TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 07/16 pp686–695
DOI: 10.26599/TST.2022.9010049
Volume 28, Number 4, August 2023

Fusion Model for Tentative Diagnosis Inference Based on Clinical Narratives

Ying Yu, Junwen Duan*, and Min Li

Abstract: In general, physicians make a preliminary diagnosis based on patients' admission narratives and admission conditions, largely depending on their experiences and professional knowledge. An automatic and accurate tentative diagnosis based on clinical narratives would be of great importance to physicians, particularly in the shortage of medical resources. Despite its great value, little work has been conducted on this diagnosis method. Thus, in this study, we propose a fusion model that integrates the semantic and symptom features contained in the clinical text. The semantic features of the input text are initially captured by an attention-based Bidirectional Long Short-Term Memory (BiLSTM) network. The symptom concepts, recognized from the input text, are then vectorized by using the term frequency-inverse document frequency method based on the relations between symptoms and diseases. Finally, two fusion strategies are utilized to recommend the most potential candidate for the international classification of diseases code. Model training and evaluation are performed on a public clinical dataset. The results show that both fusion strategies achieved a promising performance, in which the best performance obtained a top-3 accuracy of 0.7412.

Key words: tentative diagnosis; clinical narrative; Bidirectional Long Short-Term Memory (BiLSTM); Term Frequency-Inverse Document Frequency (TF-IDF); fusion strategy

1 Introduction

Tentative diagnosis[1] is a preliminary inference of patient diseases. It is based on the recognition of certain symptoms that may indicate disease presence. Accordingly, it relies heavily on the experiences and professional knowledge of physicians, making it at risk of misdiagnosis or missed diagnosis.

An automatic tentative diagnosis approach based on the understanding of clinical narratives could be a good assistant for physicians, particularly in the shortage of medical resources. Furthermore, it can guide patients for further examinations or treatments. Although clinical data analysis and application based on Electronic Medical Records (EMRs) have become focused issues and trends in clinical research[2], few studies have been conducted on this issue.

Considering the data-driven approaches used for secondary analysis of EMRs[3–6], we explored an automatic tentative diagnosis method based on EMRs. As a significant component of EMRs, clinical note records comprehensive information about patient conditions, including the narratives about the present illness based on patients' statements at admission. It can be used for tentative diagnosis inference as the input text. Capturing features that are related to diseases from clinical text is the main challenge of this method. On

• Ying Yu is with the School of Computer Science and Engineering, Central South University, Changsha 410083, China, and also with the School of Computer Science, University of South China, Hengyang 421001, China. E-mail: yuying@ mail.csu.edu.cn.
• Junwen Duan and Min Li are with Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China, E-mail: jwduan@csu.edu.cn; limin@mail.csu.edu.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2022-08-08; revised: 2022-09-04; accepted: 2022-10-10

the one hand, the clinical note has a long text, and it involves many medical terminologies and abbreviations. It is not trivial for semantic feature extraction. On the other hand, the onset of certain symptoms mentioned in the clinical note may have a direct relationship with some diseases. Another challenge is the identification of useful symptoms and their representation.

In addressing the aforementioned challenges, we proposed a fusion model that integrates semantic features and symptom features of the clinical text. The contextual semantic feature is captured by an attention-based Bidirectional Long Short-Term Memory (BiLSTM)[7] network. The feature of the symptom concept, extracted by MetaMap[8], is represented into a weight vector by the Term Frequency-Inverse Document Frequency (TF-IDF)[9] algorithm. A preliminary attempt has been explored in previous work[10]. Two fusion strategies are applied in the model to efficiently utilize two features: the concatenation of feature vectors before classification and the combination of prediction results after classification. We evaluated our proposed approach on a public database, namely, Medical Information Mart for Intensive Care (MIMIC) III. The experimental results proved the advantages of two fusion strategies, which achieve an average accuracy of 0.5072 and 0.5114 in the automatic tentative diagnosis of 114 diseases. Moreover, an average top-3 accuracy of 0.7412 indicates that our model is helpful for clinical decision-making.

2 Related Work

Recently, clinical texts have been used for a series of medical research, such as medical named entity recognition and relation extraction[11–14], phenotype classification[15], International Classification of Diseases (ICD) code assignment[16–21], and de-identification[22, 23]. Deep learning models, such as Recurrent Neural Networks (RNN)[16, 21, 24] and convolutional neural networks[18, 19], have been widely used to learn the representation of clinical texts without manual feature engineering[18–20], and such models have achieved state-of-the-art performances in most tasks. However, most studies of disease classification based on clinical text only focused on encoding text sequence[16–21], that is, capturing the semantic features by neural networks. Those studies overlook specific medical entities and their relations with diseases, which we believe could capture more disease-related features from clinical text.

Tentative diagnosis is an essential manual work based on the rich experience and knowledge of physicians. An automatic tentative diagnosis model based on clinical narratives may be used to make preliminary decisions. To the best of our knowledge, no previous work has systematically studied this issue. Thus, in this paper, we fill this gap by proposing a fusion model. Except for the semantic feature that is captured by neural networks, we tried to integrate symptom features into the classification model. We extracted symptom concepts from the clinical text and represented them based on their relations with diseases.

