参数。尽管在医学问答、推理和文献理解等任务上表现出色,但其部署和推理成本远高于传统医疗AI系统。相比之下,临床应用则更强调推理速度、稳定性和资源适配性,要求模型能够在低功耗设备上实现实时推理,而不是依赖大型GPU集群。因此,医疗大模型的轻量化设计不仅是技术挑战,更关乎其实际落地的可行性与普及性。

在模型轻量化方向上,已有部分研究工作初步探索了参数压缩、剪枝、量化和知识蒸馏等技术路径。模型剪枝[106]则通过移除冗余的权重或注意力连接来降低计算复杂度。而且,部分稀疏剪枝方法能够在保持或接近模型原始性能的前提下,减少30%~70%的参数量。知识蒸馏[107]通过将大模型的知识迁移到小模型上,使得小型模型能够在极小性能损失的情况下将大模型的推理速度提升2~5倍。为进一步提升知识迁移的效果,一些研究在蒸馏过程中引入思维链技术,以构建更为高效的知识传递机制[108],从而更充分地挖掘和保留大模型在推理过程中的认知路径。量化技术[109]通过将模型权重从32位浮点数压缩为更低的比特宽度,以此显著减少了显存的占用和推理延迟,同时对医疗任务性能的影响可通过量化感知训练进行补偿。

未来,医疗大模型轻量化有望在以下方向取得突破。一方面,可探索基于结构感知的剪枝与压缩策略,在保证诊断推理路径关键节点保留的基础上进行高效稀疏化,而非简单地全局稀疏化,以提升压缩的医学合理性。另一方面,探索动态模型技术,实现推理过程根据输入数据特征自适应地调整网络结构、计算路径或计算量,在保持性能的同时减少冗余计算量。

4.2　可推理

虽然当前的医疗大模型在医学诊断、医学问答和医学知识检索等任务中取得了优异成绩,但本质上,这些模型还是依赖于大规模的数据记忆和统计关联能力,而非拥有真正意义上的逻辑推理和科学推断能力。在面对复杂病理推演、跨模态因果推理和罕见病症分析等高阶医疗任务时,这种依赖表层模式匹配的模型常常表现出推理链断裂、事实幻觉或推理错误的问题[110]。因此,赋予医疗大模型真正可验证、可解释的推理能力,成为迈向更高水平智能医疗系统的关键突破口。

当前,医疗大模型推理能力的局限,主要源自其底层学习范式。尽管这些大模型在捕获长距离依赖和建模数据分布模式上取得了一定成功,但本质上仍是基于自回归或掩码预测的关联学习机制,缺乏显式的推理过程建模[111]。尤其在涉及疾病因果链推理或临床路径规划等任务时,现有大模型往往无法展现出连贯、透明的多步推理过程。而医学任务对推理可追溯性和可解释性有着更高要求,缺乏中间可视化推理步骤的模型难以满足临床安全标准[112]。

为了弥补大模型在推理能力上的不足,不少研究者通过引入链式思维(Chain-of-thought, CoT)等显式推理机制,显著提高了医学问答任务的准确率和中间推理流程的可解释性[113]。同时,结构化推理框架和神经符号推理方法也被应用于医学知识库推断和临床问答系统[114],展现出提升推理可验证性和因果理解的潜力。此外,MedAgents框架通过多代理协作,进一步增强了模型在零样本医疗推理任务中的表现[115]。

然而,与真实医生在临床实践中展现出的多源融合、动态调整的推理能力相比,现有方法仍存在推理链单一、缺乏动态适应的局限。医学推理不仅需要语言推理,还涉及视觉推理、时间推理与跨模态推理[116]。但是现有方法往往局限于单一模态或短链推理,难以覆盖完整的医学推理链。其次,医学推理高度依赖背景知识和因果关系理解,单纯依赖数据驱动的模型容易陷入伪关联误区,难以实现真正的因果推断[110]。

在未来的研究中,融合显式推理模块与隐式神经网络的混合架构有望成为主流发展方向。例如,可结合Transformer架构的上下文建模能力,引入专用推理引擎以实现因果链推导与多步逻辑验证。同时,构建医学领域的多模态推理数据集和基准评测体系,将有助于评估模型的推理能力并指导其持续优化。

