# 目　录

# 摘　要

中国作为胃癌（Gastric Cancer）发病率最高的国家之一，胃癌早期诊断和预防具有重要意义。然而，早期胃癌的隐匿性和病情的快速恶化使得及早发现和诊断变得更加困难。为此，本文针对胃镜诊断文本（Gastroscopy Diagnosis Text）进行了深入的分析研究，应用深度学习技术高效地处理庞大的胃镜诊断文本，实现对早期胃癌患者的快速、准确筛查，为临床医生提供一个高效的辅助工具，医生进而能够制定个性化、有效的随访策略，以期提高胃癌患者的生存率和生活质量。本文将人工智能技术与临床医学实践相结合研究胃癌早期诊断和治疗具有重要理论意义和应用价值，主要工作包括：

（1）针对单一胃镜诊断文本进行癌症分期分类和病变部位识别困难的问题，提出了一种基于 BERT 的三分支分类网络。该网络充分利用了 BERT 预训练模型在文本表示学习方面的优势，对胃镜诊断文本进行编码以提取其中的深层次特征信息。为了同时进行癌症分期分类和病变部位识别，采用了三分支解码器结构。此外，网络还引入了基于关键字的胃镜诊断文本切分方法，以提高病变部位识别的准确性。最后，采用了类 BERT-SUM 的 BERT 输入格式，重现了 BERT 在预训练阶段的任务，从而使得 BERT 预训练模型能够更高效、更准确地捕获胃镜诊断文本的局部特征。该分类网络可以为医生们提供更准确、更全面的胃镜诊断文本分析工具，从而提高诊断的效率和精度，为患者的治疗提供更有针对性的指导和建议。

（2）针对患者风险中未能充分利用患者多条胃镜诊断文本时序特征的不足，提出了一种基于时间序列特征的患者胃癌风险预测网络。首先，利用经过肠胃镜语料库后训练（post-training）的 BERT 预训练语言模型，并作为胃镜诊断文本的语义特征提取器，以捕捉文本的语义信息。随后，将长短期记忆网络（LSTM）网络应用于时间特征的提取，通过设计的编码方法将两个胃镜诊断文本之间的时间间隔转换为 One-Hot 编码，从而使得 LSTM 网络能够有效地利用时间间隔特征。最后，采用基于全连接层和具有 softmax 激活函数的分类器进行预测，以帮助医生更好地识别潜在的胃癌患者并制定个性化的随访策略。此外，为了应对数据集不平衡的问题，提出了一种针对胃镜诊断文本的数据增强方法，以提高模型的鲁棒性。实验结果表明，相较于其他现有方法，本文提出的方法在胃癌风险预测方面表

现最佳。

（3）设计开发了一套专注于胃镜诊断文本的科研数据管理平台，旨在为用户提供直观高效的数据管理和标注功能。该平台采用了 Vue 和 Django 技术，为用户提供了一个功能强大且直观易用的平台。通过该平台，用户可以方便地导入、编辑、标注和导出胃镜诊断文本，以满足深度学习模型对大量有标签训练数据处理的需求。该平台还提供了面向深度学习模型的部署接口，用户可以轻松地将训练好的模型部署到平台上，以供测试和其他程序调用。这种便捷的部署方式为医疗从业者提供了更为便利的工具，能够更迅速地将先进的算法应用于临床实践中。

**关键词：**胃镜诊断文本；文本分类；BERT；深度学习；医疗数据处理

# Abstract

As one of the countries with the highest incidence of Gastric Cancer, early diagnosis and prevention of this disease are of paramount importance in China. However, the covert nature of early-stage gastric cancer and the rapid deterioration of the condition make early detection and diagnosis increasingly challenging. To address this issue, this paper conducts an in-depth analysis of Gastroscopy Diagnosis Text (GDT), applying deep learning techniques to efficiently process vast amounts of GDTs for the rapid and accurate screening of early-stage gastric cancer patients. This provides an effective auxiliary tool for clinicians, enabling them to develop personalized and effective follow-up strategies, with the aim of improving the survival rates and quality of life for gastric cancer patients. This study integrates artificial intelligence technology with clinical medical practice in researching early diagnosis and treatment of gastric cancer, bearing significant theoretical and practical value. The main contributions include:

(1) In response to the challenges of cancer staging classification and lesion site identification in single GDTs, we propose a BERT-based tri-branch classification network. This network leverages the text representation learning capabilities of the BERT pre-trained model to encode GDTs, extracting deep features. A tri-branch decoder structure is utilized for simultaneous cancer staging classification and lesion site identification. Furthermore, the network introduces a keyword-based segmentation method for GDTs to enhance the accuracy of lesion site recognition. Lastly, a BERT-SUM-like BERT input format is adopted, replicating the tasks of the BERT pre-training phase to capture the local features of GDTs more efficiently and accurately. This classification network offers a more accurate and comprehensive tool for analyzing GDTs, thereby enhancing diagnostic efficiency and precision, and providing more targeted guidance and recommendations for patient treatment.

(2) Addressing the inadequacy of fully utilizing the temporal features of patients' multiple gastroscopy diagnosis texts for risk assessment, we introduce a patient gastric cancer risk prediction network based on time series features. Initially, a BERT pre-trained language model, post-trained on gastroenterological corpora, serves as a semantic feature extractor for GDTs to capture textual semantic information. Subsequently, Long Short-Term Memory (LSTM) networks

are applied for extracting temporal features, converting the time intervals between two GDTs into One-Hot encoding through a specially designed encoding method, enabling the LSTM network to effectively utilize time interval features. Finally, predictions are made using a fully connected layer and a classifier with a softmax activation function to assist clinicians in better identifying potential gastric cancer patients and developing personalized follow-up strategies. Additionally, to tackle the issue of dataset imbalance, a data augmentation method for GDTs is proposed, enhancing the model's robustness. Experimental results demonstrate that our method outperforms existing approaches in predicting gastric cancer risk.

(3) A dedicated research data management platform for GDTs has been developed, aimed at providing users with intuitive and efficient data management and annotation capabilities. Utilizing Vue and Django technologies, this platform offers a powerful and user-friendly interface. Through the platform, users can easily import, edit, annotate, and export GDTs, meeting the deep learning models' demand for processing large volumes of labeled training data. The platform also provides a deployment interface for deep learning models, allowing users to easily deploy trained models for testing and integration with other programs. This convenient deployment method offers medical practitioners a more accessible tool, enabling the rapid application of advanced algorithms in clinical practice.

**Keywords**: gastroscopy diagnostic text; text classification; deep learning; BERT; medical data processing

# 第一章 绪论

## 1.1 研究背景及意义

胃癌（Gastric Cancer）是指胃部发生的恶性肿瘤，主要发生在胃黏膜上皮细胞。我国是胃癌的高发国家，胃癌的发病率和死亡率分别位居恶性肿瘤的第二位和第三位[1]。由于胃癌早期无明显症状，偶尔出现的上腹不适、嗳气等非特异性症状又与胃炎、胃溃疡等胃部慢性疾病极其相似易被忽略，这导致我国胃癌的早期诊断率较低[2]。

胃镜诊断文本（Gastroscopy Diagnosis Text）是医生在为患者进行胃镜诊断后书写的文本报告。医生在书写胃镜诊断文本时常常遵照固定的文本格式，合格的胃镜诊断文本需要描述包括病变的位置、形态、大小、颜色、表面状态、边界等内容。但由于胃镜诊断文本由医生主观书写，不可避免的会存在医生习惯性用语或主观描述，并且由于病变种类繁多，同一种病变也可能具有不同的表现。尤其是对癌症相关病变的描述更是五花八门，很难使用正则表达式（Regular Expression）等基于规则的方法[3]进行分类或分析。

在医院对患者进行胃镜诊断的过程中会生成大量的胃镜诊断文本报告。然而这些报告往往只被用于向患者解释病情，而未被充分利用于更深层次的数据挖掘和病情统计。这意味着医院可能积累了大量未被发掘的宝贵信息，这些信息可以为医疗实践和研究提供巨大的帮助。因此借助自然语言处理技术对这些胃镜诊断文本进行解析变得尤为重要。通过这种方式可以有效地提取出文本中隐藏的有价值信息，如疾病类型、病情严重程度等，从而为患者治疗方案和随访策略制定提供更精准的参考依据。利用这些数据进行统计分析，医疗机构可以更全面地了解患者群体的病情分布情况，识别出潜在的健康趋势和风险因素，有助于制定针对性更强的随访和治疗策略并提高医疗服务的效率和质量。

通过自然语言处理技术的应用，医疗机构还可以实现文本数据的自动化处理和分析，从而降低人力成本，提高工作效率。这对于现代医疗系统而言，是一种创新性的进步。然而值得注意的是，尽管自然语言处理技术在其他领域已经取得了巨大成功，但在胃镜诊断文本方面的研究和应用仍相对较少，特别是基于深度神经网络的处理方法尚未被广泛探索和应用，这为本研究和实践提出了挑战和机遇。

## 1.2主要研究内容和创新点

本文旨在运用深度神经网络的方法，从胃镜诊断文本中提取深层次的文本特征，实现更精确的信息抽取和分类。以满足临床上对胃镜诊断文本高效分类的需求。

第一步着重实现单条胃镜诊断文本分类。为满足临床上对胃镜诊断文本高效分类的需求，提出了一种三分支分类网络。该网络以 BERT 预训练模型[4-6]为编码层，由三个全连接层组成的解码层来处理镜下所见文本和病理诊断文本。具体而言，该网络可对胃镜诊断文本报告进行分类，将其分为早期癌症、进展期癌症和其他三个类别，并能准确指示癌症病变的发病位置。实验结果显示，该网络在分类准确率和癌症召回率方面优于其他方法。

