TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 01/16 pp613–627
DOI: 10.26599/TST.2022.9010012
Volume 28, Number 4, August 2023

Classification of Medical Image Notes for Image Labeling by Using MinBERT

Bokai Yang, Yujie Yang, Qi Li, Denan Lin, Ye Li, Jing Zheng*, and Yunpeng Cai*

Abstract: The lack of labeled image data poses a serious challenge to the application of artificial intelligence (AI) in medical image diagnosis. Medical image notes contain valuable patient information that could be used to label images for machine learning tasks. However, most image note texts are unstructured with heterogeneity and short-paragraph characters, which fail traditional keyword-based techniques. We utilized a deep learning approach to recover missing labels for medical image notes automatically by using a combination of deep word embedding and deep neural network classifiers. Bidirectional encoder representations from transformers trained on medical image notes corpus (MinBERT) were proposed. We applied the proposed techniques to two typical classification tasks: Medical image type identification and clinical diagnosis identification. The two methods significantly outperformed baseline methods and presented high accuracies of 99.56% and 99.72% in image type identification and of 94.56% and 92.45% in clinical diagnosis identification. Visualization analysis further indicated that word embedding could efficiently capture semantic similarities and regularities across diverse expressions. Results indicated that our proposed framework could accurately recover the missing label information of medical images through the automatic extraction of electronic medical record information. Hence, it could serve as a powerful tool for exploring useful training data in various medical AI applications.

Key words: MinBERT; convolutional neural network; electronic medical record; medical image labeling; word embedding

1 Introduction

In recent years, the adoption of artificial intelligence (AI) in medical diagnosis has received increasing research attention, and many technological breakthroughs have been made through the use of deep learning. In particular, medical imaging analysis using deep learning approaches has attracted increased attention, with many successful examples being reported. To date, deep learning has been applied to diagnose eye diseases[1, 2], cancer[3–5], pediatric diseases[6], abnormalities in knee magnetic resonance imaging[7], aneurysms[8], and COVID-19[9]. In addition, deep learning has been utilized to predict key medical indicators[10]. Medical imaging is one of the most widely studied topics in medical AI research. However, most achievements in medical imaging have relied on large-scale training sets of reliably labeled image data.

Performance gains in medical imaging learning often require a tremendous amount of training data accompanied by high-quality labels[11–13]. In contrast to conventional imaging recognition tasks wherein image labels can be obtained through crowdsourcing, medical

• Bokai Yang, Yujie Yang, Qi Li, Ye Li, and Yunpeng Cai are with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: bk.yang@siat.ac.cn; yj.yang@siat.ac.cn; qili@link.cuhk.edu.hk; ye.li@siat.ac.cn; cnzhengj@163.com.
• Denan Lin and Jing Zheng are with Shenzhen Health Development Research and Data Management Center, Shenzhen 518055, China. E-mail: ldn308@163.com; cnzhengj@126.com.
∗ To whom correspondence should be addressed.
Manuscript received: 2022-01-05; revised: 2022-04-29; accepted: 2022-05-18

image annotation requires extensive clinical expertise and is thus considerably difficult to expand. Hence, high-quality, large-scale annotated image datasets for medical image analysis are lacking. The scarcity of reliably labeled image data limits the accuracy of AI–based medical image diagnosis[11]. Developing a computational method that efficiently extracts the annotation information of medical images can boost the application of AI methods in medical image diagnosis.

The rapid growth of electronic health-related information has provided new opportunities for disease diagnosis and medical investigations[14, 15]. In particular, the expert diagnostic results contained by electronic medical records (EMRs) can provide valuable annotation information for medical image mining and thus enable the full exploitation of numerous image resources. However, because of the complexity of medical image diagnosis, the EMR counterparts of medical images, which are usually called medical image notes, are mostly written in free-style text with natural language. Thus, structured annotation information is not directly available. Nevertheless, by matching and fusing multiple recording systems, key annotation information, such as major diagnosis and image features, can be recovered from the systems. However, only a small portion of medical images can be annotated by applying such an aggressive information fusion approach. Our previous findings indicated that only approximately 10%–30% of the medical images with free-text notes produced in daily hospital medical systems can be annotated through straight EMR fusion. Therefore, the text mining of free-text medical image notes through natural language processing (NLP) and machine learning may provide substantial benefit by enabling the full use of the large datasets of medical images generated in daily clinical practice.

EMRs are assemblies of heterogeneous medical documents that consist of structured and unstructured data with shared information. Medical image notes are free-style text records that describe the medical image type, observable features, main symptoms, progress notes, and possible diagnoses of patients under examination. Most of these records do not have a fixed format or key item. In addition, these records contain numerous unrecognized medical jargons, abbreviations, incomplete syntactic components, and type errors that considerably hinder the automatic parsing of the corpora. Traditional NLP and machine learning approaches, such as term frequency-inverse document frequency (TF-IDF)[16–18] and latent Dirichlet allocation (LDA)[19, 20], have been used to extract information from unstructured medical texts[21]. However, the application of these methods to medical image notes is limited because of the short paragraphs and aforementioned complex characteristics of such texts. Recently, deep neural networks have drastically improved the capacity of machine learning in sophisticated tasks, including the extraction of information and the interpretation of natural language documents[22, 23]. Although many studies have attempted to use these technologies to mine structured EMRs[24, 25], few have utilized them to extract information from free-style texts, such as medical image notes, because of the difficulties arising from their highly diverse and short-paragraph nature.

In recent years, transformer-based models[26] such as bidirectional encoder representations from transformers (BERT)[27] and generative pre-trained transformer (GPT)[28] have emerged as efficient tools for handling a wide variety of NLP tasks. Some recent works have also attempted to employ BERT (e.g., bidirectional encoder representations from transformers for biomedical text mining (BioBERT)[29] and bidirectional encoder representations from transformers for clinical text mining (ClinicalBERT)[30]) for medical text classification. However, given the lack of domain-specific training data, most current BERT models adopt a very shallow classifier layer over the transformer network pretrained on broad domain data and tasks. Such an approach restricts the accuracy of the classification task.

In this study, we proposed a classification approach for automatically recovering labels for medical image notes by applying machine learning. A large-scale benchmark dataset was developed by incorporating medical records from multiple EMR systems that connect accurate diagnostic labels with the corresponding image notes. We adopted bidirectional encoder representations from transformers trained on medical image notes corpus (MinBERT), a neural network architecture that combines deep word embedding with a deep neural network (DNN) classifier, to handle the short-paragraph nature of medical image notes. We applied the proposed techniques to two typical classification tasks, namely, image type identification and clinical diagnosis identification. Both of these techniques showed improved classification accuracy when compared with traditional NLP, machine learning techniques, biomedical-based BERT (BioBERT), and clinical BERT

(ClinicalBERT). Notably, the binary classifiers of MinBERT adopted a seven-layer feedforward neural network, which was more complex than the classifiers of BioBERT and ClinicalBERT. In Section 5, we discuss the reasons for the better performance of MinBERT than that of the baselines. Our results indicated that deep neural networks with embedding representation can accurately recover the missing label information of medical images and can be adopted for the automatic extraction of EMR information for other data mining tasks. The proposed technique can serve as a powerful tool for exploring training data in various medical AI applications.

In particular, we made the following contributions:

(1) We constructed a large-scale corpus of medical image notes by incorporating medical records from various clinics and hospitals.