3 Material and Method

3.1 Dataset

This work used an open-access dataset MIMIC III[25], which is a public, freely-available critical care database developed by the MIT Laboratory for Computational Physiology. It comprised health-related data of patients who stayed in Beth Israel Deaconess Medical Center between 2001 and 2012[26]. Over 46 000 patients are associated with 58 976 admission records in the latest version of MIMIC III v1.4[27]. Each admission record is tagged by several ICD version 9 (ICD-9) codes. A discharge summary records the whole information of a patient's hospitalizations in text, from admission to discharge. A sample of the discharge summary is illustrated in Fig. 1. The protected health information in the data, such as patient name, telephone number, address, and dates, has been de-identified in accordance with the Health Insurance Portability and Accountability Act (HIPAA) standards[27]. In addition, the text had many abbreviations, with formal medical style, e.g., obstructive sleep apnea (OSA) or informal style, e.g., years old (yo).

The main body of the text is composed of dozens of sections, such as chief complaint, major surgical or invasive procedure, and history of present illness.

3.2 Preprocessing

We extracted "chief complaint, history of present illness, and past medical history" as the input text. All three parts are primarily recorded on the basis of patient narratives about the onset, development, and present symptoms of his/her illness right at admission. The ICD codes for each admission record consisted of one primary diagnosis code and several secondary diagnosis codes. We selected the primary diagnosis code as the label, which is marked with "1" in the sequence number. For normalization, we used the first three numbers of the ICD-9 code to indicate the category of disease.

Admission Date: [**2174-2-12**]
Discharge Date: [**2174-2-14**]
Date of Birth: [**2122-4-28**]
Sex: M
Service: MEDICINE
Allergies: Patient recorded as having No Known Allergies to Drugs
Attending:[**First Name3 (LF) 2474**]
Chief Complaint: shortness of breath
Major Surgical or Invasive Procedure: None

History of Present Illness:
51 yo M with h/o asthma and right lung volume loss of unclear etiology (?congenital hypoplasia), recurrent bronchitis in winter, OSA, obesity, HTN who presented with 1 week of productive cough, progresssive SOB, and over the past 2 days weakness and fatigue to the point he was falling asleep at work. He tried increasing his albuterol use but this did not help so he came to the emergency room.
In the ED, initial vs were: T: 99.5 P109 BP101/56 R24 79% on RA on presentation, …

Past Medical History:
Asthma
Hypertension
…
Social History:
Married, no children. He works for [**Company 2475**]. Previously he worked in a printing company where he reports that he was exposed to fumes and did not wear a mask.
Tobacco: Quit.
…
Family History:
Grandmother with diabetes. Father with Alzheimer's.

Physical Exam:
General: Alert, oriented, no acute distress
HEENT: Sclera anicteric, MMM, oropharynx clear
Neck: supple, JVP not elevated, no LAD
Lungs: Decreased breath sounds at R base, no wheezes, rales, rhonchi
….

Brief Hospital Course:
In the ED, initial vs were: T: 99.5 P109 BP101/56 R24 79% on RA on presentation, …

Fig. 1 Example of a discharge summary. A discharge summary consists of several sections. The content in "[...**]" is privacy information, which has been de-identified in accordance with HIPAA standards.**

With regard to imbalance, we removed the samples with infrequent ICD codes (less than 50) or with less than three symptoms. The final dataset involved 31 213 notes correlating with 114 ICD-9 codes. Statistical analysis of the input texts is listed in Table 1. The "chief complaint" section of the discharge summary is the shortest part, with 3 words on average, whereas the "history of present illness" section is the longest part, with 200 words on average.

3.3 Symptom concept extraction

The first step of symptom feature representation is the identification of the symptom concept mentioned in clinical narratives. However, it is not suitable for simple text searches because of formal or informal medical abbreviations in clinical text. We used a widely available tool MetaMap to identify the symptom concepts in the narratives. It can provide access to standardized concepts in the unified medical language system metathesaurus from English biomedical text[8]. Biomedical concepts can be extracted automatically on the basis of the user's configuration of three options, i.e., data, output, and processing[28]. Most medical abbreviations that are often used in the clinical text can be recognized by MetaMap, such as Chronic Obstructive Pulmonary Disease (COPD), ArterioVenous Malformation (AVM), and Congestive Heart Failure (CHF). A Concept Unique Identification (CUI) is the identification code for each concept. A total of 127 semantic types in MetaMap 2016v2 can be selected to filter out symptom concepts. We selected 13 semantic types in accordance with Sondhi et al.'s work[29] of constructing SympGraph.

Table 1 Overview of input texts.

Text section	Length of text (number of words)		
	Average	Maximum	Minimum
Chief complaint	3	609	1
History of present illness	200	1497	1
Past medical history	48	1188	1
All of three sections	253	1656	11

In addition, we only focused on positive symptom concepts found in the text. The negative symptom concepts, that are found in negated contexts, such as the presence of negation-related words, including "denies", "without", and "no", are filtered out. Some symptom concepts may appear several times because of being mentioned in different parts of an input text. Such duplicated symptoms are all kept in our extraction.

