DiseaseNet: A Novel Disease Diagnosis Deep Framework via Fusing Medical Record Summarization

Rushan Long, Dan Yang, Yang Liu

Abstract—**The analysis of patients' Electronic Medical Records (EMR) data through deep learning is helpful for disease diagnosis, disease risk factor assessment, disease risk prediction, etc. The medical texts of patients in EMR contain rich information since the medical text is usually stored in the form of documents, some redundant and irrelevant content may be generated, which is unimportant to patients diagnosed with diseases, and they hurt the overall training efficiency and effectiveness of the model. At present, in the research on automatic disease diagnosis, the structured and unstructured medical text data of patients are rarely considered simultaneously, and they cannot be effectively fused. This paper effectively uses the structured and unstructured medical text data of patients in EMR and proposes a novel disease diagnosis deep framework DiseaseNet via fusing medical record summarization. Firstly, the framework generates the medical record summarization of the patient's medical text, to obtain fine-grained and more relevant patient information. The structured data of patients and medical record summarization use the BERT model for fusion and embedding, and then the deep features of patients were extracted by Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN), respectively. BiLSTM integrates the output of the self-attention layer with the feature representation of the CNN output. Many experimental results on the real intensive care patient dataset MIMIC-III showed that DiseaseNet can significantly improve the performance of disease diagnosis by fusing the features of patients' structured data with the medical record summarization generated by unstructured medical text.**

Index Terms—**Disease Diagnosis; Deep Learning; Summary Generation; BERT Model; Electronic Medical Records**

I. Introduction

EMR is an electronic condition file that records the doctor's diagnosis and development of the patient's condition. The patient information is stored in data, which usually includes the patient's test results, medication, procedures, diagnostic codes, vital signs and other structured data, unstructured medical text descriptions, and medical images. EMR is used to mine the patient's valuable feature data, for example, automatic disease diagnosis based on EMR data can assist doctors in disease diagnosis, disease risk prediction, patient similarity research [1-2], and evaluation of disease risk factors.

Most previous studies on EMR-oriented disease diagnosis tasks used deep learning methods on limited patient structured data, and they ignored rich unstructured medical text information. They didn't effectively integrate structured and unstructured data in EMR. In recent years, with the continuous development of NLP (Natural Language Processing) technology, large-scale pre-training models are also increasing, from the relatively early Word2vec [3], Glove [4], and Elmo [5] to the current BERT (Bidirectional Encoder Representations from Transformers) [6] pre-training model based on Transformer [7]. Compared with the previous models, BERT can capture the real bidirectional context semantics and fine-tune tune BERT on downstream tasks. For example, text classification, matching, reading comprehension, machine translation, and other tasks have shown strong performance. At present, relevant work has applied NLP technology to patients' unstructured text data in EMR, and then predicted medical events and made some progress. For example, related work [8] uses CNN in sentence-text classification, Multi-layer CNN was used to capture the local features of more patient text sequences. Because Long Short Term Memory (LSTM) can capture the semantic features of text sequences, related work [9] proposed cancer text classification based on LSTM, which showed good performance.

Although BERT performs well in multiple NLP tasks, the use of BERT models for disease diagnosis in patients based on EMR data still faces the following problems and challenges:

- **How to effectively extract and utilize patients' medical text information**. Usually, EMR in the patient's medical text is stored in the form of a document, due to the training model BERT cannot be directly embedded into the document level, only suitable for sentence and paragraph level tasks, if the use BERT is directly truncated or using a sliding window to embed the input document, can lead to some features of patient data information is lost or unable to capture the key information.
- **How to effectively integrate structured and unstructured text in EMR**. At present, there is little relevant research work to effectively integrate structured and unstructured text data of patients in EMR at the same time, and most pre-training models are trained on general text. If the existing BERT pre-training model is directly used to embed structured data, there will be some problems, such as the inconsistency between the encoded text and the pre-training text.

Manuscript received December 30, 2021; revised July 28, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62072084 and 67072086.

Rushan Long is a postgraduate student at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: aslongrushan@163.com).

Dan Yang is a professor at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: asyangdan@163. com).

Yang Liu is an associate professor at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (corresponding author, e-mail: liuyang_lnas@163.com).

- **How to accurately capture the deep semantic features of patients**. At present, in the work related to disease diagnosis models, large-scale pre-training models are rarely used to capture the semantic information of the patient text. In addition, in the existing processing of patient text features, LSTM, CNN, and other models are usually used to capture the features. However, using the LSTM network can only capture the features in a single direction of the sequence, which is difficult to cover the important feature level in the patient's text. When using CNN to capture the features of the text sequence, it is difficult to capture the global semantic features of the text sequence. Therefore, a single LSTM or CNN cannot fully extract the deep semantic features of the patient.

Given the above problems and challenges, this paper applied BART (Bidirectional and Auto-Regressive Transformers) pre-training model to automatically generate the summary, compress the unstructured documents in EMR, generate a medical record summarization with key information and controllable length, fuse it with the structured data of patients in EMR, and embeds it with BERT large-scale pre-training model, BiLSTM and CNN are used to further extract the deep features of patients, respectively. BiLSTM can capture the contextual semantic features. Considering the importance between context words and word features, we make the hidden layer vector output by BiLSTM pass through the self-attention layer. Finally, it is fused with the local features captured by CNN. This paper extracted the patient's medical record data from the real MIMIC-III data set for disease diagnosis. Through experimental comparison, our model is superior to other methods. The main contributions are as follows:

- A novel disease diagnosis deep framework DiseaseNet via fusing medical record summarization is proposed which has applied the BERT pre-training model to fuse and embed the features of patients. BiLSTM+self-Attention and CNN were used to further capture the deep features of patients, and the disease was diagnosed by fusion of deep feature information.
- In the task of disease diagnosis, this paper extracts patient data from the real MIMIC-III data set and compares it with the baseline model. DiseaseNet effectively integrate patient structured data and medical record summarization.

II. Related Work

A. Text Summary Generation

In the field of NLP, abstract refers to summarize a given article, that is, to keep it as concise as possible while ensuring that it can reflect the main content of the article. A good abstract can quickly carry out information retrieval, reduce redundant information in the retrieval process, and improve user experience.

Abstract generation has been an important task in the NLP field with the development of deep learning. In relevant Reference, related work [10] proposed a general end-to-end sequence learning method in which a multi-layer LSTM network is used to project the input sequence to a fixed-dimensional vector, and then another LSTM is used to decode the target sequence from the vector. Later, related work [11] proposed a local attention-based method to generate a summary for each word of the input sentence. Related work [12] further improved related work [11] and proposed to encode sentences based on a convolutional attention network and generate summaries of input sentences by Recurrent Neural Network (RNN). Related work [13] propose a selective encoding model to extend the sequence-to-sequence framework for summary generation, which consists of a sentence encoder, a selection gating network, and an attention decoder, which achieves good performance on sentence summary generation tasks. In recent years, pre-training models are usually integrated into neural network models. Related work [14] proposed the BART pre-training model, which combined bidirectional encoder and one-way autoregressive decoder, and achieved the best performance in automatic text summary generation task.

