

基于 ERNIE 分类模型的淋巴水肿智能辅助诊断研究^{*}

汤昊宸¹ 信建峰² 冀秀元¹ 常 鲲² 孙宇光² 夏 松² 徐 毅¹ 沈文彬²

(¹ 中国科学院自动化研究所 北京 100190 ² 首都医科大学附属北京世纪坛医院 北京 100038)

〔摘要〕 **目的/意义** 探讨人工智能辅助诊断技术应用于淋巴疾病的诊断效果, 阐述样本稀缺背景下模型微调训练的解决方案, 指出应用研究的难点和不足, 并提出展望。**方法/过程** 从北京世纪坛医院淋巴外科既往收治的患者中选取 755 例为研究对象, 基于患者电子病历非结构化文本数据, 利用 ERNIE 3.0 分类模型进行淋巴水肿辅助诊断应用研究, 通过两个层级分类任务, 实现智能化诊断淋巴水肿疾病并区分原发性和继发性。**结果/结论** ERNIE 3.0 分类模型在淋巴水肿的判别任务中, 准确率、精确率、召回率、平均 *F1* 值均超过 0.95。在原发性和继发性淋巴水肿的区分上, 模型的平均 Macro *F1* 值达到 0.901。两个任务模型的曲线下面积分别达到 0.97 和 0.865, 表明模型的准确分类效果。模型分类结果具有较好的可解释性, 填补了智能化方法在淋巴疾病领域的应用空白。

〔关键词〕 原发性淋巴水肿; 继发性淋巴水肿; 辅助诊断

〔中图分类号〕 R-058 **〔文献标识码〕** A **〔DOI〕** 10.3969/j.issn.1673-6036.2023.10.008

Study on Intelligent Auxiliary Diagnosis of Lymphedema Based on ERNIE Classification Model

TANG Haocheng¹, XIN Jianfeng², JI Xiuyuan¹, CHANG Kun², SUN Yuguang², XIA Song², XU Yi¹, SHEN Wenbin²

¹Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; ²Beijing Shijitan Hospital, Capital Medical University, Beijing 100038, China

〔Abstract〕 **Purpose/Significance** The paper discusses the diagnostic effectiveness of artificial intelligence (AI) auxiliary diagnosis technology applied to lymphatic diseases, expounds the solution of model fine-tuning training under the background of scarce samples, points out the difficulties and shortcomings of applied research, and puts forward prospects. **Method/Process** Based on the unstructured text data of patients' electronic medical records (EMR), the ERNIE 3.0 classification model is used to conduct a study on the application of lymphedema auxiliary diagnosis. Through two levels of classification tasks, intelligent diagnosis of lymphedema disease and distinguish between primary lymphedema and secondary lymphedema are realized. **Result/Conclusion** The ERNIE 3.0 classification model shows accuracy, precision, recall, and mean *F1* values over 0.95 in the discriminatory task of lymphedema. The mean Macro *F1* value of the model reaches 0.901 in the differentiation of primary and secondary lymphedema. The AUC of the model reaches 0.97 and 0.865 in the two tasks, respectively, indicating the accurate classification effect of the model. In addition, the model classification results have good interpretability, and fill the application gap of intelligent methods in the field of lymphatic diseases.

〔修回日期〕 2023-05-22

〔作者简介〕 汤昊宸, 中级工程师, 发表论文 4 篇; 通信作者: 徐毅, 沈文彬。

〔基金项目〕 科技创新 2030——“新一代人工智能”重大项目 (项目编号: 2020AAA0105003); 科技创新 2030——“新一代人工智能”重大项目 (项目编号: 2020AAA0105005); 北京市科学技术委员会项目 (项目编号: Z191100007619049)。