Furthermore, we focused on the appearance of symptom concepts instead of the concept term itself to represent a symptom concept by analyzing the relations between symptoms and diseases. Thus, the CUI, not the name, of the symptom was used to indicate

a symptom. A total of 9386 symptom concepts are extracted by MetaMap from all the input texts. Each sample contained 20 symptoms on average and up to 113 symptoms at most. The distribution of symptom frequency and ICD code frequency are shown in Figs. 2a and 2b, respectively. As shown in Fig. 2a, the symptom concepts occurred from one time to tens of thousands of times. Over half of the extracted symptoms occurred less than five times. In particular, nearly one-third of symptoms occurred only once. The most frequent symptom, "hypertensive disease (CUI: C0020538)", appeared 26 330 times. As shown in Fig. 2b, almost 15% of ICD codes occurred over 500 times. Half of the ICD-9 codes appeared in 100–500 samples.

3.4 Feature representation

The task is text-based multi-classification. We extracted semantic and symptom features from the input text to obtain comprehensive features for disease prediction. The model primarily consists of three parts: semantic feature representation, symptom feature representation, and feature fusion. The architecture of our model is shown in Fig. 3. Each part of the model is illustrated in detail in the following subsections.

3.4.1 Semantic feature representation

The semantic feature representation aims to learn a contextual semantic representation of the input text. We

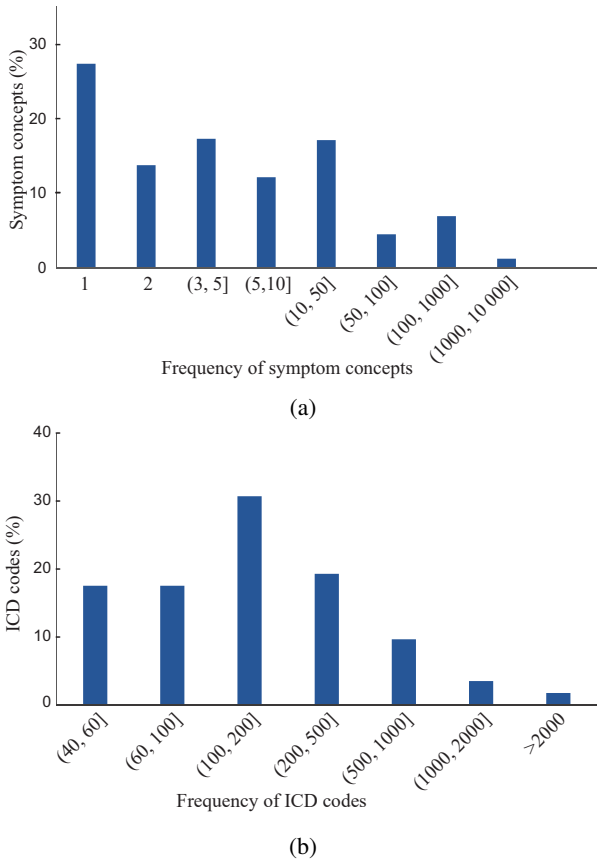


Fig. 2 Distribution of extracted symptoms and ICD codes: (a) distribution of symptoms and (b) distribution of ICD codes.

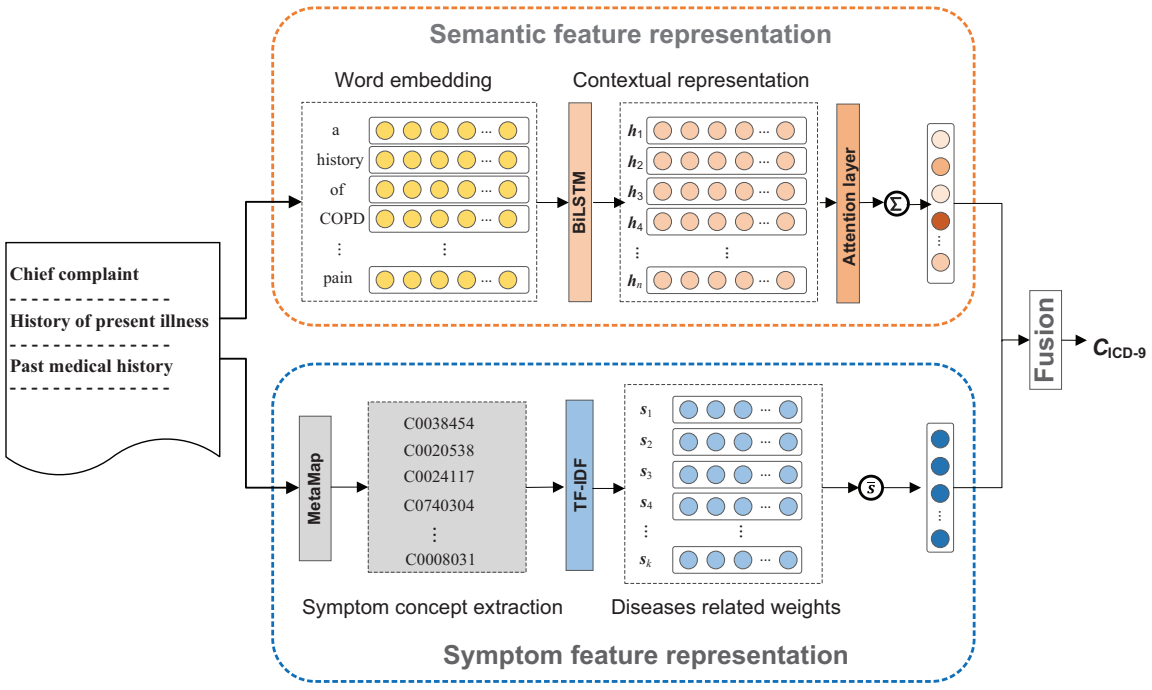


Fig. 3 Architecture of our model. The upper part in the orange rounded rectangular box indicates semantic feature extraction, whereas the bottom part in the blue rounded rectangular box indicates symptom feature extraction. $C_{\text{ICD-9}}$ indicates ICD-9 code.