(2) We built a BERT model trained on the large-scale corpus of medical image notes: MinBERT. Its binary classifier was a seven-layer neural network, which was more complex than the classifiers of BioBERT and ClinicalBERT.

(3) We demonstrated that MinBERT outperformed traditional NLP, machine learning techniques, and the domain-specific pretrained BERT model in missing label information recovery tasks.

2 Material and Method

2.1 Material

The data used in this study were obtained from the Shenzhen regional medical information system that contains medical records from various clinics and hospitals. Medical image notes and diagnostic labels were obtained from the Picture Archiving and Communication System (PACS) and Hospital Information System (HIS), respectively. Data fusion was performed across systems for patients with the same personal IDs. Six types of entities were synthetically selected from dozens of entities in the PACS and HIS to obtain a robust and complete final model: exam department, image type, Chinese name, description, examination result (from the PACS), and clinical diagnosis (from the HIS). An example of a complete note is illustrated in Table 1. The field contents were originally in Chinese, and their English translations are provided in the table.

We used the record time range as a criterion to match medical image notes with their corresponding diagnostic labels. This approach enabled us to retain only medical image notes before the diagnostic labels. The Shenzhen regional medical information system consisted of 282 740 notes describing medical image diagnosis. The data uploaded to the Shenzhen regional medical information system are considerably limited because of the uneven development of medical information in different hospitals and clinics and the inconsistent specifications and requirements of hospitals and clinics at all levels. Only 74 914 notes contained complete information regarding the six types of entities. This situation indicated that more than 70% of the notes lacked information regarding at least one entity. After processing the complete data for all six types of entities, 15 466 notes with clear diagnostic results were selected. Table 2 shows the distribution statistics of the data. In the case of binary classification, 15 466 notes (including the multilabel notes of clinical diagnosis) with complete label information were used in our research for model training and validation, and the results were applied to the remaining notes. In the case of multiclassification, no multiclassification experiment was performed for clinical diagnosis because clinical diagnosis fields may have multiple labels. Furthermore, because all medical records were collected during routine clinical activities and the obtained data were anonymous, a waive-of-consent protocol was adopted and approved by the Institutional Review Board of Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences (No. SIAT-IRB-151115-H0084), in accordance with the Guidelines of the World Medical Association (WMA) Declaration of Helsinki term 32.

Table 1 Example of a medical image note record with six free-text fields.

Field	Field content (translated from Chinese into English)
Exam department	Cardiovascular medicine
Image type	Ultrasound (US)
Chinese name	Bilateral carotid artery color doppler ultrasound
Description	Local thickening of the medial common carotid artery in the right common carotid artery.
Result	The intima-media thickness of the right common carotid artery enlargement was about 1.5 mm. There was no obvious intima-media thickness of the left common carotid artery.
Clinical diagnosis	Coronary heart disease

Table 2 Distribution statistics of the data.

Type	Distribution		
	Class	Number	Proportion (%)
Sex distribution	Male	7855	50.79
	Female	7006	45.30
	Unknown	605	3.91
Age	0–30	1334	8.62
	31–50	5255	33.98
	51–70	5927	38.32
	Over 71	2692	17.41
	Unknown	258	1.67
Clinical diagnosis	Coronary heart disease	4416	28.55
	Hypertension	3982	25.75
	Diabetes	2182	14.11
	Pulmonary tuberculosis	2289	14.80
	Abdominal pain	1157	7.48
	Others	2484	16.06
Image type	Ultrasound (US)	4364	28.22
	Computed tomography (CT)	3265	21.11
	X-ray Examination (XR)	3030	19.59
	Endoscope (GS)	2162	13.98
	Magnetic resonance (MR)	1959	12.67
	Others	686	4.44

2.2 Method

The aim of this study is to extract useful label information, such as diagnostic results, for medical image classification by using unstructured image notes derived from EMRs and a relatively small number of readily annotated cases as training examples. Traditional keyword-based NLP techniques, such as TF-IDF encoding and LDA, and machine learning techniques often fail to retrieve useful labeling information correctly because of the highly diverse and short-paragraph nature of image notes. However, in this work, we demonstrated that precise information extraction is still possible if an adequate number of annotated training instances are available. We used a deep learning approach to recover missing labels for medical image notes automatically by combining deep word embedding and deep neural network classifiers to capture semantic relationships between text features and key annotations from a large-scale dataset.

2.2.1 Our method

MinBERT. MinBERT is a BERT model trained on a large-scale corpus of medical image notes. BERT is a transformer-based machine learning technique for NLP pretraining developed by Google. BERT includes a word embedding layer in which words are transformed into vectors with respect to their relationships with other words. However, in contrast to those in previous embedding techniques, the coordinates of the vector in BERT are not defined by other words, thus enabling the representation of high-level semantic features. The basic feature of the BERT network is called a transformer. This feature is an encoder-decoder structure formed by stacking several encoders and decoders. Figure 1 presents the network architecture of the transformer. The left part of Fig. 1 shows the encoder, which uses multiheaded attention and a fully connected layer to convert the input corpus into feature vectors. The right part of Fig. 1 shows the decoder, whose input is the output of the encoder along with the predicted result. The decoder applies masked multiheaded attention and a fully connected layer to output the conditional probability of the final result. Furthermore, in contrast to the recurrent neural network and the convolutional neural network (CNN) architectures, BERT converts the distance between two words at any position into one through the attention mechanism. It thus effectively solves the long-term dependence problem of NLP.

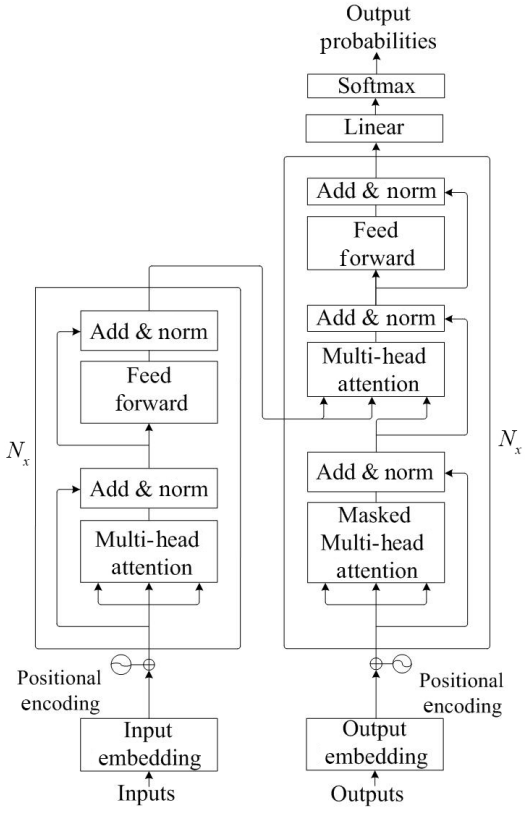


Fig. 1 Network architecture of a transformer.

BERT aims to learn a satisfactory feature representation for words through self-supervised learning on the basis of a massive corpus. Self-supervised learning refers to supervised learning that runs on data without manual annotation. In future-specific NLP tasks, the feature representation in BERT can be used as the word embedding for other tasks. Therefore, BERT provides a model for transfer learning and can be used as a feature extractor after fine-tuning in accordance with the task. In contrast to previous language representation models, BERT is designed to pretrain deep bidirectional representations from the unlabeled text by jointly conditioning the left and right contexts in all layers. Therefore, the pretrained BERT model can be fine-tuned with only one additional output layer to create state-of-the-art models for various tasks, such as question answering and language inference, without substantial task-specific architectural modifications[27].

