TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 11/18 pp117–130
DOI: 10.26599/TST.2021.9010093
Volume 28, Number 1, February 2023

KTI-RNN: Recognition of Heart Failure from Clinical Notes

Dengao Li*, Huiting Ma, Wenjing Li, Baofeng Zhao, Jumin Zhao, Yi Liu, and Jian Fu

Abstract: Although deep learning methods have recently attracted considerable attention in the medical field, analyzing large-scale electronic health record data is still a difficult task. In particular, the accurate recognition of heart failure is a key technology for doctors to make reasonable treatment decisions. This study uses data from the Medical Information Mart for Intensive Care database. Compared with structured data, unstructured data contain abundant patient information. However, this type of data has unsatisfactory characteristics, e.g., many colloquial vocabularies and sparse content. To solve these problems, we propose the KTI-RNN model for unstructured data recognition. The proposed model overcomes sparse content and obtains good classification results. The term frequency-inverse word frequency (TF-IWF) model is used to extract the keyword set. The latent dirichlet allocation (LDA) model is adopted to extract the topic word set. These models enable the expansion of the medical record text content. Finally, we embed the global attention mechanism and gating mechanism between the bidirectional recurrent neural network (BiRNN) model and the output layer. We call it gated-attention-BiRNN (GA-BiRNN) and use it to identify heart failure from extensive medical texts. Results show that the $F1$ score of the proposed KTI-RNN model is 85.57%, and the accuracy rate of the proposed KTI-RNN model is 85.59%.

Key words: heart failure; diagnosis; text classification; deep learning

1 Introduction

Cardiovascular diseases are the leading cause of the increase in global mortality. How to treat this disease has become a major public health problem worldwide[1]. According to the American College of Cardiology, more than 400 million people have been diagnosed with this disease globally, causing one-third of total global deaths[2]. In recent years, the incidence of heart failure has been increasing, and patients with heart failure have gradually become younger[3]. Between 2012 and 2030, the prevalence of heart failure in the US will increase to 46%, and patients over 18 years old with heart failure will reach more than 8 million[4]. We determine that the prevalence of heart failure is increasing, which is related to the poor quality of patients' life, hospitalization, and consumption of health resources[5]. Heart failure has the following typical clinical symptoms: increased jugular venous pressure, lung crackles, and peripheral edema[6, 7]. Shortness of breath, ankle swelling, and fatigue may also occur[8]. When the heart cannot pump enough blood to the body because of diabetes, hypertension, or other heart diseases, heart failure occurs[9]. Although heart failure can be managed through drug therapy, cardiac devices, and specific heart failure programs, its mortality rate is still high[10]. Therefore, the early detection of heart failure can reduce the enormous burden of this disease

• Dengao Li, Huiting Ma, and Jian Fu are with the College of Data Science, Taiyuan University of Technology, Jinzhong 030600, China. E-mail: lidengao@tyut.edu.cn; 741625633@qq.com; 1046019528@qq.com.
• Wenjing Li is with the Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA. E-mail: wenjingli@ucsb.edu.
• Baofeng Zhao is with the College of Mining Engineering, Taiyuan University of Technology, Taiyuan 030024, China. E-mail: zhaobf517@sohu.com.
• Jumin Zhao and Yi Liu are with the College of Information and Computer, Taiyuan University of Technology, Jinzhong 030600, China. E-mail: zhaojumin@tyut.edu.cn; yi.liuvip@foxmail.com.
∗ To whom correspondence should be addressed.

Manuscript received: 2021-11-08; accepted: 2021-11-22

on individuals and our society[11]. How to identify the disease has gradually become a research hotspot.

With the entity of the disease, the range of diagnostic tests and biomarkers and the treatment modalities have grown exponentially[12]. Medical text is an indispensable information carrier in the medical profession that provides significant data support for clinical diagnosis and pathological study. Electronic health records (EHRs) represent a large repository of electronic data points. These electronic data points represent a variety of clinical information[13]. EHR has been used in hospitals to standardize and integrate medical records written by doctors. Therefore, EHR is an abundant source of health information[14]. Medical texts include various data types, such as electronic medical records and medical and radiological reports. These texts are utilized for medical research and auxiliary diagnosis. Moreover, these texts are diverse, redundant, and heterogeneous. With the development of big data, the widespread application of EHR has led to a rapid increase in the number and diversity of medical data[15]. Because patient electronic medical records are pivotal providers of data-driven health research, understanding the information in electronic medical records is critical[16]. At the same time, with the establishment of clinical data resources, deep learning has become a valuable resource for researchers in healthcare analysis[17]. However, regardless of whether it is a doctor or a deep learning model, we determine that there is still uncertainty in the occurrence of diagnostic errors[18] as the success of the diagnosis is limited by the doctor's knowledge and experience. At present, most researchers only use structured EHR data, such as previous diagnosis and drug treatment processes, but these data may not include the patient's historical and current disease information[19, 20]. However, if combined with the data in the clinical medical record, then the prediction performance may be improved. Moreover, doctors can review the clinical information inputted by other doctors using unstructured data to better understand the patient's condition and treatment effect[21]. Therefore, the use of unstructured data for disease prediction has considerable significance.

The use of deep learning methods to process medical texts is attracting considerable attention from researchers. Nuthakki et al.[22] proposed a model that could make a final diagnosis based on unstructured data, such as the current medical history and symptoms at admission, as clinicians lack a reliable method to determine which patients with congestive heart failure need to undergo cardiac resynchronization therapy (CRT). Leiter et al.[23] trained and tested a deep natural language processing (NLP) model to identify patients with congestive heart failure undergoing CRT. We detect the high sparseness and serious colloquialization of medical texts through research. For the processing of medical texts, the introduction of external corpus texts will only reduce recognition efficiency. Thus, we need to expand the medical texts. The term frequency-inverse word frequency (TF-IWF) model[24] has high efficiency and accuracy. This algorithm can reduce the cost of filtering effectual information and accelerate the spread and circulation of information[25]. The latent dirichlet allocation (LDA) model[26] can solve the problem of sparse semantics. The LDA model does not need to introduce external corpus data and can directly use the abundant information of the original corpus. However, most researchers use a single model for text processing or only consider a single feature vector to represent the text; thus, the evaluation index is low. Therefore, we propose the KTI-RNN model, which extends the medical text based on its content. The proposed model is used to extract keyword and topic word sets. To ensure medical analysis objectivity, the proposed model can identify potential relationships between concepts and rules in medical data. Finally, we use the gated-attention bidirectional recurrent neural network (GA-BiRNN) model for classification to help doctors separate heart failure texts.

Our contributions are summarized as follows:

(1) **Clinical significance**: Heart failure is the terminal stage of cardiovascular disease. Hypertension, myocardial infarction, and other diseases may induce heart failure. The medical text contains a large amount of medically helpful information. The diagnosis of heart failure can be completed through analysing medical records. The use of deep learning methods to process medical text can accelerate the diagnosis of heart failure, thereby assisting doctors in making decisions.

(2) **Model significance**: The LDA model is used to extract topic word sets. The TF-IWF model is used to extract keyword sets. These models help complete and expand the medical record text content. The improved bidirectional recurrent neural network (BiRNN) model, i.e., the GA-BiRNN model, is used to accurately recognize medical text.

(3) **Indicator significance**: The $F1$ score of the proposed KTI-RNN model is 85.57%, and the accuracy

rate of the proposed KTI-RNN model is 85.59%. The results are significantly better than without text content expansion and without classifier improvement. This finding proves that the content expansion and classifier improvement that we proposed are effective.