第二步着重结合患者既往的胃镜诊断文本记录，实现对患者的癌症风险进行预测。本文提出了一种基于胃镜诊断文本时序特征的患者胃癌风险预测网络。该网络利用在肠胃镜语料库上后训练过的 BERT 预训练语言模型提取胃镜诊断文本的语义特征，并采用 LSTM 网络提取时序特征[7]。另外，本文还设计了一种编码方式，巧妙地将两次胃镜诊断文本的时间间隔转换为 One-Hot 编码，以使 LSTM 网络能感知时间间隔这一重要特征。该方法使用全连接层和 softmax 激活函数作为分类器输出最终结果。实验结果表明，该网络在分类准确率和召回率方面均优于其他方法。

最后，建立了一套面向胃镜诊断文本的科研数据管理平台。用户可在平台上导入、编辑、标注和导出胃镜诊断文本，以满足训练深度模型所需的有标签数据的直观高效处理。该平台还提供了面向深度模型的部署接口，用户可将训练的模型部署在平台上供测试和其他程序调用。上述基于 BERT 的三分支胃镜诊断文本的分类网络和基于胃镜诊断文本时序特征的患者胃癌风险预测网络已成功部署于科研数据管理平台，并已应用于齐鲁医院的临床诊断流程中[8, 9]。

**本文创新点包括：**

（1）提出了一种基于 BERT 的三分支胃镜诊断文本分类网络，实现了高效地对单条胃镜诊断文本进行分类，并且能够准确识别癌症病变的发病位置。

（2）提出了一种基于胃镜诊断文本时序特征的患者胃癌风险预测网络。该网络利用 GT-BERT 预训练语言模型提取胃镜诊断文本的语义特征，并采用融合时间间隔特征的 LSTM 网络提取时序特征。实现了高效准确的结合患者既往的胃镜诊断文本记录，对患者的癌症风险进行预测。

（3）建立了面向胃镜诊断文本的科研数据管理平台。用户可在平台上进行导入、编辑、标注和导出胃镜诊断文本。同时，平台还提供了面向深度模型的部署接口。

## 1.3 论文组织结构

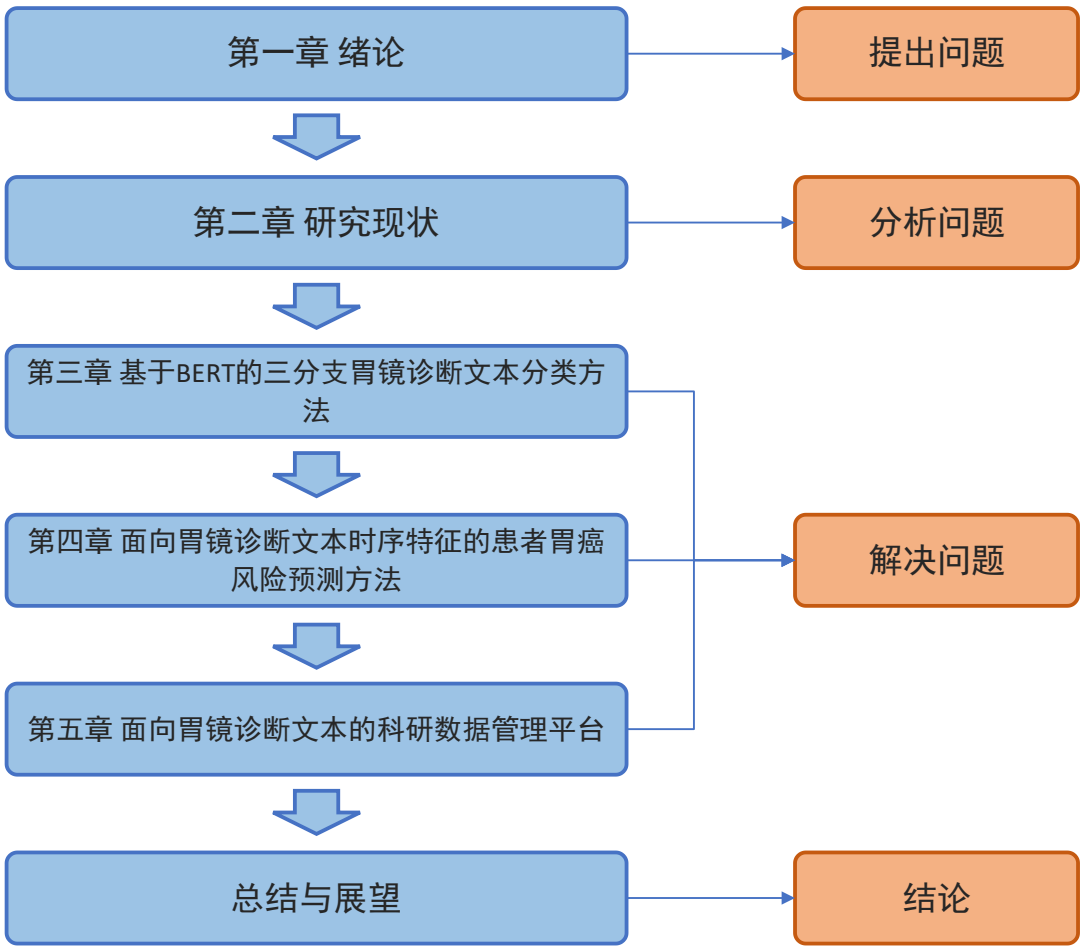


图 1-1 组织结构

本文的组织结构如图 1-1 所示，各个章节的主要内容如下所示：

第一章，主要概述了本文工作的研究背景和意义，然后提出了本文的主要研究内容和创新点；最后，运用组织结构图的形式表示了本文的主要研究内容。

第二章，对现有的通用文本分类和胃癌风险预测等方法进行了更进一步的分析和研究。

第三章，提出了一个基于 BERT 的三分支胃镜诊断文本分类方法。该方法综合考虑了胃镜诊断文本中特有的镜下所见部分和病理诊断部分对分类结果的影响，并实现了对患者癌症分期进行分类的同时识别病变所在的胃内部位。

第四章，提出了一个基于胃镜诊断文本时序特征的患者胃癌风险预测方法。该方法通过抽取患者既往多次胃镜诊断文本的时序特征，以实现更精准的预测患者的患癌风险。

第五章，本文还搭建了一套面向胃镜诊断文本的科研数据管理平台，用户可以在平台上导入、编辑、标注、导出胃镜诊断文本，以实现直观高效的处理深度模型所需要的训练数据。同时本平台还提供了面向深度模型的部署接口，用户可以将自己训练的模型部署在平台上，以供测试和其他程序调用。

第六章，总结本文主要研究内容和创新点，并对下一步工作进行展望。

# 第二章 医疗文本分类研究现状

近年来，由于深度学习的快速发展，自然语言处理（NLP）技术取得了重大进展。基于胃镜诊断文本的胃癌风险预测通常被归类为一种特殊的文本分类任务。在本章中将简要介绍在一般领域和胃癌风险预测领域进行的文本分类研究，由于本文同时也涉及时间序列预测和数据增强领域，所以也对上述两个领域的相关工作进行了探讨。

## 2.1 通用文本分类

在早期研究中，传统的机器学习方法通常被用于文本分类任务。尽管基于概率和条件独立假设的朴素贝叶斯（Naive Bayes）分类器[10]有局限性，但其也能在各种文本分类任务中表现出令人满意的性能。支持向量机(SVM)[11]能在高维空间中寻求最佳的超平面进行分类，也表现出卓越的泛化能力并经常优于朴素贝叶斯分类器。决策树方法旨在构建了一棵内部节点代表特征、叶子节点代表类别的树，在文本分类任务中提供了可解释性和易实施性，但它们在某些情况下容易出现过拟合现象。随机森林[12]是一种基于多棵决策树的集合学习方法，通过投票机制提高了分类性能，并在许多文本分类任务中表现出对噪声和过拟合的鲁棒性。k-NN 算法是一种基于实例的学习方法，根据特征空间中的 k 个近邻的类别对输入样本进行分类，尽管 k-NN 算法在大数据集中计算成本很高，但其在各种文本分类任务中仍然表现出良好的性能。

近年来，基于深度学习的方法在文本分类和分析任务中取得了丰硕的成果。许多研究将深度学习方法用于医学文本分类和分析。例如，有研究使用卷积神经网络对医学文本进行分类，结果表明卷积神经网络在临床文本分类任务中表现良好[13, 14]。

随着技术的快速发展，许多基于预训练模型的文本分类方法正在蓬勃发展。其中，谷歌于 2018 年推出的 BERT（Bidirectional Encoder Representation from Transformers）采用 Transformer[15]架构，为自然语言处理任务树立了新的基准。此后，BERT 的变体也被广泛提出并应用于自然语言处理任务，特别是 Facebook AI 提出的 RoBERTa（A Robustly Optimized BERT Pretraining Approach）[16]，进一步提高了自然语言处理任务的性能。此外，OpenAI 开发的 GPT（Generative Pre-trained Transformer）[17, 18]在文本分类任务中也取得了显著成果。

