International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 14, Issue 8 (August Ver. I 2018), PP.01-15

Study on prediction methods of stroke-related pneumonia

[1]Li Chaoli,[3]SunHaimei,[3]Bao Chongming,[1] Zhou Lihua, [1]Kong Bing

[1]*School of Information, Yunnan University, Kunming 650091*
[2]*Department of Neurology, The First Affiliated Hospital of Kunming Medical University, Kunming 650032*
[3]*School of Software, Yunnan University, Kunming 650091*

Fund Projects: National Natural Science Foundation of China (61762090, 62062066); Research Fund of the Center for Diagnosis and Treatment of Neurological Diseases of Yunnan Province (ZX2019-D3-05); Scientific Research Fund of the Department of Education of Yunnan Province (2019J0005)

Corresponding Author: Kong Bing.
*E-mail:1547516229@qq.com*kongbing@ynu.edu.cn

ABSTRACT

Objective: Establish an appropriate prediction model for stroke-related pneumonia, predict whether stroke patients will cause pneumonia, and evaluate and analyze the predictive effect of the stroke-related pneumonia prediction model. Methods: Collect 183 clinical stroke cases (79 cases in the pneumonia group and 104 cases in the non-pneumonia group) on the first day of admission and medical history data. After data processing and correlation analysis, 27 risk characteristics of stroke-related pneumonia were screened out, and logistic regression was used to verify the validity of the risk characteristics. Construct stroke-related pneumonia risk prediction models based on high-correlation features, including naive Bayes, Gaussian Bayes, neural networks, decision trees, and K-neighbor models, and conduct a comparative analysis of predictions. Results: It was identified that the percentage of lymphocytes, the ratio of neutrophils to lymphocytes, etc. are high-risk characteristics that affect the onset of stroke-related pneumonia. The prediction accuracy of the five models are: 85.4%,91%,87.6%,90.6%,88.32%. Conclusion: This study uses the correlation analysis method to screen out the high-risk characteristics that affect the incidence of pneumonia from the original data, and builds a prediction model based on the high-related risk characteristics. Through the analysis and comparison of multiple machine learning prediction models, the Gaussian Bayesian model has high prediction accuracy for stroke-related pneumonia and stable time. It is a better prediction model for stroke-related pneumonia.

Keywords: stroke pneumonia, prediction method, naive Bayes, gauss Bayes,decision-tree

---

Date of Submission: xx-xx-xxxxDate of acceptance: xx-xx-xxxx

I. INTRODUCTION

Stroke-associated pneumonia (SAP) is one of the important dangerous diseases that cause death after stroke, and the diagnosis is difficult [1-3](2020). In the neurology, intensive care unit, the prevalence of pneumonia after stroke is about 52.1%, of which early-onset pneumonia is about 70.6% [4](2010). SAP not only aggravates the neurological and cognitive dysfunction of stroke patients, but also prolongs the hospital stay, increases medical expenses, and increases the burden on families and society [5](2016). Therefore, the prediction of the early onset of stroke-related pneumonia is particularly important.

Most domestic studies on the prediction of stroke-related pneumonia are based on the prediction score table, and SAP is predicted through regression analysis. Han Lin et al [6](2020) explored the clinical predictive value of changes in the number of T lymphocyte subgroups in peripheral blood for SAP and used multivariate Logistic regression and one-way variance to analyze independent predictors of SAP. The results showed that the stroke score of the predictive score table increased. High and low CD4+ cell count has high predictive value for the occurrence of SAP; Yang Hancai et al. [7](2020) measured the C-reactive protein/albumin ratio in elderly patients with acute ischemic stroke and used multivariate Logistic regression to analyze that the increase in the

About the author: Li Chaoli (1994-), male, master's degree student, main research direction is machine learning; Sun Haimei (1972-), female, deputy chief physician, Ph.D. main research direction is neurological diseases; BaoChongming (1971-), Male, associate professor, master's degree, main research direction is multi-agent systems, social network analysis and machine learning; Wang Chongyun (1971-), male, associate professor, Ph.D. main research direction is evolutionary ecology, gene flow and population genetic structure; Zhou Lihua (1968-), female, professor, doctor, main research direction is data mining, machine learning and social network analysis; Kong Bing (corresponding author) (1968-), male, associate professor, doctor, main research direction is social network analysis and Machine learning

