Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 10, No. 3, 2023

A-BBL: A Risk Prediction Model for Patient Readmission based on Electronic Medical Records

Nan Yin*, Yong Li

College of Computer Science & Engineering, Northwest Normal University, Lan Zhou 730070, China
* **Corresponding author**: Nan Yin

Abstract: With the spread of medical digitization, electronic health record data has been accumulated in large quantities, laying the foundation for intelligent medical changes. ICU data is mined and analyzed to identify the risk of patient readmission in a timely manner, prevent and control the deterioration of patients' conditions, and reduce the burden of patient costs. However, due to the poor quality of medical data, potential information cannot be effectively mined. In view of the above problems, a patient readmission risk prediction model A-BBL is proposed. By extracting and analyzing the patient 's discharge summary information, the readmission risk of discharged patients within 30 days is predicted. The A-BBL model consists of three parts: firstly, BioBert is used to pre-train the medical text data, extract the semantic information of the medical text, and then generate the corresponding word vector. Then, the sequence model BiLSTM is used to capture the context information and model the input sequence. Finally, the self-attention mechanism is used to extract the key information in the input sequence, enhance the vector representation ability of the sequence, thereby improving the performance and accuracy of the model, so as to predict the readmission rate of patients. Based on the MIMIC-III real medical data set, the A-BBL model for patient readmission prediction proposed in this paper is verified. Compared with the baseline model, the accuracy is improved by 7.2 %. This study can help medical staff better understand and pay attention to the progression of critically ill patients, im-prove the survival rate of patients, and reduce the readmission rate of patients.

Keywords: Electronic Health Records; LSTM; BioBERT; MIMIC-III.

1. Introduction

The digital transformation of the medical field has led to the rapid growth of medical health information. Electronic health records (EHR) and electronic medical records (EMR) are important components of medical electronic data, including a large number of diagnosis and treatment information of patients, such as electrocardiogram (ECG) waveform, medical text, laboratory test results, treatment, drugs, diagnosis and population information. This heterogeneous information has become valuable re-sources for medical staff to assist clinical decision-making. Electronic health record plays an important role in improving the medical system and improving the medical and health conditions of residents. However, due to the large amount of electronic medical data, complex structure and scattered information, it is difficult for doctors to dig out more useful information.

In recent years, electronic medical data has become a new focus of academic at-tention and research, and more and more scholars have explored it in depth. Mining and analyzing medical data can help medical staff to understand the health status and condition of patients more comprehensively, discover the law and trend of disease development in time, and provide patients with the best diagnosis, treatment and prevention programs, so as to improve the quality and effect of medical services. In addition, by evaluating the therapeutic effect and predicting the survival time of pa-tients, it can reduce medical accidents and misdiagnosis rates, rationally allocate med-ical resources, and improve the efficiency and quality of medical services. These appli-cations are of great significance in clinical practice. The application of deep learning technology in the medical field is of great significance. It can make healthcare more intelligent, refined and personalized, and provide a huge boost to the development and progress of healthcare.

In the medical field, predicting and preventing patients' readmission is an im-portant issue. The readmission of patients has certain harmfulness, which is mainly manifested in the following aspects: (1) Increasing medical costs: readmission means that re-diagnosis and treatment are required, and these processes require medical re-sources and time to increase medical costs; (2) Increased treatment time: Readmission of patients usually requires longer treatment and recovery, which may affect the patient 's life and work; (3) Increased risk of complications: readmission may cause more serious health problems, such as nosocomial infection, thrombosis, etc. These complications may have a longer-term impact on the health of patients; (4) Psychological impact: re-admission may have a negative impact on the patient 's psychology, such as doubts about the treatment effect, distrust of medical institutions, etc., thereby affecting the patient 's enthusiasm and confidence in treatment [5-6]. Therefore, people pay more and more attention to the prediction and prevention of readmission in order to improve the efficiency and quality of patient care, and determine it as one of the goals of medical quality improvement. In summary, this paper selects a patient readmission risk prediction task for study based on deep learning techniques. Based on publicly available electronic health records, we use the textual information of patients' pre-discharge medical summaries to study a deep learning patient readmission risk model, which can further improve the accuracy of patient readmission prediction.

2. Related Work

Readmission risk prediction is an important medical

problem, which can be pre-dicted by a variety of methods, including statistical learning, machine learning and deep learning.

