scientific reports

OPEN

Enhancing medical text classification with GAN-based data augmentation and multi-task learning in BERT

Xinping Chen[1] & Yan Du[2]✉

With the rapid advancement of medical informatics, the accumulation of electronic medical records and clinical diagnostic data provides unprecedented opportunities for intelligent medical text classification. However, challenges such as class imbalance, semantic heterogeneity, and data sparsity limit the effectiveness of traditional classification models. In this study, we propose an enhanced medical text classification framework by integrating a self-attentive adversarial augmentation network (SAAN) for data augmentation and a disease-aware multi-task BERT (DMT-BERT) strategy. The proposed SAAN incorporates adversarial self-attention, improving the generation of high-quality minority class samples while mitigating noise. Furthermore, DMT-BERT simultaneously learns medical text representations and disease co-occurrence relationships, enhancing feature extraction from rare symptoms. Extensive experiments on the private clinical datasets and the public CCKS 2017 dataset demonstrate that our approach significantly outperforms baseline models, achieving the highest F1-score and ROC-AUC values. The proposed innovations address key limitations in medical text classification and contribute to the development of robust clinical decision-support systems.

Keywords GAN, BERT, Medical text classification, Attention, Deep learning in healthcare

With the continuous advancement of smart healthcare, medical information data such as electronic health records (EHRs) has been widely adopted, providing strong support for the application of artificial intelligence (AI) in the medical field[1]. Medical text classification, which involves the automated categorization of textual data into predefined classes, plays a crucial role in various applications such as disease diagnosis, risk assessment, and medical information retrieval[2,3].

Traditional machine learning (ML) techniques have been applied in medical text classification, yet they depend heavily on feature engineering and struggle with complex semantic representations[4]. Deep learning (DL) models[5–8] have notably enhanced classification accuracy by learning hierarchical features from raw text data[9]. However, these models require large labeled datasets and often suffer from performance degradation due to the inherent imbalance in medical datasets, where rare disease cases are underrepresented[10].

In recent years, the research on medical text classification has made remarkable progress driven by pre-trained deep learning technology. Methods based on pre-trained language models can effectively acquire prior knowledge in the medical field by pre-training on large-scale medical corpora, overcoming the limitation of classic ML methods which necessitate a large volume of labeled data. Among them, the BERT (Bidirectional Encoder Representations from Transformers) model, with its bidirectional encoding mechanism and context-awareness, has demonstrated unique advantages in medical text classification tasks. This model can not only accurately understand the contextual meaning of medical terms but also effectively identify subtle differences in symptom descriptions, providing strong technical support for improving the accuracy of disease diagnosis. For instance, Liao and Bressem et al.[11,12] proposed the medical text classification model based on BERT to significantly enhance the classification performance by fine-tuning on specialized medical datasets. Gasmi et al.[13] developed an optimized deep learning model by integrating particle swarm optimization into BERT fine-tuning, improving classification performance on medical text datasets. Zhang et al.[14] introduced KG-MTT-BERT, which incorporates medical knowledge graphs to enhance multi-task learning in clinical text classification, significantly outperforming baseline models in diagnostic-related group classification tasks. Zeberga[15] proposed the RNN_

[1]College of Artificial Intelligence and Big Data, Chongqing Polytechnic university of Electronic Technology, Chongqing 401331, China. [2]Ultrasonography Department, The Fourth People's Hospital of Chongqing, Chongqing University Central Hospital, Chongqing 400014, China. ✉email: 764425821@qq.com

Bert_Based model, integrating recurrent neural networks with BERT for superior temporal feature extraction in medical text classification, achieving state-of-the-art results on Stanford Sentiment Treebank v2 (SST-2). Dai et al.,[16] addressed Chinese polyphone disambiguation within an end-to-end framework by leveraging semantic features extracted from pre-trained BERT models. The second paper (Darraz et al.[17] integrated sentiment analysis with BERT to enhance hybrid recommendation systems, improving their performance and accuracy.

Expect the limitation of labeled data, the prevalent medical data class imbalance also mainly causes performance degradation of DL in medical text classification. Generative Adversarial Network (GANs)[18,19] optimizes the model's generalization performance by generating high-quality minority class samples, providing a technical path for class imbalance problem. To address the prevalent class imbalance issue in medical data, Suresh et al.[20] explored the application of generative adversarial networks in medical data augmentation. By optimizing the architecture design of the generator and discriminator, they successfully generated high-quality minority class samples, effectively balancing the class distribution of the dataset. Bissoto et al.[21] provided a critical review of GAN-based data augmentation and anonymization techniques for skin-lesion analysis, evaluating their potential in improving model robustness and privacy in medical image analysis. In addition to GANs, multi-task learning enhances the model's ability to learn the features of the minority class by introducing auxiliary tasks, significantly improving the recognition accuracy of rare cases while maintaining the model's generalization ability. For instance, Liu et al.[22] proposed a multi-task learning approach adapted to similar tasks for mortality prediction in rare diseases, enhancing prediction accuracy across diverse conditions. Facing the task of diagnosing the Alzheimer's disease, Suk et al.[23] combined sparse method with deep multi-task learning framework to enhance the clarity and effectiveness. Furthermore, Zhang et al.[24] appended the multi-modal information in multi-task learning framework to strengthen the diagnostic accuracy of the same disease. Zeng et al.[25] presented a deep belief network-based multi-task learning method for diagnosing Alzheimer's disease, offering improved diagnostic capabilities across multiple tasks.

Despite these advancements, existing medical text classification approaches still face notable limitations. While BERT-based models have shown great potential in capturing complex semantic structures, they often struggle with class imbalance, leading to poor generalization for rare diseases. Moreover, current multi-task learning strategies improve representation learning, but they lack explicit mechanisms to mitigate class imbalance effectively. Additionally, most existing approaches focus on either data augmentation or multi-task learning separately, without fully leveraging their complementary strengths. These gaps highlight the need for an integrated approach that simultaneously enhances representation learning and addresses class imbalance challenges.

To address these limitations, the proposed classification framework incorporating three key innovations:

1. Self-attentive adversarial augmentation network (SAAN): Existing GAN-based data augmentation methods generate synthetic samples but fail to preserve domain-specific medical knowledge. We introduce SAAN, which employs adversarial sparse self-attention to enhance the quality of generated samples, ensuring more realistic and semantically coherent minority-class representations.
2. Disease-aware multi-task BERT (DMT-BERT): Standard BERT models lack explicit medical domain knowledge integration. We extend BERT by incorporating a secondary task that learns disease co-occurrence patterns alongside classification, allowing the model to better understand medical relationships and improve classification accuracy for rare disease categories.
3. The effectiveness of our proposed approach is validated through experiments on real-world medical datasets, demonstrating significant improvements over baseline models in terms of classification performance, especially for rare disease detection.

The remainder of this paper is structured as follows: Sect. 2 gives theoretical basis. Section 3 presents the details of our proposed methodology. Section 4 discusses the experimental setup and performance evaluation. Finally, Sect. 5 concludes the study and outlines future research directions.

Background on GAN and BERT

GAN and BERT are the basic modules of the proposed method Consequently, a comprehensive background on both GAN and BERT is presented as below.