Recent empirical improvements based on transfer learning with language models have demonstrated that rich, unsupervised pretraining is an integral part of many language-understanding systems. In particular, these improvements enable even low-resource tasks to benefit from deep unidirectional architectures. The major benefit of BERT is that it further generalizes these findings to deep bidirectional architectures, thus allowing the same pretrained model to successfully tackle a broad set of NLP tasks. Figure 2 presents the MinBERT model in a text classification task. The left part of Fig. 1 shows a transformer block that corresponds to a "TRM" in the part of MinBERT provided in Fig. 2. Classifier layers for consequence-specific tasks can usually be simple single-layer or few-layer neural networks because BERT can efficiently represent text features by pretraining. MinBERT adopts a seven-layer neural network to improve the performance of the model. In Fig. 2, CLS represents the beginning of a sentence, $\boldsymbol{E}$ represents the embedding vector and $\boldsymbol{T}_i$ represents the eigenvector of token i after MinBERT processing.

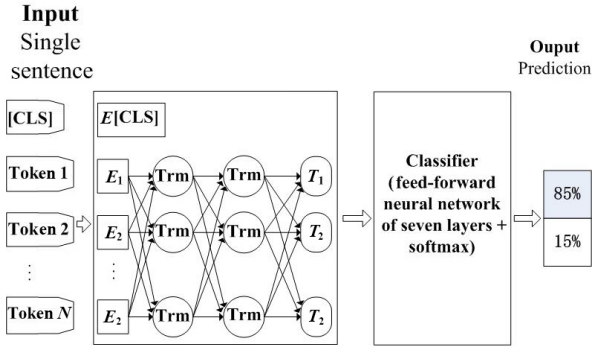


Fig. 2 Diagram of MinBERT model in a text classification task.

2.2.2 Baseline methods

Word2Vec. Word embedding is a general term for a language model and representation learning technology in NLP. Conceptually, this technique refers to embedding a high-dimensional space with the number of all words into a continuous vector space with a low dimension wherein each word or phrase is mapped to a vector on the real number field. The use of word embedding to represent phrases has considerably improved the effects of syntax and text sentiment analysis in NLP. The continuous bag-of-words (CBOW) model, a frequently used approach for word embedding, learns word representations that can be efficiently trained by using a large amount of textual data[31]. In contrast to a language model that can predict each next word in the corpus on the basis of past words, a model that is only required to produce suitable word embeddings is not subject to such restrictions. Therefore, we used n words before and after the target word w_t to predict target word. This approach is used in the CBOW model because it uses continuous representations in no particular order of importance[32]. The purpose of the CBOW model is only marginally different from that of the language model:

$$J_\theta = \frac{1}{T}\sum_{t=1}^{T} \lg p(w_t | w_{t-n}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+n}) \tag{1}$$

As shown in Eq. (1), J_θ represents loss function, T represents the number of words, p represents the probability. Rather than feeding n previous words to the model, the model receives a window of n words around the target word w_t at each time step t.

A recent study reported that the distributed representations of words capture many linguistic regularities and that many similarities between words can be expressed as linear translations[33]. In this study, we implemented Word2Vec by adopting the Python Gensim package and the following specific parameters: the trained vector length (embedding size) of 128, the minimum count of the word frequency value of 1, and the maximum distance between current and predictive words of 5.

CNN. CNNs are a specialized type of neural network for processing data that have a known grid-like topology[34]. They execute two key functions[35, 36]. First,

they perform feature extraction. The input of each neuron is connected to the local receptive domain of the previous layer, and the feature of the local region is extracted. Once the local feature is extracted, its positional relationship with other features is determined. Second, they perform classification. Each computing layer of the network consists of several feature maps. Each feature map is a two-dimensional plane. The weights of all neurons on the plane are shared. Finally, a fully connected layer is used as the classification layer to compute feature vectors and output classification results. In particular, CNNs are used to train the vector space model generated in the previous step, extract features, and classify. The TensorFlow framework can be used to build CNNs; the input can be generated by using the TF-IDF or LDA algorithms or the Word2Vec vector space model. This process produces the first layer of a CNN. Subsequently, convolutional and pooling layers are used to extract abstract features in sequence. This work used convolution kernels with sizes of 3, 4, and 5 in the first convolutional layer, and the results of the three kernels were merged into the maximum pooling layer. Finally, a fully connected layer is followed by the activation layer, which uses the softmax activation function to output a classification probability. During training, the difference between predicted and true values is evaluated by using a loss function, and the parameters of the training model are optimized using the back propagation algorithm. A suitable parameter must be set for the dropout, and the training speed (learning rate) and number of epochs should be determined. If the training model is as expected, then as the number of iterations increases, the classification performance improves, and the loss continuously decreases. If the training model is overfitting or does not converge, then the training should be stopped, and the parameters and structures should be adjusted in time to restart the training. Figure 3 shows the structure of the CNN.

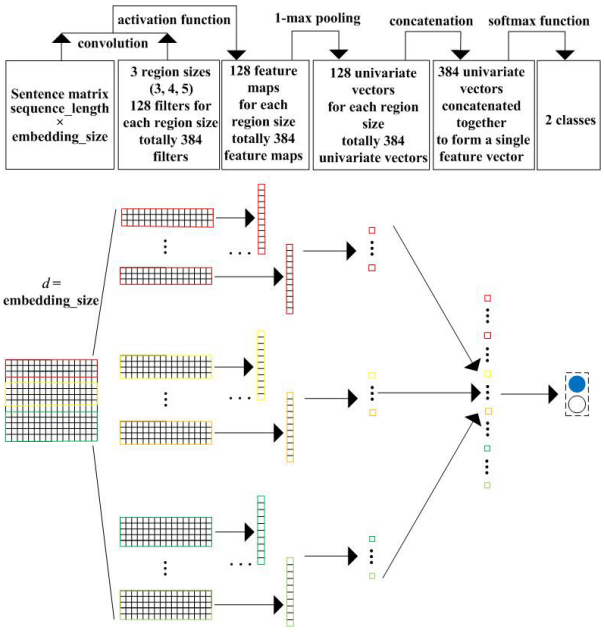


Fig. 3 Structure chart of the CNN.

TF-IDF. TF-IDF[17, 18] is a mathematical algorithm that is commonly used to assess the importance of a word for a particular piece of text in a dataset. Words that appear frequently in text segments are highly important, whereas those that appear frequently in the dataset are less important. In this work, we used TF-IDF to map texts to a matrix.

LDA. LDA[19, 20] is a powerful learning algorithm that is utilized to cluster words into topics and documents into mixtures of topics automatically and jointly. A topic model is a hierarchical Bayesian model that associates each document with a probability distribution over "topics", which are in turn distributions over words. For example, a topic in a collection of newswires might include words, such as "sports", which in turn includes "baseball," "home run", and "player", and a document on steroid use in baseball might include "sports", "drugs", and "politics". The labels "sports", "drugs", and "politics" are posthoc labels assigned by a human, and the algorithm itself only assigns probabilities to associated words. In this model, parameter estimation learns what the topics are and which documents employ them in what proportions. In this study, we used LDA to map texts to a matrix.