2.1. Statistical methods

Regression analysis and survival analysis are commonly used statistical methods. Regression analysis mainly uses the patient 's basic information (such as age, gender, etc.), clinical indicators (such as vital signs, laboratory examination indicators, etc.) and treatment methods and other variables to establish a linear regression model or Logistic regression model to predict the risk of readmission. Blecker et al. used a regression model to study the trend of hospitalization readmission rate of heart failure patients covered by Medicare insurance in the United States. Survival analysis mainly aims to establish a survival model for patients who have readmission within a certain period of time. The Kaplan-Meier curve can intuitively represent the survival probability of patients, while the Cox proportional hazard model can consider the interaction between multiple variables, so as to more comprehensively assess the risk of readmis-sion. However, statistical-based patient readmission risk prediction usually requires feature selection and variable screening to ensure the accuracy and interpretability of the prediction model. At the same time, because the statistical method relies on the assumed data distribution and model assumptions, and has high requirements on data quality, it is necessary to take certain measures to preprocess and clean the data to eliminate the influence of data noise and missing values on the prediction results.

2.2. Machine learning methods

In the field of machine learning, methods such as decision tree and support vector machine are widely used in readmission risk prediction. Kerexeta et al. used a va-riety of machine learning algorithms to construct a prediction model for the 30-day readmission rate of heart failure patients. The comparison showed that the random forest algorithm had the highest prediction accuracy. Zheng et al. proposed a pa-tient readmission rate prediction model based on ant colony algorithm meta-heuristic algorithm and data mining technology. Mortazavi et al. analyzed the application of different machine learning techniques in the prediction of readmission in patients with heart failure. A series of classification algorithms are used and their performance is compared. The experimental results show that the random forest algorithm performs best in predicting 30-day readmission of patients with heart failure, with high accuracy and sensitivity. Based on the time series information generated by patients during hospitalization, the researchers explored the application of recurrent neural networks and their variants. These models can capture the time dependence of disease devel-opment and learn the potential links between different diseases in the patient 's history. Reddy et al. used recurrent neural network and long short-term memory network (RNN-LSTM) to predict the readmission of patients with systemic lupus erythematosus. Use the patient 's previous hospitalization records to learn the potential links between the diseases and add this information to the model. The contribution of this study is to use time-serialized information to more accurately predict the risk of hospital read-mission. Lin et al. used long short-term memory (LSTM) networks and convolu-tional neural network (CNN) models to analyze and predict unplanned intensive care unit (ICU) readmissions. Using time series data from medical records, important fea-tures are extracted and modeled using LSTM to predict whether patients will be ad-mitted to ICU again.

2.3. Research methods and models

With the continuous advancement of Natural Language Processing (NLP) tech-nology, more researchers have begun to use the text data in patients ' electronic health records to predict the risk of readmission. Compared with the traditional risk prediction model based on structured data, the model based on NLP technology can obtain the patient 's health status and medical history information more comprehensively and accurately, thereby improving the prediction accuracy of readmission risk. NLP tech-nology can automatically identify and extract information such as entities, relationships, and events in medical texts. At the same time, it can also analyze unstructured infor-mation such as patient 's emotional state and language features, providing a richer data source for risk prediction. Craig et al. proposed a patient readmission prediction model based on a one-dimensional convolutional neural network structure. The model converts words into vector representations through word2vec pre-training tech-nology, extracts features by convolutional layers, retains the most important features by maximum pooling layers, and generates prediction results by fully connected output layers. Features can be automatically extracted from doctor notes to predict patient re-admission risk. Although the prediction effect of this model is not as good as some published models, it emphasizes the importance of unstructured text in medical records in predicting patients ' readmission risk, and conducts in-depth research on text feature learning and proposes new clinical insights. Based on the BERT model, Huang et al. proposed a ClinicalBERT model for medical text training. Deep representation learning of clinical tests can automatically learn the representation of words, phrases and sentences, as well as the semantic relationship between them, so as to reveal the clinical information hidden in the deep text.

At present, these methods have begun to pay attention to the importance of pa-tients' medical text data, but the application of data feature extraction is not sufficient. This paper proposes a A-BBL model for predicting patients' readmission risk, and ex-tracts more comprehensive text information to judge patients' readmission risk.

3. Research methods and models

This study explored the feasibility and effectiveness of patient readmission risk prediction based on BioBERT-BiLSTM-Attention model by summarizing medical texts 48 hours before discharge. In the study of text classification prediction, the BioBERT model is an improved version of the BERT model. It is pre-trained on a large-scale biomedical text corpus, making it perform well in text processing tasks in the biomedical field. The BiBLST-Attention model can use the BiLSTM model to model the input sequence, so that it can better capture the context information when dealing with long text. At the same time, the Attention mechanism can help the model focus on learning key information. It has good robustness when dealing with text data with certain noise, and can effectively deal with some outliers and mislabeled data. The A-BBL model is shown in Figure 1.