ratio could be used to predict the risk of SAP in elderly patients with acute ischemic stroke；ShuZhaorui et al. [8](2020)discussed the prediction scores of 162 SAP cases and concluded that the prediction scores have certain application value for SAP, but there is a lack of relevant large-scale studies to evaluate the impact of these prediction scores on clinical decision-making and prognosis, and its practicality More clinical studies are still needed to verify, and further randomized controlled clinical studies are needed to verify the independent risk factors of SAP, to develop a corresponding predictive score table that meets the characteristics of the population in the region. Foreign predictive analysis models mostly use predictive scoring tables and cytokines to predict SAP. ElżbietaGradek-Kwinta et al. [9](2020)evaluated the clinical effects of single cytokines and their combination factors on SAP prediction and evaluated the classification ability of cytokines through ROC curve analysis. The results showed that in vitro synthetic cytokines have potential effects on the occurrence of SAP; GulistanBahat et al. [10] (2020)used logistic regression analysis and found a high correlation between serum vitamin D content and SAP; PhucDuc Dang et al. [11] (2020)compared the predictive score scale stroke score with the swallowing function assessment scale to predict SAP, and the swallowing function assessment scale ROC analysis area under curvature reached 0.858. The study showed that the swallowing function assessment scale has better SAP predictive value. The prediction accuracy of foreign studies based on statistical methods is relatively high, but the differences in ethnicity, risk characteristics, vascular disease spectrum, etc. are not suitable for the domestic population [12](2017),and statistical methods affect the incidence of pneumonia with too many features, and the prediction time is long. The model is complex and requires a lot of funds to test case data [13](2019).

The use of machine learning, data mining, and other methods to establish a high-precision risk prediction model has been very well used in many medical binary classification data. For example, GopiBattineni [14](2020) and others used support vector machines, logistic regression, and K-nearest neighbors to study breast cancer risk prediction models, and the accuracy of the prediction model was as high as 98.20%; Xuemeng Li et al. [15](2019)optimized the decision tree to analyze the risk factors of stroke in daily life habits; NafizatusSalmi[16](2019)et al. used the naive Bayes classification model to predict colon cancer, and the accuracy of the prediction model was as high as 95.24%; Cao Wenzhe et al. [17] (2016)used a variety of machine learning algorithms to diagnose prostate cancer, and the results proved that the prediction effect of the multi-factor prediction model established by machine learning is better than any single-factor prediction model. Wang Meng et al. [18](2020) predicted cerebral hemorrhage-related pneumonia based on four machine learning algorithms, including four predictors of age, NIHSS score, white blood cell count, and swallowing dysfunction, and four models predicted the correct rate of the test set and training set. In 60%-81%.

This study uses statistical methods and machine learning methods to perform predictive analysis on SAP. First, perform correlation analysis and regression analysis on the real clinical ICU data to verify the effectiveness of the features, and screen out high-correlation features that affect the incidence of stroke-related pneumonia; Then, based on the five algorithms of naive Bayes, Gaussian Bayes, neural network, decision tree and K-nearest based on machine learning, a stroke-related pneumonia prediction model was established and evaluated. Through experimental comparison, it is found that in SAP forecast: (1) Several highly correlated features can predict SAP well, and the prediction effect is better than that of commonly used clinical ROC curve analysis; (2) Gaussian Bayesian model is better than other models in predicting stroke-related pneumonia it is good.

II. Research objects and methods

2.1 Research object

This study uses regression data from the Department of Neurology of a medical university. Including patients admitted to the hospital within 24 hours of acute stroke from September 1, 2017, to December 31, 2019. The basic clinical test data and medical history data of stroke ICU patients on the first day of admission are the research objects. A total of 183 stroke patients in the Department of Neurology were collected, of which 79 cases developed stroke pneumonia during hospitalization and 104 cases did not occur. The patients were divided into two groups according to whether pneumonia occurred or not, and regression analysis was performed.

Inclusion criteria: ①Adults (>18 years old) at admission; ②All patients who were clinically diagnosed with a stroke, after informed consent, all participants underwent brain magnetic resonance imaging or brain CT scan and laboratory examination within 24 hours; ③The blood parameters of the patients were measured upon admission. ④ Exclude other diseases, such as malignant tumors, hematological diseases, immunosuppressive agents, active infections within 2 weeks before admission, severe liver and kidney disease, major trauma, or surgery.

2.2 Research methods

2.2.1 Basic patient information

Collect age, gender, and stroke risk factors (hypertension, diabetes, atrial fibrillation, stroke history, coronary heart disease, current smoking, drinking), hospitalization days and hospital mortality, and other medical history data. Perform laboratory tests including blood glucose measurement and blood parameter determination within 24 hours of admission. The Glasgow Coma Center (GCC) and the National Institutes of Health stroke score were performed to assess the severity of the initial stroke.

