·综　　述·

深度学习技术在 ICD 智能编码中的应用综述*

伍祥林，刘煜民△

（重庆大学附属肿瘤医院病案管理科，重庆 400030）

［摘　要］ 该文探讨了深度学习在智能辅助国际疾病分类（ICD）编码中的应用与进展。针对传统人工编码存在的效率低、易出错等问题，深度学习技术通过构建深层神经网络模型，结合自然语言处理技术，自动从医疗文本中提取关键信息，能够有效提升 ICD 编码的准确性和效率，降低编码成本，为编码员提供决策支持，推动医疗信息化的发展。该文综述了国内外在智能辅助 ICD 编码技术方面的研究成果，深入分析当前面临的挑战，并展望未来研究方向，以期进一步推动智能辅助 ICD 编码技术的发展。

［关键词］ 深度学习；　智能编码；　国际疾病分类编码；　自然语言处理；　医疗信息化；　综述

DOI：10.3969/j.issn.1009-5519.2025.01.041　　**中图法分类号**：R197

文章编号：1009-5519(2025)01-0192-04　　**文献标识码**：A

Review of the application of deep learning technology in ICD intelligent coding*

WU Xianglin, *LIU Yuming*△

(*Department of Medical Record Management*, *Chongqing University Cancer Hospital*, *Chongqing* 400030, *China*)

［**Abstract**］ This paper discussed the application and progress of deep learning in intelligently assisted International classification of diseases code (ICD) coding. Aiming at the problems of low efficiency and error-proneness of traditional manual coding, deep learning technology can automatically extract key information from medical texts by constructing a deep neural network model and combining natural language processing technology. It can effectively improve the accuracy and efficiency of ICD coding, reduce coding costs, provide decision support for coders, and promote the development of medical informatization. This paper summarized the research results of intelligent assisted ICD coding technology at home and abroad, deeply analyzed the current challenges, and looked forward to the future research direction, in order to further promote the development of intelligent assisted ICD coding technology.

［**Key words**］ Deep learning;　Intelligent coding;　International classification of diseases code;　Natural language processing;　Medical informationization;　Review

国际疾病分类（ICD）是由世界卫生组织制定的对各类疾病编码的标准，作为患者住院治疗过程信息转化的关键工具，ICD 编码质量直接影响疾病分类的准确性，是医院进行相关统计、单病种管理的基础和前提[1]。随着疾病相关诊断（DRGs）付费制度改革在各地区的推行，作为 DRGs 正确入组的关键条件，ICD 编码工作受到越来越多医疗机构及医保部门的重视。由于诊疗人数多、工作负荷大、临床诊断书写不规范、编码规则复杂、编码员专业水平有限、临床与编码衔接错层等原因，人工编码易错且低效，迫切需要通过技术手段加以改进。随着信息技术的发展，医疗信息的电子化已经成为常态，关于运用自然语言处理、深度学习等技术完成 ICD 智能辅助编码的问题成为学术界研究的前沿焦点。

1　深度学习在 ICD 编码中的应用特征

1.1　深度学习技术的特征　深度学习[1]是机器学习领域的一个重要分支，通过构建深层神经网络模型来模拟人脑的学习过程，实现对复杂数据的自动分析和处理。深度学习技术在自然语言处理、图像识别等领域取得了显著成果[2-3]。深度学习通过构建深层次的神经网络模型，模拟并优化人脑的学习机制，从而实现对复杂、高维数据的自动分析和处理。深度学习模

*　**基金项目**：重庆市沙坪坝区科卫联合医学科研项目（2023SQKWLH015）。

△　**通信作者**，E-mail：18319881@qq.com。

型能够逐层抽象数据特征，从原始数据中提取出更为高级、抽象的信息表示，增强模型对数据的理解和泛化能力。

1.2　文本处理与图像识别　在自然语言处理（NLP）领域，深度学习技术凭借其强大的语义理解和生成能力，推动了文本分类、情感分析、机器翻译等任务的显著进步[4]。通过构建复杂的神经网络结构[3-4]，如循环神经网络（RNN）、长短时记忆网络（LSTM）、Transformer 等，深度学习模型能够有效地捕捉文本中的上下文信息，理解语言的内在规律和含义，进而实现高精度的文本处理任务。同样，在图像识别领域，深度学习技术也展现出卓越的性能，通过卷积神经网络（CNN）等模型[3]，深度学习能够自动从图像中提取出边缘、纹理、形状等底层特征，并逐步组合成更为复杂的特征表示，最终实现对图像内容的准确识别和理解。