B. Classifiers in Deep Learning

In recent years, deep learning technology has developed rapidly and has brought about revolutionary changes in NLP, CV (Computer version), and other fields. In terms of EMR data, relevant studies have used deep learning methods to establish risk prediction models for diseases, including CNN and RNN models. For example, Deep Neural Network (DNN) and LSTM models were used in related work [15], and the average performance was significantly improved in the prediction of infectious diseases. Related work [16] used the CNN network to predict future disease events based on EMR data of 300,000 patients over 4 years. Related work [17] obtained n-gram feature representation in sentences through TextCNN one-dimensional convolution and extracted shallow features of texts, which performed well in the short text classification. However, in the field of long texts, TextCNN mainly extracts features by a window, which was limited in long-distance modeling and insensitive to word order. In the prediction of heart failure, related work [18] took the lead in using the RNN network to analyze the pre-clinical time sequence relationship in EMR. To learn better word representation, related work [19] uses BiLSTM network architecture to pre-train a character-based language model to generate a contextual representation of words. To fully capture the feature information of sentences, a model based on CNN-BiGRU and attention mechanism was proposed in related work [20]. The model used CNN to extract features, and then used Bidirectional Gated Recursive Unit (BiGRU) for continuous learning. Finally, the accuracy of the model was improved in emotional text classification through the attention network. Related work [21] focuses on words that are important to the classification of emotional polarity in sentences through the attention mechanism and combines the advantages of CNN in extracting text local features and the BiGRU network in extracting semantic information of long text context to improve the text feature extraction ability of the model. In 2018, the BERT pre-training model, which consists of multi-layer bidirectional encoders, achieved excellent results in NLP text classification, machine translation, text matching, reading comprehension, and other tasks. Related work [22] adopted the BERT-BiLSTM-Attention model, which

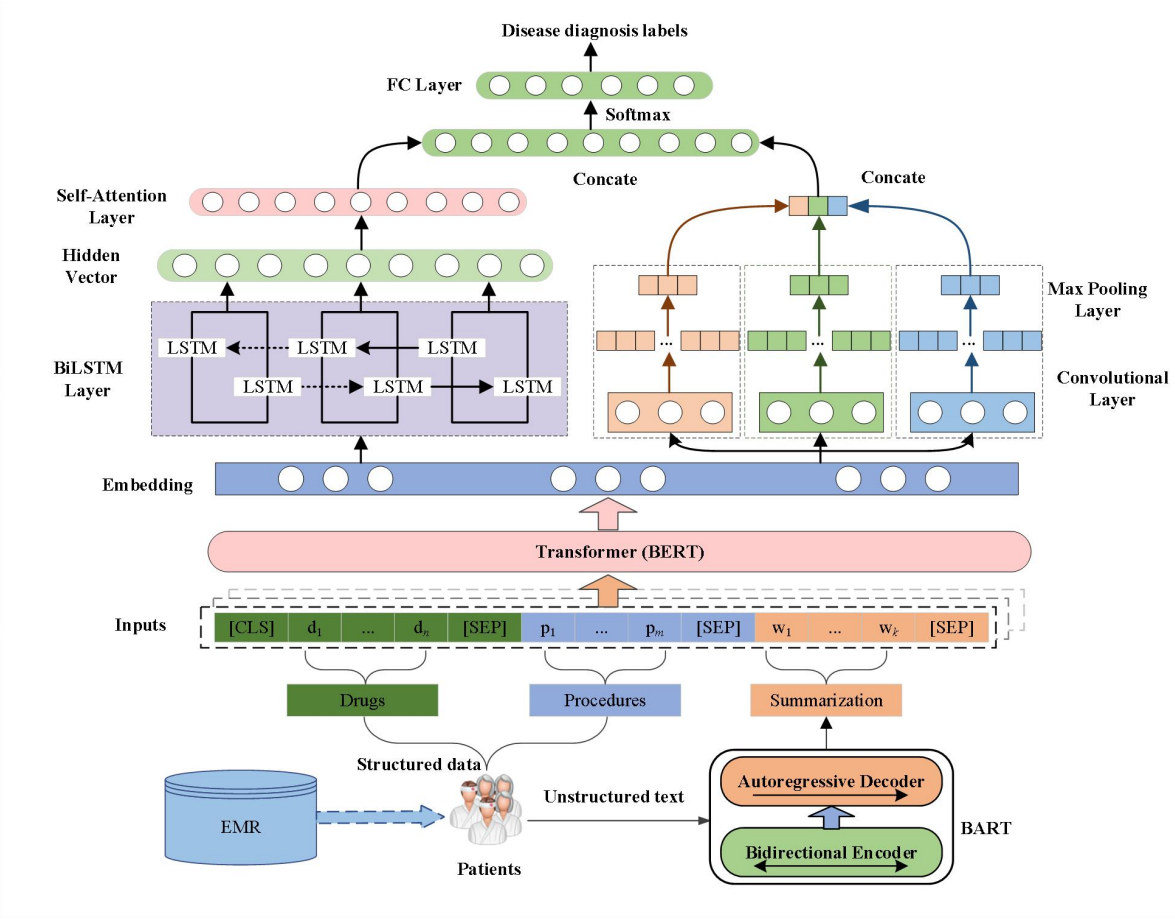


Fig. 1 Disease Diagnosis Deep Framework DiseaseNet via Fusing Medical Record Summarization

significantly improved its performance in text sentiment analysis.

C. Neural Network Model Integrating Multiple Features

To better fuse feature information, relevant studies have used neural networks to fuse external features in various tasks. Experiments show that this is conducive to improving the experimental performance of the task. In the named entity recognition task, related work [23] used CNN-LSTM-CRF neural network model to jointly train the Chinese Named Entity Recognition (CNER) and word segmentation model, it improves the recognition ability of the CNER model on the boundary of named entities. In the machine translation task, related work [24] integrates syntactic information in the machine translation task to improve performance. In-text classification, related work [25] improves the effectiveness of classification by integrating different similarity measures of citation information and collection structure content (such as title and abstract). Related work [26] mainly integrates their topic model and BERT to judge semantic similarity. The effect is more obvious when using this model in specific fields. Related work [27] proposed the TaBERT model, which combines unstructured and structured table data and is constructed on the BERT model, which can linearize the table structure to adapt to the BERT model based on the transformer. In the field of medical and health, related work [28] proposes a deep framework for detecting depression on social media by combining user behavior and their language patterns (mainly including users' social interactions) to detect user depression. The framework consists of a CNN and attention augmented GRU network and achieves good empirical performance on depression detection.