4.3　启发式架构

自Transformer架构于2017年提出以来,凭借其强大的特征建模能力和良好的扩展性,Transformer及其变体成为当前医疗大模型的主流基础架构。无论是面向医学文本生成的BioGPT[117]、BioMedGPT[118],还是多模态融合的LLaVA-Med[119],这些模型无一例外都是以Transformer为核心骨架。然而,在医疗大模型不断向更复杂的推理、决策和科学发现等高阶任务拓展的背景下,基于自注意力的纯关联学习模式缺乏显式的层次推理和对抽象概念的学习能力[120],导致以Transformer为架构的大模型难以支持深度医学推理和临床科学推断。因此,探索超越单一Transformer范式的,更接近人类认知方式的新型医疗大模型架构成为未来发展的重要方向。

从认知科学和脑启发的角度来看,人类的学习与推理过程并非简单的序列关联,而是一个高度结构化、动态调整且多策略并用的过程[121]。在面对复杂任务时,人类具备在不同抽象层次间灵活切换的能力,能够在感知、语言、记忆、注意力和推理等多个认知模块之间协调信息流动。例如,人类医师在进行医学推理时,能够在不同抽象层次上切换思考。从具体症状到潜在病理过程,能够基于已有知识生成假设,并在诊疗过程中不断验证与修正,还能灵活地在视觉、语言、记忆和推理等不同认知模块之间协调信息处理。而当前的Transformer虽然在表面上模拟了长距离依赖建模,但本质上仍然是一种静态

的、单尺度的单路径信息处理机制，缺乏人类认知系统中天然存在的多模态融合、抽象层次建模、因果推理和元认知调控等能力。

为突破上述局限，未来可借鉴人脑的多尺度处理能力，研究具备层次化结构的Transformer变体，通过引入局部与全局注意力机制，使模型能够在不同空间与语义尺度上整合信息，模拟人脑皮层中自下而上与自上而下交互的信息流动。同时，在推理机制方面，引入神经符号融合架构，将医疗大模型的感知表示能力和符号系统的逻辑因果推理能力相结合，使模型不仅能理解语言与识别图像，还能执行结构化思维与显式知识操作，从而实现更符合人类专家认知过程的诊疗决策推理。通过融合认知科学原理、神经启发机制和新型计算范式，走向多模块协作、动态推理与自适应学习的新一代智能架构。这不仅是技术发展的自然趋势，也是医疗AI真正实现从“辅助工具”到“认知伙伴”转变的关键跃迁。

5 结束语

随着人工智能技术的飞速发展，医学影像分析与数字辅助诊断正逐步从传统的任务定制型方法，迈向以大规模预训练模型为基础的“基础模型”时代。这一转变不仅显著提升了医疗AI在多任务、多模态、多场景下的泛化能力，也加速了医疗人工智能向具备推理能力与上下文理解能力的“通用智能”演进。医疗基础模型的出现，有力地推动了AI技术在临床辅助诊断、疾病早期筛查及个性化治疗等关键医疗场景中的实际应用，为智慧医疗的落地与普及提供了坚实支撑。

本文系统梳理了医疗人工智能的发展脉络，从初期的萌芽阶段，到统计学习与深度学习的快速演进，再到当前引领新一轮技术浪潮的医疗大模型阶段，全面呈现了该领域的演进轨迹与关键转折点。在此基础上，综述了大模型在多个典型医学子领域的研究进展，包括病理学、眼科学和神经系统疾病等，展现了其在复杂任务处理、跨模态理解和智能辅助决策中的强大潜力。

尽管医疗大模型展现出广阔前景，但其在大规模部署与临床落地过程中仍面临诸多挑战，例如模型可解释性不足、对高质量标注数据的高度依赖、以及对计算资源的巨大需求等。解决这些问题需要技术创新与实践探索的协同推进。展望未来，医疗AI的发展应更加聚焦于模型的轻量化设计、推理能力的增强，以及融合人类认知机制的启发式架构构建。通过持续优化模型效率与可靠性，医疗大模型有望实现从“辅助工具”向“智能伙伴”的跃迁，进一步推动智能医疗系统走向普适、精准、以患者为中心的新时代。

参考文献：

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017. DOI: 10.48550/arXiv.1706.03762.