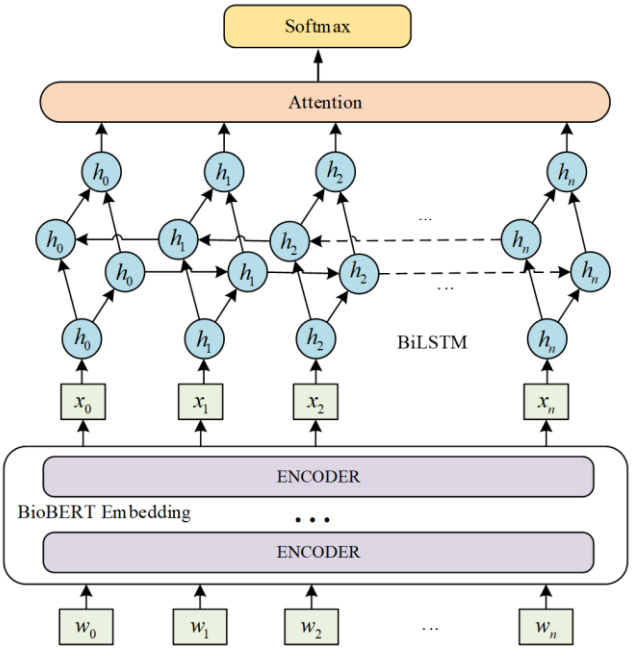

Figure 1. A-BBL model diagram

3.1. BioBERT medical text pre-training

The discharge summary text data in health records contains a large number of medical terms. The general BERT model cannot understand a large number of profes-sional terms, abbreviations, and synonyms in medical text data. BioBERT is a do-main-specific BERT model based on biomedical corpus. It is pre-trained on large-scale biomedical domain databases such as PubMed and PMC, and retains the structure and parameters of the original BERT model. The pre-trained model diagram is shown in Figure 2. BERT (Bidirectional Encoder Representations from Transformers) is a Trans-former-based pre-training model published by Google in 2018. It is an unsupervised pre-training model that trains on a large-scale corpus and can learn rich language knowledge, common natural language representations, including vocabulary, grammar, syntax and semantics.

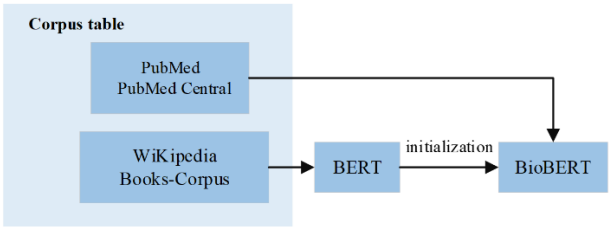

Figure 2. BioBERT pre-trained model diagram

The pre-training process of BERT includes two stages: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In the MLM phase, BERT randomly masks some words in the input text, and then predicts the masked words by the context in-formation of the remaining words through the Transformer encoder. In the NSP stage, BERT combines two input sentences into a training instance, and then uses the Trans-former encoder to predict whether the two sentences are adjacent in the original text.

The input representation of the BERT model consists of three parts. The word vector (Token Embeddings) is the first part of the input layer of the BERT model, which maps each word to a fixed-length vector representation, that is, the word vector. Seg-ment embeddings are used to distinguish different sentences or paragraphs in the input sequence; positional Embeddings are used to specify the position of each word in the input sequence. The combination of these three vectors represents the BERT model 's encoding of the input text, as shown in Figure 3. This input representation enables BERT to effectively handle the relationship between different lengths and different text seg-ments.

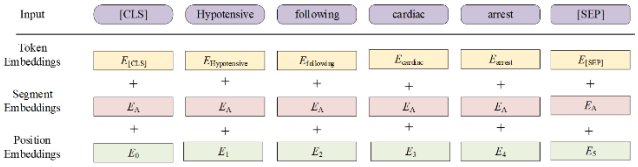

Figure 3. Input representation diagram of BERT model

The discharge summary text of each patient can be expressed as $x = \{W_1, W_2, \cdots W_l\}$, where W_i represents a word, representing a series of words composed of x, and l represents the maximum length of the discharge summary text, which limits the scope of the text and avoids the excessive consumption of computing resources caused by too long text. The output of the BioBERT word embedding layer is calculated as Equation (1), where $H_w \in R_{l \times w_d}$, w_d is the word encoding dimension, the output size of the BioBERT model is a multiple of 768, and the model uses the output of the last layer as input, with a size of 768.