III. The Proposed Framework

Based on EMR patient data, we proposed a novel disease diagnosis deep framework DiseaseNet via fusing medical record summarization. DiseaseNet is shown in Fig. 1. The goal of DiseaseNet is to input structured and unstructured medical text data, and predict the patient's disease diagnosis labels, such as diabetes, heart failure, stomach disease, etc. in EMR, although patients have rich medical texts, they are usually long and can contain unrelated information, to effectively utilize and integrate medical information of patients. DiseaseNet first applies the BART automatic summary generation pre-training model to extract the key information in the patient's medical text, and then integrates and embeds the patient's structured data such as drugs and procedures with the generated medical record summarization using the BERT model. Each patient U_i has a drug set $Drugs_i = \{d_1, d_2, ..., d_n\}$ (n represents the number of drugs used by the patient U_i), a procedure set $Procedures_i = \{p_1, p_2, ..., p_m\}$ (m represents the number of programs done by the patient U_i), and a generated $Summarization_i = \{w_1, w_2, ..., w_k\}$, where k represents that there are k words in the $Summarization_i$ of the patient U_i. To capture the deep features of the patient, the embedded representation of the BERT output sequence passed through BiLSTM and CNN respectively. To capture the importance between BiLSTM output words and word features, the BiLSTM output

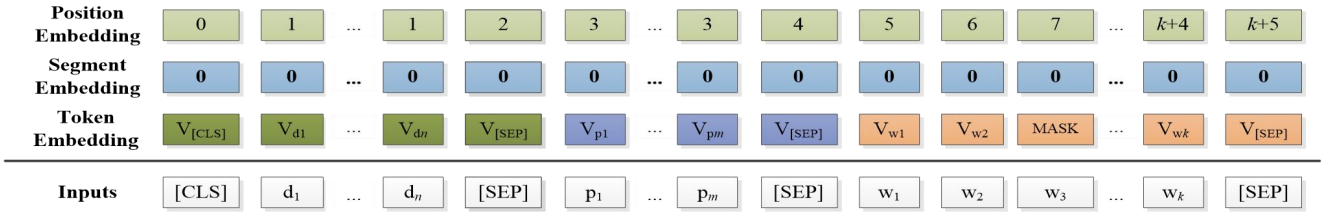


Fig. 2 Input Representation of Patient Data

sequence features passed through the self-Attention layer. And then, the features of CNN output and self-attention output are fused. Finally, the disease diagnosis $label_i$ of patient U_i is obtained through the full connection layer.

A. Generation of Patient's Medical Record Summarization Based on BART Pre-training Model

The patient's medical text in EMR contains rich feature information, which plays an important role in the patient's disease diagnosis, but it is usually lengthy and may contain irrelevant information. To extract the key information in the medical text, DiseaseNet generates the medical record summarization based on the strong automatic text summary generation pre-training model BART based on the context language. The generated medical record summarization only focuses on most non-redundant information in the patient's medical text and summarizes the key information, and the length of the generated summary can be controlled within a certain range (such as [50,150]) to facilitate the input of subsequent models. This model is a sequence-to-sequence model. It combines bidirectional encoder coding and unidirectional autoregressive decoder decoding. Firstly, the patient's medical text is encoded by a bidirectional encoder. Then, the one-way autoregressive decoder is used to generate the medical record summarization. For example, given the medical text sequence $X = \{x_1, x_2, \ldots, x_n\}$, where n represents the length of the medical text, (x_1, x_2,..., x_n) is a single word in the medical text, the sequence X is input into the BART model, encoded by the two-way coder, and then decoded by the one-way autoregressive decoder, get the medical record summarization sequence $Y = \{y_1, y_2, \ldots, y_m\}$ generated by the output of BART pre-training model, where m is the length of medical record summarization and m is less than the length of medical text n, (y_1, y_2,..., y_m) is a single word in medical record summarization.

B. Embedded Representation of Patient Feature Fusion Based on BERT Model

1) Patient Feature Fusion Based on BERT Model

BERT pre-training model not only makes full use of large-scale unmarked annotation to mine rich semantic information but also further deepens the depth of the NLP model. BERT pre-training task is composed of Masked Language Model (MLM) and Next Sentence Prediction (NSP). Firstly, we fuse the patient's drugs, procedures, and medical record summarization through linear splicing, in which the start position of the spliced sequence is marked with special characters [CLS], and the middle and end positions of drugs, procedures, and medical record summarization are marked with special characters [SEP]. Then, the sequence is transformed into the input representation vector of the BERT model. The sequence is encoded by Transformer bidirectional encoder to obtain the context semantic representation vector of patient features.

2) Input Representation of Patient Data

In BERT, a mask ratio of 15% is adopted for the patient's input sequence, that is, [mask] is used to replace the original word. To alleviate the inconsistency between the training stage and the prediction process. We choose three operations for the mask according to probability: (1) the mask holds the [mask] flag with an 80% probability. (2) The mask is randomly replaced with any word in the thesaurus with a probability of 10%. (3) The mask has a 10% probability of restoring sequence words. The input representation of patient data is shown in Fig. 2, which is composed of the sum of token embeddings, segment embeddings, and position embeddings respectively. The dimension of the three vectors is e, and the input representation vector corresponding to the following calculation sequence is used to represent the vector $\boldsymbol{v}$:

$$\boldsymbol{v} = \boldsymbol{v^t} + \boldsymbol{v^s} + \boldsymbol{v^p} \quad (1)$$

where, $\boldsymbol{v^t}$ represents word vector, $\boldsymbol{v^s}$ represents segment vector and $\boldsymbol{v^p}$ represents position vector. The sizes of the three vectors are $N \times e$. N represents the maximum length of the sequence and e represents the dimension of the word vector. The three vectors are calculated as follows:

a) Patient word vector. The spliced sequence of patient data is transformed into a real value vector representation through a word vector matrix. The patient input sequence representation vector is automatically learned in the process of model training, and the global semantic information of the sequence context can be learned and fused with the semantics of words. Specifically, supposing that the vector corresponding to the input sequence x represents $\boldsymbol{e^t} \in R^{|v| \times e}$, and its corresponding word vector is represented as $\boldsymbol{v^t}$.

$$\boldsymbol{v^t} = \boldsymbol{e^t} \cdot \boldsymbol{W^t} \quad (2)$$

Where, $\boldsymbol{W^t} \in R^{|v| \times e}$ represents the trainable word vector matrix, $|v|$ represents the size of the vocabulary, $\boldsymbol{v^t} \in R^{N \times e}$.

b) Patient segment vector. A segment vector is used to encode the segment to which the current word belongs. The segment encoding corresponding to each word in the input sequence is the sequence number of the current word in the segment. When the input sequence is a segment, the segment encoding of all words is 0. When the input sequence is two segments, the segment encoding corresponding to each word in the first segment is 0, and the segment code corresponding to each word in the second segment is 1. When the input sequence is n segments, the segment code corresponding to each word in the first segment is 0, and the segment code corresponding to each word in the second segment is 1. The segment code corresponding to each word in the nth segment is (n - 1), where (n = 1, 2...). We input the spliced patient sequence, set all the segment codes corresponding to the sequence to 0,

and use the segment vector matrix to convert the segment code into a real value vector to obtain the segment vector $\boldsymbol{v^s}$ of the patient.

$$\boldsymbol{v^s} = \boldsymbol{e^s} \cdot \boldsymbol{W^s} \tag{3}$$

Where, $\boldsymbol{W^s} \in R^{|s|\times e}$ represents the trainable segment vector matrix, $|s|$ represents the number of segments, $\boldsymbol{v^s} \in R^{N\times e}$.

c) Patient position vector. The position vector is used to encode the absolute position of each word in the patient sequence. In the structured data used by patients, it is considered that there is no order between drugs and procedures. We fixed the absolute position of the patient's drug and procedure. Each word in the input sequence is converted into a position code according to its absolute position, and then the position code $\boldsymbol{e^p} \in R^{N\times N}$ is converted into a real value vector by using the position vector-matrix $\boldsymbol{W^p}$ to obtain the patient's position vector $\boldsymbol{v^P}$. The representation of $\boldsymbol{v^P}$ is as follows:

$$\boldsymbol{v^p} = \boldsymbol{e^p} \cdot \boldsymbol{W^P} \tag{4}$$

where, $\boldsymbol{W^P} \in R^{N\times e}$ represents the trainable position vector matrix, N represents the maximum position length, $\boldsymbol{v^P} \in R^{N\times e}$.