Random forest. Random Forest (RF)[37, 38] is an algorithm in which forests of unrelated decision trees are established randomly. After training an RF model, each decision tree in the RF can classify a new input. Finally, which category is chosen to be more frequently, it is categorized into that category. In this paper, we applied RF to classify the vectorized matrix.

BioBert. BioBert[29] is a domain-specific language representation model pretrained on large-scale biomedical corpora. First, BioBERT is initialized with weights from BERT. Then, BioBERT is pretrained on biomedical domain corpora (PubMed abstracts and PMC full-text articles). Finally, BioBERT is fine-tuned and evaluated on downstream tasks. BioBERT significantly outperforms BERT in three representative biomedical text mining tasks: Biomedical named entity recognition, biomedical relation extraction, and biomedical question

answering.

ClinicalBERT. ClinicalBERT[30] is a BERT model for clinical text. ClinicalBERT is pretrained on clinical domain corpora (clinical text in MIMIC-III). ClinicalBERT outperforms various baselines in 30-day hospital readmission prediction by using discharge summaries and the first few days of notes on various clinically motivated metrics in intensive care units.

3 Experimental Setup

3.1 Experimental environment

Experiments were performed by using the prevailing deep learning framework TensorFlow version 2.3.1 with the Windows Server 2008 operating system. The TensorFlow platform was deployed on a high-performance computing server equipped with two 10-core Intel Xeon E5-4610 V2 processors with 128 GB of memory. In addition, the computing server was equipped with four NVIDIA Tesla RTX8000 GPUs with 48 GB of memory (total memory of 192 GB). The programing language used was Python 3.6.

3.2 Experimental framework

Figure 4 provides a detailed schematic of the experimental framework.

The first step was data preprocessing. Medical image notes and diagnostic labels were obtained from the PACS and HIS, respectively. First, we performed data fusion across systems for patients with the same personal IDs. Six types of entities were synthetically selected from the dozens of entities in the PACS and HIS to obtain a robust and complete final model: Exam department, image type, Chinese name, description, examination result (from the PACS), and clinical diagnosis (from the HIS). Then, we selected 15 466 complete valid instances from more than 280 000 EMRs for the final experiment by using SQL and saved the data in a file suffixed.csv. We selected positive and negative cases from the dataset and labeled them as 1 and 0, respectively. We used Python's readline function to read the data line by line. Next, we segmented the text[39, 40] by utilizing Jieba, the best Python Chinese word segmentation module[41]. Before classification, we removed some words[42], including punctuation and place names, that were not crucial to the classification. Then, we used the image type or clinical diagnosis as labels to train the models.

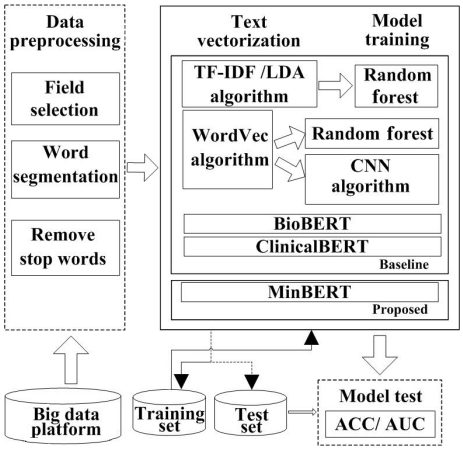


Fig. 4 Detailed schematic of the proposed experimental approach and the baseline approach for comparison. ACC represents accuracy and AUC represents the area under the receiver operating characteristic curve.

In the second step, we established a vector space model. We used MinBERT to train word embedding because natural language cannot be used by computers. In this step, we replaced MinBERT with Word2Vec, TF-IDF, and LDA as the baseline algorithms for mapping texts to a matrix. The dimensionality of the Word2Vec embedding space was empirically determined as 128[43].

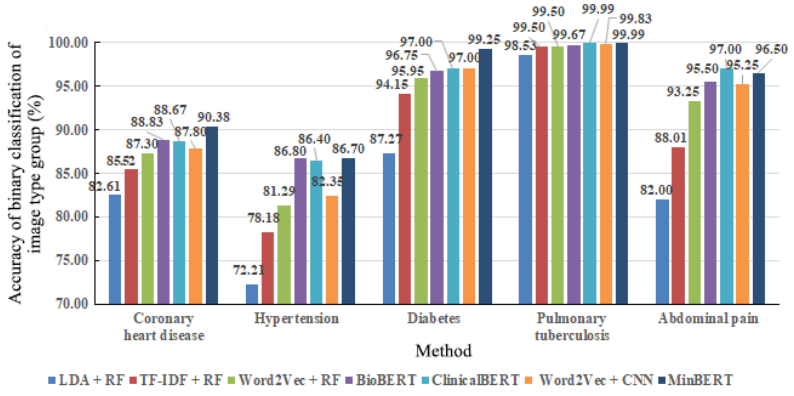
The final step involved setting up and testing a machine learning model. The aforementioned matrix was used as the features, and 80% of the dataset was used for training. The remaining 20% of the data were used to test the trained model. We applied the CNN and RF as the baseline methods to train the data.

MinBERT, BioBERT, and ClinicalBERT do not require data preprocessing and can be used directly to classify original data. We adjusted the model parameters to determine the best parameters for classification[44]. Early stopping was applied to prevent overfitting". The remaining 20% of the dataset was used to test the trained model. The 80/20 splits were stratified such that the proportions of positive and negative cases were the same as those in the complete dataset.

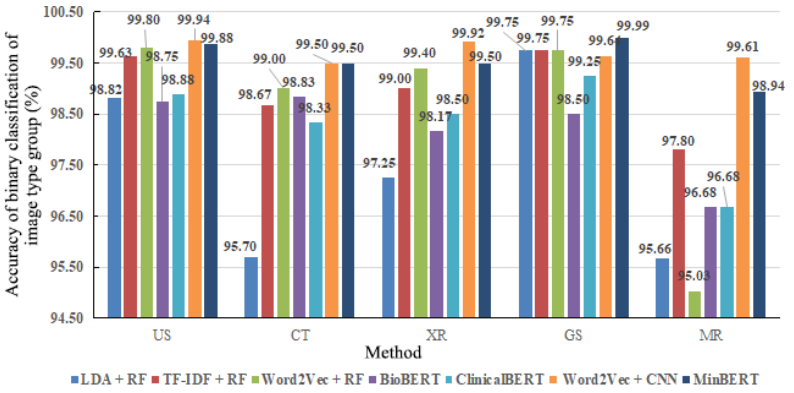
4 Result

We performed binary classification experiments with two groups, namely, clinical diagnosis and image type. Each group contained 35 small binary classification experiments. In this study, accuracy and the area under the receiver operating characteristic (ROC) curve (AUC) were determined to evaluate classification performance. Accuracy was used as a statistical measure to determine how efficiently a binary classification test could correctly identify or exclude a condition. In other words, accuracy was the proportion of correct predictions (true positives and true negatives) among the total number of cases