Because of the standard irregularities and missing data in the test results and case data, the obvious unreasonable data is discarded and the mean value completion method [19](2011) replaces the characteristic data with missing values. A total of 64 items of raw data were used for basic data analysis using SPSS. The Spearman correlation coefficient value indicates the degree of correlation between the feature and the pneumonia result. The correlation coefficients between the baseline table and Spearman of the main data obtained are shown in Table 1:

Table 1 SPSS basic data analysis

Variable	Non-SAP (n = 104)	SAP (n = 79)	p-value	Spearman |r|
Male, n(%)	66(63.5%)	50 (63.3%)	0.981	0
Stroke type, n(%)	7(6.7%)	38 (48.1%)	0.000	0.503
COPD n (%)	2(1.9%)	2(2.5%)	0.78	0
Myocardial infarction (%)	4 (3.8%)	1(1.3%)	0.289	0
Atrial fibrillation (%)	11(10.6%)	9(11.4%)	0.861	0
Coronary Heart Disease (%)	10(9.6%)	5(6.3%)	0.422	0
diabetes (%)	23 (22.1%)	16(20.3%)	0.761	0
hypertension (%)	74 (71.2%)	51 (64.6%)	0.342	0
Hyperlipidemia (%)	10(9.6%)	2(2.5%)	0.055	0
Heart failure (%)	5(4.8%)	1(1.3%)	0.183	0
gout (%)	3(2.9%)	2(2.5%)	0.885	0
Smoking (%)	35 (33.7%)	50 (63.3%)	0.000	0.312
Drinking (%)	27 (26.0%)	50 (63.3%)	0.000	0.393
PPI n (%)	62 (60.2%)	70 (88.6%)	0.000	0.321
age, mean±SD	62.18±13.12	67.40±13.82	0.01	0.192
body temperature, (°C) mean±SD	36.57±0.25	36.79±0.71	0.004	0
Heart rate mean±SD	75.46±13.42	82.93±18.49	0.002	0.221
Breathe	18.44±1.76	18.79±4.63	0.475	0
Systolic blood pressure	147.15±23.59	151.81±30.94	0.249	0
Diastolic blood pressure	84.63±14.68	86.20±18.37	0.522	0
leukocyte (×109 /L)	7.45±2.24	11.87±4.16	0.000	0.583
Neutrophil absolute value (×10 9 /L)	5.08±2.24	10.10±4.11	0.000	0.673
Absolute lymphocyte value	1.73±0.60	1.07±0.49	0.000	0.699
Absolute value of monocytes	0.49±0.17	0.63±0.28	0.000	0.251
NLR	3.65±3.37	12.28±9.62	0.000	0.698

Neutrophil to monocyte number ratio	11.49±8.92	20.16±16.36	0.000	0.451
MLR	0.322±0.194	0.698±0.479	0.000	0.592
Hemoglobin (g/dL)	146.45±18.38	136.58±21.13	0.001	0.228
Hematocrit	43.77±5.45	40.71±5.86	0.000	0.25
Eosinophil percentage	0.5±7.5	1.66±7.24	0.000	0.576
Glucose glucose mean ± SD	6.27±2.46	9.77±9.33	0.000	0.49
Red blood cell volume distribution width	13.36±1.08	13.63±1.56	0.165	0
platelet (×109 /L)	207.58±89.93	196.55±67.99	0.364	0
Mean platelet volume	10.93±1.00	11.16±1.42	0.203	0
Mean platelet volume to platelet count ratio	0.060±0.024	0.065±0.030	0.184	0
Platelet to lymphocyte ratio	131.52±63.50	239.33±211.74	0.000	0.439

As can be seen from Table 1, in the stroke case data, the percentage of lymphocytes, the Neutrophil to Lymphocyte Ratio (NLR), the percentage of neutrophils, the Monocyte to Lymphocyte Ratio (MLR), white blood cells, The percentage of eosinophils, the absolute value of lymphocytes, and the type of stroke have a high correlation with the occurrence of pneumonia.

Based on correlation analysis, regression analysis is further used to analyze the correlation degree of quantitative changes between variables to verify the validity of relevant characteristic data. The binary Logistic regression was used to further analyze the association between the above characteristics and the occurrence of pneumonia, and the characteristics with a significantly less than 0.05 were included in the model (significance less than 0.05 is considered to be statistically significant). As shown in table 2:

Table 2 Logistic regression analysis related data

Test result variable	B（Regression coefficients）	SE（Standard error）	Wald（Chi-square value）	Significance
Percent Lymphocytes	-.232	.033	49.627	.000
Eosinophil percentage	-.852	.187	20.754	.000
Absolute lymphocyte value	-2.321	.372	39.010	.000
Stroke type	-2.773	.475	34.035	.000
Albumin	-.032	.010	9.559	.002
Smoking	-24.003	5.781	17.236	.000
Percentage of monocytes	-.220	.070	9.953	.002
Red blood cell	-.829	.248	11.184	.001
Hemoglobin	-.026	.008	10.168	.001
Hematocrit	-.099	.029	12.003	.001
Body temperature	1.025	.366	7.848	.005
Creatinine	.013	.005	7.575	.006
PPI	3.613	.876	17.013	.000
Absolute value of monocytes	2.602	.709	13.492	.000
Drinking	11.969	2.314	26.748	.000

