

# Three-Branch BERT-Based Text Classification Network for Gastroscopy Diagnosis Text

Zhichao Wang<sup>1,2</sup>, Xiangwei Zheng<sup>1,2</sup> ✉, Jinsong Zhang<sup>1,2</sup>, and Mingzhe Zhang<sup>1,2</sup> ✉

## ABSTRACT

During a hospital visit, a significant volume of Gastroscopy Diagnostic Text (GDT) data are produced, representing the unstructured gastric medical records of patients undergoing gastroscopy. As such, GDTs play a crucial role in evaluating the patient's health, shaping treatment plans, and scheduling follow-up visits. However, given the free-text nature of GDTs, which lack a formal structure, physicians often find it challenging to extract meaningful insights from them. Furthermore, while deep learning has made significant strides in the medical domain, to our knowledge, there are not any readily available text-based pre-trained models tailored for GDT classification and analysis. To address this gap, we introduce a Bidirectional Encoder Representations from Transformers (BERT) based three-branch classification network tailored for GDTs. We leverage the robust representation capabilities of the BERT pre-trained model to deeply encode the texts. A unique three-branch decoder structure is employed to pinpoint lesion sites and determine cancer stages. Experimental outcomes validate the efficacy of our approach in GDT classification, with a precision of 0.993 and a recall of 0.784 in the early cancer category. In pinpointing cancer lesion sites, the weighted F1 score achieved was 0.849.

## KEYWORDS

gastroscopy diagnostic text; Bidirectional Encoder Representations from Transformers (BERT); text classification

Gastric cancer is a malignant tumor that originates mainly in the epithelial cells of the gastric mucosa. In China, a country with a high incidence of gastric cancer, the incidence and mortality rates of this cancer rank the second and third among malignant tumors, respectively<sup>[1]</sup>. Because early-stage gastric cancer often presents no distinct symptoms, and the occasional non-specific symptoms like epigastric discomfort and belching can be easily confused with conditions such as gastritis or gastric ulcers, the disease often goes unnoticed. Consequently, the diagnosis rate for early-stage gastric cancer remains relatively low in China and other countries<sup>[2]</sup>.

As shown in Fig. 1, the Gastroscopy Diagnostic Text (GDT) is a text report written by physicians following a gastroscopy

diagnosis. A comprehensive GDT typically describes the location, shape, size, color, surface condition, and borders of the lesion. However, because GDTs are manually written by physicians, they might contain customary terms or subjective descriptions. Additionally, due to the variety of lesions, the same lesion can manifest in different ways. In particular, the descriptions of cancer-related lesions are highly varied, making it challenging to extract relevant information using rule-based methods such as regular expressions.

During hospital gastroscopy, each hospital generates a large amount of GDTs. Often, these GDTs are only used to explain the condition of patients, rather than for data mining and condition statistics, which has led to the accumulation of a large number of

**Esophageal:** The mucosa is smooth, with clear markings of submucosal blood vessels and normal contractions and peristalsis. The distance between the cardia and the incisor is 43 cm. There is an ulcer on the distal posterior wall of the cardia, and the surrounding mucosa is significantly elevated and rough. Take 4 biopsies, the texture is slightly tough and brittle, which is prone to bleeding. The dentate line is slightly unclear.

**Gastric fundus:** The mucosa is congested and edematous. The gastric juice is clear with moderate volume. In the gastric fundus near to cardia, a mucosal protrusion near the lesion. A 0.4 × 0.4 cm submucosal protrusion can be seen at the bottom of the stomach, with a smooth surface.

**Gastric body:** The mucosal is congested and edematous. The folds are regular.

**Gastric antrum:** The mucosa is significantly congested and edematous, with redness and paleness alternating, and a mottled change. The peristalsis is good. The mucosa at the pyloric entrance and anterior pyloric area is congested and edematous. There is a mucosal protrusion. Take 4 biopsies, the texture is crisp and is prone to bleeding.

**Duodenal bulb and descending:** The mucosa are smooth, with small patches of mucosal congestion and edema visible.

(Cardia) Adenocarcinoma. (Pylorus) Focal intramucosal carcinoma. (Gastric angle) Chronic atrophic gastritis, mild, with intestinal metaplasia, active stage.

Fig. 1 Example of diagnostic gastroscopy text (GDT).

<sup>1</sup> School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

<sup>2</sup> State Key Laboratory of High-end Server & Storage Technology, Jinan 250013, China

Address correspondence to [Xiangwei Zheng, xwzhengcn@163.com](mailto:Xiangwei Zheng, xwzhengcn@163.com); [Mingzhe Zhang, zmz420@126.com](mailto:Mingzhe Zhang, zmz420@126.com)

© The author(s) 2024. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

underutilized GDTs in each hospital. The information on patient conditions implied in these GDTs is of high research value, which can be used in many ways, such as statistics of patients, providing personalized treatment plans. However, GDTs are free-text in nature, it is difficult to obtain valuable information from the original GDT. Therefore, using Natural Language Processing (NLP) techniques to parse these GDTs can effectively extract valuable information. In addition to being used to benefit patients, processing GDTs by NLP can also reduce the labor cost of hospitals and improve efficiency. Nevertheless, to our knowledge, there are few studies on GDT using natural language processing techniques, not to mention using deep learning techniques to process GDTs.