[**Keywords**] primary lymphedema; secondary lymphedema; auxiliary diagnosis

1 引言

淋巴水肿由淋巴系统的发育性障碍（原发性淋巴水肿）或获得性障碍（继发性淋巴水肿）导致，局部体液滞留和组织肿胀，是严重影响身体健康和生活质量的疾病之一。据国际淋巴学会统计，目前全世界患者数量约 1.7 亿人，中国的患者数量超过 1 000 万人^[1]，尤其是继发性淋巴水肿患者数量正伴随着癌症或肿瘤患者数量快速增长^[2]，每年接受乳腺肿瘤、妇科肿瘤、男性泌尿生殖系统手术治疗的患者出现淋巴水肿的比例高达 10% ~ 60%，通常在术后几个月甚至数年后才出现，开始时没有明显症状，随着病情逐步加重，多数患者将面临肢体残疾的结局。目前对淋巴水肿这类淋巴回流障碍性疾病的相关研究还不多见，对淋巴水肿的病因、发病机制及病理学方面的研究还缺少系统性的认识。因此，可开展系统性诊断治疗淋巴水肿的权威医疗机构较少，专科诊疗可及性较差。而及时准确的诊断是阻断淋巴水肿恶化、提升术后康复痊愈率的关键。辅助诊断首先要判定是否属于淋巴水肿疾病，继而判别属于原发性还是继发性分类。临床医疗中主要通过患者主诉、体格检查初步诊断，并结合放射性核素淋巴显像、电子计算机断层扫描（computed tomography, CT）、磁共振成像（magnetic resonance imaging, MRI）和诊断性穿刺等方法鉴别诊断。然而，一方面由于医疗资源不足，医生须要在治疗方面投入大量精力，例如淋巴疾病相关的手术治疗，无法支撑大量患者（或潜在患者）持续、动态的检查诊断；另一方面，淋巴水肿的病因复杂多样，如果没有丰富的临床经验，容易产生主/客观不一致的诊断结果，造成误诊或漏诊。

近年来人工智能技术取得快速发展，在大数据基础上，深度学习算法可以有效提取数据的表征，挖掘数据和任务之间的关联关系^[3]。基于文本数据^[4]、图像数据^[5]在临床疾病辅助诊断领域开展的

广泛研究，在医疗资源不足的背景下，显著提升医生的工作效率。基于假设演绎法的临床决策理论^[6]指导医生依据患者主诉和电子病历中的数据诊疗决策，是一个专业医疗知识增强的文本自然语义理解过程，为基于电子病历文本数据的辅助诊断智能方法研究提供了理论依据。但是，人工智能辅助诊断技术在淋巴水肿领域应用研究比较少，并且医院内部患者病例数不足以支撑人工智能模型的学习，是典型的少样本学习任务。

目前针对文本数据自然语言处理技术，依据大数据和大模型的学习范式，双向编码器表征（bidirectional encoder representation from transformers, BERT）^[7]（大模型有 4 810 亿个参数）和生成预训练转换器（generative pre-trained transformer, GPT）^[8]（1 750 亿个参数），在大量语料库基础上进行模型的训练学习，形成强大的先验自然语言理解能力，结合少量的专业领域知识数据对模型微调，即可完成快速有效的领域适应学习。借助上述预训练-微调学习范式，本文基于脱敏的患者电子病历主诉文本数据，以及百度公司研发的中文大模型 ERNIE（enhanced representation through knowledge integration）3.0，研发淋巴水肿智能诊断模型，实现是否为淋巴水肿疾病的诊断，以及原发性和继发性的分类判断。实验结果的准确性、可解释性等验证了模型的可行性。

2 模型介绍

2.1 ERNIE 3.0

ERNIE 3.0 是百度公司研发的知识增强语义表示模型，见图 1。首先，引入知识图谱，通过对词、实体等语义单元的掩码，学习完整概念的语义表示；然后，基于序列化建模 Transformer-XL 方法，构建通用语义层和任务语义层，对自然语言生成、自然语言理解和关系抽取等展开多任务学习训练，不同任务适配更合适的网络结构，显著提高模型在

相应任务上的表现，并且在微调阶段，可以固定通用语义表示模块，微调任务语义表示模块的参数，提高训练效率，实现少样本包含领域知识任务的快速学习，目前在语言推断、命名实体识别、情感分析、问答匹配等各类中文任务上取得最优效果。

