

# 基于超声及钼靶检查报告的乳腺癌知识图谱构建<sup>\*</sup>

胡 林<sup>1</sup> 杜雪蓓<sup>2</sup> 范 真<sup>2</sup> 周 婷<sup>2</sup>

(<sup>1</sup> 成都大学电子信息与电气工程学院 成都 610106

<sup>2</sup> 电子科技大学附属医院/四川省人民医院老年科 成都 610072)

〔摘要〕 目的/意义 提取乳腺超声及钼靶检查报告中的特征信息, 构建乳腺癌知识图谱, 促进乳腺癌诊疗数据的智能化管理, 为乳腺癌临床决策支持系统开发奠定基础。方法/过程 结合乳腺癌诊疗指南、临床专家经验构建知识图谱本体及概念层; 基于 RoBERTa 模型和 Global Pointer 进行实体关系联合抽取, 形成数据层; 选用 Neo4j 存储知识图谱。选取多种传统串联式实体关系抽取模型进行对比实验, 评估模型性能。结果/结论 基于该乳腺癌知识图谱的模型准确率、召回率及 *F1* 值均高于 95%, 优于对比模型。

〔关键词〕 乳腺癌; 知识图谱; 乳腺超声; 乳腺钼靶; 临床决策支持系统

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2025.04.010

## Construction of a Knowledge Graph for Breast Cancer Based on Ultrasound and Mammography Reports

HU Lin<sup>1</sup>, DU Xuebei<sup>2</sup>, FAN Zhen<sup>2</sup>, ZHOU Ting<sup>2</sup>

<sup>1</sup>Department of Electronic Information and Electrical Engineering, Chengdu University, Chengdu 610106, China; <sup>2</sup>Department of Geriatric, Affiliated Hospital of University of Electronic Science and Technology/Sichuan Provincial People's Hospital, Chengdu 610072, China

〔Abstract〕 **Purpose/Significance** To extract the characteristic information of breast ultrasound and mammography reports, and to construct a knowledge graph for breast cancer, so as to promote the intelligent management of breast cancer diagnosis and treatment data, and lay a foundation for the development of clinical decision support system (CDSS) for breast cancer. **Method/Process** By combining breast cancer diagnosis and treatment guidelines with clinical expert experience, the breast cancer knowledge graph ontology and conceptual layer are constructed. Then, the entity and relationship joint extraction is performed resorting to a RoBERTa model and Global Pointer so as to form the data layer. Finally, Neo4j is adopted to store the constructed knowledge graph. A variety of traditional serial entity relation extraction models are selected for comparative experiments to evaluate the performance of the models. **Result/Conclusion** The accuracy rate, the recall rate, and the *F1* value of the model based on the breast cancer knowledge graph exceed 95%, outperforming the comparison models.

〔Keywords〕 breast cancer; knowledge graph; breast ultrasound; breast mammography; clinical decision support system

〔修回日期〕 2024-10-15

〔作者简介〕 胡林, 博士, 讲师, 发表论文 10 余篇; 通信作者: 杜雪蓓。

〔基金项目〕 国家自然科学基金资助项目 (项目编号: 82301771); 医学数字影像与通讯 (DICOM) 标准国家地方联合工程实验室开放基金资助项目 (项目编号: KFKT2024002)。

1 引言

近年来乳腺癌发病率迅速上升，危害女性健康<sup>[1]</sup>。准确便捷的乳腺癌早期诊断和治疗对改善患者预后、延长生存期具有重要意义<sup>[2]</sup>。乳腺超声和钼靶检查是乳腺疾病早期筛查和诊断的重要手段，具有快捷、无创、普及度高等特点，适用于不同地区及不同收入人群<sup>[3]</sup>。然而，钼靶检查对乳腺钙化病变敏感，对不伴钙化的致密型乳腺癌敏感度及准确性不高<sup>[4]</sup>；超声检查对致密型软组织内低回声灵敏度较高<sup>[3]</sup>，对钙化组织敏感度较低，易漏诊。研究<sup>[5]</sup>表明，乳腺超声与钼靶联合检查可有效降低漏诊率。

