·智慧医疗· doi:10.3969/j.issn.1671-8348.2023.22.026
网络首发 https://link.cnki.net/urlid/50.1097.R.20231114.1008.004(2023-11-14)

基于医学大数据的临床科研模型智能化构建研究*

王 飞,汪 鹏,黄艺璠,李 颖,胡 川△
(陆军军医大学第一附属医院医学大数据与人工智能中心,重庆 400038)

[摘要] 随着大数据与人工智能技术在医学领域的深入应用,越来越多的医学研究关注于利用大规模医学数据进行模型构建,对模型构建过程提出了智能化、个性化要求。该研究结合医学模型构建流程,汇聚、处理主要医疗业务数据,将数据搜集、模型建立、统计分析等研究方法进行工程化封装,以期实现临床科研模型构建流程化、模块化。

[关键词] 医学大数据;临床研究;模型;智能化构建
[中图法分类号] R-05 [文献标识码] A [文章编号] 1671-8348(2023)22-3509-04

Research on intelligent construction of clinical scientific research model based on medical big data*

WANG Fei, WANG Peng, HUANG Yifan, LI Ying, HU Chuan△
(*Medical Big Data and Artificial Intelligence Center, First Affiliated Hospital, Army Military Medical University, Chongqing* 400038, *China*)

[**Abstract**] With the extensive use of big data and artificial intelligence technology in the medical area, more and more medical researches focus on the use of large-scale medical data for conducting the model construction, and put forward intelligent and personalized requirements for the model construction process. This study combines the medical model building process, gathers and processes the main medical business data, and engineering packages the research methods such as data collection, model building and statistical analysis to realize the process automation and modularization of medical model building.

[**Key words**] medical big data; clinical research; model; intelligent construction

近年来,随着医学多模态数据的不断汇聚和人工智能技术的迅猛发展,基于大数据的临床研究已成为医学与计算机科学交叉融合的热点方向,越来越多的临床研究趋向于使用人工智能技术对医学数据进行建模、训练、分析,形成回顾性结论或前瞻性预判[1]。在医疗领域,将人工智能算法模型的构建方法与应用工具对临床科研人员开放,对建模方法进行流程化设计与实现,使其能够通过信息化工具进行算法模型的智能化构建,将有助于提高自主化、个性化临床研究能力,提升临床科研水平[2]。

1 需求分析

基于医学大数据的临床研究过程(包括数据采集与处理、数据标注与训练、模型迭代与应用等几个典型阶段),其核心在于采集相关原始数据进行清洗、去重、标注等处理后做算法训练支撑以验证研究思路与结果[3]。然而,对医疗行业而言,存在大量文本、影像等多模态数据[4],临床研究的主要瓶颈问题聚焦在以下两个方面。

1.1 如何对汇聚的文本与影像数据进行信息关联

临床诊疗数据类型丰富、结构复杂、模态多样,将影像数据与病历、检查等文本数据打通和融合是医学大数据应用和人工智能探索的重要方向,不同模态数据之间的有效融合与内部关联展示仍未得到充分体现[5]。基于本体构建方法,在文本结构化提取模型融合术语库,实现表述归一化的基础上,如何对影像数据进行结构化处理,将病灶类型、病灶大小、解剖学位置、影像学征象等核心内容与病历、检查报告等信息进行关联,实现多维综合分析处理与跨模态检索匹配仍是临床研究在数据搜索层面须解决的首要问题[6]。

1.2 如何建立符合医学数据特征的模型构建方法

在医学领域,人工智能为临床研究与诊疗提供了算法和模型支撑,包括 Faster R-CNN、Logistic Regression、SVM、Random Forest 等[7]。但由于医学本身的复杂性,现有框架体系下的算法模型还不能完全满足个性化研究需求,需要相关技术人员配合完成标注、调参、训练等过程[8]。如何结合医学研究流程,对

* 基金项目:国家重点研发计划项目(2018YFB2101204);重庆市卫健委医学科研项目(2023WSJK062)。 作者简介:王飞(1982-),高级工程师,硕士,主要从事医学大数据与人工智能的临床应用研究。 △ 通信作者,E-mail:huchuan023@vip.qq.com。

模型自主化构建与自适应调参等进行工程化封装，实现模型自主构建、智能调参、迭代优化等功能，也是临床研究在工程化开发与应用层面提出的重要需求。

因此，对临床研究模型的构建流程进行信息化处理，将数据搜集、模型建立、统计分析等方法工具化，有助于进一步辅助科研人员灵活、高效地运用大数据与人工智能相关方法进行科学研究。