许多研究表明，预训练模型在实际应用中为该分类任务带来了性能改进。本文注意到有一项研究介绍了 EPAT-BERT[19, 20]，这是一种基于 BERT 的模型，使用两个粒度的预训练任务在电力相关文本上进行预训练。通过专注于电力特有的形态学和语义，EPAT-BERT 在

电力审计文本分类方面优于现有模型。还有一项研究重点介绍了 BiLGAT[21]，BiLGAT 是一种新的中文短文本分类模型，它通过聚合 BERT 的不同隐藏状态和使用双向格图注意力网络融合词典特征来增强字符表示。在三个数据集上的实验结果显示出优越的性能，在 THUCNEWS 上达到 94.75%的准确率，在 TNEWS 上达到 70.71%，在 CNT 上达到 86.49%。还有一项工作利用预先训练的大型语言模型（LLM）来预测由于安全问题而退出市场的可能性[22]，通过跨数据库验证实现了超过 0.75 的 AUC。并成功识别了超过 50%随后停用的药物，证明了它们在早期识别潜在药物不良反应和提高药物开发安全性方面的有效性。

## 2.2 胃癌风险预测方法

目前，一些研究正在利用胃镜诊断文本进行胃癌风险预测任务。有一项研究将多种传统的机器学习方法结合起来对胃镜诊断文本进行分类[23]，并将其与单一的传统机器学习方法进行比较。尽管这些方法取得了前所未有的优秀结果，但其依赖的传统的机器学习方法仍然因严重依赖特征工程和泛化能力弱而受到批评[24]。

除了临床文本分类任务外，一些研究还探索了基于深度学习的方法在胃镜诊断文本和胃镜诊断图像综合分析中的应用。例如， ID-GCS 采用分层注意机制来整合癌症筛查的多模式语义，并提出了新的胃镜报告筛查智能决策方法。实验结果表明，ID-GCS 在胃镜检查任务中表现良好[25]。

虽然现有的努力在医学文本和胃镜诊断文本分类和分析方面取得了一定的成功[26]，但在胃镜诊断文本的分类和分析中，文本预训练模型的使用尚未看到。因此，本文主要探讨如何有效地将预训练模型应用于胃镜诊断文本的分类和分析，以完成胃癌风险预测的任务。

## 2.3 时间序列特征预测

时间序列特征相关数据指的是具有明显时间序列标注的数据，如社交媒体数据和天气数据。许多研究表明，有效地利用数据中的时间特征对于提高预测准确性有着重要贡献。[27, 28]

在时间序列特征相关数据预测领域已经有了大量工作。有一项工作着重于通过整合网络拓扑信息和文本特征来改进股票数据的时间序列预测效果[29]。该方法涉及构建连接股票上下游行业的图形，提取有用的文本和拓扑特征，并利用机器学习来预测股票时间序列。强调了将基于行业链的股票拓扑结构纳入时间序列预测的有效性。另外，MTS-LOF[30]是一种新的医疗时间序列表示学习框架，结合了对比学习和遮罩自编码器方法。MTS-LOF 解决了标记医疗时间序列数据的挑战，通过多遮罩策略提供了复杂、上下文丰富的表示。实验

结果表明，MTS-LOF 优于其他方法，有望通过增强表示学习并探索医疗数据中时间和结构依赖关系的相互作用来显著改善医疗应用。

## 2.4 数据增强

在最近的研究中，数据增强（DA）技术在机器学习和深度学习领域引起了广泛关注。数据增强旨在通过转换和扩展原始数据来提高模型的性能和泛化能力。研究人员成功地通过应用旋转、翻转、缩放、引入随机噪声或变形等方式来改善模型在有限数据集上的训练性能。此外，先进技术如生成对抗网络（GANs）[31-33]也为数据增强提供了新的可能性，逐渐使合成数据的质量接近真实数据。这一研究方向的持续发展为解决小样本学习和模型过拟合等问题提供了有效手段，并为计算机视觉和自然语言处理的发展做出了重要的贡献。

在自然语言处理领域，也有许多优秀的数据增强研究工作。一项研究探索了数据增强方法在自然语言处理中的应用，突出了在深度学习和机器学习任务中数据增强的显著成功[34]。强调了除了在机器翻译中的回译之外，自然语言处理中数据增强技术的采用速度较慢，成功有限。该研究旨在通过对自然语言处理数据增强技术在不同设置下进行全面分析和比较，重点关注词汇多样性和语义保真度等方面，解决对近期研究中各种现有数据增强方法之间关系缺乏实际理解的问题。

Markus Bayer 等提出了用于自然语言处理中数据增强的文本生成方法[35]。该方法在短文本和长文本任务中都显示出了更有效的改进，使得在构建的低数据情景中获得了多达 15.53%和 3.56%的额外准确率提升。该研究在 11 个数据集上评估了该方法，在真实世界的低数据任务中有了显著的提升，并讨论了在不同类型数据集上成功应用该方法的影响和模式。

# 第三章 基于 BERT 的三分支胃镜诊断文本分类方法

## 3.1 引言

医院会诊中产生的大量文本数据是医生们记录诊断意见、制定治疗计划以及提供医疗建议的重要途径。这些数据承载着丰富的医疗信息，对于了解患者的病情和制定有效的治疗方案至关重要。然而，这些文本数据通常以非结构化形式存在，使得医生们难以快速准确地检索和分析其中的信息。由于文本的多样性和复杂性，需要一种有效的分类和分析方法来高效地提取其中的关键信息[36]。

目前尚未出现基于文本预训练模型对胃镜诊断文本的分类和分析的方法。因此，为了解决这一问题，本章提出了一种基于 BERT 的三分支胃镜诊断文本分类网络。这一网络旨在充分利用 BERT 预训练模型在文本表示学习方面的优势[37]，通过对胃镜诊断文本进行编码，提取其中的深层次特征信息。同时，为了使模型更全面地理解和分析文本内容，该方法采用了三分支解码器结构，实现了同时对病变部位和癌症分期进行分类。这一开创性的分类网络有望为医生们提供更准确、更全面的胃镜诊断文本分析工具，从而提高诊断的效率和精度，为患者的治疗和回访方案制定提供更有针对性的指导和建议。

## 3.2 研究材料

### 3.2.1 数据集

采用的数据集源自山东大学齐鲁医院早癌筛查数据库，共筛选使用了 2618 条胃镜诊断文本数据。这些文本数据均是由医生在对真实患者进行胃镜检查时所记录生成的。如图 3-1 所示，每条胃镜诊断文本均由镜下所见文本和病理诊断文本两部分组成，其中每一条文本至少包含镜下所见部分。

在这些胃镜诊断文本中，镜下所见部分是医生对患者进行胃内镜检查时所观察到的胃部组织病理学改变的详细描述。这些描述涵盖了组织的形态、结构和任何异常病变的特征，为医生提供了直观的视觉印象和诊断依据。部分患者在接受胃镜诊断时还进行了病理化验，因此这部分胃镜诊断文本还包含了病理诊断部分。病理诊断部分是针对于胃镜检查中发现的异常组织进行的病理学评估和诊断。通过对组织标本进行显微镜下的检查和评估，医生可以确定异常组织的性质、类型、程度和严重性，从而帮助确定疾病的类型和严重程度。这些病理诊断信息对于医生来说同样是至关重要的，因为它们提供了更准确详细的关于异常组织的性质和类型的信息，有助于进一步确认疾病的严重程度和制定更为有效的治疗方

案。因此，这份来自真实临床实践的数据集对于本研究具有重要的价值和意义。

镜下所见：食道粘膜光滑，粘膜下血管纹理清晰，收缩蠕动正常。贲门距门齿43cm，贲门远侧后壁见一溃疡，周边粘膜明显隆起粗糙，活检4块，质稍韧，质脆易出血。齿状线略欠清晰。
胃底粘膜充血水肿，粘液湖清亮，量中等，胃底近贲门见病变附近粘膜隆起。胃底可见一约0.4*0.4cm粘膜下隆起，表面光滑。
胃体粘膜充血水肿，皱襞规整。
胃窦粘膜明显充血水肿，红白相间，花斑样改变，蠕动尚可，幽门口及幽门前区粘膜充血水肿粗糙，粘膜隆起，质脆易出血，取活检4块送病理。
十二指肠球部及降段粘膜光滑，可见小片状粘膜充血水肿。

病理诊断：（贲门）腺癌。（幽门）局灶性粘膜内癌。（胃角）慢性萎缩性胃炎，轻度，并肠上皮化生，活动期。

图 3-1 胃镜诊断文本示例

### 3.2.2 数据集预处理及标注

在本章中，数据集的预处理起着至关重要的作用，其主要目的是对原始胃镜诊断文本数据进行清理、标准化、转换和格式化，以确保数据的质量和可用性。本研究采取了一系列措施来处理原始数据，以便使其适合用于模型训练和推理。首先，对原始数据进行了筛选和清洗，剔除了问诊系统测试数据等脏数据和无法使用的数据，以确保最终的数据集质量。同时也去除了 HTML 标签、多余的空格等对模型训练和推理有影响的符号，使得数据更加干净和规范。