This paper is organized as follows: In Section 2, we review the related works. In Section 3, we propose the KTI-RNN model to diagnose medical records. In Section 4, we present the results and discussion. In Section 5, we conclude this paper.

2 Related Work

2.1 Diagnosis using unstructured data

In the text classification field, the use of deep learning for NLP has become a research hotspot[27]. The use of unstructured data for disease diagnosis has also attracted considerable attention. Liang et al.[12] used deep learning models to extract pivotal clinical information from electronic medical records. Deep learning models showed high diagnostic accuracy in multiple organ systems, thereby assisting doctors in processing large amounts of data and performing diagnosis and evaluation. Goh et al.[21] developed an artificial intelligence algorithm called SERA, which can predict and diagnose sepsis by combining structured and unstructured data. Le et al.[28] used clinical data and machine learning algorithms to help diagnose children with early heart failure. Nagamine et al.[29] obtained related concepts from clinical notes to describe heart failure, build vector representation, cluster patients with heart failure based on the similarity of the vectors, and determine the salient content of each cluster using statistical tests.

At the same time, research on the topic model, classification[30, 31], and information can be found in many papers[32, 33]. In the following subsections, we will introduce the related works on the LDA, TF-IWF, and BiRNN models in text classification.

2.2 LDA model

The LDA model is used to extract the topic distribution of a document, and each document topic is given in the form of a probability distribution. Topic clustering or text classification is performed according to the topic distribution. The LDA has an important position in topic models. Chen et al.[34] combined the LDA and negative matrix factorization (NMF) on text topic mining. Several researchers made medically assisted diagnoses of electronic medical records based on word vector and LDA models[35]. Selvi et al.[36] used the LDA model to generate a topic model based on the classification of medical datasets. Zhu et al.[37] used the social network analysis (SNA) to describe the main keyword of adverse medical events and the LDA to investigate topics of different hazard levels. They combined SNA and LDA to detect common topic keywords. Then, they analyzed physicians' and nurses' potential behavior in these events of distinct injury levels. Gao et al.[38] aimed to solve the problem of sparse semantics and insufficient co-occurrence information in the special extraction of healthcare reviews. They proposed the CO-LDA model based on word co-occurrence analysis.

2.3 TF-IWF model

The weighting method of the TF-IWF model can reduce the influence of the same type of text in the corpus on the weight of words, thereby more accurately expressing the importance of words in the document. Considerable research on TF-IWF has been conducted. Yan et al.[39] proposed the TF-IWF-IDF model that combines IWF and term frequency-inverse document frequency (TF-IDF). They built a two-step single-channel algorithm based on a multi-topic center. Their algorithm achieved a high accuracy rate in the field of financial hotspot detection and tracking. Zhang et al.[40] proposed a real-time Chinese text keyword extraction method based on on-screen visual hotspots. They utilized the TF-IWF, position statistical distribution, and word distance to construct a Chinese text extraction model, and their model achieved good results. Short texts have less content and loose format. Thus, Zheng et al.[41] proposed an authentic comprehensive method. They combined the biterm topic model (BTM) topic features and improved the weight calculation methods (based on TF-IWF), thereby increasing the final $F1$ score.

2.4 BiRNN model

An advantage of the BiRNN model is that it can combine the information on the left and right sides of the current word to complete the text classification process. The BiRNN model is widely used in text classification work. Leevy et al.[42] used the RNN model and conditional random field (CRF) to explore the related work on automatic free text recognition. Yu et al.[43] proposed the multilayer attention bidirectional recurrent neural network (MA-BiRNN) model to implement disease code assignment. Chen et al.[44] used hierarchical attention BiRNN to encode the grammatical perception representation of medical reports hierarchically. Finally, they classified breast tumors by combining the grammatical and semantic perception representations of medical reports. Ma et al.[45] proposed a new model called Att-BiRNN-Att, which combines the BiRNN with two attention mechanisms to conduct medical answer selection.

The most valuable information about disease and hospitalization can only be found in clinical records or narratives. The study of medical notes can help increase the effectiveness of EHR information. Deep learning for NLP has become a research hotspot in the field of text classification[27]. The latest advances in deep learning and NLP enable machines to learn linguistic expressions in medical texts to make effective and accurate predictions[46]. Therefore, the use of deep learning models for medical text classification can assist doctors in accurately diagnosing and classifying patients' diseases and providing patients with reasonable treatment in time.

3 KTI-RNN Model

The KTI-RNN model proposed in this study expands the content of medical text based on topic words and keywords. We also use an improved classifier to complete the classification of medical texts. The KTI-RNN includes three modules, namely, input module for TF-IWF keyword extraction and LDA topic word extraction, GA-BiRNN module, and classification output module. The framework diagram is shown in Fig. 1. The processing steps are as follows:

(1) Preprocess the original text and split the training and test sets.

(2) Construct the LDA topic word model, train all of the texts in the training set, and extract the topic word set of each sample.

(3) Use the TF-IWF model to extract the category keyword set of each sample.

(4) Construct the GA-BiRNN model, load the Glove[47] pretrained word vector, splice the keywords and topic words, train the neural network model, and perform predictive classification of the text.

3.1 Text preprocessing

For the medical discharge summary processed in this study, the text needs preprocessing before model training.

(1) **Remove the non-text part of the data**

Illegal characters and labels need to be removed. We use Python's regular expressions to filter the non-text part of the data. Some special non-English characters and punctuations are also detected. We establish an illegal character and word table to filter them.

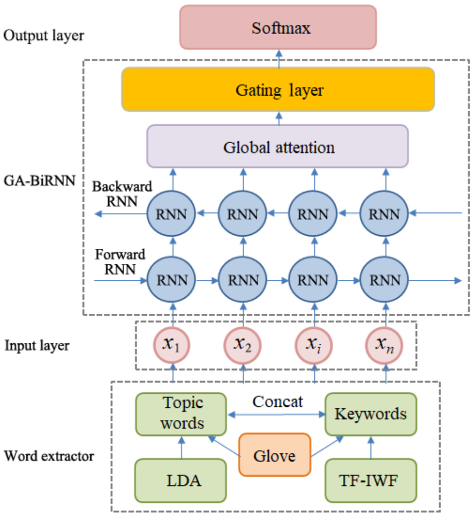


Fig. 1 KTI-RNN: The TF-IWF is used to extract the keyword set; the LDA model is used to extract the topic word set; these models help complete and expand the medical record text content; the GA-BiRNN is used for classification.

(2) **Word segmentation**

Because English words are separated by spaces, the split function is used for word segmentation.

(3) **Remove stop words**

Stop words are insignificant words in a sentence. Removing them will not affect understanding the entire sentence semantics. In this study, we filter stop words, such as function words, pronouns, and verbs and nouns without specific meaning, e.g., "a", "the", "to", and "their", using the stop word list. This step is completed by using the natural language toolkit (NLTK) library.

(4) **Stemming**

Stemming can make the words more unified and can significantly reduce the size of the word space and the difficulty of model learning. Our experiment uses NLTK's WordNet to implement the stemming process.

Finally, the text is case-converted, all of the texts are converted uniformly to lowercase characters, excessively long texts are truncated, and the maximum processing length is set to 512.

