Recurrence Prediction and Risk Classification of COPD Patients Based on Machine Learning

Xin Qi[1], Hong Chen[2]*
Academic Affairs Office, Heilongjiang University of Chinese Medicine, Harbin 150001, China[1]
Chinese Pediatrics, First Affiliated Hospital, Heilongjiang University of Chinese Medicine, Harbin 150040, China[2]

***Abstract*—In response to the frequent recurrence and readmission of patients with chronic obstructive pulmonary disease, a machine learning based recurrence risk prediction and risk classification model for patients with chronic obstructive pulmonary disease is studied and constructed. Approach: This model first utilizes the optimized long short-term memory network to recognize named entities in patient electronic medical records and extract entity features. Then, XGBoost is used to predict the probability of patient relapse and readmission, and its risk is classified. Results: These results confirm that the optimized bidirectional long short-term memory network has the best performance with an accuracy of 84.36% in electronic medical record named entity recognition. The accuracy of XGBoost is the highest on both the training and testing sets, with values of 0.8827 and 0.8514, respectively. XGBoost has the best predictive ability and effectiveness. By using k-means for layering, the workload of manual evaluation was reduced by 91%, and the overall simulation accuracy of the model was as high as 97.3% and 96.4%. Conclusions: These indicate that this method can be used to balance high-risk patients between risk, cost, and resources.**

Keywords*—Machine learning; COPD; BiLSTM; XGBoost; k-means; recurrence; risk classification

I. Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a type of lung disease characterized by airflow restriction, which is a long-term, irreversible, and progressive chronic respiratory disease [1]. COPD is a common disease with a high incidence rate, which poses a heavy burden on the physical health and disease burden of the people [2]. It has become an important public health problem that restricts China's economic and social development [3]. Due to factors such as medical quality, hygiene effectiveness, and economy, some patients may relapse and be hospitalized within a month due to the same reasons, which has become a major challenge facing the world [4-5]. Predicting preventable frequent relapse hospitalization and understanding the causes of relapse hospitalization are currently important topics of widespread concern [6]. Machine Learning (ML) is a scientific method that enables prediction, identification, and decision-making of environmental changes without human intervention [7]. ML has achieved rapid development in medicine, especially in early warning of diseases, providing convenient and fast decision support for people [8]. Currently, ML and data mining have become methods that can potentially improve the predictive ability of relapse admission risk prediction models [9]. How to use ML technology to predict the recurrence of COPD patients and classify their risk is the main problem of this study. Currently, research on recurrence prediction and risk classification for COPD patients mainly relies on the experience and professional knowledge of clinical doctors, lacking objective and systematic prediction models. Therefore, the research on using ML technology for predicting the recurrence and risk classification of COPD patients still needs to be improved. The main objective of this study is to use ML technology to establish a model that can accurately predict the recurrence of COPD patients and classify their risk, improve the management and treatment level of COPD patients, reduce the risk of recurrence, and improve their quality of life. The innovation of this study lies in the full utilization of electronic medical record text information, while also ensuring the personal information of patients.

The article consists of four sections. In Section I, there is a literature review that introduces the relevant research content of different scholars. Section II outlines the related works. Section III is the research method, which mainly introduces the Named Entity Recognition (NER) of electronic medical records of COPD patients, as well as the prediction and risk classification of relapse and readmission of COPD patients. Section IV is the result analysis, which explains the electronic medical record NER of COPD patients under different ML algorithms, as well as the results of predicting relapse, readmission, and risk classification of COPD patients. Discussion is presented in Section V. Finally, Section VI is the conclusion that summarizes the results of recurrence prediction and risk classification for COPD patients and points out the shortcomings of the research.

II. Related Works

The Bidirectional Long Short-Term Memory Network (BiLSTM) is widely used in NER and predictive analysis, and many scholars have achieved good research results. To improve the NER efficiency of product review, researchers such as Postiche H proposed a product review NER method based on BiLSTM. These experiments confirm that the research method is more efficient in identifying named entities in product reviews compared to current methods. Benali B A et al. proposed a multi-head self-attention mechanism based on structures such as BiLSTM to improve the accuracy of NER in natural language texts, to perform NER on natural language in social media. These experiments confirm that this method can combine characters and words in the embedding layer, resulting in significantly better recognition results than current naming recognition methods [10]. Long R and other scholars proposed a disease diagnosis depth framework that integrated BiLSTM to enhance the analysis of electronic

medical record data. These experiments confirm that the framework can significantly improve the performance of disease diagnosis [11]. Puh K et al. designed a deep learning model based on BiLSTM to predict the emotions in tourist comments in the tourism industry to extract and rate the emotions in tourist comments. These experiments confirm that this deep learning model has more efficient work efficiency and more accurate prediction results compared to other models [12].

ML has become an emerging and effective method for predicting the risk of disease recurrence. Matheson A M and other researchers proposed an ML-based prediction method to address the high risk of COPD recurrence. These experiments confirm that this method's predicting accuracy is superior to traditional regression analysis methods, which is beneficial for the prevention of COPD recurrence [13]. AL khadar and other scholars designed a recurrence prediction model based on decision tree classifier to enhance the prediction efficiency of recurrence and survival rate in oral cancer patients. These experiments confirm that the research method has the best predictive performance compared to other traditional ML models [14]. To compare the predictive performance of ML algorithm for early biochemical recurrence after prostatectomy, Wong NC et al. selected three supervised ML algorithms to construct models for prediction and compared them with traditional regression analysis. These experiments confirm that these three ML models have high accuracy in predicting biochemical recurrence [15]. Paredes AZ and other scholars proposed a fusion ML index prediction method to address the issue of high postoperative recurrence rates in patients with liver metastasis from colorectal cancer. This method combines resampling method with multivariate mixing effect to construct a prediction model. These experiments confirm that this method has a high accuracy in predicting indicators, which is beneficial for clinical analysis of postoperative recurrence risk [16].

In summary, BiLSTM can improve the accuracy of NER, and ML outperforms traditional analysis methods in predicting the risk of disease recurrence. However, the above studies lack the application of NER in predicting the recurrence risk of disease patients. There are few studies that combine these two and use ML to predict the risk of disease recurrence based on NER of electronic medical records. Therefore, to predict and classify the risk of recurrence in COPD patients, this study uses ML to perform NER on the patient's electronic medical record and predicts the re-admission of recurrence, providing a certain scientific basis for medical institutions.