经过预处理之后，最终得到了 2459 条可以用于模型训练的胃镜诊断文本。为了进一步确保文本标签的准确性，邀请了山东大学齐鲁医院的研究生与在职医生进行文本标记。最后得到了共计 2459 条标记完的胃镜诊断文本，每一条文本都同时具有序列标签（Sequence Label）和全文类别标签（Document Label）。这些胃镜诊断文本中，全部包含镜下所见部分，其中 1894 条还包含病理诊断部分。

序列标签着重标记描述癌症病变的文字序列。镜下所见部分的序列标记包括疑似早期癌症和疑似进展期癌症两类，而病理诊断部分的序列标记则包括病理确诊早期癌症和病理确诊进展期癌症两类。且每一个序列标签都标有在原文本中的开始位置和结束位置。

另外，为了标记每一条胃镜诊断文本所属患者的胃部病变情况，还为每一条有标记的胃镜诊断文本添加了全文类别标签。这些全文类别标签包括其他、早期胃癌和进展期胃癌三类，能够更准确地了解患者的疾病情况，为后续的研究和临床实践提供更为详尽的信息基础。

## 3.3 方法

### 3.3.1 方法概述

为解决基于胃镜诊断文本同时提取患者癌症发病部位和病情分级的问题，本章提出了一种创新性的三分支分类网络，如图 3-2 所示。该网络结构的设计旨在同时利用文本胃镜诊断文本的局部特征和全局特征，从而实现同时提取患者癌症病变所在的胃内部位和病情分级。方法的主要技术如下。

首先，为了有效地处理胃镜诊断文本，提出了一种基于关键字的文本切分方法。该方法能够将胃镜诊断文本根据其所描述的每一个胃内部位进行切分，得到若干文本单元，从而使得每个文本单元都能够更加准确地描述患者单一胃内部位的病变状况。这为后续提取出胃镜诊断文本的局部信息并分类提供了基础。

其次，采用了类似于 BERT-SUM 的 BERT 输入格式，将每个文本单元首尾分别添加[CLS]字符和[SEP]字符，并将它们依次相连。这可以帮助网络分别理解文本的整体和局部结构，并同时提取文本的全局和局部特征。

随后，利用 Google 预训练的中文 BERT-base 模型对胃镜诊断文本进行编码，提取文本的全局和局部特征。通过抽取 BERT 输出的代表全局特征的[CLS]字符所对应的特征向量，可以获得胃镜诊断文本的全局向量表示。抽取 BERT 输出的代表局部特征的[CLS]字符对应的特征向量，可以获得文本每个局部的向量表示。随后将这些特征向量分别送入三个可训练的全连接层，这三个全连接层分别用于处理来自胃镜诊断文本的全局特征、镜下所见部分和病理诊断部分的局部特征。这样的处理方式使得网络能够更好地理解不同部分的文本特征，并对其进行有效的分类。

最后，网络输出的分类结果被重新组织和处理，以输出患者的癌症进展情况及病变的胃内部位。通过这种三分支分类网络的设计，可以更准确地提取患者的癌症发病部位和病情分级，为医生的诊断和治疗提供更加全面和准确的参考信息。

### 3.3.2 基于关键字的胃镜诊断文本切分方法

为实现从单条胃镜诊断文本中同时提取患者的癌症分期和病变部位的需求，在文本预处理阶段就要求将胃镜诊断文本按照胃内部位切分为文本单元。考虑到胃镜诊断文本通常按照胃内部位进行书写的特征，相同部位的病变描述往往连续出现。这为基于关键字切分胃镜诊断文本提供了可行性。因此，设计出有效的文本单元切分方法对于算法的准确性提升至关重要。

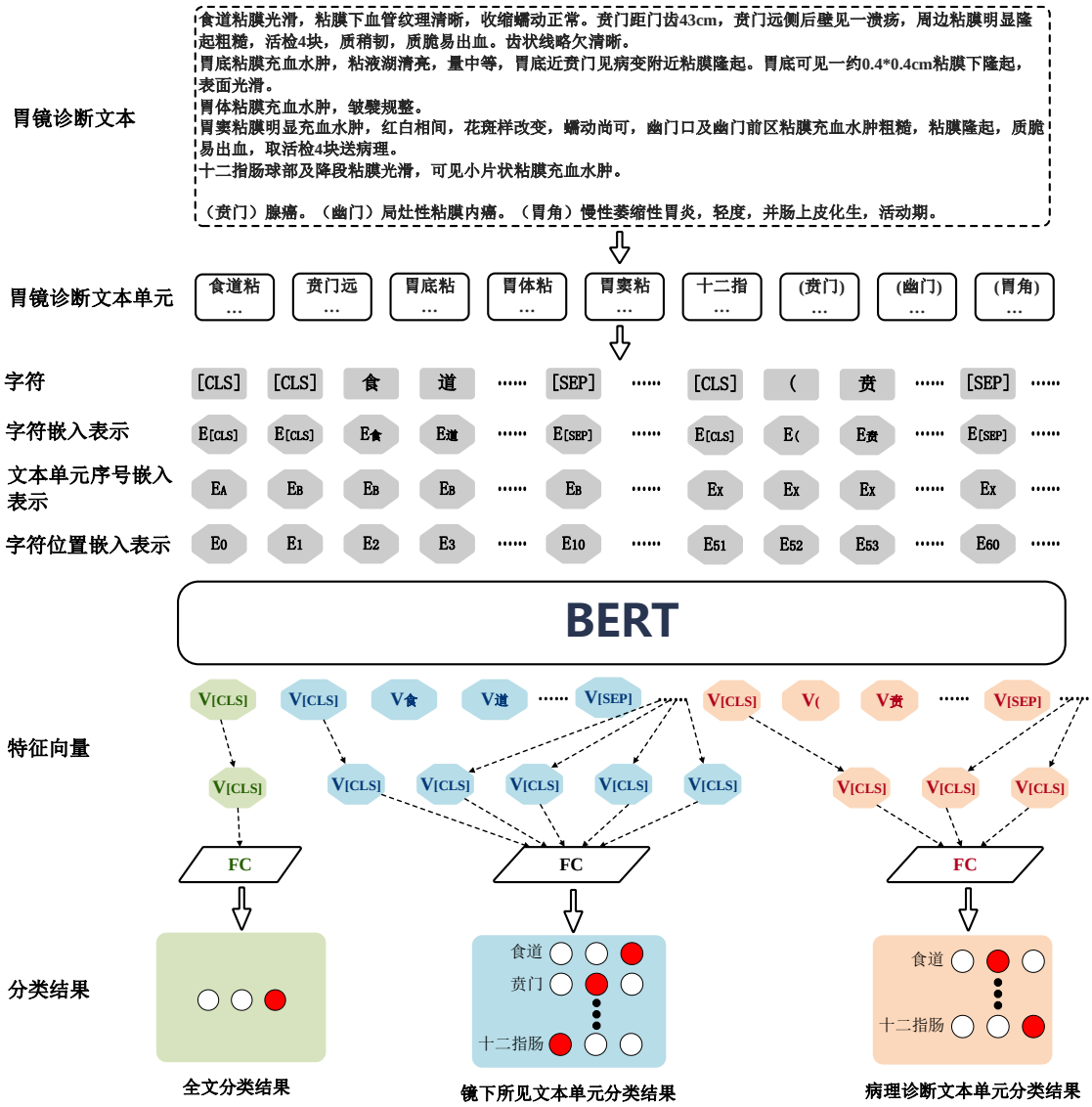


图 3-2 用于胃镜诊断文本分类的三分支分类网络

为了实现对胃镜诊断文本的有效切分，提出了一种基于胃内部位关键字的切分方法，如图 3-3 所示。这些胃内部位关键字包括：食道、贲门、胃底、胃体、胃窦、十二指肠。在胃镜诊断文本切分过程中，首先将文本按照所有的句号和换行符进行切分，得到若干文本片段。接着，分别检索这些片段中是否包含胃内部位关键字。最后，将不包含胃内部位关键字的文本片段与其在原文中位置前包含胃内部位关键字的文本片段合并在一起，形成新的文本片段。这样合并完成的每一个新文本片段即被视为一个文本单元。其中，来自镜下所见部分的文本单元被称为镜下所见文本单元，而来自病理诊断部分的文本单元被称为病理诊断文本单元。通过这种创新的切分方法，能够更准确地提取胃镜诊断文本中的关键信息，为后续的分析和分类工作奠定了坚实的基础。

为便于理解，对胃镜诊断文本切分的过程可以描述为：

$$outputTexts = mergeFragments(fargment_1, fargment_2, \ldots\ldots, fargment_n) \quad (3-1)$$

$$fargment_n = \text{containsKeyword}(split(inputTexts), keywordList) \quad (3-2)$$

其中，containsKeyword(textList, keywordList)是一个自定义的函数，用于判断输入的文本列表中是否含有$keywordList$中的关键字，并返回一个文本列表，包含一个含有关键字的文本和其随后的多个不包含关键字的文本。$split(text)$是一个自定义的文本切分函数，用于将胃镜诊断文本按照自定义的规则进行切分。$mergeFragments(fargmentList)$用于将文本列表合并成一段完整的文本。具体实现的步骤如算法 3-1 所示。