In 2019, Google innovatively used the coding layer accumulation of 12 or 24 layers Transformer<sup>[3]</sup>, and designed two language tasks to train to form a self-coding language model Bidirectional Encoder Representations from Transformers (BERT)<sup>[4]</sup>. BERT can perform concurrently compared to the originally used Recurrent Neural Network (RNN)<sup>[5]</sup>, and Long Short-Term Memory (LSTM)<sup>[6,7]</sup>. It also incorporates transformer's multi-head self-attention, which can better extract the features of sentences. Since the [CLS] tag is added to the first part of each input sentence of BERT, the vector output from [CLS] through BERT contains the semantic information of the whole sentence, and is often used in text classification tasks<sup>[8]</sup>. However, traditional BERT-based text classification algorithms can only achieve multiple classification tasks with no association between tags. The need of extracting patients' cancer sites and disease grading based on the description of endoscopic finding text and pathology report text explored in this paper can hardly be designed as a traditional text classification task based on BERT pre-trained language model.

To accurately classify GDT and pinpoint the location of lesions in the stomach, our text classification model necessitates a single input with multiple outputs. The conventional method channels the global feature vectors produced by the encoder into numerous feature decoders, yielding multiple classification outcomes. While this method has been validated across several studies, it is not without its limitations. Given that each decoder receives identical input, branches with distinct functions struggle to tap into the robust feature representation potential of the pre-trained model.

Unlike traditional natural language texts, GDTs has clear boundaries for the description of each part of the stomach. For texts with such characteristics, we can extract local and global feature vectors of the text separately to better construct a model with one input and multiple outputs.

To solve the above problems, this study proposes a three-branch classification network to deal with GDTs. For the encoding layer of the network, our method innovatively exploits the clause encoding capability of the BERT pre-trained model, using specialized input to guide the BERT pre-trained model to output both the global feature vectors and multiple local feature vectors. The decoding layer of the network consists of three fully connected layers to process the endoscopic seen text and the pathological diagnostic text, respectively. Specifically, this network can classify GDTs into three categories: early-stage cancer, progressive cancer, and others. At the same time, the network can determine the site of cancer lesions. The experimental results show that our network outperforms other existing methods in terms of classification accuracy and cancer recall, which provides an approach to help hospitals use GDTs more effectively and efficiently.

## 1 Related Work

In recent years, the field of text classification and analysis has witnessed significant advancements due to the rapid development of deep learning technology. In this section, we will present a comprehensive overview of the research studies conducted in the areas of text classification, medical text classification, and GDT classification, organized in chronological order according to the evolution from machine learning approaches to deep learning methods.

### 1.1 General text classification

In earlier studies, traditional machine learning methods were commonly employed in text classification tasks. Traditional machine learning methods like Naive Bayes<sup>[9]</sup>, Support Vector Machine (SVM)<sup>[10]</sup>, decision trees<sup>[11]</sup>, random forests<sup>[12]</sup>, and k-NN<sup>[13]</sup> relied on features like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Although these methods have achieved success in many text classification tasks, their performance has been increasingly challenged by the advancements in deep learning and pre-trained models.

In recent years, deep learning based methods have achieved great success in the field of text classification and analysis<sup>[14]</sup>. Many studies use deep learning methods for medical text classification and analysis. For example, one study used Convolutional Neural Networks (CNN) to classify medical texts<sup>[15,16]</sup>, and the results showed that the method performed well on the clinical text classification task<sup>[17]</sup>.

With the development of pre-trained models, methods based on pre-trained models have achieved remarkable success in text classification tasks. In particular, BERT was introduced by Google in 2018, employing a transformer architecture and bidirectional context for training, setting a new performance benchmark for natural language processing tasks. Following this, Facebook AI proposed Robustly optimized BERT (RoBERTa), which improved upon BERT by using larger batches, longer sequences, more data, and removing the next sentence prediction task, further enhancing the performance of natural language processing tasks<sup>[18,19]</sup>. In addition, Generative Pre-trained Transformer (GPT) developed by OpenAI has also achieved significant results in text classification tasks, albeit with a unidirectional context for training<sup>[20,21]</sup>. In these approaches, pre-trained models are first trained to learn general language representations and then applied to specific text classification tasks by adding a classification layer on top of the models and fine-tuning them using task-related data.

### 1.2 GDT classification

There are also some related works in GDT classification and analysis. Another study uses a fusion of multiple traditional machine learning methods to classify GDTs and compares them with a single traditional machine learning method, and the results show that the method using a fusion of multiple machine learning methods has better classification results.

In addition to medical text classification tasks, several studies have explored the application of deep learning based methods for the fusion analysis of GDTs and gastroscopic diagnostic images. For example, one study used multimodal semantic fusion for gastric cancer screening via a hierarchical attention mechanism<sup>[22]</sup>, which proposed a novel Intelligent Decision support method for Gastric Cancer Screening (ID-GCS), ID-GCS extracts text semantics from multimodal gastroscopy reports using a hybrid attention mechanism and performs semantic fusion to integrate the semantics of textual gastroscopy reports and images to

improve the gastroscopy. The results are interpretable. Experimental results show that this method performs well for gastroscopy screening.