人工智能技术在医疗领域的应用价值日渐突显<sup>[6]</sup>，其中，知识图谱技术逐渐受到关注。知识图谱是以视觉化方式展现实体间关系的语义网络，节点代表实体，边揭示实体联系<sup>[7]</sup>。医疗领域知识图谱可帮助筛选、整合医学信息，构建医学知识关系数据库，建立疾病辅助诊断、风险评估和决策系统<sup>[8]</sup>。但目前医学知识图谱仍存在以下不足。一是试图囊括所有疾病，而医学知识复杂多样、质量不一；二是主要基于单一数据源构建，忽视了医学领域多样化、高质量知识来源的有效融合<sup>[9]</sup>；三是多基于医学常识构建<sup>[10]</sup>，缺乏基于疾病真实世界数据的知识图谱构建研究。另外，乳腺癌专病知识图谱构建通常从临床指南、医学百科全书和电子病历中提取三元组知识并进行融合<sup>[11]</sup>，但相关研究成果较少且缺乏深度分析。因此，有必要基于深度融合多源临床数据构建高质量乳腺癌专病知识图谱。

2 乳腺癌知识图谱研究现状

靳淑雁等<sup>[12]</sup>借助 Protégé 软件构建基于乳腺癌专病库的知识图谱，李革等<sup>[13]</sup>利用 Neo4j 图数据库构建乳腺癌疾病知识图谱并进行可视化分析，李小龙等<sup>[14]</sup>提出超声检查报告的构建方法，Li X 等<sup>[15]</sup>提出基于钼靶检查报告的乳腺癌知识图谱构建流

程。以上研究采用基于 BiLSTM – CRF 模型<sup>[16]</sup>及其改进模型<sup>[14–15]</sup>的串联式实体关系抽取方法，在知识抽取模型、知识图谱构建等方面有所进展，推动了乳腺癌风险预测及诊断决策的智能化，但其不足是依据单一检查报告构建的知识图谱所含诊断知识有限。研究<sup>[17]</sup>证实基于 RoBERTa 的实体关系联合抽取模型表现更优。目前该方法已逐渐应用于各种专病知识图谱构建，如邓嘉乐等<sup>[18]</sup>基于 RoBERTa – CRF 进行肝癌电子病历实体识别研究，Cui X 等<sup>[19]</sup>提出基于 RoBERTa 和 Global Pointer 的中文生物医学命名实体识别模型，在实体关系联合抽取方面表现优异，但尚未应用于乳腺癌知识图谱构建。综上所述，本研究基于乳腺癌超声及钼靶检查报告，构建多源临床数据库，并从中提取相关知识，采用基于 RoBERTa 和 Global Pointer 的实体关系联合抽取模型，构建乳腺癌知识图谱。

3 乳腺癌知识图谱构建

3.1 整体流程

乳腺癌知识图谱构建整体流程，见图 1。第 1 步，知识图谱本体及概念层设计；第 2 步，数据收集、处理及整合，获取数据层；第 3 步，构建知识图谱图数据库。

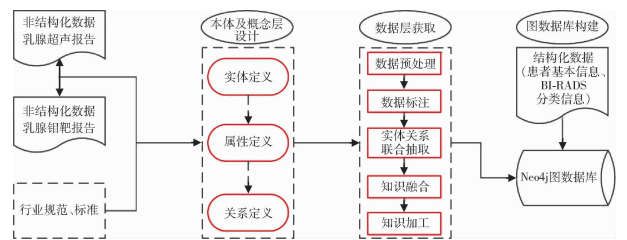


图 1 乳腺癌知识图谱构建整体流程

3.2 本体及概念层设计

知识图谱的逻辑架构分为两层：概念层（模式层）和数据层（实例层），见图 2。依据美国放射学会《乳腺影像报告与数据系统图谱》（breast imaging reporting and data system, BI – RADS）<sup>[20]</sup>和《中国抗癌协会乳腺癌诊治指南与规范（2021 年

版)》<sup>[21]</sup>，结合医学专家建议，以及多家医院报告实例，定义概念层。参考国内外文献及行业规范，修正、完善概念层，设计为 3 级分层结构。第 1 层含“常见征象”等 7 个一级实体属性分类，第 2 层含“肿块”等 17 个二级实体分类，第 3 层为相关实体属性，见表 1。概念层不同实体和属性之间主、谓、宾三要素示例，见表 2。