Background on GAN

GAN is a class of DL models designed for data generation. GAN consists of two modules: a generator (G) that produces synthetic samples, the authenticity of these samples are assessed by a discriminator (D). These two networks are trained simultaneously in a minimax game, where G aims to generate realistic samples while D strives to differentiate between authentic and generated samples. Specifically, the loss function$\mathbf{L}_G$ of G is defined as:

$$\mathbf{L}_G \;=\; -\,\mathrm{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[log\,\mathrm{D}\left(\mathrm{G}\left(\mathbf{z}\right)\right)\right] \tag{1}$$

where $\mathrm{G}\left(\mathbf{z}\right)$ represents the sample generated by G, $D\left(\right)$ denotes the output of D, and $\mathrm{E}_{\mathbf{z}\sim p(\mathbf{z})}$ indicates the expectation of the random noise $\mathbf{z}$. This loss function guides G to continuously optimize itself, enabling the samples it generates to better "deceive" D. D acts as a "judge", It receives inputs which can be real embedding vectors or the output of G, and Produces an estimated likelihood $\in[0,1]$, signifying how probable it is that the provided instances are authentic. The loss function L_D of D can be expressed as:

$$L_D=-E_{\mathbf{x}\sim pdata(\mathbf{x})}\left[\log D(\mathbf{x})\right]-E_{\mathbf{z}\sim p(\mathbf{z})}\left[\log(1-D(G(\mathbf{z})))\right] \tag{2}$$

where $\mathbf{x}$ represents the real sample, and $\mathbf{x} \sim pdata(\mathbf{x})$ denotes the distribution of the real data. D seeks to improve its capability to differentiate authentic and synthetic instances by reducing the value of L_D.

Background on BERT

The powerful training capacity of the BERT model is driven by its input design and the multi-head self-attention mechanism. BERT's primary input consists of vectors representing words, which are converted from raw text into tokens via the tokenizer function. These tokens capture the words' meanings, while the positional embeddings represent their relative positions in the sequence. This structure enables BERT to effectively capture contextual and sequential relationships within the text. Self-attention mechanism is applied to the encoder module in BERT. The essence of attention is to give higher weight to the part of attention, so as to obtain more effective information. Calculating formulas as (3) shown.

$$Attention(\mathbf{Q},\mathbf{K},\mathbf{V})_j = soft\max(\frac{\mathbf{Q}_j\mathbf{K}_j^T}{\sqrt{d_k}})\mathbf{V}_j \tag{3}$$

Where (Q,K,V) is the result of input x multiplied by different weight matrices, $d_k$is the dimension of kth head. The advantage of this design is that the model can have a more comprehensive feature extraction capability.

Proposed methodology

In this section, we present the detailed derivations of our proposed enhanced medical text classification framework, as shown in Fig. 1, which mainly consists of Self-Attentive Adversarial Augmentation Network (SAAN) and Disease-Aware Multi-Task BERT (DMT-BERT).

These models are developed to tackle the class imbalance and rare disease recognition in medical text classification tasks. First, we introduce the architecture and working principles of SAAN, which leverages a self-attentive adversarial approach to generate high-quality minority class samples for data augmentation. Following this, we delve into the DMT-BERT framework, which integrates multi-task learning with BERT to enhance feature learning, focusing specifically on disease co-occurrence relationships to improve classification performance. Both models complement each other and provide a holistic solution to enhance the overall classification accuracy, particularly for rare diseases in medical datasets.

SAAN

The SAAN is designed to address class imbalance in medical datasets by generating high-quality minority class samples. By incorporating sparse self-attention mechanisms within the GAN framework, SAAN enhances the generation process, ensuring that the synthetic samples are both contextually relevant and semantically coherent. This approach enhances the feature extraction capacity from the minority class, which is often underrepresented in medical text datasets. As shown in Fig. 2, SAAN is composed of an sparse attention generator (SAG) and an sparse attention discriminator (SAD). For better generating meaningful synthetic samples, sparse self-attention mechanisms are combined into both components to extract the long-term dependencies of data.

SAG is responsible for creating synthetic samples that resemble the minority class. Inputting random Gaussian noise $\mathbf{z}$, SAG outputs a synthetic medical text sample $\hat{\mathbf{x}}$ as demonstrated in Eq. (4), in which sparse self-attention is incorporated to selectively process input regions, ensuring contextually coherent generation.

$$\hat{\mathbf{x}} = SAG\left(\mathbf{z}\right) = \text{sparse\_}selfattention\left(\mathbf{z}\right) \tag{4}$$

The $\text{sparse\_}selfattention\left(\right)$ works by projecting the input sequence, i.e. $\mathbf{z} = \{z_i\}_{i=1}^{t}$ with z_i *w.r.t* $x_i$and $z_i \in R^d$, into several subspaces using linear transformations with respect to different parameter matrices.

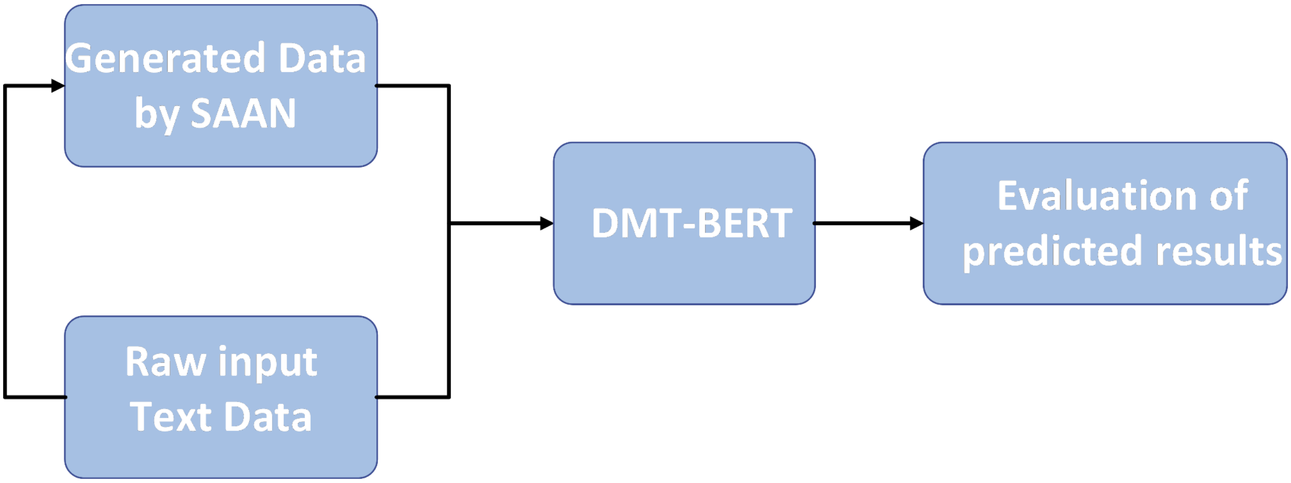


Fig. 1. Overall framework of the proposed enhanced medical text classification model.

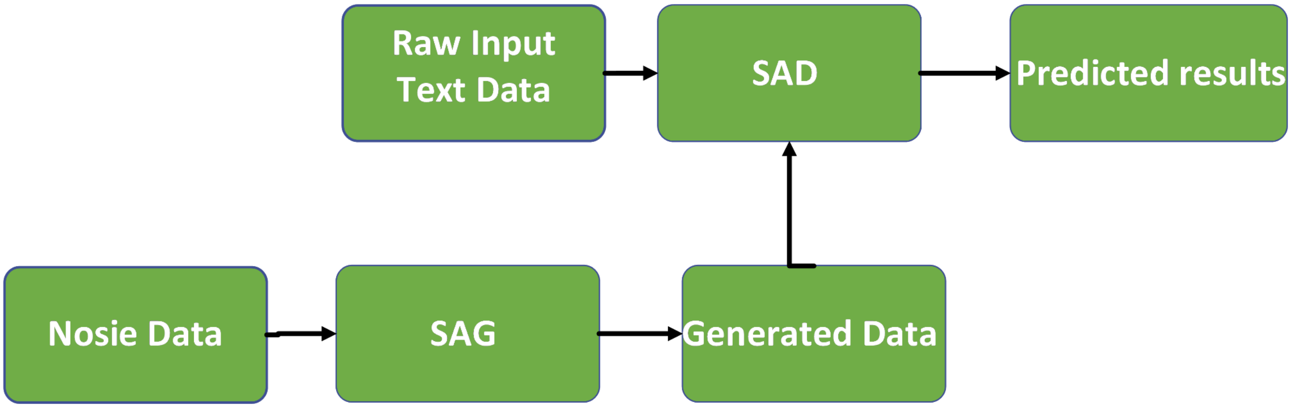


Fig. 2. The structure of the proposed SAAN.

