

基于深度学习的乳腺超声报告 BI-RADS 抽取增强型自然语言处理方法

郭雅琳, 韩宇辰*

华中科技大学物理学院, 武汉 430074, 湖北, 中国

摘要: 乳腺癌是全球致死率最高的疾病之一, 准确解读乳腺超声报告对于早期诊断与治疗至关重要。然而, 这类报告中的关键发现往往埋没在非结构化文本里, 给自动化抽取造成困难。本文提出一种基于深度学习的自然语言处理 (NLP) 方法, 用于从乳腺超声数据中抽取乳腺影像报告与数据系统 (BI-RADS) 类别。我们在一批取自沙特阿拉伯某医院、经人工标注的报告数据集上训练了循环神经网络 (RNN) 模型, 具体采用双向长短期记忆网络 (BiLSTM) 结构; 同时引入不确定性估计技术来处理歧义病例, 并借助数据增强提升模型性能。实验结果表明, 本文的深度学习方法优于传统的基于规则的方法与机器学习方法, 在分类任务中取得了令人满意的准确率。本研究对于放射学报告的自动化、辅助临床决策以及推进乳腺癌研究均具有重要意义。

关键词: BI-RADS; 深度学习; 长短期记忆网络; 命名实体识别; 自然语言处理; 循环神经网络

An Enhanced NLP Approach for BI-RADS Extraction in breast Ultrasound Reports Using Deep Learning

YaLin Guo, YuChen Han*

School of Physics, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

Abstract: Breast cancer stands as one of the top causes of death around the globe, making the accurate interpretation of breast ultrasound reports vital for early diagnosis and treatment. Unfortunately, key findings in these reports are often buried in unstructured text, complicating automated extraction. This study presents a deep learning-based natural language processing (NLP) approach to extract breast imaging reporting and data system (BI-RADS) categories from breast ultrasound data. We trained a recurrent neural network (RNN) model, specifically using a BiLSTM architecture, on a dataset of reports that were manually annotated from a hospital in Saudi Arabia. Our approach also incorporates uncertainty estimation techniques to tackle ambiguous cases and uses data augmentation to boost model performance. The experimental results indicate that our deep learning method surpasses traditional rule-based and machine-learning techniques, achieving impressive accuracy in classification tasks. This research plays a significant role in automating radiology reporting, aiding clinical decision-making, and pushing forward the field of breast cancer research.

Keywords: BI-RADS; Deep-learning; LSTM; Named entity recognition; NLP; RNN

一、引言

乳腺癌仍是全球最常见的癌症之一，无论对患者个体还是公共卫生系统都构成沉重负担[1]。早期发现对于降低死亡率、减轻经济压力至关重要。因此，医学指南普遍主张通过常规乳腺 X 线摄影筛查来评估乳腺癌风险[2]。为减少差异、规范放射科医师对乳腺 X 线摄影结果的报告，美国放射学会（ACR）推出了乳腺影像报告与数据系统（BI-RADS）[3]。该系统为乳腺 X 线摄影所见提供了一套标准化术语，并给出用于评估恶性风险的六类分级框架，见表 1[4]。BI-RADS 最初虽是为乳腺 X 线摄影设计的，但后来已被推广到磁共振成像（MRI）、乳腺超声等其他影像手段。然而，这些重要的放射学发现常以非结构化的叙述形式记录，依赖结构化数据的计算系统因此无从利用。已有大量研究采用自然语言处理（NLP）技术，从各类英文乳腺影像报告（包括乳腺 X 线摄影与乳腺超声数据）中抽取 BI-RADS 所见与最终评估类别。早期方法以基于规则的为主[5]，例如 MedLEE——最早用于从乳腺 X 线摄影报告中抽取异常所见的临床 NLP 系统之一[6]。