Platelet to lymphocyte ratio	.011	.002	22.024	.000
Heart rate beats/min	.034	.011	10.001	.002
Glucose	.343	.071	23.160	.000
Routine blood leukocytes	.487	.075	42.347	.000
Monocyte to lymphocyte ratio	6.101	1.053	33.588	.000
Neutrophil absolute value	.549	.080	47.057	.000
Percentage of Neutrophils	.171	.024	50.943	.000
Neutrophil to lymphocyte ratio	.373	.060	38.984	.000

The data in Table 2 is a regression analysis of the risk characteristics in Table 1, and the significance is consistent with the Spearman correlation results, indicating that the risk characteristics are effective.

2.2.2 Classification prediction method

There are many classification models, such as Bayesian classification, decision tree, neural network, KNN, etc. This article mainly uses the naive Bayes model, Gaussian Bayes model, decision tree model, K-neighbor algorithm and neural network model Perform classification forecasts on SAP. The discussion on this is shown in Table 3:

Table 3 Classification process and analysis of five machine learning models

Model	Classification process	Pros and cons
MultinomialNB	The training set calculates the prior probability and the posterior probability and analyzes the received input. Under the condition of the known classification probability, the category with the maximum probability value of the input to be processed belongs to a certain category is the data the type.	The classification efficiency is stable, the model is simple, the classification effect is good for small-scale data, and it is not sensitive to missing data. However, there is often a certain correlation between each attribute. When the attribute correlation is small, the Naive Bayes classifier has a good classification effec[t[17][20]-[21](2020)
DecisionTreeClassifier	The sample-set generates a decision tree, and a set of hierarchical rules is inferred from a set of disordered and irregular cases by determining the logical (branch) relationship of "if-then". The probability distribution of all possible outcomes is expressed in a tree diagram to generate a decision tree.	The decision tree is easy to cause an over-fitting phenomenon, which has good diagnostic results on training data but has no good diagnostic effect on test data [22](2019)
KNeighborsClassifier	Calculate the distance between the tested patient and the training set instance points of all known label categories, sort them in ascending order of distance, select the k training instance points with the smallest distance from the current, and finally put the highest frequency in these k instances points The category is used as the category of the new patient data point.	The choice of k value has a greater impact on the model. If the k value is small, the prediction result is very dependent on the instance points of the neighbors, and overfitting is prone to occur; if the k value is large, the instance points far away from the input instance will affect the prediction results, and underfitting is prone to occur. For small sample data, the accuracy of each k value is not much different [23](2018)
MLPClassifier	The training stage provides a series of input and output data sets to the neural network. Through numerical calculation methods and parameter optimization techniques, the weighting factors of node connections are continuously adjusted until the desired output can be produced from the given input.	The artificial neural network prediction results are relatively stable [24]-[26](2019), which describes the subjectivity and mindset of the human brain in judgment problems, and can avoid the fatigue of the human brain in the decision-making process [27](2006). Extracting more feature data and momentum constant requires the "trial-and-error" method, which is extremely time-consuming and not conducive to predicting the result label of pneumonia

		[28](2008).
GaussianNB	The principle of the maximum posterior probability is adopted, that is, the posterior probability is only proportional to the conditional probability and the prior probability, which is converted into a maximum likelihood problem. Maximum likelihood estimation obtains the parameters of Gaussian distribution-mean and variance and uses probability density to get the probability that a sample belongs to each category.	The Gaussian Bayesian model prediction process uses the vector mean and the covariance matrix to calculate the probability, and train and test the data. Each feature attribute does not need to be independent, the prediction process is simple, and the accuracy is stable, which is wider than the use conditions of Naive Bayes [29](2002).

III. Results and analysis

3.1 Determination of experimental data set

To verify the effectiveness of the proposed method, based on the initial data (79 cases of pneumonia group and 104 cases of the non-pneumonia group), 7 pneumonia data sets (shown in Table 4) were extracted to analyze and compare the naive Bayesian model and neural network. Classification performance of five machine learning models: network model, decision tree model, K-neighbor model, and Gaussian Bayes model. The prediction process test set and training set are tested at 3:7, and the common clinical ROC curve analysis and classification model are used for comparison.

The size of the Spearman correlation coefficient is used as the basis for determining the experimental data group. A total of 27 items with a Spearman coefficient greater than 0.2 (strong correlation or moderate correlation) are experimental data objects. The data are sorted according to the correlation of the characteristic data, and the top 20 Items, the first 15 items, the first 10 items, and the first 5 items are used as experimental data sets to test whether fewer high-correlation features can obtain higher prediction accuracy.