Each sparse attention head applies its own matrix for projection, generating separate representations for each subspace. To enforce sparsity, each head applies a local window mask$\mathbf{M}_j \in \{0,-\infty\}^{n\times n}$where$M_{j,ik}=0$ only if $|i-k| \leqslant w$ (window size *w*≪*n*), in which *w* is set as 86. The parallel sparse attention computations are denoted as *m*, which refers to the multi-head sparse attention mechanism:

$$\text{sparse\_}selfattention\left(\mathbf{z}\right)=Concat(\{head_j\}_{j=1}^{m})\mathbf{W}^A \tag{5}$$

$$head_j=\text{sparse\_}attention(\mathbf{Q},\mathbf{K},\mathbf{V})_j \tag{6}$$

$$\text{sparse\_}attention(\mathbf{Q},\mathbf{K},\mathbf{V})_j=\text{softmax}\left(\frac{\mathbf{Q}_j\mathbf{K}_j^T}{\sqrt{d_k}}+\mathbf{M}_j\right)\mathbf{V}_j \tag{7}$$

$$\mathbf{K}_j=\mathbf{z}\mathbf{W}_j^k \tag{8}$$

$$\mathbf{V}_j=\mathbf{z}\mathbf{W}_j^v \tag{9}$$

$$\mathbf{Q}_j=\mathbf{z}\mathbf{W}_j^q \tag{10}$$

where $\mathbf{Q}_j$,$\mathbf{K}_j$ and $\mathbf{V}_j$ represent the query, key, and value vectors, respectively. $head_j$ refers to the *j*-th attention head, and W_j^k, W_j^v, $W_j^q \in R^{d\times d_k}$are the trainable matrices for each head, besides $W^A \in R^{md_k\times d}$ and $d_k=d/m$.The noise vector **z** is sampled from a Gaussian distribution, and the self-attention mechanism helps the Generator to focus on the most relevant features during sample generation.

SAG aims to minimize the probability of the SAD correctly classifying its generated samples as fake. The SAG's objective is to "fool" the SAD, making it believe the synthetic samples are real. The SAG loss is defined as:

$$\mathbf{L}_{SAG}=-E_{\mathbf{z}\sim p(\mathbf{z})}\left[\log SAD\left(SAG\left(\mathbf{z}\right)\right)\right] \tag{11}$$

where $SAD\left(SAG\left(\mathbf{z}\right)\right)$is the SAD's output for the generated sample.

SAD is responsible for differentiating real and synthetic samples, utilizing self-attention to identify global relationships within the input sequence. SAD outputs a probability *SAD*(**x**) representing the likelihood that the input $\mathbf{X}=[\mathbf{x},\hat{\mathbf{x}}]$ is a real sample **x** (i.e., from the minority class) or a fake one **x**^ (i.e., generated by SAG).

$$SAD\left(\mathbf{X}\right)=\text{sparse\_}selfattention\left(\mathbf{X}\right) \tag{12}$$

SAD's goal is to correctly distinguish between real and fake samples. The SAD loss is defined as:

$$\mathbf{L}_{SAD}=-E_{\mathbf{x}\sim p(\mathbf{x})}\left[\log SAD\left(\mathbf{x}\right)\right]-E_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(1-SAD\left(SAG\left(\mathbf{z}\right)\right)\right)\right] \tag{13}$$

where $p(\mathbf{x})$ is the real data distribution, and $SAD\left(\mathbf{x}\right)$ is the SAD's output for a real sample.

This adversarial mechanism enables the generator to gradually master the distribution characteristics of real data and eventually generate high-quality samples of minority classes. These generated samples are then used to expand the training dataset, effectively alleviating the problem of class imbalance and enhancing the recognition ability of the improved BERT classification model for minority classes.

DMT-BERT

The DMT-BERT framework extends the BERT model by introducing multi-task learning to capture disease-related co-occurrence patterns, whose structure is shown in Fig. 3. This allows the model to better handle rare disease classification by simultaneously learning useful auxiliary tasks related to disease co-occurrences while performing the main classification task.

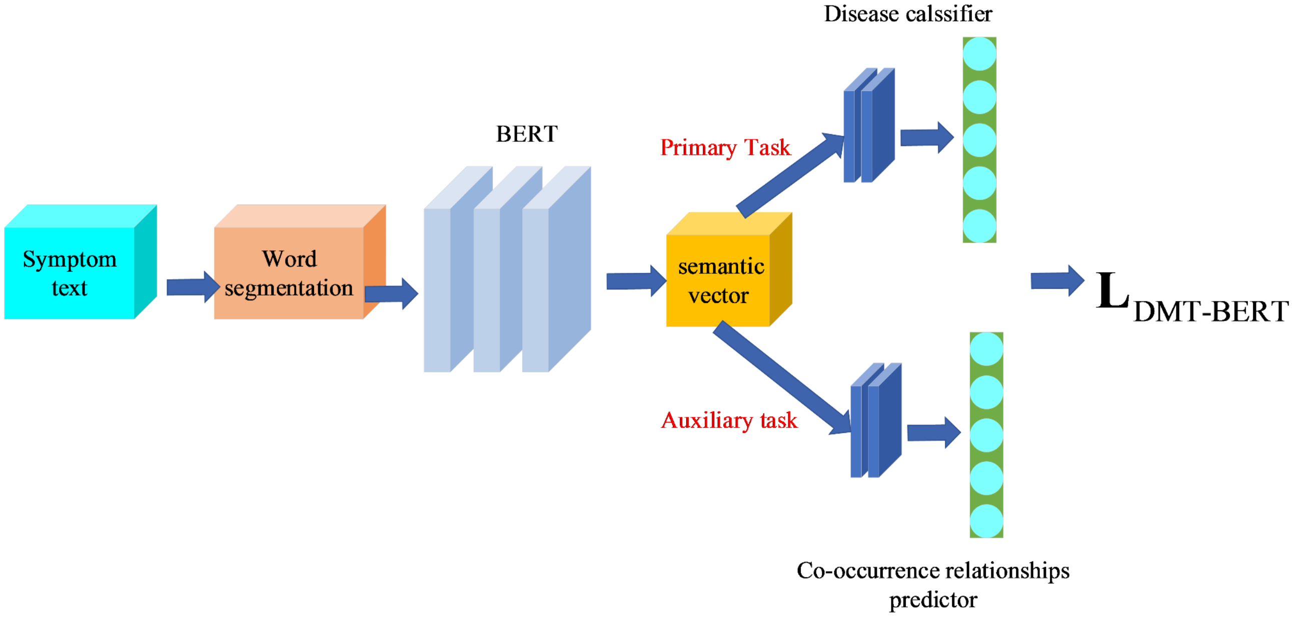


Fig. 3. The structure of the proposed DMT-BERT.

