Applied Intelligence (2022) 52:13114–13131
https://doi.org/10.1007/s10489-022-03222-y

Bi-level artificial intelligence model for risk classification of acute respiratory diseases based on Chinese clinical data

Jiewu Leng[1,2] · Dewen Wang[1] · Xin Ma[1] · Pengjiu Yu[3] · Li Wei[3] · Wenge Chen[1]

Accepted: 6 January 2022 / Published online: 22 February 2022

Abstract

Objective: The high incidence of respiratory diseases has dramatically increased the medical burden under the COVID-19 pandemic in the year 2020. It is of considerable significance to utilize a new generation of information technology to improve the artificial intelligence level of respiratory disease diagnosis. Methods: Based on the semi-structured data of Chinese Electronic Medical Records (CEMRs) from the China Hospital Pharmacovigilance System, this paper proposed a bi-level artificial intelligence model for the risk classification of acute respiratory diseases. It includes two levels. The first level is a dedicated design of the "BiLSTM+Dilated Convolution+3D Attention+CRF" deep learning model that is used for Chinese Clinical Named Entity Recognition (CCNER) to extract valuable information from the unstructured data in the CEMRs. Incorporating the transfer learning and semi-supervised learning technique into the proposed deep learning model achieves higher accuracy and efficiency in the CCNER task than the popular "Bert+BiLSTM+CRF" approach. Combining the extracted entity data with other structured data in the CEMRs, the second level is a customized XGBoost to realize the risk classification of acute respiratory diseases. Results: The empirical study shows that the proposed model could provide practical technical support for improving diagnostic accuracy. Conclusion: Our study provides a proof-of-concept for implementing a hybrid artificial intelligence-based system as a tool to aid clinicians in tackling CEMR data and enhancing the diagnostic evaluation under diagnostic uncertainty.

Keywords Acute respiratory diseases · Risk classification · Deep learning · Chinese clinical named entity recognition · Artificial intelligence

1 Introduction

With the increasing number of the aging population and pollution changes in the external environment, respiratory diseases have significantly increased in terms of morbidity, disability rate, and mortality. Respiratory disease is defined as any of the diseases and disorders of the airways and the lungs that affect human respiration. It may affect any of the structures and organs that have to do with breathing, including the nasal cavities, the pharynx (or throat), the larynx, the trachea (or windpipe), the bronchi and bronchioles, the tissues of the lungs, and the respiratory muscles of the chest cage (www.britannica.com/science/respiratory-disease). In the year 2020, acute respiratory diseases have become the focus of attention under the COVID-19 pandemic. It is found that predictive diagnosis of severe asthma, severe pneumonia, and lung cancer are critical problems in the clinical treatment of patients with acute respiratory diseases [1].

✉ Pengjiu Yu
ypj725@163.com

✉ Li Wei
runkingone@126.com

Jiewu Leng
jwleng@gdut.edu.cn

[1] State Key Laboratory of Precision Electronic Manufacturing Technology and Equipment, Guangdong University of Technology, Guangzhou 510006, Guangdong, China

[2] Department of Information Systems, Chengdu Research Institute, City University of Hong Kong, Hong Kong, China

[3] Department of Pharmacy, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510120, China

However, early warning of acute respiratory disease is still challenging because of three facts. Firstly, although the accumulated medical data that contains vast information about the patient's admission to the hospital has been increasing exponentially every year in the information age, it is still difficult to securely integrate the separated large-scale clinical data about acute respiratory diseases from the distributed database. Secondly, the medical data (e.g., the long-term/temporary prescriptions, diagnostic information, nursing records, examinations, surgical records, and transfer records) is stored in different forms, namely, unstructured text, semi-structured multi-media, and structured Entity-Relation (E-R) database [2]. Using a single analysis technique cannot identify complex factors and will lead to inaccurate classifications. Thirdly, the diagnosis of severe patients is usually performed based on the experience, medical knowledge, and skills of clinicians, which will lead to a significant difference in the speed and accuracy of treatment. Given the characteristics of acute respiratory diseases, its clinical medication is complex, and its treatment cost is prohibitive. Simultaneously, it is difficult for hospitals to judge the relationship between the comprehensive treatment of drugs and clinical outcomes. At present, effective risk classification of acute respiratory diseases is in a significant absence.

Advanced artificial intelligence models, including deep learning algorithms, are gradually introduced for clinical data mining and classification of severe diseases [3]. Based on a large number of real Chinese Electronic Medical Records (CEMRs) from the China Hospital Pharmacovigilance System, this study establishes a bi-level artificial intelligence model for an early warning platform of acute respiratory diseases. Firstly, a dedicated design of the "BiLSTM+Dilated Convolution+3D Attention+CRF" deep learning model is established to realize Chinese clinical named entity recognition from the unstructured data in CEMRs. This module achieves high accuracy through a mixed-use of bi-directional Long Short-Term Memory neural network, dilated convolution, self-attention, conditional random field, transfer learning, word vector modeling, character vector modeling, and semi-supervised learning. The proposed module reduces the workload of manual data annotation and lays the foundation for practical data mining of acute respiratory disease. Secondly, combining the extracted entity data with other structured data in the CEMRs, a comprehensive risk classification module of acute respiratory disease is established based on a customized XGBoost algorithm. This module could identify the pathogenic factors of acute respiratory diseases and make early warning of the risk. In practical application, it provides clinicians with scientific references for diagnosis and treatment decision-making.

The rest of this paper is organized as follows. After a literature review on the artificial intelligence research for disease diagnosis in Section 1, Section 3 presents a hybrid artificial intelligence framework for risk classification of acute respiratory diseases. Then, key enabling techniques, including deep learning-based Chinese clinical named entity recognition, and risk classification of acute respiratory diseases, are discussed in Section 4 and 5, respectively. Experiments and discussions are presented in Section 6 to verify the proposed model. Finally, the conclusions are presented in Section 7.

2 Literature review

Electronic medical records (EMRs) reduce the storage cost of paper-based medical records [4]. Sweeney [5] proposed an anonymized EMR system, accelerating the de-privacy of EMR in the medical field. The de-privacy of EMRs can further accelerate the research on diseases without worrying about disclosing patients' privacy. Information extraction from unstructured data in EMRs is to identify some critical entities from the text for further use. Traditionally, information extraction uses rule-based reasoning methods, but the disadvantage of these methods is that each data type requires a set of unique rules [6]. Beyond the information extraction function, the artificial intelligence-based disease diagnosis and prediction could be categorized into disease classification & diagnosis, tendency judgment, occurrence prediction, and risk classification [7]. Table 1 provides an overview of the themes, data types (e.g., EMRs and Magnetic Resonance Imaging (MRI)), and artificial intelligence models (e.g., convolutional neural network (CNN) and Long short-term memory (LSTM)) for various disease diagnoses and prediction. For instance, 1) use natural language processing technique for extracting the pathogenic factors from the unstructured medical data; 2) capture the changes and potential development direction of patients' disease by combining the diagnosis and treatment information in multiple periods; 3) find the risk factors affecting the disease from large-scale medical data, and identify the correlation of influencing factors. Although machine learning classifiers have already demonstrated excellent image-based diagnoses, analysis of diverse unstructured EMR data remains challenging [27].

Accurate information extraction from unstructured EMRs is the foundation of efficient risk classification. The clinical named entity recognition is a crucial task in information extraction, usually modeled as a sequence labeling problem. Machine learning algorithms, including the hidden Markov model, maximum entropy Markov model, bidirectional LSTM, conditional random field, and BERT embedding, are widely adopted [28]. Incorporating artificial intelligence technologies into the early warning of disease risk will help improve the

Table 1 Artificial intelligence models for disease diagnosis and prediction

Function	Theme	Disease type	Artificial intelligence models	Data Type	Ref.
Information extraction	Extract valuable information	Myocardial infarction	Gaussian naïve Bayes-based active balancing mechanism	Imbalanced electrocardiogram data	[8]
	Medical data processing	Beta-lactam allergy	Fast incremental decision tree	EMRs	[9]
Classification & diagnosis	Classification of chronic diseases	Chronic diseases	Hybrid deep learning	EMRs	[10]
	Classification of cardiac disorder	Cardiac disorder	Adaptive neuro-fuzzy inference system	Electrocardiogram signals	[11]
	Detect Covid-19 disease	Covid-19 disease	CNN	Chest X-ray images	[12]
	Infer illness and predict outcomes	Diabetes and mental health	LSTM	EMRs	[13]
	Brain disease prognosis	Brain disease	Weakly-supervised CNN	MRI and clinical scores	[14]
Tendency judgment	Infection rates of COVID-19	COVID-19	LSTM	EMRs	[15]
	Rehabilitation progress	Rehabilitation	CNN	Movement data	[16]
	Dynamic changes in disease	Congenital heart disease	Bayesian classification	Cardiopathy data	[17]
	Transcriptional effects of mutations	Mutations	Hybrid deep learning	DNA sequence	[18]
	Mortality detection in ICU	Unspecified	Deep learning and rule-based reasoning	EMRs	[19]
Occurrence prediction	Cognitive impairment conversion prediction	Dementia	Hybrid CNN	MRI	[20]
	Predict postoperative morbidity	Heart disease	Ensemble model	EMRs	[21]
	Predict the occurrence of a disease	Multicategory-multifactorial disease	Generalized artificial intelligence strategy	EMRs	[22]
Risk prediction	Predict the risk level of the disease	Multivariate disease	Deep learning model	EMRs	[23]
	Stratify the clinical risks of acute coronary syndrome	Acute coronary syndrome	Regularized stacked denoising auto-encoder model	EMRs	[24]
	Disease risk prediction	Unspecified	Multimodal data-based recurrent CNN	Semi-structured EMRs	[25]
	Multiple disease risk prediction	Multiple diseases	Directed disease network and recommendation system	EMRs	[26]