Table 4 Data feature selection and purpose

Datasets	Description	Feature number	Purpose
Dataset1	Overall data	64	Verify the prediction accuracy change value of the overall data and related feature data
Dataset2	Spearman related	27	Verification of correlation feature data has a better predictive effect on pneumonia
Dataset3	Overall ROC analysis feature items	9	Verification that machine learning has a better prediction accuracy for highly correlated feature items than ROC analysis
Dataset4	Top 20 items sorted by relevance	20	Test whether fewer highly correlated features can achieve higher prediction accuracy
Dataset5	Spearman related (first 15 items)	15	
Dataset6	Spearman related (first 10 items)	10	
Dataset7	Spearman related (first 5 items)	5	

3.2 Results and analysis of the model

In the study, 128 of the 183 samples were randomly selected as training samples, and the remaining 55 were used as test samples. Among 64 independent variables, it was found that the percentage of lymphocytes(LY), the percentage of neutrophils to lymphocytes(NLR), the percentage of neutrophils(NEUR), the percentage of monocytes to lymphocytes(MLP), the white blood cells(WBC), the percentage of eosinophils(EOSR), the type of stroke(ST), etc. are strokes. The important influencing factors of pneumonia ($|r|<0.4$ are weakly correlated or not correlated), and the characteristic risk factors with higher correlation (10 items) are shown in Figure 1:

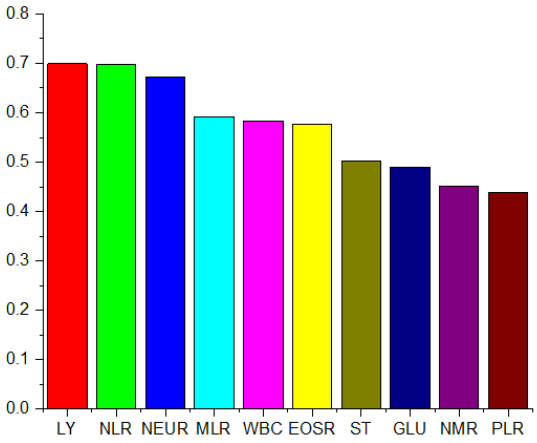

Figure 1 Main model risk attributes

3.2.1 ROC curve analysis

ROC analysis has been widely used in clinical diagnosis and treatment and population screening research. Take each test result as the possible diagnostic threshold, and draw the curve with the calculated sensitivity as the ordinate and 1-specificity as the abscissa. The size of the AUC indicates the accuracy of the diagnostic test. AUC has been generally recognized as an accurate indicator for the authenticity evaluation of a diagnostic test. For a diagnostic test, the diagnostic value is low when the AUC is between 0.5 and 0.7, which is between 0.7 and 0.9. The diagnostic value is moderate in time, and the diagnostic value is higher when it is above 0.9 [30](2006). According to the characteristic data, the ROC curve of the SAP risk assessment results of ICU patients is drawn, as shown in Figure 2. Table 5 shows the results of ROC curve evaluation characteristics versus SAP.

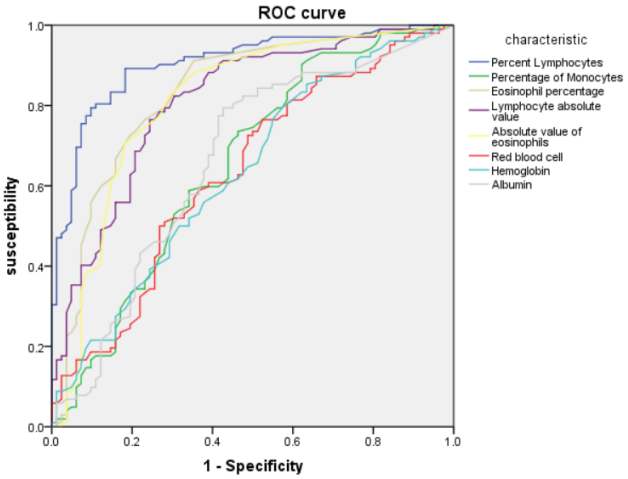

Figure 2 The diagnostic value of ROC curve evaluation features for SAP

Table 5 ROC curve evaluation characteristics versus SAP results

Test result variable	AUC	P	95CI%	Optimal cut off value	Specificity	Sensitivity	Accuracy
Percent Lymphocytes	0.91	<0.001	0.86-0.95	13.8	88.24%	81.70%	85.80%
Percentage of Monocytes	0.65	<0.001	0.57-0.74	4.42	90.20%	37.80%	69.40%
Eosinophil percentage	0.83	<0.001	0.77-0.89	0.03	82.35%	70.70%	72.70%
Red blood cell	0.63	<0.001	0.55-0.72	4.47	76.47%	47.60%	63.40%
Hemoglobin	0.63	<0.001	0.55-0.71	133.5	79.41%	42.70%	64.50%
Albumin	0.66	<0.001	0.58-0.74	35.15	79.41%	57.30%	69.40%