used the Long Short-Term Memory (LSTM) network[30] as the text encoder, which is a variant of RNN and is competitive in feature extraction from sequential data[31]. Furthermore, we adopted the bidirectional version, that is, BiLSTM, which can capture information from the forward and backward directions of the sequences. Given the noise in the narratives, we adopted the attention mechanism, which has been used successfully in many deep learning models[32–36], to focus on the most important information in sequence.

Here, a word-level attention-based BiLSTM network is applied for semantic feature extraction. As shown in Fig. 3, each word of the input text is embedded into a vector by word2vec with a dimension of 128. Then, the embedding sequences are input into the BiLSTM network. The output of the hidden layer is the concatenation of both directional hidden states for each LSTM element, that is, $\overleftrightarrow{\boldsymbol{h}}_i = [\overrightarrow{\boldsymbol{h}}_i; \overleftarrow{\boldsymbol{h}}_i]$, $i = 1, 2, \ldots, n$. A context attention mechanism is used to weigh the semantic feature after BiLSTM, and focus on informative words of the input texts. The final semantic feature vector of an input text is calculated by several steps at the attention layer. In particular, the following formulas are used:

$$\boldsymbol{u}_i = \tanh\left(\boldsymbol{W}\overleftrightarrow{\boldsymbol{h}}_i + \boldsymbol{b}\right) \tag{1}$$

$$\alpha_i = \frac{\exp\left(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_c\right)}{\sum_i \exp\left(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_c\right)} \tag{2}$$

$$\boldsymbol{v} = \sum_i \alpha_i \overleftrightarrow{\boldsymbol{h}}_i \tag{3}$$

A one-hidden-layer multi-layer perceptron with weight matrix $\boldsymbol{W}$ is used to turn the output vector $\overleftrightarrow{\boldsymbol{h}}_i$ into a hidden representation vector (denoted as $\boldsymbol{u}_i$), which is produced by the tanh( ) activation function. The weight for each word representation vector (denoted as α_i) is calculated by the softmax function with $\boldsymbol{u}_i$. A context vector $\boldsymbol{u}_c$ is randomly initialized and jointly learned during training. The final contextual semantic feature vector $\boldsymbol{v}$ is the weighted sum of $\overleftrightarrow{\boldsymbol{h}}_i$.

3.4.2 Symptom feature representation

In representing symptoms based on the relationship between symptoms and diseases, a TF-IDF weighting scheme is used to measure the occurrences of symptoms in a text. TF-IDF is a classic statistic algorithm proposed for information retrieval[37], which aims to reflect the importance of a word to a document in a corpus. It has been used for symptom-based human disease network construction[38] and symptom representation[10]. Here, each symptom can be regarded as a word, whereas the disease can be regarded as a document. TF-IDF is used to determine the importance of a symptom to a disease. Symptom $\boldsymbol{s}_i$ is represented by a vector $(w_{i1}, w_{i2}, \ldots, w_{iK})$, consisting of its TF-IDF weight values for 114 disease codes in our label set. w_{ij} indicates the strength of the correlation between symptom i and disease j. It is quantified by using the following formula:

$$w_{ij} = TF_{ij} \times \log \frac{|K|}{|D_i|} \tag{4}$$

where K is the number of all diseases mentioned in the dataset, and D_i denotes the number of diseases which is associated with symptom i; that is, symptom i occurs in the clinical texts labeled with those disease codes. TF_{ij} denotes the number of correlations between symptom i and disease j in the text. A simple neural network of Continuous-Bag-Of-Words (CBOW) is used to obtain the symptom feature vector of the input text. The mean value of symptom representation vectors is the final representation of symptom features.

3.5 Fusion strategy

Two fusion strategies are explored to take advantage of two features extracted from clinical narratives. The first strategy is feature fusion. The semantic feature vector and symptom feature vector are concatenated into one feature vector as the input of the classifier (Fig. 4a). The other strategy is decision fusion. Two models were trained on the basis of two different features. Each model produces a result vector of the probabilities for all labels by its classifier, which is a fully connected network with the softmax function. As shown in Fig. 4b, the max value of the proportional sum of two probability result vectors indicates the most likely outcome. In particular, the following formulas are used:

$$p_i' = \beta \times p_i^t + (1 - \beta) \times p_i^s \tag{5}$$

For sample i, p_i^t is the probability vector produced by the neural network based on semantic feature representation; p_i^s is the probability vector produced by the neural network based on symptom feature representation; β indicates the proportion of p_i^t in the sum vector of p_i'.