We linearize and splice each patient's U_i data to obtain a sequence X_i, which is summarized as follows:

$$\begin{aligned} X_i = {} & [CLS]d_1d_2\cdots d_n[SEP]p_1p_2\cdots p_m \\ & [SEP]w_1w_2\cdots w_k[SEP] \end{aligned} \tag{5}$$

where, [*CLS*] bit represents the start mark of patient sequence input, [*SEP*] bit represents the end mark character of each segment in the sequence, where $Drugs_i = \{d_1, d_2, ..., d_n\}$, $Drugs_i$ represents the drug set of patients U_i, $Procedures_i = \{p_1, p_2, ..., p_m\}$, $Procedures_i$ represents the assembly of patient U_i, $Summarization_i = \{w_1, w_2, ..., w_k\}$, $Summarization_i$ represents the patient's medical record summarization.

For the original input sequence X_i of patient U_i, the input representation vector $\boldsymbol{V_i}$ of the BERT model is obtained through input representation processing, and the input of all patients represents vector $\boldsymbol{V} = [\boldsymbol{V_1}, \boldsymbol{V_2}, ..., \boldsymbol{V_N}]$, where is $\boldsymbol{V_i}$:

$$\boldsymbol{V}_i = InputRepresentation(X_i) \tag{6}$$

where, $\boldsymbol{V} \in R^{N\times e}$ represents the final output result of the input layer, that is, the sum of word vector, segment vector, and position vector, and N represents the maximum sequence length.

The input representation vector $\boldsymbol{V}$ passes through the BERT model to obtain the patient feature representation vector $\boldsymbol{X}^{'}$ encoded and output by the BERT model:

$$\boldsymbol{X}^{'} = BERT(\boldsymbol{V}) \tag{7}$$

C. Extract Deep Features of Patients

To extract the patient's deep feature information, DiseaseNet uses BiLSTM and CNN to further extract the patient feature representation vector output by BERT. BiLSTM can capture the context semantic features, and CNN can capture the features between local words. First, the hidden layer vector output by BiLSTM is passed through the self-Attention layer, and then the self-Attention layer is spliced and fused with the output vector of CNN, Finally, through a full connection layer, the patient's disease diagnosis label is obtained.

1) BiLSTM Layer and self-attention in DiseaseNet

BiLSTM is composed of left-right two-way LSTM superposition, which can capture the context semantic information of patient U_i feature sequence. The patient feature representation vector $\boldsymbol{X}_i^{'}$ output by BERT, as the input vector of BiLSTM, passes through the left-right LSTM network to obtain $\boldsymbol{h}_i^L$ and $\boldsymbol{h}_i^R$ vectors respectively, and then the $\boldsymbol{h}_i^L$ and $\boldsymbol{h}_i^R$ vectors are superimposed to obtain the output vector $\boldsymbol{S}_i$ of BiLSTM. The calculation process of $\boldsymbol{S}_i$ is as follows:

$$\boldsymbol{h}_i^L = LSTM(\boldsymbol{X}_i^{'}, \boldsymbol{h}_{i+1}^L) \tag{8}$$

$$\boldsymbol{h}_i^R = LSTM(\boldsymbol{X}_i^{'}, \boldsymbol{h}_{i-1}^R) \tag{9}$$

$$\boldsymbol{S}_i = \boldsymbol{W}_i^L \cdot \boldsymbol{h}_i^L + \boldsymbol{W}_i^R \cdot \boldsymbol{h}_i^R + \boldsymbol{b}_i \tag{10}$$

where, $\boldsymbol{W^L}$ and $\boldsymbol{W^R}$ are the trainable weight vectors of left and right LSTM respectively, $\boldsymbol{b}_i$ is the bias coefficient, and $\boldsymbol{S}_i$ constitutes the feature vector $\boldsymbol{S} = [\boldsymbol{S}_1, \boldsymbol{S}_2, ..., \boldsymbol{S}_n]$ of patient U output by BiLSTM.

The self-attention mechanism can calculate the time relationship between two-word vectors far away in the patient's U_i feature vector and assign a weight to them by calculating the correlation between vectors, to notice the importance between words and word features. DiseaseNet uses the self-Attention mechanism to capture the influence degree among word vectors in the BiLSTM output patient's U_i feature vector $\boldsymbol{S}_i = [\boldsymbol{s}_i^{(1)}, \boldsymbol{s}_i^{(2)}, \ldots, \boldsymbol{s}_i^{(j)}, \ldots, \boldsymbol{s}_i^{(n)}]$, where n represents the number of word vectors. Firstly, three different learnable parameter matrices $\boldsymbol{W}^q$, $\boldsymbol{W}^k$, and $\boldsymbol{W}^v$ are used to transform the patient U_i word vector $\boldsymbol{s}_i^{(j)}$ into three new vectors $\boldsymbol{q}_i^{(j)}$, $\boldsymbol{k}_i^{(j)}$, and $\boldsymbol{v}_i^{(j)}$, which are expressed as query, key and value vectors respectively. $\boldsymbol{q}_i^{(j)}$, $\boldsymbol{k}_i^{(j)}$ and $\boldsymbol{v}_i^{(j)}$ are expressed as follows:

$$\boldsymbol{q}_i^{(j)} = \boldsymbol{W}^q \cdot \boldsymbol{s}_i^{(j)} \tag{11}$$

$$\boldsymbol{k}_i^{(j)} = \boldsymbol{W}^k \cdot \boldsymbol{s}_i^{(j)} \tag{12}$$

$$\boldsymbol{v}_i^{(j)} = \boldsymbol{W}^v \cdot \boldsymbol{s}_i^{(j)} \tag{13}$$

Then calculate the similarity between each word vector in $\boldsymbol{S}_i$, get the weight value of each word vector to each word vector, normalize the weight value of each word vector through the activation function, and finally sum the weight value with each word vector. Finally, $\boldsymbol{S}_i$ obtains the patient U_i feature vector $\boldsymbol{Z}_i$ output from the Attention layer, $\boldsymbol{Z}_i = [\boldsymbol{z}_i^{(1)}, \boldsymbol{z}_i^{(2)}, \ldots, \boldsymbol{z}_i^{(n)}]$. The calculation process is as follows:

$$\boldsymbol{z}_i^{(j)} = \sum_{m=1}^{n} \boldsymbol{\alpha}_i^{jm} \cdot \boldsymbol{v}_i^{(j)} \tag{14}$$

$$\boldsymbol{\alpha}_i^{jm} = softmax(\tilde{\boldsymbol{\alpha}}_i^j)^m \tag{15}$$

$$\tilde{\boldsymbol{\alpha}}_i^{jm} = \frac{\boldsymbol{q}_i^{(j)} \cdot (\boldsymbol{k}_i^{(m)})^T}{\sqrt{d_k}} \tag{16}$$

where, d_k is used to adjust to avoid too large a vector dot product. $\tilde{\boldsymbol{\alpha}}_i^j = [\tilde{\boldsymbol{\alpha}}_i^{j1}, \tilde{\boldsymbol{\alpha}}_i^{j2}, \ldots, \tilde{\boldsymbol{\alpha}}_i^{jm}, \ldots, \tilde{\boldsymbol{\alpha}}_i^{jn}]$, $\boldsymbol{\alpha}_i^{jm}$ is the attention weight of word vector $\boldsymbol{s}_i^{(j)}$ to word vector $\boldsymbol{s}_i^{(m)}$ in $\boldsymbol{S}_i$.