Feature encoding module of DMT-BERT
BERT mainly makes up the encoding layer of DMT-BERT. The process of encoding mainly involves the following key steps:

Word segmentation Firstly, the text needs to be processed through word segmentation, which splits consecutive words into meaningful words or sub-words to enhance the feature extraction ability. In the DMT-BERT model, word segmentation is implemented by BertTokenizer. BertTokenizer is a word segmenter based on the WordPiece algorithm, which can decompose the input text into a series of sub-words. This segmentation method can not only effectively handle common words but also solve the problem of out-of-vocabulary words. Moreover, the WordPiece algorithm achieves a good balance between the semantic integrity of words and the size of the model's vocabulary by learning sub-word units.

Encoding by BERT Bert as the encoder deals with the segmented text sequence through a multi-layer Transformer structure. Each Transformer layer contains multi-head self-attention mechanism and feed-forward neural network. By decomposing the attention mechanism into multiple heads, the model can simultaneously capture the semantic features of the text from different representation subspaces. Parallel attention computations not only enhance the model's expressive power but also improve the stability of feature extraction. For the input $\mathbf{x} \in \mathbb{R}^{n \times d}$, the calculation process of the multi-head self-attention mechanism is:

$$\text{MultiHead}(\mathbf{x}) = \text{contact}(\mathbf{head}_1, \mathbf{head}_2, \ldots\ldots, \mathbf{head}_n) \tag{14}$$

Among them, the calculation formula for each head is,

$$\mathbf{head}_i = \text{sparse}\_selfattention(\mathbf{XW}_i^Q, \mathbf{XW}_i^K, \mathbf{XW}_i^V) \tag{15}$$

And the formula of the Attention is,

$$sparse\_selfattention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{QK}^T}{\sqrt{d_k}} + \mathbf{M})\mathbf{V} \tag{16}$$

Among them, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times d}$represent Query, Key and Value respectively. Through the multi-head attention mechanism, each attention head can focus on different aspects of the input sequence. Moreover, for the scaling factor $\sqrt{d_k}$, it can effectively avoid the gradient problem in the case of large dimensions.

Extract semantic vector In each Transformer layer of DMT-BERT, a two-layer structure is adopted to construct the feed-forward neural network. The first layer uses a larger hidden dimension to expand the feature representation space. The second layer subsequently transforms the features to match the original dimension. This design enables the model to maintain the integrity of information while enhancing the nonlinear transformation ability. Meanwhile, the use of residual connections ensures that the gradients can effectively propagate in deep networks. After obtaining the hidden states $\mathbf{f} \in \mathbb{R}^{n \times d}$, the final text vector representation of DMT-BERT is obtained through mean operation:

$$\mathbf{H} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{f}_i \tag{17}$$

By adopting this pooling strategy, the model is enabled to effectively integrate all positional information within the sequence, thereby generating a compact text representation.

Multi-task learning architecture of DMT-BERT

The DMT-BERT model consists of two key components:

1. Primary task: The task of classifying medical texts into predefined categories (e.g., disease types or symptom categories) using the [CLS] token.
2. Auxiliary task: The task of predicting disease co-occurrence relationships, which captures how certain diseases or symptoms are likely to occur together.

The primary task involves classifying medical texts into predefined categories. This is done by using the BERT architecture, which generates contextualized embeddings of each token, while the output $\mathbf{h}_{cls}$associated [CLS] token is for the final classification decision. Supposed that$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_1, ?., \mathbf{x}_n]$, where$\mathbf{x}_i$ represents the tokens (words or sub-words), the output from the BERT model can be represented as $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_1, ?., \mathbf{h}_n]$, where each $\mathbf{h}_i$ is the embedding vector of the token$\mathbf{x}_i$. Then, $\mathbf{h}_{cls}$ is input a FC layer to obtain the predicted label for the input sequence. The formula for classification is:

$$\hat{\mathbf{y}} = \mathrm{softmax}(\mathbf{W}_{cls}\mathbf{h}_{cls} + \mathbf{b}_{cls}) \tag{18}$$

where $\mathbf{W}_{cls}$and $\mathbf{b}_{cls}$ are the weight matrix and bias of the FC layer, $\hat{\mathbf{y}}$ is the predicted class label (disease type or symptom category).

The auxiliary task in DMT-BERT aims to predict disease co-occurrences based on the relationship between different diseases or symptoms mentioned in the text. The co-occurrence prediction is based on the output embeddings for the tokens $\mathbf{h}_1, \mathbf{h}_1, ?., \mathbf{h}_n$. To predict disease co-occurrences, we define a binary classification for each pair of diseases $\mathbf{d}_i$ and $\mathbf{d}_j$ in the text. The task is to predict whether diseases $\mathbf{d}_i$ and $\mathbf{d}_j$ co-occur in the same medical record. The co-occurrence prediction is modeled by calculating the attention between the token representations corresponding to the diseases $\mathbf{d}_i$ and $\mathbf{d}_j$. For each pair, the attention score $\mathbf{A}_{ij}$ is computed as:

$$\mathbf{A}_{ij} = \frac{\mathbf{h}_i \cdot \mathbf{h}_j^T}{\|\mathbf{h}_i\| \, \|\mathbf{h}_j\|} \tag{19}$$

where $\mathbf{h}_i$ and $\mathbf{h}_j$ are the embeddings for the tokens corresponding to diseases $\mathbf{d}_i$ and $\mathbf{d}_j$, and $\|\cdot\|$represents the vector norm. Thus, the co-occurrence prediction loss is then computed using binary cross-entropy:

$$\mathbf{L}_{\text{co-occurrence}} = -\sum_{i,j} \left[\mathbf{y}_{ij} \log\left(\mathbf{p}_{ij}\right) + \left(1 - \mathbf{y}_{ij}\right) \log\left(1 - \mathbf{p}_{ij}\right)\right] \tag{20}$$

Where $\mathbf{y}_{ij}$ is the true label indicating whether diseases $\mathbf{d}_i$ and $\mathbf{d}_j$ co-occur, $\mathbf{p}_{ij}$is the predicted probability that diseases co-occur, computed from the attention score $\mathbf{A}_{ij}$.

The overall loss function for DMT-BERT is a weighted sum of the primary classification loss and the auxiliary co-occurrence prediction loss:

$$\mathbf{L}_{\text{DMT-BERT}} = \lambda_1 \mathbf{L}_{\text{classification}} + \lambda_2 \mathbf{L}_{\text{co-occurrence}} \tag{21}$$

Where $\mathbf{L}_{\text{classification}}$ the cross-entropy loss for the primary disease classification task, λ_1 and λ_2 are hyperparameters controlling the relative importance of the two tasks.

In DMT-BERT, the BERT model is fine-tuned with additional layers corresponding to the auxiliary task. The [CLS] token output is used for the main classification task, while the other token embeddings are used to predict disease co-occurrences in the auxiliary task. The BERT model undergoes pre-training with two tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). During fine-tuning, both the classification task and the co-occurrence task are optimized simultaneously.

Experimental analysis

Evaluation indexes

Multi-evaluation indexes are adopted in the paper. First, Confusion Matrix (Table 1) are presented. Next, the ROC curve is used to illustrate the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across various threshold values. This offers an intuitive assessment of the model's classification ability.

Besides, F1 Score is adopted as another evaluation index, whose calculation formula are,

$$\mathrm{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{22}$$

$$\text{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{23}$$

Actual	Predicted	
	Positive	Negative
Positive	TP (true positive)	FN (false negative)
Negative	FP (false positive)	TN (true negative)

Table 1. The structure of confusion matrix, making general binary classification as an example.