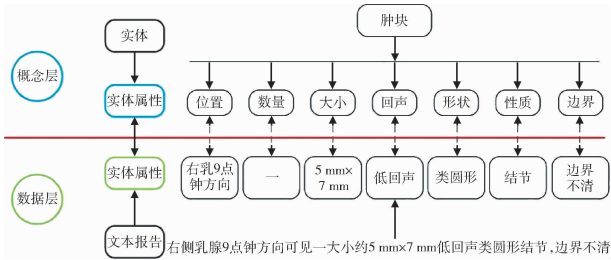


图 2 乳腺癌知识图谱概念层与数据层映射关系

表 1 乳腺癌知识图谱概念层实体及其属性

| 项目     | 一级实体 | 二级实体  | 实体属性                                |
|--------|------|-------|-------------------------------------|
| 超声检查特征 | 常见征象 | 肿块    | 形状、大小、边缘特征、内部回声、回声均匀性、后方回声、描述、部位、类型 |
|        |      | 钙化    | 数量、形态、分布、类型、位置、否定                   |
|        |      | 周围组织  | 描述                                  |
|        | 特殊征象 | 淋巴结   | 位置、描述、否定、回声                         |
|        | 伴随征象 | 血流    | 位置、否定、描述、血流速度                       |
|        |      | 皮肤    | 形态、位置、厚度                            |
|        |      | 导管变化  | 大小、描述、位置                            |
| 钼靶检查特征 | 常见征象 | 结构    | 分布                                  |
|        |      | 肿块    | 形状、大小、密度、特征、数目、位置、否定                |
|        |      | 钙化    | 位置、数目、形状、类型、大小、密度、否定、分布             |
|        | 特殊征象 | 腺体    | 分类                                  |
|        |      | 导管变化  | 描述                                  |
|        |      | 淋巴结   | 位置、分布、数目、形状、否定、描述                   |
|        | 合并征象 | 不均性   | 位置、描述、否定、回声                         |
|        |      | 血管改变  | 位置、形状、否定                            |
|        |      | 腋窝淋巴结 | 数目、形状、否定、位置、描述                      |
|        | 类别描述 | 组合异常  | 数目、位置、否定、描述                         |
|        |      | —     | 描述、位置                               |

表 2 概念层不同实体和属性之间的主、谓、宾三要素（示例）

| 主语（subject）      | 谓语（predicate） | 宾语（object）               |
|------------------|---------------|--------------------------|
| 实体（基本信息）         | select        | 属性（姓名、性别等）               |
| 实体（报告）           | has_a         | 实体（基本信息、乳腺信息、BI-RADS 信息） |
| 实体（乳腺信息）         | instance_of   | 一级实体（常见征象、合并征象等）         |
| 一级实体（常见征象、合并征象等） | part_of       | 二级实体（肿块、钙化、结构等）          |
| 二级实体（结构、肿块等）     | select        | 属性（位置、大小等）               |
| 特征对比             | analyzes      | 肿块、淋巴结、导管变化、钙化特征的差异和相似性  |
| 总体评估             | evaluates     | 综合超声和钼靶结果的诊断             |

3.3 数据层获取

传统方法通常将实体关系抽取任务分为实体识别和关系抽取两个子任务，先对句子进行实体识别，再根据实体识别结果进行实体间关系抽取，是串联式抽取<sup>[22]</sup>。本研究采用 RoBERTa 预训练模型

结合 Global Pointer 完成实体关系联合提取。

3.3.1 数据来源及预处理 以四川省人民医院放射科 2019 年 10 月—2021 年 8 月 1 084 例女性患者的乳腺钼靶和超声报告为数据源。删除重复、不完整或错误报告，经去重、缺失值处理、繁简转换、清除无效信息、去除停用词等数据清洗和必要脱敏

处理，选取 1 002 份报告用于构建知识图谱。依据知识图谱概念层设计对 1 002 份报告进行 BI - RADS 变量标注。在专业医生指导下制定标注规范，形成经医学专家审核的标注说明文档。基于 Web 框架开发标注工具，采用迭代模式预标注，经多轮标注一