examined. The ROC curve was obtained by plotting the true positive rate against the false positive rate at various threshold settings. When using normalized units, the AUC was equal to the probability that a classifier would rank a randomly chosen positive instance higher than a randomly chosen negative instance (assuming that "positive" ranks higher than "negative"). Notably, this work recorded the average performance of five repeated classification tasks on each dataset. The experimental results of clinical diagnosis and image type obtained with LDA + RF, TF-IDF + RF, Word2Vec + RF, BioBERT, ClinicalBERT, Word2Vec + CNN, and MinBERT are presented in Fig. 5. In the clinical diagnosis group, Word2Vec+RF yielded an average accuracy of 91.46%, which was higher than the average accuracies of 89.07% and 84.52% exhibited by TF-IDF + RF and LDA + RF, respectively. In the image type group, Word2Vec + RF yielded an average accuracy of 98.5%; however, LDA + RF presented an average accuracy of only 97.44%. These results indicated that Word2Vec was a more efficient method for vectorizing texts in short-paragraph situations than TFIDF and LDA. In the clinical diagnosis group, MinBERT exhibited an average accuracy of 94.56%, which was the highest among the accuracies of all methods (LDA + RF, TF-IDF + RF, Word2Vec + RF, BioBERT, ClinicalBERT, Word2Vec + CNN, and MinBERT). Moreover, MinBERT yielded an average accuracy of more than 99.5% in the image type group. This result showed that deep word embedding and deep neural network were more efficient than traditional NLP (e.g., TF-IDF encoding and LDA), machine learning techniques (e.g., Word2Vec + RF and Word2Vec + CNN), and biomedical or clinical specific contextual embedding's (e.g., BioBERT and ClinicalBERT). We performed a t-test on the results of MinBERT and ClinicalBERT (the most typical baseline) to show the difference between the performance of MinBERT and that of the baselines quantitatively. The results are provided in Table 3. Table 3 demonstrates the results of MinBERT were significantly different from those of ClinicalBERT. Meanwhile, the ROC curves of the clinical diagnosis and image type groups with MinBERT are provided in Fig. 6. Sometimes, our model could not perform the classification correctly. For example, 44 notes were wrongly classified in the binary



(a) Clinical diagnosis group

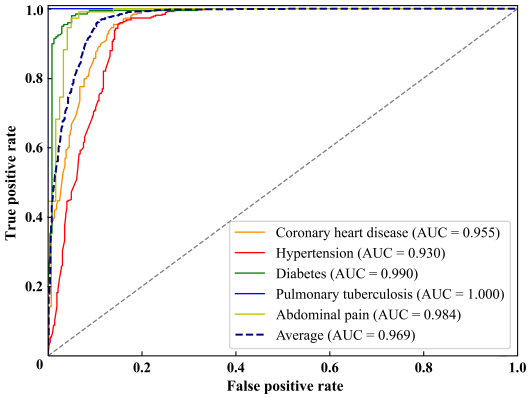


(b) Image type group

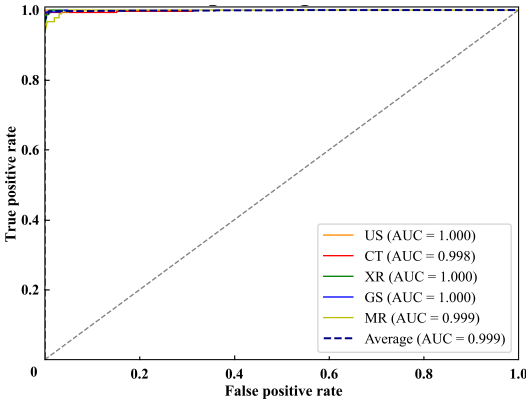
Fig. 5 Experimental results of binary classification.

Table 3 T-test on the result of MinBERT and ClinicalBERT.

Type	Distribution		
	Class	*t*-Rate	*P*-value
Clinical diagnosis	Coronary heart disease	3.4657	0.0108
	Hypertension	3.3541	0.0195
	Diabetes	6.066	0.0004
	Pulmonary tuberculosis	1.6330	0.1778
	Abdominal pain	−2.306	0.0528
Image type	US	6.4746	0.0018
	CT	5.0937	0.0015
	XR	3.7995	0.0053
	GS	3.5648	0.0235
	MR	4.9568	0.0011



(a) Clinical diagnosis group



(b) Image type group

Fig. 6 Receiver operating characteristic curves of binary classification with MinBERT.

classification of hypertension with MinBERT. These instances of misclassification were divided into the following situations: (1) the model could not find useful information because the notes were too short or had missing import fields; (2) the part of the imaging examination did not include the heart; and (3) given that hypertension lacks obvious image indications, the notes did not contain adequate medical meaning.

We used LDA + RF, TF-IDF + RF, Word2Vec + RF, BioBERT, ClinicalBERT, Word2Vec + CNN, and MinBERT for multiclassification problems in the image type group. The five types of the classification results of US, CT, XR, GS, and MR are provided in Table 4. MinBERT obtained an accuracy of more than 99.5% in multiclassification problems; this accuracy was higher than the accuracies of LDA + RF, TF-IDF + RF, Word2Vec + RF, BioBERT, ClinicalBERT, and Word2Vec + CNN. Figure 7 presents the confusion matrix results of multiclassification with MinBERT. MinBERT made few mistakes in the five types of classification of US, CT, XR, GS, and MR.

We used t-SNE[45] to visualize the word vectors generated by Word2Vec to understand how word embedding maps crucial keywords. The results are illustrated in Fig. 8 (translated from Chinese into English). Theoretically, word embedding maps each keyword into a numerical vector in a high-dimensional space, and words with similar semantic meanings would be mapped into similar vectors. T-SNE then projects the high-dimensional space into a low-dimensional one to illustrate intuitively which words are mapped into close positions. By inspecting the projected vectors, we

Table 4 Results of the five types of classification of US, CT, XR, GS, and MR.

Method	ACC	Recall	F1-score
LDA + RF	0.9045	0.9042	0.9033
TF-IDF + RF	0.9817	0.9816	0.9816
Word2Vec + RF	0.9859	0.9858	0.9858
BioBERT	0.9552	0.9551	0.9543
ClinicalBERT	0.9674	0.9670	0.9660
Word2Vec + CNN	0.9947	0.9947	0.9947
MinBERT	0.9952	0.9952	0.9952

	US	CT	XR	GS	MR
US	100%	0	0	0	0
CT	0	98.96%	1.04%	0	0
XR	0	0.55%	99.18%	0.27%	0
GS	0	0.55%	0.27%	99.18%	0
MR	0	0.81%	0	0	99.19%

Fig. 7 Confusion matrix results of the five classifications of US, CT, XR, GS, and MR with MinBERT.