3.2 Topic word extraction

This study uses LDA to extract the topic word set from the medical text. The LDA[26] is an unsupervised machine learning technology, which is usually used by researchers to identify the hidden topic information in a large-scale document set. The LDA model adopts the bag-of-words method. Each document is regarded as a word frequency vector. Finally, textual information is converted into digital information. Notably, digital information is easier to model than textual information. Each document is regarded as a different topic collection, and each topic is considered as a distinct vocabulary collection. According to document-topic-words, LDA provides the extraction result of the document topic distribution based on the probability distribution. Each document in the corpus corresponds to a polynomial distribution of n topics. Each topic corresponds to a polynomial distribution of V words in the vocabulary. The vocabulary is composed of all mutually different words in all documents of the corpus. For each word in document d, we randomly select a topic from the multinomial distribution corresponding to the document. A word from the multinomial distribution corresponding to this topic is extracted; this process is repeated N times. Then, document d is produced. In this study, the Gensim package of the Python environment is used. The training process of the LDA model is as follows:

(1) Document d_i is selected according to the prior probability $p(d_i)$.

(2) The topic distribution σ_i of d_i is obtained by sampling from the Dirichlet distribution a.

(3) A sample from the topic polynomial distribution σ_i is taken to generate the topic $E_{i,j}$ of the j-th word in the document d_i.

(4) A sample from the Dirichlet distribution b is taken to determine the word distribution $\phi_{E_{i,j}}$ corresponding to topic $E_{i,j}$.

(5) w_{ij} is sampled from the multinomial distribution $\phi_{E_{i,j}}$.

After model training is completed, the topic words generated by each document are obtained, as expressed in Eq. (1):

$$Topic\text{-}i = \{w'_1 : t_1, w'_2 : t_2, w'_3 : t_3, \ldots, w'_n : t_n\} \quad (1)$$

where $Topic\text{-}i$ is the topic word set of document i, w'_i is the topic word, and t_i is the probability of the topic word.

3.3 Keyword extraction

The TF-IWF model is used in this study to extract the keyword set from the medical text. Keyword set construction involves screening out recognizable words and completing the effective expansion of text content. The most common model used to construct the keyword set is the TF-IDF. However, the TF-IDF model has distinct shortcomings. The structure of the TF-IDF model is simple. The keywords extracted cannot effectively reflect the importance of words and the distribution of feature words. The adjustment of weights cannot be well-completed. Particularly in congener corpora, the keywords of some congener texts can be easily masked. For example, if there are many articles of category C in corpus D (where text j is an article belonging to category C), then the IDF value of the words related to category C will be small and the recall rate of the extracted text keywords will be low. For this shortcoming, this study uses the improved weighting algorithm TF-IWF[24] to extract keywords.

The TF-IWF is expressed in Eq. (2):

$$TF\text{-}IWF(w_{ij}) = TF(w_{ij}) \times IWF(w_{ij}) \quad (2)$$

where $TF(w_{ij}) = \frac{n_{ij}}{\sum_h n_{hj}}$ (where n_{ij} is the frequency of word t_i in text j and $\sum_h n_{hj}$ is the sum of all vocabularies in text j) and $IWF(w_{ij}) = \log\frac{\sum_{i=1}^m nt_i}{nt_i}$ (where $\sum_{i=1}^m nt_i$ is the sum of the frequency of all words in the corpus and the denominator nt_i is the total frequency of the word t_i in the corpus).

The keywords generated by the aforementioned method for each document is expressed in Eq. (3):

$$Keywords\text{-}i = \{w''_1 : q_1, w''_2 : q_2, w''_3 : q_3, \ldots, w''_n : q_n\} \quad (3)$$

where $Keywords\text{-}i$ is the keyword set of document i, w''_i is the extracted keyword, and q_i is the score of the keyword. Then, the keywords are sorted according to the keyword scores of the documents. Finally, the top k keyword set is considered the extended word set. All experiments in this study take the top 5% keyword set.

3.4 Text classification

The GA-BiRNN model proposed in this study is improved based on the BiRNN model. The global attention mechanism and gating mechanism are embedded between the BiRNN model and the output layer. The GA-BiRNN model is divided into four parts, namely, vectorized input layer, coding layer, gating layer, and output layer. The coding layer is composed of the BiRNN layer[48], attention layer, and fully connected layer. The following subsections will specifically introduce the four parts of the GA-BiRNN model.

3.4.1 Vectorized input layer

The input layer is the vectorized representation of the text. After expansion, each sample can be expressed as $x = \{w_1, w_2, w_3, \ldots, w_n\}$, which represents the n words after expansion. Then, the pretrained word vector of glove.6b.300d is downloaded. For the word in x, if it can be found in Glove, then the pretrained word vector is used for the initialization process. The word vector that cannot be found in Glove is initialized randomly.

3.4.2 Coding layer

The coding layer is generated in two steps.

First, the output word vector of the BiRNN layer is calculated. The input vector of BiRNN is used as the word vector in this study. The main purpose of the BiRNN layer is to extract deep-level text features from the input text vector. As shown in Fig. 1, the BiRNN model is composed of two parts, namely, forward RNN and backward RNN, as expressed in Eq. (4):

$$h_{ijt} = BiRNN(c_{ijt}), t \in [1, m] \tag{4}$$

where the word vector c_{ijt} is the t-th word of the j-th sentence input at the i-th moment and h_{ijt} is the word vector after BiRNN encoding.

Then, the probability weight that should be assigned to each word vector is calculated. The global attention mechanism is mainly added here. The input of the attention layer is the output vector h_{ijt} processed by the activation of the BiRNN layer in the previous layer. The weight coefficient of the attention layer is calculated as follows:

$$u_{ijt} = \tanh(w_w h_{ijt} + b_w) \tag{5}$$

$$a_{ijt} = \frac{\exp(u_{ijt}^{\mathrm{T}} u_w)}{\sum_t \exp(u_{ijt}^{\mathrm{T}} u_w)} \tag{6}$$

$$s_{ijt} = \sum_{i=1}^{n} a_{ijt} h_{ijt} \tag{7}$$

where w_w and b_w are the weights and biases of feature conversion, respectively; a_{ijt} is the attention weight matrix, which has different probability weights assigned by the attention mechanism and multiplication accumulation summation of each hidden state and uses the softmax function to normalize; u_w is the parameter to be learned by the attention mechanism layer; and s_{ijt} is the feature representation after adding the self-attention mechanism.

3.4.3 Gating layer

The gating mechanism transforms the input feature representation X into a new feature space Y, as expressed in Eq. (8):

$$Y = \sigma(W \times X + b) \circ X \tag{8}$$

where X is the output of the BiRNN, σ is the sigmoid activation function, $\circ$ is the dot product of the elements, $W \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^{n}$ are learnable parameters, and $\sigma(W \times X + b)$ is a gated neuron belonging to [0, 1]. When the output value is close to 1, the door is open. When the output value is close to 0, the door is closed. The shape of $\sigma(W \times X + b)$ is the same as X, and gated selection can be made for each dimension feature in X, i.e., it will act on each dimension of the input X. The specific process is shown in Fig. 2.

The main reason for adding a gating mechanism after BiRNN comes from two aspects. First, we hope to introduce a nonlinear feature crossover between the input features. Second, we also hope to control the activation degree of each neuron of the input feature through this self-attention gating unit.

The gating mechanism can build the dependency relationship between neurons. For example, for a discharge record of a heart failure category, the document contains the words related to this category, e.g., “heart”, “failure”, “chronic”, and “congestive”. However, the method used to extract keywords and topic words is an unsupervised algorithm. The expanded vocabulary will also contain some words that do not refer to the category, such as “put”, “occlusion”, and “follow up”. The weights of these irrelevant features should be reduced, and the features related to the category should be weighted. The gating mechanism proposed in this study can be automatically learned to obtain the weights of ordinary important features.