2) Convolution Layer of CNN in DiseaseNet

The convolution layer of CNN is used to capture the local features of the patient feature sequence embedded by the BERT model. BERT in DiseaseNet is encoded by words, and the patient feature output by BERT represents the vector $\boldsymbol{X}^{'}$, and each word vector is represented as $\boldsymbol{X}_i^{'}$. To capture the local features between h adjacent word vectors in $\boldsymbol{X}^{'}$, $\boldsymbol{X}^{'}$ obtains the feature matrix $\boldsymbol{T}$ through the CNN layer, $\boldsymbol{T} = [\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_i, \ldots, \boldsymbol{C}_n]$, where is $\boldsymbol{C}_i$:

$$\boldsymbol{C}_i = f(\boldsymbol{W} \cdot \boldsymbol{X}_{i:i+h-1}^{'} + b) \tag{17}$$

Admission Date: [**2118-6-2**] Discharge Date: [*2118-6-14**]
Date of Birth: ****
Sex: F
Service: M ICU and then to [**Doctor Last Name **]
Medicine HISTORY OF PRESENT ILLNESS: This is an 81-year-old female with a history of emphysema (not on home O2), who presents with three days of shortness of breath thought by her primary care doctor to be a COPD flare. Two days before admission, she started on a prednisone taper, and one day before admission she required oxygen at home to maintain oxygen saturation greater than 90%. She has also been on levofloxacin and nebulizers, and was not getting better, and presented to the [**Hospital 18**] Emergency Room. In the [**Hospital 3 **] Emergency Room, her oxygen saturation was 100% on CPAP. She was not able to be weaned off this despite nebulizer treatment and Solu-Medrol 125 mg IV x2. A review of systems is negative for the following: fever, chills, nausea, vomiting, night sweats, change in weight, gastrointestinal complaints, neurologic changes, rashes, palpitations, and orthopnea. It is positive for the following: Chest pressure occasionally with shortness of breath with exertion, some shortness of breath that is positively related but is improved with nebulizer treatment. PAST MEDICAL HISTORY: 1. COPD. Last pulmonary function tests in [**2117-11-3**] demonstrated an FVC of 52% predicted, an FEV1 of 54% predicted, an MMF of 23% predicted, and an FEV1: FVC ratio of 67% predicted, that does not improve with bronchodilator treatment. The FVC, however, does significantly improve with bronchodilator treatment consistent with her known reversible airflow obstruction in addition to an underlying restrictive ventilatory defect. The patient had never been on oxygen before this recent episode. She has never been on steroid taper or been intubated in the past. Lacunar CVA. MRI of the head in [**2114-11-4**] demonstrates a mild degree of multiple small foci of high T2.

Fig. 3 Sample Medical Text Data for Patients

where, $\boldsymbol{W} \in R^{h\times e}$, h is the height of convolution kernel $\boldsymbol{W}$, e is the dimension of word vector $\boldsymbol{X}_i^{'}$, b is the bias coefficient, and f is the nonlinear function.

Three convolution layers of CNN with different windows are selected in DiseaseNet to obtain three feature matrices $\boldsymbol{T}_1$, $\boldsymbol{T}_2$, and $\boldsymbol{T}_3$ respectively. To make the dimension of the feature matrix consistent, $\boldsymbol{T}_1^{'}$, $\boldsymbol{T}_2^{'}$, and $\boldsymbol{T}_3^{'}$ are obtained respectively through CNN's max-pooling layer. Finally, the feature matrix $\boldsymbol{F}$ is obtained through the splicing function:

$$\boldsymbol{T}^{'} = maxpooling(\boldsymbol{T}) \tag{18}$$

$$\boldsymbol{F} = concate(\boldsymbol{T}_1^{'},\ \boldsymbol{T}_2^{'},\ \boldsymbol{T}_3^{'}) \tag{19}$$

Through the above calculation, DiseaseNet fused the patient feature matrix output by self-Attention and the patient feature matrix spliced by CNN as the input of the full connection layer (FC) and use the SoftMax activation function to obtain patients' disease diagnostic labels. The calculation process of *labels* is as follows:

$$labels = SoftMax(FC(concate(\boldsymbol{Z},\ \boldsymbol{F}))) \tag{20}$$

IV. Experiments

A. Data set description and preprocessing

MIMIC-III (Medical Information Mart for Intensive Care III) [29] is a large multi-parameter intensive care medical information data set that is free and open to the public. It records the data of ICU patients from Beth Israel Dikang Medical Center (BIDMC) from 2001 to 2012. Its purpose is to develop and evaluate an advanced ICU monitoring system. The clinical data set records the clinical information of more than 60000 ICU patients, including vital signs, drugs, experimental measurements, medical texts, procedure information, diagnostic information, length of hospital stay, and survival data. All data resources in the MIMIC-III dataset are strictly identified. We select the patients' drugs, procedures, disease diagnosis categories, and medical text data from the MIMIC-III dataset for disease diagnosis. The medical text mainly includes some basic information about the patient (such as date, age, and gender), current medical history, family history, inspection reports, and other recorded information. The text data sample of the intercepted part is shown in Fig. 3.

According to the MIMIC-III dataset, the disease Diagnosis Related Groups (DRG) [30] and the International Classification of Diseases Ninth Edition (ICD-9) [31] diagnostic codes. We selected 6 common diseases: diabetes, heart failure, sepsis, respiratory failure, gastritis, and atherosclerosis. The patients' data on the drugs, procedures, medical texts, and disease diagnosis categories were extracted from these patients' records. We then removed such patients with missing patient medical texts. Secondly, if the patient's drug and procedure data are missing, such patient records are retained and filled with the special character 'SEP'. Finally, for the patient's drug use, only the type of drug is selected as the 'MAIN'. Then, the number of drugs for each patient is randomly selected and controlled within 30. The procedure data is sorted according to the importance, and the number of procedures for each patient is selected within 3. After data preprocessing, 9592 patients are finally selected. The statistics on the number of patients with different diseases are shown in Table I. The sample data of patients are shown in Table II.