Although existing work has achieved some success in medical text and GDTs classification and analysis, no studies have been seen to apply text pre-trained models to GDTs classification and analysis in today's widely used text pre-trained models. Therefore, this study focuses on how to make full use of the knowledge of pre-trained models and apply pre-trained models to the classification and analysis of GDTs.

## 2 Material and Method

In this section, we reveal some details of the data and methods used for the experiments, including the acquisition and pre-processing of the dataset, and the detailed structure of the proposed model.

### 2.1 Dataset

**Dataset status.** The dataset used in this study was obtained from the database of early-stage cancer screening cases at Qilu Hospital, Shandong University, and included 2618 GDTs recorded by physicians during gastroscopic examinations of actual patients. The GDTs comprised of endoscopic seen texts and pathological diagnostic texts. Each GDT included an endoscopic seen section, which provided a detailed description of the pathological changes in the stomach tissue observed by the physician during the gastroscopic examination of the patient. The pathological diagnosis section involved microscopic examination and evaluation of tissue specimens to determine the nature, type, extent, and severity of abnormal tissue, thereby assisting physicians in determining the type and severity of the disease.

**Dataset pre-processing and annotation.** The main purpose of dataset preprocessing is to cleanse, standardize, convert, and format the raw text data. In this study, we screened and cleaned the raw GDT data, eliminated data that were corrupted or deemed unusable, such as test data from the consultation system, stripped HTML tags, extra spaces, and other symbols that may impact model training and inference. In the end, we obtained a dataset of 2459 GDTs that were suitable for model training.

To ensure the accuracy of the text labels, graduate students and practicing doctors from Qilu Hospital of Shandong University were invited to perform text labeling. Finally, a total of 2459 GDTs were obtained, labeled with both sequence labels and document labels. Of these, 1894 GDTs contained both the pathological diagnosis section and the endoscopic seen section. The remaining 565 GDTs contained only the endoscopic seen section.

Sequence labels were used to mark the sequence of text describing the cancer lesion. The sequence labels for the endoscopic seen section were for suspected early-stage gastric cancer and suspected progressive gastric cancer. The sequence labels for the pathological diagnosis section were for pathologically confirmed early-stage gastric cancer and pathologically confirmed progressive gastric cancer. Each sequence label was marked with the beginning and end position in the original document.

Document label is designed to mark the gastric lesions of the patient to which this GDT belongs. Each marked GDT contains a unique document label, which includes three categories: other, early-stage gastric cancer, and progressive gastric cancer.

### 2.2 Method

To determine cancer lesion sites and stages from GDT, we introduce a three-branch classification network as illustrated in

Fig. 2. Initially, a keyword-driven text-slicing approach segments the GDT into distinct text units based on intragastric site keywords. We then employ a BERT-SUM-inspired encoding, adding [CLS] and [SEP] tokens to the start and end of these units to capture local GDT features. Leveraging a Chinese BERT-base model pre-trained by Google, we encode the GDT for feature extraction. The feature vectors linked to the [CLS] tokens from BERT feed into a three-branch classification network, which classifies based on the GDT's global, endoscopic, or pathological diagnostic segments. The network's final output provides detailed information on the lesion sites and cancer stages.

It is worth mentioning that the network structure proposed in this study is compatible with most BERT pre-trained models, such as ALBERT and RoBERTa. We can use different pre-trained models for different application scenarios. For example, in scenarios with high real-time requirements such as assisting doctors in diagnosis, we can use small models like ALBERT to improve inference speed.

**Keyword-based text-slicing method for GDT.** This study aims to simultaneously extract the cancer stage and lesion site requirements of patients from GDT, requiring that the GDT is divided into text units according to intragastric sites in the text pre-processing stage. Since the GDTs are written based on different intragastric sites, the descriptions of lesions at the same sites are usually continuous. This continuity allows for the possibility of segmenting GDTs using keywords. Effective text unit slicing methods are important for improving the accuracy of algorithm recognition. How to slice GDT correctly and efficiently has become an urgent problem to be solved.

In order to achieve effective slicing of GDT, a GDT slicing method based on intra-gastric sites keywords is proposed in this study, as shown in Fig.3 and Algorithm 1. Keywords for intragastric sites include esophagus, cardia, fundus, body, antrum, and duodenum. For the slicing of GDT, first, the original text was split into several text segments, with all periods and line breaks as segmentation points, and then each text segment was retrieved to determine whether they contained intragastric site key-words. Finally, the text segments that do not contain intragastric site keywords are merged into the text segments that contain intragastric site keywords before the original location. Each merged new text segment is a text unit. Among them, the text unit from the endoscopic seen section is referred to as the endoscopic seen text unit, and the text unit from the pathological diagnosis section is referred to as the pathological diagnosis text unit.

In this study, 100 GDTs were randomly selected from the GDT dataset and sliced using the above keyword-based text slicing method for GDT, and a total of 654 text units were obtained, which were judged by the graduate students and practicing doctors in Qilu Hospital. A total of 651 text units were correctly classified, with an accuracy rate of 99.54% (three of the text units were incorrectly identified due to abbreviations and other writing methods used for the keywords of intragastric sites). The experiment shows that the above keyword-based text slicing method for GDT is effective and accurate.

