第 1 章　远程会诊智能分诊研究

1.1　研 究 背 景

健康作为人民幸福生活的根本基石，以及国家繁荣发展的关键战略支撑，其重要性不言而喻。党的第二十次全国代表大会高瞻远瞩，明确提出“推进健康中国建设”，将医疗卫生事业的发展提升至国家战略的核心层面，彰显了对人民健康的高度重视与坚定承诺。在 2023 年全国两会的重要议事日程中，“智慧医疗+”体系建设成为代表委员们热烈讨论的焦点话题。这一体系的核心使命在于借助先进的信息技术，打破长期以来制约医疗发展的资源壁垒，致力于实现优质医疗服务的广泛普及，让全体人民都能从中受益。

然而，不得不正视的是，我国医疗资源领域长期深受“总量不足、配置失衡”难题的困扰[1]。依据《中国卫生健康统计年鉴（2022）》的权威数据，全国范围内的三级医院在数量上仅占医院总数的 12.6%，却承担了高达 35.2%的诊疗量。这一数据清晰地反映出优质医疗资源过度集中于直辖市、省会城市及经济高度发达地区的严峻现实。与之形成鲜明对比的是，占全国面积 70%以上的偏远地区、山区和农村地区，医疗资源匮乏的问题极为突出。在这些地区，基层医疗机构，如乡镇卫生院、社区卫生服务中心，受限于硬件设施、专业人才等因素，诊疗能力极为有限，面对稍微复杂的病症就显得力不从心，难以提供有效的医疗服务。

在这样的背景下，偏远地区的患者一旦遭遇疑难杂症，就不得不踏上漫长而艰辛的就医之路，长途跋涉前往大城市的医院寻求救治。国家卫生健康委员会的专项调研数据显示，中西部农村地区的患者平均单程就医距离超过 150 公里。这不仅意味着患者需要在路途上耗费大量的时间，还带来了沉重的经济负担。往返的交通费用、异地的住宿费用及因误工造成的收入损失，人均超过 3000 元。对于部分重症患者家庭而言，年均就医支出甚至占到家庭年收入的 40%以上。这种“看病难、看病贵”的困境，如同沉重的枷锁，不仅极大地加剧了患者家庭的经济压力，更严重的是，可能导致患者因延误最佳治疗时机，使得病情恶化，给患者的生命健康带来极大的威胁。

远程医疗，作为信息技术与医疗服务深度融合的创新性产物，为破解这一困局带来了曙光。它依托先进的音视频通信技术、高效的数据传输技术等，成功搭建起跨地域诊疗协作的桥梁，为优质医疗资源下沉到基层、偏远地区提供了具有突破性意义的解决方案[2]。在 2020 年新型冠状病毒感染疫情肆虐的特殊时期，远程医疗的巨大价值和关键作用得到了充分彰显。当时，全国远程医疗服务量呈现出爆发式增长，同比增长高达 400%。武汉方舱医院通过远程会诊系统，与全国 500 余家医院实现了实时、高效的协作。这一举措在

很大程度上有效地缓解了当地医疗资源极度紧张的压力，为疫情防控和患者救治工作发挥了不可替代的重要作用[3]，也让人们深刻认识到远程医疗在应对突发公共卫生事件及优化医疗资源配置方面的巨大潜力。

近年来，在国家政策的大力推动和技术持续进步的双重利好因素作用下，我国远程医疗平台建设取得了令人瞩目的显著进展。截至 2023 年底，全国已有 83%的三甲医院成功建成远程医疗平台，这些平台广泛覆盖了超过 2 万家基层医疗机构。然而，随着远程医疗服务的深入开展，远程会诊流程中的“分诊瓶颈”问题逐渐浮出水面，成为制约远程医疗进一步发展的关键障碍。通过对国家远程医疗中心进行深入的实地调研，发现现行的人工分诊模式存在三大核心问题，严重影响了远程医疗服务的效率和质量。

1.1.1 远程医疗中的人工分诊困境

1. 基层医生操作成本高

在提交会诊申请这一基础环节，基层医生面临着烦琐且耗时的操作流程。他们需要从上级医院复杂的科室导航栏中，手动筛选并选择拟会诊的科室。上级医院的科室设置往往极为复杂，以某三甲医院为例，其涵盖了多达 62 个临床科室及亚专科。如此繁杂的科室体系，使得基层医生在选择科室时困难重重。据实际观察和统计，基层医生平均需要耗费 5~8 分钟来仔细浏览科室列表，以确定合适的会诊科室。而在一些偏远地区，由于医生对上级医院科室职能缺乏足够的了解和熟悉，这一操作时间更是大幅延长，部分医生甚至需要花费 15 分钟以上。大量的时间耗费在科室选择上，严重影响了会诊申请的整体效率，使得患者的就医等待时间无形延长，也在一定程度上降低了远程医疗服务的及时性和便捷性。

2. 分诊准确性不足

不同医院之间，科室设置存在显著差异。基层医院的科室划分相对简单，多为内科、外科等大的类别；而一些大型医院则会根据疾病种类、治疗手段等因素，将科室进行细致的细分，如将“心血管内科”进一步细分为“冠心病科”“心律失常科”等。这种科室设置的不一致性，给基层医生准确选择会诊科室带来了很大的困难。相关数据显示，高达 34.7%的基层医生在选择会诊科室时，无法精准地挑选出最合适的科室。人工选择的科室与最终专家会诊科室不一致的比例达到 18.3%，导致大量的会诊申请需要远程医疗中心的工作人员进行二次协调。二次协调不仅增加了工作流程的复杂性和冗余度，还容易导致信息传递的偏差和延误，进一步影响远程会诊的效率和质量。

3. 中心工作人员负荷过重

国家远程医疗中心是国家卫生健康委批准的依托郑州大学第一附属医院设立的我国唯一的国家级远程医疗中心，负责全国远程医疗技术发展研究、行业应用检测、技术标准和临床规范制定、行业交流等工作，也是中国卫生信息学会远程医疗专委会依托机构。

2021 年国家远程医疗中心的数据揭示了工作人员面临的巨大工作压力。当年，总计有 11720 条会诊申请需要处理，而承担这一重任的仅有 12 名工作人员，每人每日平均需要处理 40.4 条申请。由于会诊申请内容涉及 62 个不同的科室，涵盖各种各样的疾病信息和患者情况，且专家的日程安排极为紧张，平均每位专家每周可承接的远程会诊次数不超过 8 次。工作人员需要在复杂的申请内容与紧张的专家日程之间进行艰难的匹配，这无疑需要耗费大量的时间和精力[4]。据统计，完成一次会诊申请的协调工作，工作人员需要耗费大量时间，导致会诊响应时间平均长达 2.3 天。对于一些紧急病例而言，如此漫长的协调延迟可能会使患者错过最佳的治疗时机，严重影响诊疗效果。

人工分诊模式所暴露出的这些弊端，犹如一道道紧箍咒，严重制约了远程医疗平台服务能力的提升[5]，也与当下“智慧医疗+”所倡导的高效、精准的发展目标背道而驰。经科学测算，如果能够将分诊效率提升 50%，远程医疗平台的日处理能力将从当前的 120 条申请提升至 180 条，同时，会诊响应时间也能够缩短至 1 天以内。这一数据充分显示出提升分诊效率对于远程医疗发展的巨大潜力和迫切需求。基于此，本研究聚焦于远程会诊智能分诊问题，期望借助人工智能技术的强大力量，实现科室的精准推荐，从而有效地降低基层医生的使用成本，显著提升远程医疗服务的整体效率。

1.1.2　远程会诊申请文本特征及挑战

值得注意的是，远程会诊申请文本独具特征，这也为智能分诊带来了特殊的挑战。基层医生在撰写申请时，受限于时间、经验及对患者病情的初步判断，主要侧重于描述患者的基本信息，如年龄、性别等，以及初步诊断结果，例如“胃腺癌化疗后”“支气管肺炎”等。然而，对于诊疗过程中关键的疑问，如“是否需要手术”“用药调整建议”等内容，却较少在申请文本中提及。这种文本特征使得在直接运用传统的文本分类算法进行分诊时，准确度较低。在本研究前期所进行的预实验中，采用传统文本分类算法的分诊准确度仅为 68.2%，这一结果显然无法满足实际应用中对分诊准确性的要求

1.1.3　BTB 智能分诊模型的提出

为了有效地应对上述挑战，本研究创新性地提出了融合患者信息与初步诊断文本的 BTB（BERT-TextCNN-双层 BiGRU）智能分诊模型。该模型旨在通过多维度的特征提取方式，充分挖掘申请文本中的有效信息，从而提升科室推荐的精准性。

从理论价值层面来看，本研究极大地拓展了深度学习在医疗文本分类领域的应用场景。首次将 BERT、TextCNN 与双层 BiGRU 三种先进的技术相结合，并应用于远程会诊分诊这一复杂的医疗场景中，为处理复杂医疗文本的特征提取工作提供了全新的方法和思路，丰富了医疗人工智能领域的理论研究成果。

从实践意义角度出发，BTB 智能分诊模型有望为远程医疗平台提供高效、智能化的分诊工具。预计在实际应用中，该模型能够将分诊准确度提升至 90%以上，同时减少基层医生操作时间 60%以上。这将显著降低基层医生的工作负担，提高会诊申请的提交效率。此

外，准确的分诊结果还能够减少远程医疗中心工作人员的协调工作量，优化工作流程，提升远程医疗服务的整体质量和效率，为推进医疗信息化与数智化转型提供强有力的技术支撑，为实现优质医疗资源的高效配置和远程医疗服务的可持续发展奠定坚实基础。

1.2 文 献 综 述

1.2.1 文本分类

文本分类指的是利用计算机将输入的文本信息自动划分到预定义的某个类别中。在线医疗智能分诊系统允许患者输入疾病问诊文本，并通过智能算法判断患者应挂号的科室，本质属于在线医疗领域的文本分类任务。

