Disease Prediction Model based on Neural Network ARIMA Algorithm

Kedong Li*

Jining Center for Disease Control and Prevention
Jining, China

Abstract—Because the morbidity data of infectious diseases do not only have a single linear or nonlinear characteristic, but also have both linear and nonlinear characteristics, the combination model prediction method is often used to predict the morbidity of infectious diseases in recent years. Compared with the single model prediction analysis method, the combination model can combine the advantages of a single model to extract the effective information contained in the original time series more scientifically and fully. In the context of big data, for the medical field, massive medical data is complex, and the traditional manual data processing method has been unable to meet the current needs. With the help of the computer, data mining can discover new knowledge that is potentially useful and understandable by clearing, integrating, selecting, and transforming the original data. Using data mining, we can organize and reproduce the useful medical knowledge hidden in medical big data. In this paper, an ARIMA-GRNN model is established; the fitting value and the corresponding time are used as the input of the neural network. The actual morbidity is used as the output to train the network and construct the ARIMA-GRNN combined model. Due to the different information flow of BP neural network and neural network, this study also constructed ARIMA-GRNN combined model and ARIMA model, and compared the modeling effect and prediction performance of various models. The average absolute percentage error of the experimental results in this paper is less than 8.63%, and the average absolute percentage error is less than 5%. Compared with other models, it has a better prediction effect, higher accuracy, and more obvious advantages. In this paper, the prediction of disease is dynamic and continuous. It is of great significance for disease prevention and control to use monitoring data to study the epidemic trend and periodic change law, and to make a reasonable prediction.

Keywords—Disease prevention and control; trend prediction; neural network; combination model; ARIMA algorithm

I. INTRODUCTION

At present, feature learning technology is mainly divided into two categories: domain knowledge-driven and data-driven. Domain knowledge- driven methods extract features from image data or non-image data based on domain knowledge. For image data, it is mainly to extract low-level features such as artificially designed texture and wavelet, or qualitative indicators such as capsule and peritumoral blood vessels [1]. Non-image data mainly include some clinical characterization indicators of liver tumors, such as quantitative indicators, age, and gender, or qualitative and laboratory indicators, such as protein level and liver function, quantitative and qualitative characterization indicators in non-image data, and artificial

design features in image data. The heterogeneity between the two brings difficulties to knowledge-based feature learning. From another point of view, the representation fusion between image data and non-image data brings hope to the performance improvement of the feature learning model [2]. Knowledge-driven feature learning method has good interpretability and robustness through artificial feature extraction and algorithm design, but its disadvantage is that it needs artificial participation, low efficiency, and difficult model processing. Data-driven methods are represented by deep learning techniques.

The data-driven neural network has a powerful self-learning ability, which can automatically learn features from a large number of data samples, and is the mainstream of data-driven methods [3]. With the increasing number of convolutional neural network layers and the increasing network width, the low-level pixel-level features can be gradually abstracted into high-level semantic features layer by layer to better extract the rich features hidden in large-scale image data [4]. As an effective learning method for extracting high-level semantic features, convolutional neural networks have achieved good results in many image classification and segmentation tasks. Data-driven feature learning methods are more efficient, but less interpretable and robust, and depend heavily on the number of labeled samples [5]. Especially when the convolutional neural network which has achieved better performance in the field of natural images is directly transferred to the field of medical images with small samples. The network is easy to overfit and has poor robustness [6]. Thus, in small-scale medical image feature learning, the mainstream CNN network is still difficult to meet the VMI. The structure of the network, the training mode, and the scale of the parameters need to be redesigned and adjusted to suit the specific VMI prediction task.

In 2015, Wei Wu et al. used the ARIMA model, ARIMA-GRNN combined model, ARIMA, and feedback dynamic nonlinear autoregressive neural network to establish a combined model based on monthly incidence data of hemorrhagic fever with renal syndrome. The study showed that the prediction performance of the combined model based on a dynamic neural network was higher than that of the static neural network. The prediction accuracy of the combination model is higher than that of the single model [7]. In 2016, Tian Dehong used the monthly incidence data of human brucellosis in China to establish the ARIMA model, BP neural network, ARIMA, and BPNN to establish the combination model. The study showed that the prediction accuracy of the combination

*Corresponding Author.

model was significantly higher than that of the single model [8]. In 2017, Wang Yongbin and others used the incidence data of hand, foot and mouth disease to establish ARIMA model, RBF neural network and ARIMA and RBF neural network to establish a combination model. The results showed that the prediction accuracy of the combination model was better than that of the RBF neural network model and ARIMA model [9]. The above literature analysis shows that compared with the single model prediction analysis method, the combination model can combine the advantages of a single model to extract the effective information contained in the original time series more scientifically and fully.