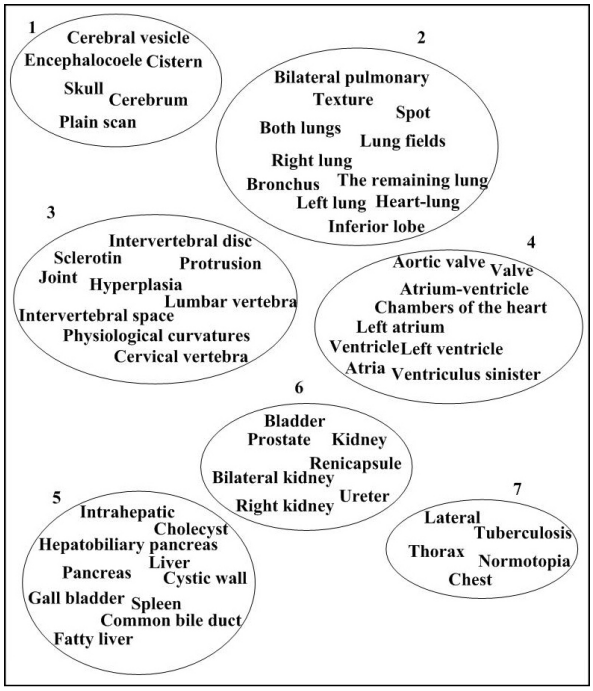


Fig. 8 Visualization results of Word2Vec.

could intuitively see that Word2Vec created clusters of words with prominent semantic meanings. For example, words that are anatomically related to cerebral structures gathered in Cluster 1 in Fig. 8. Clusters 2, 3, 4, 5, 6, and 7 represented concepts related to the respiratory tract, the skeletal system, the cardiovascular system, the abdominal organs, the urinary system, and the thorax, respectively. Notably, these concepts were learned purely by word-word associations from training data without any a priori knowledge provided. Furthermore, due to the short-paragraph nature of the medical texts, words in the same cluster rarely coexist in the same medical note. Therefore, traditional NLP approaches based on word-word or word-topic associations would likely fail to capture these concepts. Moreover, we see that the concepts learned by word embedding were somehow different from knowledge-derived ones, such that "tuberculosis" clustered with thorax words and "texture" clustered with respiratory (chest) words, implying their casual relationships in diagnostic notes.

We then explain how MinBERT used the learned embedding features to predict text labels with the help of the local interpretable model-agnostic explanations (LIME)[46] visualization tool. However, similar to that learned by most deep neural network models, the decision surface learned by MinBERT was highly nonlinear. LIME provides a simple illustration of how each feature contributes to the overall prediction probability of a given input sample and thus yields insights into the mechanism of the created model. Figure 9 shows an example of the LIME visualization result of a sample digital radiology X-ray examination note. We could determine the probability that a patient has coronary heart disease (CHD), and we could also identify the information on which the MinBERT model was based. As shown at the top of Fig. 9, the probability that the patient had CHD was 0.85. The deepening of the orange color of the text part at the bottom of Fig. 9 indicated that this part was considered crucial to the model's judgment of CHD. Furthermore, the deepening of the blue color indicated that this part was crucial to the model's judgment that CHD was absent. In the middle part, the right side represented the basis for the model's judgment if a patient had CHD, and the left side represented the basis for the model's judgment that the person did not have CHD. The features supporting CHD were mostly cardio-related descriptions, such as "aortic shadow tortuous and widened", "left ventricular enlarged", and "cardiothoracic ratio (enlarged)", which are important pre-CHD symptoms. "Pulmonary artery segment depressed" and "lung texture enhanced" are also known to be cardiac-related symptoms. The

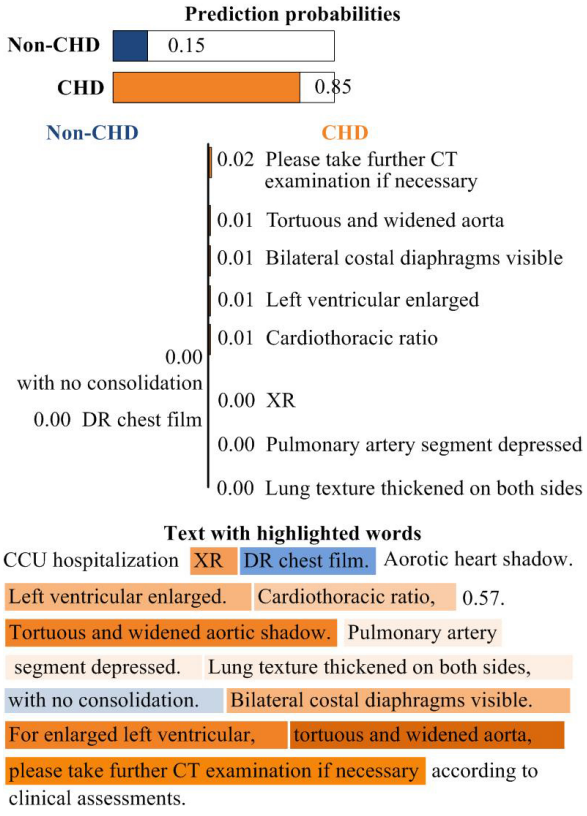


Fig. 9 Visualization of the MinBERT results.

term "further CT examination" was associated with the detection of high CHD risk, although it did not directly mention the clinical observation. This example illustrated that with the combination of word embedding and deep learning, we can combine weakly correlated keyword features for the precise inference of text content and thus overcome the challenges posed by the substantial heterogeneity of medical texts.

5 Discussion and Conclusion

The data mining of large medical information repositories accumulated during daily diagnosis activities is widely believed to provide beneficial knowledge and insights to clinical research. Medical texts and images are currently the most important types of medical information data. The HIS and PACS systems have been widely adopted for the efficient management and extraction of useful clinical information. However, despite the tremendous efforts that have been made to structuralize medical records, a considerable amount of fact information remains concealed in medical records. Such information cannot be directly extracted from structured databases due to the complexity of human language and knowledge inferences. Moreover, in practical situations, most electronic medical records are actually not yet structuralized or have severe losses of important messages. Hence, adopting artificial intelligence technologies to identify complex facts, concepts, and relationships from semistructured or unstructured medical texts has long attracted extensive research interest. Nevertheless, traditional NLP and machine learning techniques for text mining are highly dependent on the distributions of word frequencies, which are vulnerable to highly heterogeneous texts that lack stable keyword-concept (or topic) relationships. Medical texts, which are mostly produced amid hectic clinical activities in the form of quick notes, are sparsely featured with deformed sentence structures, high heterogeneity, short paragraphs, and even misspellings. As a result, despite numerous attempts and some promising examples, the accurate and efficient identification of clinical facts and concepts from medical record instances remains a crucial bottleneck in medical data mining.

In this research, we focused on the problem of classifying medical image notes for imputing EMR labels as an example of medical text mining because of two reasons: First, the automatic identification of medical images has demonstrated great potential in assisting diagnostic decisions, and imaging-based medical AI has attracted considerable interest recently. Nevertheless, most current AI models for processing image data adopt deep learning, which relies heavily on large amounts of accurately labeled training data. However, such data are scarce and difficult to accumulate in most clinical situations. Processing stock image note texts contained by EMRs to obtain high-quality patient labels would greatly benefit future medical AI studies. Second, medical image notes are typical examples of short, heterogeneous, and deformative medical texts, which are suitable for validating new computational technologies targeting the features of medical texts.

With a large corpus of 282 740 image notes, we adopted the combination of deep word embedding and deep neural networks to create accurate classification models for medical image note data. We demonstrated that Word2Vec, a typical word embedding approach, can grasp the semantic similarity between words without prior domain knowledge and despite diverse term expressions. In contrast to traditional approaches that focus on single word frequency and single word-word or word-topic links, Word2Vec converts each word into an embedding vector with each coordinate component representing its relationship with a context word. Thus, it memorizes the overall semantic location of a word within a corpus rather than learning individual links. This ability is achieved by training a two-layered neural network by using a bag of adjacent words as input. As shown in Fig. 8, Word2Vec completed the mapping of crucial keywords and clustered similar words at adjacent positions and thus basically formed knowledge about specified concepts without a priori domain knowledge or human intervention. At the same time, the mapped matrix contained complete original text information, and similar texts were gathered. This situation contributed to the efficient functioning of Word2Vec in vector creation. By contrast, the diversity of medical notes limited the effectiveness of conventional keyword matching or word frequency statistics (TF-IDF). Meanwhile, topic models such as LDA are hindered by the short-paragraph nature of the medical text dealt with in this work. This situation is consistent with the observation that word embedding outperformed the traditional approach in our experiment.