Dataset	Body	Signs	Disease	Check	Treatment	Total
Train set	10,719	7831	722	9546	1048	29,866
Test set	3021	2311	553	3143	465	9493

Table 2. Token-level distribution of five entity types.

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{24}$$

The description of the clinical dataset

The clinical data used is a statistical table containing a large number of patient cases, which records the detailed medical history, examination results, diagnosis and treatment situations of different patients. The clinical dataset is retrospectively collected from 2 tertiary cardiovascular hospitals in China between January 2023 and December 2023, comprising 317 de-identified patient records. This paper focuses on the four columns: chief complaint, present illness history, past medical history, and the effect of right heart angiography. The three columns of chief complaint, present illness history and past medical history contain rich symptom characteristics, while the column of the effect of right heart angiography contains five different results, namely no obvious right-to-left shunt at the atrial septum plane, right-to-left shunt at the atrial septum plane (grade I), right-to-left shunt at the atrial septum plane (grade II), right-to-left shunt at the atrial septum plane (grade III), and others. In this experiment, the above results are classified into 0 class, 1 class, 2 class, 3 class and 4 class respectively.

The description of the CCKS 2017 datasets

To evaluate our proposed model, we conducted experiments on the CCKS 2017 Task 2 dataset[26], which comprises 1,198 training samples and 796 test samples. Table 2 summarizes the statistics of five distinct entity types. Both datasets were annotated using the BIO scheme, where the "B-" prefix indicates the first character of an entity and the "I-" prefix denotes subsequent characters. Specifically, B-BODY/I-BODY denote the initial and non-initial characters for body parts; B-SIGNS/I-SIGNS correspond to those for symptoms; B-CHECK/I-CHECK refer to medical examinations; B-DISEASE/I-DISEASE indicate disease entities; and B-TREATMENT/I-TREATMENT represent treatment entities. The label "O" is used for characters that do not belong to any named entity.

Data preprocessing

A multi-stage preprocessing framework is implemented to ensure data quality and consistency. In the first phase, regular expressions are employed to clean the raw text by removing special characters, redundant spaces, punctuation, and other irrelevant elements while correcting common formatting errors such as repeated characters and improper line breaks. At the same time, a specialized medical term dictionary is integrated to maintain the professional integrity of the text; term matching and standardization processes ensure accurate extraction and consistent representation of medical terminology. Next, the open-source Chinese word segmentation tool jieba is utilized for precise tokenization. Given the unique characteristics of medical texts, the jieba segmenter is optimized for the medical field through dictionary expansion and adjustment of word frequency weights, which improves the recognition accuracy of domain-specific terms. This step lays a solid foundation for subsequent feature extraction. Finally, the representational capabilities of a pre-trained BERT model are leveraged to convert the processed text into high-dimensional embedding vectors. This transformation captures the semantic nuances of the text while preserving contextual relationships, thereby providing high-quality feature representations for downstream classification tasks.

The detailed description of Stratified K-Fold Cross-Validation has been appended. To ensure the model's stability and generalization capability, we implement a stratified K-fold cross-validation approach with K set to 5. In this method, the dataset D is partitioned into five subsets D1, D2, …, D5, each reflecting the original class distribution. For each fold, one subset is designated as the validation set while the remaining four subsets are used for training. This process is repeated until every subset has served as the validation set once. Performance is evaluated using metrics such as precision, recall, F1 score, and accuracy, with the final results obtained by averaging the outcomes across all folds. This detailed strategy minimizes evaluation bias and provides a robust estimate of the model's performance.

Data enhancement

To address the issue of class imbalance in the dataset and enhance the generalization ability of the model, we have devised a dual data augmentation strategy: The first strategy is based on the SAAN as described previously.

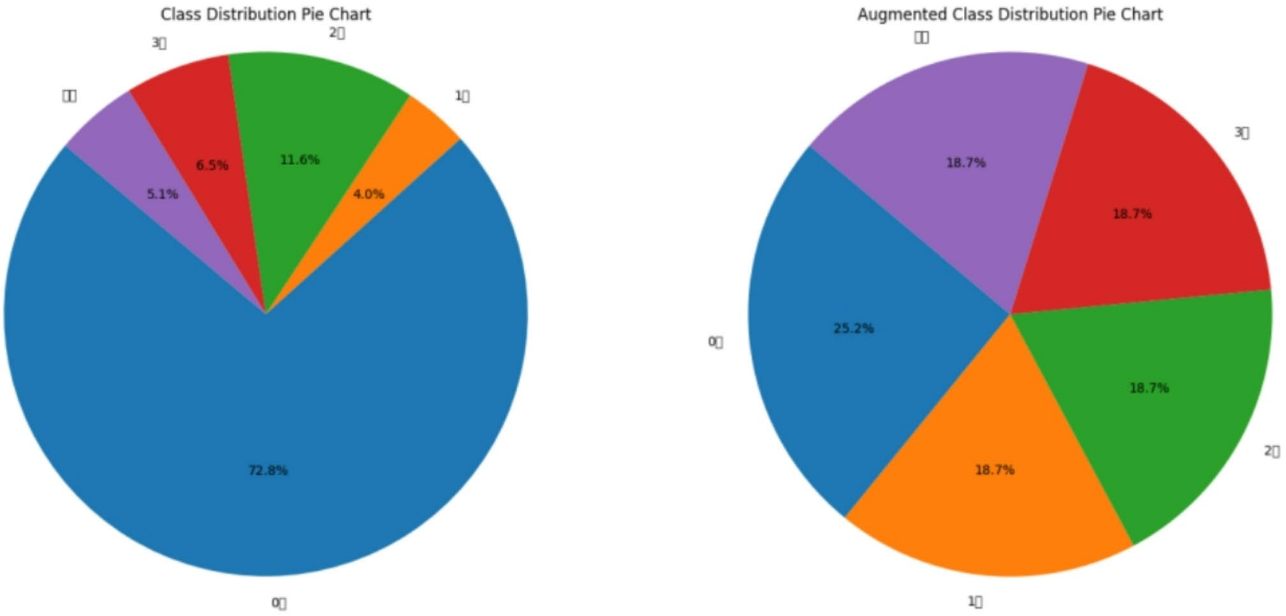


Fig. 4. The proportions of various data before and after enhancement.

Hyperparamete	Value	Hyperparameter	Value
Maximum text length	512	Number of encoder module	2
batch size	32	Encoder of DMT-BERT	256
learning rate	2e-5	Disease classifier/ Predictor	768-256-4/256-2
Epochs	200	Dropout	0.2
AG module	50	AD module	50

Table 3. The hyper-parameters of the prosed model.

Through the generator of SAAN, the features of minority class samples are simulated, and new samples with similar statistical characteristics are generated, thereby increasing the proportion of the minority class in the dataset. This method not only balances the class distribution but also maintains the semantic coherence of the generated samples.

The second strategy adopts a vocabulary-based enhancement approach, mainly modifying the text through two operations: synonym substitution and random insertion. For the original text T, an enhanced text T' is generated by replacing the words in it with a certain probability p, and the probability model is expressed as:

$$\mathrm{P}\left(\mathrm{T}'|\mathrm{T}\right) = \mathrm{p}^{k}(1-\mathrm{p})^{(n-k)} \tag{25}$$

where p represents the probability of substitution, *k* denotes the number of words to be replaced, and *n* indicates the total number of words in the text. Meanwhile, through random insertion operations, words related to the semantic context are inserted at appropriate positions in the text to further increase the expression diversity of the text. The insertion positions and the selection of words are based on the calculation of the semantic relevance to the context. Figure 4 show the proportions of various data before and after enhancement.