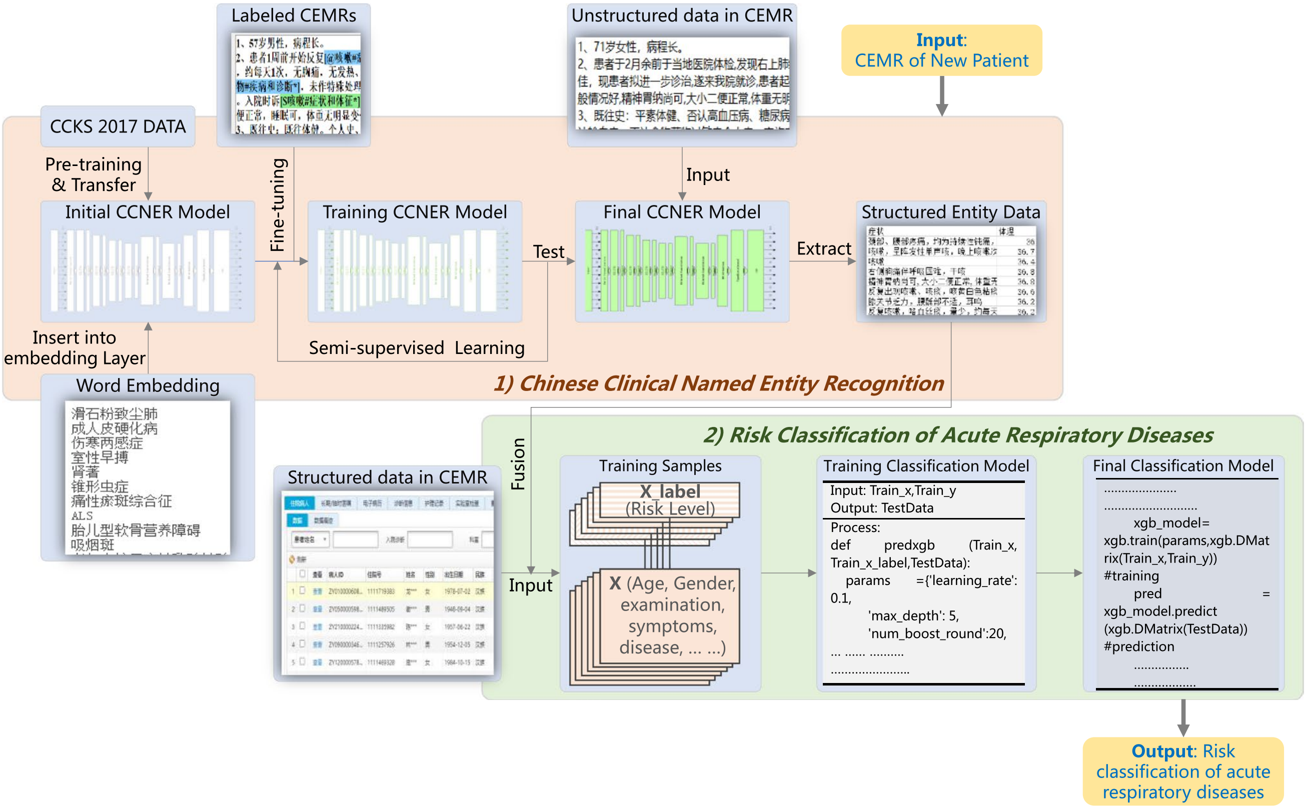
efficiency and accuracy of medical diagnosis. These models have achieved excellent results in improving medical quality, reducing mortality and morbidity, and reducing medical expenses. This research presents a mixed-use of bi-directional LSTM, dilated convolution, self-attention, conditional random field, transfer learning, and semi-supervised learning for risk classification of acute respiratory diseases.

3 Framework of the bi-level artificial intelligence model

The data size of unstructured text in the CEMR of each patient varies largely from a dozen of words to thousands of words. We have tried directly to build a deep learning model based on unstructured text data in the CEMR. However, the classification accuracy is poor and far from satisfactory because there exist distracting information and useless information in unstructured text data in the CEMR, besides it is quite difficult to establish an effective input data modeling due to the wide range of the data size of unstructured text. Therefore, based on a large number of semi-structured data of CEMRs from the China Hospital Pharmacovigilance System, this study proposes a bi-level artificial intelligence model for risk classification of acute respiratory diseases. As shown in Fig. 1, the proposed model includes two modules. Firstly, a hybrid deep learning module is established to realize Chinese clinical named entity recognition from the unstructured data in CEMRs. To improve the efficiency and accuracy of named entity recognition, key techniques such as self-attention, transfer learning, and semi-supervised learning mechanism are introduced. Integrating the extracted entity data with other structured data in the CEMRs, the second part is a customized XGBoost [29] module to mine the structured CEMRs data to realize the risk classification of acute respiratory diseases.

The highlight of the proposed model is threefold:

a) Combination of Chinese word vector (trigram and bigram) and Chinese character/token vector (unigram): The word vector is introduced to avoid ambiguity in the Chinese language model, while the character vector contains more semantic information.
b) Elaborated design of neural network: The structure of the neural network is crucial to information extraction. The bi-directional Long Short-Term Memory neural network, conditional random field, and dilated convolution are integrated.
c) Transfer learning strategy: To achieve better precision without massive data, the labeled data from the China Conference on Knowledge Graph and Semantic Computing (CCKS) 2017 (https://biendata.xyz/competition/CCKS2017_2/) is used to pretrain parameters to make the neural network more suitable for the CMER data. The purpose of transfer learning is to perform pretraining of the model and then transfer to the new data in the same field, which can speed up the training process.
d) Semi-supervised strategy: The semi-supervised technique is introduced into the neural network model to achieve a reasonable accuracy rate in a small amount of EMR data. The training set is expanded with self-semi-supervised learning. Semi-supervised learning can


Labeled CEMRs
1、57岁男性，病程长。
2、患者1周前开始反复[@咳嗽#
，约每天1次，无胸痛，无发热，
物#疾病和诊断*]，未作特殊处理
。入院时诉[$咳嗽#症状和体征*]
便正常，睡眠可，体重无明显变
3、既往史：既往体健，个人史，
Unstructured data in CEMR
1、71岁女性，病程长。
2、患者于2月余前于当地医院体检,发现右上肺
佳，现患者拟进一步诊治,遂来我院就诊,患者起
般情况好,精神胃纳尚可,大小二便正常,体重无明
3、既往史：平素体健、否认高血压病、糖尿病
Input:
CEMR of New Patient
CCKS 2017 DATA
Pre-training
& Transfer
Fine-tuning
Input
Initial CCNER Model
Training CCNER Model
Test
Final CCNER Model
Extract
Structured Entity Data
Insert into
embedding Layer
Semi-supervised Learning
Word Embedding
滑石粉致尘肺
成人皮硬化病
伤寒两感症
室性早搏
肾著
锥形虫症
痛性瘀斑综合征
ALS
胎儿型软骨营养障碍
吸烟斑
1) Chinese Clinical Named Entity Recognition
2) Risk Classification of Acute Respiratory Diseases
Fusion
Structured data in CEMR
Input
Training Samples
X_label
(Risk Level)
X (Age, Gender, examination, symptoms, disease, … …)
Training Classification Model
Final Classification Model
Output: Risk classification of acute respiratory diseases

```
Input: Train_x,Train_y
Output: TestData
Process:
def     predxgb     (Train_x,
Train_x_label,TestData):
   params     ={'learning_rate':
0.1,
         'max_depth': 5,
         'num_boost_round':20,
... ...... ..........
......................
```

```
....................
..........................
        xgb_model=
xgb.train(params,xgb.DMat
rix(Train_x,Train_y))
#training
        pred                    =
xgb_model.predict
(xgb.DMatrix(TestData))
#prediction
        ...............
        .................
```

Fig. 1 The bi-level workflow of the risk classification of acute respiratory diseases

reduce the dependence on human labeling data and realize real artificial intelligence.

4 Deep learning-based Chinese clinical named entity recognition

A dedicated design of the "BiLSTM+Dilated Convolution+3D Attention+CRF" deep learning model is presented for Chinese clinical named entity recognition (CCNER) of CEMR. The CCNER is to identify and extract the individual entities (including entity type and entity boundary) about the patient's physical condition, clinician's diagnosis, and treatment from the unstructured data of CEMRs.

A. *Data modeling for deep learning*

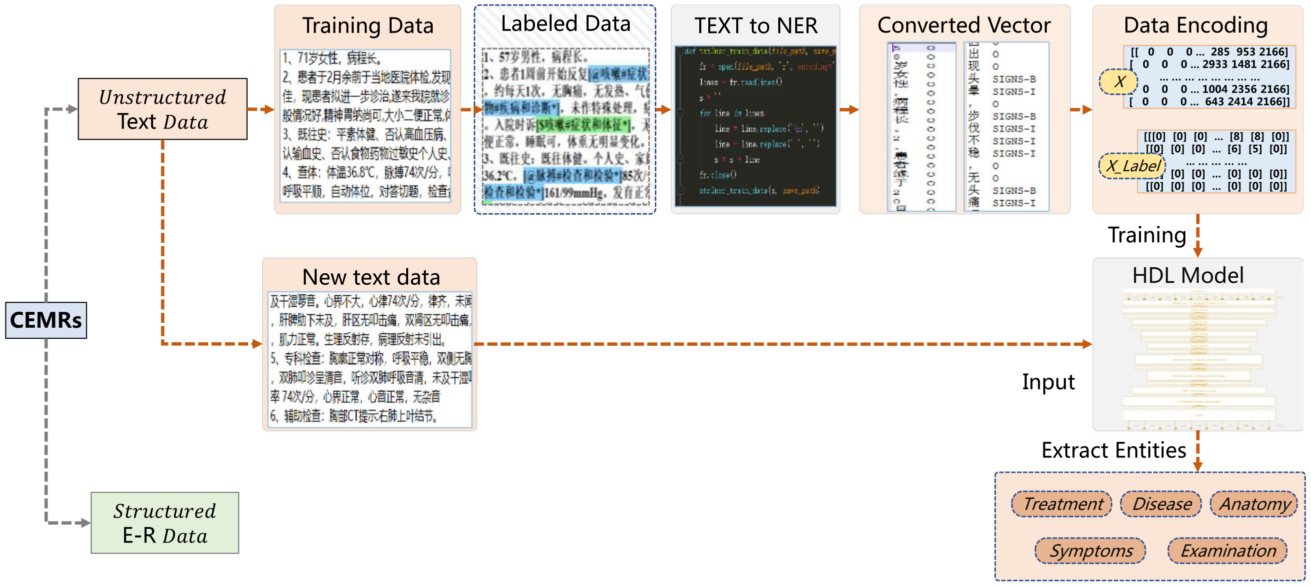
The data under research comes from the semi-structured CEMRs (i.e., Chinese Electronic Medical Records) of real patients with acute respiratory diseases in China Hospital Pharmacovigilance System (CHPS) established based on the National Adverse Drug Reaction Reporting System of China. CEMRs contain the patient's diagnosis, time of admission, description of the illness, description of physical condition, and other relevant clinical information. As shown in Fig. 2, the CEMRs in CHPS could be categorized into unstructured text data and structured E-R data (e.g., age, gender, temperature) [30].

1) *Chinese electronic medical records*

Through the analysis of CEMR data, the entities to be extracted are divided into the following five commonly-used categories:

1) Examination (检查和检验). It includes biopsy, chest Computed Tomography, MRI, and other examination that are conducted to detect the patient's physical condition.
2) Symptom (症状与体征). It represents a patient's current physical state and often contains the manifestation situation or precursor of some diseases. For example, dizziness, vomiting, coughing, sputum, and other symptoms as vital disease signs, from which the clinician can infer the disease and severity of these symptoms.
3) Disease (疾病与诊断). It records the patient's health status, which is an essential feature of data mining in the later stage, such as tumor shadow, lung cancer, differentiated adenocarcinoma, and other diseases.
4) Treatment (治疗). It refers to the treatment measures for the disease. The effectiveness of the drug (e.g., chemotherapy, pemetrexacin, libitol, and carboplatin chemotherapy) and the accuracy of the application can be compared after the treatment.
5) Anatomy (身体部位). Some descriptions of the patient's body parts correlate with the disease, such as alveolar walls, lungs, right maxillary sinus, and other body parts.