致性评价及专家审核，完成正式标注，见图 3。标注后，构建映射关系，进行数据对齐、标签序列转换，其中 B 为实体开始标记，I 为实体内标记，O 为非实体标记。

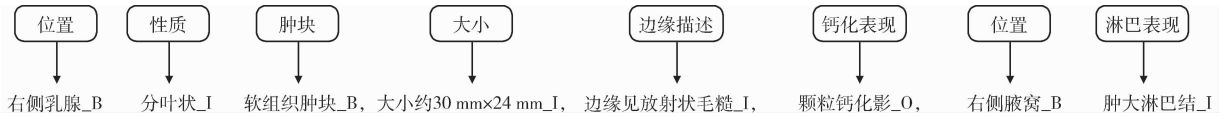


图 3 数据标注示例

3.3.2 实体关系联合抽取 构建基于 RoBERTa 的预训练模型，结合 Global Pointer 执行实体关系联合抽取。RoBERTa<sup>[23-25]</sup> 是双向编码器表征（bidirectional encoder representations from transformers, BERT）的改进版，专为自然语言处理设计，通过优化预训练细节，提高模型理解能力和泛化能力。Global Pointer 是基于 span 分类的解码方法。模型构架，见图 4。输入层：输入乳腺癌超声和钼靶报告文本，含患者信息、病变特征、病变位置等信息。Transformer 层（RoBERTa）：包含 12 层 Transformer 的 RoBERTa 模型，将文本编码为高维特征表示，捕获词汇、语法和上下文信息，生成词语或标记的向

量表示。Global Pointer 解码层：对 RoBERTa 输出解码，处理实体重叠，预测文本中任意两位置间是否存在特定关系，实现实体及关系联合抽取。输出层：输出识别出的实体及其关系，如患者信息、病变特征、病变位置及其关联。选用命名实体识别任务中常用的性能指标评估模型性能。精确率（precision, P）表示预测为正例的实体中正例样本的比例。召回率（recall, R）表示所有正例样本中被正确识别为正例的比例。F1 值为 1 时（即精确率和召回率均为完美），模型表现最佳；F1 值为 0 时，模型表现最差。

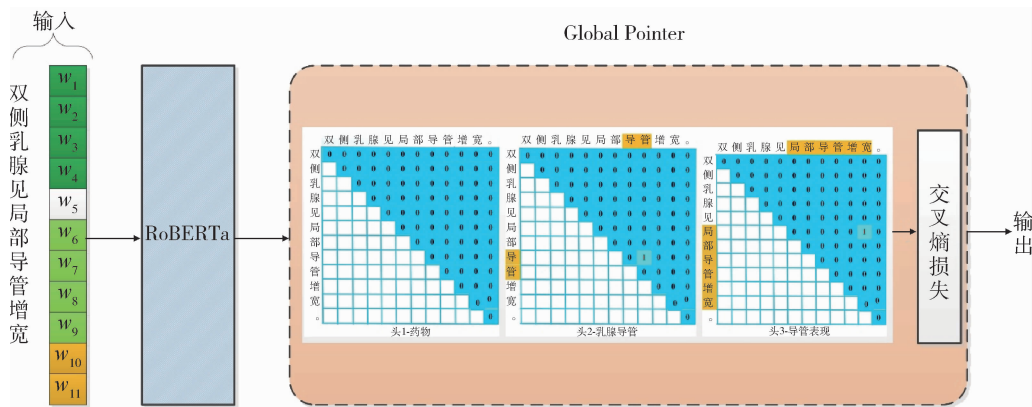


图 4 实体关系联合抽取模型构架

3.4 知识融合与知识加工

知识融合指融合各层知识，包括不同知识库中的同一实体类别、多源异构知识等，以升级和维护知识图谱。为避免医学实体命名不统一导致的信息冗余和误解，本研究依据乳腺癌相关指南、规范和

行业标准开展知识融合，通过文本相似度计算和标准化词汇库，替换同义词为标准术语，实现同义描述和实体消歧。知识加工是形成高质量知识图谱不可或缺的部分，便于知识的大规模统一管理。采用文本匹配和语义相似计算方法进行加工，将数据与概念层中的实体和实体属性一一对应，完成实体与

属性配对。最终，依据本体构建，完成乳腺癌知识图谱“实体 - 属性 - 关系”对应后，利用 Neo4j 构建知识图谱图数据库。

3.5 实验设计与结果分析

3.5.1 实验设计 将标注数据按 8:1:1 比例划分为训练集、测试集和验证集。引入联合抽取模型，用训练集进行首轮训练；训练完成后用验证集验证，观察结果，确定模型调优方向。模型训练流程，见图 5。模型初始参数设置，见表 3。模型迭代过程采用 Adam 优化器，并使用 FGM 对抗训练方式。