**BERT-SUM-like text encoding format.** BERT-SUM is an advanced text summarization model that leverages the power of BERT to generate high-quality extractive summaries. By incorporating BERT, a state-of-the-art pre-trained language model, BERT-SUM can better understand the context and semantics of the input text to produce more coherent and richer summaries<sup>[23]</sup>. To achieve the need of extracting both global features of GDT and local features of each text unit, this study was inspired by BERT-SUM and used the text encoding format of BERT-SUM type for the GDT classification task.

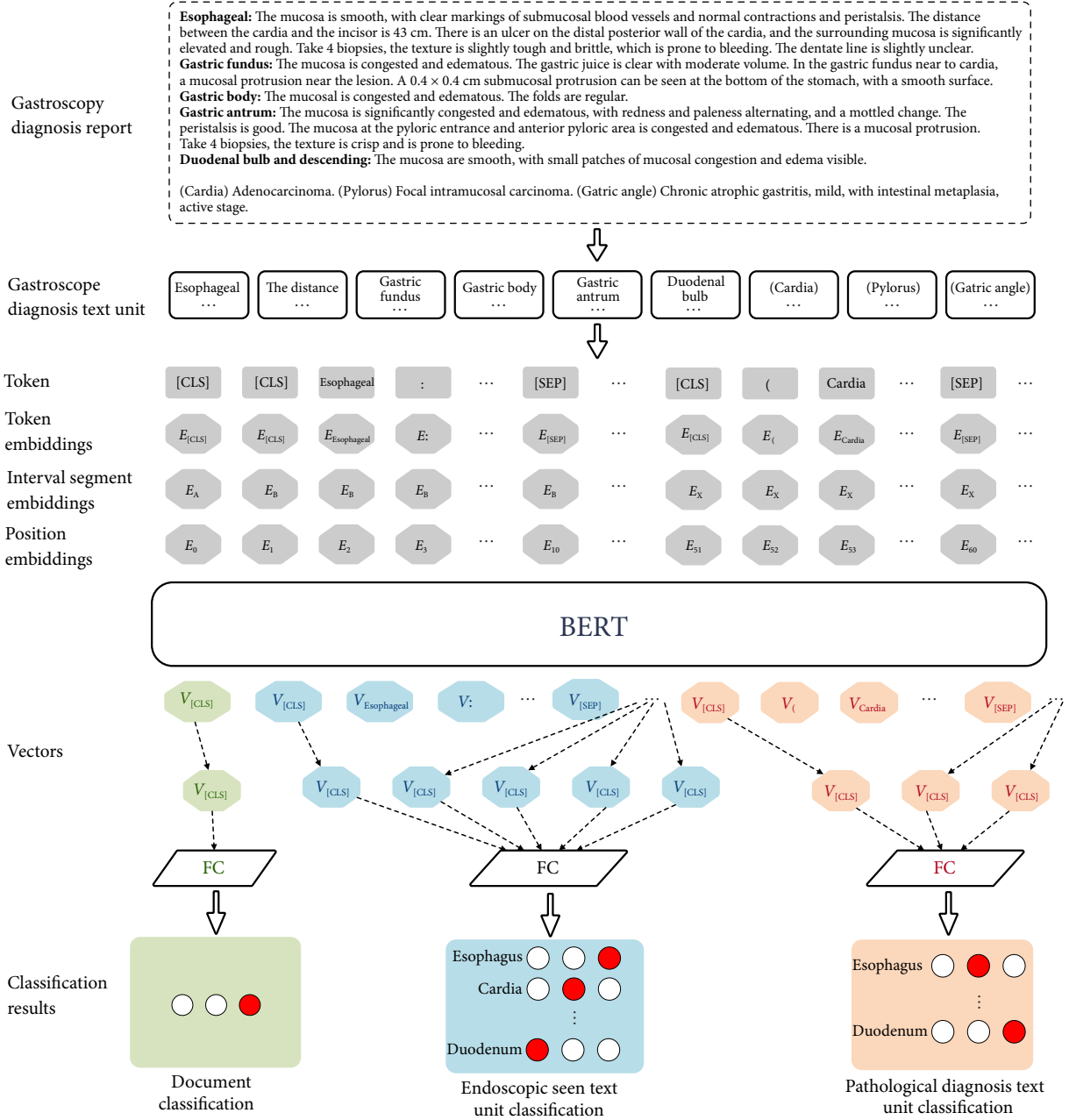


Fig. 2 Three-branch classification network for GDT classification.

We incorporated the [CLS] token at the beginning of each text unit and the [SEP] token at the end of each text unit. In the BERT model, the [CLS] token serves to extract features from the entire text; in this study, we utilize it to extract features from individual text units. Subsequently, we concatenate all text units and introduced another [CLS] token at the beginning of the sequence to extract global features from the combined sentence.

**Three-branch classification network.** The decoder part of the network uses three linear layers to downscale and classify three types of feature vectors: global feature vectors, endoscopic seen feature vectors, and pathological diagnosis feature vectors, respectively.