In terms of entity categories, it is necessary to ensure no duplication/ambiguity or mutual inclusion among entities. In the CCNER, two kinds of entities, namely, disease and symptom, attract more attention. There are some pre-negative words and uncertain modifiers, such as untouched (未



1、71岁女性，病程长。
2、患者于2月余前于当地医院体检,发现
佳，现患者拟进一步诊治,遂来我院就诊
般情况好,精神胃纳尚可,大小二便正常,
3、既往史：平素体健、否认高血压病、
认输血史、否认食物药物过敏史个人史、
4、查体：体温36.8℃，脉搏74次/分，
呼吸平顺，自动体位，对答切题，检查

1、57岁男性，病程长。
2、患者1周前开始反复[@咳嗽#症状
，约每天1次，无胸痛，无发热、气
物#疾病和诊断*]，未作特殊处理，
，入院时诉[$咳嗽#症状和体征*]，
便正常，睡眠可，体重无明显变化。
3、既往史：既往体健。个人史、家
36.2℃，[@脉搏#检查和检验*]85次/
检查和检验*]161/99mmHg，发育正

```
8    O        出  O
岁   O        现  O
女   O        头  SIGNS-B
性   O        晕  SIGNS-I
，   O        ，  O
病   O        步  SIGNS-B
程   O        伐  SIGNS-I
长   O        不  SIGNS-I
。   O        稳  SIGNS-I
2    O        ，  O
、   O        无  O
患   O        头  SIGNS-B
者   O        痛  SIGNS-I
```

```
[[   0    0    0 ...  285  953 2166]
 [   0    0    0 ... 2933 1481 2166]
 ... ... ... ... ... ... ... ...
 [   0    0    0 ... 1004 2356 2166]
 [   0    0    0 ...  643 2414 2166]]

[[[0]  [0]  [0]  ...  [8]  [8]  [0]]
 [[0]  [0]  [0]  ...  [6]  [5]  [0]]
 ... ... ... ... ...
 [[0]  [0]  [0]  ...  [0]  [0]  [0]]
 [[0]  [0]  [0]  ...  [0]  [0]  [0]]]
```

及干湿啰音，心界不大，心律74次/分，律齐，未闻
，肝脾肋下未及，肝区无叩击痛，双肾区无叩击痛，
，肌力正常，生理反射存，病理反射未引出。
5、专科检查：胸廓正常对称，呼吸平稳，双侧无胸
，双肺叩诊呈清音，听诊双肺呼吸音清，未及干湿
率 74次/分，心界正常，心音正常，无杂音
6、辅助检查：胸部CT提示右肺上叶结节。

Fig. 2 Data modeling and encoding of the unstructured text data in the CEMRs

触及) and undiscovered (未发现), will change the semantic connotation. It is also critical to capture the features and meaning of such pre-negative words or modifiers.

2) *Data annotation and encoding for generating training samples*

The annotation of the patient's CEMR is to identify and mark the critical entities of diseases, symptoms, examinations, anatomy, and treatment. The annotation algorithm is developed based on the YEDDA (https://github.com/johnzhaoxiao/YEDDA) in the Python3.5 environment. The annotation algorithm CEMR will generate a ".ann" file. A conversion script is also written to convert it to the training set of the deep learning algorithm with Python3.5. Figure 2 shows how to convert the text-type clinical data into training samples.

A critical step in data preprocessing of training samples is encoding the word into the vector. There are no orthographic boundaries between words in Chinese, which is the main difficulty of working with Chinese computationally (in addition to the bewildering array of encodings used for Chinese and the simplified/traditional script controversy). A Chinese word (e.g., "咳嗽") frequently consists of two, three, or more Chinese characters (e.g., "咳", "嗽"). The Chinese character/token features include the semantic information of the word. The word-based method may cause mistakes of segmentation, which will result in lexical-semantic ambiguity. The character-based method can avoid semantic deviation brought by the segmentation error in the word-based method. A combination of Chinese word vector and Chinese character vector brings better semantic information of the training CEMR data to the deep neural network.

Word embedding is used to preserve the semantic information of the high-dimensional sparse vector through low-dimensional space mapping in the context [31]. Traditional one-hot encoding of word and character is of high dimension and sparse vector. Other encoding methods, such as word2vec (code.google.com/archive/p/word2vec/), Fastext (fasttext.cc/), GloVe (github.com/stanfordnlp/GloVe), Emlo (allennlp.org/elmo), and Bert (github.com/google-research/bert), can effectively map words into the low-dimensional vector space as well as capture the semantic information. This study conducts comparative experiments among various word embedding methods, and FastText is selected to search for a number of related medical term corpus for unsupervised word vector training in CCNER.

B. *Design of deep neural network*

CCNER is a multi-label classification task [10] that needs a feature extraction algorithm to capture entity type and entity boundary, and thus to capture context semantics and word semantics. A hybrid deep learning module is presented to identify: 1) whether the word is an entity or not, 2) which entity type it is, and 3) where the boundary (the beginning and the end) of the recognized entity is.

1) *Architecture of hybrid deep learning module for CCNER*

In the named entity recognition task, it is a common practice to use a neural network to extract context semantic information, then classify it through a full connection layer, and finally, use the conditional random field (CRF) layer for sequence constraint to improve accuracy [32]. The neural network predicts the entity category and boundary of the word through the full connection layer. The CRF layer learns the transfer probability in the text, calculates the loss function, and then performs the backpropagation to update the weight of each layer of the network. Figure 3 illustrates the network structure of an innovative "BiLSTM+Dilated Convolution+3D Attention+CRF" hybrid deep learning (HDL) module for CCNER designed in this paper.

The proposed deep learning-based CCNER module contains a four-layer bi-directional Long Short-Term Memory neural network (BiLSTM), bi-layer Dilated Convolution, a Self-Attention Layer, and a CRF Layer. The BiLSTM is more suitable for extracting features of long sequences than the Transformer [33] (which ultimately abandoned the circulation mechanism of RNN and adopted a way of self-attention for global processing), because the computation of Transformer increases when the sequence changes to length [34]. The absolute position-coding of the Transformer cannot capture word order, leading to a reduction in the capturing ability of long sequences. In the sequence problem of the named entity recognition task, the transmission of time step in BiLSTM takes more consideration of the sequential relation of words than the absolute position-coding of Transformer. Two layers of Dilated Convolution are added for expanding receptive fields at the end of BiLSTM to capture both the local context semantic information and correlation feature among the words.

The use of BiLSTM enhances the ability to capture the context semantics than a single LSTM. Traditional LSTM neural network usually suffers from the locality in the semantic extraction of text because of its long-term dependence. The CCNER requires global information in solving the task of named entity recognition. A feature-weighted classification using an attention mechanism can better capture semantic information for a named entity recognition task. Therefore, the self-attention mechanism and CRF layer are used in the proposed HDL-based CCNER module to capture and repair the correlation between global information in feature extraction.

On the one hand, a 3D-attention layer is built to conduct a global weighting to find the importance of the features

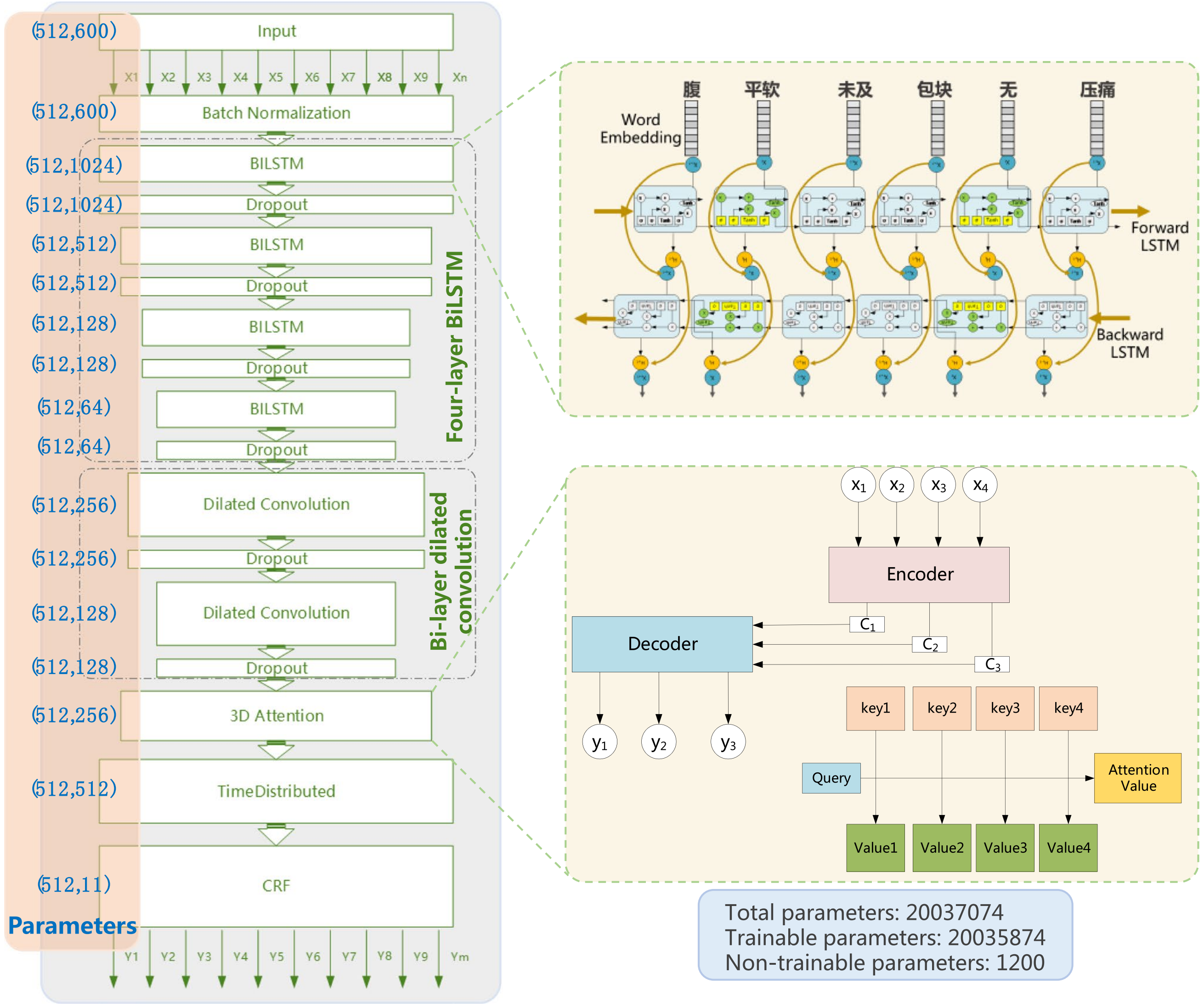


Fig. 3 The network structure of the "BiLSTM+Dilated Convolution+3D Attention+CRF" model for CCNER

extracted from the words. It does not rely on the fixed influence of words, which avoids the shortcomings of the local information extraction in BiLSTM. On the other hand, the SoftMax function is conventionally used as mapping to 0–1 interval in the last layer of the neural network, which will lead to a weak correlation between labels. To obtain the correlation between labels in the CCNER task, CRF is used in HDL to perform normalization to remove a constraint on the use of global information to repair correlation for labels, and consequently, the algorithm robustness will be improved.

Notably, as shown in Fig. 3, the number of parameters in the proposed HDL module is about 20 million, which is much less than the 130 million parameters of the widely-used Bert+BiLSTM+CRF model. Thus, the forward reasoning time of HDL is much faster than the Bert+BiLSTM+CRF model, which is more suitable for online usage than the Bert+BiLSTM+CRF model.

2) *Four-layer BiLSTM*

BiLSTM is a combination of the forward directional LSTM and the backward directional LSTM. LSTM substantially alleviates the gradient disappearance of traditional Recursive Neural Network (RNN), and LSTM can effectively alleviate the long-term dependence by improving the implicit structure of traditional RNN and thus could capture more extended sequence semantics. LSTM has three gated structures, namely, the input gate, the forget gate, and the output gate.