3.4.4 Output layer

The softmax function is used to calculate the output category probability for the input of the output layer. Thus, text classification can be performed, as expressed in Eq. (9):

$$y_j = \mathrm{softmax}(w s_j + b) \tag{9}$$

where w and b are the weight and bias to be trained, respectively, s_j is the feature representation of the output layer, and y_j is the predicted label of the document j output.

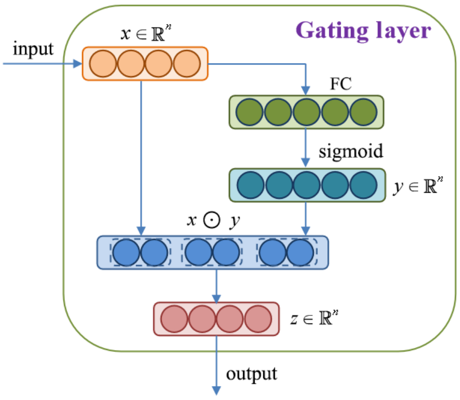


Fig. 2 Gating mechanism.

3.5 Loss function

According to the predicted and actual labels of the output layer, the loss function training network is constructed. The model gradually approaches the real distribution with the training prediction distribution. The loss function used in this study is a cross-entropy loss because it can measure the difference between the true distribution and the predicted distribution. The smaller the value is, the better the predicted result. Thus, the purpose is to minimize their distance, which can be expressed mathematically in Eq. (10).

$$loss = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} y(x_{ij})\log(q(x_{ij})) \qquad (10)$$

where $y(x_{ij})$ is the predicted value of the j-th category of the i-th sample, $q(x_{ij})$ is the actual value of the j-th category of the i-th sample, m is the number of samples, and n is the number of categories.

3.6 Pseudocode of KTI-RNN

To better illustrate the KTI-RNN model, we will use pseudocode. The specific algorithm process is shown in Algorithm 1.

(1) The training dataset D is constructed, and the Glove embedding of the pretraining number is downloaded. The version downloaded in this experiment is glove.6b.300d. The dimension of the embedding is 300, and the corpus used for training is 6 billion.

Algorithm 1 KTI-RNN training procedure

```
Input: Training data $D$
Load pretrained word embeddings glove.6b.300d;
Initialize GA-BiRNN weights $W$ using Xavier method;
Cut training sentences to tokens as $M$ for input $D$;
LDA:
Put input segmented corpus $M$ to gensim package;
Load stopwords to remove nonsensial words;
Set parameters: num-topics = 70, $\alpha = 0.15$;
Training model saved as LDA.mm;
Output topic words list $LW_i$;
TF-IWF:
For each sentence $D_i$, calculate term frequency $TF(w_{ij})$;
Calculate inverse word frequency $IWF(w_{ij})$;
Then $TF\text{-}IWF(w_{ij}) = TF(w_{ij}) \times IWF(w_{ij})$; see (Eq. (2));
GA-BiRNN:
while not converge do
    Sample $(sequence, y)$ from data $D$;
    Concat tokens, extended keywords, and extended topic words into $x$;
    Pass $x$ through Glove embeddings to produce $s$;
    ▷ Standard forward-propagation;
    $\widehat{y} \leftarrow F(s; \omega)$;
    Compute $L(s, y; \theta)$ using Eq. (10);
    Update parameters $W = W + \eta \times \partial L(s', y; \theta)/\partial W$;
end while
```

(2) The training corpus is segmented. The LDA topic word expansion and TF-IWF keyword expansion are performed.

(3) The expanded words and original texts are spliced together, inputted into the GA-BiRNN, and subjected to BiRNN semantic coding, global attention mechanism, gating mechanism, and softmax classification output layer. The cross-entropy loss function is calculated, and the network parameters are updated through backpropagation and iteratively trained until the model converges.

4 Results and Discussion

4.1 Data description and preprocessing

The data used in this study are obtained from the Medical Information Mart for Intensive Care (MIMIC-III v1.4)[49], which is a large, single-center, and freely available database. From the MIMIC-III database, 10 436 patients are extracted according to 25 types of ICD-9 codes corresponding to heart failure. Similarly, 4557 patients are extracted according to 23 types of ICD-9 codes corresponding to myocardial infarction, and 18 066 patients are extracted according to 19 types of ICD-9 codes corresponding to hypertension. On this basis, we screened whether patients had discharge summary records. Finally, we obtained 10 270 patients with discharge summary records for heart failure. The number of patients with myocardial infarction was 4464, and the number of patients with hypertension was 17 644. The specific data volume is shown in Table 1. The ratio of the training set to the test set is 7:3. The training set uses the first 70% of the data, and the test set uses the remaining 30% of the data.

4.2 Experimental environment and parameter settings

Hardware environment: Intel® Xeon® CPU 3.66 GHz, memory 16 GB, GPU NVIDIA Tesla V100, 32 G.

Software environment: CentOS 7.0, TensorFlow

Table 1 Distribution of experimental datasets.

Category of text	Total number	Number of patients with discharge records
Heart failure	10 436	10 270
Myocardial infarction	4557	4464
Hypertension	18 066	17 644

1.15.0, Keras 2.3.1.

Parameter settings:

TextCNN/BiRNN/TextRNN/Attention-BiRNN: The maximum length of the input sequence is 100, the training batch size is 64, the number of training epochs is 20, and the embedding dimension is 300.

HAN: The maximum feature is 5000, the training batch size is 32, the number of training epochs is 10, and the embedding dimension is 50.

RCNN: The maximum length of the input sequence is 100, the training batch size is 64, the number of training epochs is 10, and the embedding dimension is 20.

FastText: The n-gram range is equal to 1, the maximum length of the input sequence is 100, the training batch size is 64, the number of training epochs is 20, and the embedding dimension is 300.

4.3 Model evaluation criteria

For model evaluation, this study adopts four criteria, namely, precision, recall, $F1$ score, and accuracy (ACC). The calculation formulas of the four model evaluation criteria are expressed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = \frac{2\times TP}{FP + FN + 2\times TP} \tag{13}$$

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \tag{14}$$

where *TP* is the number of true-positive samples, *TN* is the number of true-negative samples, *FP* is the number of false-positive samples, and *FN* is the number of false-negative samples.

4.4 Experimental results and analysis

Table 2 shows the effect comparison between the KTI-RNN model proposed in this study and other baseline models. Because of the addition of the global attention and gating mechanisms, the KTI-RNN exhibits a significant improvement compared with the seven other groups of models. The $F1$ score of the KTI-RNN model is 85.57%, and the accuracy rate is 85.59%; this result is 12.26% and 12.47% higher than that of FastText, respectively. Compared with the best model, i.e., Attention-BiRNN, the $F1$ score and accuracy rate of KTI-RNN have increased by 2.13% and 2.17%, respectively.

Table 2 KTI-RNN and baseline model comparison results.

(%)

Model	Precision	Recall	$F1$ score	Accuracy
FastText	72.52	74.11	73.31	73.12
TextCNN	81.46	79.91	80.67	80.93
BiRNN	87.65	73.68	80.02	81.68
TextRNN	86.95	67.48	75.99	78.76
RCNN	79.79	82.05	80.90	78.41
HAN	83.15	81.59	82.63	82.62
Attention-BiRNN	83.04	83.84	83.44	83.42
KTI-RNN	85.37	**85.77**	**85.57**	**85.59**

Table 3 Analysis of the effect of the submodules of the KTI-RNN model.