文本分类技术作为自然语言处理（NLP）的关键技术之一，被广泛应用于在线医疗领域，如科室推荐和疾病预测等。早期的文本分类方法大多基于人工制定规则，准确度和效率都较低。在 19 世纪 60 年代，学者们通常通过计算文本归属某类别的概率来进行分类。20 世纪 70 年代，Salton 等[6]提出向量空间模型，为文本分配不同权重并进行加权计算，这一方法为后续模型发展奠定了基础。进入 20 世纪 90 年代后，文本分类技术逐步采用 TF-IDF、词袋模型和支持向量机等统计与机器学习算法，使分类的准确性和效率得到显著提升。然而这些方法只能捕捉到浅层次的语义特征，难以捕捉文本中的深层语义关系，也无法处理上下文依赖问题，导致模型仍然难以实现由输入到输出变量的复杂映射。

近年来，数字化的发展使社会各领域积累了大量文本数据，而人工智能和大数据技术的进步推动了深度学习在文本分类任务中的应用。作为机器学习的延伸，深度学习在应对海量数据、学习深层语义关系和减少人工特征工程依赖方面展现出明显优势。随着神经网络算法的不断改进创新，CNN 和 RNN 逐步被引入文本分类任务。Kim 等[7]提出 TextCNN 模型，利用不同大小的卷积核提取文本不同层次的局部特征。随后，该方法在在线医疗文本分类领域得到进一步应用。例如，郑承宇等[8]提出基于 ALBERT 和 TextCNN 的医疗文本分类模型，通过结合 ALBERT 的上下文语义理解能力和 TextCNN 的局部特征提取能力，显著提升了分类的准确性。杨杰等[9]提出了融合多级语义信息的文本分类模型，并通过注意力池化技术和胶囊网络提取和强调文本中的关键语义，最后通过对抗训练保持其分类的稳定性和准确性。在提升模型性能方面，部分研究者探索了知识图谱和多模态数据的融合策略。He 等[10]和 Li 等[11]提出了结合医学知识图谱的深度学习模型，以改善分类效果。与此同时，迁移学习和预训练模型的结合也成为提升文本分类性能的重要方向。Benzorgat 等[12]提出了一种由 DenseNet201、GoogleNet（InceptionV3）和 InceptionResNetV2 组成的混合模型，该模型在多个公开数据集上均取得了优异结果，表明模型的多层特征融合能够有效地增强分类性能。尽管深度学习推动了在线医疗文本分类的进步，但仍面临诸多挑战，如高质量大规模数据集的缺乏、医疗文本中大量专业术语及口语化表达带来的理解难度等[13-15]。

针对这些问题，国内外学者提出了一系列优化策略。例如，Shorten 等[16]利用数据增

强技术扩充数据集规模，显著改善了模型的泛化能力，Wang 等[17]通过知识图谱增强模型对医疗术语的理解能力，提升了分类的准确性。此外，为了应对在线医疗场景中的计算资源限制，提高模型的计算效率，一些研究者开发了轻量化模型。例如，2020 年由华为和华中科技大学提出的 TinyBERT[18]，在大幅减少模型参数量的同时，保持了较高的分类性能，使其适用于资源受限的在线医疗环境。国内的研究主要围绕深度学习与医疗数据的结合进行优化。例如，臧志栋等[19]将提示学习与深度学习模型融合，提升了在线医疗问答社区的短文本分类准确度。王若佳等[20]基于从春雨医生在线医疗网站爬取的数据，验证了支持向量机、随机森林、集成分类等机器学习分类模型在分诊领域的有效性。白思萌等[21]使用 BioBERT 模型和交叉注意力机制将语义信息进行特征融合，提高了医学文本分类效果。赵楠等[22]引入多任务学习框架，结合领域自适应和元学习策略，提高了小样本文本分类的性能表现。

国外学者更多关注文本数据与其他模态数据，如图像、语音等结合的应用，例如，Chaddad 等[23]提出了一种可解释的 AI 模型，成功将医学图像和文本结合用于分类任务，并揭示了深度学习黑盒模型的部分内在机制。Geo 等[24]提出了场景分类模型与深度语义匹配模型，将其应用于医院就诊全流程中的机器人辅助诊疗系统。Sakib 等[25]设计了一种系统，利用卷积神经网络处理处方图像，并结合 BERT 模型对文本进行分类，实验结果表明该系统在病史构建任务中表现优异。

1.2.2　远程会诊平台智能分诊研究

传统的分诊是指分诊护士根据患者的症状及体征，判断患者病情的严重程度及隶属专科，并合理安排其就诊的过程[26]。近年来，随着计算机技术的发展，智能分诊技术得到了极大发展，是提高远程医疗平台服务能力和会诊效率的有效手段。现有的文献研究将智能分诊主要分为两类：病情分诊和学科分诊。在病情分诊中，研究者通常使用医疗传感器、CT 影片等数据，通过仿真对比实验的方法，验证智能分诊系统的性能。Hamid 等[27]利用可穿戴传感器数据为使用远程医疗的冠心病患者开发了一种智能分诊系统。Tschandl 等[28]开发和测试了基于人工智能技术的皮肤癌识别系统，获得的结果表明该模型在辅助专家诊断方面具有潜力。Xie 等[29]通过实验模拟得出人工智能与眼科专家在评估远程医疗患者糖尿病视网膜是否病变时没有显著性差异。Saiteja 等[30]为监测使用远程医疗系统的慢性病患者的健康状态，构建了基于机器学习的分诊框架，根据糖尿病和高血压状态将患者分为危急和正常两类，结果表明该框架的提出和开发可促进远程患者健康监测。在学科分诊中，研究者通常依托于某个远程医疗平台，阐述该平台实现智能分诊的方式。董天舒等[4]提出使用远程会诊申请的关键字段对专家科室和专家特长进行模糊匹配后匹配专家，再根据会诊科室的忙闲状态匹配会诊科室，最后结合专家交互后建立会诊。史嘉兴等[31]依据公众发热、症状、旅居史进行判断，智能将使用远程医疗的患者分到发热门诊和普通门诊。

梳理文献发现，有较多学者使用深度学习算法对互联网医疗智能分诊及专家推荐进行了研究[32-33]，但是针对远程医疗情形关注较少[34]。有部分学者研究了机器学习方法辅助远程会诊分诊，但是研究对象主要集中于病情分诊，鲜少有学者关注学科分诊，且现有的研究不能解决调研中发现的实际问题。现阶段一些远程医疗平台仍依靠人工手动分诊。近年

来，随着参与远程会诊的基层医生与患者的数量逐年增长，人工分诊乃至传统的辅助分诊已不能满足使用者的需求。因此，探索深度学习算法在远程会诊智能学科分诊中的应用，是远程医疗平台智能化发展的重要方向。

1.2.3 基于深度学习模型的分类算法研究

深度学习模型能够捕获到文本更深层次的语义信息，在自然语言处理中取得了巨大的成功。许多学者致力于使用深度学习模型进行文本分类研究，最具有代表性的是卷积神经网络（CNN）和递归神经网络（RNN）及它们的变体。Bello 等[35]使用 CNN、RNN 和 BiLSTM 模型对推特的文本进行情感分类，研究发现 BERT-CNN、BERT-RNN、BERT-BiLSTM 的组合在准确度、精准率、召回率和 F1-score 方面均优于与 Word2vec 和基于预训练词向量的组合。李文亮等[36]提出一种多特征融合的神经网络模型，有效地提升了文本情感分类效果。Zhou 等[37]为提高短文本分类的准确度，开发了一种基于语意扩展的文本卷积神经网络。还有学者[38]尝试将 CNN、RNN 及它们的变体融合使用，进一步捕获文本的表征信息，从而优化现有的分类模型。杨文涛等[39]结合两种传统神经网络的优势，提出了一种长文本分类模型，实验结果表明，混合模型能够提高文本分类的准确度。Jiang 等[40]考虑文本特征的表达方式，使用预训练模型 BERT 得到网民评论的特征，并将获取到的特征输入 TextCNN-BiGRU 模型中获得更深层次的语义特征，最终使用 Softmax 函数判断文本的情感。然而，模型的串联结构使得后者模型的性能受限于前者，后者模型的性能并未得到充分发挥。赵宏等[41]则将嵌入层的文本特征分别输入 TextCNN 模型和 BiGRU 模型中，提取文本的局部语义信息和上下文信息，然后进行特征融合判断文本情感，结果表明，模型的性能优于 BiGRU-TextCNN 模型。Zhou 等[42]则认为 BiGRU 模型不能完全度量上下文信息。因此，本章充分考虑 BiGRU 模型的隐藏状态和细胞状态，并利用双层 BiGRU 模型表征文本的上下文信息。

上述研究为远程会诊智能分诊提供了一定的建模思路。本章拟使用 TextCNN 模型和双层 BiGRU 模型分别提取远程会诊申请的特征，然后与 BERT 的特征进行融合，从而进行远程会诊智能分诊。但是考虑到远程会诊的现实情景，直接采用分类算法进行分诊可能会频繁出现会诊安排与基层医生需求不匹配的现象。因此，本章使用深度学习算法为基层医生推荐多个合适的会诊科室，最终，由基层医生选择所需的会诊科室。基于此，本章将 BTB 模型引入远程会诊的分科分诊中，为远程医疗平台智能分诊系统的建设提供理论支撑和实践指导。

1.3 研 究 模 型

1.3.1 BERT 模型背景

在双向编码器表征法（BERT）于 2018 年由谷歌提出之前，语言模型领域存在诸多限制。传统的语言模型，如基于 RNN 及其变体［长短期记忆网络（LSTM）和门控循环单元

（GRU）］的模型，大多采用单向的序列处理方式。它们在处理文本时，只能按照从前往后或从后往前的顺序理解文本，这就导致无法同时捕捉单词在句子中的左右上下文信息。例如，当遇到“他在银行附近等待”这样的句子时，单向模型可能由于缺乏完整上下文，无法准确判断“银行”一词到底是指金融机构还是河岸。而 BERT 创新性地采用了双向 Transformer 架构，能够同时考虑句子中每个单词的前后文信息，极大地丰富了语义理解的深度和广度，使得模型能够更精准地把握语义。

BERT 另一重大贡献是开创了“预训练+微调”的全新范式。在 BERT 之前，NLP 任务往往需要针对每个具体任务收集和标注大量的数据来训练模型，这不仅耗费大量的时间和人力成本，而且模型的泛化能力常常受限。BERT 则先在海量的无监督文本数据上进行预训练，通过这种方式学习通用的语言知识和语义模式。之后，针对特定的 NLP 任务，如情感分析、文本分类、问答系统等，只需在预训练好的 BERT 模型基础上进行微调，利用少量的特定任务标注数据对模型进行优化即可。这种范式大大减少了对大规模标注数据的依赖，降低了任务开发成本，同时显著提升了模型在不同任务上的泛化能力。例如，在情感分析任务中，以往可能需要收集和标注成千上万条带有情感标签的文本数据来训练模型，而现在基于预训练的 BERT 模型，可能只需几百条标注数据进行微调，就能构建出高精度的情感分析模型。