2 流程设计

基于医学大数据的临床科研模型构建的流程需经过临床问题定义、数据收集、数据处理、模型构建等几个主要步骤。临床问题定义：把建模解决的问题进行细化，让一个无法进行建模或分析的任务变成可执行的内容。数据收集：根据定义的临床目标，对数据进行采集与汇聚，主要涉及电子病历、检查、检验、移动医护、病理等主要业务信息系统数据[9-10]。数据处理：包括数据标签、数据清洗、缺失值填充、变量处理、归一化等。对重复值、异常值进行处理，以完善数据质量；根据模型对数据特征的要求，对空值、缺失值进行填充，按比例缩放使之落入特定的区间。模型构建：通过特征之间的关系，补充特征、过滤特征及选择特征，对模型进行超参数优化与迭代，对准确率、灵敏度、特异度和马太相关系数等指标进行验证与评价。模型构建流程见图 1。

3 功能构建

3.1 图文结合的多模态数据库与搜索引擎

基于大数据技术框架，采集整合患者诊疗数据，包括病历文书、医学影像、实验室检验、病理检测等，应用大数据存储搜索、大数据治理、多模态特征结构化提取等技术，构建多模态数据库与搜索引擎[11]，见图 2。

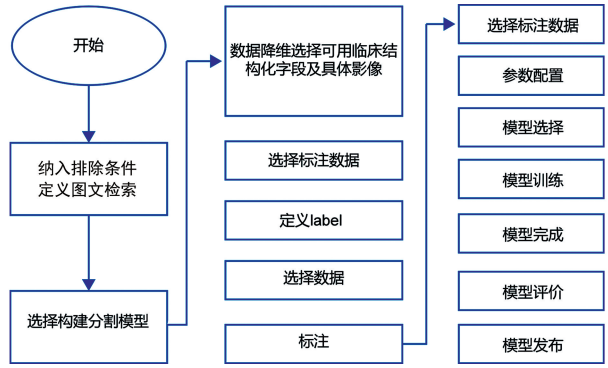

图 1 模型构建流程图

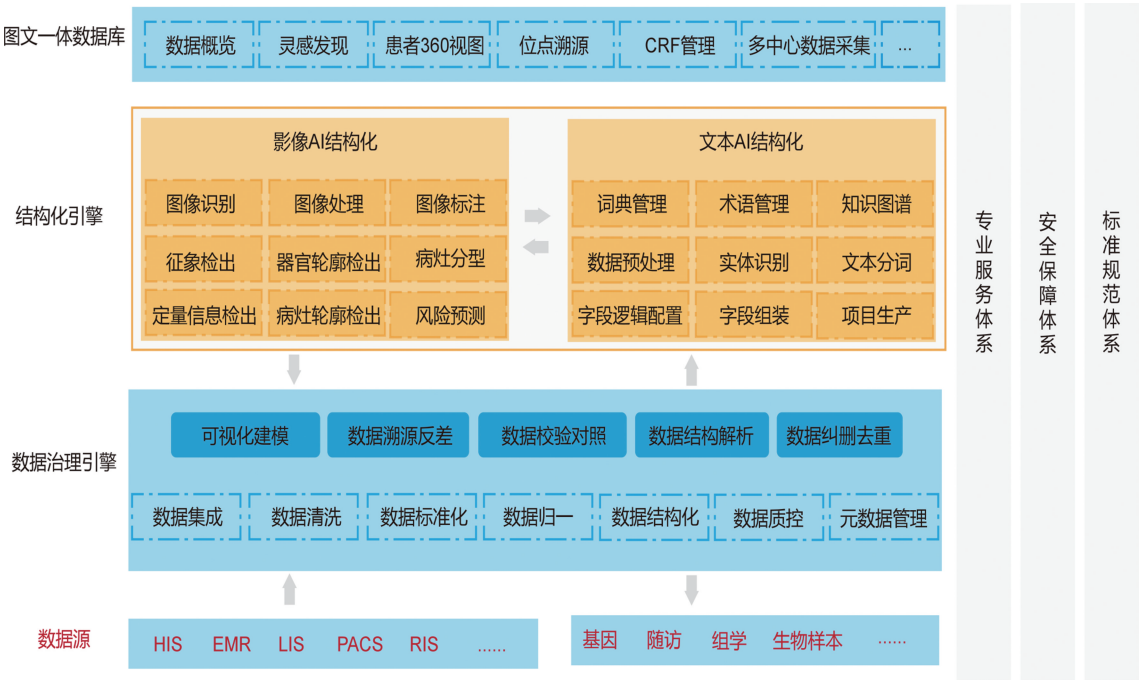

图 2 数据采集与治理框架图

文本数据标准化与结构化：基于已建立的医学大数据治理平台，结合患者真实入院、出院记录中主诉、手术、影像报告、病理等医学文本数据，使用多个信息抽取模型提取症状、病灶信息、诊断结果等信息。同时，提取疾病原词、手术原词，将其归一化处理[12]。