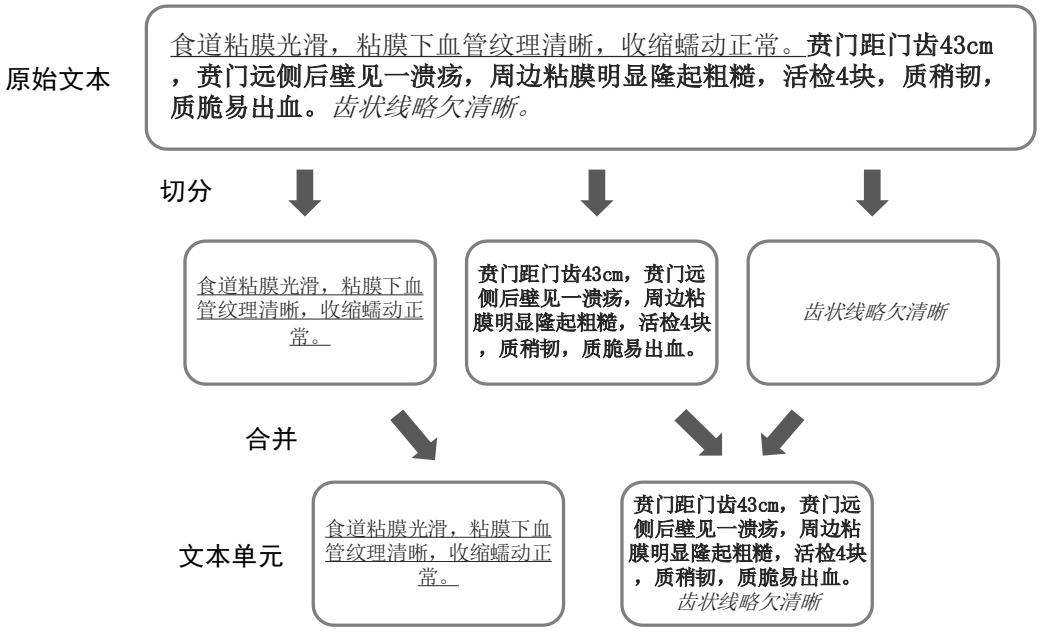


图 3-3 基于关键字的胃镜诊断文本切分方法

本研究从胃镜诊断文本数据集中随机抽取了 100 条胃镜诊断文本，经过上述基于关键字的胃镜诊断文本切分方法的处理，共得到了 654 个文本单元。随后邀请了齐鲁医院的研究生和在职医生对这些文本单元的切分效果进行评判。结果显示，被正确划分的文本单元共计 651 个，准确率达到了 99.54%。这表明，基于关键字的胃镜诊断文本切分方法在划分文本单元时表现出了较高的准确性和有效性。为后续的数据处理和分析工作提供了坚实的基础。

### 3.3.3 类 BERT-SUM 的 BERT 输入格式

为了满足同时提取胃镜诊断文本全局特征和每个文本单元局部特征的需求，本章深受 BERT-SUM 启发，采用了类似 BERT-SUM 的 BERT 输入格式来处理胃镜诊断文本分类任务。

在处理每个文本单元时，首先在其首部插入一个特殊的[CLS]字符，然后在文本单元的尾部插入另一个特殊的[SEP]字符。在传统的用于文本分类任务的 BERT 模型中，[CLS]字

符通常被用于抽取整个文本序列的特征，但在本章中将其用于抽取每个文本单元的局部特征。这种方式巧妙的还原了 BERT 在预训练阶段时的预测下句任务使得模型在学习每个文本单元的局部特征时更加有效。

随后，将所有的文本单元连接在一起，形成一个整体的文本序列，并在该序列的首部再次添加一个用于抽取全局特征的[CLS]字符。这样做的目的是让模型能够在学习局部特征时同时考虑整个胃镜诊断文本的内容。通过这种方式可以同时获取胃镜诊断文本的全局特征和每个文本单元的局部特征，为后续的分类任务提供更加丰富和准确的信息。

**算法 3-1：基于关键字的胃镜诊断文本切分方法**

**输入**：胃镜诊断文本 *gdt*

**输出**：文本单元列表 *text_unit_list*

```
gdt = "食道黏膜光滑……"
stomach_sites = ['食道', '贲门', '胃底', '胃体', '胃窦', '十二指肠']
text_unit_list = []
使用换行符和句号对 gdt 进行切分
for text_part in gdt:
    for site in stomach_sites:
        if site in text_part:
            // 文本单元如包含胃内部位关键字
            将 `text_part` 添加到 `text_unit_list`
        else if text_unit_list:
            // 文本片段如不包含胃内部位关键字
            将 `text_part` 连接到 `text_unit_list` 的最后一个元素
        end if
    end for
end for
```

### 3.3.4 三分支分类网络

网络的解码器部分采用三个线性层分别对全局特征向量、镜下所见文本单元特征向量、病理诊断特征向量三种类型的特征向量进行降维和分类。

首先，针对全局特征向量，采用传统的简单线性分类器。假设 BERT 的输出为$T \in \mathbb{R}^{L\times D}$，其中 $L$ 表示文本序列的长度，$D$ 表示 BERT 输出的维度。使用一个大小为 $D \times 3$的可训练参数矩阵 $W_a$ 来表示线性层，分类器的输出 $Y_a$ 可以通过以下公式计算得到：

$$Y_a = \sigma(W_a T_a + b_a) \qquad (3-3)$$

其中，$\sigma$ 表示 $softmax$ 函数，$T_a$表示全局特征向量。这个分类器的作用是对全局特征进行降维和分类，以便后续的处理和分析。

其次，对于镜下所见文本单元特征向量，在传统的简单线性分类器的基础上嵌套了时序层，以便同时对二维特征向量进行分类。假设镜下所见文本单元对应的[CLS]字符在原文中的索引为$I_g \in 1,2,\dots,L$。同样使用一个大小为 $D \times 3$ 的可训练参数矩阵 $W_b$ 来表示线性层，分类器的输出 $Y_b$ 可以通过以下公式计算得到：

$$V_g = gather\left(T, I_g\right) \in \mathbb{R}^{K \times D} \tag{3-4}$$

$$Y_b = timedistributed\left[V_g, f(x) = \sigma(W_b x + b_b)\right] \in \mathbb{R}^{K \times 3} \tag{3-5}$$

其中，K 为镜下所见文本单元的数量，$gather$ 函数的作用是从矩阵 $T$ 中按照抽取索引指定的行，并将它们排成一个新的矩阵；$timedistributed$ 函数的作用是按照时间维度分别对二维矩阵的每一行进行操作，并返回新的二维矩阵。这个分类器的目的是对镜下所见文本单元的特征进行分类，以便对胃镜诊断文本的局部信息进行更细致的分析。

最后，对于病理诊断文本单元特征向量的降维分类，采用了与镜下所见文本单元特征向量相似的方法。病理诊断文本单元对应的[CLS]字符在原文中的索引为 $I_p \in 1,2,\dots,L$。$M$ 为病理诊断文本单元的数量。同样地使用一个大小为 $D \times 3$ 的可训练参数矩阵$W_c$来表示线性层，分类器的输出$Y_c$可以通过以下公式计算得到：

$$V_p = gather\left(T, I_g\right) \in \mathbb{R}^{M \times D} \tag{3-6}$$

$$Y_c = timedistributed\left[V_p, f(x) = \sigma(W_c x + b_c)\right] \in \mathbb{R}^{M \times 3} \tag{3-7}$$

### 3.3.5 分类器

在本方法中采用了一个具有三个分支的线性层作为分类器，将模型编码层的输出降维转换为一组类别得分的向量。这些向量中的每一个元素表示了模型对每个类别的置信度。然后通过 softmax 函数将该向量转换为所有类别的概率分布。预测的类别标签是 softmax 函数输出向量中概率最高的类别，即具有最高概率值的类别标签。

对于给定的向量$x \in \mathbb{R}^n$，softmax 函数定义为：

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{3-8}$$

其中，$n$为向量$x$的维度，在本章中指癌症分期的类别数。

基于 BERT 的三分支胃镜诊断文本分类方法能够同时确定癌症的分期和病变组织在胃内的位置。其中癌症的分期结果是通过对代表胃镜诊断文本整体特征的向量进行降维分类得到的，即根据图 3-2 中的全文分类结果。而病变组织在胃内的位置则是通过对代表镜下

所见和病理诊断的局部特征向量进行降维分类得到的，即根据图 3-2 中的镜下所见和病理诊断文本单元分类结果。由于镜下所见和病理诊断中代表同一胃内部位的特征向量分别经过两个独立的全连接层降维，因此可以得到两个结果。根据临床经验，在实验中将两个结果中更严重的癌症分期级别作为最终的分类结果。

模型在训练过程中使用分类交叉熵（Categorical Crossentropy）[38]作为计算损失的函数。对训练数据的真实标签进行 One-Hot 编码，以便于对损失进行计算。在本章中，三个网络分支的分类器输出为$Y_a^i$，$Y_b^i$，$Y_c^i$。对真实标签进行 One-Hot 编码后的标签向量为$P_a^i$，$P_b^i$，$P_c^i$。$i$的范围是 1-3，表示三个分类类别。损失 loss 定义为：

$$Loss = -\sum_{i=1}^{3} \begin{bmatrix} P_a^i \\ P_b^i \\ P_c^i \end{bmatrix} \cdot \log\left( \begin{bmatrix} Y_a^i \\ Y_b^i \\ Y_c^i \end{bmatrix} \right) \tag{3-9}$$