TABLE I
STATISTICS ON THE NUMBER OF PATIENTS WITH DIFFERENT DISEASES

Disease label	Number of patients
Diabetes	660
Heart Failure	1266
Septicemia	2212
Atherosclerosis	3564
Gastritis	427
Respiratory Failure	1463
Total	9592

B. Experimental Parameter Setting

We applied the BART-Large-CNN text summary generation pre-training model, set the maximum length of the patient's medical text to 1024, set the length interval of the medical record summarization automatically generated by the BART model to [50, 150], and used the BERT-Base-Uncased pre-training model with 12 layers, 768 hidden units, and 110 M parameters. For the BiLSTM layer, the hidden unit is set to 128, the number of attention layer hidden units is set to 64, the number of three CNN windows is 128, and the height of each CNN window is 3, 4, and 5 respectively, and the hidden layer unit of maximum pool layer output is 128. The Adam optimizer [32] is used in the training stage of DiseaseNet, and the learning rate is set to 3e-5, dropout is 0.5, the epoch is 5, and batch size is set to 16, the loss function uses the cross-entropy loss function, and the dataset was divided by 5-fold cross-validation.

C. Evaluation Metrics

To reduce the impact of the imbalance in the number of

TABLE II
SAMPLE DATA OF PATIENTS

Subject_id	Hadm_id	Drugs	Procedures	Medical record summarization	Diagnostic result
12765	165123	Magnesium sulfate. Potassium chloride. Iso-osmotic dextrose. Oxycodone-acetaminophen. Sodium chloride 0.9% flush. Amiodarone HCL. Isosorbide dinitrate.	*ICD9_CODE:3774 Insertion or replacement of epicardial lead [electrode] into epicardium. *ICD9_CODE:3783 Initial insertion of the dual-chamber device.	The patient was admitted for epicardial LV lead placement, which he underwent successfully without complication. He still has an R chest tube in place and his pacer appears to be functioning appropriately.	Heart Failure
24995	167081	Insulin. Vancomycin HCL. Iso-osmotic dextrose. Sodium bicarbonate. Insulin human regular. Calcium gluconate. Sevelamer. Ciprofloxacin.	*ICD9_CODE:3995 Hemodialysis. *ICD9_CODE:3895 Venous catheterization for renal dialysis.	The patient was admitted to the hospital with altered mental status. She was diagnosed with DM type 1 x 35 years, chronic renal failure, peripheral neuropathy, proliferative retinopathy (left eye blindness), and Carpal Tunnel Syndrome. The patient has no history of illicit drug use.	Diabetes

disease categories in the dataset on the performance evaluation of the model, macro average precision (Macro_P), macro average recall (Macro_R), and macro average F1 (Macro_F1) are used to evaluate the performance of the model.

The meanings of evaluation metrics parameters are shown in Table III. The real disease category and predicted disease category of patients are divided into:

- True Positive (TP): the real patient's disease category is positive, and the predicted patient's disease category is positive.
- False Positive (FP): the real patient's disease category is negative, and the predicted patient's disease category is positive.
- False Negative (FN) the real patient's disease category was positive, and the predicted patient's disease category was negative.
- True Negative (TN): the real patient's disease category is negative, and the predicted patient's disease category is negative.

Among them, the calculation formulas of precision (P), recall (R), and F1 are as follows:

TABLE III
MEANING OF EVALUATION METRICS PARAMETERS

True Value	Predictive Value	
	Positive Example	Negative Example
Positive Example	TP	FN
Negative Example	FP	TN

$$P = \frac{TP}{TP + FP} \tag{21}$$

$$R = \frac{TP}{TP + FN} \tag{22}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{23}$$

Macro_P, Macro_R, Macro_F1 is the arithmetic mean of P, R, and F1 values of each disease category, and the formula is:

$$Macro\_P = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{24}$$

$$Macro\_R = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{25}$$

$$Macro\_F = \frac{1}{n}\sum_{i=1}^{n} F_i \tag{26}$$

D. Experimental Results and Analysis

Firstly, we used structured data of patients for disease diagnosis through DiseaseNet. We compare them with CNN, BiLSTM, BERT, BERT+BiLSTM, BERT+ BiLSTM+att (Attention), and BERT+CNN models respectively. The comparison results are shown in Table IV. The BERT+BiLSTM+att model used in Macro_P reached the highest evaluation metrics, 78.3%, while DiseaseNet in Macro_R and Macro_F1 reached the highest evaluation metrics, 78.6%, and 78.2% respectively. DiseaseNet has achieved good performance in disease diagnosis on patients' structured data.

TABLE IV
COMPARISON RESULTS OF DIFFERENT MODELS USING STRUCTURED DATA FOR DISEASE DIAGNOSIS

Model	**Macro_P**	**Macro_R**	**Macro_F1**
CNN	0.732	0.646	0.676
BiLSTM	0.747	0.731	0.737
BERT	0.758	0.768	0.761
BERT+BiLSTM	0.777	0.765	0.766
BERT+BiLSTM+att	**0.783**	0.761	0.768
BERT+CNN	0.768	0.774	0.768
DiseaseNet	0.781	**0.786**	**0.782**

Using the patient's medical record summarization, the above model is still used for disease diagnosis. Through comparison, it can be seen from Table V that DiseaseNet performs disease diagnosis on the patient's medical record summarization, and all metrics reach the highest, Macro_P, Macro_R, Macro_F1 was 75.6%, 68.8%, and 71.3% respectively. Our DiseaseNet model achieved the best performance in disease diagnosis based on generated medical record summarization.

TABLE VI
COMPARISON RESULTS OF DIFFERENT PATIENT FEATURES IN DISEASE DIAGNOSIS

Feature	Model	Macro_P	Macro_R	Macro_F1
Structured Data	CNN	0.732	0.646	0.676
	BiLSTM	0.747	0.731	0.737
	BERT	0.758	0.768	0.761
	BERT+BiLSTM	0.777	0.765	0.766
	BERT+BiLSTM+att	**0.783**	0.761	0.768
	BERT+CNN	0.768	0.774	0.768
	DiseaseNet	0.781	**0.786**	**0.782**
Medical Record Summarization	CNN	0.530	0.535	0.529
	BiLSTM	0.619	0.574	0.591
	BERT	0.745	0.672	0.694
	BERT+BiLSTM	0.730	0.667	0.691
	BERT+BiLSTM+att	0.732	0.675	0.696
	BERT+CNN	0.743	0.685	0.709
	DiseaseNet	**0.756**	**0.688**	**0.713**
Structured Data and Medical Record Summarization	CNN	0.736	0.709	0.719
	BiLSTM	0.755	0.771	0.759
	BERT	0.830	0.818	0.821
	BERT+BiLSTM	0.826	0.814	0.816
	BERT+BiLSTM+att	0.828	0.830	0.827
	BERT+CNN	0.827	0.820	0.823
	DiseaseNet	**0.832**	**0.832**	**0.831**