(%)

Model	Precision	Recall	$F1$ score	Accuracy
KTI-RNN	85.37	**85.77**	**85.57**	**85.59**
Without LDA	79.11	79.24	79.17	79.50
Without TF-IWF	80.45	82.49	81.46	81.49
Without LDA + TF-IWF	76.71	72.31	74.46	75.87
Without LDA + TF-IWF + Glove	73.00	70.81	71.88	73.08
Without Glove	82.45	78.53	80.44	80.93

Table 3 shows the contribution of each submodule of the KTI-RNN model, which includes the LDA topic word model, TF-IWF keyword model, and Glove pretrained word vector. We determine that the LDA topic word model has the most significant impact on the results. Without the LDA model, the $F1$ score and accuracy rate have decreased by 6.4% and 6.09%, respectively. The impact of the Glove pretrained word vector is relatively small. The $F1$ score and accuracy have decreased by 5.13% and 4.66%, respectively, without the initialization of the Glove pretrained word vector initialization. The removal of the LDA word topic model, TF-IWF keyword model, and Glove pretrained word vector at the same time has the most significant impact on the results, and the effect is reduced by approximately 13%.

Figure 3 shows the accuracy curve of the deep learning model subjected to the training process. Figure 3 shows the comparison of the accuracy and convergence rates of the KTI-RNN model and the five other models. Both the accuracy and convergence rates of the KTI-RNN model are the highest. Both the convergence and accuracy rates of the models without LDA + TF-IWF and without LDA + TF-IWF + Glove are the lowest. Both the convergence and accuracy rates of the three other groups, namely, without LDA, without TF-IWF, and without Glove, are not much different.

Figure 4 shows the comparison and analysis of the change process of the perplexity of the probability

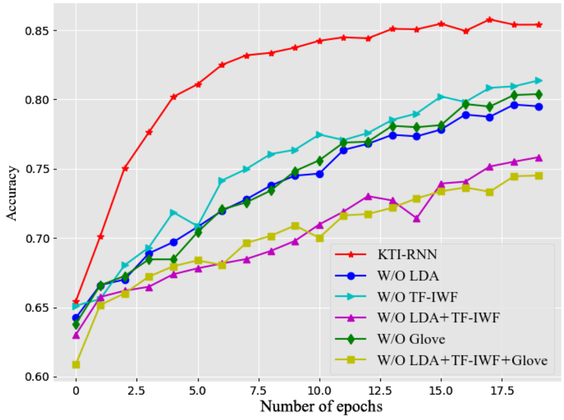


Fig. 3 Curves of epoch and accuracy.

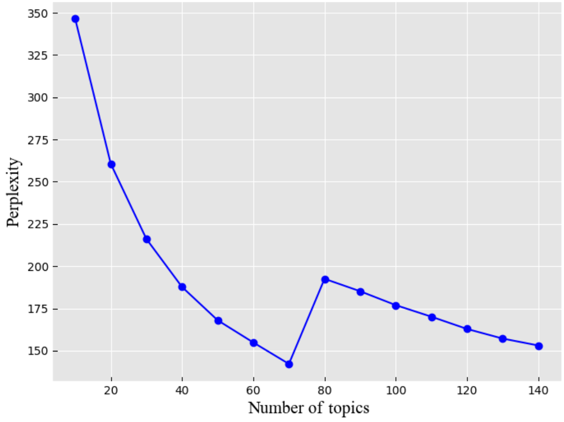


Fig. 4 Relationship between the number of topics and the value of perplexity.

distribution matrix returned by the LDA model when different topics are selected. The experiment compared the value of perplexity between 10 and 140 topics with an interval of 10. When the number of topics is 70, the value of perplexity is the smallest. The value of perplexity first increases and then decreases when the number of topics is over 70. Therefore, this study selected 70 as the number of topics for all experiments.

Tables 4 and 5 show the topic word distribution of the trained LDA model and the TF-IWF keyword distribution, respectively. Table 4 shows that the words under the same topic are relatively similar and belong to the same field, e.g., “anemia” and “hematocrit” under Topic 1, indicating that the LDA model converges normally. The effect of the sampled topic word distribution has reached expectations. Table 5 shows that most of the extracted keywords under the same document are related to the classified documents. The TF-IWF algorithm suppresses the scores of vocabularies that are unrelated to the category, and the keywords with high scores have high quality.

Table 4 Topic word distribution of the LDA model.

Topic	Topic word and probability value
Topic 1	anemia : 0.025 + hct : 0.021 + hematocrit : 0.017 + transfusion : 0.015
Topic 2	pharmacy : 0.084 + chest : 0.049 + anterior : 0.040 + apical : 0.014
Topic 3	chloride : 0.074 + glucoseneg : 0.039 + chamber : 0.024 + rales : 0.013
Topic 4	angioplasty : 0.059 + intraaortic : 0.036 + pump : 0.031 + failure : 0.019
…	…
Topic 70	bilaterally : 0.054 + mildly : 0.027 + saturating : 0.018 + nad : 0.016

Table 5 TF-IWF keywords.

Doc	Keyword and score
Doc 1	rocaltrol : 0.015 + provera : 0.015 + premarin : 0.011 + nephrocaps : 0.009
Doc 2	saphenectomy : 0.017 + piperacillin : 0.013 + hemodialysis : 0.011 + affected : 0.008
Doc 3	femoroperoneal : 0.034 + chronit : 0.027 + hiphx : 0.018 + procede : 0.014
Doc 4	segment : 0.007 + inferior : 0.004 + stents : 0.003 + atrial : 0.003
…	…
Doc *n*	methadone : 0.026 + opioid : 0.014 + headache : 0.011 + diffusion : 0.006

The setting of the embedding dimension of the embedding layer is important for the RNN. The larger the dimension is, the stronger the capability of the model to express the importance of words. However, large setting dimensions will lead to the risk of model overfitting. Figures 5 and 6 show the accuracy rates and $F1$ scores of the comparison of the embedding size of the proposed KTI-RNN model and without LDA, without IWF, and without Glove. Figures 5 and 6 show that when the dimension is less than 300 and gradually increases, the accuracy rate and $F1$ score of the model also gradually increase. When the dimension is greater than 300, the models exhibit different degrees of overfitting. Therefore, in this study, the embedding dimension is set to 300 in all experiments.

Notably, the RNN can handle a long sequence input. When the sequence is short, the model will have a high risk of underfitting because of insufficient information. If the sequence is long, then the problem of “catastrophic forgetting” will occur. In other words, after learning the latest sequence information, the model will forget the historical information that is far from the current moment. Therefore, the length of the input sequence

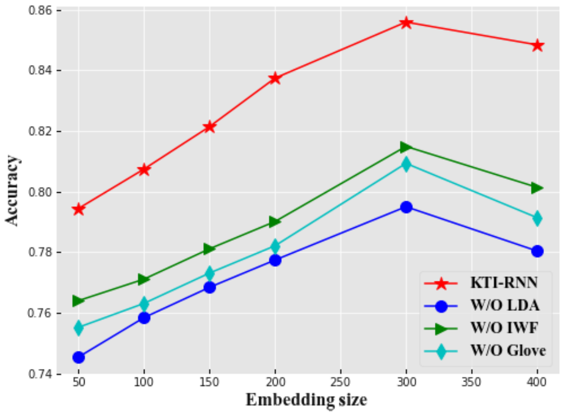


Fig. 5 Impact of different embedding dimension settings on the accuracy rate.