ROC curve analysis results show that the percentage of lymphocytes and the percentage of eosinophils produce higher AUC values than smoking, percentage of monocytes, red blood cells, hemoglobin, and albumin. Besides, the best cut-off value of lymphocyte percentage for SAP is 13.8, specificity is 88.24%, and sensitivity is 81.7%. After calculation, the optimal cut-off value of lymphocytes has a prediction accuracy of 85.8%% for SAP, which has a high diagnostic value for SAP patients.

3.2.2 Machine learning model analysis

Based on seven data sets, five machine learning algorithms: Naive Bayes, Gaussian Bayes, decision tree, K-nearest, and neural network, are used to construct probabilistic prediction models and compare the results. Test patients internally verify the model built on the training set. The results show that the prediction accuracy is 0.854, 0.91, 0.906, 0.883, 0.876. As shown in Figure 3:

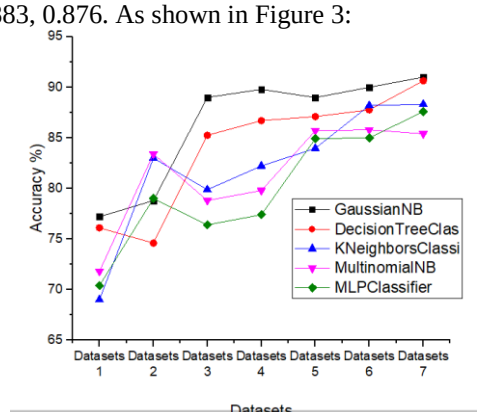

Figure 3 Comparison of classification accuracy of five models on seven data sets

As can be seen from Figure 3, (1) For the data set: ① When the model predicts the data with relevant characteristics after statistical data processing, the prediction accuracy is significantly improved compared to the overall data, indicating the relevant characteristic data Have an impact on SAP's forecast accuracy;②Traditional ROC analysis of overall feature data and Spearman-related (top 10) feature data have similar prediction accuracy, but the model prediction accuracy is less than that of Spearman's related (top 5) feature data, indicating that traditional ROC analysis has overall feature data and SAP Very strong correlation, but the correlation is weaker than Spearman correlation (top 5) characteristic data;③As the selection of related feature data feature items decreases, the accuracy of the model becomes higher and higher. For feature data with higher correlation, the prediction accuracy of the five models reaches the highest value, and the prediction accuracy of the five models differs by three at the highest value. About a percentage point; (2) Regarding models: K-nearest neighbor model, naive Bayes model, and artificial neural network model show weakness in processing SAP data, and the accuracy of the model has been lower than that of Gaussian Bayes model and decision tree model. Especially on the data set with high data dimension, the performance is more obvious. Therefore, it is not suitable to directly use the K-nearest neighbor model, the Naive Bayes model, and the artificial neural network model when processing data with small sample features. (3) Overall:①The Gaussian Bayes model and the decision tree model have very good processing capabilities for small sample data. The prediction effects on the seven data sets are better than the other three models, but the decision tree model has a prediction accuracy Has been weaker than the Gaussian Bayes model;② The Gaussian Bayesian model can reach a high accuracy rate for each data set, and there will be no "unsuitability" phenomenon, and the results prove that the accuracy of the multi-factor prediction model established by Gaussian Bayes is better than that of any ROC curve analysis sheet. The predictive accuracy of factor features. As shown in Table 6, the Gaussian Bayesian model has the highest precision, recall, and F1 scores, which shows that the actual application value of SAP data set prediction is high.

Table 6 Multi-index evaluation of classification performance of each model

Model	Index			
	Accuracy（%）	Precision（%）	Recall rate（%）	F1 score（%）
MultinomialNB	0.854	0.818	0.844	0.822
DecisionTreeClassifier	0.906	0.809	0.84	0.823
KNeighborsClassifier	0.883	0.863	0.817	0.841
MLPClassifier	0.876	0.82	0.755	0.791

GaussianNB	0.91	0.915	0.891	0.893

Analyzing the performance of a model should not only consider the prediction accuracy, but also the prediction time. The experimental prediction time is shown in Figure 4:

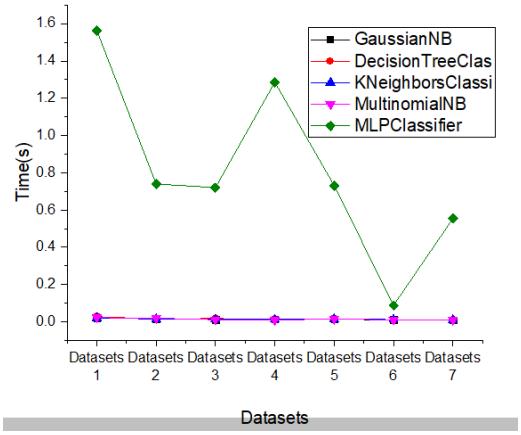

Figure 4 Comparison of classification time of five models on seven data sets

It can be seen from Figure 4 that in the experiment, except the artificial neural network model is extremely unstable in time, the other four models are all very stable. Gaussian Bayesian model prediction results are more reasonable under the conditions of ensuring high prediction accuracy, strong stability, less time-consuming prediction, and avoiding wrong judgments.

IV. CONCLUSION

Combined with the test data and medical history data obtained on the first day of admission, the experiments are performed using common medical statistical methods and artificial intelligence predictive analysis models. Through data processing and data set analysis, the important risk characteristics of stroke-related pneumonia are screened out, and five models of Naive Bayes, Gaussian Bayes, neural network, decision tree, and K-neighborhood are used to construct stroke-related pneumonia predictions The model can be used to predict patients with stroke-related pneumonia and assist doctors in diagnosing the occurrence of pneumonia.

The contributions of this research are as follows: (1) Using a combination of statistics and machine learning. At present, there are few methods to predict SAP using machine learning models in China. This study attempts to use machine learning models combined with statistics to predict the risk of stroke-related pneumonia; (2) Incorporating high-correlation feature factors into the prediction model can predict stroke-related pneumonia well. Through experimental comparison, it is found that the five features with the highest correlation can be used to construct a good prediction model. (3) The use of routine laboratory testing indicators to construct a predictive model does not require expensive and time-consuming special tests, which facilitates the practical application of clinicians and does not increase the burden on patients; (4) In the experiment, the machine learning Gaussian Bayesian prediction model combined with statistical methods has a maximum prediction accuracy of 91% for multi-feature data, which is 5 percentage points higher than the clinical single feature ROC curve analysis accuracy, indicating that the machine learning prediction model is more effective than ROC analysis better. At the same time, due to the limited amount of data, the research process has limitations. In the future, more risk characteristics will be incorporated and the amount of data will be increased, and a variety of machine learning models will be introduced for comparative analysis to obtain a better model while ensuring the reliability and promotion of the model.

In summary, the stroke-associated pneumonia prediction model based on machine learning methods has a high diagnostic value. According to routinely collected data, the key risk characteristics that affect the onset of stroke-associated pneumonia can be quickly and effectively screened from numerous data. And predict its risk of disease, so it has good application value. We plan to integrate this technology into a medical point of care clinical decision support system to provide more accurate stroke-related pneumonia risk prediction.

REFERENCES

[1] Urvish K. Patel, NishanthKodumuri, Mihir Dave, et al. Stroke-Associated Pneumonia: A Retrospective Study of Risk Factors and Outcomes. 2020, 25(3):39-48..

[2] Li X,Wu M,Sun C,Zhao Z,Wang F,Zheng X,Ge W,Zhou J,Zou J. Using machine learning to predict stroke-

