2018 IEEE International Conference on Systems, Man, and Cybernetics

# Utilizing Narrative Text from Electronic Health Records for Early Warning Model of Chronic Disease

Jie Meng, Runtong Zhang*, *Senior Member*, *IEEE*, Donghua Chen
School of Economics and Management
Beijing Jiaotong University
Beijing, China
16120614@bjtu.edu.cn, rtzhang@bjtu.edu.cn, 15113181@bjtu.edu.cn

***Abstract*—Chronic diseases are associated with high morbidity and mortality, and they cannot be quickly determined and identified during their early stages. Therefore, early identification and warning of chronic diseases are important. This study proposed an early warning model (EWM) based on electronic health records (EHRs). The model included comprehensive methods that identify whether the patient is expected to suffer from chronic disease and provide early warning based on the undiagnosed narrative text of EHRs. A professional medical terminology library called Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) was utilized to improve the accuracy of early warning. The model utilized semantic relationships in the SNOMED CT to expand the effects of related terms of renal cancer. We utilized 1,300 medical records and more than 1,400 progress notes in the EHRs in our experiments. The proposed EWM of renal cancer achieved a precision of 90%, recall of 91%, and F-measure of 91%.**

***Keywords—electronic health record; biomedical named entity recognition; Systematized Nomenclature of Medicine—Clinical Terms; early warning model; convolution neural network.***

## I. INTRODUCTION

Electronic health records (EHRs) contain a large amount of medical information that are important data sources used for disease prediction [1], personalized information recommendation [2], clinical decision support [3], and drug discovery [4]. However, the valuable medical knowledge in EHRs are difficult to automatically obtain, express, update, and use in computer-aided analysis. The main reasons [5] are expressed as follows. On one hand, EHRs contain structured data and also semi-structured and non-structured narrative information. Many studies have focused on the analysis of medical structured data, in which the narrative texts in EHRs are ineffectively and incompletely utilized [6]-[7]. On the other hand, EHRs from different medical institutions may contain various spelling errors, abbreviated medical terms, and different narrative styles in different medical areas. However, the narrative text still preserves extremely detailed diagnosis and treatment information of patients. Therefore, biomedical named entity recognition is first conducted before analyzing these texts. Biomedical named entity recognition has attracted considerable attention in the identification of drugs [8], genes, and proteins [9]. Research methods of biomedical named entity recognition have mainly focused on the improvement of machine learning [10]-[11].

Studies on early warning and risk prediction of diseases are developed earlier than the utilization of EHRs. These studies are frequently based on the experimental data and vital sign data directly obtained by doctors [12]-[13]. However, researchers can achieve an early warning model (EWM) for specific diseases with the development of EHRs [14]-[15]. Several researchers have investigated the use of topic modeling [16]-[17] and convolution neural network (CNN) [18] to aid in the classification of patients. A chronic disease is considered as a disease with high morbidity and mortality, high-risk, and high-frequency clinical adverse events [19]. However, medical experts cannot frequently identify chronic diseases in their early stages. Therefore, the patients with these diseases are unable to receive proper healthcare services. Thus, early warning and recognition of chronic disease for patients can promote the rapid initiation, detection, and professional aid of better treatment.

The rest of this paper utilized a large amount of narrative text of EHRs related to chronic patients to address the aforementioned issues. We establish a novel EWM of chronic disease to improve the accuracy of early warning of disease. Section II described the methods and algorithms of EWM. Section III evaluated the performance of EWM through experiments. Section IV provided the conclusion.

## II. MATERIALS AND METHODS

The detailed methods of our EWM are expressed as follows. The input data of the model are the undiagnosed narrative text from EHRs of the patient. First, we establish a biomedical terminology library of single disease. Then, we extract and analyzed the medical terms from the single-disease EHRs. We establish a Chinese–English medical terminology mapping table that maps the Chinese medical concepts to English concepts to apply the semantic relationships from Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) in analyzing the text. Subsequently, we implement knowledge discovery based on SNOMED CT to acquire the biomedical terms within the text where the biomedical terms extracted from the narrative text are combined with the medical terms in SNOMED CT; these terms form the knowledge-based text for

---

This work was partially supported by a key project of National Natural Science Foundation of China under grant number 71532002 and the Fundamental Research Funds for the Central Universities under grant number 2017YJS069.
*Corresponding author: Prof. Runtong Zhang, email: rtzhang@bjtu.edu.cn

2577-1655/18/$31.00 ©2018 IEEE
DOI 10.1109/SMC.2018.00713

EWM. With the aim of achieving knowledge-based text, we establish a CNN-based text classification model to classify knowledge-based text into two classes, namely, ill or non-ill. The patient is warned when the model determines that he/she may be suffering from the disease. Figure 1 shows the overall methods of the EWM. The steps of the experimental method will be explained in detail below.