III. ML Based Recurrence Prediction and Risk Classification Model for COPD Patients

A recurrence prediction and risk classification model for COPD patients is constructed for the prediction of recurrence. Firstly, BiLSTM-CRF is used to perform NER on electronic medical records. Then, XGBoost is used to predict the recurrence and readmission of COPD patients. Finally, k-means is used to classify the recurrence risk of COPD patients, as well as early prevention and effective intervention.

A. Electronic Medical Record NER for COPD Patients Based on Optimized BiLSTM

The electronic medical record collects rich diagnosis and treatment information from patients, including detailed historical data of patients seeking medical treatment and treatment in the hospital. It is the foundation and core of medical and health big data. There are various forms of electronic medical record storage, including structured and textual data. If people only focus on easily processed structured data and ignore textual data, it will lead to the inability to obtain certain characteristics of patient real data [17]. When designing a method for predicting the recurrence and readmission of COPD patients, it is first necessary to extract medical entities with more features from text data in electronic medical records and structurally process them. Fig. 1 shows the NER of an electronic medical record.

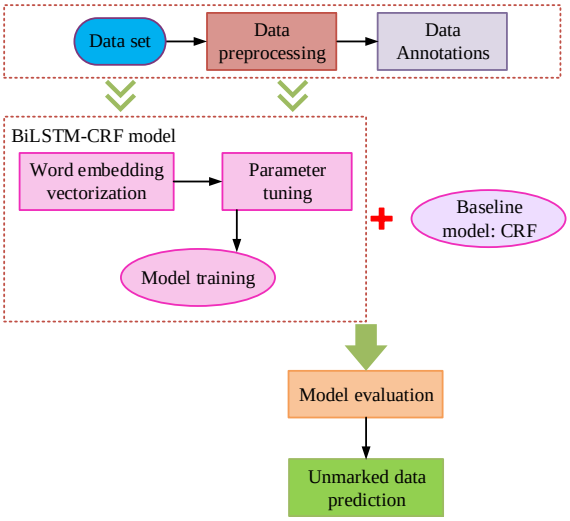

Fig. 1. The recognition process of named entities in electronic medical records.

According to Fig. 1, in electronic medical records' NER, it should first obtain data, and then preprocess and annotate the data. In the experiment, an electronic medical record NER method based on BiLSTM-CRF is designed, followed by word embedding vectorization, parameter tuning, and model training. Due to the fact that CRF is a commonly used method superior to other methods, it is used as a baseline method for method evaluation and prediction of unlabeled data.

The electronic medical record data in this study are sourced from the hospitalization records of patients in a tertiary hospital in Zaozhuang City. COPD patients who were hospitalized in the hospital from January 2020 to January 2021 were selected. With consent, the researchers logged into hospital's patient information management system and searched for target patient information. And it was deprivileged to obtain the inpatient information table and inpatient case text. Among them, textual information written by doctors, such as admission and exit records, can promote the design of predictive methods for patient relapse and readmission, and needs to be processed using the NER method based on ML. Before performing NER on inpatient electronic medical records, it is necessary to standardize the electronic

medical records, including desensitization of patient data, detection and processing of sick sentences, and removal of irrelevant symbols [18]. Useful entity information was extracted for predicting the recurrence and readmission of COPD patients, mainly divided into comorbidities, examination indicators, indicator values, disease course, past physical fitness, and lifestyle habits. Afterwards, the data entities were annotated to construct the entity annotation dataset for the training model.

When designing the NER method based on BiLSTM, it is necessary to first perform word vector embedding transformation. Word embedding can solve the problem of vector encoded word similarity by converting text into numerical values. On this basis, the BiLSTM-CRF system is used to identify entities defined in the electronic medical records of hospitalized COPD patients. This system mainly includes two parts: BiLSTM layer and CRF loss layer. The former is essentially a traditional recurrent neural network. Due to its risk of gradient vanishing or exploding, it utilizes Long and Short-Term Memory (LSTM) for long-distance dependency capture [19]. Since LSTM cannot encode information from back to front, it is optimized to obtain BiLSTM to capture bidirectional semantic dependencies. The outputs of BiLSTM are all independently selected markers with higher scores, and there is no significant correlation between each marker. CRF can extract a certain number of constraint rules from samples and adjust them at the syntactic level. Therefore, this can enhance the constraints on samples, reduce the possibility of illegal sequences appearing, improve the prediction accuracy of labeled sequences, and ensure the generation of globally optimal labeled sequences. A BiLSTM-CRF system was constructed by combining BiLSTM and CRF. BiLSTM was used to learn features in the training set, including sub embedding and other features. After feedback to CRF layer, the tag sequence with the highest score was output. For a certain input sentence $[x]_1^T$, $[f\theta]i,t$ means the score of the i-th tag at time t. $[A]_{i,j}$ represents the probability of the ij-th position in the transition probability matrix of CRF, that is, the probability of transitioning from state i to state j. So for a certain input sentence and its label sequence $[i]_1^T$, Eq. (1) is the final score.

$$S\left([x]_1^T,[i]_1^T,\theta\right)=\sum_{t=1}^{T}\left([A]_{[i]_{t-1},[i]_t}+[f\theta]_{[i]_t,t}\right) \quad (1)$$

In Eq. (1), S represents the final score. A BiLSTM-CRF system has been successfully constructed to recognize entity information in electronic medical record texts. This model mainly consists of a word vector, BiLSTM, and CRF layers in Fig. 2.

Through Fig. 2, the system takes the text sequence of medical records as input and the relevant entities in the medical records as output. In the implementation of the system, each character input is first converted into a vector form through Word2Vec, which is used as the input of BiLSTM to extract contextual features. The output feature vector is used as the input of the CRF layer, and the input is normalized and the final label sequence is output. BiLSTM can provide users with more complete contextual information, thereby better understanding contextual dependencies. On this basis, an additional CRF layer has been added to further improve the existing BiLSTM system. By comprehensively optimizing the recognition results at the sentence level, the shortcomings of BiLSTM system are compensated.