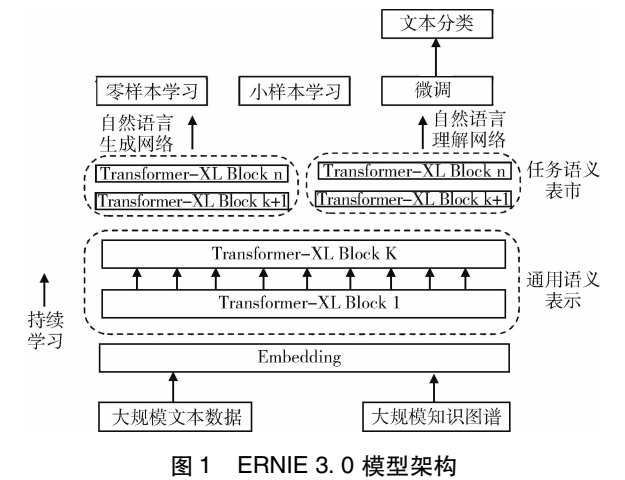


图 1 ERNIE 3.0 模型架构

序列化建模的核心模块 Transformer - XL 衍生于 Transformer 模型，后者主要依赖于自注意力机制，计算任意序列位置输入信息之间的关系，注意力值计算方式如下：

Attention (Q, K, V) = softmax (QK^T / sqrt(d_k)) V (1)

利用多头注意力机制，建模序列整体输入信息编码，多头注意力值计算方式如下：

head_i = Attention (QW_i^Q, KW_i^K, VW_i^V) (2)

MultiHead (Q, K, V) = Concat (head_1, . . . , head_h) W^o (3)

其中，映射参数矩阵分别为 $W_i^Q \in R^{d_{model} * d_k}$ ， $W_i^K \in R^{d_{model} * d_k}$ ， $W_i^V \in R^{d_{model} * d_v}$ ， $W^o \in R^{hd_v * d_{model}}$ ， Q 是目标词矩阵， K 是关键词矩阵， V 是原始特征。与 Transformer 相比，Transformer - XL 中输入序列和位置编码不再共享权值，并且引入相对位置编码和新的可学习参数，解决算法无法对超过固定长度的依赖关系建模的限制，使模型天然具备长文本处理能力，适用于本研究的电子病历长文本数据。

2.2 模型推断

在已有的 ERNIE 3.0 大模型基础上，根据少量的患者数据，学习两个层级任务模型，即首先判断

是否为淋巴水肿，然后判别原发性和继发性疾病，本质上均为分类模型，模型的推断过程，见图 2。

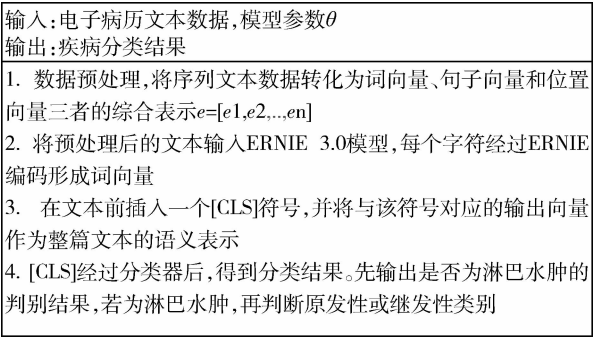


图 2 面向电子病历文本数据的淋巴水肿智能辅助诊断过程

3 实验设置

3.1 数据介绍

实验数据来自医院脱敏数据，具体数据统计描述，见表 1。淋巴水肿电子病历文本数据 300 例，非淋巴水肿（乳糜病）电子病历文本数据 475 例。淋巴水肿患者女性居多，淋巴水肿和非淋巴水肿平均主诉文本长度接近。与此同时，原发性淋巴水肿 49 例，继发性淋巴水肿 251 例，样本出现不平衡，因此在训练原发性或继发性的淋巴水肿判别器时，采用合成少数类过采样技术（synthetic minority over-sampling，SMOTE）对少样本数据过采样，以达到训练时的样本均衡。