In this study, the ARIMA model was chosen to establish neural network and Elman neural network, considering that the predicted value of ARIMA model can fit the seasonality and periodicity, and has a similar trend to the measured value. Therefore, the fitting value and time of the disease Arima model are used as the input of the network. The actual morbidity is used as the output to train the network. Its core is to use the nonlinear mapping ability of the neural network to correct the random effect part to improve its prediction accuracy. The realization of the prediction model can realize the prediction of the disease epidemic situation and assist the medical staff to predict and manage the disease epidemic situation. Therefore, this study can effectively assist in the diagnosis and prognosis of the disease, and play an unlimited clinical and social value on the basis of the use of limited medical resources, especially in critically ill patients.

The main contents of this paper are as follows:

1) The background and significance of the research are introduced.

2) The basic theory of the prediction model is introduced.

3) The modeling steps of the BP neural network are analyzed.

4) The empirical part of the combination model is done.

5) Comparison of the fitting diagram of the combined model, comparison of the actual prediction effect and conclusion are done.

6) Conclusions and prospects for the whole paper are made.

II. RELATED WORK

A. Artificial Neural Network

ANN (Artificial Neural Network) is an artificial model based on the function of the human brain and connected with various problems in real life with the relationship between mathematics and physics [10]. The research of artificial neural networks is based on the structure of the biological nervous system. The smallest element in the nervous system of the biological world is the neuron, which consists of nerve cells and multiple processes [11]. The artificial neural network is constructed with reference to biological neurons, and the composition of a single artificial neuron is shown in Fig. 1.
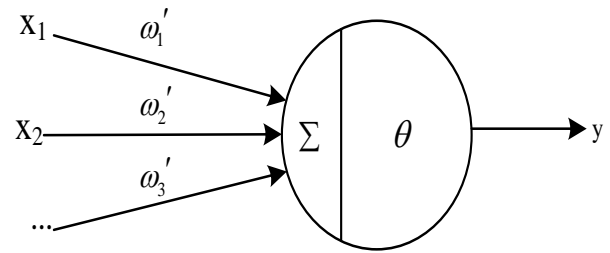

Fig. 1. Diagram of Single Neuron Model.

The composition of artificial neurons mainly includes three elements:

1) A series of input signals and the weight at each connection point represent the strength of each signal. When the connection weight is positive, the neuron is activated; and when the connection weight is negative, the neuron is inhibited.

2) A summation module for integrating all the input signals was used [12].

3) A nonlinear activation function, which acts as a nonlinear mapping by limiting the output interval of the neuron is adopted.

In addition, there is a deviation, namely the threshold θ.

All of the above processes are expressed mathematically as follows:

$$net = \sum_{i=1}^{n} X_i \omega_i; o = f(net - \theta) \tag{1}$$

Where *net* represents the cumulative sum of the input neurons; *o* represents the sum of the neuron outputs; X_i represents the input quantity of the *ith* input neuron, and ω_i represents the connection weight of the *ith* input neuron of this neuron; $f(x)$ is the activation function, which describes the connection between the neuron input and output [13]. The selection of these parameters depends on the size of the training data, the characteristics of the studied sequence and some subjective experience. The quality of parameter selection will play a key role in the final prediction results.

B. Types of Neural Network Structures

The activation functions in the neural network structure are as follows:

1) Hard limit function: The expression for the hard limit function is as follows:

$$y = f(u) = \begin{cases} 1, u \geq 0 \\ 0, u < 0 \end{cases} \tag{2}$$

Or:

$$y = f(u) = \text{sgn}(u) = \begin{cases} 1, u \geq 0 \\ -1, u < 0 \end{cases} \tag{3}$$

Where $\text{sgn}(\square)$ is the sign function. The hard limit function in Equation 2 is also called the single limit function, and the hard limit function in Equation 3 is also called the double limit function.

2) Linear function: The expression for the linear function is as follows:

$$y = f(u) = u \qquad (4)$$

The output neurons of neural networks with linear functions realize the function approximation [14].