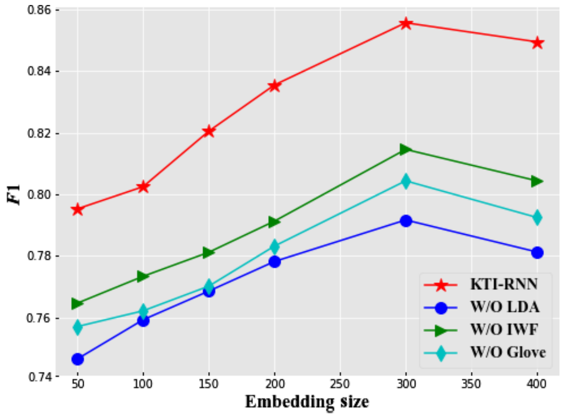


Fig. 6 Influence of different embedding dimension settings on the $F1$ score.

has a direct impact on the effect of model learning. The optimal sequence length setting should match the memory capability of the designed model. The following experiments are designed to determine the effect of different input sequence lengths on the model indicators. The length of the input sequence varies from [50, 100, 150, 200, 300, 400]. Figures 7 and 8 show that, as the sequence length increases, the accuracy rate and $F1$ score first increase and then decrease. In particular, when the length is 100, the model effect is optimal. The effect of the model gradually decreases when the length is greater than 100. However, when the length is between 300 and 400, the indicators of the model change only slightly. For the KTI-RNN model, when the input sequence reaches a certain length, the model will appear "saturated". The sequence length continues to increase such that the index values will not decrease. This effect gradually stabilizes to a certain level.

In the KTI-RNN model, two modules called the global attention and gating mechanisms are embedded

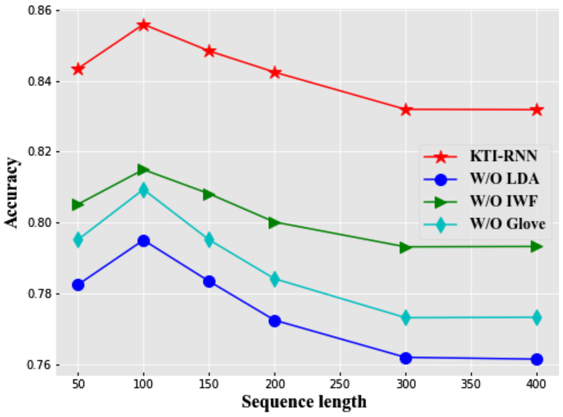


Fig. 7 Relationship between input token length and accuracy.

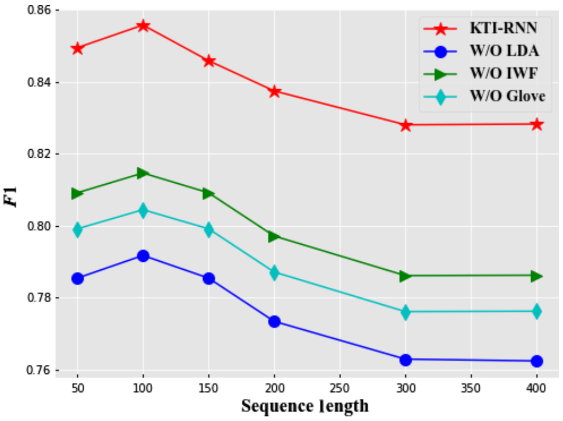


Fig. 8 Relationship between input token length and $F1$.

between the BiRNN model and the output layer. The global attention mechanism can help capture the interdependence between features. For example, in the input token sequence, the subsequent tokens can focus on the previous part of the information related to it and help infer and predict the label of the document. Because of the long sequence length, the global attention mechanism should be used. In this scenario, the characteristics of the global attention mechanism are better than those of the local attention mechanism. The gating mechanism is added to help the model filter redundant features and weigh important features that are helpful to the tags. The subsequent experiment analyzes the importance of these two components through an ablation study. Figures 9 and 10 show that the role of the gating mechanism is more significant than that of the global attention mechanism among the four groups of analysis models, i.e., KTI-RNN, KTI-RNN without LDA, KTI-RNN without TF-IWF, and KTI-RNN without Glove. With the removal of the gating mechanism, the accuracy rate and $F1$ score of the model decrease by approximately 4%. The attention

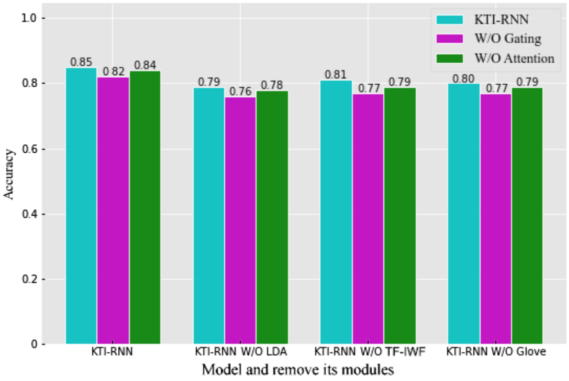


Fig. 9 Influence of the gating mechanism and attention mechanism on the accuracy rate of the model.

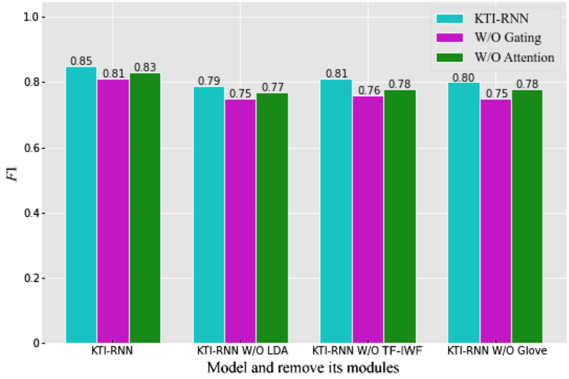


Fig. 10 Influence of the gating mechanism and attention mechanism on the $F1$ score of the model.

mechanism has a relatively small effect. Without the attention mechanism, the accuracy rate of the model decreases by approximately 2%. This finding indicates that, for long text classification, a large amount of redundant information is detected. The proportion of information helpful to label classification is relatively small. A reasonable feature filtering technology will help improve the classification effect of the model.

5 Conclusion

Unstructured data contain a large amount of relevant patient information. To assist doctors in making more reasonable decisions, we use this type of data to identify heart failure. In contrast to traditional methods that use structured data and feature engineering, the novel method proposed in this study utilizes unstructured data to process medical texts. We propose the KTI-RNN model, which uses the TF-IWF and LDA to extract keyword and topic word sets from clinical notes, respectively, and complete the expansion of medical text content. Based on the improved BiRNN model called the GA-BiRNN, we embed the global attention mechanism and gating mechanism between the BiRNN model and the output layer and use it to train and classify medical texts. The final accuracy rate is 85.59%, and the $F1$ score is 85.57%. Through ablation experiments, we determine that when the two extended word sets are not used, the final accuracy rate is 75.87%, and the $F1$ score is 74.46%. When only the BiRNN model is used to classify medical texts, the final accuracy rate is 81.68%, and the $F1$ score is 80.02%. The KTI-RNN model that we proposed is relatively accurate in medical text diagnosis and can be easily applied to other fields with appropriate datasets. Future work may include combining structured and unstructured data to diagnose heart failure, testing different word embedding methods for clinical notes, and other NLP applications using KTI-RNN.