## 3.4 实验

在本章中将详细讨论胃镜文本分析所采用的三分支分类网络的训练和预测过程。并将深入探讨网络的训练策略、参数调整以及优化算法的选择，以确保模型能够充分地学习和捕获胃镜诊断文本中的关键特征。

此外还将模型与基线模型进行比较，基线模型采用了基于 BERT 的简单文本分类方法。通过与基线模型的对比，可以评估三分支分类网络在胃镜文本分类任务中的性能优势，并探讨其在不同方面的表现差异。这有助于更全面地了解和评价所提出方法的有效性和可行性。

最后将对实验结果进行详尽的分析，探讨模型在不同指标下的表现，分析实验结果的稳定性和一致性。为进一步的研究和应用提供有益的启示。

### 3.4.1 实验环境

本方法使用了 TensorFlow[39]、Keras[40]和 BERT4keras 框架来实现。BERT 预训练模型是由 Google 训练的，具有 12 层隐藏层和 12 个注意力头，输出维度为 768，适用于中文文本的 RoBERTa。实验环境配置为 Ubuntu 20.04 操作系统，Intel 至强银牌 4208 CPU，Python 版本为 3.7，内存容量为 128GB。此外，使用了 NVIDIA RTX3090 GPU 加速计算。

### 3.4.2 评价指标

本章采用了多个标准的评价指标，包括准确率、精确率、召回率和 F1 值等。为了更全

面地评估模型的实用性，还引入了早癌召回率作为额外的评价指标，以更好地反映模型在临床应用中的表现。早癌召回率的引入是考虑到临床上对于早期癌症的准确诊断至关重要，因此该指标具有实际临床意义。

此外，本研究着重强调了基于 BERT 的三分支分类网络的另一个优势，即在对胃镜诊断文本进行癌症分期分类的同时，能够准确提取文本中描述患者病变的胃内部位的能力。这一点对于医学临床实践具有重要意义，因为准确确定病变的位置对于制定治疗计划和预后评估至关重要。为了评估模型在提取病变位置方面的性能，本研究与基于 BERT 的多标签分类网络进行对比，并利用常用的$F1_{\omega}$（加权 F1 分数）作为性能评估指标。

设$\omega_i$是标签$i$的权重，即标签$i$在整个数据集中出现的频率，$P_i$ 是标签$i$的精确率，$R_i$ 是标签$i$的召回率，$N$ 是标签的总数，在本章中 N 为 3。则$F1_{\omega}$可以表示为

$$F1_{\omega} = \frac{1}{\sum_{i=1}^{N} \omega_i} \sum_{i=1}^{N} \omega_i \cdot \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \qquad (3-10)$$

通过这些综合的评估和分析，可以全面了解模型的性能表现，并为未来的临床应用和医学研究提供有力支持。

### 3.4.3 模型训练与微调

对 BERT 模型进行全面的再训练和微调，以便将其应用于胃镜诊断文本分类任务。本章在 BERT 模型的微调过程中对其全部参数均进行了更新。同时，为了更好地适应任务，对网络中的线性层进行了随机初始化，这些线性层在整个训练过程中将与 BERT 模型的参数一起进行训练，从而使得整个网络能够更好地捕捉文本数据的特征并进行降维分类。

在模型的训练过程中，采用了 Adam 优化器作为优化算法，这是一种常用的自适应学习率优化算法，能够在训练过程中自动调整学习率，从而更有效地更新模型参数。还设置了一些常见的超参数，如学习率、批量大小和最大序列长度等，这些超参数的选择对于模型的训练效果至关重要。在实验中选择了初始学习率为 $5e^{-5}$。此外还设置了批量大小为 8 和最大序列长度为 512，以充分利用计算资源并提高模型的训练效率。

为了进一步提高模型的泛化能力和防止过拟合，引入了一些正则化技术，如 L2 正则化和 dropout。这些技术能够有效地减少模型在训练数据上的过拟合现象，并提高其对未见过的数据的泛化能力。通过在训练过程中对正则化项进行加权，能够有效地控制模型的复杂度，从而提高其在实际应用中的性能表现。

将训练数据集划分为训练集和验证集，并在每个训练轮次结束时对模型在验证集上的性能进行评估。这种训练和验证的交替过程有助于及时发现模型训练过程中的问题，并及

时调整模型的参数和超参数，以保证模型的训练效果和性能稳定性。最终，在模型性能达到一定水平并且不再提升时停止训练，以避免模型的过度拟合和性能下降。

表 3-1 实验数据统计

| | 训练集 | 验证集 | 共计 |
|---|---|---|---|
| **早期胃癌** | 246 | 53 | 299 |
| **进展期胃癌** | 352 | 82 | 434 |
| **其他** | 1369 | 357 | 1726 |
| **共计** | 1967 | 492 | 2459 |

### 3.4.4 实验过程与结果

训练过程在一块 RTX3090 GPU 上进行，经过了多次迭代，神经网络模型在第 18 次循环时显示出了收敛的迹象，损失函数值降至 0.023，并且在验证集上达到了最佳的准确率。接着对模型的癌症分期分类性能进行了广泛的评估，

随着深度学习技术的不断发展，大型语言模型如 chatGPT 已经成为了当今领域的重要代表之一[41, 42]。chatGPT 所采用的生成语言模型结构以及其训练数据的多样性使其在各个领域的任务中都展现出了出色的性能。chatGPT 无需额外的领域特定训练，就能够在多个任务中达到 SOTA（State of the Arts）水平，这使其成为了许多研究者和从业者的首选工具之一。本章利用了 OpenAI 提供的 API（Application Programming Interface），将 chatGPT 的核心模型 GPT3.5 与本方法进行了比较。关于 GPT3.5 的提示词，设计了一系列针对胃镜诊断文本分析任务的提示，并将其提供给 GPT3.5 模型。具体而言设计了如下提示词：“你是一个用于分析胃镜诊断文本的人工智能模型。我会提供给你一段胃镜诊断文本。请提供我的患者的胃癌分期（早期癌症、进展期癌症、其他）和受影响部位（食管、贲门、胃底、胃体、胃窦、十二指肠）。请以以下形式返回给我：（癌症分期：早期癌症，受影响部位：食管、贲门）。”然后，将 GPT3.5 返回的结果与测试数据的真实值进行比较，并使用相同的评估指标对模型的性能进行评估。通过这一实验设计，可以客观地评估 GPT3.5 在胃镜诊断文本分类和分析任务中的表现，并进一步了解其在医学领域中的适用性和局限性。这同时可以为未来利用 GPT3.5 进行医学文本分析提供参考。

方法采用了类似于 BERT-SUM 的 BERT 输入格式，对 BERT 的输入进行了编码，以实现单输入多输出的机制。这种编码格式的设计旨在充分利用 BERT 模型的表示学习能力，同时实现对文本中多个信息的有效提取。为了评估这种输入格式的有效性，将其与传统的基于 BERT 预训练模型的单输入多输出分类网络（SIMO-BERTCN）进行对比。在 SIMO-BERTCN 模型中，首先从 BERT 的输出中获取全局特征向量，然后将其送入七个全连接层中。这些全连接层的输出分别用于提取癌症分期和六个不同胃区域的病变状态。具体而言，

从第一个全连接层的输出中推导出患者癌症分期信息，并从其余六个全连接层的输出中获取对应各个胃区域病变状态的信息。随后将 SIMO-BERTCN 模型的输出与测试数据的真实值进行比较，并使用相同的评估指标对模型进行了全面的评估[43, 44]。

本次实验旨在评估不同模型在癌症分期和胃内部位识别方面的分类性能。本研究考察了多种基于单个文本类别输出的模型，其中包括卷积神经网络（CNN）、长短期记忆神经网络（LSTM）、传统的基于 BERT 的文本分类网络（BERTCN）以及本研究提出的新模型。除此之外还评估了这些模型在从描述患者病变的文本中识别胃内部位方面的能力。具体而言，比较了基于 BERT 的多标签文本分类网络（ML-BERTCN）、GPT-3.5 Turbo、SIMO-BERTCN 和本研究提出的模型的性能。这些模型各自采用不同的方法和架构，并以解决癌症分期和胃内部位识别的任务为目标。通过对比实验结果可以全面了解各模型在任务上的表现。具体的实验结果见表 3-2。

表 3-2：实验结果

| | 精确率 | 召回率 | F1 | 早癌召回率 | $F1_\omega$ |
|---|---|---|---|---|---|
| CNN | 0.934 | 0.926 | 0.930 | 0.496 | - |
| LSTM | 0.842 | 0.855 | 0.848 | 0.428 | - |
| BERTCN (RoBERTa) | **0.993** | **0.993** | **0.993** | 0.765 | - |
| ML-BERTCN (RoBERTa) | - | - | - | - | 0.724 |
| GPT-3.5 turbo (API) | 0.557 | 0.659 | 0.637 | 0.679 | 0.707 |
| SIMO-BERTCN(RoBERTa) | 0.962 | 0.933 | 0.947 | 0.679 | 0.591 |
| **OUR (RoBERTa)** | **0.993** | **0.993** | **0.993** | **0.784** | **0.849** |