$$H_w = BioBERT(s) \tag{1}$$

3.2. BiLSTM learns text context features

In the electronic health record summary text processing, BiLSTM is used as a text summarizer, which is responsible for capturing the contextual semantic information of the input text sequence, extracting the deeper features of the input text vector, and avoiding the influence of the latter words in the RNN. Specifically, BiLSTM is a deep recurrent neural network that can consider both the context information before and after the current word. By learning the context information, the input text is modeled and the corresponding semantic representation vector is generated. BiLSTM is com-posed of two one-way, opposite-direction LSTMs, with multiple shared weights, and ultimately connected to the same layer of output, which has the ability to remember past and future information, as shown in Figure 4.

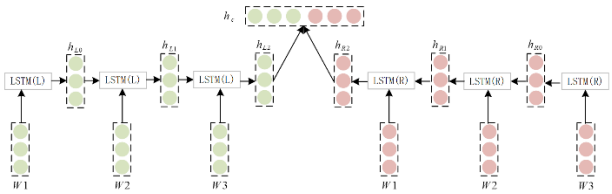

Figure 4. BiLSTM model diagram

Compared with the traditional one-way LSTM, BiLSTM can capture the depend-encies between words more comprehensively and further improve the expression abil-ity and accuracy of the model.LSTM improves the hidden layer on the basis of RNN, and adds three gates to it, which are forget gate, input gate, output gate, and a new hidden state (cell state). The LSTM model diagram is shown in Figure 5.

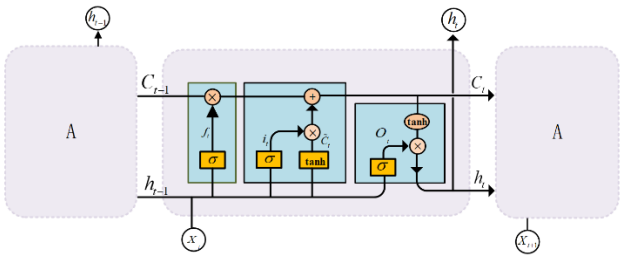

Figure 5. LSTM model structure diagram.

i_t is the output of the input gate, c_t is the output of the candidate memory unit, where x_t is the current input, h_{t-1} is the output of the previous neuron, W_{ix}、U_{ih}、W_{cx} and W_{ch} are weight matrices, b_i and b_c are bias matrices. Input x_t and hidden state h_{t-1} selectively record the information into the cell state through the tanh function and sigmoid function in equations (2) and (3).

$$i_t = \sigma\left(W_{ix}x_t + U_{ih}h_{t-1} + b_i\right) \tag{2}$$

$$\tilde{c}_t = \tanh\left(W_{cx}x_t + W_{ch}h_{t-1} + b_c\right) \tag{3}$$

f_t is the output of the forgetting gate, where W_{fx} and W_{fh} are the weight matrices, and b_f is the bias matrix. The input x_t and the hidden state h_{t-1} are calculated by the sigmoid function in Formula (4) to obtain the f_t value, and the information is selectively forgotten according to the f_t value.

$$f_t = \sigma\left(W_{fx}x_t + W_{fh}h_{t-1} + b_f\right) \tag{4}$$

o_t is the output of the output gate, h_t is the output of the hidden state, where W_{ox} and W_{oh} are weight matrices, and b_o is a paranoid matrix. The input x_t and the hidden state h_{t-1} are calculated by the sigmoid function of the formula (5). The formula (6) multiplys o_t with the cell state $tanh\left(c_t\right)$ to obtain the final output h_t .

$$o_t = \sigma\left(W_{ox}x_t + W_{oh}h_{t-1} + b_o\right) \tag{5}$$

$$h_t = o_t \tanh\left(c_t\right) \tag{6}$$

c_t is the output of the memory unit, multiplying c_{t-1} with the forgetting gate f_t, and multiplying the candidate memory unit c_t with the input gate i_t, the calculation process is like the formula (7).

$$c_t = fc_{t-1} + i_t\tilde{c}_t \tag{7}$$

Similarly, the calculation process of backward LSTM is similar to that of forward LSTM, but the input sequence is calculated in reverse order, and the calculation formula is no longer introduced in detail. The final BiLSTM calculation formula is shown in Formula (8).