4 Experiment and Result

4.1 Metrics and training

We used accuracy to evaluate the performance of models. The Area Under the Precision-Recall curve (AUPR) is another metric. Experiments were performed on a server with an NVIDIA GeForce Titan X Pascal CUDA GPU

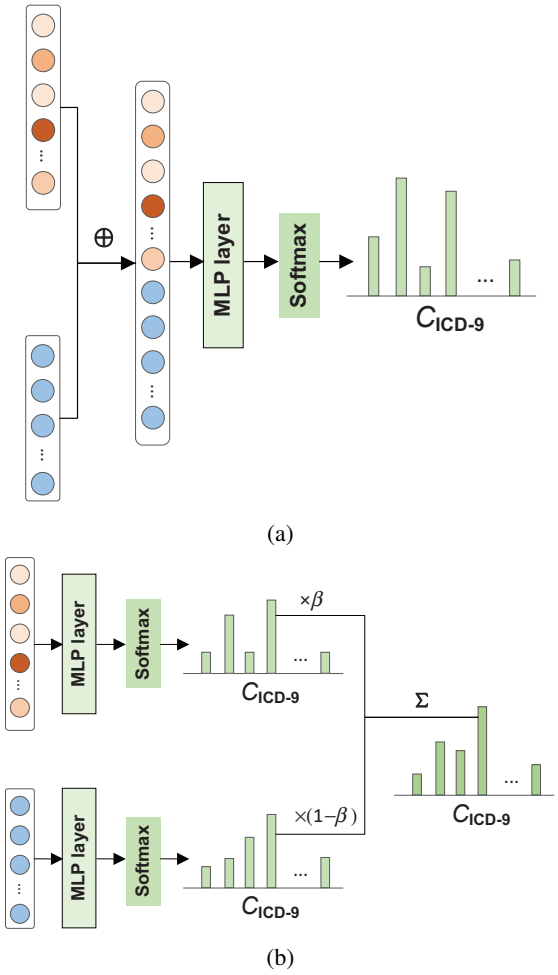


Fig. 4 Two fusion strategies. (a) shows the feature fusion strategy (Strategy 1); the annotation "⊕" indicates the concatenation operation, and (b) shows the decision fusion strategy (Strategy 2).

processor. The dataset was split into the training set and testing set at the ratio of 9:1, consisting of 28 091 and 3122 records, respectively. The testing set remained unused until the final evaluation. We ran five times (t1−t5) with the same hyper-parameters and recorded the results on the testing set. The training set was shuffled each time.

4.2 Comparison of two fusion strategies

The results of five tests are shown in Table 2. The model based on the semantic feature shows over 10% more advantages in average accuracy than the model based on the symptom concept feature. Although the model based on only symptom features does not achieve a remarkable results, the good accuracy and AUPR achieved by both fusion strategies indicate the effective supplementary role of the symptom feature, compared with the results of the model based on either single feature. Therefore, the fusion model has advantages for this task. Based on the standard deviation of results, the improvement of Strategy 2 seems to be more stable than that of Strategy 1. In the case of decision support applications, we also reported the top-3 accuracy (denoted as Acc.@3). Therefore, three candidate codes (out of 114 codes) with the highest scores are presented for review. The average Acc.@3 is approximately 0.74, which is inspiring and auxiliary for clinical decision-making.

4.3 Hyper-parameter tuning and training

The hyper-parameters of our model are searched for the best performance. The learning and dropout rates of BiLSTM are 0.005 and 0.300, respectively. The value of β (0.55) in decision fusion strategy is searched for the best performance. Figure 5 shows the performance at different values of β in five tests. Notably, the optimal value of β for each test is concentrated in an interval between 0.55 and 0.65. The results of the model based on the semantic feature do not account for a great proportion during fusion, although the performance of the semantic-based model is better than that of the symptom-based model. Therefore, the symptom feature plays a role in fusion.

4.4 Contribution of different parts of input text

The input text consists of three parts; however, the part of the model with the most important contextual

Table 2 Performance of five tests.

Test	Semantic feature only			Symptom feature only			Strategy 1			Strategy 2 (β=0.55)		
	Acc.	AUPR	Acc.@3	Acc.	AUPR	Acc.@3	Acc.	AUPR	Acc.@3	Acc.	AUPR	Acc.@3
t1	0.498	0.395	0.727	0.390	0.272	0.633	0.506	0.413	0.731	0.509	0.406	0.741
t2	0.497	0.377	0.715	0.394	0.274	0.632	0.514	0.399	0.723	0.511	0.399	0.739
t3	0.499	0.384	0.721	0.393	0.273	0.635	0.501	0.395	0.741	0.514	0.401	0.740
t4	0.498	0.393	0.734	0.393	0.275	0.630	0.508	0.401	0.734	0.512	0.406	0.744
t5	0.501	0.388	0.728	0.395	0.273	0.631	0.507	0.400	0.740	0.511	0.409	0.742
Average	0.4986	0.3874	0.7250	0.3930	0.2734	0.6322	0.5072	0.4016	0.7338	**0.5114**	**0.4042**	**0.7412**
Standard deviation ($\times 10^{-3}$)	1.36	6.47	6.48	1.67	1.02	1.72	4.17	6.05	6.55	1.62	3.66	1.72