Table 1 A Description of BI-RADS Six-Category Mammography Findings

表 1 BI-RADS 乳腺 X 线摄影所见的六类分级说明

BI-RADS 类别	所见	乳腺癌可能性	处理建议
0	需补充影像检查或既往检查资料	不适用	召回补充影像检查和 / 或等待既往检查资料
1	阴性	可忽略	常规筛查
2	良性	可忽略	常规筛查
3	可能良性	<2%	短期随访
4	可疑	23%~34%	组织学诊断
5	高度提示恶性	>95%	组织学诊断
6	活检证实为恶性	100%	临床适宜时行手术切除

另有一些基于规则的方法被用于从乳腺 X 线摄影记录中划分 BI-RADS 乳腺组织构成类别[7]，以及借助 GATE 自然语言处理框架考察 BI-RADS 评估类别中的歧义问题[8]。此后出现了基于机器学习的 NLP 系统，例如用支持向量机（SVM）与朴素贝叶斯（NB）抽取 BI-RADS 类别与侧别信息，F1 值达到 0.95，性能超过基于规则的方法[9]。此外，还有一些 NLP 流水线被用于抽取 BI-RADS 评估类别，并借助统计检验，用抽取到的所见来支持临床决策。

随着电子健康档案的普及，沙特阿拉伯已积累起海量的临床文书数字档案，亟需以 NLP 手段从中提炼有价值的信息[10-12]。已有若干研究探索了在沙特医疗记录中做症状抽取与疾病进展分析的 NLP 技术。近年来，针对沙特临床文本的研究已扩展到更广泛的 NLP 问题，包括否定检测以及从手术记录中抽取肿瘤相关信息。尽管如此，此前尚无研究专门针对沙特乳腺超声报告的 BI-RADS 所见做结构化抽取，这正是该领域的一处关键空白。

结构化的 BI-RADS 抽取对临床决策至关重要。为填补这一空白，本文针对沙特阿拉伯的乳腺超声数据提出一种能够抽取全部 BI-RADS 所见分类的深度学习方法。鉴于学界对深度学习模型的兴趣日增，本文明确考察了双向长短期记忆（BiLSTM）网络在该任务上的应用，具体探索了以 BiLSTM 为基础的循环神经网络（RNN）用于从乳腺超声报告中抽取 BI-RADS 所见。我们用一套包含 465 份报告的标注数据集表明，深度学习方法优于传统的基于条件随机场（CRF）的机器学习方法，说明其在推进乳腺癌研究与临床决策支持方面具有潜力。

近年的 NLP 研究表明, 在命名实体识别 (NER) 任务上, 基于 RNN 的模型优于 CRF 模型, 在结合人工构造特征与领域词典时尤其如此。在临床领域, RNN 已被成功用于医学事件检测医学概念抽取临床语境下的时间信息抽取以及疾病名称识别。然而, 在沙特医疗场景中运用深度学习与机器学习抽取 BI-RADS 的研究仍属空白。本文余下部分的结构如下: 首先介绍方法与实验; 其次给出结果并讨论所得发现; 最后在第四节对全文作总结。

二、方法与实验

(一) 数据集与标注

数据来源: 本文使用的乳腺超声报告来自沙特阿拉伯阿西尔省海米斯穆谢特妇产医院, 时间跨度为 2015 年至 2020 年。所有报告均已做匿名化处理, 把患者姓名、住址、电话号码、医务人员姓名等敏感信息替换为代号或化名, 以保护患者隐私, 同时保证文本读起来不出现异常的空缺。需要说明的是, 本文的数据集只有文本报告, 不含配套的超声图像。