[2] HAN X, ZHANG Z, DING N, et al. Pre-trained models: Past, present and future[J]. AI Open, 2021, 2: 225-250.

[3] RYDNING D R J G J, REINSEL J, GANTZ J. The digitization of the world from edge to core[J]. Framingham: International Data Corporation, 2018, 16: 1-28.

[4] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.]: ACL, 2019: 4171-4186.

[5] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[EB/OL]. (2018-01-01). https://api.semanticscholar.org/CorpusID:49313245.

[6] DeepSeek-AI, GUO D, YANG D, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning [EB/OL]. (2025-01-12). https://arxiv.org/abs/2501.12948v1.

[7] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[8] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the

IEEE, 1998, 86(11): 2278-2324.

[9] SHORTLIFFE E, SCOTT A, BISCHOFF M B, et al. ONCOCIN: An expert system for oncology protocol management [C]//Proceedings of the 7th International Joint Conference on Artificial Intelligence. [S.l.]: ACM, 1997: 876-881.

[10] MELLE W, SHORTLIFFE E, BUCHANAN B. EMYCIN: A knowledge engineer's tool for constructing rule-based expert systems[EB/OL].(1999-01-01). https://people.dbmi.columbia.edu/~ehs7001/Buchanan-Shortliffe-1984/Chapter-15.pdf.

[11] YU J, GOODMAN S R. Syndeins: The spectrin-binding protein(s) of the human erythrocyte membrane[J]. Proceedings of the National Academy of Sciences of the United States of America, 1979, 76(5): 2340-2344.

[12] MUSEN M A. Dimensions of knowledge sharing and reuse[J]. Computers and Biomedical Research, 1992, 25(5): 435-467.

[13] COIERA E. Guide to medical informatics, the internet and telemedicine[M]. London: Chapman & Hall, Ltd., 1997.

[14] KONONENKO I. Machine learning for medical diagnosis: History, state of the art and perspective[J]. Artificial Intelligence in Medicine, 2001, 23(1): 89-109.

[15] ALPAYDIN E, KAYNAK C, ALIMOGLU F. Cascading multiple classifiers and representations for optical and pen-based handwritten digit recognition[C]//Proceedings of International Workshop on Frontiers in Handwriting Recognition. [S.l.]: [s.n.], 2000.

[16] HASTIE T, TIBSHIRANI R, FRIEDMAN J H, et al. The elements of statistical learning: Data mining, inference, and prediction[M]. New York: Springer, 2009.

[17] KOUROU K, EXARCHOS T P, EXARCHOS K P, et al. Machine learning applications in cancer prognosis and prediction [J]. Computational and Structural Biotechnology Journal, 2015, 13: 8-17.

[18] OBERMEYER Z, EMANUEL E J. Predicting the future—Big data, machine learning, and clinical medicine[J]. The New England Journal of Medicine, 2016, 375(13): 1216-1219.

[19] DELEN D, WALKER G, KADAM A. Predicting breast cancer survivability: A comparison of three data mining methods[J]. Artificial Intelligence in Medicine, 2005, 34(2): 113-127.

[20] WILLIAMS K, ADEBAYO IDOWU P, ADEMOLA BALOGUN J, et al. Breast cancer risk prediction using data mining classification techniques[J]. Transactions on Networks and Communications, 2015. DOI: 10.14738/tnc.32.662.

[21] RAJKOMAR A, DEAN J, KOHANE I. Machine learning in medicine[J]. New England Journal of Medicine, 2019, 380(14): 1347-1358.

[22] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[EB/OL]. (2018-01-01). https://web.cs.ucdavis.edu/~yjlee/teaching/ecs289g-winter2018/alexnet.pdf.

[23] SETIO A A A, CIOMPI F, LITJENS G, et al. Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks[J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1160-1169.

[24] ESTEVA A, KUPREL B, NOVOA R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115-118.

[25] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[M]// Medical Image Computing and Computer-Assisted Intervention. Cham: Springer International Publishing, 2015: 234-241.

[26] 郝小可, 何子龙, 卢欣楚, 等. 基于多尺度残差融合图卷积网络的脑疾病诊断研究[J]. 数据采集与处理, 2024, 39(4): 827-842.
HAO Xiaoke, HE Zilong, LU Xinchu, et al. Research on brain disease diagnosis based on multi-scale residual fusion graph convolutional network[J]. Journal of Data Acquisition and Processing, 2024, 39(4): 827-842.