Model parameter settings

After ten parallel trials, based on the optimal test results, the hyperparameter Settings of the model are shown in the Table 3.

Experimental performance evaluation of the clinical dataset

Prediction evaluation of the proposed method

Table 4 outlines the model's performance metrics for each of the five cross-validation folds in detail. From the data, it can be observed that the model performs stably across all folds, with an average precision rate of 0.93 and an average recall rate of 0.91. Particularly noteworthy is that even on the minority class samples, the model still maintains a relatively high recognition accuracy rate, which validates the effectiveness of the GAN data augmentation strategy in balancing the class distribution.

Sub task	Precision	Recall	F1 score
Fold1	0.94	0.92	0.92
Fold2	0.92	0.88	0.89
Fold3	0.93	0.91	0.91
Fold4	0.93	0.91	0.91
Fold5	0.93	0.91	0.91
Mean value	0.93	0.91	0.91

Table 4. Accuracy rate, recall rate, F1 score on each fold.

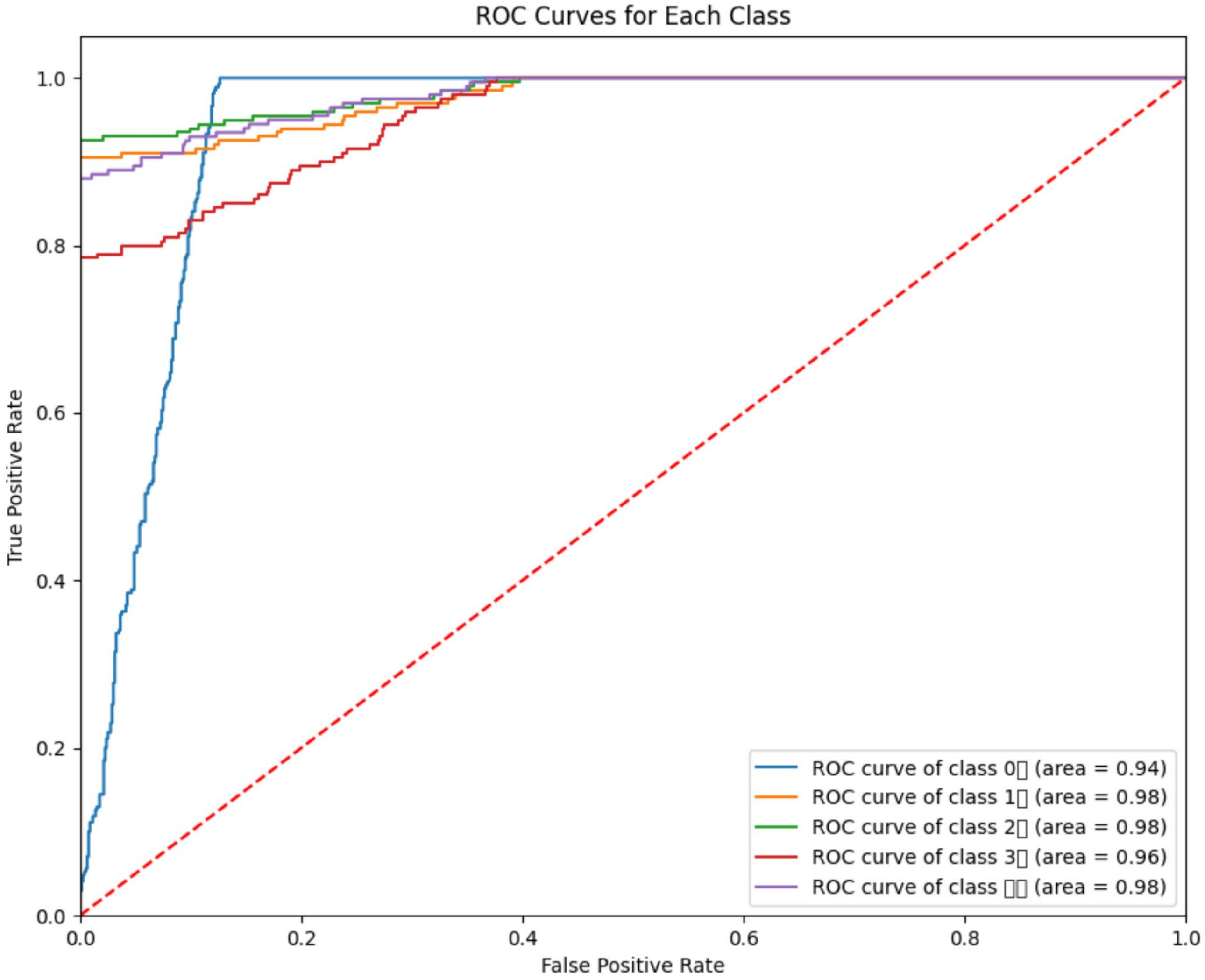
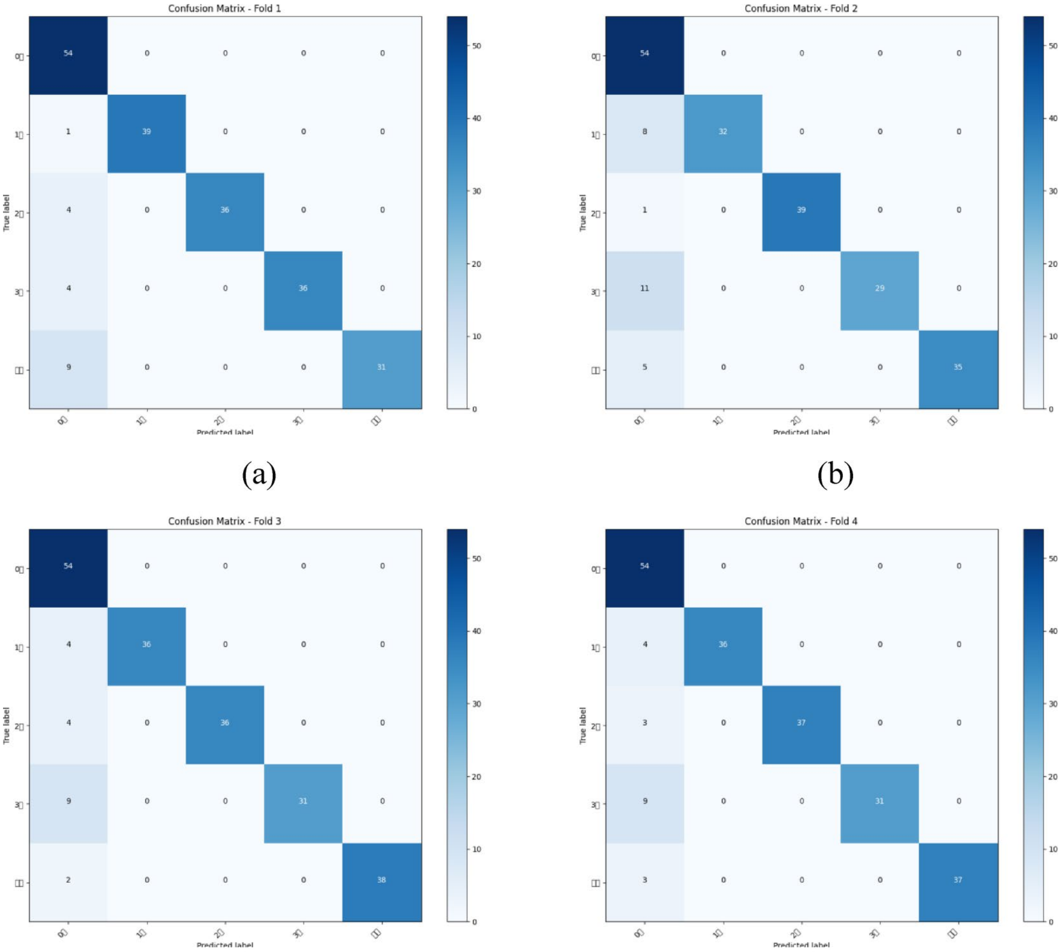


Fig. 5. ROC-AUC curves for each category.