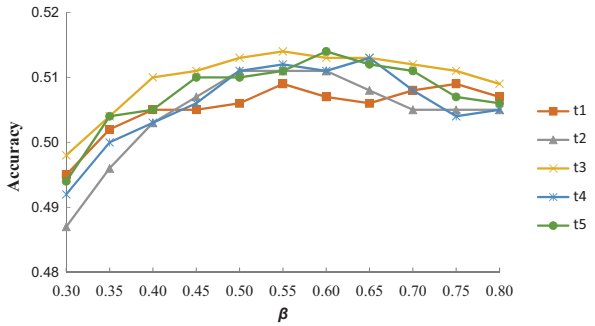


Fig. 5 Combination performance at different values of β.

semantic features remains unknown. The experiments based on each text section were performed, and the average accuracy is listed in Table 3. The section "history of present illness" shows its dominant position in the three parts, whose accuracy value is only 4% lower than the value based on all three sections. The average length of this section is the longest among the three parts. High accuracy indicates that the section "history of present illness" includes informative text, and indicates diseases inference at present admission. Although the "chief complaint" is the shortest section, it may involve keywords or symptom terms related to diseases. Based on the input text of two parts, namely, "chief complaint" and "history of present illness", a 2% accuracy promotion is achieved. Although the "past medical history" section is long, it shows the worst performance in results. However, it seems helpful for our task, considering the best result of all three parts.

5 Discussion

5.1 Analysis of combined feature representation

We projected the feature vector into a 2D space by using the t-SNE method to qualitatively and evidently assess the combined feature representation in Strategy 1. The scatterplots of three feature representations of those samples related to top-5 and top-3 diseases in the testing set are shown in Fig. 6. The scatterplot of symptom feature representation shows an unclear clustering margin compared with that of semantic feature representation.

Table 3 Prediction performance of different sections of patient narratives.

Text section	Average accuracy
Chief complaint	0.3390
History of present illness	0.4598
Past medical history	0.2224
Chief complaint and history of present illness	0.4795
All of three sections	0.4986

However, the combined feature representation is more evident than the semantic feature representation, which indicates the advantage of Strategy 1.

5.2 Contribution of symptom feature

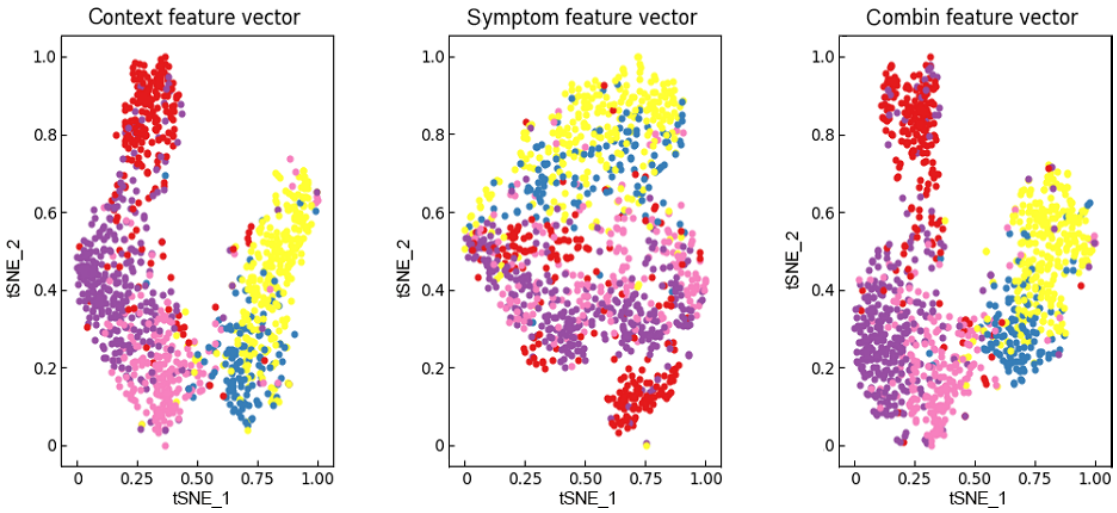
Based on the performance of the single feature-based model, the symptom-feature-based model is not as good as the semantic-feature-based model. In revealing the contribution of the symptom feature, we analyzed the results of the decision fusion strategy by observing the number of true-positive samples in the union of two classifier results. Such results were compared with those of the single classifier and their intersection. As shown in Table 4, 924 true-positive samples, on average, were found in the intersection set. However, over 1800 true-positive samples were found in the union set. Based on the union results, around 300 true-positive samples were increased, compared with the results of the model based on the semantic feature. Those labels undoubtedly are replenished by the model based on symptom feature. Therefore, the semantic feature and symptom feature are complementary.