### *A. Establishment of biomedical terminology library*

The experimental data used in this study are Chinese EHR texts. Usually, the biomedical named entity recognition is based on medical terminology library. However, no open-source Chinese professional medical terminology library currently exists. Therefore, we develop a biomedical terminology library before applying biomedical named entity recognition in the text.

The biomedical terminology library included two parts. The first part is the basic terminology set where a Chinese version of the International Classification of Diseases 10th version (ICD-10) is established because many Chinese hospitals currently use ICD-10 to exchange medical terminologies in different medical systems.

The second part is the single-disease professional terminology set. We develop a professional biomedical terminology library for single disease through the Conditional Random Field (CRF) method. The training corpus contains the previous basic terminology set when the CRF is used for model training. The key algorithm of CRF for building the library is shown in Algorithm 1.

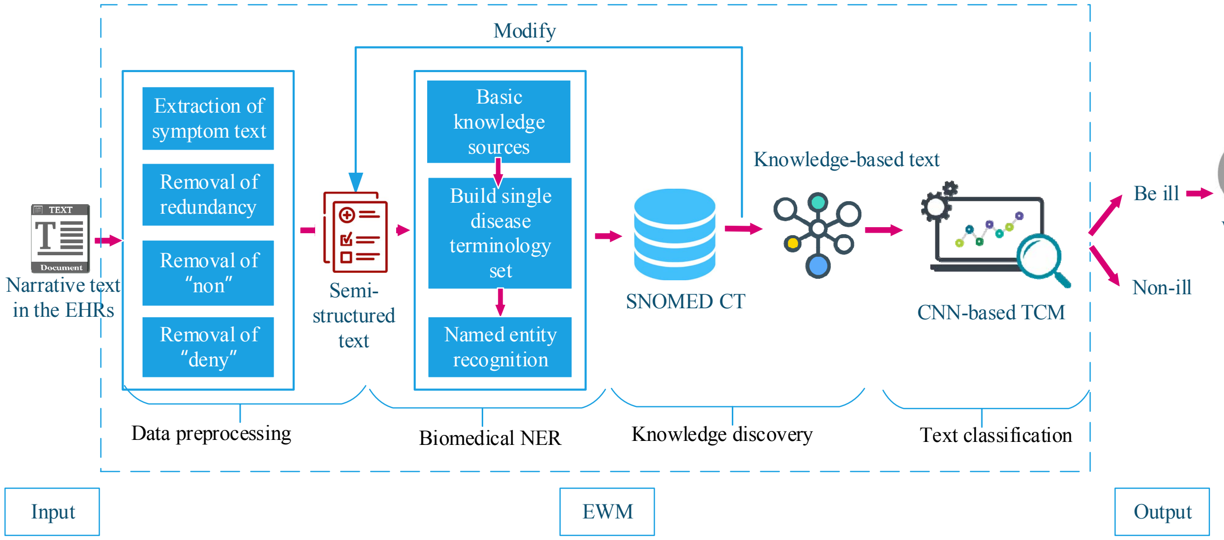

Modify
Extraction of symptom text
Removal of redundancy
Removal of "non"
Removal of "deny"
TEXT
Document
Narrative text in the EHRs
Semi-structured text
Basic knowledge sources
Build single disease terminology set
Named entity recognition
SNOMED CT
Knowledge-based text
CNN-based TCM
Be ill
Warning
Non-ill
Data preprocessing
Biomedical NER
Knowledge discovery
Text classification
Input
EWM
Output