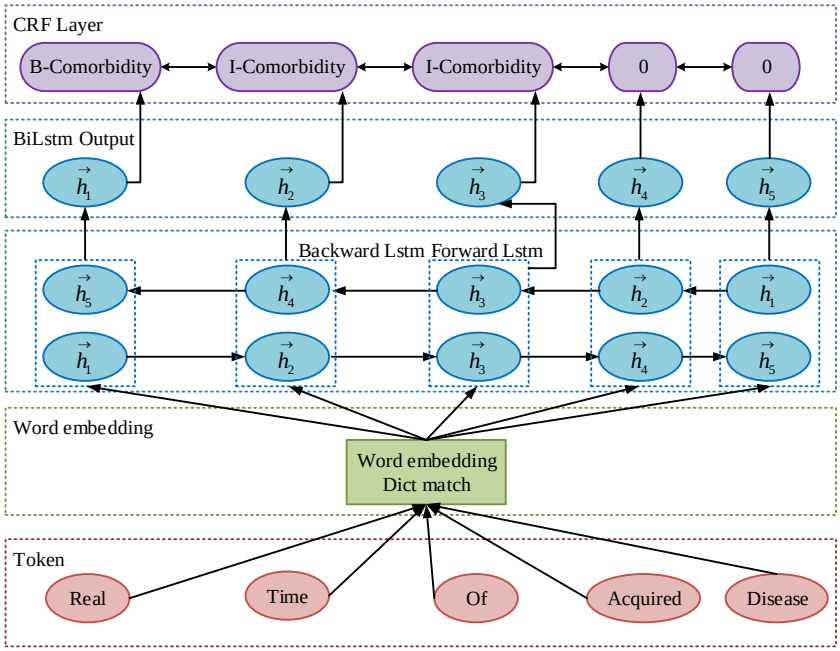

Fig. 2. BiLSTM-CRF network architecture.

B. Prediction of Recurrence in COPD Patients Based on XGBoost

This model's second layer is the prediction of relapse and readmission. That is, based on the medical information generated by the patient during hospitalization, it determines whether they will relapse and be re-admitted due to their physical condition for a period of time after discharge. 30 days are usually the most commonly used readmission threshold. Therefore, in this study, the readmission threshold is set to 30 days when predicting the recurrence and readmission of COPD patients [20]. With medical big data accumulating and ML progressing continuously, the prediction and prediction technology for relapse and readmission can effectively identify potential risk factors, providing a basis for the diagnosis and treatment of relapse and readmission patients. Fig. 3 shows the ML-based prediction of relapse and readmission in COPD patients.

According to Fig. 3, in the ML-based prediction of relapse and readmission of COPD patients, it is necessary to first integrate geometric features and then clean the data. Afterwards, a recurrence and readmission prediction system for COPD patients is designed and evaluated. Based on the previous section of NER, the extracted entity information and features are preserved. Due to the presence of noise and missing raw data, it is necessary to preprocess it to carry out predictions and obtain accurate prediction results. Fig. 4 shows the preprocessing content.

According to Fig. 4, preprocessing includes feature processing for comorbidities, assignment of discrete features, handling of outliers and missing values, descriptive statistics of variables, and handling of classification imbalance. Using ML algorithm to construct a relapse readmission prediction system for COPD patients, considering the system applicability, it should find the optimal classification method to achieve the best prediction effect. XGBoost is similar to a decision tree, in which each tree is divided by a special interval in its columns. On each tree shape, special attention should be paid to the sample with the highest error rate on the previous tree shape, so that there will be more optimization samples on the next tree shape. Finally, the results of each tree are combined and the weighted average method is generally used to obtain the results of each tree [21]. When adjusting XGBoost parameters, it should first adjust the learning rate of the system, followed by the total number of estimators used, the depth of each estimator, and finally the L1 and L2 regularization parameters [22]. To evaluate the classification performance, it should use five indicators: accuracy, accuracy, recall, F1 value, and ROC. Eq. (2) represents the accuracy.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

In Eq. (2), ACC represents accuracy. TP, FN, TN, and FP represent positive cases that are correctly classified, positive cases that are misclassified as negative cases, negative cases that are correctly classified, and negative cases that are misclassified as positive cases, respectively. Eq. (3) represents the precision.

$$P = \frac{TP}{TP + FP} \tag{3}$$

In Eq. (3), P represents precision. Eq. (4) represents the recall rate.

$$R = \frac{TP}{TP + FN} \tag{4}$$

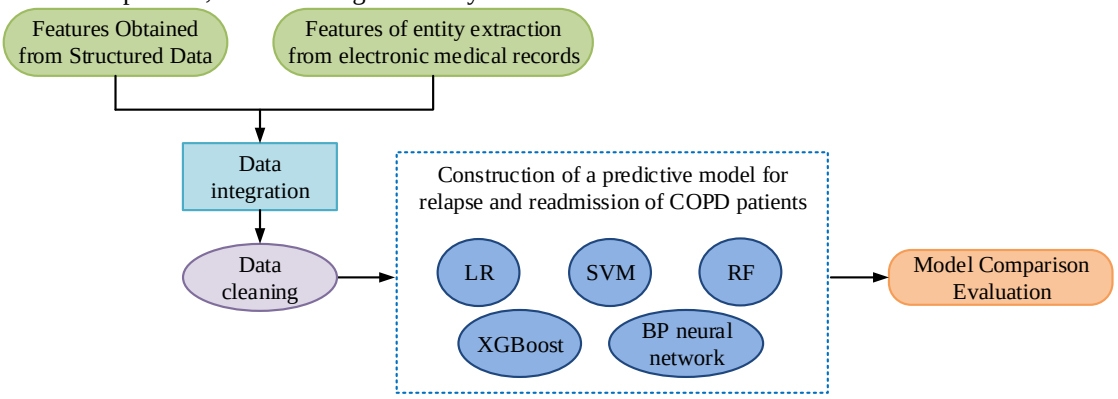

Fig. 3. Machine learning based prediction process for relapse and readmission of COPD patients.