TABLE V
COMPARISON RESULTS OF DIFFERENT MODELS USING MEDICAL RECORD SUMMARIZATION FOR DISEASE DIAGNOSIS

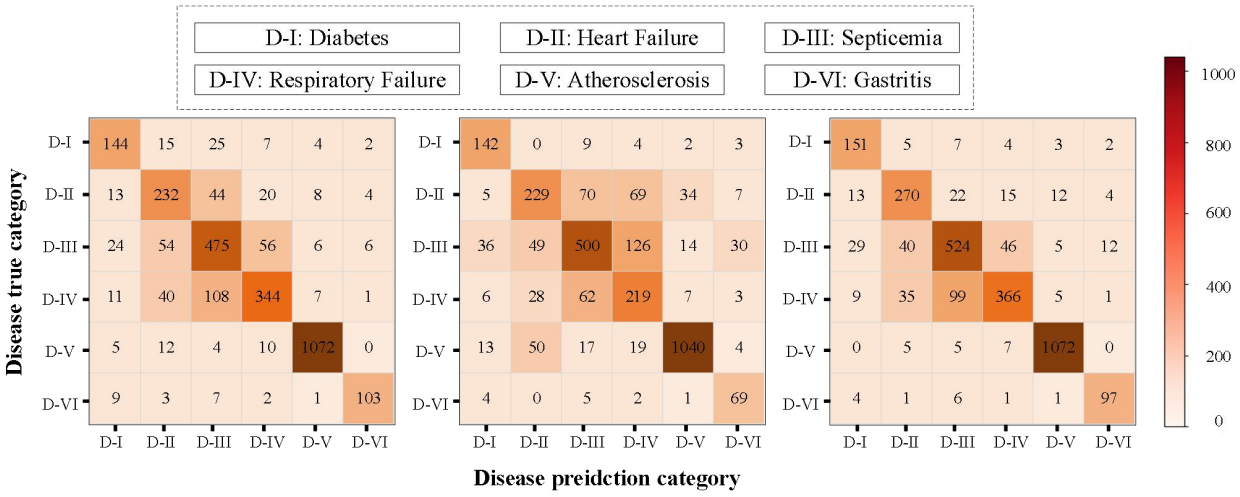
Model	Macro_P	Macro_R	Macro_F1
CNN	0.530	0.535	0.529
BiLSTM	0.619	0.574	0.591
BERT	0.745	0.672	0.694
BERT+BiLSTM	0.730	0.667	0.691
BERT+BiLSTM+att	0.732	0.675	0.696
BERT+CNN	0.743	0.685	0.709
DiseaseNet	**0.756**	**0.688**	**0.713**

Tables IV and V showed that using structured patient data alone results in better disease diagnostic performance than using generated medical record summaries. However, both structured and unstructured patient data are important in disease diagnosis. Therefore, to further improve the performance of disease diagnosis, the structured data of patients and the generated medical record summarization were fused for disease diagnosis.

We integrated the patient's drugs, main procedures, and medical record summarization. As shown in Table VI, the DiseaseNet model achieved better results than other models, Macro_P, Macro_R, and Macro_F1 are 83.2%, 83.2%, and 83.1% respectively. Comparing the disease diagnosis from different data features of patients, only the patient's structured data and the medical record summarization generated by medical text are used for disease diagnosis, and there is no fusion of the patient's structured data and the generated medical record summarization, which has a high evaluation metrics for disease diagnosis. Our DiseaseNet model can effectively integrate patients' structured data and medical record summary for disease diagnosis.

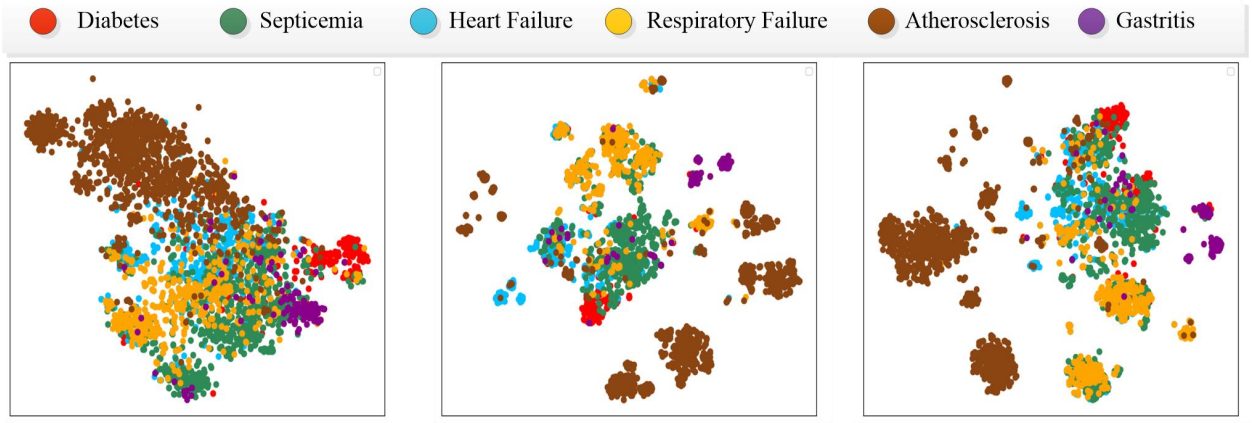
E. Visualization of confusion matrix in disease diagnosis

Fig. 4 shows the visualization of patient disease diagnosis results using the confusion matrix on the test set. As shown in Fig. 4 (a), (b), and (c) show that DiseaseNet uses patient structured data, and medical record summarization and



(a) Using Structured Data (b) Using Medical Record Summarization (c) Fusing Structured Data and Medical Record Summarization

Fig. 4 Visualization of Patient's Disease Classification Confusion Matrix



(a) Using Structured Data (b) Using Medical Record Summarization (c) Fusing Structured Data and Medical Record Summarization

Fig. 5 Visualization of Patient Disease Diagnosis

integrates structured data and medical record summarization to diagnose patients respectively. The horizontal axis in the figure represents the disease prediction category, the vertical axis represents the true category of the disease. The correct results predicted by the confusion matrix are on the diagonal. The values on the diagonal represent the number of correct diagnoses of the disease, that is, the more the number, the darker the color. According to the mutual comparison of (a), (b), and (c) in Fig. 4, Fig. 4 (c) shows that DiseaseNet is significantly more accurate in disease diagnosis than using only structured data of patients or medical record summarization.

F. Visualization of disease diagnosis

To observe the disease diagnosis effect of DiseaseNet more intuitively, on the test set, we output the vector fused by BiLSTM+self-attention layer and CNN in the DiseaseNet model, reduced the dimension and visualized the vector representation of patients on the two-dimensional plane. The visualization effect is shown in Fig. 5. We used t-SNE (t-distributed Stochastic Neighbor Embedding) [33] dimensionality reduction technology. t-SNE can map the high-dimensional data representation in space to two-dimensional or three-dimensional low-dimensional space, to realize the visualization of disease diagnosis data representation. DiseaseNet uses patient structured data, and medical record summarization and integrates structured data and medical record summarization. Visualization of patient disease diagnosis results, as shown in Fig. 5 (a), (b), and (c). Each color point in the figure represents patients with a class of diseases. A good disease diagnosis model can map the feature representation of patients with similar diseases to close high-dimensional space. When the dimension is reduced to a two-dimensional plane, the points with the same color of similar diseases gather, and the points with different colors of diseases are far away from each other. It can be observed that in Fig. 5 (c) compared with (a), (b), the (c) diagram that fuses the patient structured data and medical record summarization, points of the same color are more concentrated, which means that the diagnosis of the patient's disease is more accurate.