The input gate decides to let how much new information get into the BiLSTM cell state updated from S_{t-1} state to

S_t. The forget gate is a smart design in BiLSTM, which can alleviate the long-term dependence and enable the CCNER model to capture more extended context semantics. The forget gate is responsible for getting rid of some unimportant information. The characteristics of the Sigmoid function also determine the function of the forget gate: 1 is to retain all information ultimately, and vice versa. For example, in terms of predicting the next word in CCNER, the current BiLSTM cell already contains subject information, where the pronoun is picked out; and the model can capture a more extended sequence by trying to choose to forget the old subject when it captures the new subject to alleviate long-term dependency. The output gate controls the final output and will also remove some information of the cell state that does not need to be output. It maps the state information through the Tanh layer to the values between −1 and 1, and multiplies these values by the previous Sigmoid function to get the output information.

$$i_t = \sigma\left(W_i \bullet \left[h_{t-1}, X_t\right] + b_i\right) \tag{1}$$

$$S_t = \tanh\left(W_S \bullet \left[h_{t-1}, X_t\right] + b_S\right) \tag{2}$$

$$f_t = \sigma\left(W_f \bullet \left[h_{t-1}, X_t\right] + b_f\right) \tag{3}$$

$$O_t = \sigma\left(W_0\left[h_{t-1}, X_t\right] + b_0\right) \tag{4}$$

$$h_t = O_t * \tanh\left(S_t\right) \tag{5}$$

In BiLSTM, the forward LSTM network computes the hidden state $\overrightarrow{h_t}$ of the left context of the sentence at the word X_t, while a backward LSTM network reads the same sentence in reverse and outputs $\overleftarrow{h_t}$ given the right context. These two vectors are concatenated to form the hidden state of a BiLSTM network, i.e., $h_t = \left[\overrightarrow{h_t};\overleftarrow{h_t}\right]$, which can make use of more sentence-level information. Table 2 provides an overview of the notations used in the four-layer BiLSTM module.

Table 2 Notations used in the four-layer BiLSTM module

Notations	Implications
S_t	The state of the current BiLSTM cell
i_t	The new information getting into the BiLSTM cell state
σ	The Sigmoid function
f_t	The forgotten information in the BiLSTM cell state
X_t	The input of the current BiLSTM cell
b_i, b_S, b_f, b_0	The bias in three gates of the current BiLSTM cell
W_i, W_S, W_f	The weights in three gates of the current BiLSTM cell
O_t	The output information in the BiLSTM cell state
h_{t-1}, h_t	The output of the previous and current BiLSTM cell
$h_t = \left[\overrightarrow{h_t};\overleftarrow{h_t}\right]$	Two vectors (forward and backward) to form the hidden state of a BiLSTM network

To better capture the contextual semantic information, we need to add more neural layers to capture the context. After dozens of experiments in the CCNER module for information extraction, the results show that the loss will dramatically increase when the layer number of BiLSTM is greater than 4. The dimension of the first layer is 512, which is the maximum word width of the Bert module. The dimensions of the hidden layers were usually halved layer by layer. It is tuned based on performance on the validation dataset (also called a trial-and-error approach). Moreover, the computation time increases exponentially with the increase of the parameter number. Along with the increasing layer number of the HDL model, the large-scale parameters are more likely to cause an overfitting issue. To avoid the overfitting problem, regularized Dropout layers need to be added to the BiLSTM to reduce the complexity of parameter adjustment and thereby speed up the model convergence.

3) *Bi-layer dilated convolution*

In the CCNER task, each word of a sentence is mapped into a vector. Each word will affect the accuracy of the named entity recognition task. Extracting each entity needs the analysis of contextual information. To avoid information loss in the sampling process, a bi-layer of dilated convolution is incorporated into the HDL module to capture more extensive context information of the next sentence.

In the conventional convolution, the sliding window scanning of the input position is used to extract local correlation, and the convolution translation is performed by sliding window movement. The dilated convolution is used in the HDL module to make the receptive field of convolution increase via a filling of 0 between the convolution spaces, while the size of the convolution matrix itself remains unchanged [35]. Different sizes of kernels (e.g., 3 and 7) are tested to capture context information of various granularities. In this way, convolution captures a broader range of data features to scan and find the relationship between sentences.

The bi-layer of dilated convolution entrusts probability information to each word of the input sentence, which is consistent with the position information assigned to each word by the BiLSTM layer. Then, it is transmitted to the 3D-attention layer, which can be decoded using the Viterbi algorithm to find the optimal path.

4) *Self-attention layer*

The CCNER needs to not only focus on local information but also to capture holistic context information. BiLSTM relieves long-term dependence to a certain extent, but a more prolonged dependence will make the BiLSTM inefficient. The learning and training in the BiLSTM model are not parallel because the current state of BiLSTM depends on the previous state. The attention mechanism can achieve parallel training to cut down dependence on the previous information and thus improve the efficiency of training [20]. A widely-used encoder-decoder self-attention mechanism [36] is introduced to extract some crucial features of data and to mitigate longer-term dependence in the HDL module. The encoder will translate sentences into vectors, and then the decoder decrypts the vector, which reduces the loss of information.

The computation logic of the self-attention mechanism is shown in Fig. 4. The bottom layer of the self-attention module of HDL is a mapping of Query, Key, and Value. Phase 1 input the Query multiplied by Key to get the weight of each Key's Value. In Phase 2, a SoftMax function is used for 0–1 interval mapping. Phase 3 do product between Query and Key to calculate the similarity (the higher the similarity and the matching degree, the greater the weight), and then the weighted sum is multiplied by Value to calculate the Value of Attention, which can calculate the weight of each word in the text and give more weight to essential feature words.

The advantage of the self-attention mechanism is to capture global information. However, the disadvantage of the attention mechanism is that it cannot consider the characteristics of a sequential relationship because all features are parallel processing. The absolute cosine positional coding used in Bert also suffers from some shortcomings [37]. Therefore, a FastText-based complex embedding technique is used in HDL to get rid of a pretrain-finetune discrepancy.

5) *Output CRF layer*

To obtain the correlation between labels in the CCNER task, CRF is used in HDL to perform normalization to remove a constraint and repair correlation for labels, and consequently, the model robustness will be improved. The CRF is a kind of undirected probabilistic graph model. Under the condition of P(Y|X) (X as the input and Y as the output), Y can be used as a sequence label. Using (the regularization of) the maximum likelihood estimation on the training data, the maximal Y could be calculated under the conditional probability P(Y|X). The conditional probability of a conditional random field is calculated using the following formula:

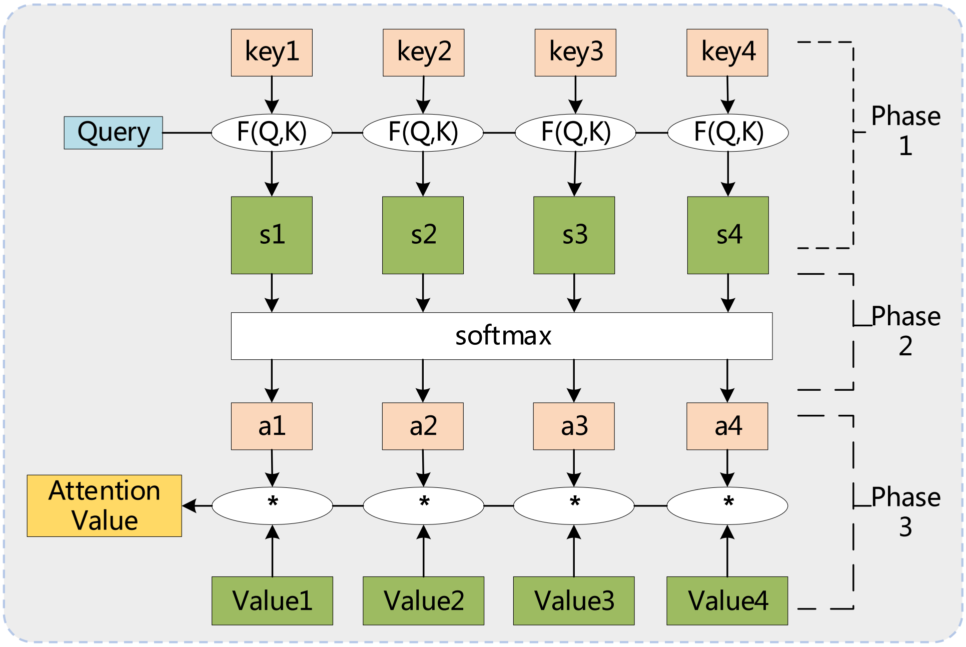
$$\mathrm{p(y|x)} = \frac{1}{Z(x)}\prod_{i=1}^{m}\varphi_i\left(y_{C_i},x_{C_i}\right) \tag{6}$$

$$\mathrm{Z(x)} = \sum\nolimits_y \prod_{i=1}^{m}\varphi_i\left(y_{C_i},x_{C_i}\right) \tag{7}$$

where m represents the maximum number of clusters. C_i represents the ith largest cluster. x_{C_i} and y_{C_i} represent the random vector corresponding to the cluster vertex, respectively. φ denotes the potential function. Z(x) represents the normalized function.

There is a correlation among the observation sequence in a large number of training data, and thus a simple feature function cannot capture all the entities in CCNER tasks. Compared to the hidden Markov chain model that quickly falls into the local optimal results, CRF calculates conditional probability through the input data and makes normalization of global information to solve the

Fig. 4 The computation logic of the self-attention mechanism



annotation bias problem, which is suitable for named entity recognition task.

III. *Key techniques for achieving high accuracy in HDL*

The transfer learning and semi-supervised learning technique are used to improve the performance of the proposed HDL module.

1) *Transfer learning mechanism*

The pretraining and finetuning is a popular trick in deep learning models, which require massive data to be trained. Generating the training data requires too much workforce to do the labeling job. Moreover, high-quality labeled data is challenging to obtain. Transfer learning can achieve excellent results even with a small amount of annotation data [38]. Pretraining the model weights based on a similar public dataset can achieve reasonable accuracy. For instance, Bert's pretraining is a time-consuming task, and Google uses a large amount of Tensor Processing Unit (TPU) for pretraining Bert in millions of high-quality corpus. Ordinary companies use transfer learning on these pretrained models in finetuning their data for the weights, which can achieve outstanding accuracy on a small amount of labeled data and significantly reduce the computation complexity. At present, there are many pretraining models in the field of natural language processing, such as EMLO (allennlp.org/elmo), BERT (github.com/google-research/bert), GPT (github.com/openai/gpt-3), and XLNET (github.com/zihangdai/xlnet). When doing some downstream tasks, higher accuracy can be achieved by finetuning models.

This study uses the Chinese electronic medical record dataset from CCKS 2017 (biendata.xyz/competition/CCKS2017_2/) to perform a pretraining of the proposed HDL module. The test evaluation shows the accuracy of the pretrained HDL module is up to 92%, which is comparable in accuracy with other algorithms [39, 40]. Then, the pretrained HDL is transferred using labeled CCNER data from CHPS.