The classifier used for dimensionality reduction classification of the global feature vectors is a conventional simple linear classifier.  $T \in \mathbb{R}^{L \times D}$  is the output of BERT,  $W_a \in \mathbb{R}^{D \times 3}$  denotes the trainable parameters of the linear layer, and the output of the classifier  $Y_a$  is

$$Y_a = \sigma(W_a T_1 + b_a) \quad (1)$$

where  $\sigma$  is the softmax function.  $b_a$  is the bias of this linear layer.

The classifier used for dimensionality reduction classification of the endoscopic seen feature vectors is a temporal layer nested on top of the traditional simple linear classifier, allowing it to classify two-dimensional feature vectors simultaneously. The [CLS] token corresponding to the endoscopic seen text is indexed in the original text as  $I_g \in 1, 2, \dots, L^k$ ,  $K$  is the number of endoscopic seen text units,  $W_b \in \mathbb{R}^{D \times 3}$  denotes the trainable parameters of the linear layer, and the output  $Y_b$  of the classifier is

$$V_g = \text{gather}(T, I_g) \in \mathbb{R}^{K \times D} \quad (2)$$

$$Y_b = \text{timedistributed}[V_g, f(x) = \sigma(W_b x + b_b)] \in \mathbb{R}^{K \times 3} \quad (3)$$

The role of gather is to extract the rows specified by the index from the matrix and arrange them into a new matrix, timedistributed is to operate on each row of the two-dimensional matrix separately according to the time dimension and return a

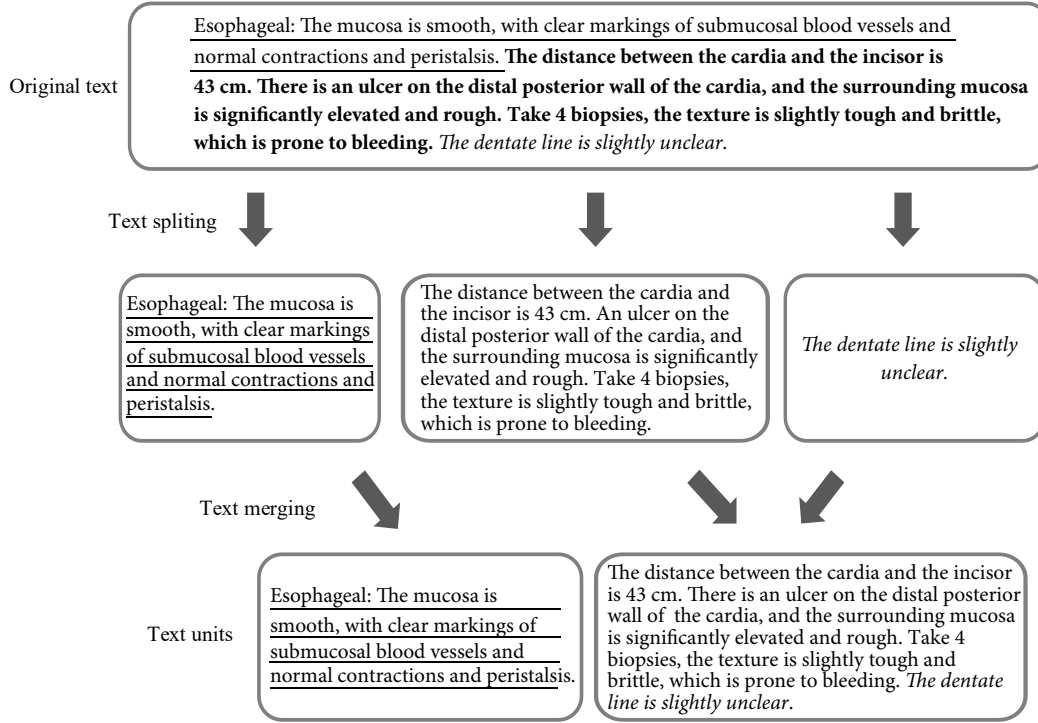


Fig. 3 Keyword-based text slicing method for GDT.

**Algorithm 1** Keyword-based text slicing method for GDT

```

Input: gastroscopie diagnosis text gdt
Output: text unit list text_unit_list

1 gdt = "Esophageal: The mucosa is smooth..."
2 stomach_sites = ["Esophageal", "cardia", "Gastric fundus",
  "Gastric body", "Gastric antrum", "Duodenal bulb and descending"]
3 text_unit_list = []

  // Split the original GDT text by periods and line breaks
4 gdt = split(["\n", "."], gdt)
5 for text_part in gdt:
6   for site in stomach_sites:
7     if site in text_part:
        // Text segment containing the intragastric site keyword
8       text_unit_list.append(text_part) #
9     else if text_unit_list:
        // The text segment does not contain the intragastric site
        keyword
10      text_unit_list[-1].concat(text_part)
11    end if
12  end for
13 end for

```

new two-dimensional matrix.