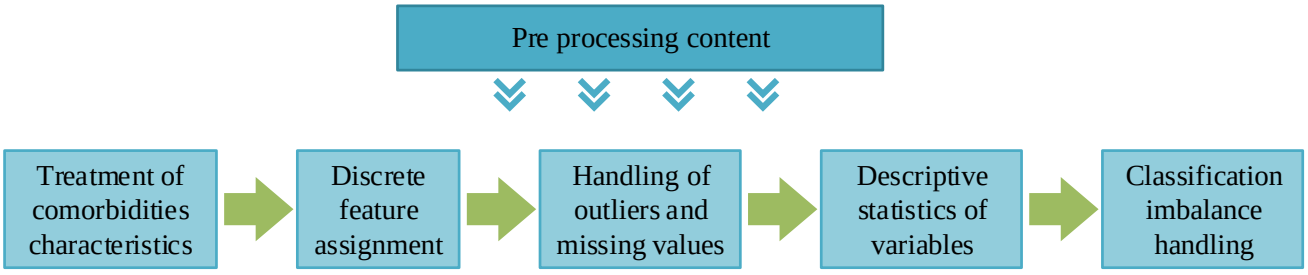

Fig. 4. Preprocessing content.

In Eq. (4), R represents the recall rate. Eq. (5) is the expression for F1.

$$F1 = \frac{2P*R}{P+R} \tag{5}$$

Eq. (6) is the calculation of ROC.

$$FPR = \frac{FP}{FP+TN} \tag{6}$$

In Eq. (6), FPR represents the ROC value. To further explain the output results of the system, this study will use Shapley addition to explain the output results of the prediction system. As a visualization tool, it calculates each feature and explains the system's prediction results by contributing to the prediction. Eq. (7) is the calculation of the Shapley value.

$$y_i = y_{base} + f\left(x_{i1}\right) + f\left(x_{i2}\right) + ... + f\left(x_{ik}\right) \tag{7}$$

In Eq. (7), x_{ik} represents the k-the feature of the i-the sample. The system baseline is y_{base}. The predicted value of the system for this sample is y_i. $f\left(x_{ik}\right)$ represents the Shapley value of x_{ik} [23]. Generally speaking, $f\left(x_{ik}\right)$ is the contribution of x_{ik} to the final predicted value y_i. When $f\left(x_{ik}\right) > 0$, it indicates that the feature can improve the predicted value and has a positive effect. On the contrary, it indicates that the feature value leads to a decrease in the predicted value, which has a reverse effect. This value can reflect the influence and positive and negative shapes of sample features and has a powerful data visualization function, which is conducive to visually displaying the prediction results.

C. Recurrence Risk Classification of COPD Patients Based on K-means

The last layer of the model is the classification of recurrence risk for COPD patients. After selecting the best recurrence and readmission prediction method XGBoost, the selected method can be arranged and used. And combined with the predicted recurrence and readmission results of the patient, the discharge decision was obtained together. On this basis, a risk scoring system is constructed to divide recurrent COPD patients into different risk groups based on the obtained risk scores, providing decision support for doctors. The probability of XGBoost output on the test dataset was stratified using k-means and the results were validated in Fig. 5.

According to Fig. 5, the risk stratification process requires first using a complete prediction dataset to perform predictions on XGBoost. Then, the prediction probabilities output by XGBoost on the training dataset are grouped according to the equidistant method, and samples that do not meet the required probabilities are discarded. Then, the retained samples are input into XGBoost, and the prediction probability is clustered according to k-means. Samples that do not meet the required probability are eliminated, and samples that meet the probability continue to be predicted in XGBoost [24-25]. Finally, the evaluation index values obtained from each method training are calculated, and patients are stratified according to risk. In the k-means clustering, a dataset with d dimensions, whose number is n, is divided into k clusters. The purpose is to minimize the sum of squares of the distances between each observation in the same cluster and the center of the cluster to which it belongs in Eq. (8).

$$J\left(c,u\right) = \sum_{i=1}^{M} \left\| x_i - u_{ci} \right\|^2 \tag{8}$$

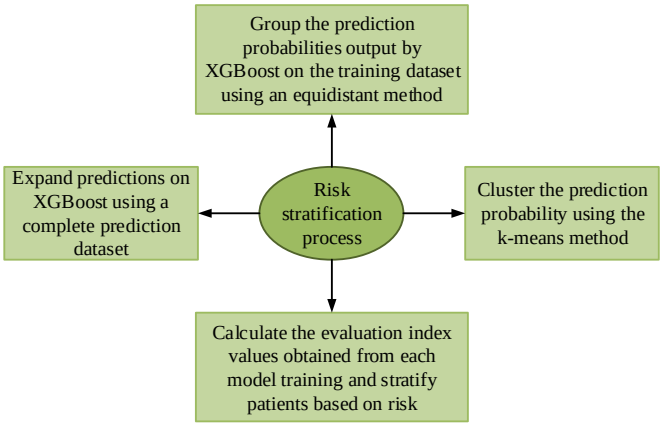

Fig. 5. Risk stratification process.

In Eq. (8), x_i represents the i-the sample. c_i is the cluster to which x_i belongs. u_{ci} represents the center point of the cluster pair. M is the total number of samples [26]. By using the k in the initial conditions, multiple different choices are made for k class, and the optimal clustering is obtained.

D. Methodology

In this study, the design of data collection and analysis programs is crucial. Data are collected from hospitals or medical databases, including electronic health records and clinical trial data. Information is collected from various aspects such as medical history, lifestyle, treatment response, and laboratory test results. The data types include patient basic information, clinical data, lifestyle data, and treatment information. Patients' basic information features include age, gender, weight, etc. Clinical data features include lung function test results, blood gas analysis, imaging examination results, etc. Treatment information features include drug treatment, non-drug treatment, etc. Lifestyle data features include smoking history, dietary habits, activity levels, etc.

Then the data are preprocessed and cleaned, including handling missing values and outliers. Standardization processing, such as normalizing numerical data, is conducted. Feature engineering, such as selecting features with high relevance, is performed. Statistical analysis of basic characteristics, such as age distribution and gender ratio, is carried out. The recurrence of COPD is analyzed, including frequency, severity, etc.