Figure 5 displays the ROC curves for the classification model across five different categories. The ROC curve is a key metric for assessing classification models, displaying the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) to highlight the model's effectiveness in classification. The red dashed line represents the baseline for random guessing, while the five solid lines, each in a different color, correspond to the ROC curves of the five categories. The legend for each curve shows the corresponding AUC (Area Under the Curve) values, with class 0 having an AUC of 0.94, and the other three categories achieving an AUC of 0.98. An AUC value closer to 1 signifies superior model performance. From the figure, it is evident that all categories exhibit high AUC values, demonstrating the model's excellent classification performance across these categories. The model has shown high accuracy and reliability in handling classification tasks for these five categories.

Figure 6 contains four subplots, each displaying a confusion matrix for a classification model across different folds (Fold 1 to Fold 4). The confusion matrix provides a visual representation of model performance by



(a)

(b)

(c)

(d)

Fig. 6. Confusion matrices on different folds.

comparing true labels with predicted labels. In each subplot, the x-axis shows the true labels, and the y-axis displays the predicted labels. Diagonal values indicate correctly classified samples, while off-diagonal values represent misclassifications. The color intensity indicates the sample size, with darker colors indicating higher quantities. The model's performance is relatively stable across all folds, with most samples being correctly classified, as reflected in the dark-colored blocks along the diagonal. For instance, in Fold 1, classes 0, 1, 2, and 3 had 54, 38, 36, and 31 samples correctly classified, respectively. In Fold 2, classes 0, 1, 2, and 3 had 54, 27, 24, and 35 samples correctly classified, respectively. Similar trends were observed in other folds, although there were minor misclassifications in certain folds (e.g., Fold 3 and Fold 4). However, the model's classification performance remained strong overall. These confusion matrices indicate that, despite some misclassifications, the model can accurately recognize most categories, demonstrating high classification accuracy.

Study of ablation

A series of ablation experiments are performed to assess the effectiveness of the proposed approach. In these experiments, we compared different versions of the model by systematically removing or modifying key components. Model m1 represents the complete SAAN and DMT-BERT framework, incorporating both SAAN-based augmentation and multi-task learning. Model m2 adopts the identical architecture as m1, except it excludes the SAAN component, relying solely on DMT-BERT with multi-task learning. In this setup, the model does not benefit from the synthetic minority class samples generated by SAAN. Model m3 adopts the full SAAN architecture but uses a single-task BERT model for classification instead of multi-task learning. In this case, the model performs only the primary classification task, omitting the auxiliary disease co-occurrence prediction task. Model m4 represents a baseline version, where both the SAAN and DMT-BERT components are removed, and the model relies on the standard BERT architecture with no augmentation or multi-task learning. Each model is optimized through fine-tuning, and ten parallel experiments are run for each version. The average F1-score, Recall, and Precision accuracy across all experiments were calculated and summarized in Table 5.

The results clearly demonstrate that m1, the full model, consistently outperforms all other variants. Specifically, Model m2 (DMT-BERT only) shows a noticeable decrease in performance, especially in the recognition of rare

Model	Precision	Recall	F1 score
m4	0.85	0.80	0.82
m3	0.90	0.88	0.89
m2	0.88	0.85	0.86
m1	0.93	0.91	0.91

Table 5. Ablation experiment results of different methods on clinical dataset.

Model	Precision	Recall	F1 score
BERT	0.89	0.85	0.87
RoBERTa[27]	0.89	0.88	0.87
XLNet[28]	0.90	0.86	0.86
Wind2vec-BERT[29]	0.90	0.87	0.88
Bert-pin[30]	0.91	0.88	0.89
Mask-guided BERT[11]	0.92	0.88	0.90
Proposed	0.93	0.91	0.91

Table 6. Comparison results with state-of-arts.

diseases, while Model m3 (SAAN only) performs better but still lacks the benefit of multi-task learning. Model m4, the baseline model, performs the worst, as it neither incorporates the augmentation from SAAN nor benefits from the auxiliary task of disease co-occurrence learning. These results demonstrate that both SAAN and DMT-BERT are essential for enhancing the model's performance. The SAAN-based augmentation significantly handles class imbalance, while the multi-task learning helps generalize ability to rare diseases by learning disease co-occurrence patterns.

Study of compared results with recent methods

To highlight the advantages of the proposed approach, a comparative study was conducted against several recent deep learning models, including BERT, RoBERTa, XLNet, Wind2vec-BERT, Bert-pin, and Mask-guided BERT. Ten parallel experiments are performed to fine-tune each model on the medical dataset, and the performance metrics (precision, recall, and F1-score) are summarized in Table 6. As shown, the highest performance across all metrics is achieved by the proposed model, with a precision of 0.93, recall of 0.91, and F1-score of 0.91, surpassing the best-performing baseline, Mask-guided BERT, by 1% in F1-score.Wind2vec-BERT and Bert-pin both exhibit improvements over the baseline BERT, with Bert-pin attaining slightly higher precision and recall than Wind2vec-BERT. The marked gains in the proposed method can be attributed to the integration of SAAN for data augmentation and DMT-BERT for multi-task learning, which together address class imbalance and capture disease co-occurrence patterns more effectively. This approach is particularly beneficial for the detection of rare diseases in medical text. In conclusion, superior performance is demonstrated by the proposed model in comparison to existing state-of-the-art methods, and robust generalization capabilities are exhibited, making it highly effective for medical text classification tasks.

Besides, the proposed approach, combining SAAN-based data augmentation and multi-task learning, offers significant accuracy improvements for medical text classification, particularly for rare diseases. While this integration does introduce a computational burden due to the complexity of the models, the resulting precision enhancements are crucial for medical applications. These improvements enable more accurate disease diagnosis and risk assessment, ultimately contributing to better patient care. Thus, the trade-off between computational complexity and accuracy is justified, as the benefits of improved classification outweigh the additional computational cost. Noted that this research has achieved remarkable results on specific clinical datasets, it may encounter some challenges when generalized to different types and domains of medical texts.

Experimental performance evaluation of the CCKS 2017 datasets

Table 7 presents the precision, recall, and F1-scores for all five entity types. Notably, the SIGNS and CHECK categories exhibit superior F1 performance relative to DISEASE and TREATMENT. This discrepancy may be due to the highly variable lengths of TREATMENT annotations and a relatively lower volume of annotated DISEASE instances.

In order to evaluate the performance of our proposed model, we compare our model with several baseline models. Table 8 provides a performance comparison among several models on the CCKS 2017 dataset in terms of precision, recall, and F1-score. The experimental results demonstrate the superior performance of the proposed method compared to existing approaches. Comparative analysis with architectures including RoBERTa[27] and XLNet[28] confirms the proposed method's superior capability in capturing domain-specific linguistic patterns. Our model achieves state-of-the-art precision of 94.52% and recall of 93.24%, surpassing the best baseline (Mask-guided BERT[11]) by 1.40% and 1.90% respectively in these two metrics. While the F1-score (91.80%) slightly trails Wind2vec-BERT[29] (90.89%) and Bert-pin[30] (91.54%) in absolute value, this discrepancy primarily stems from our method's balanced optimization strategy that prioritizes clinical decision-making requirements

Model	Types	Precision	Recall	F-score
Proposed method	BODY	94.52	93.44	94.10
	SIGNS	98.00	97.80	97.50
	CHECK	96.26	95.62	91.80
	DISEASE	92.78	90.86	90.70
	TREATMENT	91.04	88.48	85.90

Table 7. Different types of named entity recognition results.