BERT 的出现给 NLP 领域带来了革命性的变化，在多项基准测试中取得了前所未有的成绩。在通用语言理解评价（general language understanding evaluation，GLUE）基准测试中，BERT 刷新了多项任务的最高分，几乎在所有任务上都超越了以往的最佳模型。它的成功激发了学术界和工业界对预训练语言模型的深入研究和广泛应用，促使众多基于 BERT 的变体和改进模型不断涌现。像 RoBERTa 对训练数据和训练方法进行优化，ALBERT 通过参数共享等技术提高模型效率等。这些后续模型进一步推动了 NLP 技术的创新和发展，使得机器对语言的理解和生成能力提升到了新的高度，让 NLP 技术在更多实际场景中得以应用。

BERT 采用了 Transformer 的编码器部分，由多个 Transformer 块堆叠而成。每个 Transformer 块主要包含两个子层：多头注意力子层（multi-head attention sub-layer）和前馈神经网络子层（feed-forward neural network sub-layer）。在多头注意力子层中，输入被分成多个头，每个头分别计算注意力权重，然后将结果拼接起来，这种机制让模型能够从不同角度捕捉输入序列中的语义信息，提高了模型的表达能力。前馈神经网络子层则对多头注意力子层的输出进行进一步处理，它由两个全连接层组成，中间使用 ReLU 激活函数，负责对语义信息进行非线性变换，增强模型的学习能力。通过多个 Transformer 块的堆叠，BERT 能够不断对输入文本进行深度语义编码，学习到丰富的语义表示。BERT 主要有 BERT-BASE 和 BERT-LARGE 两种预训练模型结构。BERT-BASE 包含 12 个 Transformer 块，隐藏层维度为 768，注意力头数为 12 个，总参数数量约为 1.17 亿，在模型大小和性能之间取得了较好的平衡，适用于大多数场景，在计算资源有限时也能快速部署和应用。BERT-LARGE 具有 24 层编码器、16 个注意力头和 340M 参数，在计算资源允许的情况下，能学习更复杂的语义模式，在复杂的问答或文本分类任务中，能够更好地捕捉文本中的细微语

义差异，提供更准确的预测结果，但训练和推理过程更耗时，对硬件设备要求更高。

1.3.2 TextCNN 的演进之路

TextCNN 的发展得益于 CNN 在图像领域的巨大成功。CNN 最初是为处理图像数据而设计的，其核心思想是通过卷积层中的卷积核在图像上滑动，自动提取图像的局部特征。这种局部感知和权值共享的机制使得 CNN 在图像识别、目标检测等任务中表现出色，能够高效地处理大规模的图像数据并取得高精度的结果。研究者们受到 CNN 在图像领域成功的启发，尝试将其应用于文本数据处理。虽然文本数据与图像数据在形式上有很大差异，但从信息处理的角度看，两者有一定的相似性。文本可以看作由单词或字符组成的序列，类似于图像中的像素矩阵，因此可以借鉴 CNN 的局部特征提取思想来处理文本。

当 CNN 被引入文本处理领域时，需要针对文本的特性进行一些改进和调整。在图像中，卷积核通常以固定的大小在二维图像上滑动，而对于文本，由于句子的长度和单词的表示方式不同，需要设计合适的卷积核大小和滑动方式。通常会使用不同大小的卷积核来捕捉不同长度的 n-gram 特征，例如，使用大小为 3、4、5 的卷积核，分别对应于连续 3 个、4 个、5 个单词组成的文本片段。这些不同大小的卷积核能够提取文本中不同粒度的局部特征，小卷积核可以捕捉到单词之间紧密的语义关系，大卷积核则可以获取更广泛的上下文信息。与图像中的卷积操作类似，文本卷积操作也是通过卷积核与文本向量进行卷积运算，得到特征图。然后通过池化操作，如最大池化，从特征图中提取最重要的特征。最大池化在文本处理中的作用是能够保留文本中最显著的特征，而忽略特征在文本中的具体位置信息，这对于处理文本中特征位置不固定的情况非常有效。经过池化后的特征再通过全连接层进行分类或其他任务的预测。

TextCNN 在文本分类任务中得到了广泛的应用，并取得了良好的效果。与传统的文本分类方法相比，TextCNN 能够自动从文本中提取有效的特征，无需人工手动设计特征，大大提高了特征提取的效率和准确性。在新闻分类场景中，TextCNN 可以通过学习大量新闻文本的特征，准确地将新闻归类到不同的主题类别，如政治、经济、体育、娱乐等。在情感分析任务中，TextCNN 能够分析文本中所表达的情感倾向，判断文本是积极、消极还是中性情感。它在短文本分类任务中优势更为明显，由于短文本通常包含的信息较少，传统方法难以提取足够的特征，而 TextCNN 能够通过不同大小的卷积核快速捕捉短文本中的关键特征，实现准确分类。

1.3.3 双层 BiGRU 的发展轨迹

RNN 是一种能够处理序列数据的神经网络模型，它通过隐藏层状态来保存序列中的上下文信息，使得模型能够对序列中的每个元素进行处理时考虑到之前的信息。然而，RNN 在处理长序列时存在梯度消失和梯度爆炸的问题，导致其难以学习到长距离的依赖关系。为了解决这些问题，提出了 GRU 这一概念。GRU 在 RNN 的基础上引入了门控机制，主要包括更新门和重置门。更新门决定了前一时刻的隐藏状态有多少信息需要保留到当前时刻，重置门则控制了当前输入与前一时刻隐藏状态的结合程度。通过这种门控机制，GRU

能够更好地处理长序列数据，选择性地记忆和遗忘信息，从而在一定程度上缓解了梯度消失和梯度爆炸的问题，提高了模型对序列中长距离依赖关系的学习能力。

虽然 GRU 在处理序列数据方面有了一定的改进，但它仍然只能从前向后或者从后向前单向地处理序列信息。而在很多自然语言处理任务中，同时考虑序列中前后两个方向的信息对于准确地理解语义非常重要。BiGRU 就是为了满足这一需求而产生的。BiGRU 由两个 GRU 组成，一个按顺序从前向后处理输入序列，另一个按逆序从后向前处理输入序列。这两个 GRU 的输出再进行拼接或其他操作，使得最终的输出能够同时包含序列中前后两个方向的信息。例如在处理句子“我喜欢苹果，因为它很美味”时，前向的 GRU 在处理“因为它很美味”时，能够结合前面“我喜欢苹果”的信息，而后向的 GRU 在处理“我喜欢苹果”时，能够结合后面“因为它很美味”的原因信息，通过这种双向信息的融合，BiGRU 能够更全面、准确地理解句子的语义。

为了进一步提升 BiGRU 的特征学习能力，研究者们提出了双层 BiGRU 结构。在双层 BiGRU 中，第一层 BiGRU 对输入序列进行初步的特征提取和信息融合，得到一个包含双向上下文信息的特征表示。然后，将这个特征表示作为第二层 BiGRU 的输入，第二层 BiGRU 再次对这些特征进行处理和融合。通过这种双层结构，模型能够对输入序列进行更深入的特征挖掘，学习到更复杂、更抽象的语义特征。在处理长篇文章或复杂文本时，双层 BiGRU 能更好地捕捉文本中不同层次的语义关系，将句子之间、段落之间的信息进行整合，从而在诸如文本分类、情感分析、机器翻译等自然语言处理任务中取得更好的效果。例如，在文本分类任务中，对于一篇包含多个段落、主题较为复杂的文档，双层 BiGRU 能够通过层层处理，准确地提取出文档的关键主题特征，实现更精准的分类。

1.3.4　BERT-TextCNN-双层 BiGRU 模型的融合

在自然语言处理任务中，单一的 BERT、TextCNN 或 BiGRU 模型都存在一定的局限性。虽然 BERT 能够通过双向 Transformer 架构捕捉丰富的上下文语义信息，并且在预训练后对多种任务有较好的适应性，但它在处理一些局部特征明显的文本时，可能无法像 TextCNN 那样快速有效地提取特定的局部模式。例如，在一些需要快速识别文本中特定关键词组合或固定搭配的任务中，TextCNN 的卷积操作能够更直接地定位和提取这些局部特征，而 BERT 可能需要更复杂的处理过程。TextCNN 虽然擅长提取文本的局部特征，但它对文本整体的上下文信息利用不够充分，在处理长文本或需要综合考虑全局语义的任务时，表现可能不如 BERT 或 BiGRU。BiGRU 能够捕捉文本的双向上下文信息，但在面对大规模无监督数据的预训练和复杂语义理解方面，其能力相对 BERT 较弱。在一些需要理解复杂语义关系和进行知识迁移的任务中，BiGRU 可能需要更多的训练数据和更复杂的模型结构才能获得较好的效果。

BERT-TextCNN-双层 BiGRU 模型正是为了克服单一模型的局限性，充分发挥各个模型的优势而设计的。BERT 作为预训练模型，能从大规模无监督文本中学习到丰富的语言知识和语义表示，为后续的模型提供高质量的文本特征输入。TextCNN 则可以在 BERT 输出的特征基础上，进一步挖掘文本中的局部特征，捕捉那些对任务有重要影响的关键信息

片段，与 BERT 的全局语义信息形成互补。双层 BiGRU 通过双向信息捕捉和多层结构，能够对文本进行更深入的上下文分析，进一步融合局部与全局信息，提高模型对文本语义的理解和表达能力。在文本分类任务中，首先 BERT 对文本进行预训练和初步的特征提取，得到包含丰富语义的文本表示；其次 TextCNN 从中提取局部的关键特征，如特定的词汇组合或短语模式；最后双层 BiGRU 对这些特征进行整合和深入分析，考虑文本的前后文关系，从而更准确地判断文本的类别。通过这种优势互补的融合方式，BERT-TextCNN-双层 BiGRU 模型能在多种自然语言处理任务中取得比单一模型更好的性能表现，为自然语言处理的实际应用提供更强大的工具。