associated pneumonia in Chinese acute ischaemic stroke patients.[J]. European journal of neurology,2020.
[3] Sabrina A. Eltringham, Karen Kilner, Melanie Gee, et al. Factors Associated with Risk of Stroke-Associated Pneumonia in Patients with Dysphagia: A Systematic Review. 2020, 35(5):735-744.
[4] Han Jie, Wang Junping, Li Ming. Analysis of risk characteristics of stroke-related pneumonia[J]. Chinese Journal of General Practitioners,2010(09):635-637.
[5] ZHANG X, YU S, WEI L, et la.The A2DS2 Score as a Predictor of Pneumonia and In-Hospital Death after Acute Ischemic Stroke in Chinese Populations[J].PLoS One,2016,11(3):e0150298
[6] Han Lin,Zhang Ying,Gao Yu, Zhang Li.The study of the predictive value of T lymphocyte subsets changes in stroke-related pneumonia[J].Journal of Clinical and Experimental Medicine,2020,19(06):614-617.
[7] Yang Hancai, Xu Zhiding, Peng Keke, OuJiangang, Dong Wentao. The predictive value of CRP/Alb ratio for stroke-related pneumonia in elderly patients with acute ischemic stroke[J]. Chinese Medical Journal,2020,32(05):94-98.
[8] Shu Zhaorui, Wang Bing, Zhang Kefei. The application value of the predictive scoring scale for stroke-related pneumonia[J]. Medical Information,2020,33(01):124-126.
[9] ElżbietaGradek-Kwinta, Mateusz Czyzycki, Kazimierz Weglarczyk, et al. Ex vivo synthesized cytokines as a biomarker of stroke-associated pneumonia. 2020, 510:260-263.
[10] GulistanBahat, SerdarOzkok, SavaşOzturk, et al. Some Comments for Better Understanding of the Study Entitled "Reduced Vitamin D Levels are Associated with Stroke-Associated Pneumonia in Patients with Acute Ischemic Stroke" [Letter]. 2020, 2020(default):159-160.
[11] PhucDuc Dang, Minh Hien Nguyen, Xuan Khan Mai, et al. A Comparison of the National Institutes of Health Stroke Scale and the Gugging Swallowing Screen in Predicting Stroke-Associated Pneumonia. 2020, 2020(default):445-450.
[12] GuDongfeng. Application of cardiovascular disease risk prediction model in the Chinese population[J]. Chinese Medical Information Guide，2017,32（22）：17-17
[13] Wu Juhua, Zhang Shuo, Tao Lei, Jiang Shunjun. Research on the prediction model of stroke risk based on the neural network[J]. Data analysis and knowledge discovery,2019,3(12):70-75.
[14] GopiBattineni, NaliniChintalapudi, Francesco Amenta. Performance analysis of different machine learning algorithms in breast cancer predictions. 2020, 6(23).
[15] Xuemeng Li, Di Bian, Jinghui Yu, et al. Using machine learning models to improve stroke risk level classification methods of China national stroke screening. 2019, 19(6):3651-3654.
[16] NafizatusSalmi, ZuhermanRustam. Naïve Bayes Classifier Models for Predicting the Colon Cancer. 2019, 546(5)
[17] Cao Wenzhe, Ying Jun, Zhang Yahui, et al. Research on prostate cancer diagnosis model based on a machine learning algorithm [J]. China Medical Equipment,2016，31（4）：30-35.
[18] Wang Meng, Qin Lu, Wang Chunjuan, Li Jiao, Wang Yilong, Zhao Xingquan, Wang Yongjun, Li Zixiao. Research on the prediction model of cerebral hemorrhage associated pneumonia based on the machine learning algorithm[J]. Chinese Journal of Stroke,2020,15(03):243-249.
[19] Shen Xue. Research on Missing Data Completion Based on Bayesian Method [D]. Chongqing University,2011
[20] Yi Xia, Li Sheng, Qin Lihua, Chen Xiaoyang, Wang Xiaoyun. The application of Bayesian classifier in TCM syndromes[J]. TCM Research,2013,26(06):4-6.
[21] Yue Xi, Tang Mengxuan. An Improved Naive Bayesian Classification Model Based on Attribute Weighting. 2020, 1550(2):022017.
[22] Feng Yunxia, Zhang Run. Decision tree algorithm for lung cancer diagnosis based on electronic medical records[J]. Computer system application,2019,28(10):257-263.
[23] Wang Haorui. Cardiology online auxiliary diagnosis system based on k-nearest neighboralgorithm[J]. Electronic production, 2018, 000(020):49-51.
[24] Wu Juhua, Zhang Shuo, Tao Lei, Jiang Shunjun. Research on the prediction model of stroke risk based on the neural network[J]. Data analysis and knowledge discovery,2019,3(12):70-75.
[25] Zhang Yuyuan, Zhang Yansong, Liu Bo, et al. Prediction of the length of service at the onset of coal workers' pneumoconiosis based on the neural network.. 2020, 75(4):242-250.
[26] Hongping Hu,Haiyan Wang,Feng Wang,Daniel Langley,Adrian Avram,Maoxing Liu. Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network[J]. Scientific Reports,2018,8(5).
[27] Tang Shaokai, Xu Bin, Li Changxing. Epidemiological characteristics of sexually transmitted diseases in a city and the application of mathematical models in the trend of incidence [J]. Modern preventive medicine,2006,33(1):81- 83.
[28] Zhang Rui, Jiang Pan, Qu Bing. Discussion on the advantages and disadvantages of BP neural network [EB/OL]. Beijing: China Science Paper Online [2008-12-01]
[29] Gordon N J, Salmond D J, Smith A F M.Novel approach to nonlinear/non-Gaussian Bayesian state estimation[J].IEE Proceedings F – Radar and Signal Processing,2002,140(2):107-133
[30] Song Hualing, He Jia, Huang Pinxian, et al. Application research of parametric and non-parametric methods for estimation of area under ROC curve[J]. Journal of Second Military Medical University，2006,27:726-728.