We then applied MinBERT, a recently developed deep embedding technique, to the designated problem and determined whether it further improved classification performance. MiniBERT introduced more favorable features than previous embedding models, such as

Word2Vec. First, MinBERT is context-aware and would thus produce multiple embedding results for the same word with polysemantic meanings, whereas previous embedding approaches are context-independent. Hence, unlike those obtained by previous methods, the embedding vector coordinates obtained by MinBERT were no longer associated with a fixed context word, thus conferring increased flexibility to the model and helping reduce the dimension of the embedded space. Second, MinBERT adopts a special mechanism called attention that selectively assigns different link weights to adjacent words. This mechanism enables the discovery of distant semantic connections in a long sentence. Therefore, the semantic association model obtained by using MinBERT is more refined than those obtained with previous embedding approaches. Compared with Word2Vec+CNN, MinBERT has better classification performance and requires fewer network layers in the classification phase because it learns embedded representations better, indicating that the model may potentially have better generalization capacity. MinBERT's binary classifier adopts a seven-layer feedforward neural network and is more complex than the classifiers of BioBERT and ClinicalBERT. The shallow classifiers likely hurt the performances of BioBERT and ClinicalBERT. BioBERT is pretrained on biomedical domain corpora (PubMed abstracts and PMC full-text articles). ClinicalBERT is pretrained on clinical domain corpora (clinical text in MIMIC-III). A deep classifier network could not be used given the relatively small amount of domain-specific data in BioBERT and ClinicalBERT. In addition, MinBERT was trained completely on domain-specific data. MinBERT used general corpus data for pretraining to compensate for the lack of domain-specific data. Figure 9 shows that MinBERT not only helped achieve increased classification accuracy but also provided explainable information on how the machine-learned concept provided support to the neural network classifier model.

Another favorable feature resulting from the use of word embedding is that the training of the embedding model requires only plain texts, which can be neither necessarily labeled nor directly related to the target task. Although the labeled corpus in our study was sufficiently large for training and unlabeled notes were not used, the decomposition of the pretrained embedding model and classification model allowed the transfer of the trained embedded model to other tasks and the fusion of multiple datasets, which would further enhance model performance in highly complex tasks.

In summary, we focused on the challenge of efficient information extraction from medical images and note text data, which are typically unstructured and heterogeneous with brief paragraphs. We proposed using the combination of word embedding and deep neural networks to learn a context-aware semantic representation of medical concepts solely from a medical text corpus without prior domain knowledge or human intervention and then to identify key information, such as diagnosis results, related diseases, or examination settings, with high accuracy. We demonstrated that the proposed approach is powerful in accurate information extraction with respect to the two target tasks of recovering disease diagnosis labels and examination type labels for medical images and significantly outperformed traditional methods. The proposed method further supports the efficient generation of large, high-quality medical image datasets from daily hospital information systems, given that it provides a powerful, fast, accurate, and low-cost solution for recovering missing labels in diagnostic medical images. Moreover, the proposed methods are highly extensible and transferrable to general medical text mining tasks, such as query answering, disease risk prediction, and clinical decision support, and would serve as powerful tools in various medical AI applications.

Acknowledgment

This work was supported in part by the Shenzhen Science and Technology Program (No. JCYJ20180703145002040), the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDB38050100) and the Shenzhen Science and Technology Program (No. JCYJ20180507182818013).

References

[1] D. S. Kermany, M. Goldbaum, W. J. Cai, C. C. S. Valentim, H. Y. Liang, S. L. Baxter, A. Mckeown, G. Yang, X. K. Wu, F. B. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

[2] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, et al, Clinically applicable deep learning for diagnosis and referral in retinal disease, *Nat. Med.*, vol. 24, no. 9, pp. 1342–1350, 2018.

[3] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, and A. Tsirigos, Classification and mutation prediction from non-small cell lung cancer histopathology images using

deep learning, *Nat. Med.*, vol. 24, no. 10, pp. 1559–1567, 2018.

[4] R. H. Xu, W. Wei, M. Krawczyk, W. Q. Wang, H. Y. Luo, K. Flagg, S. H. Yi, W. Shi, Q. L. Quan, K. Li, et al., Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma, *Nat. Mater.*, vol. 16, no. 11, pp. 1155–1161, 2017.

[5] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, et al., International evaluation of an AI system for breast cancer screening, *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

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

[7] N. Bien, P. Rajpurkar, R. L. Ball, J. Irvin, A. Park, E. Jones, M. Bereket, B. N. Patel, K. W. Yeom, K. Shpanskaya, et al., Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet, *PLoS Med.*, vol. 15, no. 11, p. e1002699, 2018.

[8] A. Park, C. Chute, P. Rajpurkar, J. Lou, R. L. Ball, K. Shpanskaya, R. Jabarkheel, L. H. Kim, E. Mckenna, J. Tseng, et al., Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model, *JAMA Netw. Open*, vol. 2, no. 6, p. e195600, 2019.

[9] J. Chen, L. L. Wu, J. Zhang, L. Zhang, D. X. Gong, Y. L. Zhao, Q. X. Chen, S. L. Huang, M. Yang, X. Yang, et al., Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography, *Sci. Rep.*, vol. 10, no. 1, p. 19196, 2020.

[10] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. B. Liu, J. Marcus, M. M. Sun, et al., Scalable and accurate deep learning with electronic health records, *npj Digital Med.*, vol. 1, no. 1, p. 18, 2018.

[11] K. Yan, X. S. Wang, L. Lu, and R. M. Summers, DeepLesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning, *J. Med. Imaging*, vol. 5, no. 3, p. 036501, 2018.

[12] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, ImageNet: A large-scale hierarchical image database, in *Proc. 2009 IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 1175–1181.

[13] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, Microsoft COCO: Common objects in context, in *Proc. 13th European Conf. on Computer Vision*, Zurich, Switzerland, 2014, pp. 740–755.

[14] G. Mujtaba, L. Shuib, R. G. Raj, R. Rajandram, K. Shaikh, and M. A. Al-Garadi, Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection, *PLoS One*, vol. 12, no. 2, p. e0170242, 2017.

[15] M. Li, Z. H. Fei, M. Zeng, F. X. Wu, Y. H. Li, Y. Pan, and J. X. Wang, Automated ICD-9 coding via a deep learning approach, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 16, no. 4, pp. 1193–1202, 2019.

[16] J. Martineau and T. Finin, Delta TFIDF: An improved feature space for sentiment analysis, in *Proc.3rd IEEE Int. Conf. on Weblogs and Social Media*, San Jose, CA, USA, 2009.

[17] P. Soucy and G. W. Mineau, Beyond TFIDF weighting for text categorization in the vector space model, in *Proc. 19th Int. Joint Conf. on Artificial Intelligence*, Edinburgh, UK, 2005, pp. 1130–1135.