The first step of analyzing the program is to select a suitable ML model based on the characteristics of the data. The second step is to train and validate the model, dividing the

dataset into training and testing sets. The model is trained on the training set and its performance is verified on the testing set. The model parameters are optimized using methods such as cross validation. The third step is to conduct feature importance analysis, analyzing the impact of each feature on predicting COPD recurrence and determining the most influential risk factors. The fourth step is to carry out risk classification, classify patients according to the predicted results, and set clinical decision guidelines corresponding to different risk levels. These steps need to be carried out in compliance with relevant data protection regulations and ethical standards.

IV. Result Analysis of Recurrence Prediction and Risk Classification Model for COPD Patients

The results of the ML-based recurrence prediction and risk classification model for COPD patients were analyzed, thereby promoting the improvement of medical structure and the utilization of medical resources. These results include the NER results of COPD patient electronic medical records based on optimized BiLSTM, the recurrence prediction results of COPD patients based on XGBoost, and the recurrence risk classification of COPD patients based on k-means.

A. NER Results of Electronic Medical Records for COPD Patients Based on Optimized BiLSTM

A suitable programming language was selected to implement the BiLSTM-CRF system and conduct the experiment. Appropriate development tools and libraries were installed to support system development. Electronic medical record data were preprocessed, including word segmentation, feature extraction, and label format conversion. It was ensured that the data format met the model input requirements and useful features were extracted for system training. When setting up the experimental environment, it ensures that all development tools, libraries, and data can be installed and configured correctly, enabling it to run smoothly. The BiLSTM-CRF system is built, and Table I shows the experimental environment settings.

Statistical analysis was conducted on the results of three recognition methods, BiLSTM, CRF, and BiLSTM-CRF, on the test set. All data are the average obtained from a 10-fold cross test. The overall experimental results were calculated in Table II.

According to Table II among these three methods, BiLSTM-CRF has the highest accuracy, regression rate, and F1 value, followed by BiLSTM and CRF. Overall, the accuracy rates of the three recognition methods BiLSTM, CRF, and BiLSTM-CRF are 84.36%, 83.67%, and 87.40%, respectively. Among these three methods, BiLSTM-CRF has the best performance. Subsequently, the recognition performance of different types of entities using the BiLSTM-CRF method was analyzed in Fig. 6.

According to Fig. 6, there is no significant difference in P and R values between these six entities naming recognition results. BiLSTM-CRF has the best recognition effect for both "comorbidities" and "examination indicators", with F1 greater than 90%. For F1, "lifestyle habit" is 74.00% and "course of disease" is 74.39%, indicating the difficulty in distinguishing between these two types. Based on the analysis of experimental results and annotated corpora, the effectiveness of entity recognition is closely related to the number of annotations and the degree of differentiation between different types of entities.

TABLE I. Experimental Environment

Number	Name	Type
(1)	Operating system	Windows 10
(2)	Development language	Python
(3)	CPU	TeslaP40
(4)	GPU	Intel Core i5-4210M @2.50GHz
(5)	Learning framework	Kera’s
(6)	Development platform	Jupiter Notebook 5.78

TABLE II. Performance of Three Models

Entity Category	BiLSTM			CRF			BiLSTM-CRF		
	P (%)	R (%)	F1(%)	P (%)	R (%)	F1(%)	P (%)	R (%)	F1(%)
Complication	90.25	87.43	87.24	87.75	85.06	86.37	92.09	91.84	91.97
Inspection indicators	87.68	88.36	87.76	85.81	93.86	84.31	91.35	92.09	90.22
Habits and customs	70.35	60.19	62.73	66.42	41.79	50.89	74.83	75.47	74.00
Course of disease	62.76	53.86	62.77	52.66	36.06	41.39	72.21	73.78	74.39
Indicator value	80.39	78.64	68.37	77.47	65.75	53.89	84.14	84.17	88.82
Previous constitution	81.27	78.26	72.43	79.74	73.16	77.79	83.05	83.26	84.63
Whole	84.36	81.58	79.07	83.67	74.31	79.07	87.40	88.28	85.83

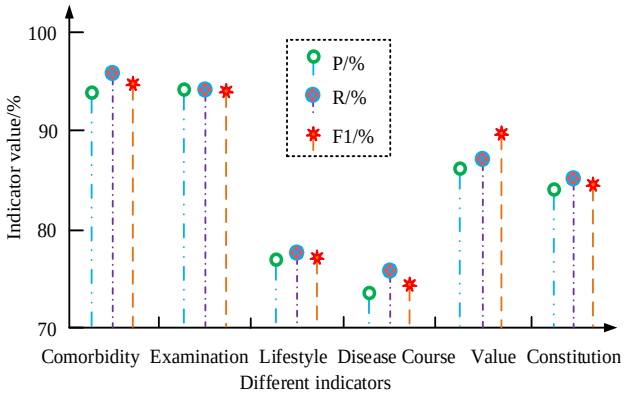

Fig. 6. The recognition effect of different types of entities in BiLSTM-CRF method.

B. *Prediction of Recurrence in COPD Patients Based on XGBoost*

The experimental data are divided into training and testing sets, with a ratio of 7:3. The performance of individual patterns in the training set was evaluated using a grid search method with 10-fold cross validation. On this basis, the optimal hyperparameter combination for each method was selected, corresponding hyperparameter combinations were established, and fitting was conducted on a complete training dataset. Finally, the method was validated through independent experimental data. Due to the fact that this method is not affected by the experimental set during parameter adjustment and learning, the experimental set is only used for final evaluation. A comparison was made between five common ML constructed prediction systems: Logistic Regression (LR), Support Vector Machine (SVM), BP neural network, Random Forest (RF), and XGBoost in Fig. 7.

Fig. 7(a) and Fig. 7(b) show the predictive performance results of the five systems on the training and test sets, respectively. According to Fig. 7, five systems' accuracy on the training set is XGBoost, BP, RF, SVM, and LR, in descending order, with values of 0.8827, 0.8559, 0.8512, 0.8239, and 0.7244, respectively. The accuracy of these five systems on the test set is in descending order of XGBoost, RF, BP, SVM, and LR, with values of 0.8514, 0.8254, 0.7896, 0.7687, and 0.6978, respectively. Among these five systems, XGBoost has the best performance in terms of prediction accuracy, recall, and F1 value. Subsequently, the ROCs of five systems were analyzed in Fig. 8.