1.3　深度学习在 ICD 编码中的应用特征　在基于深度学习的智能辅助 ICD 编码技术研究中，深度学习技术同样发挥着核心作用。运用深度学习算法和 NLP 技术对病历中检验检查、诊断信息、医嘱、收费项目等医学文本进行识别与处理[5-8]，通过机器深度学习技术从病情描述文本中找出与患者该次就诊时的病情相关度最高的部分提取出来，作为支持该次 ICD 编码的依据，寻求较优的模型架构和模型参数，从而构建深度学习 ICD 智能化编码模型，并设计更加高效、可靠的编码辅助工具。这一过程不仅大大提高了编码准确性和编码效率，还为编码员提供了有力的辅助决策支持，推动了医疗信息化在 ICD 编码领域的发展。

2　国外深度学习在 ICD 编码中的研究与实践概览

2.1　国外智能辅助编码处理方法分类　目前，国外智能辅助编码任务的处理方法研究大致可以分为 3 类。第 1 类是基于规则的方法，即基于专家知识生成一系列基于规则的分类器，这一类方法极大地依赖于医疗专家的人工干预，规则库的维护和扩充需要耗费大量资源，并且很容易出现应用到大规模编码空间时规则覆盖面不够等问题[1]；第 2 类是基于实例的方法，但是由于诊断属于自由文本，“高度重复性”这一假设过于严格；第 3 类是基于学习的方法，抛开了对领域知识的需求，让计算机自动学习完成医疗文本编码的任务，如运用各种传统的机器学习算法或运用 NLP 技术[9]对医疗文本进行语义解析[10]、词性标注、命名实体识别[11]等。

2.2　智能辅助编码技术算法百家争鸣　WANG 等[12]设计了 4 组不同的特征作为朴素贝叶斯分类器的输入，根据模型的分类性能来比较分析不同特征的效用，实验结果表明对原始关键词特征进行词干抽取（stem）可以大幅提高模型的分类表现。KAUR 等[13]采用 k 近邻算法，将医疗记录中字段缺失的 ICD 字段推测问题转化为根据该病例与其他病例在其他未缺失维度上的相似性匹配相应的编码。ZHANG 等[14]提出可以通过关键词索引从 PubMeb 获取相关文章来扩充数据集以解决智能辅助编码任务中的数据不平衡问题。此外，还有研究是运用 NLP 技术计算医疗文本之间的相似度，如 WANG 等[10]通过依存句法分析将医疗文本解析成语法树，并赋予语法树中各节点相应的权重，最终根据树与树之间的匹配度来为测试集中的医疗文本匹配对应的编码。WU 等[15]在心脏衰竭的早期预测运用了 RNN。鲁汶大学应用科学领域的一项研究报告中显示，临床记录 ICD 编码的深度学习方法比较研究中使用 LSTM，通过分析从患者的电子病历中提取的多元时间序列数据来确定诊断[4]。TENG 等[16]则针对非结构化的自由医疗文本采用 LSTM 来确定文本关系类别。

2.3　深度学习在智能编码应用领域增长迅速　将深度学习具体应用于医疗文本智能辅助编码的研究近来也呈现快速增长的趋势：MOONS 等[3]确定了 CNN 应用于医疗文本智能辅助编码时的最优参数；DUARTE 等[17]构建了层次神经网络来处理死亡证明的 ICD-10 编码任务；LI 等[18]提出了 Deeplabeller 的神经网络模型来分别提取文本的局部和全局特征。深度学习的方法能够克服应用基于规则的方法和基于传统机器学习方法时需要准确描述规则或特征的局限性，通过构建模型自动学习特征表达，实现端到端的分类系统。