Fig. 1. Overview of our EWM

**Algorithm I.** CRF Algorithm for building the SDPTS

```
Input: Training set TS; Test set TS*
Output: STS
Let $W$ be a list of words in a text t:$\{ w_i | i = 1,2,\dots,N_w\}$
Let $T$ be a list of TS
Let C and C* be a list of characters in a word of TS and TS*, respectively
Let NER be a function for assigning location tags and part-of-speech tags for character
Let NR be the results of NER
Let STS be a list of segmented TS*
for each $c$ in C where $c_i \in w_i \in t_k \in$ TS, $i = 1,2,\dots,N_c$; $j = 1,2,\dots,N_w$ and $k = 1,2,\dots,N_t$ do
    $r$ = ($c$, part-of-speech tag, location tag) = NER($c$)
    $r \rightarrow$NR
CRF Model = CRF_Learn (NR)
tagged TS* = CRF_Test c* ∈ TS*
STS = exacting terms from TS* based on tags of c*
return STS
```

In Algorithm I, the text is labeled with tags, including location and part-of-speech tags. We use a Begin/End labeling method for location tagging. This method describes the components of the lexical chunk based on the relative position information of the words in a group of blocks. Five labels are used to represent a block, as follows: B—at the beginning, the block is made up of two or more characters. E—end, the block is made up of two characters or more. M—the block is composed of three or more characters. S—word entity, the current character is a complete block. O—non-entity, the current character is outside of any block. For part-of-speech tag, we define five types, namely, punctuation (PUNC), time (TIM), number (NUM), preposition (P), and noun (N).

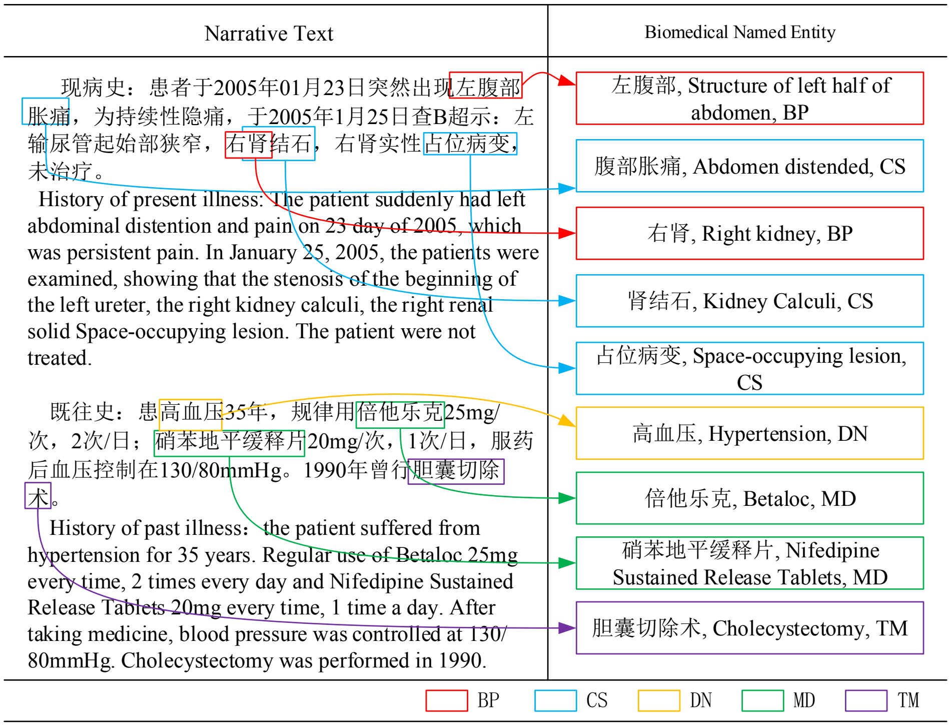
After obtaining a segmented text, we can filter out other words with only nouns remaining and count the word frequency of nouns. Then, words with high frequency are selected, which are highly correlated with the single disease. Subsequently, we define a category for the entities. In our study, the body parts, disease names, medicine names, treatments, and clinical symptoms are recognized in preparing the EWM. We define BP as a body part, such as the right kidney, DN as a disease name, such as kidney cancer, MN as a medicine name, such as Aspirin, TM as treatment, such as biopsy, and CS as clinic symptoms, such as uneven density. All the terms are marked with tags, which compose our biomedical terminology library combined with the biomedical terminology set.

## *B. Biomedical named entity recognition*

All biomedical named entities of EHR texts can be identified based on the aforementioned structured biomedical terminology library. This study used the Chinese Words Segmentation tool "Jieba," which uses dynamic programming to determine the maximum probability path and maximum combined segmentation based on the word frequency. However, this tool cannot be accurately segmented for medical text. Therefore, a custom dictionary that contains the biomedical terminology library built in the previous step should be used in segmenting words, which can accurately extract the medical terms in the text and classify them properly. Figure 2 shows an example of biomedical named entity recognition.