Model	Precision	Recall	F1 score
BERT	87.90	89.23	89.93
RoBERTa[27]	89.40	89.18	87.90
XLNet[28]	90.24	88.89	87.39
Wind2vec-BERT[29]	90.67	89.90	90.89
Bert-pin[30]	92.14	89.23	91.54
Mask-guided BERT[11]	93.12	91.34	91.23
Proposed	94.52	93.24	91.80

Table 8. Comparison results with state-of-arts of the CCKS 2017 datasets.

through precision-recall tradeoff calibration. Notably, the proposed approach maintains consistent performance improvements across all evaluation dimensions, particularly excelling in precision-critical medical entity recognition tasks where it outperforms conventional BERT variants by 6.62–14.06%. This enhanced capability in minimizing false positives while preserving recall sensitivity suggests effective integration of domain-specific linguistic patterns through our novel architectural modifications.

Conclusion

This paper proposes an enhanced medical text classification framework integrated with SAAN and DMT-BERT to address class imbalance and improve the classification of rare diseases in medical texts. The main contributions are as follows: First, the introduction of SAAN enhances high-quality synthetic samples generation for underrepresented classes, addressing the class imbalance issue. Second, DMT-BERT leverages multi-task learning to capture disease co-occurrence patterns, which improves the model's performance, particularly in recognizing rare diseases. The effectiveness of these innovations was confirmed through ablation experiments, which demonstrated the significant contribution of both components to the overall performance. The enhanced medical text classification framework outperforms recent models in terms of key metrics such as F1-score, Precision, and Recall, showing improvements for disease detection. The results highlight the effectiveness of the proposed model in handling medical text classification tasks, particularly in overcoming the challenges of class imbalance and rare disease recognition. In the future, the explain-ability strategies will be researched to bridge AI models with clinical reasoning.

Data availability

To improve the transparency and reproducibility of our work, raw data are publicly available at https://github.com/XS-RUL/clinical-datasets.git.

Received: 7 February 2025; Accepted: 10 April 2025
Published online: 22 April 2025

References

1. Martínez-García, M. & Hernández-Lemus, E. Data integration challenges for machine learning in precision medicine. *Front. Med.* **8**, 784455 (2022).
2. Sun, W. et al. Data processing and text mining technologies on electronic medical records: a review. *J. Healthc. Eng.* **2018** (1), 4302425 (2018).
3. Wu, S. et al. Deep learning in clinical natural Language processing: a methodical review. *J. Am. Med. Inform. Assoc.* **27** (3), 457–470 (2020).
4. Edara, D. C., Vanukuri, L. P., Sistla, V. & Kolli, V. K. K. Sentiment analysis and text categorization of cancer medical records with LSTM. *J. Ambient Intell. Humaniz. Comput.* **14** (5), 5309–5325 (2023).
5. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521** (7553)), 436–444 (2015).
6. Xiang, S. et al. Dynamic Self-Learning neural network and its application for rotating equipment RUL prediction. *IEEE Internet Things J.* (2024).
7. Xiang, S., Li, P., Huang, Y., Luo, J. & Qin, Y. Single gated RNN with differential weighted information storage mechanism and its application to machine RUL prediction. *Reliab. Eng. Syst. Saf.* **242**, 109741 (2024).
8. Li, P. et al. Applying neural network to health Estimation and lifetime prediction of Lithium-ion batteries. *IEEE Trans. Transp. Electrification* (2024).
9. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6** (1), 1–54 (2019).

10. Shaikhina, T. & Khovanova, N. A. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artif. Intell. Med.* **75**, 51–63 (2017).
11. Liao, W. et al. Mask-guided BERT for few-shot text classification. *Neurocomputing* **610**, 128576 (2024).
12. Bressem, K. K. et al. Medbert. De: A comprehensive German Bert model for the medical domain. *Expert Syst. Appl.* **237**, 121598 (2024).
13. Gasmi, K. In *Improving bert-based model for medical text classification with an optimization algorithm, International Conference on Computational Collective Intelligence*, 101–111 (Springer, 2022).
14. Zhang, Y., Wu, X., Fang, Q., Qian, S. & Xu, C. Knowledge-enhanced attributed multi-task learning for medicine recommendation. *ACM Trans. Inform. Syst.* **41** (1), 1–24 (2023).
15. Zeberga, K. et al. [Retracted] A novel text mining approach for mental health prediction using Bi-LSTM and BERT model. *Comput. Intell. Neurosci.* **2022** (1)), 7893775 (2022).
16. Dai, D. et al. Disambiguation of chinese polyphones in an end-to-end framework with semantic features extracted by pre-trained bert. Preprint at https://arXiv.org/2501.01102 (2025).
17. Darraz, N., Karabila, I., El-Ansari, A., Alami, N. & El Mallahi, M. Integrated sentiment analysis with BERT for enhanced hybrid recommendation systems. *Expert Syst. Appl.* **261**, 125533 (2025).
18. Xia, W. et al. Gan inversion: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **45** (3), 3121–3138 (2022).
19. Frid-Adar, M. et al. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **321**, 321–331 (2018).
20. 20 Suresh, T., Brijet, Z. & Subha, T. Imbalanced medical disease dataset classification using enhanced generative adversarial network. *Comput. Methods Biomech. BioMed. Eng.* **26** (14), 1702–1718 (2023).
21. Bissoto, A., Valle, E. & Avila, S. In Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1847–1856 (2021).
22. Liu, L. et al. In *Multi-task learning via adaptation to similar tasks for mortality prediction of diverse rare diseases, Amia annual symposium proceedings*, 763 (2021).
23. Suk, H. I., Lee, S. W., Shen, D. & Initiative, A. s. D. N., deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. *Brain Struct. Function.* **221**, 2569–2587 (2016).
24. Zhang, D., Shen, D. & Initiative, A. s. D. N., Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage* **59** (2), 895–907 (2012).
25. Zeng, N., Li, H. & Peng, Y. A new deep belief network-based multi-task learning for diagnosis of Alzheimer's disease. *Neural Comput. Appl.* **35** (16), 11599–11610 (2023).
26. An, Y., Xia, X., Chen, X., Wu, F. X. & Wang, J. Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF. *Artif. Intell. Med.* **127**, 102282 (2022).
27. Liu, Y. et al. Roberta: A robustly optimized bert pretraining approach. Preprint at https://arXiv.org/1907.11692 (2019).
28. Yang, Z. et al. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems* **32** (2019).
29. Yu, M., Han, J., Wu, H., Yan, J. & Zeng, R. Short-Term wind power prediction based on wind2vec-BERT model. *IEEE Trans. Sustain. Energy* (2024).
30. Hu, Y., Ye, K., Kim, H. & Lu, N. Bert-pin: A bert-based framework for recovering missing data segments in time-series load profiles. *IEEE Trans. Industr. Inf.* (2024).

Acknowledgements

This paper was supported by Chongqing Technical Innovation and Application Development Special General Project (CSTB2024TIAD-GPX0054), and partially supported by school level research projects (120777), supported by the experimental conditions of Chongqing University of Posts and Telecommunications and Chongqing University.

Author contributions

Xinping Chen. wrote the main manuscript text and Yan Du prepared the data. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.