6 Conclusion

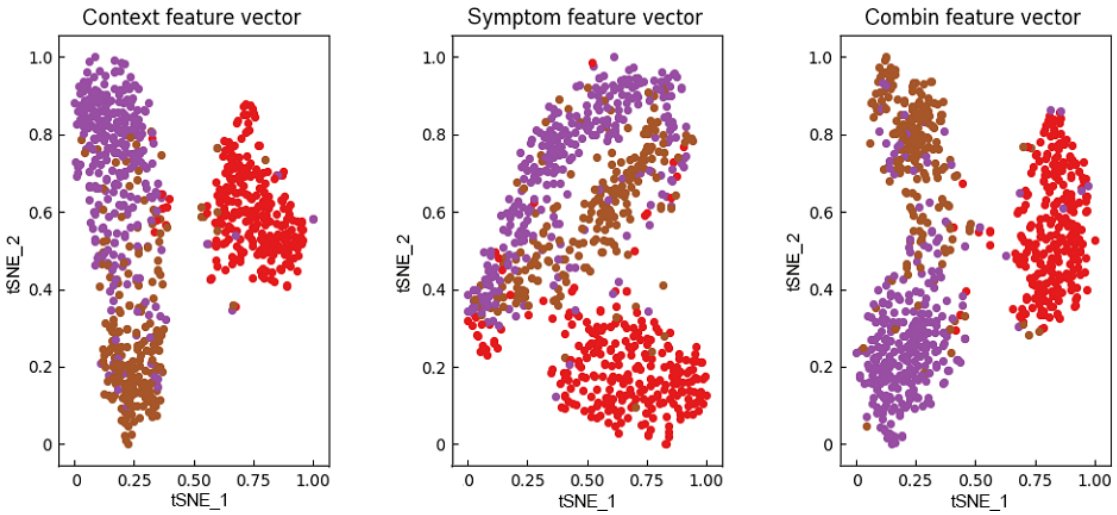
Tentative diagnosis prediction based on clinical narratives is a significant and challenging task. In capturing disease-related features from the input texts, we proposed a fusion model based on two kinds of features: the semantic feature and the symptom feature. Feature-level fusion and decision-level fusion are utilized to predict the candidate ICD code. The two fusion models achieved better experimental results on the MIMIC III dataset, which indicate that the fusion of two kinds of features enriches the representation of clinical text. Result analysis of the decision fusion strategy proves that the representation of symptoms is a useful supplement for semantic features. However, our work also has some limitations. A better representation method for symptoms is needed when some diseases are correlated with the same symptoms. In addition, the records of MIMIC III are about inpatients, particularly patients in intensive care units. For more general cases, the exploration of outpatients' records is our future direction.

Acknowledgment

We thank the anonymous reviewers for their helpful comments. This work was supported in part by the Science and Technology Major Project of Changsha (No.



(a) Scatterplots of the final representations of the samples related to the top-5 diseases in the testing set



(b) Scatterplots of the final representations of the samples related to top-3 diseases in the testing set

Fig. 6 t-SNE scatterplots of three final representations. In Figs. 6a and 6b, the left chart is the scatterplot of contextual semantic feature representation, the middle chart is the scatterplot of symptom feature representation, and the right chart is the scatterplot of combined feature representation.

Table 4 True positive sample comparison of Strategy 2.

Model	t1	t2	t3	t4	t5	Average
Semantic feature only	1556	1553	1559	1556	1563	1557
Symptom feature only	1218	1229	1226	1226	1232	1226
Intersection of results	922	925	919	929	917	924
Union of results	1852	1857	1870	1853	1878	1862

kh2202004) and the National Natural Science Foundation of China (No. 62006251). We are grateful for resources from the High-Performance Computing Center of Central South University.

References

[1] I. Boas, Early and tentative diagnosis of gastrointestinal carcinoma, *Am. J. Cancer*, vol. 15, no. 3, pp. 1586–1589, 1931.

[2] Y. Yu, M. Li, L. Liu, Y. Li, and J. Wang, Clinical big data and deep learning: Applications, challenges, and future outlooks, *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 288–305, 2019.

[3] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism, in *Proc. 30th Int. Conf. on Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 3512–3520.

[4] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks, in *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 1903–1911.

[5] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, Risk prediction on electronic health records with prior medical knowledge, in *Proc. 24th ACM SIGKDD Int. Conf. on*

Knowledge Discovery & Data, London, UK, 2018, pp. 1910–1919.

[6] 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, *Nat. Med.*, vol. 25, no. 3, pp. 433–438, 2019.

[7] A. Graves and J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Netw.*, vol. 18, nos. 5&6, pp. 602–610, 2005.

[8] A. R. Aronson, Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program, in *Proc. AMIA 2001*, Washington, DC, USA, 2001, p. 17.

[9] G. Salton, A. Wong, and C. S. Yang, A vector space model for automatic indexing, *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[10] Y. Yu, M. Li, L. Liu, F. X. Wu, and J. Wang, Tentative diagnosis prediction via deep understanding of patient narratives, in *Proc. 2019 IEEE Int. Conf. on Bioinformatics and Biomedicine* (*BIBM*), San Diego, CA, USA, 2019, pp. 1000–1003.

[11] A. N. Jagannatha and H. Yu, Bidirectional RNN for medical event detection in electronic health records, in *Proc. 2016 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA, 2016, pp. 473–482.

[12] Y. Luo, Recurrent neural networks for classifying relations in clinical notes, *J. Biomed. Inform.*, vol. 72, pp. 85–95, 2017.

[13] S. Gao, M. T. Young, J. X. Qiu, H. J. Yoon, J. B. Christian, P. A. Fearn, G. D. Tourassi, and A. Ramanthan, Hierarchical attention networks for information extraction from cancer pathology reports, *J. Am. Med. Inform. Assoc.*, vol. 25, no. 3, pp. 321–330, 2018.