影像数据集建立：以影像特征建立数据集，提取影像信息 tag 值进行结构化处理，影像信息提取结果可与检查报告文本信息进行关联。影像病灶识别支持对病灶自动进行检测，圈定 ROI 区域，并能够对检测的 ROI 进行分类，甄别病灶 ROI 及非病灶 ROI，以及对病灶 ROI 进行检测结果及定位的输出，支持 ROI 区域的快速访问[13]。病灶特征描述包括大小、密度、位置、征象、历史配准、倍增时间等。

多模态数据搜索引擎构建：将影像高维数据集与临床信息数据集进行整合，进而实现临床文本数据与影像数据的高精融合，构建图文结合的多模态数据库。基于深度学习的多模态影像数据和文本资料表示方法，实现基于图像深度特征与局部语义感知的语

义层次模型，通过注意力机制等方式获取细粒度语义区分能力，有效解决多语义、图描述和细粒度交互等问题，进而提高图文匹配的准确度[14]。基于混合迁移学习策略和深度重构哈希机制的高层语义特性图像标注方法，引入标签偏置正则约束机制，建立基于双向特征学习的“图像-文本”跨模态检索方法和支持系统。

3.2　基于机器学习的流程化建模方法

根据影像数据特点，提供合适的医学影像模型框架，建立机器学习模型流程化建立方法[15]。影像研究的每一个任务都可以看成一个机器学习，其目的就是建立输入的影像数据与临床结论之间的关系，并将这种关系以机器学习模型的形式存储起来，当有新影像数据的时候，可以使用机器学习模型进行结论预测，见图 3。