3) Saturation linear function: The saturation linear function is expressed as follows:

$$y = f(u) = \frac{1}{2}(|u+1| - |u-1|) \qquad (5)$$

The curve of the saturation linear function is shown in Fig. 2. This activation function is also commonly used in classification problems.
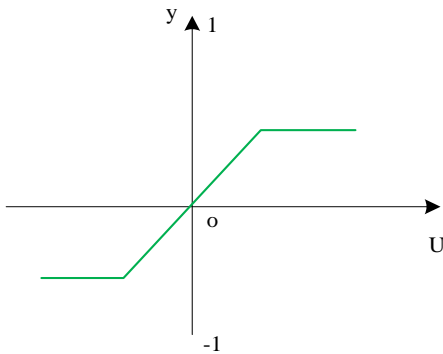

Fig. 2. Saturation Linear Function Diagram.

4) Sigmoid function: Sigmoid function is the most frequently used activation function in neural network algorithms. The Sigmoid function is a strictly monotonically increasing smooth function and has the asymptotic property [15]. The logarithmic tangent function is a form of the Sigmoid function, and its functional form is:

$$y = f(u) = \frac{1}{1 + e^{-\lambda u}} \qquad (6)$$

In the equation, the parameter λ is called the gain of Sigmoid function, which is the slope parameter of sigmoid function. By changing this parameter, sigmoid functions with different slopes can be obtained. The larger the value of λ, the steeper the curve [16]. The logarithmic tangent function, also known as the unipolar Sigmoid function, is differentiable and varies continuously from 0 to 1 [17]. The plot of the unipolar Sigmoid function is shown in Fig. 3.
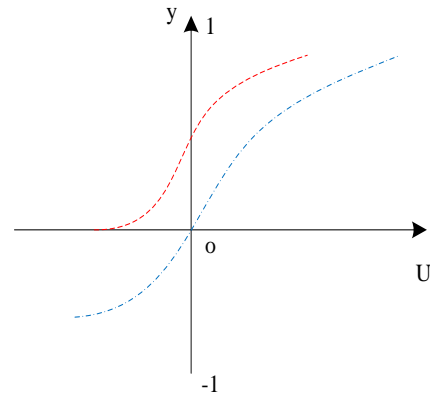

```
-------- Unipolar Sigmoid function
-·-·-·-· Bipolar sigmoid function
```

Fig. 3. Sigmoid Function Diagram.

The sigmoid activation function defined in the equation has a range of 0 to 1. Sometimes, the range of the activation function needs to vary from -1 to 1 and be odd symmetric about the origin. A double tangent sigmoid activation function can be used for this purpose [18]. Its functional form is:

$$y = f(u) = \tanh(\lambda u) = \frac{e^{\lambda u} - e^{-\lambda u}}{e^{\lambda u} + e^{-\lambda u}} \qquad (7)$$

C. BP Neural Network

In the BP neural network, backward propagation is a learning method requiring supervised learning, which is mainly reflected in the training process of BP neural network. Feedforward network is a structure, which is reflected in the network architecture of BP neural network. A typical feedforward neural network is shown in Fig. 4.

The BP neural network has the characteristics of simple structure, easy to use and high efficiency, which is why more and more studies use the neural network [19]. The algorithm of error backward propagation gradually optimizes the connection weights between neurons by iterative processing so that the error between the final output result and the expected result tends to be stable and minimum.
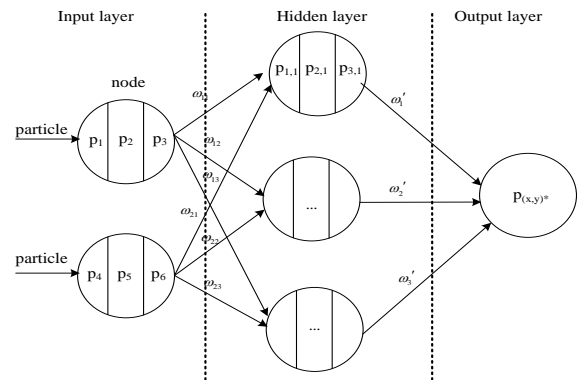

Fig. 4. Typical Feedforward Neural Network Diagram.

III. MODELING STEPS OF BP NEURAL NETWORK

The steps of BP neural network modeling and prediction are as follows:

1) Select samples and construct a training set. The use of appropriate samples is an important prerequisite for model construction [20]. Select the appropriate structure according to the actual situation, and try to make the selected structure contain the maximum information.