According to Fig. 8, the areas under the ROC of these five systems are XGBoost, RF, BP, SVM, and LR in descending order, with AUC values of 0.9173, 0.8256, 0.8252, 0.7682, and 0.6993, respectively. XGBoost has the best predictive ability and effectiveness, verifying its efficiency as a classification method for predicting the recurrence and readmission of COPD patients.

C. *Classification of Recurrence Risk in COPD Patients Based on K-means*

On the basis of XGBoost-based patient relapse readmission prediction system, k-means was selected for risk classification. To verify the superiority of this method, equidistant partitioning was used for comparative analysis. Firstly, patients are divided into five risk groups, namely low risk high reliability, low risk medium reliability, low risk or high-risk low reliability, high risk medium reliability, and high-risk high reliability. Among them, for patients with a lower risk of relapse and higher credibility, the classification probability tends to be closer to zero. For patients with high reliability and high risk of relapse and readmission, the classification probability tends to be closer to 1. Table III shows the results.

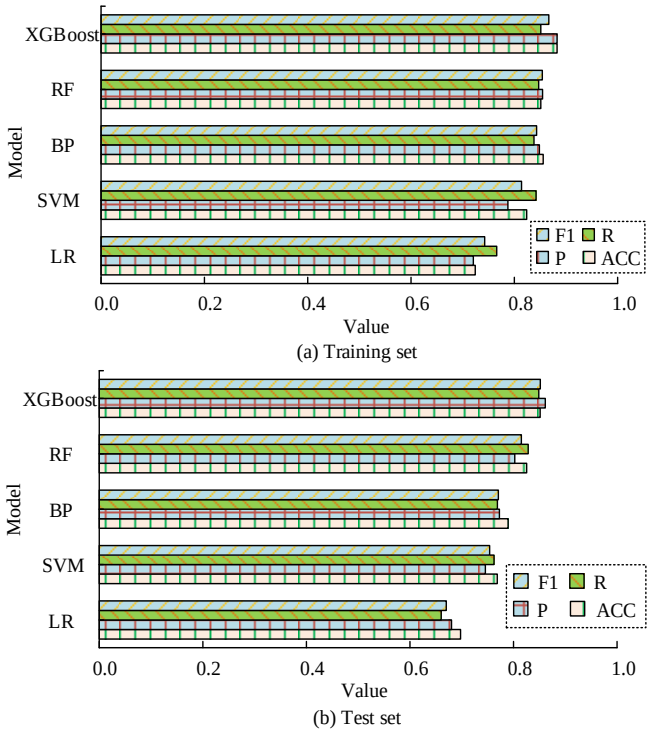

Fig. 7. Performance results of five common machine learning prediction system.

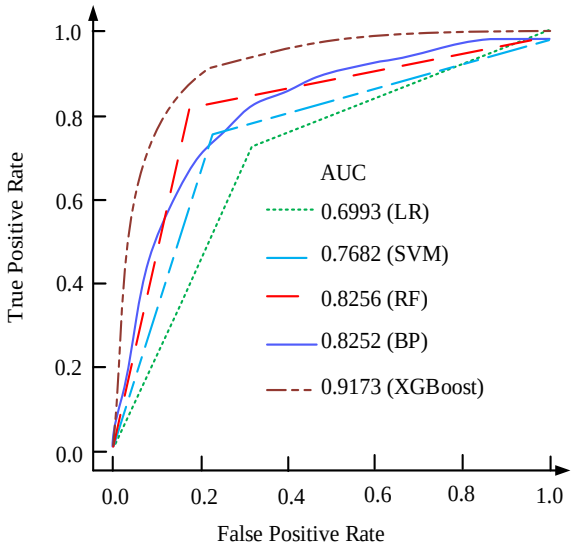

Fig. 8. ROC curves of 5 systems.

According to Table III, in equidistant stratification, only 17% of patients were classified into the low confidence group, which meant that the workload of manual evaluation was reduced by 83%. Among patients stratified using k-means, only 9% were classified into the low confidence group, which meant that the workload of manual evaluation was reduced by 91%. Compared with equidistant layering methods, k-means stratification required less manual evaluation workload. Therefore, k-means was selected as an evaluation indicator for the risk of relapse and readmission. In the current situation, if the balance point between different risk types is determined as a low risk or high-risk patient layer with a prediction probability of [0.286, 0.507] and low reliability, decision-makers only need to spend their limited resources to study these cases. That is, only 9% of the total cases need to be further manually evaluated. Through this approach, high-risk patients can achieve a balance between risk, cost, and resources.

Experimental validation confirms that among the three layers of the overall model for predicting recurrence and risk classification in COPD patients, each layer has the best performance. These three layers were integrated, and simulation analysis was conducted on the recurrence prediction and risk classification of COPD patients to verify the overall model's prediction and classification results in Fig. 9.

Fig. 9(a) and Fig. 9(b) show the fitting of the overall model for patient recurrence prediction and risk classification, respectively. According to Fig. 9, the simulation accuracy of the overall model for predicting patient recurrence and risk classification is 97.3% and 96.4%, respectively. This model can accurately predict the recurrence probability and risk classification of COPD patients, which is beneficial for providing decision-making assistance suggestions for hospitals.

TABLE III. PREDICTED RESULTS OF RISK GROUP FOR RECURRENT READMISSION PATIENTS

Method	Confidence level	Probability	Discard data (%)	Retained data	Accuracy (%)	Sensitivity (%)	Specificity (%)	Class I error	Class II error
Equidistant partition	All	All	0	1335	85.09	71.39	90.03	82	121
	Low	<0.4,>0.6	4	1287	83.92	71.82	91.63	70	115
	Medium	<0.3,>0.7	16	1110	86.48	73.24	94.75	46	103
	High	<0.2,>0.8	37	830	89.05	73.13	74.76	23	68
	Highest	<0.1,>0.9	76	332	94.32	76.52	99.62	2	19
K-means clustering and partitioning	All	All	0	1345	85.26	71.53	90.35	82	117
	Medium	<0.286,>0.507	10	1220	87.53	77.64	87.96	91	62
	Highest	<0.156,>0.759	46	751	90.21	82.94	93.56	33	42