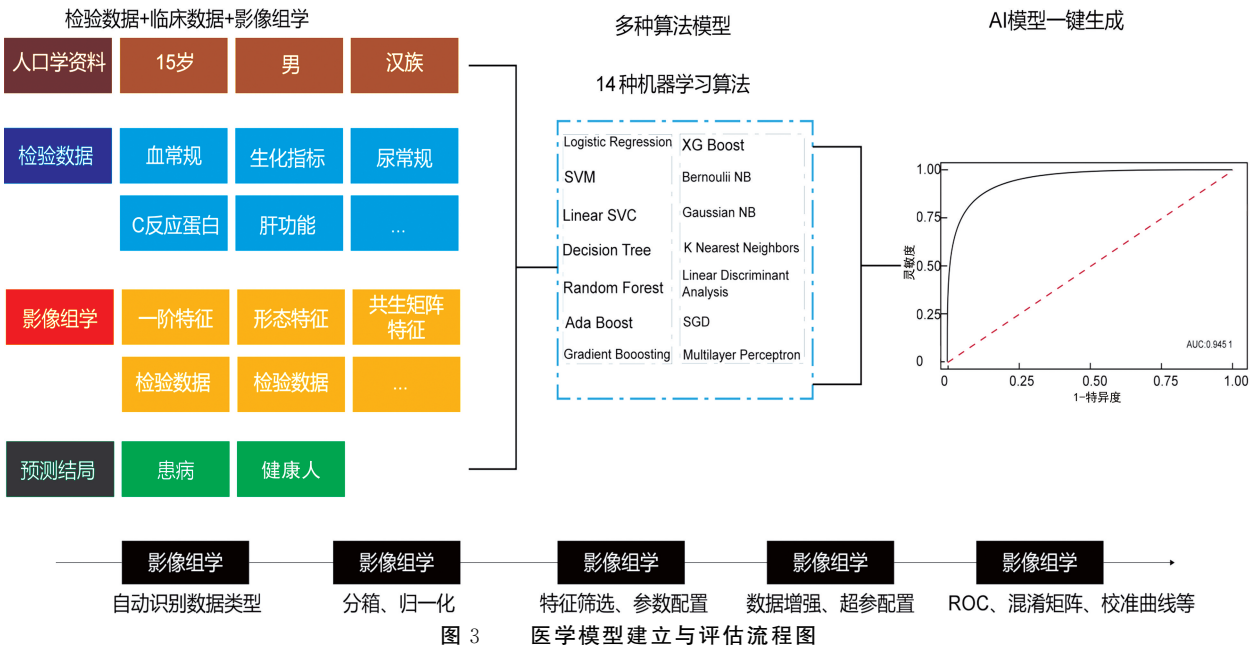

图 3　医学模型建立与评估流程图

影像多维度特征提取：针对多模态影像数据，采用基于 multi-phase FCN 的深度学习算法，提取并比较分析多种模态下肿瘤的高维影像学特征，实现病变区域的精准分割；针对分割出的病灶，实现基于 multimodal and multiscale CNN 的深度学习算法充分提取不同维度的影像特征，并通过神经网络算法分析多模态影像学特征和肿瘤的关联[16]。

特征选择：主要包括两个部分，第一部分是从影像数据中高通量地提取影像特征，第二部分是应用这些特征建立有价值的预测模型。特征选择需要挑选与临床结论相关的、不冗余的特征信息，可提供多种单因素、多因素和联合方法，包括 F-test、皮尔逊、互信息、L1 正则、树模型、循环特征删除等 6 大类特征选择方法，并支持数值特征自动选择。

机器学习模型算法：在获得了高通量影像特征并经过选择后，提供常用的多种机器学习模型算法，包括 Logistic Regression、Random Forest、SVM 等，以支持完成机器学习建模流程。为了更好地掌控模型训练过程，将开放机器学习算法在建模过程中的参数调节接口，研究人员可通过选择合适的算法模型和调整模型的参数来改变算法的效能，从而获得更好的模型[17]。

3.3　临床研究流程化实现

文本与影像智能搜索：应用自然语义处理技术，结合医疗专业术语的语义结构，实现医疗语义信息从原始的自然语言表达扩展分析为结构化的 Key-Value 模式[18]。对原文使用关键词进行搜索，满足多个搜索条件间的逻辑关系，支持多层逻辑关系或与非的嵌套组合，快速完成科研纳入排除过程。对影像数据进行清洗筛选和重采样，支持影像高维度量化变量提取，将影像特征提取的数据纳入数据库，形成影像 tag、影像病灶轮廓信息检索。

医学模型流程化构建：集成影像处理算法包，建立影像数据标注、多维度特征提取、特征选择、模型建立、研究结果分析全流程的医学模型构建功能。其中数据标注包括手动分割、半自动分割、扩展标注、插值标注等功能，可自动检测出病灶并分割，实现自动勾勒标注；同时，临床研究人员可完成像素级图像标注，支持淋巴结的自动识别、自动分区，自动完成影像标注。多维度特征提取可从影像数据中提取出＞1 000 维度的影像特征，可对数据的缺失值和异常值进行补全或删除处理。特征选择提供多种单因素、多因素和联合方法，支持与临床结论关联程度较高的特征参数。模型建立提供包括 Logistic Regression、Random Forest、SVM 等模型算法，实现个性化选择。研究结果分析可以输出一般研究所需要的 ROC 曲线下面积（AUC）、灵敏度、特异度等值，并展示模型训练和校验结果[19]。

4　小　　结

通过对临床科研模型智能化构建的研究，建立流

程化、模块化、智能化的数据采集、处理、标注、建模、分析等功能，满足研究人员对数据检索、数据标注、模型构建等方面的个性化需求，在一定程度上解决了人工智能算法应用在临床研究上的技术和工程阻碍，提升了临床科研与应用能力。但在实际应用中，根据模型构建的几个关键步骤，还需关注以下方面。

4.1 数据质量

高质量的数据集是支撑医学研究的前提条件，在应用过程中需考虑从数据获取、处理、业务适配等方面进行差异性分析。要对汇聚的数据按层级进行质控规则校验，对不满足值域设定的要进行排除，按照规则的分类进行分类筛选[20]。