2) *Semi-supervised learning mechanism*

Another technique for achieving high accuracy in HDL is the semi-supervised learning mechanism. Semi-supervised learning is to use unlabeled data for marking, and then train with the labeled data to increase the scale of training data [41]. Semi-supervised learning can be summarized into three categories: 1) pretraining with unlabeled data (including unsupervised or pseudo-supervised pretraining), then finetuning with label data; 2) Semi-supervised algorithm based on network features (use labeled data to train the network, then use the trained network features to classify the unlabeled data, and finally select some useful unlabeled data to add to the training set); and 3) make the network work in self-training semi-supervised fashion (use the trained network to predict the label of unlabeled data).

The semi-supervised learning strategy in the HDL module is a kind of pseudo-supervised pretraining on unlabeled data. The principle is to inject the pseudo-label data that could obtain a high level of confidence in the training set. It can improve a few percentage points of classification accuracy on the named entity recognition task of CEMR and reduce workforce to expand the training set.

5 Risk classification of acute respiratory diseases

This section introduces a machine learning module for the risk classification of acute respiratory diseases.

A. *Data modeling for risk classification*

A changing or deteriorating process from minor/mild respiratory disease to acute respiratory disease will happen if the patient doesn't take countermeasures. Therefore, the occurrence of acute respiratory disease deteriorated from minor/mild respiratory disease should be predicted in advance. The Risk Level is defined as illustrated in Fig. 6 and will be predicted as an early warning of acute respiratory disease. The data label for classification is the patient's risk level. The risk classification is a multi-grade classification [42] from the patient's various indicators and symptoms description as well as previous treatment information. Capturing the relationship between the clinical data and the final illness from multiple dimensions is critical to find the potential factors related to the disease and thus to predict the risk level of the patient's disease. As shown in Fig. 5, the data modeling for risk classification is based on the extracted entity data and other structured data in the CEMRs.

There exists some scoring method for making a comprehensive judgment of the patient's ability to act illness severity. Figure 6 presents a definition of risk level based on the well-known Zubrod-Ecog-WHO Performance Score (ZPS) method and the Karnofsky Performance Score (KPS) method [43]. The KPS is a functional status score method. The higher the score, the better the health status. This scoring method is designed based on an investigation in The First Affiliated Hospital of Guangzhou Medical University, which is a well-known hospital in the respiratory disease treatment area in China. This scoring method has been embedded in the database of the China Hospital Pharmacovigilance System.

Generally, a score above 80 is considered non-dependent and can live on its own without the care of others. A score from 50 to 70 is classified as semi-dependent and

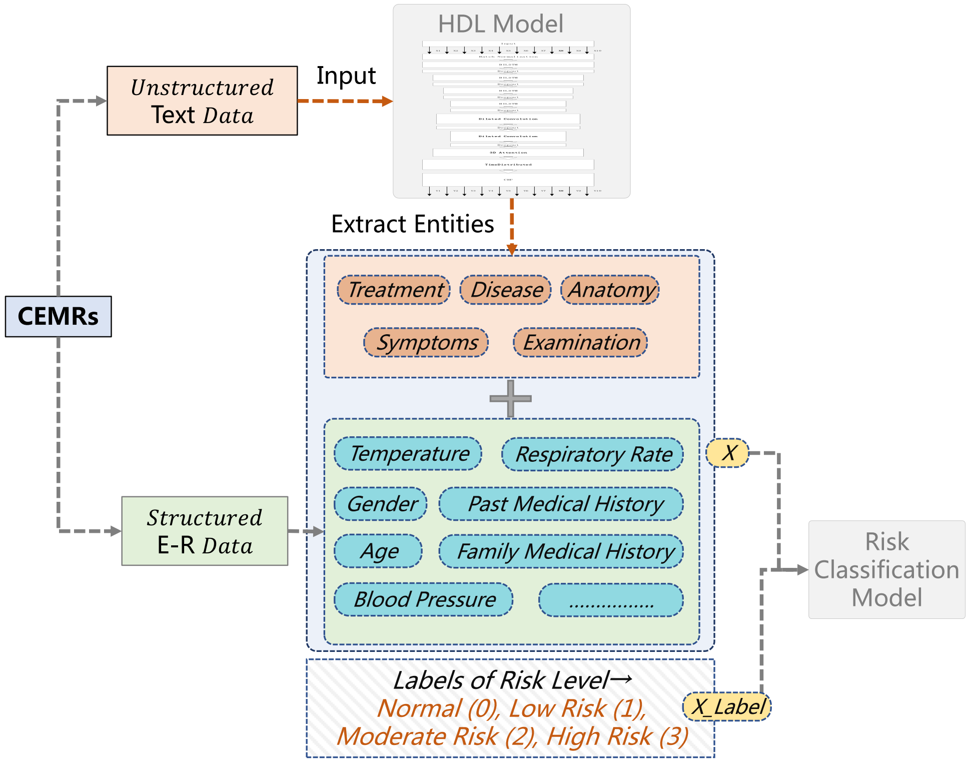
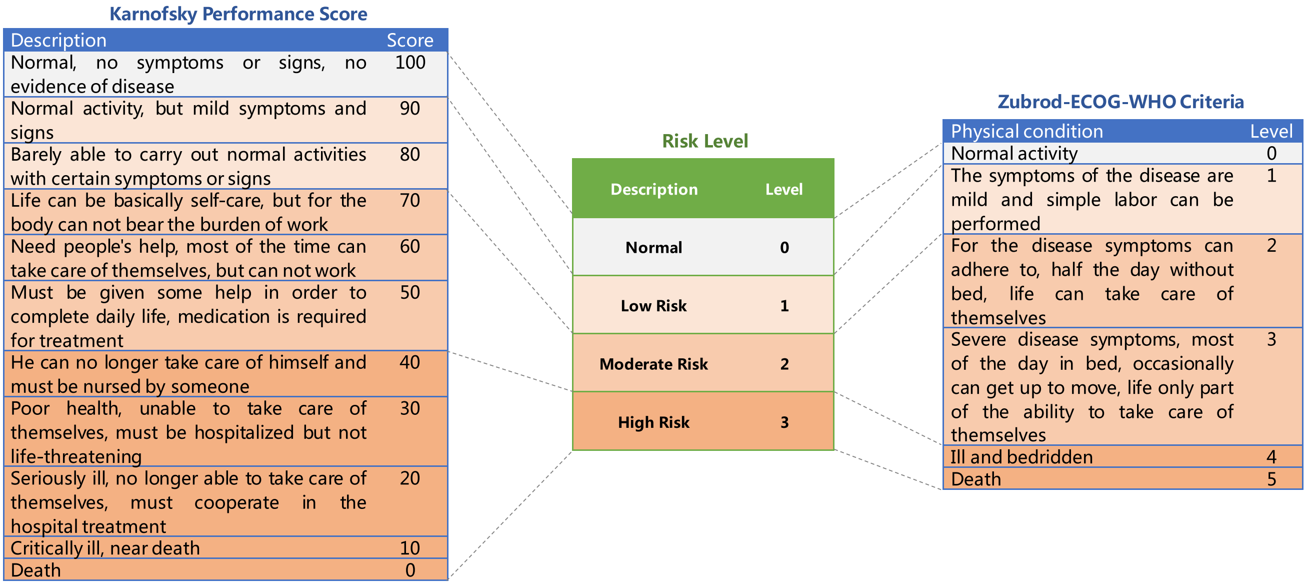


Fig. 5 The data structure of training samples in the risk classification model


Karnofsky Performance Score

Description	Score
Normal, no symptoms or signs, no evidence of disease	100
Normal activity, but mild symptoms and signs	90
Barely able to carry out normal activities with certain symptoms or signs	80
Life can be basically self-care, but for the body can not bear the burden of work	70
Need people's help, most of the time can take care of themselves, but can not work	60
Must be given some help in order to complete daily life, medication is required for treatment	50
He can no longer take care of himself and must be nursed by someone	40
Poor health, unable to take care of themselves, must be hospitalized but not life-threatening	30
Seriously ill, no longer able to take care of themselves, must cooperate in the hospital treatment	20
Critically ill, near death	10
Death	0

Risk Level

Description	Level
Normal	0
Low Risk	1
Moderate Risk	2
High Risk	3

Zubrod-ECOG-WHO Criteria

Physical condition	Level
Normal activity	0
The symptoms of the disease are mild and simple labor can be performed	1
For the disease symptoms can adhere to, half the day without bed, life can take care of themselves	2
Severe disease symptoms, most of the day in bed, occasionally can get up to move, life only part of the ability to take care of themselves	3
Ill and bedridden	4
Death	5

Fig. 6 Definition of risk level for early warning of acute respiratory disease

occasionally requires the care of others. A score below 50 is considered as unable to live on their own. The ZPS score shows the behavioral ability score in the treatment of severe diseases such as lung cancer. By referring to the KPS and ZPS method, this study defines four risk levels of acute respiratory diseases, namely, Normal (Level 0), Low Risk (Level 1), Moderate Risk (Level 2), and High Risk (Level 3) to indicate the patient's illness. Level 0 could be mapped into score 100 in KPS and 0 in ZPS. Level 1 could be mapped into a score of 80–100 in KPS and 1 in ZPS. Level 2 could be mapped into a score of 50–80 in KPS and 2–3 in ZPS, which needs to be tracked to see if it goes to a high-risk level, and further intensive treatment is needed. Level 3 could be mapped into scores 0–50 in

KPS and 4–5 in ZPS, which needs the highest level of individual attention.

B. *Customized XGBoost for risk classification*

Risk warning of disease is essentially a multi-label classification of risk levels. There are many classification algorithms in machine learning, such as logistic regression, support vector machine, random forest, and XGBoost algorithm [29]. Logistic regression is a widely used algorithm that calculates output variables by discretizing the characteristics of input variables. Logistic Regression is of low computation complexity and is commonly used as a benchmark model. Support Vector Machine (SVM) algorithm has excellent generalization performance in the case of a small number of samples. SVM algorithm cannot be trained in a parallel manner, so it is not suitable for large-scale sample training. Random Forest algorithm is a robust classifier formed by several weak classifier groups. It not only improves the classification ability of the decision tree but also avoids the problem that the decision tree is easy to overfit.

XGBoost algorithm is an integrated learning method similar to the random forest algorithm, which uses integrated learning to upgrade the weak classifier to the robust classifier. XGBoost is designed to use Boost thinking to promote and reduce data bias, and the model is extremely robust. Therefore, the XGBoost algorithm is used as the kernel of the disease risk classification model in this paper. Table 3 provides the XGBoost, together with the parameters used in the risk warning of HDL.

The reason why the XGBoost performs training quickly is that the continuous processing of features can find the best segmentation point for feature pre-sorting, and block storage can store the data in memory respectively for each column. XGBoost algorithm is an improvement of the Gradient Boosting Decision Tree (GBDT) algorithm [44]. XGBoost includes a level-wise undifferentiated splitting of all subnodes of the same layer each time. Instead of adopting GBDT's violent method, the XGBoost algorithm uses the feature pre-sorting technique and then performs the multi-threaded parallel computation to improve the training speed. Since pre-sorting is column storage, it is stored in persistent memory for accelerating the reads and writes in the training phase, making XGBoost perform excellently.