2) Data preprocessing: the BP neural network has special requirements for the data of training samples. If the data interval changes too much, it cannot be used as output data.

3) Network structure design: this process includes the selection of the number of network layers, the number of hidden layer nodes, the number of input layer nodes and output layer nodes, learning rate, training function, the selection of hidden layer activation function and the output layer activation function.

4) Initialize that network, and randomly distribute the weight value and the threshold value of each connection.

5) Input the divided data.

6) Recalculate and adjust that weight value and the threshold value of each connection in accord to the error.

7) Obtain the latest parameters before proceeding from step.

8) At the beginning, when the given training times are reached or the output error is not higher than the given error standard, terminate the training.

9) Predict that time series by using the model and obtaining a prediction result.
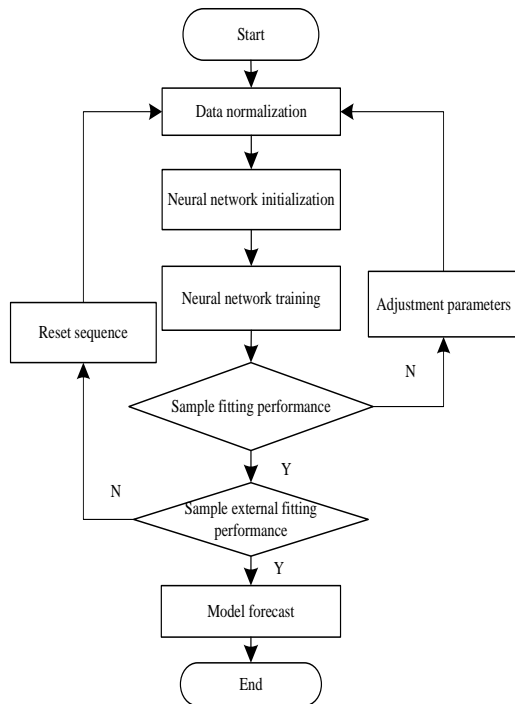
The general process is shown in Fig. 5.

Fig. 5. FLOW Chart of BPNN Model.

IV. EMPIRICAL ANALYSIS OF BP NEURAL NETWORK

The raw data is split into two data sets: a training set and a test set. The training set is used to train the model and select the optimal network model; the test set is used to evaluate the performance of the selected optimal network model [21]. The data before January 2021 is used as the training set, and the data for the whole year of 2021 is used as the test set. By dividing the data set in this way, most of the information contained in the original data can be retained so that the network model can be better learned and trained [22]. At the same time, the problems of over-learning and over-fitting can be avoided.

1) Activation Function: in this study, the Sigmoid function is selected as the activation function of the hidden layer, which can well increase the nonlinear mapping ability of the network [23]. Although this study belongs to the regression algorithm, the Sigmoid function is also selected as the activation function of the output layer because the monthly incidence rate of major diseases is between 0 and 1.

2) The number of neural nodes in each layer, the number of iterations and the learning rate: the number of neural nodes in the input layer and output layer is usually determined by referring to the characteristics of their own data. In this study, according to the splitting of the data set, the number of neural nodes in the input layer is 3, and the number of neural nodes in the output layer is 1 [24]. The empirical formula used in this study to estimate the number of neural nodes in the hidden layer is as follows:

$$m = \sqrt{M + N} + a \tag{8}$$

Where, m represents the number of neural nodes in the hidden layer; M represents the number of neural nodes in the input layer; N represents the number in the output layer; and *a* is an adjustment constant ranging from 1 to 10 [25].

Since the amount of data in this study is not too large, the number of iterations is fixed at 1000. The learning rate is 0.15, and the hidden layer is 3-12 for comparison. The data set is shown in Table I.

TABLE I. TRAINING SET MSE AND TEST SET MSE WITH DIFFERENT NUMBER OF HIDDEN LAYERS

Number	Training set MSE	Test set MSE
3	0.0026	0.0038
4	0.0026	0.0032
5	0.0026	0.0032
6	0.0027	0.0032
7	0.0026	0.0033
8	0.0026	0.0032
9	0.0026	0.0031
10	0.0026	0.0030
11	0.0026	0.0031
12	0.0027	0.0031

In Table I, the selection of the number of hidden layers from 3 to 12 has no obvious effect on the MSE of the training set of data. When the number of hidden layers is from 7 to 11, the MSE of the training set is relatively small. When the number of layers is 10, the test set MSE is the smallest, and the MSE at this time is 0.0030.