4.2 数据标注

当前的数据标注方式仍由临床科研人员进行手动标注，标注耗时长、精度低、一致性差，不同人员对同一份数据的标注，同一个人对同一份数据在不同时间的标注都会略有差异，不利于后续用数据进行模型训练。

4.3 模型迭代

医学影像数据的处理关键是提取数据的特征和属性，根据特征值与属性值进行分类和关联分析。而使用的分类模型和分析方法都对最终的分析结果产生影响，如何选取最佳数据挖掘分析方法，需要不断地进行参数调整与验证。

4.4 数据安全

医学大数据研究涉及病历文本、影像等大量诊疗数据，须重点防范出现数据安全问题。要建立多角度的防控措施，包括大数据集群建立单独的管理区，在核心交换机划分专用 VLAN，部署数据库审计、堡垒机的安全设备等，对数据进行实时监控，确保数据安全。

参考文献

[1] 薛万国，乔岫，车贺宾，等. 临床科研数据库系统的现状与未来[J]. 中国数字医学，2021，16(1)：2-6.

[2] 张知非，杨郑鑫，黄运有，等. 医学大数据与人工智能标准体系：现状、机遇与挑战[J]. 协和医学杂志，2021，12(5)：614-620.

[3] 潘逸航，郑子龙，张国庆，等. 临床研究大数据治理平台的建设与实践[J]. 中国卫生信息管理杂志，2020，19(6)：918-924.

[4] 萧锘，曹磊，叶琪，等. 医院临床科研大数据平台数据资源分层设计研究[J]. 中国数字医学，2022，17(9)：90-94.

[5] 曹晓均，韦晓燕，毛钤镶. 医院专病数据治理实践[J]. 中国数字医学，2021，16(11)：17-20.

[6] 朱文珍，吕文志. 医疗人工智能发展现状及展望[J]. 放射学实践，2022，37(1)：1-3.

[7] 席韩旭，孙邦凯，张晨，等. 临床智能研究平台建设及相关问题探讨[J]. 医院管理论坛，2021，38(9)：88-90.

[8] 宋若齐，吴疏桐，王闯世. 医学研究中常见动态预测模型方法介绍[J]. 中国循证医学杂志，2022，22(10)：1224-1232.

[9] 牛艳艳，于洋，段芳芳，等. 全流程思维能力训练助力临床医生科研能力提升[J]. 中国医院，2023，27(8)：88-91.

[10] 李君，李晓东，宋淑洁，等. 中医临床肝病大数据知识工程研究[J]. 中西医结合肝病杂志，2023，33(5)：385-388.

[11] 曾伟，曾小琴，冉露，等. 搜索引擎技术在急诊知识库中的研究与应用[J]. 现代医药卫生，2022，38(20)：3585-3587.

[12] 龚军，孙喆，向天雨，等. 医疗大数据平台研究与实践[J]. 重庆医学，2019，48(14)：2504-2507.

[13] 胡佳迎，钟臻，侯佳音，等. 面向区域医学影像共享平台的 AI 服务建设[J]. 中国数字医学，2020，15(12)：110-112.

[14] 文含，赵莹，蔡秀定，等. 一种具有域自适应反标准化的多模态医学图像对比学习算法[J]. 生物医学工程学杂志，2023，40(3)：482-491.

[15] 张瑞萍，刘应龙，张文静，等. 基于人工智能的多模态影像辅助海马体自动勾画研究[J]. 中国医学物理学杂志，2022，39(3)：390-396.

[16] 王芳，夏雨薇，尹训涛. 影像组学分析流程及临床应用的研究进展[J]. 中华解剖与临床杂志，2021，26(2)：236-241.

[17] 林玉萍，郑尧月，郑好洁，等. 基于医学影像分割方法的多模态语料库构建[J]. 模式识别与人工智能，2021，34(4)：353-360.

[18] 庞晓燕，尹思艺，蔡秀军，等. 构建基于大数据的人工智能临床辅助决策系统方法与效果研究[J]. 中国数字医学，2020，15(9)：49-52.

[19] 胡振生，杨瑞，朱嘉豪，等. 大语言模型在医学领域的研究与应用发展[J]. 人工智能，2023，7(4)：10-19.

[20] 周琳，王飞，赵浩宇. 医院数据存在问题与管理对策研究[J]. 重庆医学，2022，51(24)：4300-4303.

（收稿日期：2023-04-18 修回日期：2023-09-22）
（编辑：唐 璞）