[27] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[EB/OL]. (2014-09-23). https://arxiv.org/abs/1409.2329v5.

[28] HANIN B. Which neural net architectures give rise to exploding and vanishing gradients? [EB/OL]. (2018-01-03). https://arxiv.org/abs/1801.03744v3.

[29] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[30] HENIGHAN T, KAPLAN J, KATZ M, et al. Scaling laws for autoregressive generative modeling[EB/OL]. (2020-10-14).

https://arxiv.org/abs/2010.14701v2.

[31] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620(7972): 172-180.

[32] LIU J, ZHOU P, HUA Y, et al. Benchmarking large language models on CMExam: A comprehensive Chinese medical exam dataset[EB/OL]. (2023-06-03). https://arxiv.org/abs/2306.03030v3.

[33] HUANG Y, BAI Y, ZHU Z, et al. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models[EB/OL]. (2023-05-08). https://arxiv.org/abs/2305.08322.

[34] JOHNSON A E W, POLLARD T J, GREENBAUM N R, et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs[EB/OL]. (2019-01-07). https://arxiv.org/abs/1901.07042v5.

[35] IRVIN J, RAJPURKAR P, KO M, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2019: 590-597.

[36] NETWORK C G A. Comprehensive molecular portraits of human breast tumours[J]. Nature, 2012, 490(7418): 61-70.

[37] LIU M, HU W, DING J, et al. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating Chinese medical large language models[J]. Big Data Mining and Analytics, 2024, 7(4): 1116-1128.

[38] KIM Y J, JANG H, LEE K, et al. PAIP 2019: Liver cancer segmentation challenge[J]. Medical Image Analysis, 2021, 67: 101854.

[39] GULSHAN V, PENG L, CORAM M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs[J]. JAMA, 2016, 316(22): 2402-2410.

[40] ORLANDO J I, FU H, BARBOSA BREDA J, et al. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs[J]. Medical Image Analysis, 2020, 59: 101570.

[41] MENZE B H, JAKAB A, BAUER S, et al. The multimodal brain tumor image segmentation benchmark (BRATS)[J]. IEEE Transactions on Medical Imaging, 2015, 34(10): 1993-2024.

[42] JACK JR C R, BERNSTEIN M A, FOX N C, et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods [J]. Journal of Magnetic Resonance Imaging, 2008, 27(4): 685-691.

[43] LAMONTAGNE P, BENZINGER T, MORRIS J, et al. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease[J]. MedRxiv, 2022. DOI: https://doi.org/10.1101/2019.12.13.19014902.

[44] TSENG L J, MATSUYAMA A, MACDONALD-DICKINSON V. Histology: The gold standard for diagnosis?[J]. The Canadian Veterinary Journal, 2023, 64(4): 389.

[45] CHEN R J, CHEN C, LI Y, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 16123-16134.

[46] HE K, CHEN X, XIE S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE, 2022: 15979-15988.

[47] HE Y, HUANG F, JIANG X, et al. Foundation model for advancing healthcare: Challenges, opportunities and future directions[J]. IEEE Reviews in Biomedical Engineering, 2024. DOI: https://doi.org/10.48550/arXiv.2404.03264.

[48] WANG X, ZHAO J, MAROSTICA E, et al. A pathology foundation model for cancer diagnosis and prognosis prediction[J]. Nature, 2024, 634(8035): 970-978.

[49] VORONTSOV E, BOZKURT A, CASSON A, et al. A foundation model for clinical-grade computational pathology and rare cancers detection[J]. Nature Medicine, 2024, 30(10): 2924-2935.

[50] OQUAB M, DARCET T, MOUTAKANNI T, et al. DINOv2: Learning robust visual features without supervision[EB/OL]. (2023-04-07). https://arxiv.org/abs/2304.07193v2.

[51] HUANG Z, BIANCHI F, YUKSEKGONUL M, et al. A visual-language foundation model for pathology image analysis using medical Twitter[J]. Nature Medicine, 2023, 29(9): 2307-2316.