## *C. SNOMED CT-based knowledge discovery*

We obtain the biomedical named entities from the text based on the previous steps. However, most of them are Chinese medical terms. To perform knowledge discovery by utilizing the semantic network in SNOMED CT that does not have a general Chinese version, We establish a Chinese–English medical terminology mapping table (CE-MTMT), which is used to gain the English terms from the extracted Chinese medical terminology. We use a tool named MetaMap [20] to automatically correct and map the medical terms from English concept to the term that is most similar and most likely in SNOMED CT. We used the semantic relationships of SNOMED CT that connect the concepts between the same level and different hierarchical structures, including the IS-A and attribute relationships. Then, we obtained the terms related to medical terms, which strengthen the information on single chronic disease in the text and make the characteristics of single chronic disease obvious. An example of knowledge discovery is shown in Figure 3.


Narrative Text
Biomedical Named Entity

现病史：患者于2005年01月23日突然出现左腹部胀痛，为持续性隐痛，于2005年1月25日查B超示：左输尿管起始部狭窄，右肾结石，右肾实性占位病变，未治疗。

History of present illness: The patient suddenly had left abdominal distention and pain on 23 day of 2005, which was persistent pain. In January 25, 2005, the patients were examined, showing that the stenosis of the beginning of the left ureter, the right kidney calculi, the right renal solid Space-occupying lesion. The patient were not treated.

既往史：患高血压35年，规律用倍他乐克25mg/次，2次/日；硝苯地平缓释片20mg/次，1次/日，服药后血压控制在130/80mmHg。1990年曾行胆囊切除术。

History of past illness：the patient suffered from hypertension for 35 years. Regular use of Betaloc 25mg every time, 2 times every day and Nifedipine Sustained Release Tablets 20mg every time, 1 time a day. After taking medicine, blood pressure was controlled at 130/80mmHg. Cholecystectomy was performed in 1990.

左腹部, Structure of left half of abdomen, BP
腹部胀痛, Abdomen distended, CS
右肾, Right kidney, BP
肾结石, Kidney Calculi, CS
占位病变, Space-occupying lesion, CS
高血压, Hypertension, DN
倍他乐克, Betaloc, MD
硝苯地平缓释片, Nifedipine Sustained Release Tablets, MD
胆囊切除术, Cholecystectomy, TM
BP
CS
DN
MD
TM

Fig. 2. Example of biomedical named entity recognition

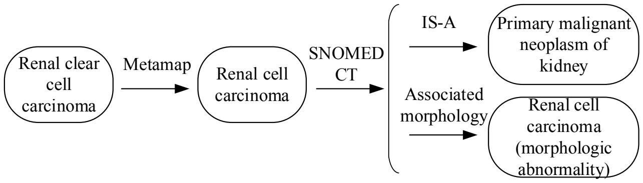

Renal clear cell carcinoma
Metamap
Renal cell carcinoma
SNOMED CT
IS-A
Primary malignant neoplasm of kidney
Associated morphology
Renal cell carcinoma (morphologic abnormality)

Fig. 3. An example of knowledge discovery based on SNOMED CT

## *D. CNN-based text classification model*

We apply the CNN in establishing the text classification model. Compared with traditional text classification techniques, we use distributed metrics to represent the text classification model based on CNN. In the classification stage, the CNN, which is a deep learning model, can automatically extract and learn sentence features on the basis of word vector, which reduces the dependence on artificial feature selection and optimizes the effect of feature selection. The detailed steps are expressed as follows:

### *1) Text preprocessing*

The main task of knowledge-based text preprocessing focuses on the word segmentation of the text. This process repeats the process of word segmentation and segments $T$. This process establishes the corresponding dictionary $D$ at the same time, uses word2vec model to initialize the word representation in $D$, and adds it to $D$. Then, $T$ is used as the input of the participle tool, output $T$ is divided into $D$ words $w_1$, $w_2$, …, $w_d$, and $T$ is expressed as $T: \{w_1, w_2, \dots, w_d\}$.