表 1 淋巴水肿和非淋巴水肿病例统计

统计项	淋巴水肿	非淋巴水肿
总例数（例）	300	475
原发性（例）	59	0
继发性（例）	241	0
男性（例）	49	230
女性（例）	251	245
最大年龄（岁）	80	77
最小年龄（岁）	2	1
平均年龄（岁）	52.79	30.71
平均单词数（个）	1 754	1 846
最大单词数（个）	1 985	1 957
最小单词数（个）	1 352	1 279

3.2 实验参数设置

实验中模型的超参数包括完整数据集训练次数 (epoch)、学习率 (learning rate)、文本最大长度 (max_len)、批量大小 (batch_size)。由于显存限制, max_len 最大长度设置为 2 048, 文本超出部分将截断。本文数据集经过数据处理 (去除患者基本信息、诊断结果信息) 后, 文本长度均小于 2 048, 因此无需截断, 保证数据的完整性。同样由于显存限制, batch_size 设置为 4。学习率一般为 $e-5$ 级别, 从常用的 $2e-5$ 、 $3e-5$ 和 $5e-5$ 中利用网格搜索法确定, 实验结果表明学习率为 $3e-5$, epoch 为 5 时模型效果最优。

3.3 模型微调过程

预训练-微调的学习范式要求固定预训练模型的通用语义表示部分、微调任务语义表示部分, 根据少量的样本数据即可实现最终模型的快速学习。本文将 ERNIE 3.0 分类任务中的任务语义网络模块微调, 实现两个学习模型的自适应学习过程, 见图 3。

输入:模型初始化参数 θ ,学习率 λ
输出:更新后的参数
1. 数据预处理,将序列文本数据转化为词向量、句子向量和位置向量三者的综合表示 $e=[e1,e2,...,en]$
2. 将预处理后的文本输入ERNIE 3.0模型,每个字符经过ERNIE编码形成词向量
3. 在文本前插入一个[CLS]符号,并将与该符号对应的输出向量作为整篇文本的语义表示
4. [CLS]经过分类器后,得到分类结果
5. 计算损失函数 $L_e(\theta=-y\log p(y x;\theta), p(y x)$ 代表预测标签为 y 的概率
6. $\theta=\theta-\lambda\nabla L_e(\theta)$

图3 面向电子病历文本数据的淋巴水肿智能辅助诊断模型微调训练

3.4 模型评估指标

实验中使用的评价指标如下。其中, 真正 (true positive, TP) 表示实际为正样本且预测为正样本的个数, 假正 (false positive, FP) 表示实际为

负样本但预测为正样本的个数, 真负 (true negative, TN) 表示实际为负样本且预测为负样本的个数, 假负 (false negative, FN) 表示实际为正样本但预测为负样本的个数, 准确率 (accuracy) 表示全部预测结果中正确预测结果的占比, 精确率 (precision) 表示全部正样本的预测结果中正确预测的占比, 召回率 (recall) 表示全部正样本中正确预测的占比, $F1$ 值是精确率和召回率的调和平均值, M 代表正样本数, N 代表负样本数, $rank_i$ 表示预测为正样本的得分, 曲线下面积 (area under curve, AUC) 表示受试者工作特征曲线 (receiver operating characteristic curve, ROC) 下方的面积, AUC 越接近 1, 表明方法真实性越高、分类能力越好。

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

(4)

$$precision = \frac{TP}{TP + FP}$$

(5)

$$recall = \frac{TP}{TP + FN}$$

(6)

$$F1 = \frac{2TP}{2TP + FP + FN}$$

(7)

$$MacroF1 = \frac{1}{n} \sum_{i=1}^n F1^i$$

(8)

$$AUC = \frac{\sum_{i \in positiveClass} rank_i - \frac{M(M+1)}{2}}{M \times N}$$

(9)