1.3.5 研究框架

本章提出了基于深度学习的远程会诊智能分诊模型，依据基层医生会诊申请和上级医院科室的职责，为远程会诊申请推荐合适的会诊科室。本部分技术研究框架如图 1-1 所示。①数据准备：通过国家远程医疗中心获取远程会诊申请数据。②数据预处理：对获取到的远程会诊申请进行预处理。③模型训练与效果评价：将预处理好的数据导入 BTB 模型中进行训练，并通过评价标准，对模型进行评价。④结果分析：通过多个维度对 BTB 模型的结果进行分析，验证 BTB 模型的有效性。

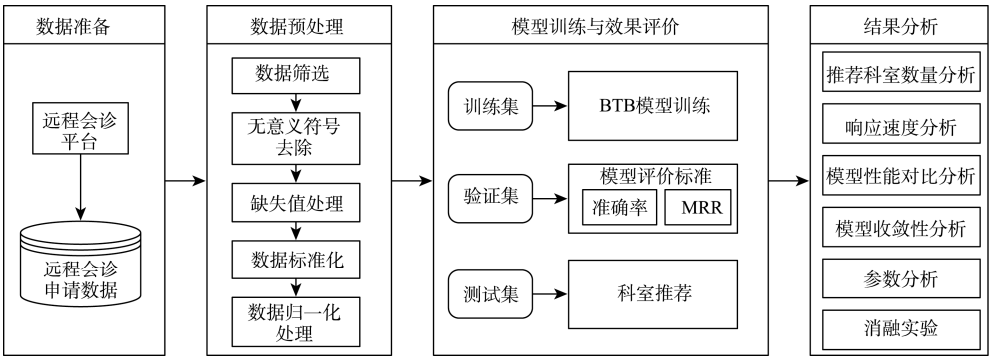

图 1-1 远程会诊智能分诊技术研究框架

1.3.6 BTB 模型

BTB 智能分诊推荐模型的结构图如图 1-2 所示。模型主要包括 BERT 文本特征表示层、特征抽取层、特征融合层和输出层。模型的整体运作流程如下：给定远程会诊申请，经预处理后输入 BERT 模型中进行文本向量化表示，获取融合上下文语义的词嵌入层，取出 token 级特征 $T_{[Token]}$和句子级特征 $T_{[CLS]}$，并分别传递到特征抽取层和特征融合层。在特征抽取层中，一方面，利用不同的卷积核提取文本的局部特征，将经过最大池化层处理的局部特征进行融合，形成特征 C；另一方面，利用 BiGRU 模型进一步获得远程会诊申请的上下文特征，得到所有 BiGRU 单元的输出后，再将其输入 BiGRU 模型中，提取最后单元的输出 H。在特征融合层，将远程会诊申请的句子级 BERT 特征 $T_{[CLS]}$、TextCNN 特征 C

和双层 BiGRU 特征 H 进行融合。最后，将融合后的特征送入 Sigmoid 分类器中，为基层医生推荐合适的会诊科室，从而达到智能分诊的目的。

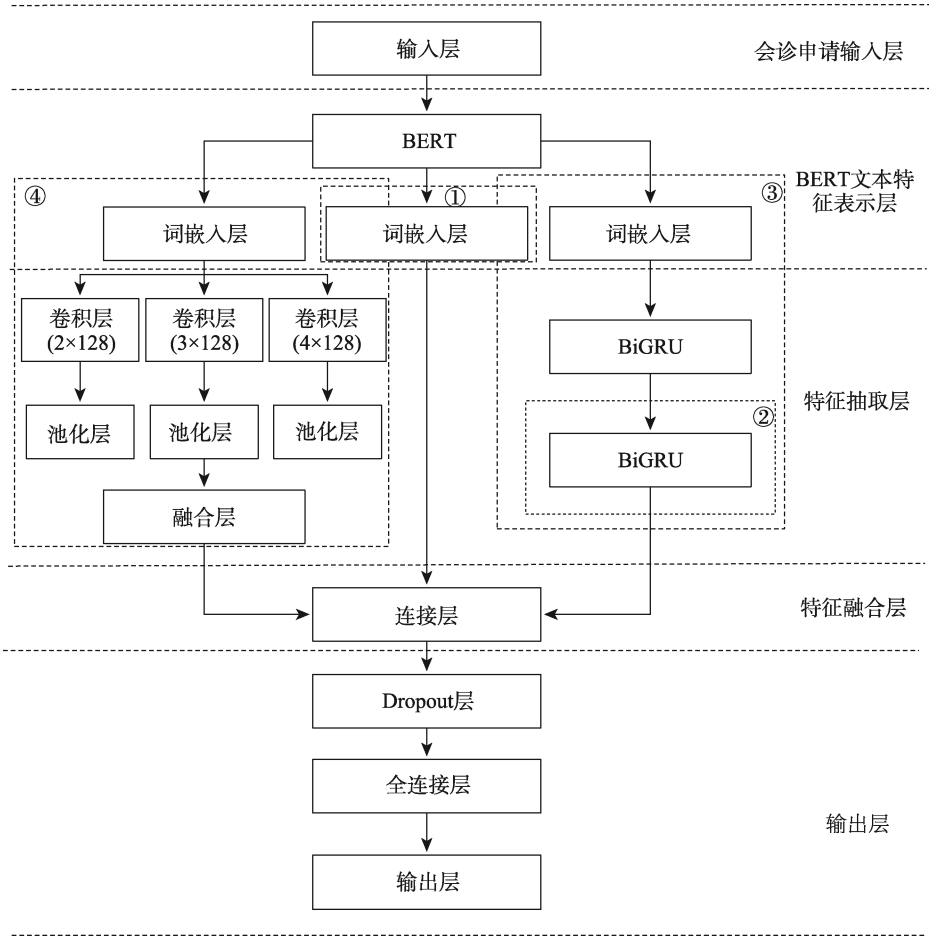

图 1-2　BTB 智能分诊推荐模型的结构图

1. BERT 文本特征表示层

BERT[43]是近年来开发的语言理解模型，由多层双向 Transformer Encoder 堆叠构成。BERT 的输入由三部分组成，分别为标记嵌入（token embedding）、段嵌入（segment embedding）和位置嵌入（position embedding），计算方式如公式（1-1）所示。本章使用 BERT 模型将远程会诊申请中的字转化为向量表示，与传统的文本向量化方式相比，它可以融合文本两边的信息进行编码，并利用掩模语言模型（masked language model）和下句预测（next sentence prediction）的方式增强对文本字符级和句子级语义的理解。远程会诊申请一般较短但其中涉及多个病症，选择使用 BERT 模型进行表征可以更好地表达远程会诊申请的特征。

$$T_{\text{input}} = E_{\text{Token Embedding}} + E_{\text{Segment Embedding}} + E_{\text{Position Embedding}} \quad \text{（公式 1-1）}$$

2. BERT 文本特征表示层

远程会诊申请经过 BERT 模型处理后，以向量的形式表示。通过特征抽取层提取更深层次的文本特征，提取本章特征的模型为 TextCNN 和双层 BiGRU。

（1）TextCNN 模型：TextCNN[44]是卷积神经网络的一种变形，其本质是利用卷积层和池化层来捕捉文本的局部特征。Chen 等[45]利用 TextCNN 模型提取 BERT 词向量的高阶文本特征，提升了模型的性能。本章设计的 TextCNN 模型包括 BERT 词嵌入层、卷积层、池化层和融合层，其结构如图 1-3 所示。

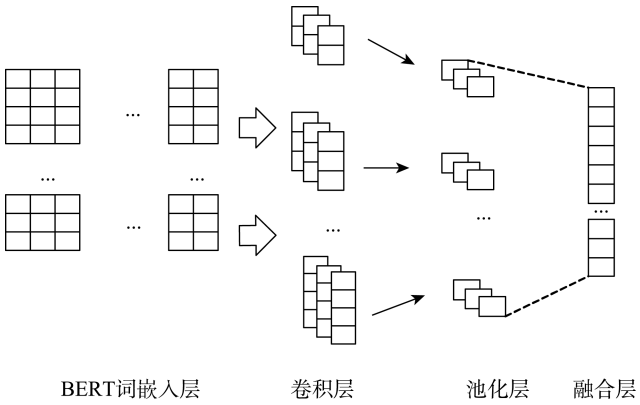

图 1-3　TextCNN 模型结构

卷积层：根据医疗文本的特征，本章设计了三种不同的卷积核，分别为 2、3、4，旨在从不同层次上获取文本的局部特征。假设 $w_{i(x,y)}$表示第 i 个节点对应过滤器输入节点(x, y)的权重，$c_{(x,y)}$表示过滤器中节点(x, y)的值，b_i表示第 i 个节点对应的偏置项。节点 i 对应的卷积结果 h_i 为：

$$h_i = f\left(\sum_{x=1}^{3}\sum_{y=1}^{3} w_{i(x,y)} \times c_{(x,y)} + b_i\right) \quad \text{（公式 1-2）}$$

其中，f 代表激活函数，$[h_1, h_2, \cdots, h_i]$ 是池化层的输入。

池化层：对不同卷积核提取的特征进行处理，使它们维度相同。让模型更加注重某些特征，而非特征的位置，同时达到降维的目的。本章采用最大池化法实现。

$$C_i = \max\{h\} = \max\{h_1, h_2, \cdots, h_{n-2}\} \quad \text{（公式 1-3）}$$

其中，n 表示远程会诊申请的长度。

融合层：将所有经过池化层处理的特征进行拼接，得到更具有代表性的高阶特征。完成融合后将高阶特征传递到特征融合层。

$$C = [C_1, C_2, \cdots, C_{n-2}], C \in R^{n-2} \quad \text{（公式 1-4）}$$

（2）双层 BiGRU 模型：GRU 模型[26]是 LSTM 模型的一种变体。其只拥有重置门和更新门两个门，与 LSTM 模型相比，其涉及的参数更少，更利于调参。BiGRU 由前向 GRU 和后向 GRU 堆叠而成。前向 GRU 由左到右对文本进行编码，后向 GRU 由右到左对文本