### *2) Input layer*

The input layer is a matrix that corresponds to text classification. Each line of the matrix represents the vector that corresponds to each word. The row number is the number of words, whereas the number of columns is the dimension of the vector. Specifically, the word representation of each word $w_i$ of $T$ in (1) is obtained from $D$, and $w_i$ in (2) is expressed as a vector $w_i\colon (m_{i1}, m_{i2}, \dots, m_{iv})$ $i \in [1，d]$ in space $R^{v\times 1}$. The matrix of $T$ indicates that the word expression of all the words in $T$ is arranged from the upper to the lower order in accordance with the sequence of words. $T$ is converted to $R^{d\times v}$ featured space, which is expressed as follows:

$$T = \begin{bmatrix} m_{11} & \cdots & m_{1v} \\ \vdots & \ddots & \vdots \\ m_{d1} & \cdots & m_{dv} \end{bmatrix}. \tag{1}$$

### *3) Convolution layer*

The feature matrix of $T$ in (1) is used as input, and the convolution kernel is consistent with the dimension of row vector. Meanwhile, the movement in the column direction can obtain the inherent characteristics of the original text. This feature is determined by the grammatical features of the original text without manual intervention. Thus, the automatic feature extraction of the text is realized.

In the convolution layer, multiple convolution kernels are established to comprehensively obtain the text feature and reduce the possibility of feature extraction. $C \in R^{k\times v}$ is an arbitrary convolution kernel, which is expressed as follows:

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1v} \\ \vdots & \ddots & \vdots \\ c_{k1} & \cdots & c_{kv} \end{bmatrix}. \tag{2}$$

The operation of the convolution layer uses $d - k + 1$ submatrices $T_p \in R^{k\times v}$ of feature matrix $T$ for the EHRs from upper to lower, $p \in [1，\mathrm{d} - \mathrm{k} + 1]$, as shown in (3), to perform (*) operations with C in (2).

$$T_p = \begin{bmatrix} m_{p1} & \cdots & m_{pv} \\ \vdots & \ddots & \vdots \\ m_{p+k-1,\ 1} & \cdots & m_{p+k-1,\ v} \end{bmatrix}. \tag{3}$$

The (*) operation is defined as

$A_{m\times n} * B_{m\times n} = sigmoid\left(\sum_{i,\ j\in(1,\ m)}(a_{ij} \times b_{ij} + bias_{ij})\right)$.(4)

$bias_{ij}$ is a component of bias matrix $BIAS$, that is, the deviation value that adjusts the weight matrix. The sigmoid is an activation function used to normalize the eigenvalues. Thus, the expression of $T$ and any convolution kernel C are all performed by *d-k+1* times (*) operation, and the output of each (*) operation is sequentially arranged. Convolution matrix $S \in R^{(d-k+1)\times 1}$ is obtained. Similarly, $T$ is convoluted with the convolution kernel of the remaining $m - 1$, and output $S_1$, $S_2, \dots, S_m$ of all convolution layers is obtained and passed in the pool layer of the CNN.

The convolution layer of CNN is a feature extraction process. A characteristic of a convolution kernel is extracted, and a characteristic matrix is obtained. CNN performs the same information transformation through the same convolution to the different regions of the original input, generalizes the local features, and preserves the entire characteristics when extracting a feature.

### *4) Pool layer*

After the convolution operation of $m$ convolution kernel, the expression of T is transmitted from the convolution layer to the matrix expression of $m$ $R^{(d-k+1)\times 1}$ space. These features are aggregated by the pool layer to simplify the expression of the features. The pooling operation is defined in the pool layer, which is expressed as

$$pooling\left(S_{(d-k+1)\times 1}\right) = \alpha(s_1, s_2, \dots, s_{d-k+1}). \tag{5}$$

$\alpha$ is selected as the maximization operator in this study, that is, the maximum element of the convolution layer vector is selected as the eigenvalue. In the pool operation, each convolution matrix is converted to 1D eigenvalue. Thus, feature vector $P\colon (P_1, P_2, \dots, P_m)$ of a $R^{m\times 1}$ space is obtained. $P$ is the feature vector after pooling. The pooling process realizes the extraction of features and also reduces the dimension of the feature and improves the classification efficiency.

### *5) Output layer*

The output layer is connected to the pool layer, pool layer vector $P$ is used as the input, a softmax classifier is used to classify vector $P$, and the final classification results are the output.