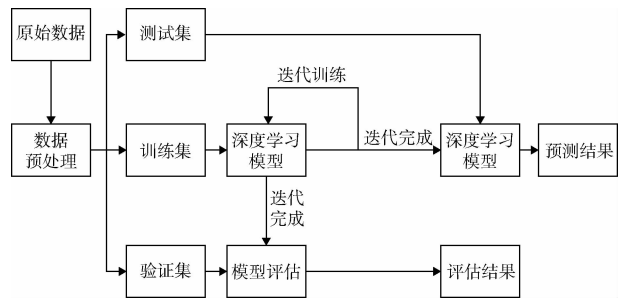


图 5 模型训练流程

表 3 模型参数设置

| 参数名称   | 具体设置               |
|--------|--------------------|
| 最大句子长度 | 256                |
| 批次大小   | 8                  |
| 学习率    | 2e-5               |
| 优化器    | Adam               |
| 损失函数   | Cross Entropy Loss |

3.5.2 实体关系联合抽取结果 对乳腺超声、钼靶报告中的实体及其关系进行联合抽取，形成乳腺癌知识图谱，其中实体类别及其数量，见图 6。腺体、侧别、BI - RADS 等级等实体数量较多，形状、手术治疗、预后等实体数量较少，大部分类别实体数量少于 20 个。该图展示了乳腺癌知识图谱构建的重点数据来源。

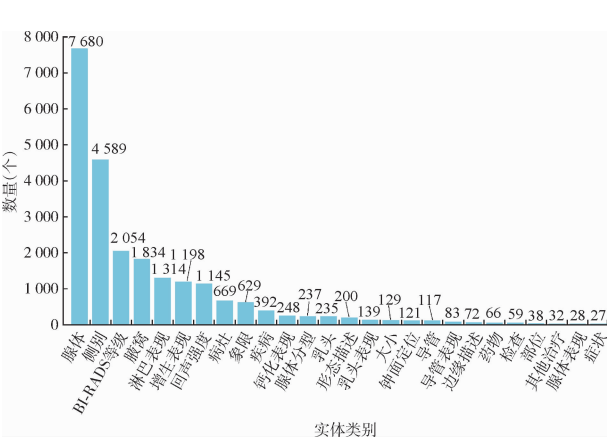


图 6 乳腺癌知识图谱实体分布（部分）

乳腺癌知识图谱关系分布，见图 7，其中腺体\_侧别、腺体\_BI - RADS 等级和腋窝\_侧别是 3 个最主要的关系，数量均超过 1 000 个，表明在乳腺超声及钼靶报告中，腺体与侧别、BI - RADS 等级、腋窝等的关系非常核心，可能是乳腺检查的关键指标。以外，腺体\_增生表现、腺体\_回声强度、腺体\_钙化表现等特性与现象之间更详细的关系也比较多见，体现了乳腺癌知识图谱的多样性和复杂性。

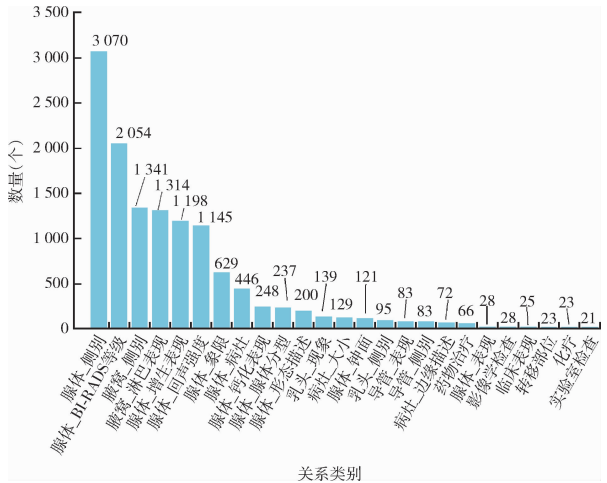


图 7 乳腺癌知识图谱关系类别分布（部分）

使用 FGM 对抗训练的实体关系抽取模型在初期学习阶段经历了一些震荡，但随着时间的推移，准确率逐渐改善，召回率越来越高并逐渐稳定，且两者波动性小，见图 8。显示本文模型在后期学习