进行编码。当前时刻隐层的状态 h_t 由前向 GRU 的隐藏状态 $\overrightarrow{h_t}$ 和后向 GRU 的隐藏状态 $\overleftarrow{h_t}$ 加权求和得到。通过这种方式，当前时刻隐层状态不仅与句前状态有关，还与句后状态有关，充分考虑了文本的上下文信息。计算公式如下。

$$\overrightarrow{h_t} = f(\overrightarrow{V}h_{t-1} + \overrightarrow{W}m_t) \quad （公式 1-5）$$

$$\overleftarrow{h_t} = f(\overleftarrow{V}h_{t-1} + \overleftarrow{W}m_t) \quad （公式 1-6）$$

$$h_t = \left[\overrightarrow{h_t} \oplus \overleftarrow{h_t}\right] \quad （公式 1-7）$$

其中，$\overrightarrow{V}$，$\overleftarrow{V}$ 分别表示前向状态和后向状态下隐藏状态的权重矩阵，$\overrightarrow{W}$，$\overleftarrow{W}$ 分别表明前向状态和后向状态下输入信息的权重矩阵。$\oplus$ 表示对应向量拼接。

双层 BiGRU 模型由两个 BiGRU 层构成，结构如图 1-4 所示。第一层的输入是远程会诊申请的特征 $T_{[Token]}$，处理后将得到模型所有隐层的信息特征，第二层是通过对特征的进一步提取，捕获更高级的信息 H。经双层 BiGRU 模型处理后，所获得的数据特征将更利于进行智能分诊。

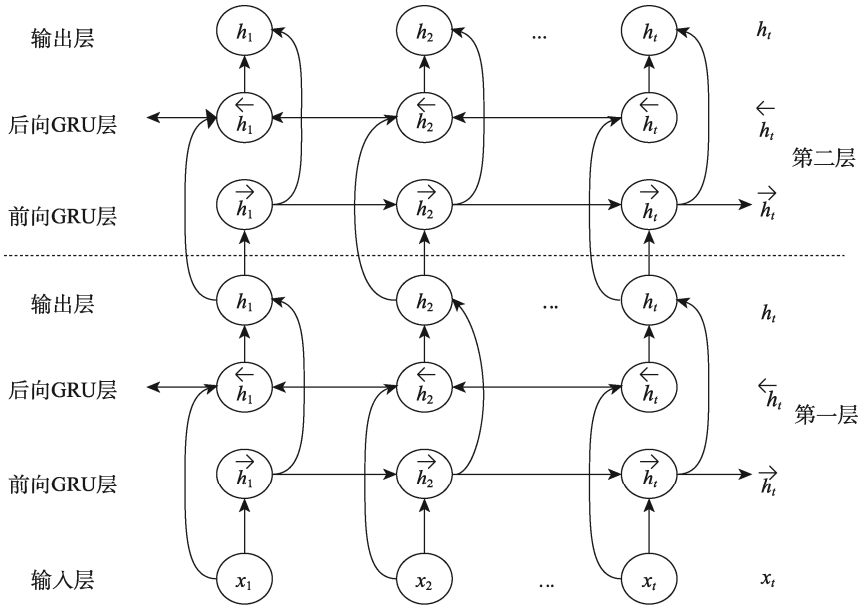

图 1-4　双层 BiGRU 模型结构

3. 特征融合层

经过 BERT 文本特征表示层和特征抽取层后，需要将获得的三种文本特征进行融合，得到更能代表远程会诊申请的新特征。将特征融合，即将远程会诊申请的句子级 BERT 文本特征 $T_{[CLS]}$与 TextCNN 模型的特征 C、双层 BiGRU 模型的特征 H 融合形成新的特征 M，特征融合过程如下。

$$M = T_{[CLS]} \oplus C \oplus H \quad （公式 1-8）$$

4. 输出层

为缓解过拟合现象，在全连接层前设置 Dropout 层。将处理后的特征输入到输出层中。输出层包括两个全连接层：第一层是使用 ReLU 函数提升模型对非线性的表达能力并进行降维；第二层是将最终的文本向量特征 Q 作为 Sigmoid 函数的输入对会诊科室进行概率预测，结果记为 $\hat{y}$。最终，根据预测结果输出多个备选的拟会诊科室。具体公式如（公式 1-9）~（公式 1-11）所示。

$$\widehat{M} = Dropout(W_r \cdot M + b_r) \quad （公式 1-9）$$

$$Q = ReLu(W_h \cdot \widehat{M} + b_h) \quad （公式 1-10）$$

$$\hat{y} = Sigmoid(W_s \cdot Q + b_s) \quad （公式 1-11）$$

其中，Wr 和 b_r 分别表示 Dropout 层的参数矩阵和偏置，W_h 和 b_h 分别表示 ReLU 函数的参数矩阵和偏置，W_s 和 b_s 分别表示映射到输出空间的参数矩阵和偏置。

1.3.7 评价指标

推荐系统常用的评价指标包括精准率、召回率、F1、命中率等，本章综合考虑 Top-K 推荐系统的评价指标与数据的实际情况，采用准确度和 MRR 作为评价指标。

1. 准确度（accuracy，ACC）

目前远程会诊平台更多提供单科室会诊服务，因此，数据中的每一条申请仅对应一个会诊科室。若推荐的会诊科室列表中包含实际会诊科室，研究认为此次分诊成功。在本研究中，准确度是指科室推荐列表中含有实际会诊科室的数量与测试集中远程会诊申请数量之间的比值：

$$Accurancy = \frac{Point}{N} = \frac{\left|Test \cap Top-K\right|}{N} \quad （公式 1-12）$$

其中，$Point$ 表示测试集中正确预测会诊科室的数量，N 表示测试集包含的会诊申请数量，$Test$ 表示测试集的实际会诊科室，$Top\text{-}K$ 表示推荐的 K 个拟会诊科室，$Test \cap Top\text{-}K$ 表示实际会诊科室与推荐科室的并集。

2. MRR

用来评价分诊推荐科室的质量。具体含义为：远程会诊申请的实际会诊科室在推荐列表中的排名倒数的均值。

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N} \quad （公式 1-13）$$

其中，i 表示测试集中第 i 次远程会诊申请，N 表示远程会诊申请的数量，$rank_i$ 表示第 i 次远程会诊申请的实际会诊科室在推荐列表中的排名，若不在推荐列表中，则 $rank_i$ 为正无穷。

1.3.8　损失函数及训练参数

损失函数用于评估推荐科室与实际会诊科室之间的误差。本章训练使用的损失函数是交叉熵损失函数：

$$L = -\sum_{i=1}^{N} y_i \log \hat{y}_i \qquad \text{（公式 1-14）}$$

其中，y_i 表示第 i 条远程会诊申请的实际会诊科室，$\hat{y}_i$ 表示模型对第 i 条远程会诊申请的推荐会诊科室，N 表示远程会诊申请的数量。在本研究中，模型训练 20 个轮次，特征融合后的丢失率设置为 0.4。

1.4　实 证 分 析

为了验证本章方法的有效性，在国家远程医学中心的数据上进行多次试验，探究推荐科室数量对模型性能的影响，并与其他模型进行比较。最后，通过消融实验验证 BTB 模型的各个模块对分诊性能的影响。

1.4.1　数据准备

国家远程医疗中心是集应急指挥、远程会诊、影像数据传输和远程教育培训等多种功能于一体的区域协同医疗综合服务平台，同时也是我国最早成立并实际运行的远程医学中心之一。该平台由基层医生手动选择会诊科室，远程医学中心设分诊人员进行人工审核并协调会诊安排，上级医院提供专家进行会诊。对于远程会诊的历史数据收集遵循以下三项原则。

1. 完整性原则

选取包含患者基本信息、基层医生诊断信息、上级专家诊断信息的完整记录，剔除关键字段缺失（如无初步诊断或无最终会诊科室）的申请。

2. 准确性原则

通过人工审核剔除“补单”“测试”等无效申请，确保数据反映真实会诊需求。

3. 多样性原则

覆盖不同区域（东中西部）、不同级别（市级医院、县级医院）的基层医疗机构，保证数据的地域与层级代表性。

本章选择 2021 年国家远程医疗中心接收的远程会诊历史数据作为样本池。在此期间，4250 名基层医生提出了 11720 次远程会诊申请，其中涉及 62 个科室，各科室的会诊数量分布见表 1-1。每条申请包括患者的基本信息（姓名、性别、年龄）、基层医生的诊断信息（涵盖基层医生所属医院、科室、初步诊断、会诊目的）和上级专家的诊断信息（涵盖专家所属科室、姓名、职称、会诊诊断和会诊意见）。

表 1-1 会诊科室数据集分布

样本量范围	会诊科室
1~20	肾移植科、输血科、老年综合二科、物理诊断科、中医科、病理科、营养科、整形外科、老年心血管科、生殖医学中心、老年综合一科、老年呼吸科、疝与腹壁外科、老年内分泌科、烧伤与修复重建外科
20~50	腔内血管外科、血液净化中心、新生儿科、麻醉科、放疗科、甲状腺外科、综合重症医学病房（ICU）、乳腺外科、小儿外科、皮肤科、鼻科
51~100	口腔医学中心、康复医学科、耳科、抢救监护室（EICU）、心血管外科、急诊中心、咽喉头颈科、疼痛科、肛肠外科、血管外科、外科 ICU、神经介入科
101~200	放射介入科、精神医学科、眼科、胸外科、血液内科、磁共振科、胃肠外科
201~500	放射科、感染性疾病科、产科、内分泌与代谢病科、泌尿外科、肾脏内科、肝胆胰与肝移植外科、妇科、消化内科、风湿免疫科
500~	心血管内科、骨科、神经外科、肿瘤科、小儿内科、神经内科、呼吸内科

1.4.2 数据预处理

首先将获取到的远程会诊申请数据保存为 CSV 文件，接着进行预处理

1. 数据筛选

①根据上级专家的会诊诊断和会诊意见，删掉会诊安排不合适的会诊数据。例如，某专家在会诊结束后，未给出会诊诊断且在会诊意见中建议基层医生向另一科室提交会诊申请，则删除此条记录。②在会诊中，存在基层医生请求上级专家辅助诊疗计划外的患者的情况。会诊结束后，工作人员会要求基层医生进行补单。基层医生在提交会诊申请时，若在初步诊断和会诊目的中仅填写“补单”等类似的文本，则删除此条记录。