### *6) Model training*

The model uses a backward propagation algorithm to train the parameter values. First, error $E$ between the output value of the model and the expected value is defined as

$$E = \tfrac{1}{2}(Y - O)^2, \tag{6}$$

where, $Y$ and $O$ have 1D values. The upper layer is expanded to the pool layer:

$$E = \tfrac{1}{2}[y - f_s(P)]^2 = \tfrac{1}{2}[y - f_s(\textstyle\sum_{j=1}^{m} f_p(S_j))]^2. \tag{7}$$

In (7), $f_s$ is the activation function used by the classifier, and $f_p$ is the activation function of the pool layer. The upper layer is expanded to the coiling layer:

$$E = \tfrac{1}{2}\{y - f_s[\textstyle\sum_{j=1}^{m} f_p(\bigcap_{t\in(1,\ d-k+1)} S_t)]\}^2. \tag{8}$$

Among the functions, the expansion of $S_t$ is expressed as

$$S_t = f_s[\textstyle\sum_{i\in(1,\ d),h\in(1,\ v)}(c_{ih} tp_{ih} + bias_{ih})]. \tag{9}$$

The training principle of the model adjusts the weights and deviations to reduce $E$. The solution is by determining the partial derivatives of $E$ to $c_{ij}$ and $bias_{ij}$ separately and obtain the descending gradient of $c_{ij}$ and $bias_{ij}$. Then, we obtain the optimal convolution kernel and bias matrix through a gradient descent method to realize the classification model of narrative text from EHRs.

## III. EXPERIMENT

### A. Data preparation

The experimental data used in our study are obtained from the renal cancer EHRs in our cooperative Chinese hospital. This dataset contains 1300 medical records and more than 1400 progress notes in the EHRs. Among these data, we select the medical narrative text that contains the necessary patients' information, including the complaints and history of present illness of patients, descriptions of disease progression, and discussions of medical experts from the progress notes.

The important patient information in the EHRs include the textual description of the disease and symptoms of patients. The information contain several important modifiers that should be considered. The results of extracted medical knowledge from the text may be incorrect when these modifiers are not recognized. Thus, the disease-related terms and their adjuncts should be identified in processing the text in the EHRs. To ensure the accuracy of the EWM of disease, redundant sentences and negative sentences, such as "no" and "denial," should be removed from the EHR.

### B. Evaluation on the library of renal cancer

We select 1000 medical records and 1400 progress notes to build the biomedical terminology library. The data of the CRF experiment includes 700 records for the training set and 300 records for the test set.

In total, 700 records are randomly selected as the training set of CRF. The biomedical terminology set is added, including ICD-10, Chinese herbal medicine, Chinese traditional medicine, and Western medicine. The CRF model was obtained after training and was used to segment the test set to obtain a segmented text. Then, the model was compared with the text marked by the doctor. The metrics of text split by CRF are shown in Table I.

TABLE I EVALUATION METRICS OF TEXT SPLITTING BY CRF

| Metric | Value | Metric | Value | Metric | Value |
|---|---|---|---|---|---|
| Number of standard words | 50990 | Number of correct words | 50455 | Number of mistaken words | 990 |
| Recall | 98.95% | Precision | 98.08% | F Measure | 98.51% |

Table I indicates that the accuracy of the CRF is high in terms of the precision metric. Therefore, we split all the records of the progress notes based on the existing CRF model. The results are combined with text marked by the doctor. Then, we classify the nouns with frequencies (>5000) based on their part-of-speech into five types and obtain doctors' confirmation. In this way, the establishment of the biomedical terminology library of kidney cancer experiment is completed. The examples of biomedical terminology base of kidney cancer are shown in Table II.

### C. Evaluation on knowledge discovery

In this step, we evaluated the performance of EWM. We used 300 records as the input of EWM. The biomedical named entities in the text should be identified. We used "Jieba" to divide the narrative text and imported the biomedical terminology library as a custom dictionary. Meanwhile, we extracted the medical terms and their category attributes.

TABLE II EXAMPLES OF BIOMEDICAL TERMINOLOGY BASE OF KIDNEY CANCER

| Type | Examples |
|---|---|
| BP | Abdomen, calyx, kidney, coronary artery |
| CS | Difficulty in passing urine, visible hematuria, soft tissue density shadow, solid tumors |
| DN | Renal carcinoma, hysteromyoma, coronary heart disease, gallstone, hypertension |
| MN | Aspirin, betaloc, ibuprofen, metoprolol tartrate, nifuda, pemetrexed disodium |
| TM | Laparoscopic cholecystectomy, immuno-biological cancer therapy |

Then, we evaluate the performance of knowledge discovery in our method. In the previous step, we also obtain the English version of each term to establish a CE-MTMT in building the biomedical terminology library of kidney cancer. The preparation of this step includes the configuration of the MetaMap API, download of offline data of SNOMED CT, and their importation into the database. All concepts related to the concept are obtained through Structured Query Language. Then, we obtain the Chinese name that was written in the original text. When the related English terms are translated into Chinese terms, they may not be accurate, but the translation is consistent. Examples of knowledge discovery are shown in Table III.