3　国内深度学习在 ICD 编码领域的探索与发展

国内关于中文医疗文本智能辅助编码的研究晚于国外，中文医疗文本智能辅助编码技术滞后于国外。中文语境下成熟和完备的标准医学术语库、医学语料库几乎处于空缺的状态，这给中文医疗文本的自动处理带来极大挑战。

3.1　基于规则的智能辅助编码技术应用　大多数中文 ICD 智能辅助编码研究还停留在基于规则的方法阶段。王成尧[19]在基于深度学习的病案 ICD 自动编码研究中建立了常用诊断与 ICD-10 编码的对照表；张润彤等[20]增加了疾病的别名字段，然后通过文本的精确匹配进行诊断的编码；肖涵月[21]模拟编码员的编码流程，将繁琐的纸质版 ICD 字典查询工作转换为基于计算机的电子字典查询。

3.2　文本挖掘技术在智能编码中的应用　近几年，部分研究学者在文本挖掘技术领域做出了探索。王天罡等[22]基于分布式语义相似度为医生的中文诊断自动匹配 ICD-10 编码；余颖[23]在面向电子病历文本

的疾病编码自动标注与预测方法研究中用文本建模方法，然后借助文本相关性度量，获取与待编码疾病诊断名称相关的 ICD 编码，该方法在四位亚目码的级别获得 79%的准确率；罗长江[24]基于改进后最长公共子序列(LCS)算法计算诊断的语义相似度，依据自由诊断和标准诊断的相似度来完成诊断到 ICD 编码的映射。

3.3 深度学习在中文 NLP 中的应用　深度学习技术在中文 NLP 领域的突破，为中文医疗文本智能辅助编码带来新机遇。通过 RNN、LSTM 等模型[3-4,25]，深度学习能够捕捉医疗文本中的长期依赖和复杂语义，提升编码准确性。同时，CNN 在提取局部特征方面表现出色，有助于快速筛选关键信息。另外，多任务学习和迁移学习技术的应用，进一步提高模型的泛化能力和效率，解决大部分医疗文本数据稀缺的问题。深度学习技术在中文 NLP 领域的应用[26-27]，为中文医疗文本辅助编码提供强大的技术支撑，推动了智能编码的发展。但是，尽管有研究学者将 NLP 的技术运用到中文智能辅助编码的任务中，但相对于英文智能辅助编码的研究及整个文本分类任务的发展，在中文语境下医疗文本的智能辅助编码仍然还是一个亟待探索的领域。

4　深度学习在 ICD 编码中的效果评估

4.1 提高编码准确性　深度学习技术通过自动学习医疗文本的特征表示和编码规则，能够显著提高 ICD 编码的准确性。有研究表明，基于深度学习的智能辅助 ICD 编码系统在部分数据集上的准确率超过了传统的人工编码方式[28]。这一优势主要得益于深度学习模型能够处理和理解复杂的医疗文本，从中提取出与疾病分类高度相关的关键信息，并据此进行准确地编码。例如，有研究表明，CNN 在医疗文本智能辅助编码任务中表现出色，其准确率显著高于传统方法[3]。

4.2 降低编码成本　智能辅助 ICD 编码技术通过自动化处理医疗文本和自动分配编码，极大降低了编码成本[29]。传统的人工编码方式需要编码员耗费大量时间和精力去阅读、理解和分析医疗文本，再根据复杂的编码规则进行编码。而深度学习技术则能够自动完成这一过程，减少人工干预，提高了编码效率。这不仅降低了医院的人力成本，还缩短了编码周期，使得医疗机构能够更快地获取准确的疾病分类信息，为临床决策和医疗管理提供有力支持。

4.3 提供决策支持　深度学习智能辅助 ICD 编码系统不仅能够提供编码结果，还能为编码员提供决策支持[29-30]。通过可视化展示模型的学习过程和编码依据，编码员可以更加直观地理解模型的决策逻辑，从而更加信任和使用系统。其次，系统还可以根据编码员的反馈进行持续优化和改进，进一步提高编码的准确性和效率。这种交互式的学习方式使得智能辅助 ICD 编码系统更加符合临床实际需求，为医疗机构提供了更加全面和高效的编码解决方案。