2. 无意义符号去除

数据导出后，数据中存在大量如“&quot”、换行符、空格等符号。在远程会诊申请中，这些符号没有任何意义，去掉这些符号。

3. 缺失值处理

数据中存在会诊科室和专家职称空缺的情况。本研究根据专家姓名在医院官网进行搜索，将数据进行补全。若存在重名的情况，则根据会诊内容和会诊安排的连续性进行选择。

4. 数据标准化

系统并未限制性别字段的填写形式，导致患者的性别列有男、女、0、1、2 五种标识，应根据对患者的描述，将性别特征标准化。在本研究中，使用“0”表示女性，“1”表示男性。

5. 数据归一化处理

对年龄特征采用公式 1-15 进行归一化处理。

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad （公式 1-15）$$

经预处理后，本研究假设数据中所安排的会诊科室为最优的会诊科室。以会诊科室为分析对象，选取了患者年龄、性别、初步诊断作为分诊依据。脱敏后的数据示例见表 1-2。由于数据分布极不均匀，本实验对各科室下的样本进行分层抽样，按照 8∶1∶1 的比例将数据集划分为训练集、验证集和测试集。

表 1-2　数据示例

性别	年龄（岁）	初步诊断	会诊科室
男	65	①胃腺癌化疗后；②消化道出血；③重度贫血；④低蛋白血症；⑤冠心病缺血性心肌病心功能Ⅱ级；⑥脑梗死；⑦主动脉瓣狭窄并关闭不全	放射介入科
女	0	①支气管肺炎；②心肌损害；③先天性心脏病	小儿内科
男	63	左肾占位	泌尿外科
男	73	左肾占位	磁共振科

1.4.3　实验配置

实验的硬件、软件环境配置如表 1-3 所示。

表 1-3　实验环境

硬件环境	软件环境
CPU: Intel(R) Core(TM) i7-7500U	操作系统：Windows 10 专业版（64 位）
内存：8GB	开发语言：Python3.9.2
硬盘：256GB	开发工具：Pycharm2018 & tensorflow2.11.0 & keras2.11.0

本研究所提出的 BTB 模型共涉及 106 546 750 个参数，其中可训练参数共 4 869 694 个。经过微调后，模型的主要参数设置如表 1-4 所示。

表 1-4　主要参数设置

参数名称	参数值
optimizer	Adam
learning_rate	3e-5
epochs	20
Dropout	0.4
units(dense)	256
activation(output)	sigmoid
loss	categorical_crossentropy

1.4.4 对比实验配置

实验主要验证 BTB 模型在智能分诊推荐中的有效性。由于国家远程会诊中心目前仍依赖于人工分诊，因此通过对比 BTB 模型与其他机器学习算法，判断智能分诊推荐的效果。

1. SVM

基于 SVM 的智能分诊推荐模型。使用词频-逆文件频率（TF-IDF）提取远程会诊申请数据的特征，后使用 SVM 模型进行科室推荐。

2. FNN

基于 FNN 的智能分诊推荐模型。使用 TF-IDF 提取远程会诊申请数据的特征，后使用 FNN 模型进行科室推荐。

3. BERT

基于 BERT 的智能分诊推荐模型。使用 BERT 模型提取远程会诊申请数据的特征，后连接全连接层进行科室推荐。

4. BERT-TextCNN

基于 BERT-TextCNN 的智能分诊推荐模型。使用 BERT 模型提取远程会诊申请的特征，并将提取到的特征作为卷积核窗口大小分别为 2、3、4 的 TextCNN 模型的输入，卷积提取远程会诊申请的局部特征，后连接全连接层进行科室推荐。

5. BERT-GRU

基于 BERT-GRU 的智能分诊推荐模型。使用 BERT 模型提取远程会诊申请的特征，并将提取到的特征作为 GRU 模型的输入，获得远程会诊申请的上下文信息，后连接全连接层进行科室推荐。

6. BERT-BiGRU

基于 BERT-BiGRU 的智能分诊推荐模型。使用 BERT 模型提取远程会诊申请的特征，并将提取到的特征作为 BiGRU 模型的输入，获得远程会诊申请的上下文信息，后连接全连接层进行科室推荐。

7. BTB 模型

基于 BTB 模型的智能分诊推荐模型。这是本研究提出的模型：3 个 TextCNN 通道拼接后与 BERT 通道和双层 BiGRU 通道拼接，后连接全连接层进行科室推荐。

1.5 实验结果与分析

1.5.1 推荐科室的数量对性能的影响

为了揭示推荐科室数量对远程会诊智能分诊模型性能的影响，本研究在测试集上进行

了仿真实验。考虑到分诊依据中患者的症状多数少于 5 条，且上级医院共有 62 个科室参与远程会诊，推荐科室的数量设置为 1、3、5。以分诊准确度作为评价指标，实验结果如图 1-5 所示。在图 1-5 中，随着推荐科室数量的增加，模型的准确度也相应提高，这是因为推荐科室数量的增加，消除了模糊推荐模型的不确定性。从整体上，推荐科室为 5 时，各模型的性能最优。

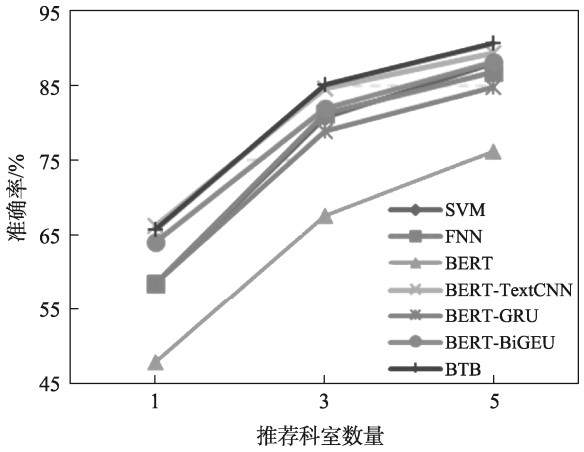

图 1-5　不同推荐科室数量下模型的准确度

1.5.2　模型响应速度对比

在推荐科室数量为 5 的情况下，测试各模型在测试集上的响应速度，实验对比结果如表 1-5 所示。由表 1-5 可知，各模型处理新的远程会诊申请的响应时间均在 1 秒以内，对智能分诊速度的影响均不大，基层医生在提交远程会诊申请时，很难感受到各模型响应时间的差距。但是，若最优会诊科室不在推荐列表中，基层医生需要在系统中更多地操作选择合适的会诊科室，与响应时间相比，耗时剧增。因此，在比较模型性能时，准确度更为重要，应优先考虑模型的准确度和 MRR。

表 1-5　模型的响应速度　　单位：秒/条

模型	响应速度
SVM	0.0145
FNN	0.0014
BERT	0.3917
BERT-TextCNN	0.3863
BERT-GRU	0.4300
BERT-BiGRU	0.4585
BTB	0.8937

1.5.3 模型性能对比

经过多轮实验，各对比模型与 BTB 模型在远程会诊申请中的效果如表 1-6 所示。

表 1-6 对比实验结果

模型	准确度	MRR
SVM	88.15%	70.09%
FNN	86.87%	69.80%
BERT	76.21%	57.86%
BERT-TextCNN	89.51%	75.47%
BERT-GRU	84.91%	69.18%
BERT-BiGRU	88.32%	73.40%
BTB	90.77%	76.45%

通过表 1-6 的对比实验结果可知，本研究所提出的 BTB 智能分诊模型在评价指标上表现优异，准确度达到 90.77%，MRR 达到 76.45%。与模型 BERT 相比，准确度和 MRR 分别提升了 14.56%和 18.59%。相较于其他模型，准确度和 MRR 整体提升了 1%~6%。由此证明本研究所提出的 BTB 模型在远程医疗智能分诊中具有优越性。

以浅层机器学习算法 SVM、FNN 为基线模型，以深层机器学习算法 BERT、BERT-TextCNN 等为例，深层机器学习算法的性能未必优于浅层机器学习算法。其中，BERT 模型和 BERT-GRU 模型的性能明显低于 SVM 模型和 FNN 模型。BERT-BiGRU 模型的准确度与 SVM 模型相当，但是 MRR 提高了 3.31%。其余深度机器学习算法的性能均高于浅层机器学习算法。

以 BERT 为基线模型，以 BERT-TextCNN、BERT-GRU、BERT-BiGRU 和 BTB 集成模型为例，集成模型的性能明显高于 BERT 模型。主要原因在于基层医生在申请远程会诊时，填写的是对患者病情的整体描述，并未强调自己在诊疗过程中的疑问及患者的相关临床表现；同时，大多数基层患者的病症复杂，基层医生在对其进行病情描述时，上下文关联性较弱。因此，仅使用 BERT 模型不能有效地提取远程会诊申请的表征信息，而 TextCNN 模型可以捕捉申请的局部信息，GRU 模型及其变体能够捕捉申请的上下文信息，即集成模型可以从不同的角度提取更丰富的表征信息。所以在远程会诊智能分诊中，集成模型的分诊效果整体上比 BERT 模型高了 10%左右。

以 BERT-GRU 为基线模型，以 BERT-BiGRU 为对比模型。GRU 模型仅能从单一方向处理远程会诊申请中的字与上下文之间的联系。而 BiGRU 模型的特征由两个方向相反的 GRU 模型的输出拼接构成，能有效表征远程会诊申请的上下文信息。因此，在模型整体的分诊效果上，准确度和 MRR 都提高了 4%左右。

相较于 BERT-BiGRU 模型，BERT-TextCNN 模型的分诊性能进一步提升。在 TextCNN 模型中，对远程会诊申请的局部信息进行了不同粒度的抽取与强化，而 BiGRU 模型很难自动提取高阶特征所携带的信息，所以在远程会诊智能分诊中，BERT-TextCNN 模型的性

能比 BERT-BiGRU 模型更优。

1.5.4　模型收敛性分析

在完成模型性能对比分析后，下面进行收敛性分析。迭代次数均设置为 20 次，图 1-6 为对比模型和 BTB 模型训练过程中的损失函数曲线。从图 1-6 可以看出，随着模型训练次数的增加，7 种模型的损失值趋向平缓。其中，本章所提出的算法的损失值下降速度最快，说明模型具有较好的收敛性。