阶段，性能趋于稳定，在实体识别准确性和完整性之间达到平衡，即模型达到了其在当前数据和架构下的最佳性能。

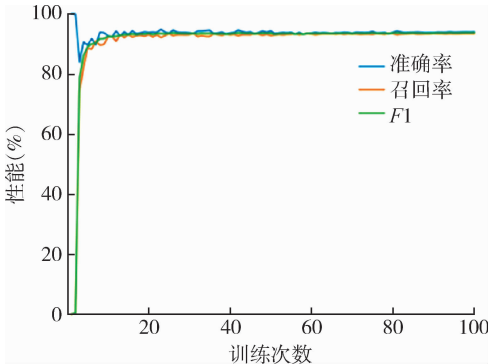


图8 乳腺癌实体关系抽取模型训练准确率变化趋势

选取传统串联式实体关系抽取模型 BiLSTM - CRF<sup>[16]</sup> 及其最新改进模型 BiLSTM - Attention - CRF<sup>[14]</sup>、BiLSTM - Highway - CRF<sup>[15]</sup> 和主流的 BERT<sup>[24]</sup> 联合抽取模型，对比不同模型在实体关系联合抽取任务中的表现，见表 4。本文模型的准确率、召回率、F1 均最高，充分体现了 RoBERTa 模型强大的语言理解能力以及 Global Pointer 的优势。

| 表 4 各模型测试集性能对比 (%)                       |       |       |       |
|------------------------------------------|-------|-------|-------|
| 模型                                       | 准确率   | 召回率   | F1    |
| BiLSTM - CRF <sup>[16]</sup>             | 91.90 | 89.64 | 88.75 |
| BiLSTM - Attention - CRF <sup>[14]</sup> | 92.00 | 90.36 | 91.04 |
| BiLSTM - Highway - CRF <sup>[15]</sup>   | 92.50 | 91.12 | 91.47 |
| BERT <sup>[24]</sup>                     | 93.80 | 92.29 | 92.51 |
| 本文模型                                     | 95.90 | 95.30 | 95.60 |

3.5.3 知识融合结果 通过知识融合可以解决乳腺癌医学文本中临床术语写法、描述多样的问题，实现临床术语标准化，以避免实体命名不一致导致的信息重复和误解。如原始词“BI - RADS: 3 类”“BI - RADS 3”“BI - RADS 3 类”“BI - RADS III 类”通过知识融合表达为标准词“BI - RADS 3 类”。对抽取出的医学实体进行对齐处理，得到实体合计 23 421 个，关系合计 13 000 个。

4 结语

本研究提出一种准确高效构建乳腺癌知识图谱的系统性流程，包括设计本体及概念层，构建 RoBERTa 预训练模型，结合 Global Pointer，对乳腺超声和钼靶报告进行实体关系联合抽取，再进行知识融合、加工，利用 Neo4j 构建知识图谱数据库。实验证明了模型联合抽取的高效性和准确性。本研究仍存在不足，一方面，在数据集选取方面存在局限，后续将结合更多数据集，优化知识图谱，提高模型的通用性和鲁棒性；另一方面，未来将开展乳腺癌知识图谱可用性研究，搭建数据共享管理平台，以期实现风险预测和辅助临床决策。

作者贡献：胡林负责研究设计、论文修订；杜雪蓓负责具体实验、数据分析、论文撰写；范真、周婷负责论文审核与修订。

利益声明：所有作者均声明不存在利益冲突。

参考文献

- 1 LEI S, ZHENG R, ZHANG S, et al. Breast cancer incidence and mortality in women in China: temporal trends and projections to 2030 [J]. Cancer biology & medicine, 2021, 18 (3): 900 - 909.
- 2 HARBECK N, GNANT M. Breast cancer [J]. Lancet, 2017, 389 (10074): 1134 - 1150.
- 3 MANN R M, HOOLEY R, BARR R G, et al. Novel approaches to screening for breast cancer [J]. Radiology, 2020, 297 (2): 266 - 285.
- 4 NAM K J, HAN B K, KO E S, et al. Comparison of full - field digital mammography and digital breast tomosynthesis in ultrasonography - detected breast cancers [J]. The breast, 2015, 24 (5): 649 - 655.
- 5 ACCIACATTI R J, LEE S H, REIG B, et al. Beyond breast density: risk measures for breast cancer in multiple imaging modalities [J]. Radiology, 2023, 306 (3): e222575.
- 6 SHEIKH A, ANDERSON M, ALBALA S, et al. Health information technology and digital innovation for national learning health and care systems [J]. The lancet digital health, 2021, 3 (6): e383 - e396.