### 3.4.5 讨论

除了在癌症分期分类任务中取得了显著的准确率外，基于 BERT 的三分支胃镜诊断文本分类方法还展现出了对癌症病变部位的高效提取能力。研究着眼于利用改进的输入结构，使 BERT 在识别胃病变位置的同时进行癌症分期分类。相较于传统的基于 BERT 的文本分类方法，本研究提出的方法在癌症分期分类任务中取得了与之相当的准确率。这一结果清晰地表明，模型在提取癌症病变部位的同时不会牺牲对癌症分期的准确性和分类性能。为了验证方法的有效性，采用了齐鲁医院提供的胃镜诊断文本数据集进行了实验。在这个数据集上，方法的精确率、召回率和 F1 指数均达到了 99.3%。这一结果为模型在处理胃镜诊断文本时的可靠性和有效性提供了有力支持。

除了对癌症分期的分类能力外，还对基于 BERT 的三分支胃镜诊断文本分类方法在癌症病灶提取任务中的性能进行了详细评估。实验结果显示，方法在癌症部位提取任务中明显优于基于 BERT 的多标签文本分类方法（ML-BERTCN）。具体来说使用了多标签分类任务中常用的加权 F1 分数作为评估指标。结果显示基于 BERT 的三分支胃镜诊断文本分

类方法在癌症部位提取中获得了 0.849 的加权 F1 得分，而基于 BERT 的多标签文本分类方法仅获得了 0.724 的加权 F1 分数。这表明方法能够更准确地提取癌症的病变部位，从而有希望在实际应用中为医生更好地确定患者的病情提供可靠的辅助。

另外，本研究也对基于 BERT 的三分支胃镜诊断文本分类方法与 GPT3.5 进行了比较。作为一种生成式的预训练语言模型，GPT3.5 在阅读理解和问答等任务中取得了 SOTA 的效果。然而，对于胃镜诊断文本分类和分析任务，GPT3.5 由于无法有效地微调以适应特定任务的需求，因此未能展现出令人满意的结果。

消融实验采取一系列措施来评估模型中各个组成部分的有效性。单输入多输出分类网络（SIMO-BERTCN）从模型架构中删除了针对胃镜诊断文本的专用输入模块，直接使用 BERT 生成的全局特征向量，并对这些向量进行了降维并产生多个输出。这一设计旨在评估类 BERT-SUM 的 BERT 输入格式和基于关键字的胃镜诊断文本切分方法在模型中的效果。实验结果揭示了单个 BERT 全局向量在胃癌分期分类和胃病变部位识别任务中的局限性。这进一步证明了引入的组件能够在整个模型中发挥重要作用。

## 3.5 本章小结

本章提出了一种基于 BERT 的三分支分类网络，该网络旨在解决单一胃镜诊断文本同时进行癌症分期分类和病变部位识别的挑战。这一网络结构充分发挥了 BERT 预训练模型在文本表示学习方面的优势，对胃镜诊断文本进行编码，并提取其中的深层次特征信息。为了同时进行癌症分期分类和病变部位识别，该方法采用了三分支解码器结构，能够有效地对病变部位和癌症分期进行分类。此外，提出了一种基于关键字的胃镜诊断文本切分方法，以提高病变部位识别的准确性。同时，研究采用了类 BERT-SUM 的 BERT 输入格式，巧妙地重现了 BERT 在预训练阶段的任务，使得 BERT 预训练模型能够更高效、更准确地对胃镜诊断文本的局部特征进行编码。这一综合性的分类网络有望为医生们提供更准确、更全面的胃镜诊断文本分析工具，从而提高诊断的效率和精度，并为患者的治疗提供更具针对性的指导和建议。

# 第四章 基于胃镜诊断文本时序特征的患者胃癌风险预测方法

## 4.1 引言

患者的胃癌风险评估对于医生向患者提供治疗建议和制定随访策略具有重要意义。在医生对患者进行胃癌风险评估时，胃镜诊断文本扮演了关键的角色。这些文本记录了医生对患者胃部观察到的详细情况和病变组织的活检记录，为医生提供了重要的参考依据。然而，由于医生评估患者胃癌风险时通常需要结合患者过往的诊断报告，目前尚未出现一种自动化方法能够有效地结合患者过往的诊断记录，以辅助医生进行胃癌风险预测。

上一章主要探讨了如何对单条胃镜诊断文本进行分类，通过训练一种基于 BERT 的三分支胃镜诊断文本分类方法，实现了对单条胃镜诊断文本的高效准确分类。然而，在临床实践中，医生通常需要参考患者过往的多条胃镜诊断记录，因为患者的癌症风险评估往往需要考虑病灶的发展趋势。因此，本章为了更准确地预测患者的胃癌风险，提出了一套基于胃镜诊断文本时序特征的患者胃癌风险预测网络。

提到能够处理时序特征的深度学习网络，就不能不提 LSTM（Long Short-Term Memory），LSTM 网络作为一种改进的循环神经网络，能够有效解决传统 RNN 难以处理长距离依赖的问题[45, 46]，其核心概念是细胞状态和“门”结构。细胞状态等同于信息传输的路径，允许信息在串行链中传递。因此，即使是较早时间步的信息也可以传递到较晚时间步的单元中，这克服了短期记忆的影响，并使 LSTM 在提取胃镜诊断文本的时间序列特征方面具有很大的优势[47]。

为了满足通过结合患者先前的胃镜检查诊断文本序列来预测患者的胃癌风险的需求，本章提出了一种基于时间序列特征的患者胃癌风险预测网络。首先，将经过胃镜语料库后训练的 BERT 预训练语言模型用作胃镜诊断文本的语义特征提取器，并将 LSTM 网络应用为基于时间的特征提取器。此外，设计了一种编码方法巧妙地将两个胃镜诊断文本之间的时间间隔转换为 One-Hot 编码，使 LSTM 网络能够将时间间隔视为重要特征。最后，利用基于全连接层和具有 softmax 激活函数的分类器来降维分类并获得最终结果。为了解决数据集不平衡的问题，本章还提出了一种胃镜诊断文本的数据增强方法。实验结果表明，与其他现有方法相比，所提出的方法在分类准确性和召回率方面表现最佳。基于时间序列特征的患者胃癌风险预测方法帮助医生更科学地利用胃镜诊断文本，并为他们的患者制定随访策略。

## 4.2 研究材料

### 4.2.1 数据集

本章所使用的数据集来自齐鲁医院早癌筛查数据库，其中包含了 4191 位病人的 15055 条胃镜诊断文本。每位患者的胃镜诊断记录均包括三次或以上的诊断文本。这些文本均由医生在对真实患者进行胃镜检查时所撰写，是临床实践中的真实记录。每条胃镜诊断文本都包含了“镜下所见”部分，其详细描述了医生在胃镜检查中观察到的胃部组织的病理学变化。另外，部分文本还包含了“病理诊断”部分，用于对发现的异常组织进行病理学评估和诊断。这些病理诊断是通过对异常组织标本进行显微镜下检查和评估得出的，有助于医生确定异常组织的性质、类型、程度和严重性，进而帮助确定疾病的类型和严重程度。这些丰富的数据提供了深入研究患者病情发展趋势和预测胃癌风险所需的重要信息基础。

### 4.2.2 数据集标注

本章所提出的模型学习过程采用监督学习的方式，因此对每一条训练数据都需要进行标注。首先利用了本文第三章的成果，即基于 BERT 的三分支胃镜诊断文本分类网络，对所有的胃镜诊断文本进行了预标记。具体操作是为每一条胃镜诊断文本赋予癌症置信度标签，并依照山东大学齐鲁医院研究生和在职医生的建议设定了癌症高风险和癌症中风险的阈值。根据这些阈值，将每一位患者的平均癌症置信度超过高风险阈值的文本预标记为高风险患者，将平均癌症置信度超过中风险阈值但未超过高风险阈值的文本预标记为中风险患者，其余则标记为低风险患者。

这些预标记的标签经过了齐鲁医院在职医生和研究生的逐一核对和确认，确保了标注的准确性和可靠性。最终得到了具有准确标签标注的训练集、验证集和测试集。这一过程不仅为模型的学习提供了高质量的训练数据，还为后续实验和性能评估提供了可靠的基础。通过这样的标注流程能够提高模型在癌症风险评估中的准确性和可靠性，从而更好地帮助医生做出临床决策。

## 4.3 方法

### 4.3.1 方法概述

为了实现结合患者过往的胃镜诊断文本预测患者胃癌风险的需求，本章提出了一种全新的胃癌风险预测模型，即基于胃镜诊断文本时序特征的患者胃癌风险预测网络，如图 4-

1 所示。该网络的设计充分考虑了胃镜诊断文本序列的时序特性，旨在利用这些文本中的时序信息来更准确地预测患者的胃癌风险。

首先采用了在肠胃镜语料库上后训练过的 BERT 预训练语言模型 GT-BERT 作为胃镜诊断文本的语义特征提取器。GT-BERT 能够更好的捕获胃镜诊断文本中的丰富语义信息，为后续的时序特征提取奠定了基础。其次，为了更好地捕捉文本序列中的时序特征，引入了 LSTM 网络作为时序特征提取器。LSTM 网络以其优秀的记忆能力和对序列数据的良好建模能力而闻名，能够有效地学习和表示文本序列中的时序信息。

关键的一步是设计一种能够编码两次胃镜诊断文本之间时间间隔的方法，以使得 LSTM 网络能够感知到这一重要特征。为此采用了一种巧妙的方法将时间间隔转换为 One-Hot 编码形式，并将其作为 LSTM 网络的输入之一。这样神经网络就能够在学习时序特征时同时考虑到文本之间的时间间隔，从而更全面地把握文本序列中的时序关系。