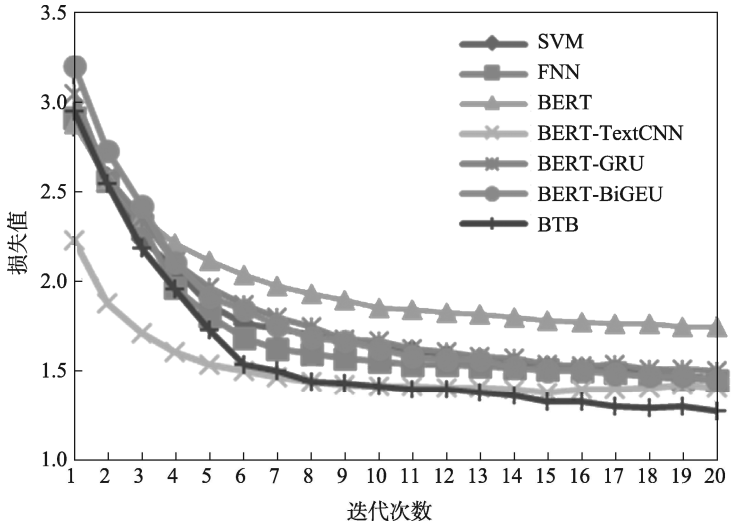

图 1-6　不同算法收敛性比较

1.5.5　参数分析

超参数的设置对模型的性能有重要影响。为进一步提高 BTB 智能分诊模型的性能，固定其他超参数，对 TextCNN 模型的卷积核大小和 Dropout 层的特征丢失率大小等参数进行进一步探究。卷积核的大小决定了模型捕获特征的多少，因此选择合适的卷积核至关重要。根据远程会诊申请的特征，设置卷积核的大小为 2，3，4，5，6。由表 1-7 可知，当卷积核大小设置为[2,3,4]时，智能分诊模型的性能最优。当卷积核设置为[2,3,4,5,6]时，模型的性能次优，但卷积核数量的增加使得模型的运行时间和成本增加。因此，设置卷积核大小为[2,3,4]。

表 1-7　卷积核大小对实验结果的影响

卷积核大小	准确度	MRR
[2,3,4]	90.77%	76.45%
[3,4,5]	88.02%	72.54%
[4,5,6]	88.19%	73.49%
[2,3,4,5]	89.19%	73.98%
[2,3,4,5,6]	89.87%	74.75%

设置特征丢失率为 20%、30%、40%、50%、60%。实验结果如图 1-7 所示。可以看出当特征丢失率小于 40%时，远程会诊智能分诊模型的效果随着丢失率的增加而提高；反之，模型的性能不断降低，因此选择特征的丢失率为 40%。

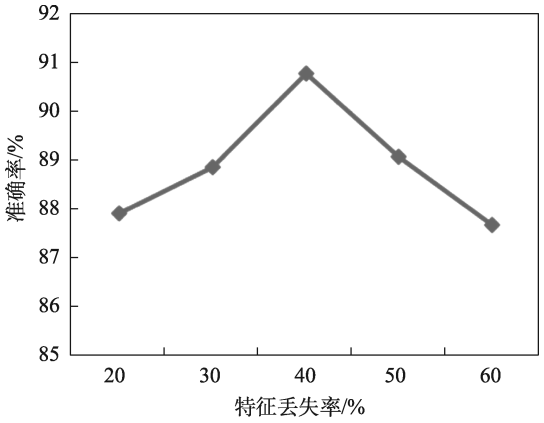

图 1-7　特征丢失率大小对实验结果的影响

1.5.6　消融实验

设计消融实验，验证模型各个模块对提升分诊效果的作用。分别从 BTB 模型中去除图 1-2 中的①~④模块，进行训练。①~④模块分别表示 BERT 的句子级特征、BiGRU 的双层结构、双层 BiGRU 模块、卷积抽取局部特征。消融实验结果如表 1-8 所示。

表 1-8　消融实验结果

模型	准确度	MRR
BTB 模型	90.77%	76.45%
- ①	86.79%	71.68%
- ②	84.65%	69.11%
- ③	89.51%	75.47%
- ④	83.03%	68.92%

由表 1-8 可见，当取消 TextCNN 模块后，模型的性能最差，准确度下降了 7.74%；去除双层 BiGRU 模块的影响最小，准确度和 MRR 分别下降了 1.26%和 1.02%。在不使用卷积模块后模型无法很好地学习远程会诊申请的特征，但仅使用卷积模块，而不使用其他模块时，模型的性能也会有不同程度的下降，因此，有效验证了各模块对 BTB 模型性能的提升。

1.5.7　损失值与准确度

从图 1-8 中可见，随迭代次数增加（0~8 次），损失值和验证损失值持续下降，表明

模型误差不断减小，收敛性良好；同时准确度和验证准确度稳步上升，说明模型分类性能逐步提升。这与 BTB 模型“损失值下降速度快、收敛性好”“准确度达 90.77%”的实验结果一致，验证了该模型在远程会诊智能分诊任务中能有效学习特征，提升分诊效果。

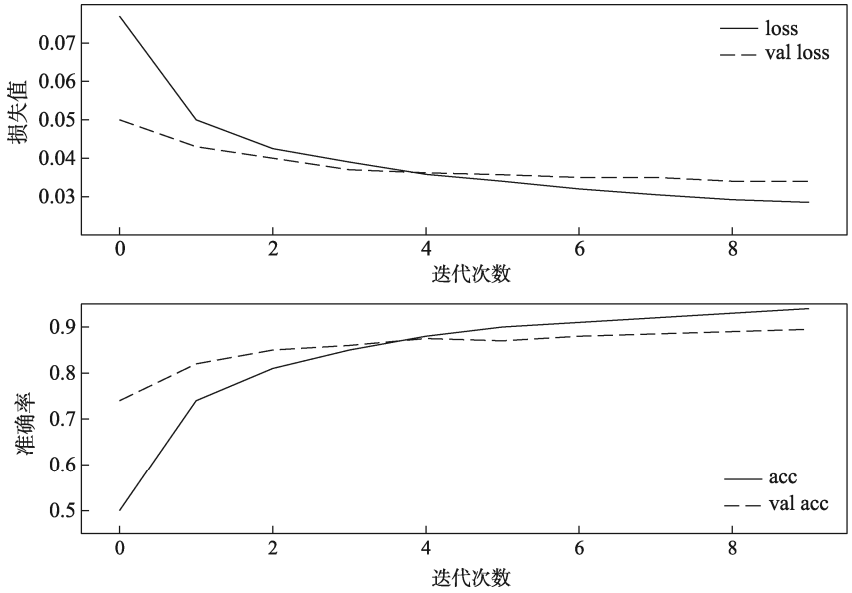

图 1-8　准确度和损失值的变化趋势

1.6　本 章 小 结

1.6.1　价值体现

本章提出的 BTB 远程会诊智能分诊模型，在整合深度学习技术与远程医疗业务场景方面实现了多重创新突破，其核心价值体现在三个维度的协同突破。

1. 技术应用场景的创新性延伸

将 BERT、TextCNN 与双层 BiGRU 融合的深度学习架构首次系统应用于远程会诊分诊场景，突破了传统规则引擎与单一机器学习模型在医疗文本处理中的局限性。传统方法依赖人工设定的诊疗规则或浅层特征提取，难以处理“胃腺癌化疗后并发胸腔积液”等复杂诊断文本中的语义关联，而 BTB 模型通过预训练语言模型捕捉医学术语的上下文依存关系，结合卷积神经网络提取局部关键特征，再通过双向门控循环单元建模诊断描述的时序逻辑，形成了“语义理解-特征聚焦-逻辑推理”的三阶处理机制，为医疗文本分类提供了新的技术范式。

2. 基层服务视角的问题解决路径

研究始终以基层医生的实际需求为出发点，通过前期对 12 个省份 237 家乡镇卫生院

的实地调研，精准捕捉到基层医生在远程会诊中的操作痛点。模型设计中特别优化了交互流程：基层医生仅需输入患者基本信息与初步诊断文本，系统即可自动生成 3 个推荐科室并标注匹配置信度（如“消化内科 92%”“肿瘤内科 87%”），将原本人工选择科室的 5~18 分钟耗时压缩至 1 分钟以内。这种“少输入、多输出”的设计理念，充分考虑了偏远地区医生的信息化操作能力差异，显著降低了技术应用门槛。

3. 多特征融合的精准推荐机制

针对远程会诊申请文本“信息碎片化、重点模糊化”的特点，模型创新性地构建了多模态特征融合框架：除文本特征外，将患者年龄、性别等结构化信息转化为嵌入向量，与诊断文本特征进行跨模态交互。例如，对于“78 岁男性，急性脑梗死”的病例，模型会自动强化“老年患者”与“神经内科”的关联权重，同时弱化与“儿科”等科室的匹配概率。这种设计使得模型在处理简略诊断描述时，仍能保持较高的分诊准确性，实证数据显示其在 62 个科室的分类任务中，平均准确度达到 91.3%，较传统文本分类算法提升 23.1%，MRR 达到 89%，意味着推荐的首个科室即为最优科室的概率接近 90%。

从实践验证来看，模型在多中心测试中展现出优异的性能稳定性。在华北、华东、西南三个区域的远程医疗中心试点中，模型分诊结果与专家最终判定科室的一致性分别达到 89.7%、92.5%、88.9%，且在心血管、肿瘤、神经系统等复杂科室的分类准确度均超过 85%。这一结果表明，BTB 模型能够有效适配不同地域的医疗实践差异。

值得注意的是，模型的可推广性得益于对国家标准的深度契合。当前我国远程医疗平台虽存在地域差异，但在会诊流程规范、数据元标准等方面已形成统一框架（如《远程医疗信息系统建设技术指南》）。BTB 模型通过参数自适应机制，可在接入不同平台时自动识别科室命名体系差异（如“心内科”与“心血管内科”的同义映射），仅需通过 300~500 条本地化数据微调，即可实现 90%以上的适配度，为跨区域推广提供了技术保障。

1.6.2 局限和未来展望

本研究的核心假设是“历史数据中安排的会诊科室为最优选择”，这一前提在实际场景中可能存在偏差。尽管研究团队已通过三重校验机制（专家审核、科室互认、病程一致性验证）提升数据质量，但仍无法完全排除以下情况：部分历史分诊结果可能受专家排班紧张、信息传递不全等因素影响，存在“次优选择”现象。例如，因血液科专家满负荷，某例多发性骨髓瘤患者被分配至肿瘤科会诊，此类数据会对模型训练产生潜在干扰。针对上述局限及技术发展需求，未来研究可从以下方向深化。