- 7 PENG C, FENG X, NASERIPARSA M, et al. Knowledge graphs: opportunities and challenges [J]. Artificial intelligence review, 2023, 56 (11): 13071 – 13102.
- 8 YANG Y, LU Y, YAN W. A comprehensive review on knowledge graphs for complex diseases [J]. Briefings in bioinformatics, 2023, 24 (1): 1 – 20.
- 9 WANG H, ZU Q, LU M, et al. Application of medical knowledge graphs in cardiology and cardiovascular medicine: a brief literature review [J]. Advances in therapy, 2022, 39 (9): 4052 – 4060.
- 10 XUE Q, WANG B, MIAO K, et al. Construction of knowledge base of Chinese medicine manufacturing [J]. China journal of Chinese materia medica, 2022, 47 (12): 3402 – 3408.
- 11 AN B. Construction and application of Chinese breast cancer knowledge graph based on multi – source heterogeneous data [J]. Mathematical biosciences and engineering, 2023, 20 (4): 6776 – 6799.
- 12 靳淑雁, 王爽, 黄琼, 等. 基于乳腺癌专病库的知识图谱构建研究 [J]. 医学信息学杂志, 2023, 44 (12): 65 – 70.
- 13 李革, 张云锁, 丁腊春, 等. 基于图数据库的乳腺疾病知识图谱构建与应用 [J]. 中国现代医生, 2023, 61 (23): 106 – 109.
- 14 李小龙, 孙水发, 唐庭龙, 等. 基于超声检查报告的乳腺癌诊断知识图谱构建 [J]. 武汉大学学报 (理学版), 2023, 69 (1): 69 – 78.
- 15 LI X, SUN S, TANG T, et al. Construction of a knowledge graph for breast cancer diagnosis based on Chinese electronic medical records: development and usability study [J]. BMC medical informatics and decision making, 2023, 23 (1): 210.
- 16 HUANG Z, XU W, YU K. Bidirectional LSTM – CRF models for sequence tagging [EB/OL]. [2024 – 09 – 22]. <http://arxiv.org/abs/1508.01991>.
- 17 王华峰, 王久阳. 一种基于 Roberta 的中文实体关系联合抽取模型 [J]. 北方工业大学学报, 2020, 32 (2): 90 – 98.
- 18 邓嘉乐, 胡振生, 连万民, 等. 基于 RoBERTa – CRF 的肝癌电子病历实体识别研究 [J]. 医学信息学杂志, 2023, 44 (6): 42 – 47.
- 19 CUI X, SONG C, LI D, et al. RoBGP: a Chinese nested biomedical named entity recognition model based on Roberta and global pointer [J]. Computers, materials & continua, 2024, 78 (3): 3603 – 3617.
- 20 美国放射学院. 乳腺影像报告与数据系统图谱 2013 版 [M]. 王殊, 洪楠, 译. 北京: 北京大学医学出版社, 2016.
- 21 中国抗癌协会乳腺癌专业委员会. 中国抗癌协会乳腺癌诊治指南与规范 (2021 年版) [J]. 中国癌症杂志, 2021, 31 (10): 954 – 1040.
- 22 GOYAL A, GUPTA V, KUMAR M. Recent named entity recognition and classification techniques: a systematic review [J]. Computer science review, 2018, 29 (8): 21 – 43.
- 23 LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach [EB/OL]. [2024 – 09 – 22]. <http://arxiv.org/abs/1907.11692v1>.
- 24 DEVLIN J, CHANG M, LEE K, et al. Bert: pre – training of deep bidirectional transformers for language understanding [EB/OL]. [2024 – 09 – 22]. <http://arxiv.org/abs/1810.04805>.
- 25 VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30 (8): 5998 – 6008.

## 《医学信息学杂志》开通微信公众号

《医学信息学杂志》微信公众号现已开通, 作者可通过该平台查阅稿件状态; 读者可阅览当期最新内容、过刊等; 同时提供国内外最新医学信息研究动态、发展前沿等, 搭建编者、作者、读者之间沟通、交流的平台。可在微信添加中找到公众号, 输入“医学信息学杂志”进行确认, 也可扫描右侧二维码添加, 敬请关注!



《医学信息学杂志》编辑部