V. EXPERIMENTAL ANALYSIS

A. Model Evaluation Comparison Index

Three error evaluation indexes are used to evaluate and compare the prediction effect of each prediction model.

1) Mean Square Error (MSE): The mean square error is the square of the difference between the true value and the predicted value and then averaged over the range $[0, +\infty)$, which is equal to 0 when the predicted value exactly matches the true value, that is, the perfect model; the larger the error, the larger the value. In this study, it refers to the average value of the square sum of the error between the real value and the predicted value of the monthly incidence rate of major diseases. Its calculation formula is formally close to the variance. The formula expression is:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

(9)

2) Mean absolute error: The mean absolute error is the average of the absolute errors. The absolute error represents the absolute value of the deviation between all observed values and the true value. Its range is $[0, +\infty)$. When the predicted value is completely consistent with the true value, it is equal to 0, that is, the perfect model. The larger the error, the larger is the value. Relative to the average error, because the deviation of the average absolute error is absolute, there is no problem of positive and negative offset between the errors, which can better show the true level of the predicted value error than the average error. In this study, the mean absolute error refers to the mean value of the sum of the absolute value of the deviation between the true value and the predicted value of monthly incidence of major diseases, and its formula expression is:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

(10)

3) Mean absolute percentage error: The mean absolute percentage error is the absolute percentage deviation of all individual observations from the true value, and its value range is $[0, +\infty)$. A MAPE of 0% indicates that the model is a perfect model, and a MAPE greater than 100% indicates that the model is an inferior model. The mean absolute percentage error has one more denominator than the mean absolute error. The specific expression is as follows:

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\%$$

(11)

Mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) are all affected by the difference between the real value and the predicted value, but all the deviations are squared or absolute in the calculation process, so there will be no negative number, and there will be no positive and negative offset of errors. It can more accurately reflect the error between the predicted value and the real value.

B. Fitting Effect of each Prediction Model on Monthly Morbidity of Major Diseases

The fitting effect between the monthly incidence rate of major diseases obtained by each prediction model and the national monthly incidence rate of major diseases is shown in Fig. 6. Since the fitting data obtained by these prediction models are not much different from the real data, it is not convenient to display them in the same figure, so each model is compared with the real value separately.
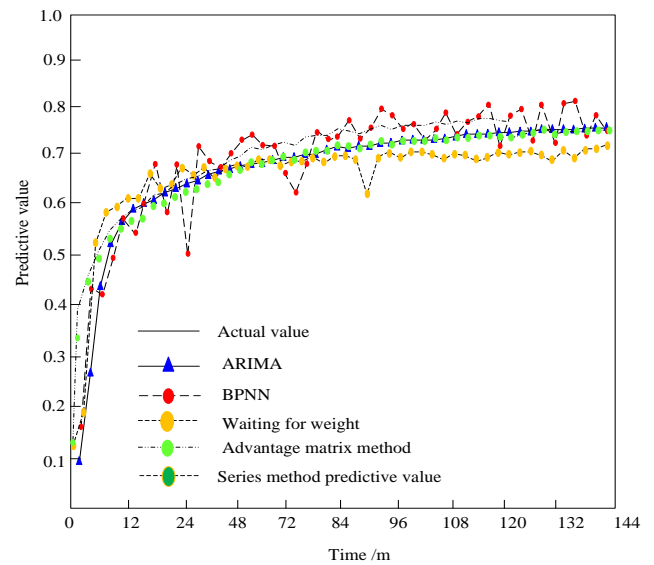

Fig. 6. Comparison between Predicted Value and Real Value by Series Method.

In Fig. 6, the fitting effect of each model is relatively good, which can reflect the change trend of the true value. The fitting effect of the series combination forecasting model is better, especially in the later period.

C. Comparison of Prediction Effect among Different Prediction Models

The annual morbidity data of major diseases in 2021 predicted by all the models used in this study are put into the same table with the real data, as shown in Table II.