4 实验结果分析

4.1 准确性分析

鉴于数据集样本量较少, 采用 5 折交叉验证方法对模型进行训练和评估, 实验结果, 见表 2, 在淋巴水肿的判别任务中, 准确率、精确率、召回率等均超过 0.95, 模型的平均 $F1$ 值达到 0.953, 方差为 0.000 3, 表明模型诊断分类的精准性。在原发性和继发性淋巴水肿的区分方面, 模型的平均 Macro $F1$ 值达到 0.901, 方差为 0.000 2。为了进一步判断分类器在不同阈值下的分类能力, 绘制 ROC 曲线, 见图 4, 在两个任务中模型的 AUC 分别达到 0.97 和 0.865, 表明模型的准确分类效果。

表 2 实验结果

评估指标	判别是否为淋巴水肿				判别淋巴水肿原发性或继发性			
	准确率	精确率	召回率	F1	准确率	精确率	召回率	Macro F1
均值	0.956	0.955	0.960	0.953	0.944	0.909	0.896	0.901
方差	0.000 0	0.000 1	0.000 1	0.000 3	0.000 2	0.000 5	0.000 3	0.000 2

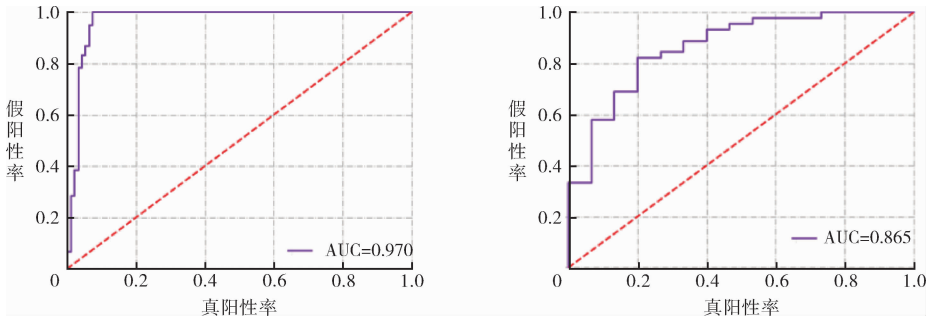


图 4 ROC 曲线

注：左图为判别是否为淋巴水肿，右图为判别原发性/继发性淋巴水肿。

4.2 可解释性分析

人工智能模型应用于临床环境需要严格依赖模型的可解释性。利用积分梯度算法对已训练好的机器学习模型归因分析，定位对分类效果影响比较大的关键词。沿着基线对梯度积分，将重要性建模为平均梯度与输入单元值的逐元素乘积，利用模型的梯度挑选对测试结果产生影响的数据，分析待预测

样本中哪些命名实体对模型预测结果具有重要作用。从 300 例样本数据中对命名实体抽取、统计，汇总 1 700 余个实体词汇，部分举例，见图 5、图 6，用颜色深浅表示命名实体对预测结果的重要性，无关或关联较小词用省略号代替，实验结果表明在关键词分析上，基于 ERNIE 3.0 的微调模型能够捕获疾病诊断相关的关键词。

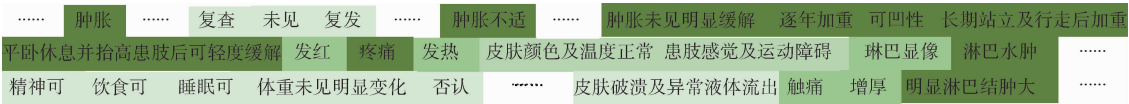


图 5 淋巴水肿判别关键词

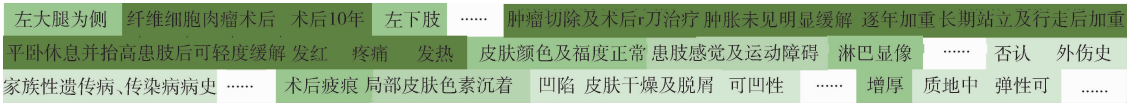


图 6 原发性或继发性判别关键词

通过积分梯度算法筛选的判别关键词，经 19 名淋巴外科临床医师确认，与实际诊断经验基本吻合，一定程度验证了本实验的可解释性。参与该任务的临床医师中主任医师 1 人，副主任医师 3 人，