标注流程: 标注指南的制定采用了迭代式的做法。

i) 初稿拟定: 与领域专家合作, 拟出标注指南初稿。

ii) 预标注: 两名标注者依据初版指南, 各自独立标注了 65 份报告构成的子集。标注者间一致性的 F 值为 0.821, 说明该任务相当复杂。

iii) 指南修订: 针对观察到的分歧, 对指南作了澄清与补充。

iv) 重新标注: 用更新后的指南对同样这 65 份报告重新标注, F 值提升至 0.942。

v) 全量标注: 把定稿的指南用于其余 400 份报告, 最终得到 465 份标注报告。整体流程见图 1。

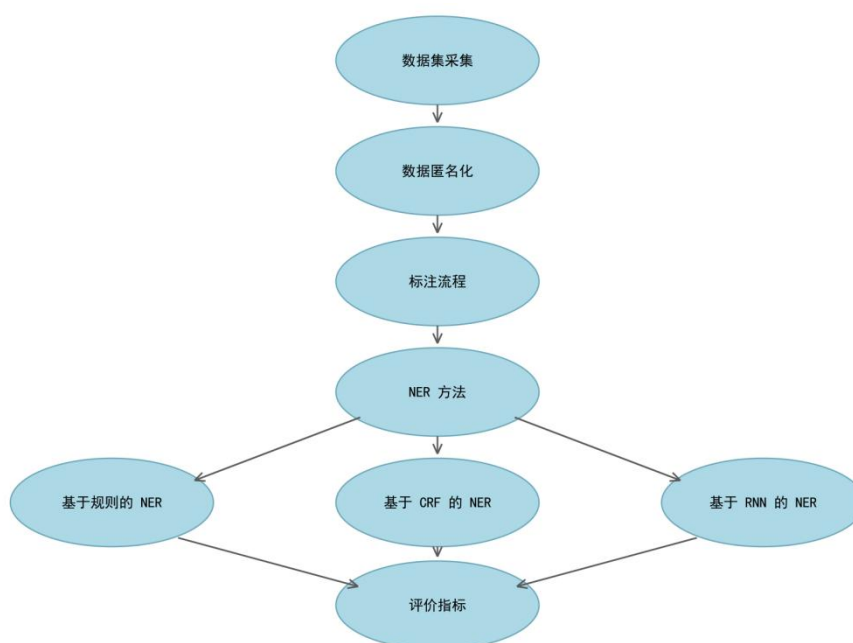


Figure 1 The Method Flowchart is the Entire Methodology Pipeline

图 1 方法流程图, 即整条方法学流水线

标注工作借助开源的网页版标注工具 BRAT 完成。除实体标注外, 我们还标注了每个实体的否定状态, 示例见图 2。随后把标注语料切分为两部分: 训练集 310 份报告 (约三分之二), 测试集 155 份报告 (约三分

之一)。

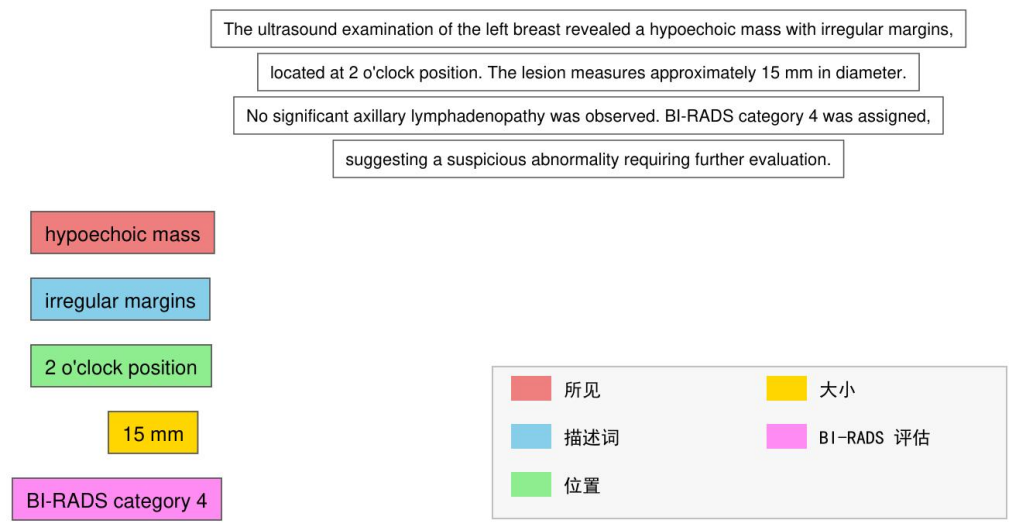


Figure 2 BART Annotation for a Breast Cancer Radiology Report
图 2 一份乳腺癌放射学报告的 BRAT 标注