The classifier used for dimensionality reduction classification of feature vectors of pathological diagnostic text units is similar to the classifier used for feature vectors of endoscopic seen text. The [CLS] token corresponding to the pathological diagnostic text units is indexed in the original text as  $I_p \in 1, 2, \dots, L^M$ ,  $M$  is the number of pathological diagnostic text units,  $W_c \in \mathbb{R}^{D \times 3}$  denotes the trainable parameters of the linear layer, and the output  $Y_c$  of the classifier is

$$V_p = \text{gather}(T, I_g) \in \mathbb{R}^{M \times D} \quad (4)$$

$$Y_c = \text{timedistributed}[V_p, f(x) = \sigma(W_c x + b_c)] \in \mathbb{R}^{M \times 3} \quad (5)$$

**Output of classification network.** In our method, we use a linear layer as a classifier in all three branches to transform the output of the model into a set of vectors of category scores. In the testing phase, the linear layer takes the input data and generates a vector of scores for each category, which can be interpreted as the confidence level of the model for each category. We then transform this vector into a probability distribution for all categories by means of a softmax function. The predicted category labels are the categories with the highest probability in the output vector of the softmax function, the category labels with the highest probability values. For a given vector  $x \in \mathbb{R}^n$ , the softmax function is defined as

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}} \quad (6)$$

where  $n$  is the dimension of vector  $x$ . In this paper, it refers to the number of categories of cancer staging.

The model uses categorical crossentropy as a function to calculate the loss during training. We use one-hot encoder to encode the true labels of the training data in order to facilitate the calculation of the loss. In this study, the classifier outputs of the three network branches are  $Y_a^i$ ,  $Y_b^i$ , and  $Y_c^i$ . The label vectors after one-hot encoding of the true labels are  $P_a^i$ ,  $P_b^i$ , and  $P_c^i$ .  $i$  ranges from 1–3, indicating the three categorical categories. Loss is formalized as

$$\text{Loss} = - \sum_{i=1}^3 \begin{bmatrix} P_a^i \\ P_b^i \\ P_c^i \end{bmatrix} \cdot \log \left( \begin{bmatrix} Y_a^i \\ Y_b^i \\ Y_c^i \end{bmatrix} \right) \quad (7)$$

### 3 Experiment

In this section, we will discuss the training and prediction of a

three-branch BERT-based text classification network for GDT, as well as a comparison with a simple BERT-based text classification method for the baseline model, and we will conclude with an analysis of our experimental results.

### 3.1 Experimental environment

The proposed method is implemented using Tensorflow, keras<sup>[24]</sup>, and bert4keras, where the BERT pre-trained model is a Chinese RoBERTa with 24 hidden layers, 768 dimensions in the output layer, with Ubuntu 20.04 as the operating system, Intel Xeon Silver CPU 4208, Python version 3.7, 128 G memory, and GPU computing card NVIDIA RTX3090.

### 3.2 Model training and fine-tuning

In this study, we used the BERT pre-trained model for training and fine-tuning. Specifically, we fine-tuned all the parameters in the BERT model to fit our GDT classification and analysis task. Also, we randomly initialized the parameters of the linear layers of the three network branches and involved them in the training process. This allows the network to be better adapted to our dataset and improves the classification performance of the proposed model.

In the training process, we used the common Adam optimizer. Some common hyperparameter settings, such as learning rate, batch size, and maximum sequence length, were used to improve the training of the proposed model. Specifically, we used an initial learning rate of  $5 \times 10^{-5}$ , a batch size of 8, and a maximum sequence length of 512. Also, we set some regularization terms, such as L2 regularization, and dropout to avoid overfitting the proposed model.

We split the training dataset into a training set and a validation set as shown in Table 1, where the training set is used to train the model, and the validation set is used to monitor the training effectiveness of the model and prevent overfitting. At the end of each training round, we evaluate the performance of the model on the validation set in order to stop training when the model performance stops improving.

### 3.3 Experimental procedure and result

**Evaluation metrics.** We trained on an NVIDIA RTX3090 GPU and the proposed model converged at the 18th epoch with a loss of 0.023, achieving the best precision on the validation set. We evaluated the cancer staging classification ability of proposed method on the validation set and used some common evaluation metrics, such as precision (depending on early-stage cancer and progressive cancer as positive samples), recall (depending on early-stage cancer and progressive cancer as positive samples), and F1 value.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

$$F1 = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (10)$$

True Positive (TP) represents the number of instances that the model accurately identifies as relevant or positive outcomes. False Positive (FP) occurs when the model incorrectly predicts instances as positive, meaning it classifies irrelevant items or outcomes as relevant. Conversely, False Negative (FN) represent instances where the model incorrectly predicts a negative outcome, failing to identify relevant items or outcomes.

Precision represents the proportion of the sample for which the model predicts positive cases that are true positive. Recall indicates the proportion of samples that are true positive cases that are predicted by the model to be true positive. F1 is the summed average of the precision and recall rates, which is used to evaluate the precision and recall ability of the model in a comprehensive way. In addition, we added early-stage cancer recall as an additional evaluation metric for practical applications. Early-stage cancer recall is defined as the ratio of data labeled as early-stage cancer and correctly predicted to all data labeled as early-stage cancer, and this evaluation index has high practical clinical significance due to the low early identification rate of gastric cancer in clinical practice.