$$h_t = [\overrightarrow{h_t} + \overleftarrow{h_t}] \tag{8}$$

3.3. Attention mechanism

The original intention of the attention mechanism is the application of biological attention in artificial intelligence. It is a technology for weighted aggregation. It can give different weights according to different parts of the input. These weights reflect the contribution of each element in the input sequence to the output. When calculating the attention weight, we represent each element in the input sequence as a vector and the target representation (or query vector) as another vector. By calculating the similarity between the target representation and each element vector in the input sequence, we can get a weight vector. By multiplying and adding the weight vector to each element vector in the input sequence, a weighted sum vector is obtained, which represents the most relevant element in the input sequence to the target representation. The Attention mechanism structure diagram is shown in Figure 6.

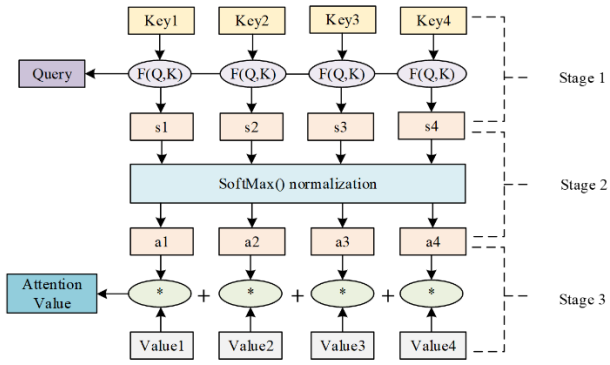

Figure 6. Attention mechanism structure diagram

In the attention mechanism, the elements in the input sequence can be regarded as a set of data pairs, where Key is used to calculate the attention weight, and Value is the value to be weighted and summed. Query represents the content that needs attention, and the attention weight coefficient corresponding to each Key is obtained by calculating the similarity or correlation between Query and each Key. These weight coefficients are used to weight Value and get the final Attention value. Therefore, the essence of the attention mechanism is to weight and sum the Value values of the elements in the input sequence, as shown in Formula (9).

$$Attention\left(Query, source\right) = \sum_{i=1}^{L_x} Similarity\left(Query, Key_i\right) * Value_i \tag{9}$$

The most commonly used attention mechanism is self-attention, in which the input sequence itself is used as a query, key, and value. The weight coefficient of the corresponding Value is obtained by calculating the similarity between the query vector Query and Key, and then the weight vector is multiplied by the value vector to obtain the attention vector, and then the output vector of each time step is calculated.

The specific calculation process of the Attention mechanism, according to the similarity score between Query and Key, first obtains the query vector Q_i, the key vector K_j and the value vector V_j through linear transformation, where i represents the current position and j represents other positions in the sequence. Then, the similarity between the query vector Q_i and each key vector K_j is calculated to obtain the original similarity score e, which is usually calculated by dot product, as shown in Formula (10).

$$e_{i,j} = Q_i \bullet K_j \tag{10}$$

In order to avoid scoring too large or too small, the original score is divided by the dot product result $\sqrt{d_k}$ is scaled, and the calculation is shown in Formula (11), where d_k is the dimension of the query vector and the key vector.

$$e_{i,j} = \frac{Q_i \bullet K_j}{\sqrt{d_k}} \tag{11}$$

After obtaining the original similarity score $e_{i,j}$, the softmax function is used to normalize it, and the score is converted into a probability form to obtain the attention weight coefficient $\alpha_{i,j}$, which is calculated as shown in the formula (12).

$$\alpha_{i,j} = \frac{\exp\left(e_{i,j}\right)}{\sum_{k=1}^{n} \exp\left(e_{i,k}\right)} \tag{12}$$

Here n is the length of the sequence, $\alpha_{i,j}$ which represents the attention weight of the i th element to the j th element, that is, the relationship strength, which can be regarded as the attention distribution of the i th element to the j th element. The attention weight coefficient $\alpha_{i,j}$ and the value vector V_j are weighted and summed to obtain the final representation V_i', which is calculated as shown in the formula (13).

$$V_i' = \sum_{j=1}^{n} \alpha_{i,j} V_j \tag{13}$$

Where V_j is the value vector of the j th element, and V_i' is the representation of the weighted sum of the i th element. Weighted summation is the core of the attention mechanism. By weighting and integrating the value vectors of different elements, a more comprehensive and comprehensive representation can be obtained.