(二) 命名实体识别方法

命名实体识别是自然语言处理中的一项关键任务，即在文本中识别实体并把它们归入既定类别，如人名、机构名、地名以及某一领域内的专门术语，见图 3。本文考察了三种 NER 方法。

1. 基于规则的方法

作为基线，该方法把人工编写的规则与实体词典结合起来，通过 UIMA Ruta 框架实现。实体词典由标注好的开发集构建，并设计正则表达式来识别特定模式。Ruta 规则可处理较为复杂的组合情形，例如把数字与单位拼在一起的写法。

2. 基于条件随机场的方法

条件随机场（CRF）是擅长序列标注任务的概率模型。本文使用 CRF++ 工具包，引入词袋、n 元语法等关键的 NER 特征来构建 NER 模型。CRF 能够纳入上下文信息，适合本文这类任务。

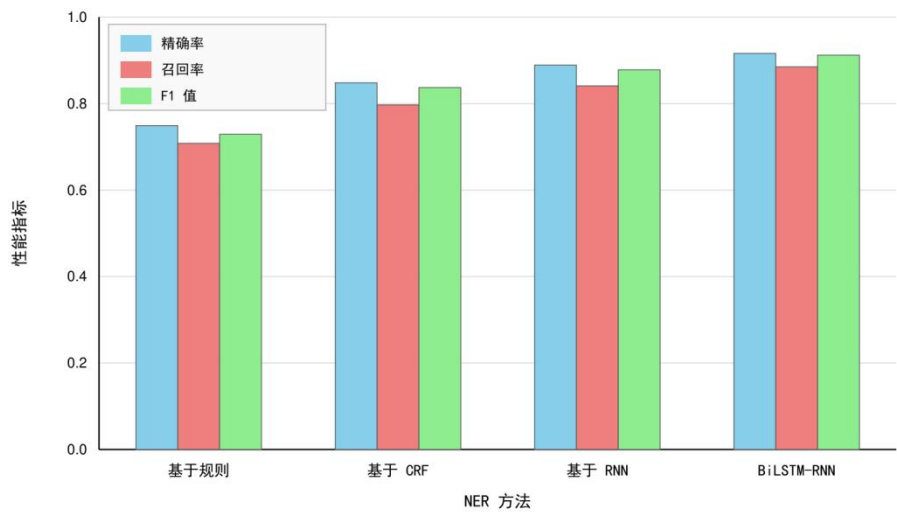


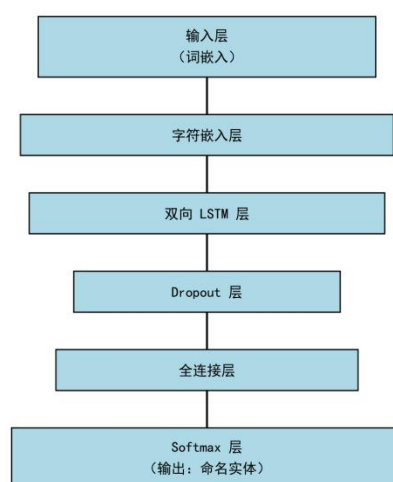
Figure 3 Performance Comparison on NER Methods**图 3** 各 NER 方法的性能对比

3. 基于循环神经网络的深度学习方法

捕捉序列数据中的长程依赖是 RNN——尤其是 LSTM 网络——的强项。本文实现了带 LSTM 单元的 RNN 结构，其设计取自 Lample 等工作，并作了如下改进：

—字符嵌入：捕捉词的形态学特征。

—双向 LSTM：放射学报告是有结构的序贯医学文本，BiLSTM 能有效捕捉文本模式中的长程依赖；它沿正向与反向两个方向处理序列，从而同时利用上文与下文信息，见图 4。