The BERT-based three-branch classification network proposed in this study was able to extract the intragastric sites in the text describing the patient's lesions while classifying the cancer stage of the GDT. In this study, we used the BERT-based multi-label classification network to compare the ability of the present model in extracting the intragastric sites in the text describing the patient's lesions. We use the  $F1_w$  (weighted F1 scores) of the multi-label classification task as an evaluation metric to evaluate the performance of the model.

$$F1_w = \frac{1}{\sum_{i=1}^N \omega_i} \sum_{i=1}^N \omega_i \cdot \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \quad (11)$$

where  $N$  is the number of tags,  $P_i$  and  $R_i$  denote the precision and recall of the  $i$ -th tag, respectively, and  $\omega_i$  denotes the frequency of the  $i$ -th tag in the dataset, the proportion of the number of samples containing the tag in the sample to the total number of samples.

**GPT-3.5 turbo.** With the development of deep learning technology, Large Language Models (LLMs) represented by ChatGPT have emerged. Thanks to the generative language model structure used by ChatGPT and the integration of training data from various fields, ChatGPT can achieve state-of-the-art performance in tasks across multiple domains without additional training. This study compared the core model GPT-3.5 of ChatGPT with our method by calling the Application Programming Interface (API) of openAI. We have designed the following prompt: "You are an AI model for analyzing gastroscopy diagnostic text. I will provide you with a section of gastroscopy diagnostic text. Please provide my patient's gastric cancer staging (early-stage cancer, progressive cancer, others) and

Table 1 Dataset detail.

| Data category      | Number of training sets | Number of validation sets | Total |
|--------------------|-------------------------|---------------------------|-------|
| Early-stage cancer | 246                     | 53                        | 299   |
| Progressive cancer | 352                     | 82                        | 434   |
| Other              | 1369                    | 357                       | 1726  |
| Total              | 1967                    | 492                       | 2459  |

affected sites (esophageal, cardia, gastric fundus, gastric body, gastric antrum, duodenal bulb, and descending). Please return it to me in the following form: {cancer staging: early-stage cancer, affected site: esophageal, cardia}.” After obtaining the return, compare the output of GPT-3.5 with the true value of the test data, and evaluate the model using the same evaluation indicators.

**Ablation experiment.** In this study, we encode the input to BERT using the BERT-SUM-like text encoding format to achieve a single-input-multi-output mechanism. To investigate the effectiveness of such an input format, we compare it with the conventional Single-Input-Multi-Output Classification Network based on BERT pre-trained models (SIMO-BERTCN). For SIMO-BERTCN, after obtaining the global feature vector from BERT’s output, we feed it into seven fully connected layers. We derive the cancer staging from the output of the first fully connected layer and obtain the lesion status of six different stomach regions from the remaining six fully connected layer. After obtaining the output, compare the output of SIMO-BERTCN with the true value of the test data, and evaluate the model using the same evaluation indicators.

**Experimental result.** In this experiment, we compare the classification abilities of various models in determining gastric cancer staging based on a single text category output. The evaluated models include CNN and LSTM for text classification, a traditional BERT-based text Classification Network (BERTCN), and our proposed model. We also evaluated the capability of various models to identify intragastric sites from texts describing patient lesions. Specifically, we compare the performance of the BERT-based Multi-Label Text Classification Network (ML-BERTCN), GPT-3.5 turbo, SIMO-BERTCN, and our proposed model. The experimental results are shown in Table 2.

The model proposed in this study uses an improved input structure to achieve BERT’s ability to classify gastric cancer staging in GDTs while recognizing the location of gastric lesions. We compare the performance with the traditional BERT-based text classification method. In the experiments, we used the GDT dataset provided by Qilu Hospital. The experimental results show that the precision of our method on the cancer staging classification task is 0.993, which is on par with the traditional BERT-based text classification method. This indicates that our method can extract cancer lesion sites without affecting the performance of cancer staging and classification.

We also compared the performance of our method and ML-BERTCN on the lesion sites extraction task of cancer. The experimental results show that our method outperforms the BERT-based multi-label text classification method on the cancer site extraction task and achieves better performance. Specifically, we use the weighted F1 score commonly used in multi-label

classification tasks as the evaluation metric and show that our method achieves a weighted F1 score of 0.849 for cancer site extraction, compared with 0.724 for BERT-based multi-label text classification methods. This indicates that our method can better extract the lesion sites of cancer, which can help doctors to better determine the patients’ conditions in practical applications.

At the same time, we compared our model with GPT-3.5. GPT-3.5, as a generative pre-trained LLM, achieved state-of-the-art effects in tasks such as reading comprehension and question answering. However, for GDT classification and analysis tasks, GPT-3.5 has not shown encouraging results due to its inability to achieve effective fine-tuning.

In our ablation study, we omitted the specialized input module for GDT from our architecture and instead utilized the global feature vectors produced by BERT, subsequently reducing their dimensionality to yield multiple outputs. This was done to assess the efficacy of the BERT-SUM-like text encoding format. The experimental results indicate that a single BERT global vector is difficult to achieve both cancer stage classification and gastric lesion location recognition. The enhancements we introduced proved to be both effective and dependable.