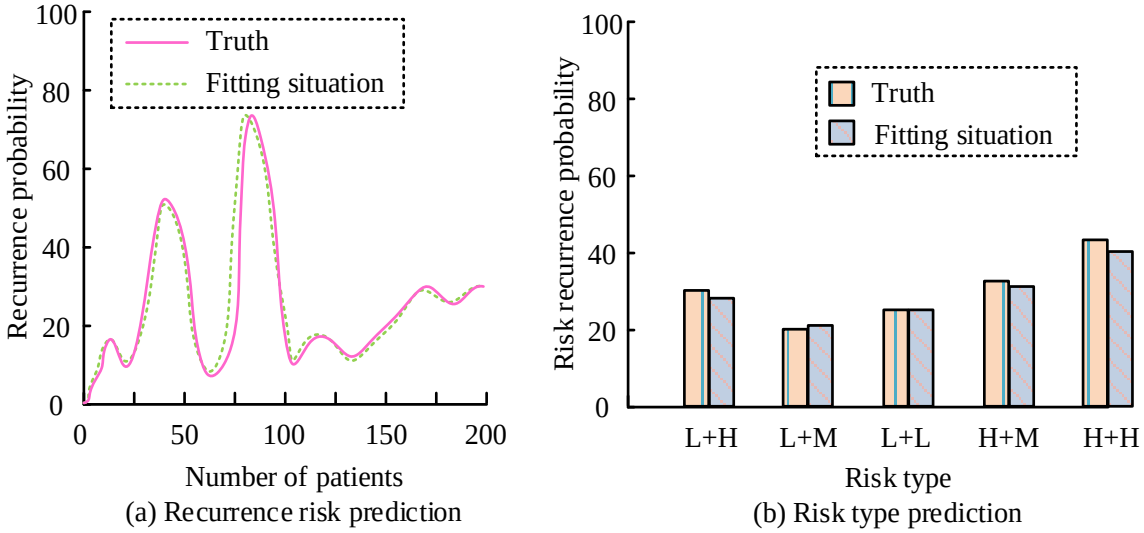

Fig. 9. Simulation effect of the overall model.

V. Discussion

Through the analysis of the results, the BiLSTM-CRF model performs the best in entity recognition tasks, followed by BiLSTM, and CRF performs the weakest. The BiLSTM-CRF model combines BiLSTM with CRF, the former can capture long-range dependent features of data, and the latter can consider constraints between labels in sequence prediction problems. This combination enables the model to perform well in entity recognition tasks with strong dependencies. Although BiLSTM performs well in capturing the contextual information of data, the lack of CRF layers makes it unable to utilize global information in label sequence prediction, resulting in weaker performance than the BiLSTM-CRF model. The simple CRF model only considers the relationship between labels and ignores contextual information, so its performance is the weakest. Entity recognition shows good recognition performance for "comorbidities" and "examination indicators", which may be due to the standardized expression of these entity types in the text. Entities such as "lifestyle habits" and "disease course" have poor recognition performance due to their diverse or vague expressions. The effectiveness of entity recognition is closely related to the quantity and quality of annotations, and more annotated samples and high-quality annotations can improve the accuracy of entity recognition.

The XGBoost model performs the best in predicting recurrence in COPD patients. XGBoost is an efficient gradient boosting algorithm that enhances model performance by continuously reducing errors during iterations. The high or low AUC value also reveals XGBoost's classification ability. In terms of risk classification, the k-means clustering algorithm reduces the workload of manual evaluation compared to equidistant partitioning, indicating that k-means can more effectively classify patients into different risk groups. This method supports hospitals to allocate resources more accurately, focusing on patients with moderate recurrence risk and lower credibility. Finally, the accuracy of the overall three-layer model in predicting recurrence and risk classification of COPD patients is verified through simulation analysis. This indicates that combining different analytical methods can comprehensively evaluate and predict the recurrence risk of patients, thereby providing more accurate support for hospital decision-making.

The entity recognition results indicate that the quantity and quality of annotated samples can significantly affect the accuracy of recognition. In the recurrence prediction model, XGBoost performs excellently with its efficient iterative performance. The selection of risk classification methods has a significant impact on reducing the workload of manual assessment, and the k-means method has shown advantages. Combining different methods can improve the accuracy of predictions and help hospitals make more precise decisions in resource allocation and decision support.

VI. Conclusion

Predicting the recurrence and hospitalization of frequent COPD patients that can be prevented, as well as understanding the reasons for recurrence and hospitalization, are important tasks to ensure the health of the people. This study uses ML to predict the recurrence probability of COPD patients and classify their risks, providing a scientific basis for medical institutions. These studies have confirmed that the accuracy rates of these three recognition methods BiLSTM, CRF, and BiLSTM-CRF are 84.36%, 83.67%, and 87.40%, respectively. BiLSTM-CRF has the best recognition effect for "comorbidities" and "examination indicators", with F1 greater than 90%. For F1, the "lifestyle habit" is 74.00% and the "course of disease" is 74.39%, indicating the difficulty in distinguishing between these two types. In the prediction of recurrence, the accuracy rates of XGBoost, BP, RF, SVM, and LR on the training set are 0.8827, 0.8559, 0.8512, 0.8239, and 0.7244, respectively. Their accuracy rates on the test set are 0.8514, 0.8254, 0.7896, 0.7687, and 0.6978, respectively. Among patients stratified using k-means, only 9% are classified into the low confidence group, which means that the workload of manual evaluation is reduced by 91%. Compared with the equidistant stratification method, the k-means stratification requires less manual evaluation workload, and the overall model has a more accurate fit for predicting recurrence and risk classification of COPD patients. There are still many shortcomings in this study, such as insufficient data scale and lack of data diversity. Future work will expand the dataset size to include more data from COPD patients to improve the model's generalization ability. And the diversity of data should be increased to ensure that patient data cover different geographical regions, races, genders, and age groups to enhance the comprehensiveness and adaptability of the model.

References

[1] Wang P X, Xu Y, Sun Y F, Cheng JW, Zhou K Q, Wu SY, Hu B, Zhang ZF, Guo W, Cao Y. Detection of circulating tumor cells enables early recurrence prediction in hepatocellular carcinoma patients undergoing liver transplantation. Liv. Int, 41(3):562-573(2021).