**Figure 4** BiLSTM-Based Model Architecture for BI-RADS Extraction**图 4** 用于 BI-RADS 抽取的 BiLSTM 模型结构

最终 RNN 模型所需的设置为：

—字符嵌入维度：50

—LSTM 层规模：词级 100 个单元

—学习率：0.005

—丢弃概率：0.5

训练损失与验证损失要经过若干轮之后才开始变化。BiLSTM 模型的损失收敛曲线，说明训练过程稳定而有效。虽然 BERT 等基于 Transformer 的模型在 NLP 任务上表现出色，本文仍选择基于 BiLSTM 的方案，原因在于它处理序贯医学文本效果好、算力需求低，并且在中等规模数据集上泛化性更佳。

评价指标：本文用以下标准指标评估 NER 系统的效能。

—精确率（PRE）：预测为正的样本中实际为正的比率。

—召回率（REC）：实际为正的样本中被正确预测出来的比率。

—F1 值（F1）：精确率与召回率的加权综合。

—准确率（ACC）：预测正确的样本占全部样本的比例。

三、结果与讨论

(一) 结果

按照 BI-RADS 标准, 专家在乳腺超声报告中确定了 20 类实体, 另设「其他」一类归集出现频次很低的情形。465 份报告共标注出 9 132 个实体。位置、回声、大小、血供这四类实体最为常见, 各自出现 1 000 次以上; 相反, 结构扭曲、钙化、组织构成等十类实体的出现次数不足 100 次, 详见表 2。

Table 2 Distribution of BI-RADS Entity Categories in the Annotated Corpus

表 2 标注语料中 BI-RADS 实体类别的分布

实体类别	实体数量
Alder 分级 (Alder)	15
结构扭曲 (Architectural-distortion)	22
钙化 (Calcifications)	30
导管改变 (Ductchanges)	52
回声 (Echo)	1 308
弹性评估 (Elasticity-assessment)	60
硬度比 (Hardness-ratio)	19
位置 (Location)	2 301
淋巴结 (LymphNode)	399
边缘 (Margin)	801
肿块 (Masses)	158
否定 (Negation)	655
方位 (Orientation)	26
后方特征 (Posterior-features)	31
阻力指数 (Resistance-index)	231
形态 (Shape)	461
大小 (Size)	1 440
皮肤 (Skin)	131
组织构成 (Tissue-composition)	14
血供 (Vascularity)	958
其他 (Other)	20
合计	9 132

本文希望确定基于规则的方法、基于 CRF 的模型与基于 RNN 的模型 (BiLSTM) 三者中, 哪一种在 NER 评估中最为有效。采用 BiLSTM 的深度学习模型取得了最高的 F1 值 0.908, 其次是基于 CRF 的模型 (F1 值 0.885), 基于规则的方法最低 (F1 值 0.864)。与较为传统的方法相比, 这一结果表明基于深度学习的系统从乳腺超声记录中抽取 BI-RADS 结论方面更为有效。

(二) 讨论

1. 弥补既有研究的空白

以往研究虽已用基于规则与统计的方法探索过 BI-RADS 实体抽取, 但把深度学习用于乳腺超声报告的结构化信息抽取仍缺乏充分研究。此外, 多数既有研究关注的是乳腺 X 线摄影而非超声影像。为填补这些空白,

本文采用基于 BiLSTM 的深度学习 NER 系统, 从乳腺超声数据中找出 BI-RADS 信息。

2. 主要发现与创新点

本文表明, 基于 RNN 的深度学习模型 (BiLSTM) 在抽取 BI-RADS 实体方面优于传统的 CRF 与基于规则的方法。基于 RNN 的模型取得的最高 F1 值 0.908, 印证了深度学习在该领域的有效性。此外, 本文的标注流程涵盖 18 类 BI-RADS 实体, 所提供的标注数据集比以往工作更为全面, 这也是本研究特色所在。