V. Conclusions and Future Work

Previous studies on patient disease diagnosis rarely consider both structured data and unstructured medical text at the same time and cannot effectively integrate them. We effectively used the structured and unstructured medical text data of patients in EMR and proposed a novel disease diagnosis deep framework DiseaseNet via fusing medical record summarization. DiseaseNet uses the BART pre-training model to generate the patient's unstructured medical text in EMR, and effectively fuses and embeds the patient's structured data and medical record summarization through the BERT model. To fully capture the patient's feature information, DiseaseNet introduces the BiLSTM+self-attention and CNN, to extract the context semantic information and local features of the patient's deep feature sequence. We validate that DiseaseNet effectively fuses patient structured data and medical record summarization, achieving the best performance in disease diagnosis, through a series of comparative experiments and analyses.

The next stage of research focuses on how to effectively use and integrate various types of patient feature data and consider the relationship between patients and other medical entities, such as establishing patients, disease association diagrams, and medical knowledge maps for disease diagnosis. The main work includes fusing various types and multimodal patient feature data, such as structured data, text descriptions, images, etc. Fully considering the relationship between patients and fusing the multimodal features of patients, a graph neural network is used to analyze and mine the graph structure.

References

[1] Z. Lin, and D. Yang, "Medical Concept Embedding with Variable Temporal Scopes for Patient Similarity," *Engineering Letters*, vol. 28, no. 3, pp651-662, 2020.

[2] H. Jiang, and D. Yang, "Learning Graph-based Embedding from EHRs for Time-aware Patient Similarity," *Engineering Letters,* vol. 28, no. 4, pp1254-1262, 2020.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Jan. 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781.

[4] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532-1543.

[5] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," Feb. 2018, *arXiv:1802.05365*, vol. 12, 1802. [Online]. Available: https://arxiv.org/abs/1802.05365.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language

understanding," Oct. 2018, *arXiv:1810.04805*. [Online]. Available: https://arxiv.org/abs/1810.04805.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Jun. 2017, *arXiv:1706.03762*. [Online]. Available: https://arxiv.org/abs/1706.03762.

[8] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical Text Classification using Convolutional Neural Networks," *Studies in Health Technology & Informatics,* vol. 235, p. 246, 2017.

[9] D. C. Edara, L. P. Vanukuri, V. Sistla, and V. Kolli, "Sentiment analysis and text categorization of cancer medical records with LSTM," *Journal of Ambient Intelligence and Humanized Computing,* 2019, pp. 1-17. doi: 10.1007/s12652-019-01399-8.

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, 2014, pp. 3104-3112.

[11] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," Sep. 2015, *arXiv:1509.00685*. [Online]. Available: https://arxiv.org/abs/1509.00685.

[12] S. Chopra, M. Auli, and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks," *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 93-98.

[13] Q. Zhou, N. Yang, F. Wei, and M. Zhou, "Selective encoding for abstractive sentence summarization," Apr. 2017, *arXiv:1704.07073*. [Online]. Available: https://arxiv.org/abs/1704.07073.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," Oct. 2019, *arXiv:1910.13461*. [Online]. Available: https://arxiv.org/abs/1910.13461.

[15] S. Chae, S. Kwon, and D. Lee, "Predicting infectious disease using deep learning and big data," *International Journal of Environmental Research and Public Health,* vol. 15, no. 8, p. 1596, 2018.

[16] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," *Proceedings of the 2016 SIAM International Conference on Data Mining*, 2016, pp. 432-440. doi: 10.1137/1.9781611974348.49.

[17] Y. Kim, "Convolutional Neural Networks for Sentence Classification," Aug. 2014, *arXiv:1408.5882*. [Online]. Available: https://arxiv.org/abs/1408.5882.

[18] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *Journal of the American Medical Informatics Association*, vol. 24, no. 2, pp. 361-370, 2017.

[19] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," Apr. 2017, *arXiv:1705.00108*. [Online]. Available: https://arxiv.org/abs/1705.00108.

[20] L. Y. Wang, C. H. Liu, H. B. Cai, T. Lu, and J. Z. Zhang, "Chinese text sentiment analysis based on CNN-BiGRU network with attention mechanism," *Journal of Computer Applications*, vol. 39, no. 10, pp. 2841, 2019.

[21] Y. Cheng, L. Yao, G. Xiang, G. Zhang, T. Tang, and L. Zhong, "Text sentiment orientation analysis of multi-channels CNN and BIGRU based on attention mechanism," *Journal of Computer Research and Development,* vol. 57, no. 12, p. 2583, 2020.

[22] H. Ge, S. Zheng, and Q. Wang, "Based BERT-BiLSTM-ATT Model of Commodity Commentary on The Emotional Tendency Analysis," in *2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI)*, 2021, pp. 130-133. doi: 10.1109/BDAI52447.2021.9515273.

[23] F. Wu, J. Liu, C. Wu, Y. Huang, and X. Xie, "Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation," *The World Wide Web Conference*, 2019, pp. 3342-3348. doi: 10.1145/3308558.3313743.

[24] V. Fossum, K. Knight, and S. Abney, "Using syntax to improve word alignment precision for syntax-based machine translation," *Proceedings of the Third Workshop on Statistical Machine Translation*, 2008, pp. 44-52.

[25] B. Zhang, Y. Chen, W. Fan, E. A. Fox, M. A. Goncalves, M. Cristo, and P. Calado, "Intelligent fusion of structural and citation-based evidence for text classification," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005, pp. 667-668. doi: 10.1145/1076034.107618.

[26] N. Peinelt, D. Nguyen, and M. Liakata, "tBERT: Topic models and BERT joining forces for semantic similarity detection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7047-7055. doi: 10.18653/v1/2020.acl-main.630.

[27] P. Yin, G. Neubig, W. T. Yih, and S. Riedel, "TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data," May. 2020, arXiv:2005.08314. [Online]. Available: https://arxiv.org/abs/2005.08314.

[28] H. Zogan, I. Razzak, S. Jameel, and G. Xu, "DepressionNet: A Novel Summarization Boosted Deep Framework for Depression Detection on Social Media," May. 2021, *arXiv:2105.10878*. [Online]. Available: https://arxiv.org/abs/2105.10878.

[29] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data,* vol. 3, no. 1, pp. 1-9, 2016.

[30] M. M. Wiley, "Diagnosis Related Groups (DRGs): Measuring Hospital Case Mix," *Wiley StatsRef: Statistics Reference Online,* 2014. doi: 10.1002/9781118445112.stat05313.

[31] SLEE and N. Vergil, "The International Classification of Diseases: Ninth Revision (ICD-9)," *Annals of Internal Medicine,* vol. 88, no. 3, pp. 424-426, 1978. doi: 10.7326/0003-4819-88-3-424.

[32] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Dec. 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980.

[33] S. Arora, W. Hu, and P. K. Kothari, "An analysis of the t-sne algorithm for data visualization," in *Conference on Learning Theory*, 2018, pp. 1455-1462.