Acknowledgment

This work was supported by the National Major Scientific Research Instrument Development Project (No. 62027819): High-Speed Real-Time Analyzer for Laser Chip's Optical Catastrophic Damage Process; the General Object of the National Natural Science Foundation (No. 62076177): Study on the Risk Assessment Model of Heart Failure by Integrating Multi-Modal Big Data; and Shanxi Province Key Technology and Generic Technology R&D Project (No. 2020XXX007): Energy Internet Integrated Intelligent Data Management and Decision Support Platform.

References

[1] G. A. Roth, C. Johnson, A. Abajobir, F. Abd-Allah, S. F. Abera, G. Abyu, M. Ahmed, B. Aksut, T. Alam, K. Alam, et al., Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015, *Journal of the American College of Cardiology*, vol. 70, no. 1, pp. 1–25, 2017.

[2] C. W. Yancy, M. Jessup, B. Bozkurt, J. Butler, D. E. Casey Jr, M. M. Colvin, M. H. Drazner, G. S. Filippatos, G. C. Fonarow, M. M. Givertz, et al., 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: A report of the American College of Cardiology/American heart association task force on clinical practice guidelines and the heart failure society of America, *Journal of the American College of Cardiology*, vol. 70, no. 6, pp. 776–803, 2017.

[3] R. A. Nishimura, C. M. Otto, R. O. Bonow, B. A. Carabello, J. P. Erwin, L. A. Fleisher, H. Jneid, M. J. Mack, C. J. McLeod, P. T. O'Gara, et al., 2017 AHA/ACC focused update of the 2014 AHA/ACC guideline for the management of patients with valvular heart disease: A report of the American College of Cardiology/American heart association task force on clinical practice guidelines,

Journal of the American College of Cardiology, vol. 70, no. 2, pp. 252–289, 2017.

[4] T. Lagu, P. S. Pekow, M. -S. Shieh, M. Stefan, Q. R. Pack, M. A. Kashef, A. R. Atreya, G. Valania, M. T. Slawsky, and P. K. Lindenauer, Validation and comparison of seven mortality prediction models for hospitalized patients with acute decompensated heart failure, *Circulation: Heart Failure*, vol. 9, no. 8, p. e002912, 2016.

[5] N. Farré, E. Vela, M. Clèries, M. Bustins, M. Cainzos-Achirica, C. Enjuanes, P. Moliner, S. Ruiz, J. M. Verdú-Rotellar, and J. Comín-Colet, Medical resource use and expenditure in patients with chronic heart failure: A population-based analysis of 88 195 patients, *European Journal of Heart Failure*, vol. 18, no. 9, pp. 1132–1140, 2016.

[6] V. Carubelli, G. Cotter, B. Davison, J. Gishe, S. Senger, I. Bonadei, E. Gorga, V. Lazzarini, C. Lombardi, and M. Metra, In-hospital worsening heart failure in patients admitted for acute heart failure, *International Journal of Cardiology*, vol. 225, pp. 353–361, 2016.

[7] J. S. Gordin and G. C. Fonarow, New medications for heart failure, *Trends in Cardiovascular Medicine*, vol. 26, no. 6, pp. 485–492, 2016.

[8] A. Triantafyllidis, C. Velardo, T. Chantler, S. A. Shah, C. Paton, R. Khorshidi, L. Tarassenko, K. Rahimi, and on behalf of the SUPPORT-HF Investigators, A personalized mobile-based home monitoring system for heart failure: The support-HF study, *International Journal of Medical Informatics*, vol. 84, no. 10, pp. 743–753, 2015.

[9] National Heart, Lung, and Blood Institute (NHLBI), Heart failure, https://www.nhlbi.nih.gov/health-topics/heart-failure, 2019.

[10] I. Sayago-Silva, F. García-López, and J. Segovia-Cubero, Epidemiology of heart failure in Spain over the last 20 years, *Revista Española de Cardiología (English Edition)*, vol. 66, no. 8, pp. 649–656, 2013.

[11] R. J. Byrd, S. R. Steinhubl, J. Sun, S. Ebadollahi, and W. F. Stewart, Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records, *International Journal of Medical Informatics*, vol. 83, no. 12, pp. 983–992, 2014.

[12] H. Liang, B. Y. Tsui, H. Ni, C. C. S. Valentim, S. L. Baxter, G. Liu, W. Cai, D. S. Kermany, X. Sun, J. Chen, et al., Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence, *Nature Medicine*, vol. 25, no. 3, pp. 433–438, 2019.

[13] M. Z. Nezhad, D. Zhu, N. Sadati, K. Yang, and P. Levi, SUBIC: A supervised bi-clustering approach for precision medicine, arXiv preprint arXiv: 1709.09929, 2017.

[14] E. A. Wang, J. B. Long, K. A. McGinnis, K. H. Wang, C. J. Wildeman, C. Kim, K. B. Bucklen, D. A. Fiellin, J. Bates, C. Brandt, et al., Measuring exposure to incarceration using the electronic health record, *Medical Care*, vol. 57, pp. S157–S163, 2019.

[15] M. Jamei, A. Nisnevich, E. Wetchler, S. Sudat, and E. Liu, Predicting all-cause risk of 30-day hospital readmission using artificial neural networks, *PloS ONE*, vol. 12, no. 7, p. e0181173, 2017.

[16] C. Xiao, E. Choi, and J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review, *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 2018.

[17] F. Li, W. Liu, and H. Yu, Extraction of information related to adverse drug events from electronic health record notes: Design of an end-to-end model based on deep learning, *JMIR Medical Informatics*, vol. 6, no. 4, p. e12159, 2018.

[18] M. S. Sajid, T. Hollingsworth, M. McGlue, and W. F. Miles, Factors influencing the diagnostic accuracy and management in acute surgical patients, *World Journal of Gastrointestinal Surgery*, vol. 6, no. 11, pp. 229–234, 2014.

[19] G. E. Simon, E. Johnson, J. M. Lawrence, R. C. Rossom, B. Ahmedani, F. L. Lynch, A. Beck, B. Waitzfelder, R. Ziebell, R. B. Penfold, et al., Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records, *American Journal of Psychiatry*, vol. 175, no. 10, pp. 951–960, 2018.

[20] G. E. Simon, S. M. Shortreed, E. Johnson, R. C. Rossom, F. L. Lynch, R. Ziebell, and R. B. Penfold, What health records data are required for accurate prediction of suicidal behavior? *Journal of the American Medical Informatics Association*, vol. 26, no. 12, pp. 1458–1465, 2019.

[21] K. H. Goh, L. Wang, A. Y. K. Yeow, H. Poh, K. Li, J. J. L. Yeow, and G. Y. H. Tan, Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare, *Nature Communications*, vol. 12, no. 1, pp. 1–10, 2021.

[22] S. Nuthakki, S. Neela, J. W. Gichoya, and S. Purkayastha, Natural language processing of MIMIC-Ⅲclinical notes for identifying diagnosis and procedures with neural networks, arXiv preprint arXiv: 1912.12397, 2019.

[23] R. E. Leiter, E. Santus, Z. Jin, K. C. Lee, M. Yusufov, I. Chien, A. Ramaswamy, E. T. Moseley, Y. Qian, D. Schrag, et al., Deep natural language processing to identify symptom documentation in clinical notes for patients with heart failure undergoing cardiac resynchronization therapy, *Journal of Pain and Symptom Management*, vol. 60, no. 5, pp. 948–958, 2020.