Table 3 The implementation detail of the customized XGBoost model

```
Input: Train_x,Train_y
Output: TestData
Process:
def predxgb (Train_x, Train_x_label,TestData):
   params ={'learning_rate': 0.1,
        'max_depth': 5,
        'num_boost_round':20,
        'objective': 'multi:softmax',
        'random_state': 27,
        'silent':0,
        'num_class':4
       }
        xgb_model= xgb.train(params,xgb.DMatrix(Train_x,Train_x_label))
#training
        pred = xgb_model.predict (xgb.DMatrix(TestData)) #classification
        prob = [ ]
        l = len(pred)
        for i in range(l):
           prob.append(pred[i][1])
        return prob  # Returns the predicted probability value of Level 0-3
```

6 Experiments and discussions

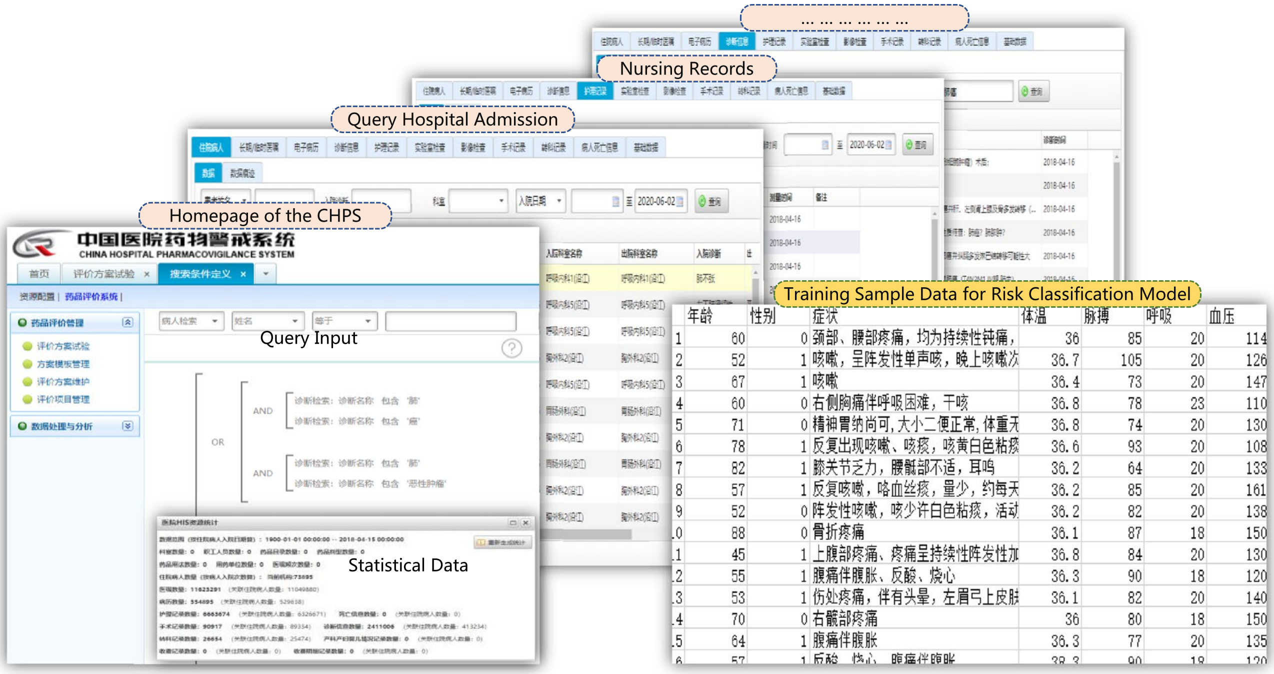
To the best of our knowledge, there is no similar hybrid system for risk classification of acute respiratory diseases based on semi-structured CEMRs. Therefore, the performance analysis of the presented system against state-of-the-art architectures is conducted from two aspects independently. The CCNER module is evaluated based on a comparison with Bert+BiLSTM+CRF and other basic models. The risk-predictor is evaluated based on a comparison with Logistic regression, SVM, and Random Forest.

A. *Data source and pre-processing*

This research is driven by the increasing need to predict the risk level of acute respiratory diseases under the COVID-19 pandemic. The research was carried out for three types of acute respiratory diseases, namely lung cancer, severe pneumonia, and severe asthma. The high mortality rate of these three types of acute respiratory diseases, if detected and treated early, will be significantly reduced.

The data under research comes from the CHPS (i.e., China Hospital Pharmacovigilance System) established based on the National Adverse Drug Reaction Reporting System (Fig. 7). It includes a massive medical database combined with artificial intelligence models to promote the prevention and control of severe respiratory diseases and reduce mortality. All CEMRs are unified in the writing form and include information such as age, gender, symptoms, past medical history, family medical history, and examination. Figure 7 also shows some examples of training data labeled with the risk level of the disease. Figure 8 provides the screenshots in the CCNER implementation designed in this paper.

The deep learning algorithm is susceptible to the missing values of the data. The missing values must be filled in before importing algorithm training. For the numerical value type, this paper uses the median filling method to avoid the interference effect brought by extreme value. For text types, special symbols or similar semantics based on the model classification results are used to fill the missing values. If there is no suitable value to fill or there is too much missing, this piece of information will be omitted directly to ensure



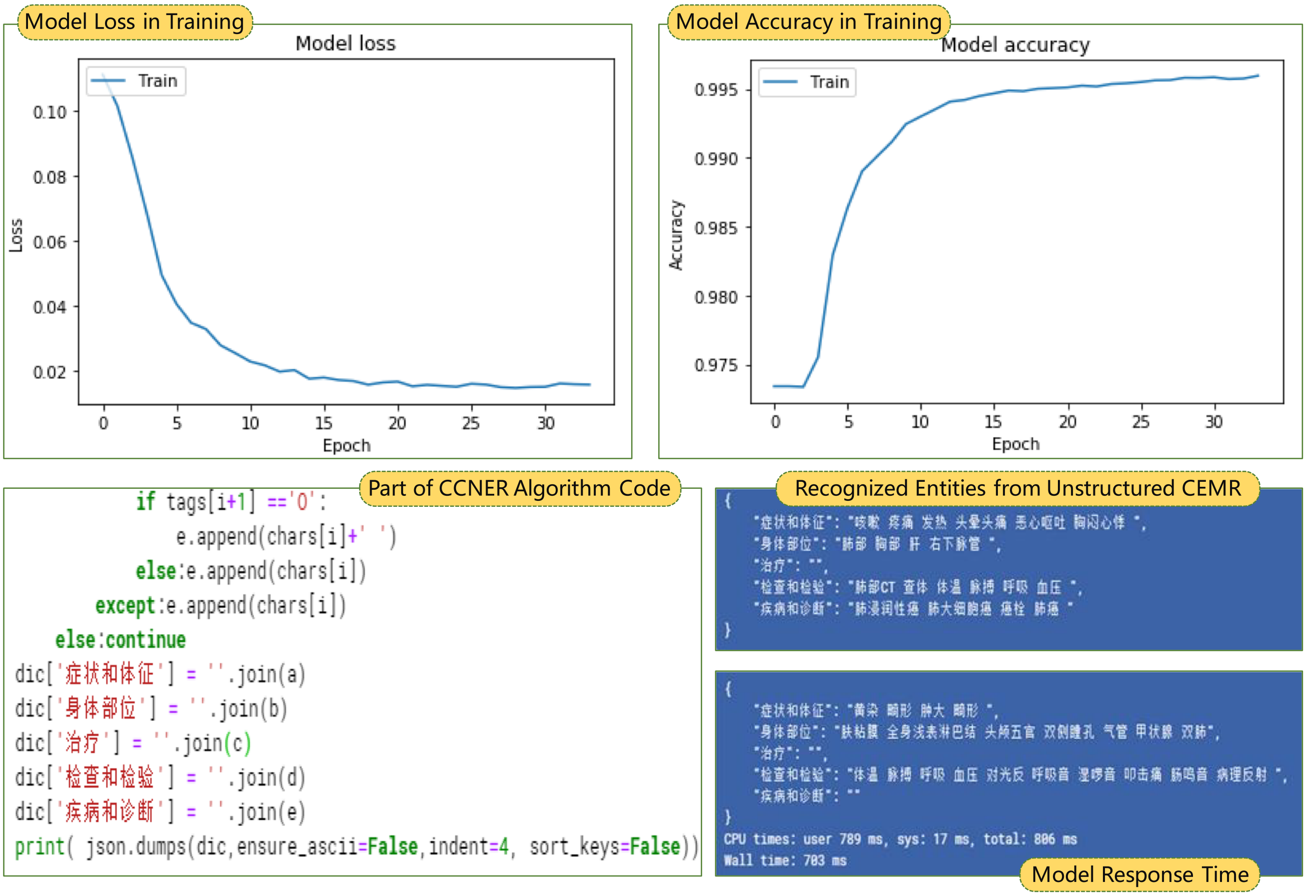
诊断检索：诊断名称 包含 '肺'
诊断检索：诊断名称 包含 '癌'
诊断检索：诊断名称 包含 '肺'
诊断检索：诊断名称 包含 '恶性肿瘤'

医院HIS资源统计

数据范围（按住院病人入院日期算）：1900-01-01 00:00:00 -- 2018-04-15 00:00:00
科室数量：0 职工人员数量：0 药品目录数量：0 药品剂型数量：0
药品用法数量：0 用药单位数量：0 医嘱频次数量：0
住院病人数量（按病人入院次数算）：当前机构:73695
医嘱数量：11623291（关联住院病人数量：11049880）
病历数量：554695（关联住院病人数量：529638）
护理记录数量：6663674（关联住院病人数量：6326671） 死亡信息数量：0（关联住院病人数量：0）
手术记录数量：90917（关联住院病人数量：89354） 诊断信息数量：2411006（关联住院病人数量：413234）
输科记录数量：26654（关联住院病人数量：25474） 产科产妇婴儿情况记录数量：0（关联住院病人数量：0）
收藏记录数量：0（关联住院病人数量：0） 收藏明细记录数量：0（关联住院病人数量：0）

	年龄	性别	症状	体温	脉搏	呼吸	血压
1	60	0	颈部、腰部疼痛，均为持续性钝痛，	36	85	20	114
2	52	1	咳嗽，呈阵发性单声咳，晚上咳嗽次	36.7	105	20	126
3	67	1	咳嗽	36.4	73	20	147
4	60	0	右侧胸痛伴呼吸困难，干咳	36.8	78	23	110
5	71	0	精神胃纳尚可，大小二便正常，体重无	36.8	74	20	130
6	78	1	反复出现咳嗽、咳痰，咳黄白色粘痰	36.6	93	20	108
7	82	1	膝关节乏力，腰骶部不适，耳鸣	36.2	64	20	133
8	57	1	反复咳嗽，咯血丝痰，量少，约每天	36.2	85	20	161
9	52	0	阵发性咳嗽，咳少许白色粘痰，活动	36.2	82	20	138
10	88	0	骨折疼痛	36.1	87	18	150
11	45	1	上腹部疼痛、疼痛呈持续性阵发性加	36.8	84	20	130
12	55	1	腹痛伴腹胀、反酸、烧心	36.3	90	18	120
13	53	1	伤处疼痛，伴有头晕，左眉弓上皮肤	36.1	82	20	140
14	70	0	右髋部疼痛	36	80	18	150
15	64	1	腹痛伴腹胀	36.3	77	20	135
16	57	1	反酸、烧心、腹痛伴腹胀	36.3	90	18	120