[52] LU M Y, CHEN B, WILLIAMSON D F K, et al. A visual-language foundation model for computational pathology[J]. Nature Medicine, 2024, 30(3): 863-874.

[53] XIANG J, WANG X, ZHANG X, et al. A vision-language foundation model for precision oncology[J]. Nature, 2025, 638(8051): 769-778.

[54] ZHOU Y, CHIA M A, WAGNER S K, et al. A foundation model for generalizable disease detection from retinal images[J]. Nature, 2023, 622(7981): 156-163.

[55] SILVA-RODRÍGUEZ J, CHAKOR H, KOBBI R, et al. A foundation language-image model of the retina (FLAIR): Encoding expert knowledge in text supervision[J]. Medical Image Analysis, 2025, 99: 103357.

[56] SHI D, ZHANG W, YANG J, et al. EyeCLIP: A visual-language foundation model for multi-modal ophthalmic image analysis [EB/OL]. (2024-09-06). https://arxiv.org/abs/2409.06644v2.

[57] DU J, GUO J, ZHANG W, et al. RET-CLIP: A retinal image foundation model pre-trained with clinical diagnostic reports [M]//Medical Image Computing and Computer Assisted Intervention. Cham: Springer Nature Switzerland, 2024: 709-719.

[58] LUO Y, SHI M, KHAN M O, et al. FairCLIP: Harnessing fairness in vision-language learning[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 12289-12301.

[59] JIANG H, GAO M, LIU Z, et al. GlanceSeg: Real-time microaneurysm lesion segmentation with gaze-map-guided foundation model for early detection of diabetic retinopathy[J]. IEEE Journal of Biomedical and Health Informatics, 2024(99): 1-14.

[60] QIU J, LI L, SUN J, et al. Large AI models in health informatics: Applications, challenges, and the future[J]. IEEE Journal of Biomedical and Health Informatics, 2023, 27(12): 6074-6087.

[61] CHEN M, ZHANG M, YIN L, et al. Medical image foundation models in assisting diagnosis of brain tumors: A pilot study[J]. European Radiology, 2024, 34(10): 6667-6679.

[62] BARBANO C A, BRUNELLO M, DUFUMIER B, et al. Anatomical foundation models for brain MRIs[EB/OL]. (2024-08-07). https://arxiv.org/abs/2408.07079v3.

[63] SUN Y, WANG L, LI G, et al. A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks[J]. Nature Biomedical Engineering, 2025, 9: 521-538.

[64] COX J, LIU P, STOLTE S E, et al. BrainSegFounder: Towards 3D foundation models for neuroimage segmentation[J]. Medical Image Analysis, 2024, 97: 103301.

[65] CARO J O, FONSECA A H O, AVERILL C, et al. BrainLM: A foundation model for brain activity recordings[J]. bioRxiv, 2023. DOI: 2023.09.12.557460.

[66] JIANG W B, ZHAO L M, LU B L. Large brain model for learning generic representations with tremendous EEG data in BCI [EB/OL]. (2024-05-18). https://arxiv.org/abs/2405.18765v1.

[67] ZHANG D, YUAN Z, YANG Y, et al. Brant: Foundation model for intracranial neural signal[J]. Advances in Neural Information Processing Systems, 2023, 36: 26304-26321.

[68] LAI H, YAO Q, JIANG Z, et al. CARZero: Cross-attention alignment for radiology zero-shot classification[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 11137-11146.

[69] PHAN V M H, XIE Y, QI Y, et al. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework[C]//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2024: 11492-11501.

[70] ZHANG X, WU C, ZHANG Y, et al. Knowledge-enhanced visual-language pre-training on chest radiology images[J]. Nature Communications, 2023, 14(1): 4542.

[71] WU C, ZHANG X, ZHANG Y, et al. MedKLIP: Medical knowledge enhanced language-image pre-training for X-ray diagnosis[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 21315-21326.

[72] BANNUR S, HYLAND S, LIU Q, et al. Learning to exploit temporal structure for biomedical vision-language processing [C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 15016-15027.

[73] LIU K, MA Z, KANG X, et al. Enhanced contrastive learning with multi-view longitudinal data for chest X-ray report generation[EB/OL]. (2025-02-20). https://arxiv.org/abs/2502.20056v1.