4. Experimental Analysis

4.1. Experimental data

The experimental data are based on the 1.4 version of the MIMIC-III multi-parameter intelligent monitoring database for intensive care published by the computational physiology laboratory of the Massachusetts Institute of Technology. MIMIC-III is a very important publicly available medical information database, which is jointly maintained by MIT Computer Science and Artificial Intelligence Laboratory (MIT CSAIL) and Massachusetts General Hospital. The database includes electronic health records (EHRs) of nearly 50,000 patients who were hospitalized in the ICU ward of Beth Israel Moral Education Medical Center in Boston, Massachusetts from 2001 to 2012.

In this experiment, a subset of MIMIC-III V1.4 was used, and six tables were used, namely PATIENTS, ADMISSIONS, ICUSTAYS, DIAGNOSES _ ICD, D _ ICD _ DIAGNOSES and NOTEEVENTS. The information is shown in table 1.

Because this experiment is based on the medical text (patient discharge summary) in the electronic health record, these data are manually recorded by the medical staff, there may be spelling errors, missing words, different formats, writing irregularities and medical terminology abbreviations and other quality problems. Before the model training, the medical text needs to be preprocessed. The specific operations are as follows:

Table 1. Explanation of experimental data.

Data Table Name	Data Table Contains Attributes
PATIENTS	row_id, subject_id, gender, dob, dod, dod_hosp, dod_ssn, expire_flag
ADMISSIONS	row_id, subject_id, hadm_id, admittime, diachtime, deathtime, admission_type, admission_location, diacharge_location, insurance, language, religion, maeital_status, ethnicity, edregtime, edouttime, diagnosis, hospital_expire_flag, has_chartevents_data
ICUSTAYS	row_id, subject_id, hadm_id, icustay_id, dbsource, first_careunit, last_careunit, first_wardid, last_wardid, intime, outtime, los
DIAGNOSES_ICD	row_id, subject_id, hadm_id, seq_num, icd9_code
D_ICD_DIAGNOSES	row_id, icd9_code, short_title, long_title
NOTEEVENTS	row_id, subject_id, hadm_id, chartdata, charttime, category, description, cgid, iserror, text

(1) Remove non-text content in medical data, such as illegal characters and labels. This experiment uses Python 's regular expression (re) to complete the filtering work, and also establishes an illegal character vocabulary to filter out some punctuation marks and special non-English characters.

(2)The NLTK (Natural Language Toolkit) natural language processing library is used to segment the medical text.

(3) Spelling check correction. There may be spelling errors in medical texts. Use Python 's third-party library pyenchant to complete the spelling check function.

(4) Stem extraction and morphological restoration. The form of English words is changeable, such as single and plural nouns, verb tenses and so on. It needs to be restored to the basic form, and the WordNetLemmatizer class based on wordnet dictionary in NLTK is used to restore the form.

(5) Convert to lowercase. Due to the case problem in English, all words are converted to lowercase by using python 's API, so that statistics like 'Heart ' and 'heart ' are one word.

(6) Introduce stop words. Stop words are words with high frequency but no actual meaning in English text, such as ' a ', ' to ' and some short words. These words do not contain information about the theme of medical text. Filter them out using the list of stop words provided by the NLTK package.

Through the above steps, the noise in the medical text can be reduced and the data can be more clean and standardized. In addition, the key information in the data is preliminarily extracted to reduce unnecessary calculation and storage, thereby reducing the calculation cost and processing time. Finally, it helps to make the data more readable and interpretable, facilitate the understanding of the working principle of the data and algorithm, and improve the accuracy of the model.

4.2. Evaluation index

The patient readmission risk prediction task in this paper is essentially a binary classification task. The goal is to predict whether patients have the risk of readmission for treatment within 30 days of discharge, and to intervene in patients with higher risks in advance, thereby reducing the risk of readmission. Therefore, the evaluation model of the four indicators of Accuracy, Precision, Recall and F1-Measure commonly used in the classification task selected in this paper.

Accuracy: The number of correctly predicted samples divided by the total number of samples. This indicator performs better for data sets with uniform category distribution, but is susceptible to smaller categories for unbalanced data sets. The calculation is shown in Formula (14).

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

Precision: The proportion of the number of positive samples correctly predicted by the classifier to the number of positive samples predicted by the classifier. This indicator focuses on the accuracy of the classifier 's prediction of positive samples, that is, to avoid incorrectly predicting negative samples as positive samples. The calculation is shown in Formula (15).