Fig. 7 Screenshots and training samples in the China Hospital Pharmacovigilance System



```
            if tags[i+1] =='0':
                e.append(chars[i]+' ')
            else:e.append(chars[i])
        except:e.append(chars[i])
    else:continue
dic['症状和体征'] = ''.join(a)
dic['身体部位'] = ''.join(b)
dic['治疗'] = ''.join(c)
dic['检查和检验'] = ''.join(d)
dic['疾病和诊断'] = ''.join(e)
print( json.dumps(dic,ensure_ascii=False,indent=4, sort_keys=False))
```

```
{
    "症状和体征": "咳嗽 疼痛 发热 头晕头痛 恶心呕吐 胸闷心悸 ",
    "身体部位": "肺部 胸部 肝 右下脉管 ",
    "治疗": "",
    "检查和检验": "肺部CT 查体 体温 脉搏 呼吸 血压 ",
    "疾病和诊断": "肺浸润性癌 肺大细胞癌 癌栓 肺癌 "
}
```

```
{
    "症状和体征": "黄染 畸形 肿大 畸形 ",
    "身体部位": "肤粘膜 全身浅表淋巴结 头颅五官 双侧瞳孔 气管 甲状腺 双肺",
    "治疗": "",
    "检查和检验": "体温 脉搏 呼吸 血压 对光反 呼吸音 湿啰音 叩击痛 肠鸣音 病理反射 ",
    "疾病和诊断": ""
}
CPU times: user 789 ms, sys: 17 ms, total: 806 ms
Wall time: 703 ms
```

Fig. 8 Recognized entities from unstructured CEMR and the model response time

data quality. Simple category information can be directly converted into numerical data. The pre-process of Chinese word segmentation is conducted using a tool named jieba (github.com/fxsjy/jieba). For text information (e.g., symptom description), text vectorization such as TF-IDF, FastText, and Word2Vec can be used. Here, TF-IDF (i.e., term frequency-inverse document frequency, github.com/jkern/tf-idf) is used for text vectorization. Because TF-IDF produces a sparse matrix that needs to be reduced to a dense matrix, the Truncated Singular Value Decomposition (SVD, github.com/ethz-asl/truncated_svd_solver) method is used to reduce the dimension of the sparse matrix.

The proposed algorithm is written based on Python3.5 and Keras deep learning framework (keras.io/), including libraries such as Numpy, Pandas, and Sklearn. The FastText algorithm (fasttext.cc/) in the Gensim library is called to train the professional word library word vector in CCNER. The hardware platform for training is a server with Intel Xeon CPU and Tesla P100 16G GPU.

B. *Experiments and discussions on CCNER*

A total of 13,786 labeled medical records are collected for conducting the experiments on the proposed CCNER modes. The widely-used "Bert + BiLSTM + CRF" (BBLC) model [40] is introduced as the benchmark. 11,000 records are used as a training set and the rest 2786 records as the test set. The accuracy of the CCNER model is evaluated by identifying whether the entity types and entity boundary is right or not. The confusion matrix is built for the distinction between false positive and negative errors. However, in the face of a large number of data, it is difficult to measure the model's merits just by counting the number. It will also take too much space to include the confusion matrix. Therefore, for a concise reason, this paper directly presents the results on the Precision rate, Recall rate, and the F1 index to evaluate the performance of CCNER models. Table 4 shows the results of experimental CCNER models.

The results show that the proposed HDL module (i.e., BiLSTM + Dilated Convolution +3D Attention + CRF + transfer learning + word vector + character vector + semi-supervised learning) work best among all CCNER models. The basic model BLDAC (i.e., BiLSTM + Dilated Convolution +3D Attention + CRF) is underperformed than the BBLC model because the neural network scale of the former model (20 million parameters) is smaller than the later model (130 million parameters). The performance of the BBLC model is lower than the "BLDAC + transfer learning + word vector + character vector" method because the preliminary training of Bert Chinese model uses general Chinese language text. Thus, the obtained character vector is underperformed than that of the "BLDAC + transfer learning + word vector + character vector" obtained in professional medical vocabulary. The combination of word vector and character vector can mine the corpus of semantic information in-depth. The semi-supervised learning technique can automate the annotation to a large number of unlabeled CEMRs, and thereby expanding the scale of training samples for the HDL module. The semi-supervised learning technique improves the basic model to enhance the ability to capture more semantic information, and thus achieve higher precision of entity extraction.

Table 5 provides the precision rate, recall rate, F1 results of the proposed HDL module under five categories of the entity identification of acute respiratory diseases. The results show that the recall rate of the *Disease* entity type is low. The main reason is that some entities of disease and diagnosis are not covered in the training set, leading to some diseases not being well recognized.

Table 4 shows the precision rate, recall rate, F1 index of the extraction of three diseases, namely lung cancer, severe pulmonary, and severe asthma, which are included in the *Disease* entity categories. The average precision rate, recall rate, and F1 values of these three respiratory diseases are mapped into the *Disease* entity row in Table 4. Based on the analysis of the results in Table 6, the reasons why the

Table 5 CCNER results for different entity types

Entity Type	Precision	Recall	F1
Examination	93.40	91.30	92.34
Symptoms	92.50	92.72	92.61
Disease	93.25	85.97	89.46
Anatomy	95.88	90.45	93.1
Treatment	95.57	91.46	93.47

Table 4 Comparison of experimental CCNER models

Models	Precision	Recall	F1
BBLC	88.40	86.71	87.55
BLDAC	81.45	79.60	80.51
BLDAC+transfer learning	90.20	87.42	88.78
BLDAC+transfer learning+word vector+character vector	92.42	88.55	90.44
HDL (BLDAC +transfer learning+word vector+character vector+semi-supervised learning)	94.12	90.38	92.21

Table 6 Identification results of three types of acute respiratory diseases

Disease Type	Precision	Recall	F1
Lung cancer	94.30	87.10	90.54
Severe pneumonia	92.50	85.14	88.67
Severe asthma	92.95	85.69	89.17

accuracy of lung cancer is better than that of severe pneumonia and severe asthma are explored. It is found that the data scale of lung cancer in the training set is much more than the other two categories. In the process of training, the model learns more about lung cancer.

As shown in Fig. 8, the information extraction speed of CCNER is 17 ms due to the fast speed of forward-reasoning of the HDL module, which lays a solid foundation for practical usage. The advantages of the proposed HDL module are three-fold:

a) Reduction in forward-reasoning time. The model parameters of the HDL module were 20 million, while the model parameters of the "Bert + BiLSTM + CRF" model were as high as 130 million. The response time of HDL was guaranteed to be 20 milliseconds when only using CPU for the forward reasoning, while the response time of the "Bert + BiLSTM + CRF" model was more than 1 s. The epochs threshold of training is 100, and it usually takes no more than 35 epochs to finish the training process. The stopping criteria for training the CCNER is that the loss value is less than 10^{-5} for five batches in a row.
b) Reduction in model training time. Bert Chinese model is trained three days on Chinese Wikipedia corpora by Google using four hundred TPU clusters. It will consume more than ten hours for finetuning one epoch of the "Bert + BiLSTM + CRF" model if using the Telas P100 GPU. The proposed HDL module needs only three hours to complete the training on the Telas P100 GPU.
c) Improvement in model accuracy. The HDL module achieves the highest F1 index among all the compared models because it introduces the transfer learning in the professional medical training corpus for obtaining prior knowledge, semi-supervised learning to enlarge the training dataset, and a combination of Chinese word vector and Chinese character vector for better capturing the semantics and context.

III. *Experiments and discussions on the risk classification model*

The extracted entities are combined with the existing structured data (Age, Gender, Temperature, Pulse, Respiratory Rate, Blood Pressure) to form the input data of the risk classification model. Especially, the extracted Disease and Anatomy are combined to form new dimensions of input data, namely, Past Medical History and the Family Medical History (if any) of the patient. Finally, the input data includes the Age, Gender, Temperature, Pulse, Respiratory Rate (RR), Blood Pressure (BP), Family Medical History (FMH), Past Medical History (PMH), Symptom, Examination, Treatment.

Table 7 Performance comparison of different model combinations

Type	Methods	Error rate	AUC	F1
BBLC-based	Logistic Regression	0.3563	0.7005	0.7465
	Support Vector Machine	0.2706	0.7292	0.7568
	Random Forest	0.2419	0.7534	0.7815
	Customized XGBoost	0.2131	0.7992	0.8075
HDL-based	Logistic Regression	0.2546	0.7378	0.7863
	Support Vector Machine	0.2234	0.7681	0.8077
	Random Forest	0.1623	0.8357	0.8548
	Customized XGBoost	0.1012	0.8639	0.8927

The performance of risk classification models combined with different CCNER approaches is evaluated to see if the extracted information is meaningful. This disease risk classification model uses three indicators, namely, error rate, F1 index, and Area Under Curve (AUC), to comprehensively evaluate the performance of these models, including the Logistic Regression [45], Support Vector Machine [46], Random Forest [47], and the Customized XGBoost. The F1 value and AUC are two indicators that reasonably reflect the performance of model classification and generalization. AUC could avoid the evaluation deviation caused by the unbalanced datasets [48]. A total of 1235 pieces of data were labeled with the risk level. The data were divided into 1000 pieces of training&verification set and 235 pieces of the test set, respectively. To avoid the influence of data variance and data distribution, ten-fold cross-validation was adopted to split the training set and verification set, and stratified sampling was carried out for the training set.

Table 7 provides an overview of the performance of these four models before and after the proposed HDL model is used. The lower the error rate, the better the performance of the risk classification model.

The results show that the performance of these four models improves a lot after the proposed HDL model is used, and the error rate has been lowered a lot. Secondly, the customized XGBoost is better compared with Logistic Regression, Support Vector Machine, and Random Forest. Thirdly, the customized XGBoost could mine the implicit information by a feature-importance ranking mechanism. The advantage of using the XGBoost algorithm is that it is relatively straightforward to get the importance score for each feature after the decision tree is created. Feature importance is calculated by the number of improved performance measures at each feature split point in a single decision tree, with nodes responsible for weighting and recording times. The greater the performance measure of a feature to improve the split point (the closer to the root node), the greater the weight. The more selected by the decision tree, the more important the feature becomes [29]. Finally, the weighted sum of the results of a feature in all the decision trees is averaged to obtain the importance score. The results of the ranking of factors influencing the risk level using the XGBoost show that symptoms (0.31), age (0.17), body temperature (0.18), FMH (0.16), and PMH (0.16) have a significant impact on a patient's disease risk. Clinicians need to pay more attention to these patients and give timely treatment to prevent disease progression.

Table 8 further shows the results of risk classification based on HDL and Customized XGBoost in different acute respiratory disease subgroups, namely, lung cancer, severe pulmonary, and severe asthma. Comparing the last row data in Table 7 with Table 8, the performance in discriminating patients' risk levels with different risks within a disease subgroup is slightly lower than that in the holistic model (mixing all three disease subgroups). In another word, the good performance of the proposed model benefits from the outstanding capability in discriminating patients with different acute respiratory diseases, as well as discriminating patients' risk levels within a disease subgroup. The reason is that although the data modeling of acute respiratory disease data is unified in this study, the risk factors in different disease subgroups are different, which results in difficulty in generalizing different underlying risk classification patterns into a single neural network. Establishing multiple risk classification models for different acute respiratory disease subgroups may be a good choice in the practical implementation of the proposed model.