[18] X. M. Ye, X. M. Mao, J. C. Xia, and B. Wang, Improved approach to TF-IDF algorithm in text classification, (in Chinese), *Comput. Eng. Appl.*, vol. 55, no. 2, pp. 104–109, 161, 2019.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[20] L. Li and Y. Zhang, An empirical study of text classification using latent dirichlet allocation, *Cs. Cmu. Edu.*, no. 1, 2018.

[21] T. François and E. Miltsakaki, Do NLP and machine learning improve traditional readability formulas? in *Proc. 1st Workshop on Predicting and Improving Text Readability for Target Reader Populations*, Montréal, Canada, 2012, pp. 49–57.

[22] X. E. Liu, X. X. You, X. Zhang, J. Wu, and P. Lv, Tensor graph convolutional networks for text classification, in *Proc. 34th AAAI Conf. on Artificial Intelligence*, New York, NY, USA, 2020, pp. 8409–8416.

[23] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[24] W. C. Sun, Z. P. Cai, Y. Y. Li, F. Liu, S. Q. Fang, and G. Y. Wang, Data processing and text mining technologies on electronic medical records: A review, *J. Healthc. Eng.*, vol. 2018, p. 4302425, 2018.

[25] W. C. Sun, Z. P. Cai, F. Liu, S. Q. Fang, and G. Y. Wang, A survey of data mining technology on electronic medical records, in *Proc. IEEE 19th Int. Conf. on e-Health Networking, Applications and Services (Healthcom)*, Dalian, China, 2017, pp. 1–6.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[27] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[28] L. Floridi and M. Chiriatti, GPT-3: Its nature, scope, limits, and consequences, *Minds Mach.*, vol. 30, no. 4, pp. 681–694, 2020.

[29] J. Lee, W. J. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, BioBERT: A pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[30] K. X. Huang, J. Altosaar, and R. Ranganath, ClinicalBERT: Modeling clinical notes and predicting hospital readmission, arXiv preprint arXiv: 1904.05342, 2019.

[31] X. Rong, word2vec parameter learning explained, arXiv preprint arXiv: 1411.2738, 2014.
[32] Y. Liu, T. Ge, K. Mathews, H. Ji, and D. McGuinness, Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion, in *Proc. 2015 Workshop on Biomedical Natural Language Processing*, Beijing, China, 2015, pp. 92–97.
[33] T. Mikolov, Q. V. Le, and I. Sutskever, Exploiting similarities among languages for machine translation, arXiv preprint arXiv: 1309.4168, 2013.
[34] P. Kim, Convolutional neural network, in *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence*, P. Kim, ed. Berkeley, CA, USA: Springer, 2017, pp. 121–147.
[35] I. Freeman, L. Roese-Koerner, and A. Kummert, Effnet: An efficient structure for convolutional neural networks, in *Proc. 25th IEEE Int. Conf. on Image Processing*, Athens, Greece, 2018, pp. 6–10.
[36] Y. Kim, Convolutional neural networks for sentence classification, in *Proc. 2014 Conf. on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1746–1751.
[37] L. Breiman, Random forests, *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[38] A. Cutler, D. R. Cutler, and J. R. Stevens, Random forests, in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Q. Ma, eds. Boston, MA, USA: Springer, 2012, pp. 157–176.
[39] E. Shelhamer, J. Long, and T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Trans. Patt. Analy. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.
[40] Y. Doval and C. Gómez-Rodríguez, Comparing neural- and N-gram-based language models for word segmentation, *J. Assoc. Inform. Sci. Technol.*, vol. 70, no. 2, pp. 187–197, 2019.
[41] K. C. Chang and H. T. Chang, Is it possible to use chatbot for the Chinese word segmentation? in *Proc. 3rd Int. Conf. on Natural Language Processing and Information Retrieval*, Tokushima, Japan, 2019, pp. 20–24.
[42] H. Saif, M. Fernández, Y. L. He, and H. Alani, On stopwords, filtering and data sparsity for sentiment analysis of twitter, in *Proc. 9th Int. Language Resources and Evaluation Conf.*, Reykjavik, Iceland, 2014, pp. 810–817.
[43] B. K. Yang, G. Z. Dai, Y. J. Yang, D. R. Tang, Q. Li, D. N. Lin, J. Zheng, and Y. P. Cai, Automatic text classification for label imputation of medical diagnosis notes based on random forest, in *Proc. 7th Int. Conf. on Health Information Science*, Cairns, Australia, 2018, pp. 87–97.
[44] D. Kostrzewa and R. Brzeski, Adjusting parameters of the classifiers in multiclass classification, in *Proc. 13th Int. Conf. on Beyond Databases, Architectures and Structures. Towards Efficient Solutions for Data Analysis and Knowledge Representation*, Ustroń, Poland, 2017, pp. 89–101.
[45] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.
[46] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.

Bokai Yang received the BS degree in automation from Hebei University of Science and Technology, Shijiazhuang, China, in 2015, and the MS degree in biomedical engineering from Northeastern University, Shenyang, China, in 2019. He is currently pursuing the PhD degree in computer application technology at University of Chinese Academy of Sciences, Beijing, China. His research interests include bioinformatics and medical big data mining.

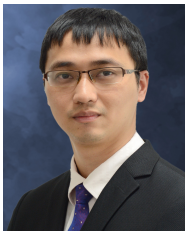
Yujie Yang received the BS degree in information and computing science from Harbin Institute of Technology, Harbin, China, in 2007, the MS degree in applied mathematics from University of Science and Technology of China, Hefei, China, in 2010, and the PhD degree in computer application technology from University of Chinese Academy of Sciences, Beijing, China, in 2021. She is currently an engineer with Shenzhen Institute of Advanced Technology, Shenzhen, China, and also with University of Chinese Academy of Sciences, Beijing, China. Her research interests include bioinformatics and medical big data mining.

Denan Lin received the BS degree in clinical medicine from Jinan University, Guangzhou, China, in 1987, and the MS degree in economic management from Central Party School of the Communist Party of China, Beijing, China, in 2006. He is the director of Shenzhen Health Development Research and Data Management Center, Shenzhen, China, majoring in population health information management.

Ye Li received the BS and MS degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 1999 and 2002, respectively, and the PhD degree from Arizona State University, Tempe, USA, in 2006, all in electrical engineering. Since 2008, he has been the director of the Research Center for Biomedical Information Technology, Shenzhen Institute of Advanced Technology, Shenzhen, China, also he is currently a professor with University of Chinese Academy of Sciences, Beijing, China. His research interests include medical big data, wearable sensors, and mobile health.

Qi Li received the BS degree in communication engineering from Hainan University, Haikou, China, in 2010. He is currently an engineer with Shenzhen Institute of Advanced Technology, Shenzhen, China. His research interests include bioinformatics, health informatics, computational genomics, and virus evolution.

Jing Zheng received the BS degree in health service management from Anhui Medical University, Hefei, China, in 2001, the MS degree in epidemiology and health statistics from University of Sun Yat-sen University, Guangzhou, China, in 2004, and the PhD degree in public administration from Beijing Normal University, Beijing, China, in 2010. He is the deputy director of Shenzhen Health Development Research and Data Management Center, Shenzhen, China, majoring in health information management and medical big data.



Yunpeng Cai received the PhD degree in computer science and technology from Tsinghua University, Beijing, China, in 2007. He is currently a professor at Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His research interests include health big data, health informatics, bioinformatics, and machine learning.