[74] ENNAB M, MCHEICK H. Enhancing interpretability and accuracy of AI models in healthcare: A comprehensive review on challenges and future directions[J]. Frontiers in Robotics and AI, 2024, 11: 1444763.

[75] TENG Q, LIU Z, SONG Y, et al. A survey on the interpretability of deep learning in medical diagnosis[J]. Multimedia Systems, 2022, 28(6): 2335-2355.

[76] XU B, YANG G. Interpretability research of deep learning: A literature survey[J]. Information Fusion, 2025, 115: 102721.

[77] DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning[EB/OL]. (2017-02-08). https://arxiv.org/abs/1702.08608v2.

[78] LOH H W, OOI C P, SEONI S, et al. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011—2022)[J]. Computer Methods and Programs in Biomedicine, 2022, 226: 107161.

[79] ABBASI-ASL R, YU B. Structural compression of convolutional neural networks[EB/OL]. (2017-05-07). https://arxiv.org/abs/1705.07356v4.

[80] OLDEN J D, JOY M K, DEATH R G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data[J]. Ecological Modelling, 2004, 178(3/4): 389-397.

[81] TSANG M, CHENG D, LIU Y. Detecting statistical interactions from neural network weights[EB/OL]. (2017-05-04). https://arxiv.org/abs/1705.04977v4.

[82] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: Visualising image classification models and saliency maps[EB/OL]. (2013-12-06). https://arxiv.org/abs/1312.6034v2.

[83] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, et al. Striving for simplicity: The all convolutional net[EB/OL]. (2014-12-06). https://arxiv.org/abs/1412.6806v3.

[84] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//Proceedings of Computer Vision — ECCV 2014. Cham: Springer International Publishing, 2014: 818-833.

[85] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 2921-2929.

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

[87] CHATTOPADHAY A, SARKAR A, HOWLADER P, et al. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks[C]//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). [S.l.]: IEEE, 2018: 839-847.

[88] WINDISCH P, WEBER P, FÜRWEGER C, et al. Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices[J]. Neuroradiology, 2020, 62(11): 1515-1518.

[89] BÖHLE M, EITEL F, WEYGANDT M, et al. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification[J]. Frontiers in Aging Neuroscience, 2019, 11: 194.

[90] LIPTON Z C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery[J]. Queue, 2018, 16(3): 31-57.

[91] BYGRAVE L. Automated individual decision-making, including profiling[M]//EU GDPR. Cham: Springer, 2021.

[92] TAYEFI M, NGO P, CHOMUTARE T, et al. Challenges and opportunities beyond structured data in analysis of electronic health records[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2021, 13(6): e1549.

[93] WANG Y, KUNG L, BYRD T A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations[J]. Technological Forecasting and Social Change, 2018, 126: 3-13.

[94] BEAULIEU-JONES B K, LAVAGE D R, SNYDER J W, et al. Characterizing and managing missing structured data in electronic health records: Data analysis[J]. JMIR Medical Informatics, 2018, 6(1): e11.

[95] WELLS B J, CHAGIN K M, NOWACKI A S, et al. Strategies for handling missing data in electronic health record derived data[J]. EGEMS, 2013, 1(3): 1035.

[96] SU Z, GUO J, YANG X, et al. Navigating distribution shifts in medical image analysis: A survey[EB/OL]. (2024-11-05). https://arxiv.org/abs/2411.05824v1.

[97] ZECH J R, BADGELEY M A, LIU M, et al. Confounding variables can degrade generalization performance of radiological deep learning models[EB/OL]. (2018-07-04). https://arxiv.org/abs/1807.00431v2.

[98] MEHRABI N, MORSTATTER F, SAXENA N, et al. A survey on bias and fairness in machine learning[J]. ACM Computing Surveys, 2022, 54(6): 1-35.

[99] OBERMEYER Z, POWERS B, VOGELI C, et al. Dissecting racial bias in an algorithm used to manage the health of populations[J]. Science, 2019, 366(6464): 447-453.

[100] DANESHJOU R, VODRAHALLI K, LIANG W, et al. Disparities in dermatology AI: Assessments using diverse clinical images[EB/OL]. (2021-11-08). https://arxiv.org/abs/2111.08006v1.

[101] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and efficient foundation language models[EB/OL]. (2023-02-13). https://arxiv.org/abs/2302.13971v1.