主治医师 5 人，住院医师 10 人。其中，12 人拥有博士学位，5 人拥有硕士学位，2 人拥有本科学位，保证了关键词判别结果的可靠性。

5 结论与展望

5.1 小结

本文主要开展人工智能方法在淋巴疾病辅助诊断方面的应用研究, 基于少量电子病历中非结构化文本数据, 采用以自然语言理解大模型为基础的预训练-微调模型学习范式, 快速学习淋巴水肿的疾病诊断模型和原发性、继发性分类模型。实验结果从准确性和可解释性方面表明基于非结构化文本进行淋巴水肿辅助诊断的有效性; 真实临床病历数据的构建和预处理为后续研究奠定了基础, 数据和算法均填补了智能化方法在淋巴疾病领域的应用空白。

5.2 局限性

本文所用数据来自非结构电子病历数据中的主诉部分, 包含大量非疾病症状的描述, 如果能进一步识别主诉中的关键信息, 排除噪声干扰, 可以进一步提升模型表现。更深入、细化的可解释性研究需要基于更庞大的临床数据支撑, 以及临床随机对照试验随机分组设计, 因本实验条件有限, 无法进一步深入研究。

5.3 展望

目前原发性和继发性的识别只是满足初步的诊断需求, 而淋巴水肿的疾病类型复杂, 在上肢、下肢、足踝、腹壁、会阴等多处位置产生症状, 且表现为动态变化过程, 因此, 淋巴水肿的精准分期判断是未来亟待解决的难点, 是下一步的研究方向。

世界卫生组织针对淋巴水肿的严重程度分为Ⅶ个等级, 主要依据患者患肢皮肤状态、水肿可自行消退、是否存在细菌性感染及患肢运动功能情况实证判断, 但对患者术后可否康复痊愈、是否会导致残疾等未有预见性借鉴意义。本实验目前停留在对淋巴水肿疾病分类诊断的初级阶段, 未来将逐步

整合患者反复入院记录、患者术后随访等文本数据, 将文本信息的分类要素结合时序模型, 形成淋巴水肿先验诊断分类和术后诊断预测验证相结合的精细化分类标准, 填补淋巴循环医学领域的空白。

随着人工智能技术的发展, 可解释性和个性化辅助诊断也是未来真实场景中的必然需求, 需要在模型研发时融入更多的领域知识, 建模疾病关键信息之间的因果关系; 既要识别电子病历数据中的共性特征, 又要挖掘患者的个性化特征, 实现个体自适应的模型学习和判断。从淋巴智能化诊疗技术的发展远景看, 引入数据规范和模型应用验证的标准是推动技术研发和临床应用转化协同发展的必经之路。

参考文献

- 1 岳云龙. 积极推进淋巴循环障碍性疾病的 MR 研究 [J]. 实用放射学杂志, 2017, 33 (3): 3.
- 2 李滨, 岳云龙, 金延方, 等. MR 体积指数测量对继发性下肢淋巴水肿分期的价值研究 [J]. 实用放射学杂志, 2017, 33 (3): 327-330.
- 3 田雪晴, 汤昊成, 程龙, 等. 辅助诊疗类医学人工智能应用评估体系架构研究 [J]. 医学信息学杂志, 2020, 41 (10): 1-8.
- 4 李文锋, 朱威, 王晓玲, 等. Text2DT: 面向临床诊疗文本的决策规则抽取技术 [J]. 医学信息学杂志, 2022, 43 (12): 16-22.
- 5 李甜, 李晓东, 刘敬禹. 人工智能辅助诊断肺结节的临床价值研究 [J]. 中国全科医学, 2020, 23 (7): 5.
- 6 POST D L. Medical problem solving: an analysis of clinical reasoning [J]. Bulletin of the New York academy of medicine, 1979, 55 (9): 886.
- 7 DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2023-05-19]. <https://arxiv.org/pdf/1810.04805.pdf>.
- 8 BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems, 2020, 33 (7): 1877-1901.