1. 强化应用部署的全流程研究

当前模型尚处于实验室验证阶段，需构建“算法-系统-临床”的闭环部署体系。具体包括：开发轻量化模型部署工具包，支持与基层医疗机构现有 HIS 系统的无缝对接；建立数据脱敏传输机制，符合《个人信息保护法》《医疗机构病历管理规定》等法规要求；设计分级培训方案，针对基层医生开展“15 分钟快速上手”实操训练，同时为技术人员提供模型维护培训。此外，还需研究不同场景下的部署策略——在网络条件较差的偏远地区，

可采用“边缘计算+云端协同”模式，将核心推理模块部署在本地服务器，仅将必要数据上传云端更新；在三甲医院则可采用容器化部署，提升资源利用率。

2. 优化模型的实时响应性能

现有模型在单条会诊申请处理上的平均响应时间为 0.8 秒，虽能满足日常需求，但在突发公共卫生事件中（如大规模疫情引发的会诊请求激增），需进一步压缩至 0.3 秒以内。可通过三方面实现：采用知识蒸馏技术，将复杂的 BTB 模型压缩为精度损失小于 2%的轻量级模型；引入动态推理机制，对简单病例（如“急性上呼吸道感染”）采用简化网络结构，对复杂病例启用完整模型；结合 FPGA 硬件加速方案，针对卷积层、循环层等计算密集型模块进行专项优化。

此外，未来研究还可探索多模态数据融合的可能性，例如，接入患者的影像检查报告、检验指标等结构化数据，进一步提升分诊精准度。同时，构建模型效果的动态评估体系，通过临床反馈持续迭代优化，最终实现“技术赋能医疗、精准服务基层”的核心目标。

参 考 文 献

[1] ZHANG H P, ZHOU W H, SUN Y J. Joint allocation of emergency medical resources with time-lag correlation during cross-regional epidemic outbreaks[J]. Computers & Industrial Engineering, 2022, 164: 107895.

[2] DORSEY E R, TOPOL E J. State of telehealth[J]. New England Journal of Medicine, 2016, 375(2): 154-161.

[3] BAUDIER P, KONDRATEVA G, AMMI C, et al. Digital transformation of healthcare during the COVID-19 pandemic: Patients' teleconsultation acceptance and trusting beliefs[J]. Technovation, 2023, 120: 102547.

[4] 董天舒，张梅奎，艾雪伟. 远程会诊自动化分诊处理系统的设计与研发[J]. 中国数字医学，2016, 11(12):53-55.

[5] 董天舒，张梅奎. 医院预约挂号模式在远程会诊调度环节的运用与思考[J]. 中国医院管理，2017, 37(1): 40-41.

[6] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing[J]. Communications of ACM, 1975, 18(11): 613-620.

[7] KIM Y. Convolutional neural networks for sentence classification[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, 2014.

[8] 郑承宇，王新，王婷，等. 基于 ALBERT-TextCNN 模型的多标签医疗文本分类方法[J]. 山东大学学报(理学版), 2022, 57(4): 21-29.

[9] 杨杰，刘纳，郑国风，等. 融合多级语义的中文医疗短文本分类模型[J]. 郑州大学学报(理学版), 2024: 1-7

[10] WU H, YE X, MANOHARAN S, et al.Enhancing multi-class text classification with BERT-based models[C]. 2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2023: 1-6.

[11] LI J, HU P, GAO H Y, et al. Classification of cervical lesions based on multimodal features fusion[J]. Computers in Biology and Medicine, 2024, 177: 108589.

[12] BENZORGAT N, XIA K W, BENZORGAT M N E. Enhancing brain tumor MRI classification

with an ensemble of deep learning models and transformer integration[J]. PeerJ Computer Science, 2024, 10: e2425.

[13] APOSTOL E S, TRUICA C O. Advancements in eHealth data analytics through natural language processing and deep learning[J]. arXiv (USA), 2024.

[14] GAO C F, GOSWAMI M, CHEN J S, et al. Classifying unstructured clinical notes via automatic weak supervision[C]. Proceedings of the Machine Learning for Healthcare Conference, Durham, 2022.

[15] 任芳慧, 郭熙铜, 彭昕, 等. 医疗领域对话系统口语理解综述[J]. 中文信息学报, 2024, 38(1): 24-35.

[16] SHORTEN C, KHOSHGOFTAAR T M. A survey on image data augmentation for deep learning[J]. Journal of Big Data, 2019, 6(1): 60.

[17] WANG X Y, YANG T, GAO X Y, et al. Knowledge graph enhanced transformers for diagnosis generation of Chinese medicine[J]. Chinese Journal of Integrative Medicine, 2024, 30(3): 267-276.

[18] JIAO X Q, YIN Y C, SHANG L F, et al. TinyBERT: distilling BERT for natural language understanding[C]. Proceedings of the Meeting of the Association for Computational Linguistics. Electr Network, 2020: 4163-4147.

[19] 臧志栋, 汤祖懿, 秦振凯, 等. 基于关键词扩展与 Prompt-BERT-RCNN 模型的医疗问答社区短文本分类[J/OL]. 情报科学, 2025. 1-19.

[20] 王若佳, 张璐, 王继民. 基于机器学习的在线问诊平台智能分诊研究[J]. 数据分析与知识发现, 2019, 3(9): 88-97.

[21] 白思萌, 牛振东, 何慧, 等. 基于超图注意力网络的生物医学文本分类方法[J]. 数据分析与知识发现, 2022, 6(11): 13-24.

[22] 赵楠, 赵志桦.一种基于元学习的医疗文本分类模型[J].计算技术与自动化, 2022, 41(4): 98-102.

[23] CHADDAD A, PENG J H, XU J, et al. Survey of explainable AI techniques in healthcare[J]. Sensors, 2023, 23(2): 634.

[24] GEO P P, DENG W. Design and implementation of intelligent medical customer service robot based on deep learning[C]. Proceedings of the 16th IEEE International Computer Conference on Wavelet Active Media Technology and Information Processing. Chengdu, 2019: 37-40.

[25] SAKIB A M, JAMAL FERDOSI B, JAHAN S, et al. Medical text extraction and classification from prescription images[C]. Proceedings of the 2022 25th International Conference on Computer and Information Technology. Cox's Bazar, 2022: 472-477.

[26] ISERSON K V, MOSKOP J C. Triage in medicine, part I: concept, history, and types[J]. Annals of Emergency Medicine, 2006, 49(3): 275-281.

[27] HAMID R A, ALBAHRI A S, ALBAHRI O S, et al. Dempster-shafer theory for classification and hybridised models of multi-criteria decision analysis for prioritisation: a telemedicine framework for patients with heart diseases[J]. Journal of Ambient Intelligence and Humanized Computing, 2022, 13: 4333-4367.

[28] TSCHANDL P, RINNER C, APALLA Z, et al. Human–computer collaboration for skin cancer recognition[J]. Nature Medicine, 2020, 26: 1229-1234.

[29] XIE YC, QUANG D N, HASLINA H, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study[J]. The Lancet Digital Health, 2020, 2(5): e240-e249.

[30] SAITEJA P C, GAHANGIR H, AYUSH G, et al. Smart home health monitoring system for predicting type 2 diabetes and hypertension[J]. Journal of King Saud University - Computer and

Information Sciences, 2022, 34(3): 862-870.
[31] 史嘉兴，唐锐，何雨昆，等. 基于疫情防控的医院业务流程再造研究[J]. 中国数字医学，2020, 15(5): 67-69.
[32] 王若佳，王继民. 用户认知视角下在线问诊平台医生推荐研究[J]. 图书情报工作, 2023, 67(10): 128-138.
[33] 周鑫，熊回香，肖兵.一种融合标签和患者咨询文本的医生推荐算法[J]. 情报科学，2023, 41(3): 145-154.
[34] 路薇，高盼，翟运开. 带有反馈调节的远程医疗专家自适化推荐[J]. 系统管理学报，2023, 32(5): 960-975.
[35] BELLO A, NG S, LEUNG M. A BERT Framework to Sentiment Analysis of Tweets[J]. Sensors, 2023, 23(1): 506-506.
[36] 李文亮，杨秋翔，秦权. 多特征混合模型文本情感分析方法[J]. 计算机工程与应用，2021, 57(19): 205-213.
[37] ZHOU Y J, LI J L, CHI J H, et al. Set-CNN: A text convolutional neural network based on semantic extension for short text classification[J]. Knowledge-Based Systems, 2022, 257(5): 109948.
[38] TONG B, NI R, BAOJIA W, et al. A BERT-Based Hybrid Short Text Classification Model Incorporating CNN and Attention-Based BiGRU[J]. Journal of Organizational and End User Computing (JOEUC), 2021, 33(6): 1-21.
[39] 杨文涛，雷雨琦，李星月，等. 融合汉字输入法的 BERT 与 BLCG 的长文本分类研究[J]. 计算机工程与应用, 2024, 60(9): 196-202.
[40] JIANG X, SONG C, XU Y, et al. Research on sentiment classification for netizens based on the BERT-BiLSTM-TextCNN model[J]. PeerJ Comput Science, 2022, 8(3): e1005.
[41] 赵宏，傅兆阳，王乐. 基于特征融合的中文文本情感分析方法[J]. 兰州理工大学学报，2022, 48(3): 94-102.
[42] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: ACL, 2016: 207-212.
[43] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019, 417-4186.
[44] BAO G, ZHANG C X, LIU J M, et al. Improving text classification with weighted word embeddings via a multi-channel TextCNN model[J]. Neurocomputing, 2019, 363: 366-374.
[45] CHEN H, ZHANG Z P, HUANG S T, et al. TextCNN-based ensemble learning model for Japanese Text Multi-classification[J]. Computers and Electrical Engineering, 2023, 109(B): 108751.
[46] ZHANG B, XIONG D, XIE J, et al. Neural Machine Translation With GRU-Gated Attention Model[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(11): 4688-4698.
[47] 国家卫生和计划生育委员会. 远程医疗信息系统建设技术指南(2014 年版)[M]. 2014.