[102] DING N, QIN Y, YANG G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models[J]. Nature Machine Intelligence, 2023, 5: 220-235.

[103] WANG X, NA C, STRUBELL E, et al. Energy and carbon considerations of fine-tuning BERT[EB/OL]. (2023-11-17). https://arxiv.org/abs/2311.10267.

[104] PETER H, HU J, JOSHUA R, et al. Towards the systematic reporting of the energy and carbon footprints of machine learning [J]. Journal of Machine Learning Research, 2020, 21: 10039-10081.

[105] MOOR M, BANERJEE O, ABAD Z S H, et al. Foundation models for generalist medical artificial intelligence[J]. Nature, 2023, 616(7956): 259-265.

[106] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural network[C]//Proceedings of the 29th International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2021.

[107] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-02). https://arxiv.org/abs/1503.02531v1.

[108] 李荣涵,浦荣成,沈佳楠,等.基于思维链的大语言模型知识蒸馏[J].数据采集与处理, 2024, 39(3): 547-558.
LI Ronghan, PU Rongcheng, SHEN Jianan, et al. Knowledge distillation of large language model based on thought chain[J]. Journal of Data Acquisition and Processing, 2024, 39(3): 547-558.

[109] JACOB B, KLIGYS S, CHEN B, et al. Quantization and training of neural networks for efficient integer-arithmetic-only inference[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 2704-2713.

[110] KIM J, PODLASEK A, SHIDARA K, et al. Limitations of large language models in clinical problem-solving arising from inflexible reasoning[EB/OL]. (2025-02-04). https://arxiv.org/abs/2502.04381v1.

[111] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[EB/OL]. (2022-01-11). https://arxiv.org/abs/2201.11903v6.

[112] WU C K, CHEN W L, CHEN H H. Large language models perform diagnostic reasoning[EB/OL]. (2023-07-08). https://arxiv.org/abs/2307.08922v1.

[113] LIÉVIN V, HOTHER C E, MOTZFELDT A G, et al. Can large language models reason about medical questions? [J]. Patterns, 2024, 5(3): 100943.

[114] SHI J, DING X, HUI S C, et al. Final: Combining first-order logic with natural logic for question answering[J]. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(6): 3103-3117.

[115] TANG X, ZOU A, ZHANG Z, et al. MedAgents: Large language models as collaborators for zero-shot medical reasoning [EB/OL]. (2023-11-10). https://arxiv.org/abs/2311.10537v4.

[116] ZHANG Z, ZHANG A, LI M, et al. Multimodal chain-of-thought reasoning in language models[EB/OL]. (2023-02-09).

https://arxiv.org/abs/2302.00923v5.

[117] LUO R, SUN L, XIA Y, et al. BioGPT: Generative pre-trained Transformer for biomedical text generation and mining[J]. Briefings in Bioinformatics, 2022, 23(6): bbac409.

[118] LUO Y, ZHANG J, FAN S, et al. BioMedGPT: Open multimodal generative pre-trained Transformer for BioMedicine[EB/OL]. (2023-08-09). https://arxiv.org/abs/2308.09442v2.

[119] LI C, WONG C, ZHANG S, et al. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day [EB/OL]. (2023-06-08). https://arxiv.org/abs/2306.00890v1.

[120] NERELLA S, BANDYOPADHYAY S, ZHANG J, et al. Transformers and large language models in healthcare: A review [J]. Artificial Intelligence in Medicine, 2024, 154: 102900.

[121] PREZENSKI S, BRECHMANN A, WOLFF S, et al. A cognitive modeling approach to strategy formation in dynamic decision making[J]. Frontiers in Psychology, 2017, 8: 1335.

作者简介:

钱波(1991-),男,助理教授,研究方向:医学影像分析、计算机视觉、人工智能等,E-mail: qiannbo@nuaa.edu.cn。

李富江(2000-),男,硕士研究生,研究方向:医学影像分析、人工智能等。

郑常乐(2001-),男,硕士研究生,研究方向:医学影像分析、人工智能等。

张道强(1978-),**通信作者**,男,教授,研究方向:医学影像分析、计算机视觉、人工智能等,E-mail:dqzhang@nuaa.edu.cn。

(编辑:张黄群)