4.4 推动医疗信息化发展　深度学习在智能辅助 ICD 编码技术中的应用，不仅提高了编码的准确性和效率，还推动了医疗信息化的发展[31]。随着医疗信息的电子化和智能化程度不断提高，医疗机构对高质量、高效率的编码需求也日益增长。深度学习技术的引入，为医疗机构提供了一种全新的编码方式，使得医疗信息的管理和利用更加便捷和高效。这有助于医疗机构实现数据驱动的决策和管理，提高医疗服务的质量和效率，推动医疗行业的整体发展。

5　深度学习在 ICD 编码应用中面临的挑战

尽管深度学习在智能辅助 ICD 编码技术中取得了显著进展，但仍面临一些挑战。首先，医疗文本的复杂性和多样性给深度学习模型的训练和优化带来较大困难[32]。医疗文本包含大量的专业术语、缩写、模糊表述及不同医生间的语言习惯差异，这些都增加了模型理解和处理的难度。此外，中文语境下医学术语库和语料库的匮乏也限制了深度学习模型的训练效果[32]。现有的中文医学术语库和语料库规模相对较小，且质量参差不齐，难以满足深度学习模型对大规模、高质量数据的需求。再次，如何确保深度学习模型的透明度和可解释性，使编码员能够理解和信任模型的决策逻辑，也是一个需要解决的问题。模型的“黑箱”特性往往让编码员对其决策过程产生疑惑，从而影响模型的实际应用效果。

6　深度学习在 ICD 编码领域的潜力与趋势

深度学习在智能辅助 ICD 编码技术中的应用已取得显著进展，不仅可以提高编码的准确性和效率，还可以降低编码成本，为医疗机构提供重要的决策支持，推动医疗信息化的发展。然而，该领域仍面临诸多挑战，包括医疗文本的复杂性和多样性、数据集的局限性和不平衡性，以及中文语境下的独特难题。未来，随着深度学习技术的不断发展和完善，智能辅助 ICD 编码技术有望在以下几个方面取得突破：(1)通过构建更加复杂和精细的深度学习模型，进一步提高编码的准确性和效率；(2)加强中文医学术语库和语料库的建设，为中文智能辅助编码技术的发展提供有力支持；(3)探索深度学习模型的透明度和可解释性提升方法，增强编码员对模型的信任和使用意愿。总之，随着智慧医疗的深入发展，智能辅助 ICD 编码技术将在医疗信息化领域发挥更加重要的作用，为医疗行业的整体发展注入新的动力。

参考文献

[1] 卢心笛. 中文诊断文本的 ICD 自动编码实证研究[D]. 北京:清华大学,2021.

[2] 耿飙,梁成全,魏炜,等. 基于深度学习的非结构化医学文本知识抽取[J]. 计算机工程与设计,2024,45(1):177－186.

[3] MOONS E, KHANNA A, AKKASI A, et al. A comparison of deep learning methods for ICD coding of clinical records[J]. Appl Sci, 2020, 10(15): 5262.

[4] Science-Applied Sciences, Studies in the Area of Applied Sciences Reported from University of Leuven (KU Leuven): a comparison of deep learning methods for ICD coding of clinical records[J]. Science Letter, 2020: 23-42.

[5] 李博. 深度学习在自然语言处理中的应用[J]. 电子技术,2024,53(4):425-427.

[6] 王晴川. 基于深度学习的中文医学命名实体识别算法研究与标注系统实现[D]. 北京:北京邮电大学,2022.

[7] 姚凌峰. 基于深度学习识别医学文本中的 PICO 成分[D]. 南京:南京邮电大学,2021.

[8] 林玉萍,龙红,李彪,等. 基于医学影像和病历文本的甲状腺多模态语料库构建与应用[J]. 西北大学学报(自然科学版),2021,51(2):198－206.

[9] MELOT B, DROUET F, ALVAREZ C, et al. Automated ICD10-coding of teleconsultations conclusions in primary care[J]. Stud Health Technol Inform, 2023, 309: 135-136.