[24] C. -H. Huang, J. Yin, and F. Hou, A text similarity measurement combining word semantic information with TF-IDF method, *Chinese Journal of Computers*, vol. 34, no. 5, pp. 856–864, 2011.

[25] A. Xiong, D. Liu, H. Tian, Z. Liu, P. Yu, and M. Kadoch, News keyword extraction algorithm based on semantic clustering and word graph model, *Tsinghua Science and Technology*, vol. 26, no. 6, pp. 886–893, 2021.

[26] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[27] X. Han, B. Li, and Z. Wang, An attention-based neural framework for uncertainty identification on social media texts, *Tsinghua Science and Technology*, vol. 25, no. 1, pp. 117–126, 2019.

[28] T. -D. Le, R. Noumeir, J. Rambaud, G. Sans, and P. Jouvet, Detecting of a patient's condition from clinical

narratives using natural language representation, arXiv preprint arXiv: 2104.03969, 2021.

[29] T. Nagamine, B. Gillette, A. Pakhomov, J. Kahoun, H. Mayer, R. Burghaus, J. Lippert, and M. Saxena, Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data, *Scientific Reports*, vol. 10, no. 1, pp. 1–13, 2020.

[30] N. Leema, H. K. Nehemiah, and A. Kannan, Neural network classifier optimization using differential evolution with global information and back propagation algorithm for clinical datasets, *Applied Soft Computing*, vol. 49, pp. 834–844, 2016.

[31] Y. N. Jane, H. K. Nehemiah, and K. Arputharaj, A Q-backpropagated time delay neural network for diagnosing severity of gait disturbances in Parkinson's disease, *Journal of Biomedical Informatics*, vol. 60, pp. 169–176, 2016.

[32] M. S. Saranya, M. Selvi, S. Ganapathy, S. Muthurajkumar, L. S. Ramesh, and A. Kannan, Intelligent medical data storage system using machine learning approach, in *Proc. 2016 Eighth International Conference on Advanced Computing (ICoAC)*, Chennai, India, 2017, pp. 191–195.

[33] D. M. Blei, Probabilistic topic models, *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[34] Y. Chen, H. Zhang, R. Liu, Z. Ye, and J. Lin, Experimental explorations on short text topic mining between LDA and NMF based schemes, *Knowledge-Based Systems*, vol. 163, pp. 1–13, 2019.

[35] X. -Y. Jin, D. -X. Pu, Y. -Z. Lan, and L. -J. Li, Medical aided diagnosis using electronic medical records based on LDA and word vector model, in *Proc. 2017 4^{th} International Conference on Information Science and Control Engineering (ICISCE)*, Changsha, China, 2017, pp. 443–445.

[36] M. Selvi, K. Thangaramya, M. S. Saranya, K. Kulothungan, S. Ganapathy, and A. Kannan, Classification of medical dataset along with topic modeling using LDA, in *Nanoelectronics, Circuits and Communication Systems*, V. Nath and J. K. Mandal, eds. Singapore: Springer, 2019, pp. 1–11.

[37] L. Zhu, I. Reychav, R. McHaney, A. Broda, Y. Tal, and O. Manor, Combined SNA and LDA methods to understand adverse medical events, *International Journal of Risk & Safety in Medicine*, vol. 30, no. 3, pp. 129–153, 2019.

[38] H. Y. Gao, J. W. Liu, and S. X. Yang, Identifying topics of online healthcare reviews based on improved LDA, (in Chinese), *Transactions of Beijing Institute of Technology*, vol. 39, no. 4, pp. 427–434, 2019.

[39] D. Yan, E. Hua, and B. Hu, An improved single-pass algorithm for Chinese microblog topic detection and tracking, in *Proc. 2016 IEEE International Congress on Big Data (BigData Congress)*, San Francisco, CA, USA, 2016, pp. 251–258.

[40] T. Zhang, W. Wang, Y. Huang, K. Liu, and X. Hu, Method of real-time keyword extraction from Chinese short-text based on visual hotspot on screen, (in Chinese), *Journal of the China Society for Scientific and Technical Information*, vol. 35, no. 12, pp. 1313–1322, 2016.

[41] C. Zheng, W. Wu, and N. Dai, Improved short text classification method based on BTM topic features, *Computer Engineering and Applications*, vol. 52, no. 13, pp. 95–100, 2016.

[42] J. L. Leevy, T. M. Khoshgoftaar, and F. Villanustre, Survey on RNN and CRF models for de-identification of medical free text, *Journal of Big Data*, vol. 7, no. 1, pp. 1–22, 2020.

[43] Y. Yu, M. Li, L. Liu, Z. Fei, F. Wu, and J. Wang, Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN, *Journal of Biomedical Informatics*, vol. 91, p. 103114, 2019.

[44] D. Chen, M. Huang, and W. Li, Knowledge-powered deep breast tumor classification with multiple medical reports, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 891–901, 2019.

[45] J. Ma, C. Che, and Q. Zhang, Medical answer selection based on two attention mechanisms with BiRNN, *MATEC Web of Conferences*, vol. 176, no. 8, p. 01024, 2018.

[46] M. Jiang, T. Sanger, and X. Liu, Combining contextualized embeddings and prior knowledge for clinical named entity recognition: Evaluation study, *JMIR Medical Informatics*, vol. 7, no. 4, p. e14850, 2019.

[47] J. Pennington, R. Socher, and C. D. Manning, GloVE: Global vectors for word representation, in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532–1543.

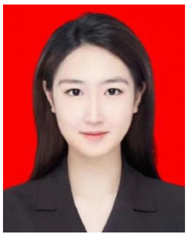
[48] M. Schuster and K. K. Paliwal, Bidirectional recurrent neural networks, *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[49] A. E. W. Johnson, T. J. Pollard, L. Shen, L. -W. H. Lehman, M. Feng, 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.

Dengao Li received the PhD degree from Taiyuan University of Technology in 2010. He is a professor of the College of Data Science at Taiyuan University of Technology. His main research interests include intelligent perception and internet of things technology, air-space-ground integration and navigation technology, and big data analysis technology and application. He is the chairman of Taiyuan Branch of International Computer Society (ACM), and a member of CCF Internet of Things Committee.

Jumin Zhao received the PhD degree from Taiyuan University of Technology in 2008. She is a professor of the College of Information and Computer at Taiyuan University of Technology. Her main research interests include intelligent perception and internet of things technology. She is the deputy secretary-general of Taiyuan Branch of International Computer Society (ACM) and a member of CCF Internet of Things Committee.

Huiting Ma received the BEng degree from Polytechnic Institute, Taiyuan University of Technology in 2018. She is currently pursuing the MS degree in Taiyuan University of Technology. Her research interest is medical data processing.



Wenjing Li received the MS and BS degrees from University of California, Santa Barbara, majoring in actuarial science in 2021. Her research interest is financial/data optimization.

Baofeng Zhao received the PhD degree from Taiyuan University of Technology in 2014. He is an associate professor of the College of Mining Engineering at Taiyuan University of Technology. His main research interests include intelligent perception and internet of things technology.

Yi Liu received the MS degree from Ningxia University in 2018. He is currently pursuing the PhD degree in Taiyuan University of Technology. His research interest is medical datasets.

Jian Fu received the BEng degree from Xi'an University of Science and Technology in 2018 and the MS degree from Taiyuan University of Technology in 2021. Her research interest is medical data mining and processing.