最后为了输出最终的患者胃癌风险预测结果，采用了一个全连接层和 softmax 激活函数作为分类器。这一分类器能够将模型提取的语义特征和时序特征有效地结合起来，从而准确地预测患者的胃癌风险等级。在模型的训练阶段，通过端到端的学习过程模型能够充分利用胃镜诊断文本序列中的丰富信息，为医生提供更准确、可靠的胃癌风险预测结果，从而辅助其进行临床决策和治疗规划。

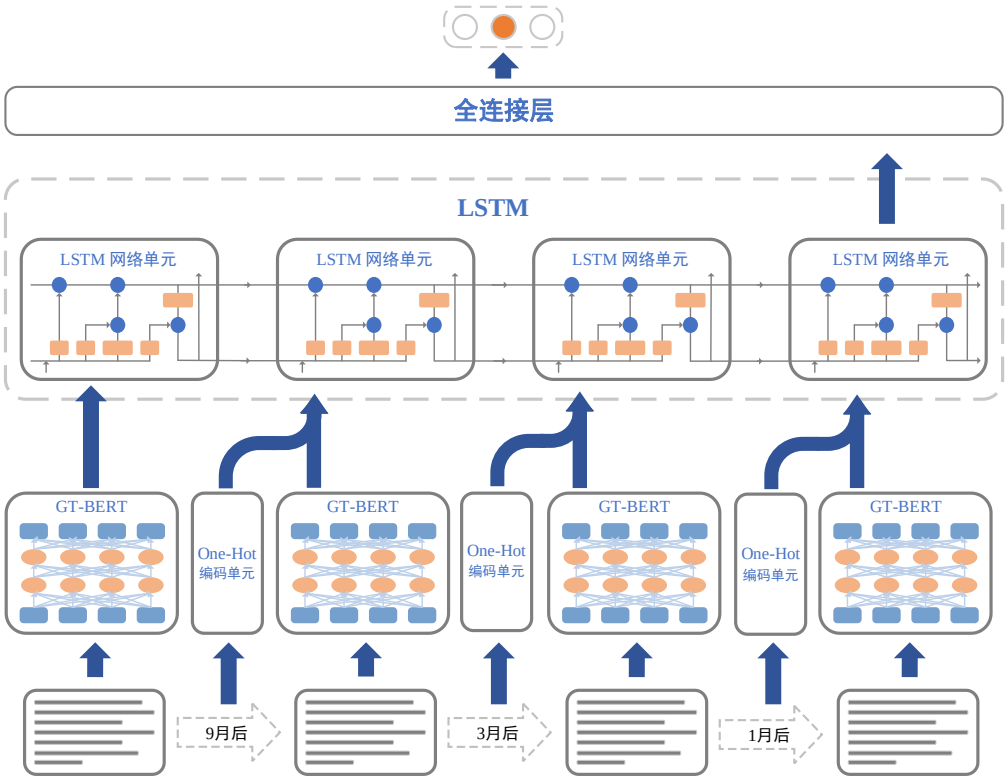


图 4-1 基于胃镜诊断文本时序特征的患者胃癌风险预测网络

### 4.3.2 面向胃镜诊断文本的 BERT 预训练模型 GT-BERT

本章所涉及的领域是自然语言处理中的医学文本处理，这一领域对于模型的准确性和专业性要求极高。然而，由于 BERT 的预训练语料主要来自通用领域，通用的 BERT 预训练语言模型在处理医学领域的文本时可能效果不佳[48, 49]。

为了解决这一问题，本章充分利用了齐鲁医院自 2012 年至 2022 年间累积的大量肠胃镜文本数据。首先将这些文本数据按照句号和换行符进行切分，得到了大量独立的句子作为 BERT 的后训练语料。接着对通用的 BERT 预训练语言模型进行了后训练，以适应医学领域的特殊需求。在后训练的过程中，设置了训练批量大小为 8，并将学习率设置为$2e^{-5}$，其余参数与谷歌在进行 BERT 预训练时设置的参数保持一致。经过 25000 步的后训练，得到了一个在肠胃镜文本领域具有更好适应性的预训练语言模型，即 GT-BERT。

设通用 BERT 预训练语言模型为$BERT$，经过后训练得到的适应医学领域的 BERT 为$GT-BERT$给定齐鲁医院肠胃镜文本数据集$D$，其中包含了$n$个样本，将其表示为：

$$D = \{s_1, s_2, \dots\dots, s_n\} \tag{4-1}$$

其中，$s_i$表示第$i$个样本，每个样本由多个句子组成。将每个样本$s_i$按句号和换行符进行切分，得到独立的句子集合$S_i$，即：

$$S_i = \{s_{i1}, s_{i2}, \dots\dots, s_{iM_i}\} \tag{4-2}$$

其中，$M_i$表示第$i$个样本中的句子数。接下来使用$BERT$模型在$D$上进行训练，以得到后训练的语言模型。经过$T$步的后训练后，得到了适应医学领域的后训练语言模型$GT-BERT$。后训练的过程可以表示为：

$$GT-BERT = Pretrain(BERT, D, T) \tag{4-3}$$

GT-BERT 的出现使得模型能够更准确地理解和处理医学领域的肠胃镜文本，为后续的胃癌风险预测任务提供了更可靠的基础。图 4-2 展示了 GT-BERT 在语料库上的后训练过程。这一创新为本章的实验结果提供了坚实的基础。

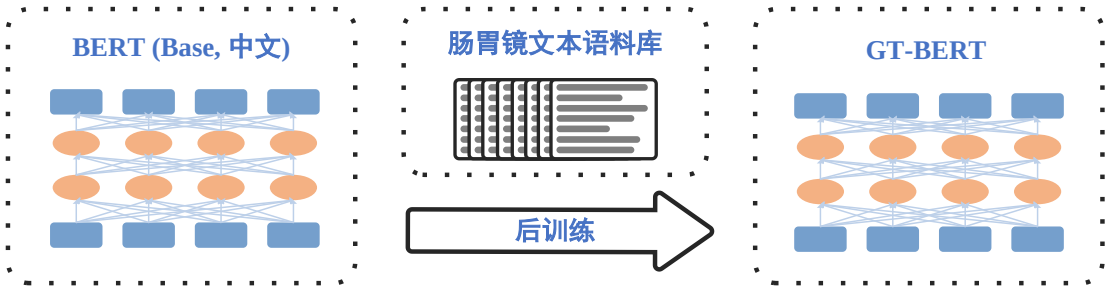


图 4-2 BERT 预训练模型的后训练

### 4.3.3 基于时间间隔特征的 One-Hot 编码模块

在时序数据分析领域，通常假设时间间隔是一致的，因此很少有研究将时间间隔特征直接融入到编码器中。然而，胃镜诊断文本数据与常见的时序数据有所不同。在这些数据中，患者多次接受胃镜检查的时间间隔通常是不一致的，而这些间隔又是医生评估患者病情进展的重要依据。因此，本章设计了一种专门用于处理时间间隔特征的 One-Hot 编码模块，以将这一信息融入到编码器中[50]。

该 One-Hot 编码模块旨在将连续的时间间隔特征映射到 8 个固定离散的时间点。在这项研究中，采纳了齐鲁医院医生的建议，并结合了医生们的临床诊断经验。将这 8 个固定的时间点设置为一个月、两个月、三个月、六个月、一年、两年、三年和五年。对于每位患者的每两次胃镜检查记录，将其时间间隔划分到与其最近的固定时间点，并将对应位置的编码设置为 1，其他位置设置为 0。具体的实现方法如算法 4-1 所示。

给定时间间隔$\Delta t$其被映射到固定时间点的编码器$One-Hot(\Delta t)$可以表示为：

$$One-Hot(\Delta t)_i = \begin{cases} 1 & \text{if } i = argmin_j\left(\left|\Delta t - t_j\right|\right) \\ 0 & \textit{其他情况} \end{cases} \quad (4-4)$$

其中，$t_j$表示第$j$个固定的时间点，$|\Delta t - t_j|$表示时间间隔$\Delta t$与固定时间点$t_j$之间的绝对差值，$argmin_j$表示绝对差值最小的$j$。这样对于给定的时间间隔$\Delta t$，其对应的 One-Hot 编码向量$One-Hot(\Delta t)$的长度为固定时间点的个数，其中只有一个元素的值为 1，其余元素的值为 0。

这样，模型可以更好地利用时间间隔信息，从而提高了对患者病情变化的理解和预测能力。

### 4.3.4长短期记忆网络 LSTM

尽管近年来针对时序数据的特征提取任务，Bi-LSTM（双向长短期记忆网络）取得了一定的成功，然而，考虑到本章所面向的胃镜诊断文本数据仅具有单向的依赖特征，而不是双向依赖，因此选择了传统的单向 LSTM 网络作为时序特征的提取器。LSTM 网络以其出色的时序特征感知能力而闻名，适用于许多时序数据的处理任务。如图 4-3 所示，两条胃镜诊断文本的间隔时间被 One-Hot 编码模块编码后，与胃镜诊断文本的特征向量连接并输入到 LSTM 网络中。这样，LSTM 网络不仅能够感知每条胃镜诊断文本的语义特征，还能同时感知到两条胃镜诊断文本之间的时间间隔特征。这种设计能够使得模型更全面地理解患者胃部病变的发展趋势，从而更准确地预测患者的胃癌风险。