## 4 Conclusion

In this paper, we have presented a three-branch BERT-based text classification network for GDT to classify and analyze GDTs. The proposed method achieves two tasks gastric cancer staging classification and intragastric sites information extraction. The experimental results show that our proposed method achieves 0.993 precision in gastric cancer staging classification, which is on par with the effect of the traditional BERT method. But in the intragastric sites information extraction task, our method outperformed the BERT-based multi-label text classification algorithm with a weighted F1 score of 0.849.

In our future work, we plan to incorporate the model with temporal features to improve the extraction of intragastric sites information. We believe that our research results can provide strong support for intelligent diagnosis and treatment in the medical field.

## Acknowledgment

This work was supported by the Key R&D Program of Shandong Province, China (Major Scientific and Technological Innovation Project) (No. 2021CXGC010506), Key R&D Program of Shandong Province, China (No. 2021SFGC0104), Natural Science Foundation of Shandong Province China (Nos. ZR2020LZH008, ZR2021MF118, and ZR2022LZH003) and National Natural Science Foundation of China (No. 62101311).

**Table 2** Experimental result of three-branch BERT-based text classification network for GDT.

| Model                | Precision    | Recall       | F1           | Early cancer recall | $F1_{\omega}$ |
|----------------------|--------------|--------------|--------------|---------------------|---------------|
| CNN                  | 0.934        | 0.926        | 0.930        | 0.496               | —             |
| LSTM                 | 0.842        | 0.855        | 0.848        | 0.428               | —             |
| BERTCN (RoBERTa)     | <b>0.993</b> | <b>0.993</b> | <b>0.993</b> | 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         |
| <b>Our (RoBERTa)</b> | <b>0.993</b> | <b>0.993</b> | <b>0.993</b> | <b>0.784</b>        | <b>0.849</b>  |

## Dates

Received: 8 September 2023; Revised: 2 November 2023;  
Accepted: 16 November 2023

## References

- [1] Y. Cui, G. Cheng, G. Tian, S. He, and Y. Yan, Secular trends in the mortality of gastrointestinal cancers across China, Japan, the US, and India: An age-period-cohort, joinpoint analyses, and Holt forecasts, *Front. Public Health*, vol. 10, p. 925011, 2022.
- [2] C. Herrera-Pariente, S. Montori, J. Llach, A. Bofill, E. Albeniz, and L. Moreira, Biomarkers for gastric cancer screening and early diagnosis, *Biomedicines*, vol. 9, no. 10, p. 1448, 2021.
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.
- [4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv: 1810.04805, 2019.
- [5] W. Zaremba, I. Sutskever, and O. Vinyals, Recurrent neural network regularization, arXiv preprint arXiv: 1409.2329, 2014.
- [6] A. Graves, Long short-term memory, in *Supervised Sequence Labelling with Recurrent Neural Networks*, A. Graves, ed. Berlin, Germany: Springer, 2012, pp. 37–45.
- [7] A. Onan and M. A. Toçoğlu, A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification, *IEEE Access*, vol. 9, pp. 7701–7722, 2021.
- [8] W. Zhou, J. Du, and X. Ren, Improving BERT fine-tuning with embedding normalization, arXiv preprint arXiv: 1911.03918, 2019.
- [9] L. Jiang, C. Li, S. Wang, and L. Zhang, Deep feature weighting for naive Bayes and its application to text classification, *Eng. Appl. Artif. Intell.*, vol. 52, pp. 26–39, 2016.
- [10] T. Joachims, Making large scale SVM learning practical, <https://api.semanticscholar.org/CorpusID:61116019>, 1998.
- [11] J. R. Quinlan, Induction of decision trees, *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
- [12] L. Breiman, Random forests, *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [13] Y. Yang and X. Liu, A re-examination of text categorization methods, in *Proc. 22nd Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Berkeley, CA, USA, 1999, pp. 42–49.
- [14] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, Deep learning: Based text classification, *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, 2022.
- [15] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, Medical text classification using convolutional neural networks, *Stud. Health Technol. Inform.*, vol. 235, pp. 246–250, 2017.
- [16] A. Onan, Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification, *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 5, pp. 2098–2117, 2022.
- [17] A. Onan, S. Korukoğlu, and H. Bulut, Ensemble of keyword extraction methods and classifiers in text classification, *Expert Syst. Appl.*, vol. 57, pp. 232–247, 2016.
- [18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv: 1907.11692, 2019.
- [19] A. Onan, Hierarchical graph-based text classification framework with contextual node embedding and BERT-based dynamic fusion, *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 7, p. 101610, 2023.
- [20] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, Improving language understanding by generative pre-training, <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>, 2018.
- [21] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 1877–1901.
- [22] S. Ding, S. Hu, X. Li, Y. Zhang, and D. D. Wu, Leveraging multimodal semantic fusion for gastric cancer screening via hierarchical attention mechanism, *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 52, no. 7, pp. 4286–4299, 2022.
- [23] Y. Liu, Fine-tune BERT for extractive summarization, arXiv preprint arXiv: 1903.10318, 2019.
- [24] N. Ketkar, Introduction to keras, in *Deep Learning with Python*, N. Ketkar, ed. Berkeley, CA, USA: Apress, 2017, pp. 97–111.