TABLE III EXAMPLES OF KNOWLEDGE DISCOVERY BASED ON SNOMED CT

| Primitive terms | SNOMED CT concept | Knowledge discovery |
|---|---|---|
| Abdominal discomfort | Abdominal discomfort | AP - Abdominal pain, finding of abdomen, discomfort (finding) |
| Abdominal mass | O/E - abdominal mass palpated | Abdominal lump, clinical finding present (context-dependent category) |
| Renal hemosiderosis | Renal hemosiderosis | Disease of kidney. nephropathy, NOS, toxic nephropathy (disorder) |
| Urgency | Urgency - urination (urgent) | Lower urinary tract structure, impaired urinary elimination, lower urinary tract |

### D. Performance of CNN-based text classification model of renal cancel

The experiment is based on the deep learning framework TensorFlow of Google. First, the data set is divided into three parts, namely, the training set (800 pieces out of 1000), verification set (200 pieces out of 1000), and test set (300 reserved pieces). The model is constructed based on the training set, the selection of optimal parameters in the model is measured based on the verification set, and the test set is used to evaluate the recognition rate of the model. The parameters involved in the experiment are expressed as follows: the gradient descent learning rate (ADA Alpha) is 0.001, the convolution kernel vector dimension (filter height) is 5, and the number of each convolution kernel (filter number) is 256. We perform two experiments as follows:

In experiment I, the experimental data used do not utilize SNOMED CT in performing knowledge discovery. TensorBoard, a visualization tool provided by TensorFlow, is used to track the accuracy of the model in the validation set during training. TensorBoard produces the results with the increase of the number of iteration steps, as presented in Fig. 4

(smoothness of 0.6). Table IV shows the metrics of the model training on the test set.

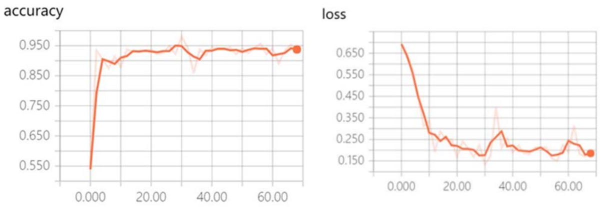


Fig. 4. Accuracy and loss of the training model in Experiments I and II

TABLE IV METRICS OF TRAINING ON THE TEST SET

| **Type** | **Metric** | | | |
|---|---|---|---|---|
| | ***Precision*** | ***Recall*** | ***F1-score*** | ***Support*** |
| Renal cancer | 0.90 | 0.91 | 0.91 | 300 |
| Non-renal cancer | 0.91 | 0.90 | 0.90 | 300 |

In Experiment II, the data used are knowledge-based text. The experimental results are shown in Table V.

TABLE V METRICS OF TRAINING ON THE TEST SET

| **Type** | **Metric** | | | |
|---|---|---|---|---|
| | ***Precision*** | ***Recall*** | ***F1-score*** | ***Support*** |
| Renal cancer | 0.82 | 0.66 | 0.73 | 300 |
| Non-renal cancer | 0.71 | 0.86 | 0.78 | 300 |

The text classification after knowledge discovery results in high accuracy and improvement of other metrics, according to the comparison of the results of the two experiments. In summary, we can accurately determine whether or not the patient has kidney cancer by performing knowledge discovery on the text from the EHRs when the constructed knowledge-based text for the CNN-based text classification model is used.

## IV. CONCLUSION

We propose an EWM that supports early warning to identify whether a patient is expected to suffer from chronic diseases through the undiagnosed narrative text of EHRs. First, we applied CRF in constructing a biomedical terminology library to extract the relevant medical terms from the narrative text, which achieves better performance by utilizing Chinese text. Combined with custom expert dictionaries, we accurately identified biomedical named entities from the text. The use of semantic relationships in SNOMED CT enhanced knowledge discovery within the text and improved the accuracy of CNN-based text classification model for proper EWM of chronic disease. The experiments prove that our method is feasible, and the warning rate can reach 90%. In summary, the model effectively enhances the clinical decision support in early warning of chronic diseases from non-English EHRs.