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

Recall: the proportion of the number of positive samples correctly predicted by the classifier to the number of actual positive samples. This indicator focuses on the ability of the classifier to predict the actual positive samples, that is, to avoid incorrectly predicting the actual positive samples as negative samples. The calculation is shown in Formula (16).

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

F1-score: the harmonic mean of precision and recall. This index considers the accuracy of the classifier 's prediction of positive samples and the ability to predict the actual positive samples. The calculation is shown in Formula (17).

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{17}$$

Among them, the number of samples whose TP is actually a positive sample is predicted by the classifier as the number of positive samples; the number of samples whose FP is actually negative samples is predicted by the classifier as positive samples. The number of samples whose FN is actually a positive sample is predicted by the classifier as a negative sample; the samples whose TN is actually negative are predicted by the classifier as the number of negative samples.

4.3. Comparison method

The performance of the A-BBL model in the patient readmission risk prediction task was evaluated by comparing with other baseline model methods. The experimental comparison is mainly divided into two aspects: (1) Verify the effectiveness of the feature representation method based on BioBERT; (2) Verify the effectiveness of the BiLSTM-Attention classification model. The comparative models include:

1.BioBERT: The BioBERT model is pre-trained to obtain the text feature representation of the patient 's discharge summary, and then directly input into the Softmax classifier through a fully connected layer.

2.Word2Vec-BiLSTM: Word2Vec is used to train the word vector representation of the patient 's discharge summary text and input it as a feature into BLSTM for classification.

3.BioBERT-RNN: The BioBERT model is pre-trained to obtain the text feature representation of the patient 's discharge summary, which is input into the RNN to complete the feature training and classification.

4.BioBERT-CNN: The BioBERT model is pre-trained to obtain the text feature representation of the patient 's discharge summary, which is input into CNN to complete the feature training and classification.

5.BioBERT-BiLSTM: The BioBERT model is pre-trained to obtain the text feature representation of the patient 's discharge summary, which is input into BiLSTM to complete feature training and classification.

4.4. Results analysis

The patient discharge summary text extracted from the MIMIC-III dataset is used to verify the patient readmission risk prediction model A-BBL proposed in this paper. The experimental results compared with other model methods are shown in Table 2.

Table 2. Comparison results of accuracy rate and F1 value of the model.

Models	Precision	F1
BioBERT	0.763	0.761
BioBERT-CNN	0.794	0.795
BioBERT-RNN	0.803	0.799
BioBERT-BiLSTM	0.818	0.813
Word2Vec-BiLSTM	0.725	0.721
A-BBL	0.835	0.836

It can be seen from Table 2 that in the task of predicting the risk of readmission within 30 days after discharge, the accuracy of A-BBL model reached 83.5 %, and Word2Vec-BiLSTM performed the worst, with an accuracy of 72.5 %. The A-BBL model has the highest F1 value of 83.6 %, and the F1 value of Word2Vec-BiLSTM is the worst of 72.1 %. The BioBERT-BiLSTM model was compared with Word2Vec-BiLSTM to verify the effectiveness of BioBERT pre-training. Compared with BioBERT-CNN and BioBERT-RNN, the feature representation uses BioBERT model for medical text pre-training to ensure a single amount, which proves the advantages of BiLSTM model in learning text context semantics. Finally, the comparison between the A-BBL model proposed in this paper and the BioBERT-BiLSTM model proves that the attention mechanism can extract important features and has the best performance on the patient readmission risk prediction data set.

The test set is used to verify the model, and the accuracy of A-BBL and other comparison models is obtained. The results are shown in Figure 7.

It can be seen from Figure 7 that the A-BBL model has the highest accuracy, with an accuracy of 82.6 %, followed by BioBERT-BiLSTM and BioBERT-RNN, and the worst performance is Word2Vec-BiLSTM, with an accuracy of 79.5 %. The text representation method based on Word2Vec has the lowest indicators, mainly because Word2Vec can well express the semantic relationship between words, but ignores the long-distance semantic association information. Overall, the patient readmission risk prediction model A-BBL

proposed in this paper has certain advantages compared with other comparative models.