[10] WANG Z Q, WANG Y Q, ZHANG H Y, et al. ICDXML: enhancing ICD coding with probabilistic label trees and dynamic semantic representations[J]. Sci Rep, 2024, 14(1): 18319.

[11] SHUAI Z, XIAOLIN D, YUN X, et al. Automated ICD coding for coronary heart diseases by a deep learning method[J]. Heliyon, 2023, 9(3): e14037.

[12] WANG Y Q, HAN X, HAO X C, et al. A Curriculum batching strategy for automatic ICD coding with deep multi-label classification models[J]. Healthcare, 2022, 10(12): 2397.

[13] KAUR R, GINIGE J A, OBST O. AI-based ICD coding and classification approaches using discharge summaries: a systematic literature review[J]. Exp Syst Applicat, 2023, 213(PB): 118997.

[14] ZHANG D C, HE D Q, ZHAO S Q, et al. Enhancing automatic ICD-9-cm code assignment for medical texts with pubmed[J]. BioNLP 2017, 2017: 263-271.

[15] WU Y F, ZENG M, FEI Z H, et al. KAICD: A knowledge attention-based deep learning framework for automatic ICD coding[J]. Neurocomputing, 2020, 469(prepublish): 376-383.

[16] TENG F, YANG W, CHEN L et al. Explainable prediction of medical codes with knowledge graphs[J]. Front Bioeng Biotechnol, 2020, 8: 867.

[17] DUARTE F, MARTINS B, PINTO S C, et al. Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text[J]. J Biomed Inform, 2018: 8064-8077.

[18] LI M, FEI Z, ZENG M, et al. Automated icd-9 coding via a deep learning approach[J]. IEEE/ACM Trans Comput Biol Bioinform, 2019, 16(4): 1193-1202.

[19] 王成尧. 基于深度学习的病案 ICD 自动编码研究[D]. 重庆:重庆大学,2021.

[20] 张润彤,陈东华,赵红梅,等. 基于中文语义分析的计算机辅助 ICD-11 编码方法研究[J]. 数据分析与知识发现,2020,4(4):44-55.

[21] 肖涵月. 基于 Seq2Seq 的中文诊断自动 ICD 编码研究与实现[D]. 重庆:重庆大学,2021.

[22] 王天罡,李晓亮,张晓滨,等. 基于预训练表征模型的自动 ICD 编码[J]. 中国数字医学,2020,15(7):53-56.

[23] 余颖. 面向电子病历文本的疾病编码自动标注与预测方法[D]. 长沙:中南大学,2022.

[24] 罗长江. 基于强化学习的 ICD 自动合并编码研究与实现[D]. 重庆:重庆大学,2022.

[25] 闫婧,赵迪,孟佳娜,等. 基于数据增强和扩张卷积的 ICD 编码分类[J]. 计算机应用研究,2024,41(11):3329-3336.

[26] 宋凡,杨鑫,王毅,等. 基于双向记忆传导的 ICD 自动编码方法[J]. 中国卫生信息管理杂志,2023,20(6):977-984.

[27] 唐灵逸,郑涛,邵维君. 基于深度学习与自然语言处理的手术室智慧化管理创新研究[J]. 中国数字医学,2023,18(8):12-17.

[28] 张晓娜,杨卫林. 基于大数据技术的病案智能编码系统的功能设计与应用探究[J]. 科学技术创新,2021(3):82-83.

[29] 李强,尤心心,周佳雯,等. 基于人工智能的病案首页智能编码技术研究与应用[J]. 中国数字医学,2022,17(10):59-63.

[30] 张艺,滕飞,胡节. 基于多任务学习的国际疾病分类自动编码模型[J]. 广西科学,2023,30(1):114-120.

[31] 王阳阳,郑西川. 基于自注意力机制的 CNN-LSTM 模型在 ICD 智能编码系统中的应用研究[J]. 中国数字医学,2020,15(11):20-24.

[32] 张述睿,张伯政,张福鑫,等. 面向 ICD 疾病分类的深度学习方法研究[J]. 计算机工程与应用,2021,57(18):172-180.

(收稿日期:2024-09-12　修回日期:2024-10-28)