## REFERENCES

[1] V. Taslimitehrani, G. Dong, N. L. Pereira, et al., “Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function,” J. Biomed. Inform, vol. 60, no. C, pp. 260-269, 2016.

[2] A. Sitapati, H. Kim, B. Berkovich, et al., “Integrated precision medicine: the role of electronic health records in delivering personalized treatment,” Wiley Interdisciplinary Reviews Systems Biology & Medicine, vol. 9, no. 3, pp. e1378, 2017.

[3] K. Jensen, K. M. Augestad, R. O. Lindsetmo, et al., “From unstructured EHR text to data-driven clinical decision support,” Int. J. Integrated Care, vol. 15, no. 7, 2015.

[4] H. Xu, M. C. Aldrich, Q. Chen, H. Liu, N. Peterson, et al., “Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality,” J. Amer. Med. Inform. Assoc., 2015, vol. 22, no. 1, pp. 179-91, 2015.

[5] M. Hoogendoorn, P. Szolovits, L. M. Moons, and M. E. Numans, “Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer,” Art. Intell. Med., vol. 69, no. C, pp. 53-61, 2016.

[6] A. Perotte, R. Ranganath, J. S., Hirsch, D. Blei, and N. Elhadad, “Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis,” JAMIA, vol. 22, no. 4, pp. 872-880, 2015.

[7] K. Gao, H. Xu, and J. Wang, “A rule-based approach to emotion cause detection for Chinese micro-blogs,” Pergamon Press, Inc. 2015.

[8] A. B. Abacha, M. F. M. Chowdhury, A. Karanasiou, Y. Mrabet, et al., “Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug-drug interaction extraction and classification,” J. Biomed. Inform., vol. 58, pp. 122-132, 2015.

[9] B. Tang, H. Cao, X. Wang, et al., “Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks,” Biomed Res Int, vol. 2014, no. 2, pp. 240403, 2014.

[10] Y. Zhang, J. Xu, H. Chen, J. Wang, and Y. Wu, “Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning,” J. Bio. Database & Curation, pp. baw049, 2016.

[11] G. Crichton, S. Pyysalo, B. Chiu, et al., “A neural network multi-task learning approach to biomedical named entity recognition,” Bmc Bioinformatics, vol. 18, no. 1, pp. 368, 2017.

[12] A. Larcher, V. Trudeau, P. Dell'Oglio, et al., “Prediction of Competing Mortality for Decision-making Between Surgery or Observation in Elderly Patients With T1 Kidney Cancer,” Urology, vol. 102, pp. 130–137, 2016.

[13] H. Keijzer, M. A. Snitselaar, M. G. Smits, et al., “Precision medicine in circadian rhythm sleep-wake disorders: Current state and future perspectives,” Personalized Med., vol. 14, no. 2, 2017.

[14] N. Werbeloff, O. Dpj, R. Patel, et al., “The Camden & Islington Research Database: Using electronic mental health records for research,”. Plos One, vol. 13, no. 1, pp. e0190703, 2018,.

[15] A. Rumshisky, M. Ghassemi, T. Naumann, et al., “Predicting early psychiatric readmission with natural language processing of narrative discharge summaries,” Transl Psychiatry, vol. 6, no. 10, pp. e921, 2016.

[16] L. W. Lehman, M. Saeed, W. Long, et al., “Risk stratification of ICU patients using topic models inferred from unstructured progress notes,” AMIA, vol. 2012, no. 10, pp. 505, 2012.

[17] A. Perotte, R. Ranganath, J. S. Hirsch, et al., “Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis,” JAMIA, vol. 22, no. 4, pp. 872-880, 2015.

[18] Q. Suo, F. Ma, Y. Yuan, et al., “Personalized disease prediction using a CNN-based similarity learning method,” IEEE Computer Society, pp. 811-816, 2017.

[19] A. Neelakantan and M. Collins. “Learning dictionaries for named entity recognition using minimal supervision,” Computer Science, 2015.

[20] D. A. Hanauer, M. Saeed, K. Zheng, et al., “Applying MetaMap to Medline for identifying novel associations in a large clinical dataset: a feasibility analysis,” JAMIA, vol. 21, no. 5, pp. 925, 2014

.