IV. *Discussions on the integrated system*

For the unstructured text-type CEMR data, the critical information is extracted through named entity recognition, which lays a foundation for collecting more related data for the follow-up risk classification task. Since the PMH and FMH that contain the extracted Disease and Anatomy are highly correlated with the risk (0.55 for PMH and 0.51 for FMH), the additional information extracted by the CCNER is critical for the risk classification.

Table 8 Risk classification in different disease subgroups

Disease Type	Error rate	AUC	F1
Lung cancer	0.1011	0.8576	0.8940
Severe pneumonia	0.1028	0.8380	0.8597
Severe asthma	0.1041	0.8058	0.8315

Early warning of acute respiratory diseases is critical for preventing lung cancer and severe asthma. Integrating cutting-edge deep learning algorithms, this study automates the accurate extraction of critical information from unstructured medical data for risk classification of severe respiratory diseases. It provides technical support for relieving them from heavy work intensity in analyzing the unstructured medical records. Moreover, this study also discovered the critical factors related to the risk of severe respiratory diseases, which provides clinicians with scientific references for diagnosis and treatment decisions. With the increase of follow-up CEMR data, the classification model will become more accurate for the disease risk classification.

7 Conclusions

This paper proposed a hybrid artificial intelligence system to extract and analyze multiple types of Chinese clinical data for the risk classification of acute respiratory diseases. The contribution of this paper is that a dedicated design of the "BiLSTM+Dilated Convolution+3D Attention+CRF" deep learning model is proposed to extract entities from unstructured medical data, which achieves higher accuracy and efficiency in the CCNER task than the popular "Bert+BiLSTM+CRF" approach. The cutting-edge artificial intelligence techniques, including transfer learning and semi-supervised learning, are introduced to improve the accuracy of named entity recognition. Combining the extracted entity data with other structured data in the CEMRs, a customized XGBoost is used to predict the risk of respiratory disease. The empirical study shows that the proposed model could provide practical technical support for improving diagnostic accuracy. Our study provides a proof-of-concept for implementing a hybrid artificial intelligence-based system as a tool to aid clinicians in tackling CEMR data and enhancing the diagnostic evaluation under diagnostic uncertainty.

With the increase of medical data, the proposed hybrid artificial intelligent system will be more accurate for disease risk classification. Here are three directions for further research. Firstly, the data modeling of the risk classification model for acute respiratory diseases is not comprehensive enough, and more dimensions of medical records should be introduced so that the model can learn more features and make a more accurate classification. Secondly, either in the task of named entity recognition of medical data or in the risk classification of disease, the amount of training data needs to be further enlarged to achieve better classification accuracy. Thirdly, in the information age, the privacy issue is a social concern. Incorporating the secure multi-party computing technique into the transfer learning based on the encrypted CMER data from the distributed database is also a potential research direction. Fourthly, more research effort will be paid in knowledge discovery of disease diagnosis based on a preliminary analysis of the correlation between influencing factors of patients' disease.

Acknowledgments The deep learning module for Chinese clinical named entity recognition presented in this paper is developed based on our practices in winning two silver medals from the Kaggle Quora Insincere Questions Classification in 2019 and Kaggle Jigsaw Multilingual Toxic Comment Classification in 2020. We want to express gratitude to the Kaggle competition organizer (https://www.kaggle.com/) for providing excellent hardware to train the neural network.

Funding This work was supported by the Guangzhou Municipal Science and Technology Bureau under Grant No. 201803010063; the Chengdu Science and Technology Program under Grant No. 2019-YF05-02050-GX; and the Science and Technology Planning Project of Guangdong Province of China under Grant No. 2019A050503010.

References

1. Perrotta DM, Decker M, Glezen WP (1985) Acute respiratory disease hospitalizations as a measure of impact of epidemic influenza. Am J Epidemiol 122:468
2. Mansmann S, Ur Rehman N, Weiler A, Scholl MH (2014) Discovering OLAP dimensions in semi-structured data. Inf Syst 44:120
3. Wong ZSY, Zhou J, Zhang Q (2019) Artificial intelligence for infectious disease big data analytics. Infect Dis Health 24:44
4. Hillestad R, Bigelow J, Bower A, Girosi F, Meili R, Scoville R, Taylor R (2005) Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Aff (Millwood) 24:1103
5. Sweeney L (1996) Replacing personally-identifying information in medical records, the scrub system. Proc AMIA Annu Fall Symp 333
6. Sarker A, Mollá D, Paris C (2016) Query-oriented evidence extraction to support evidence-based medicine practice. J Biomed Inform 59:169
7. Mohamadou Y, Halidou A, Kapen PT (2020) A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. Appl Intell 50:3913
8. Zhang H, Zhang H, Pirbhulal S, Wu W (2020) Albuquerque V.H.C.D.: Active Balancing Mechanism for Imbalanced Medical Data in Deep Learning–Based Classification Models. ACM Trans Multimed Comput Commun Appl 16:1
9. Chiriac AM, Wang Y, Schrijvers R, Bousquet PJ, Mura T, Molinari N, Demoly P (2018) Designing predictive models for Beta-lactam allergy using the drug allergy and hypersensitivity database. The journal of allergy and clinical immunology. In Practice 6:139
10. Maxwell A, Li R, Yang B, Weng H, Ou A, Hong H, Zhou Z, Gong P, Zhang C (2017) Deep learning architectures for multi-label classification of intelligent health risk prediction. BMC Bioinformatics 18:523
11. Sumathi S, Beaulah HL, Vanithamani R (2014) A wavelet transform based feature extraction and classification of cardiac disorder. J Med Syst 38:98
12. Hira S, Bai A, Hira S (2021) An automatic approach based on CNN architecture to detect Covid-19 disease from chest X-ray images. Appl Intell 51:2864
13. Pham T, Tran T, Phung D, Venkatesh S (2017) Predicting healthcare trajectories from medical records: a deep learning approach. J Biomed Inform 69:218
14. Liu M, Zhang J, Lian C, Shen D (2020) Weakly supervised deep learning for brain disease prognosis using MRI and incomplete clinical scores. IEEE Trans Cybern 50:3381
15. Zheng N, Du S, Wang J, Zhang H, Cui W, Kang Z, Yang T, Lou B, Chi Y, Long H, Ma M, Yuan Q, Zhang S, Zhang D, Ye F, Xin J (2020) Predicting COVID-19 in China using hybrid AI model. IEEE Trans Cybern 50:2891
16. Panwar M., Biswas D., Bajaj H., Jobges M., Turk R., Maharatna K., Acharyya A.: Rehab-Net: Deep learning framework for arm movement classification using wearable sensors for stroke rehabilitation. IEEE Trans Biomed Eng 66. 3026 (2019)
17. Upadhyay J, Tiwari N, Rana M, Rana A, Durgapal S, Bisht SS (2019) Pathophysiology, etiology, and recent advancement in the treatment of congenital heart disease. J Indian Coll Cardiol 9:67
18. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG (2018) Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet 50:1171
19. Celi L.A., Davidzon G., Johnson A.E., Komorowski M., Marshall D.C., Nair S.S., Phillips C.T., Pollard T.J., Raffa J.D., Salciccioli J.D., Salgueiro F.M., Stone D.J.: Bridging the Health Data Divide. J Med Internet Res 18. e325 (2016)
20. Lian C, Liu M, Pan Y, Shen D (2020) Attention-guided hybrid network for dementia diagnosis with structural MR images. IEEE Trans Cybern. PP
21. Hsieh N, Hsieh N, Hung L, Hung L, Shih C, Shih C, Keh H, Keh H, Chan C, Chan C (2012) Intelligent postoperative morbidity prediction of heart disease using artificial intelligence techniques. J Med Syst 36:1809
22. Phegley JW, Perkins K, Gupta L, Hughes LF (2005) Multicategory prediction of multifactorial diseases through risk factor fusion and rank-sum selection. IEEE Trans Syst Man Cybern Syst Hum 35:718
23. Hewson PJ, Bailey TC (2010) Modelling multivariate disease rates with a latent structure mixture model. Stat Model 10:241
24. Huang Z, Dong W, Duan H, Liu J (2018) A regularized deep learning approach for clinical risk prediction of acute coronary syndrome using electronic health records. IEEE Trans Biomed Eng 65:956
25. Hao Y, Usama M, Yang J, Hossain MS, Ghoneim A (2019) Recurrent convolutional neural network based multimodal disease risk prediction. Futur Gener Comput Syst 92:76
26. Wang T, Qiu RG, Yu M, Zhang R (2020) Directed disease networks to facilitate multiple-disease risk assessment modeling. Decis Support Syst 129:113171
27. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, Cai W, Kermany DS, Sun X, Chen J, He L, Zhu J, Tian P, Shao H, Zheng L, Hou R, Hewett S, Li G, Liang P et al (2019) Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med 25:433
28. Ohsaki M, Abe H, Tsumoto S, Yokoi H, Yamaguchi T (2007) Evaluation of rule interestingness measures in medical knowledge discovery in databases. Artif Intell Med 41:177
29. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 785. Association for Computing Machinery,San Francisco California USA
30. Chang C, Hsu C, Lui S (2003) Automatic information extraction from semi-structured web pages by pattern discovery. Decis Support Syst 35:129

31. Leng J, Jiang P (2016) A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm. Knowl-Based Syst 100:188
32. Liu Z, Tang B, Wang X, Chen Q (2017) De-identification of clinical notes via recurrent neural network and conditional random field. J Biomed Inform 75:S34
33. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention Is All You Need. In: Advances in Neural Information Processing Systems 30, vol 5998
34. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R (2019) Transformer-xl: attentive language models beyond a fixed-length context
35. Wei Y, Xiao H, Shi H, Jie Z, Feng J, Huang TS (2018) Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 7268. Salt Lake City, USA. June 18–22, 2018
36. Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B, Rueckert D (2018) Attention u-net: learning where to look for the pancreas
37. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding
38. Sun Q, Liu Y, Chua T, Schiele B (2019) Meta-Transfer Learning for Few-Shot Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), vol. 403. California, USA. June 16–20, 2019
39. Qiu J, Zhou Y, Wang Q, Ruan T, Gao J (2019) Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field. IEEE Trans Nanobiosci 18:306
40. Li X, Zhang H, Zhou X (2020) Chinese clinical named entity recognition with variant neural structures based on BERT methods. J Biomed Inform 107:103422
41. Leng J, Jiang P (2017) Mining and matching relationships from interaction contexts in a social manufacturing paradigm. IEEE Trans Syst Man Cybern Syst 47:1
42. Muhammad K, Khan S, Ser JD, de Albuquerque V (2020) Deep learning for multigrade brain tumor classification in smart healthcare systems: a prospective survey. IEEE Trans Neural Netw Learn Syst. PP
43. Karnofsky. (2008) Karnofsky performance score. Encyclopedia of Cancer. Springer, Berlin
44. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T (2017) Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, vol. Long Beach, CA, USA. Dec 4–9, 2017
45. Yang L, Qian Y (2016) A sparse logistic regression framework by difference of convex functions programming. Appl Intell 45:241
46. Mehmood Z, Mahmood T, Javid MA (2018) Content-based image retrieval and semantic automatic image annotation based on the weighted average of triangular histograms using support vector machine. Appl Intell 48:166
47. Kim S, Jeong M, Ko BC (2021) Lightweight surrogate random forest support for model simplification and feature relevance. Appl Intell 10
48. Leng J, Chen Q, Mao N, Jiang P (2018) Combining granular computing technique with deep learning for service planning under social manufacturing contexts. Knowl-Based Syst 143:295

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