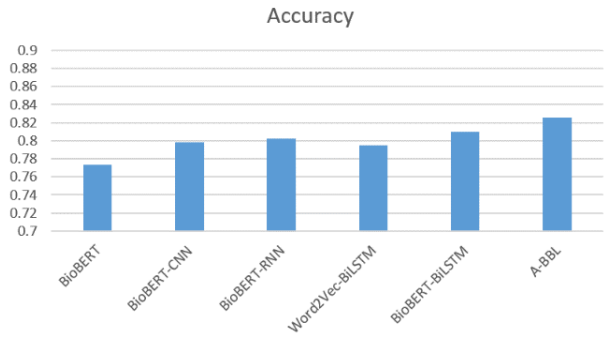

Figure 7. Accuracy Comparison Results

5. Conclusion

This paper studies the patient 's readmission risk prediction task, combined with the conventional judgment index of medical research, defines the problem as judging whether the discharged patient will be readmitted for treatment within 30 days. Based on the discharge summary text in the MIMIC-III data set of electronic health records, a patient readmission risk prediction A-BBL (BioBERT-BiLSTM-Attention) model was constructed to learn the commonality between patients who were readmitted for treatment, and to predict the readmission risk of discharged patients. A large number of experiments show that the A-BBL model proposed in this paper has higher recall rate, accuracy rate and F1 value in the prediction task of ICU patients ' readmission, which is significantly better than other models. However, due to the limitations of the data, the prediction model did not achieve the expected results. In the future, more data sets will be found to verify the model, and new patient readmission risk prediction models will be studied to obtain better prediction results and provide better diagnosis and treatment services for patients.

References

[1] Toscano F, O'Donnell E, Unruh M A, et al. Electronic health records implementation: can the European Union learn from the us. EJPH 28, cky213-401(2018).

[2] Huchang Liao, Fan Liu, Keyu Lu, et al. A review of patient behavior mining based on online medical reviews and its application in medical decision making and management[J]. Journal of the University of Electronic Science and Technology (Social Science Edition), 2022, 24(03):1-22.

DOI:10.14071/j.1008-8105(2022)-1100.

[3] Apra B, Wei D C, Bs D, et al. Subcategorizing EHR diagnosis codes to improve clinical application of machine learning models [J]. International Journal of Medical Informatics, 2021, 156, 104588. DOI: 10.1016/j.ijmedinf.2021.104588

[4] Doddavarapu V N S, Kande G B, Rao B P. Differential diagnosis of Interstitial Lung Diseases using Deep Learning net-works[J]. Imaging Science Journal The, 2020(1):1-9.

[5] Zaya M, Phan A, Schwarz ER. Predictors of re-hospitalization in patients with chronic heart failure[J]. World J Cardiol. 2012, 4(2): 23-30.

[6] Lesyuk W, Kriza C, Kolominsky-Rabas P.Cost-of-illness studies in heart failure: a systematic review 2004-2016[J]. BMC Cardiovasc Disord. 2018, 18(1): 74.

[7] Mortele KJ, Wiesner W, Intriere L, et al. A modified CT severity index for evaluating acute pancreatitis: improved correlation with patient outcome[J]. AJR Am J Roentgenol, 2004, 183(5):1261–1265.

[8] Blecker S, Herrin J, Li L, et al. Trends in Hospital Readmission of Medicare-Covered Patients With Heart Failure[J]. J Am Coll Cardiol,2019,73(9):1004-1012.

[9] Edward L. Kaplan, Paul Meier. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 1958, 53(282): 457-481.

[10] Cox, D.R. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 1972, 34 (2): 187–220.

[11] KEREXETA J, ARTETXE A, ESCOLAR V, et al. Predicting 30-day Readmission in Heart Failure using Machine Learning Techniques[J]. HEALTHINF, 2018: 308-315.

[12] ZHENG B, ZHANG J, YOON S W, et al. Predictive modeling of hospital readmissions using metaheuristics and data mining[J]. Expert Systems with Applications, 2015,42(20): 7110-7120.

[13] MORTAZAVI B J, DOWNING N S, BUCHOLZ E M, et al. Analysis of Machine Learning Techniques for Heart Failure Re-admissions[J]. Circulation: Cardiovascular Quality and Outcomes, 2016,9(6): 629-640.

[14] REDDY B K, DELEN D. Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology[J]. Computers in Biology and Medicine, 2018,101: 199-209.

[15] LIN Y, ZHOU Y, FAGHRI F, et al. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory[J]. PLOS ONE, 2019,14(7): e218942.

[16] Craig E, Arias C, Gillman D. Predicting readmission risk from doctors' notes[J]. 2017.

[17] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[18] HUANG K, ALTOSAAR J, RANGANATH R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission [J]. 2019.

[19] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. 2018.

[20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, Volume 36, Issue 4, February 2020, Pages 1234–1240.

[21] Johnson A E, Pollard T J, Shen L, et al. MMIC-III, a freely accessible critical care database[J]. Scientific data, 2016,3:160035.