[14] L. Gligic, A. Kormilitzin, P. Goldberg, and A. Nevado-Holgado, Named entity recognition in electronic health records using transfer learning bootstrapped neural networks, *Neural Netw.*, vol. 121, pp. 132–139, 2020.

[15] S. Gehrmann, F. Dernoncourt, Y. Li, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Jr. Foote, E. T. Moseley, D. W. Grant, et al., Comparing rule-based and deep learning models for patient phenotyping, arXiv preprint arXiv: 1703.08705, 2017.

[16] H. Shi, P. Xie, Z. Hu, M. Zhang, and E. P. Xing, Towards automated ICD coding using deep learning, arXiv preprint arXiv: 1711.04075, 2017.

[17] W. Ning, M. Yu, and R. Zhang, A hierarchical method to automatically encode Chinese diagnoses through semantic similarity estimation, *BMC Med. Inform. Decis. Mak.*, vol. 16, p. 30, 2016.

[18] T. Baumel, J. Nassour-Kassis, R. Cohen, M. Elhadad, and N. Elhadad, Multi-label classification of patient notes a case study on ICD code assignment, arXiv preprint arXiv: 1709.09587, 2017.

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

[20] M. Zeng, M. Li, Z. Fei, Y. Yu, Y. Pan, and J. Wang, Automatic ICD-9 coding via deep transfer learning, *Neurocomputing*, vol. 324, pp. 43–50, 2019.

[21] Y. Wu, M. Zeng, Z. Fei, Y. Yu, F. X. Wu, and M. Li, KAICD: A knowledge attention-based deep learning framework for automatic ICD coding, *Neurocomputing*, vol. 469, pp. 376–383, 2022.

[22] Z. Liu, B. Tang, X. Wang, and Q. Chen, De-identification of clinical notes via recurrent neural network and conditional random field, *J. Biomed. Inform.*, vol. 75, pp. S34–S42, 2017.

[23] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, De-identification of patient notes with recurrent neural networks, *J. Am. Med. Inform. Assoc.*, vol. 24, no. 3, pp. 596–606, 2017.

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

[25] G. B. Moody and R. G. Mark, A database to support development and evaluation of intelligent intensive care monitoring, in *Proc. of Computers in Cardiology 1996*, Indianapolis, IN, USA, 2002, pp. 657–660.

[26] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L. W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database, *Crit. Care Med.*, vol. 39, no. 5, pp. 952–960, 2011.

[27] 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, *Sci. Data*, vol. 3, p. 160035, 2016.

[28] A. R. Aronson and F. M. Lang, An overview of MetaMap: Historical perspective and recent advances, *J. Am. Med. Inform. Assoc.*, vol. 17, no. 3, pp. 229–236, 2010.

[29] P. Sondhi, J. Sun, H. Tong, and C. Zhai, SympGraph: A framework for mining clinical notes through symptom relation graphs, in *Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 1167–1175.

[30] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] A. Graves and N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, in *Proc. 31st Int. Conf. on Int. Conf. on Machine Learning*, Beijing, China, 2014, pp. 1764–1772.

[32] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, Attention-based models for speech recognition, in *Proc. 28th Int Conf on Neural Information Processing Systems*, Montreal, Canada, 2015, pp. 577–585.

[33] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, Image captioning with semantic attention, in *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 4651–4659.

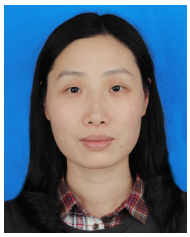
[34] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in *Proc. 54th Annu. Meeting of the Association for Computational Linguistics* (*Volume 2: Short Papers*), Berlin, Germany, 2016, pp. 207–212.

[35] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition, *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388 2017.

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

[37] K. S. Jones, A statistical interpretation of term specificity and its application in retrieval, *J. Doc.*, vol. 28, no. 1, pp. 11–21, 1972.

[38] X. Zhou, J. Menche, A. L. Barabási, and A. Sharma, Human symptoms-disease network, *Nat. Commun.*, vol. 5, p. 4212, 2014.



Ying Yu received the BEng and MEng degrees in computer science from the University of South, China in 2002 and 2009, respectively. She is currently a PhD candidate at the School of Computer Science and Engineering, Central South University, China. Her current research focuses on the analysis of clinical texts and electronic health records by using machine learning or deep learning methods.

Junwen Duan received the BEng degree in information security and the PhD degree in computer science from Harbin Institute of Technology, Harbin, China in 2013 and 2020, respectively. He is currently a lecturer at the School of Computer Science and Engineering, Central South University. His main research interests include natural language processing and text mining.

Min Li received the PhD degree in computer science from Central South University, China in 2008. At present, she is a professor and the vice dean at the School of Computer Science and Engineering, Central South University, China. Her main research interests include bioinformatics and systems biology. She has published more than 100 technical papers in refereed journals, such as *Genome Biology*, *Genome Research*, *Bioinformatics*, *Briefings in Bioinformatics*, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, and conference proceedings, such as BIBM, GIW, and ISBRA. Based on Google scholar, her paper citations are more than 8000, and the h-index is 48.