3. 与既有文献的比较

以往研究已经表明深度学习在医学文本处理中行之有效, 本文的结果与之一致。例如, An 等证明 BiLSTM 模型能够提升临床文本中的实体识别效果。但本文更进一步, 把深度学习准确地用到了乳腺超声报告上, 而不是更宽泛的临床叙述文本。与以往只依赖手工规则或统计模型的方法不同, 本文的深度学习模型能有效捕捉上下文依赖, 从而提高了准确率。

4. 本研究的局限

尽管结果令人鼓舞, 本研究仍有若干局限。首先, 数据集虽然涵盖了多样的报告, 但部分实体类别出现频次偏低, 可能影响了模型的泛化能力。其次, 由于缺乏图文对齐, 实体抽取只能在文本数据上进行, 无法获得多模态的信息。后续研究应考虑把影像特征与文本分析结合起来。

5. 对后续研究的启示

后续研究可探索把深度学习与基于规则的技术结合起来的混合模型, 以改善低频实体类别上的表现。用来自多家机构的报告扩充数据集, 有望增强模型的稳健性与泛化能力。探索 BERT、BioBERT 等基于 Transformer 的结构, 也可能提高实体抽取的准确率。

6. 小结

本文的结果证实, 深度学习方法——尤其是基于 RNN 的模型——能更有效地从乳腺超声报告中抽取 BI-RADS 结果。本研究提供了一套完整的标注数据集, 并论证了深度学习用于放射学结构化信息抽取的可行性, 从而为该领域作出贡献。多模态学习与数据集扩充方面的后续进展, 有望进一步提升 BI-RADS 分类的自动化水平与临床决策支持能力。

四、结论

本研究成功解决了借助深度学习从非结构化的乳腺超声报告中抽取结构化 BI-RADS 所见这一难题。正如引言中所预期的, 本文的 BiLSTM 模型在 BI-RADS 抽取上优于基于规则与基于 CRF 的方法, 说明它能够有效捕捉放射学报告中的语言模式。该模型取得了 0.908 的 F1 值, 印证了其用于自动化临床决策支持的潜力。

本文的发现与既有的医学文本处理深度学习研究一致, 并把应用范围拓展到了此前研究不足的乳腺超声领域。虽然模型准确率较高, 但数据集多样性不足、缺乏图文对齐等局限也指出了需要进一步研究的方向。后续工作应探索深度学习与规则相结合的混合模型, 跨机构扩充数据集, 并引入 BERT 等 Transformer 模型以提高准确率。引入多模态学习——把文本与超声图像对齐——还可以进一步增强人工智能 (AI) 驱动的放射学决策支持。

参考文献

- [1] 李爱群. 老旧建筑抗震加固改造技术. 工程力学, 2022, 39(04): 1-10.
- [2] 刘吉臻. 新能源火电耦合调峰系统. 动力工程学报, 2022, 42(04): 265-273.

- [3] 张友军. 船舶动力系统减振降噪技术. 船舶工程, 2022, 44(03): 1-7.
- [4] 杨孟飞. 星载计算机容错设计技术. 中国空间科学技术, 2022, 42(02): 1-8.
- [5] 王树国. 机器人关节驱动器高精度控制. 机器人, 2022, 44(02): 1-8.
- [6] 郝吉明. 工业烟气多污染物协同治理. 中国环境科学, 2022, 42(04): 1501-1508.
- [7] Berg A O. Screening for breast cancer: recommendations and rationale. *Annals of Internal Medicine*, 2002, 137(5): 344-346.
- [8] 肖立业. 超导输电线路工程化关键问题. 电工电能新技术, 2022, 41(02): 1-10.
- [9] 朱合华. 城市地下管网数字孪生建模. 地下空间与工程学报, 2022, 18(02): 369-378.
- [10] 李元昊. 数控机床热误差补偿技术. 中国机械工程, 2022, 33(08): 905-912.
- [11] 陈发虎. 高寒冻土路基稳定控制技术. 冰川冻土, 2022, 44(02): 478-486.
- [12] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 2018, 68(6): 394-424.