[2] Ye Z, Zhang Y, Liang Y, Lang J, Yang Cervical Cancer Metastasis and Recurrence Risk Prediction Based on Deep Convolutional Neural Network. Cur. Bio, 2022, 17(2):164-173.

[3] Chen Z. Research on internet security situation awareness prediction technology based on improved RBF neural network algorithm. Jou. Com. Cog. Eng, 1(3): 103-108(2022).

[4] Kawahara D, Nishibuchi I, Kawamura M, Yoshida T, Nagata Y. Radiomic Analysis for Pretreatment Prediction of Recurrence after Radiotherapy in Locally Advanced Cervical Cancer. International Journal of Radiation Oncology, Biology, Physics, 2021,111 (3S):E93-E93.

[5] Xiaoke Z, Yu H, Liang Z, Tao L, Zhang M .A prognostic nomogram for predicting risk of recurrence in laryngeal squamous cell carcinoma patients after tumor resection to assist decision making for postoperative adjuvant treatment. Journal of Surgical Oncology, 2019,120(4):698-706.

[6] Chan L, Sadahiro S, Suzuki T, Okada K, Miyakita H, Yamamoto S, Kajiwara H.Tissue-Infiltrating Lymphocytes as a Predictive Factor for Recurrence in Patients with Curatively Resected Colon Cancer: A Propensity Score Matching Analysis.. Oncology, 2020, 98(10):680-688.

[7] Lafaie L, Thomas Célarier, Goethals L, Pozzetto B, Botelho km Evers E. Recurrence or Relapse of COVID-19 in Older Patients: A Description of Three Cases. Journal of the American Geriatrics Society, 2020,68(10):2179-2183.

[8] Dowsett M. Integration of Clinical Variables for the Prediction of Late Distant Recurrence in Patients with Estrogen Receptor- Positive Breast Cancer Treated With 5 Years of Endocrine Therapy: CTS5 (vol 14, pg 234, 2019). Journal of Clinical Oncology, 2020,38(6):656-656.

[9] Postiche H, Piccard M. BiLSTM-SSVM: Training the BiLSTM with a Structured Hinge Loss for Named- Entity Recognition. IEEE transactions on big data, 8(1):203-212(2022).

[10] Benali B A, Mihi S, Moku A, Bazi I EI, Yakhouba N. Arabic named entity recognition in social media based on BiLSTM-CRF using an attention mechanism. Journal of Intelligent & Fuzzy Systems: App. Eng. Tec, 42(6):5427-5436(2022).

[11] Long R, Yang D, Liu Y. Disease Net: A Novel Disease Diagnosis Deep Framework via Fusing Medical Record Summarization. IAE. Int. Jou. Com. Sci, 49(3 Pt.2):808-817(2022).

[12] Puh K, Babac M B. Predicting sentiment and rating of tourist reviews using machine learning. Jou. Hos. Tou. Ins, 6(3):1188-1204(2023).

[13] Matheson A M, Parraga G. Machine Learning Predictions of COPD Mortality: Com.Hid. See. Che, 158(3):846-847(2020).

[14] AL khadar H, Meluskey M, White S, Ellis I, Gardner A. Comparison of machine learning algorithms for the prediction of five-year survival in oral squamous cell carcinoma. Jou. Ora. Pat& Med, 50(4):378-384(2021).

[15] Wong N C, Lam C, Patterson L, Shay Egan B. Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy.BJU International, 123(1):51-57(2019).

[16] Paredes A Z, Hyer J M, Salimgarh D I, Moro A, Pawlik TM. A Novel Machine-Learning Approach to Predict Recurrence After Resection of Colorectal Liver Metastases. Ann. Sur. Onc, 27 (13):5139-5147(2020).

[17] Zhang S, Zhu H, Xu H, Zhu G, Li K C. A named entity recognition method towards product reviews based on BiLSTM-attention-CRF. Int. Jou. Com. Sci. Eng, 2022,25(5):479- 489.

[18] Li D, Dong C, Chen Z, Dong Y, Liu J. A combinatorial machine-learning-driven approach for predicting glass transition temperature based on numerous molecular descriptors. Mol. Sim, 49 (6):617-627(2023).

[19] Guo Y, Mustafa Z, & Kaunda D. Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. Jou. Com. Cog. Eng, 2(1), 5–9(2022).

[20] Liu M, Stella F, Homeroom A, Lucas, Peter J F, Lonneke B, Bischoff E.A comparison between discrete and continuous time Bayesian networks in learning from clinical time series data with irregularity. Art. Int. Med, 95(APR.):104-117(2019).

[21] Reito A, Karola K, Pekkanen L, Palomera J. 30-day recurrence, readmission rate, and clinical outcome after emergency lumbar discectomy. Spine, 45(18): 1253-1259(2020).

[22] Lao Y, Yu V, Pham A, Wang T, Sheng K. Quantitative Characterization of Tumor Proximity to Stem Cell Niches: Implications on Recurrence and Survival in GBM Patients. Int. Jou. Rad. Ons. Bio. Phys, 110(4):1180-1188(2021).

[23] Meng T, Huang R, Hu P, Yin H, Song D. Novel Nomograms as Aids for Predicting Recurrence and Survival in Chordoma Patients: A Retrospective Multicenter Study in mainland China. Spine, 46(1): E37-E47(2020).

[24] Lei M, Han Z, Wang S, Han T, Fang S, Lin F, Huang T. A machine learning-based prediction model for in-hospital mortality among critically ill patients with hip fracture: An internal and external validated study. Injury, 54(2):636-644(2023).

[25] Holtkamp L H J, Lo S N, Thompson J F, Spillane A J, Stretch J R, Saw R P M, Shannon K F, Newegg O E, Hong A M. Adjuvant radiotherapy after salvage surgery for melanoma recurrence in a node field following a previous lymph node dissection. Jou. Sur. Ons, 128(1):97-104(2023).

[26] FangY, LuoB, Zhao T, He D, Jiang B, Liu Q. ST-SIGMA: Spatial-temporal semantics and interaction graph aggregation for multi-agent perception and trajectory forecasting. CAAI Transactions on Intelligence Technology, 7(4):744-757(2022).