TABLE II.   PREDICTED VALUES OF EACH MODEL

Time	Real value	BPNN	ARIMA	Equal weight method	Dominance matrix method	Series connection method
2021.1	0.2380	0.2513	0.2367	0.2440	0.2415	0.2273
2021.2	0.1840	0.2483	0.2696	0.2589	0.2625	0.2596
2021.3	0.3835	0.3801	0.4037	0.3919	0.3958	0.4121
2021.4	0.3339	0.3528	0.3471	0.3500	0.3490	0.3446
2021.5	0.4023	0.3731	0.3944	0.3838	0.3873	0.4008
2021.6	0.4178	0.4216	0.4458	0.4337	0.4377	0.4632
2021.7	0.3804	0.3657	0.4078	0.3868	0.3938	0.4170
2021.8	0.4136	0.3778	0.4119	0.3949	0.4006	0.4221
2021.9	0.4427	0.3886	0.4138	0.4012	0.4054	0.4243
2021.10	0.4188	0.3454	0.3614	0.3534	0.3560	0.3613
2021.11	0.5483	0.4343	0.4613	0.4478	0.4523	0.4819
2021.12	0.5680	0.4798	0.4992	0.4895	0.4927	0.5272

By comparing the monthly prediction results of each model with the real data, it can be seen that the prediction effect of the single ARIMA and series combination model is better than that of other models. The advantage of the single ARIMA model in the first few months is more obvious, which is similar to the real value. In the last few months, the prediction effect of the series combination model is better than that of other models. But on the whole, the prediction effect of all models in the last three months from October is obviously not as good as that of the previous months.

The overall evaluation indicators of each model are shown in Table III.

TABLE III.   EVALUATION INDEXES OF EACH MODEL

	MSE	MAE	MAPE(%)
ARIMA	0.0021	0.0356	10.0636
BPNN	0.0030	0.0427	11.1487
Equal weight method	0.0024	0.0375	10.1644
Dominance matrix method	0.0023	0.0368	10.0996
Series connection method	0.0016	0.0334	9.6914

The MSE, MAE and MAPE of the series combination model are the smallest. The MSE, MAE and MAPE of the single BP neural network are the largest. Therefore, on the whole, the series combination model shows its advantages.

The average absolute percentage error of the single ARIMA model was 10.06%. The average absolute percentage error of the single BP neural network was 11.15%, and the average absolute percentage errors of equal weight method, dominance matrix method and series method were 10.16%, 10.10% and 9.69% respectively. Only when the mean absolute percentage error of the prediction results obtained by the series method is less than 10%, it can be considered that the prediction accuracy of the series method is higher and superior to other models.

If we only want to predict the national epidemic morbidity in the short term, the single ARIMa model has good prediction effect and high accuracy. However, if we want to apply it to the long-term prediction, the advantages of the series combination model are more obvious, and its overall prediction effect is the best.

VI.   CONCLUSION

The disease prediction model studied in this paper has both linear and nonlinear time series characteristics. However, BP neural network just has an excellent performance in nonlinear prediction, so the combination model can make up for the shortcomings of a single model in practical application. It can also make full use of the advantages of ARIMA and BP neural networks, and greatly improve the accuracy of prediction. The experimental results show that the average absolute percentage error of the prediction results obtained by the series method in this paper is less than 10%, and its prediction accuracy is higher and better than other models. It can provide scientific references for disease prevention and control measures.

In this paper, the sample size of the training set is not large enough, and the high dimensionality of the data brings a lot of redundant information, which easily leads to over-fitting of the model and the reduction of prediction performance. To minimize the problems caused by small sample size and high dimensionality data, future work will continue to accumulate and supplement the complete samples of clinical follow-up data, expand the sample size of the training set, and use the new samples to evaluate the two classifiers constructed in this paper.

REFERENCES

[1]   XU S, CHAN H K, ZHANG T. Forecasting the demand of the aviation industry using hybrid time series SARIMA-SVR approach[J]. Transportation Research Part E, 2019, 122.

[2] Witteveen E, Wieske L, Sommers J, et al. Early Prediction of Intensive Care Unit–Acquired Weakness:A Multicenter External Validation Study[J]. Journal of Intensive Care Medicine, 2020,35(6):595-605.

[3] M M K G, G A D, B R J, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration[J]. Annals of internal medicine, 2015, 162(10): 735-736.

[4] Qingui C, Lishan Z, Shanhui G, et al. Prognosis predictive value of the Oxford Acute Severity of Illness Score for sepsis: a retrospective cohort study[J]. PeerJ, 2019,7: e7083.

[5] Ho K M, Williams T A, Harahsheh Y, et al. Using patient admission characteristics alone to predict mortality of critically ill patients: a comparison of three prognostic scores[J]. Journal of critical care, 2015, 31(1):21-25.

[6] Sedloň P, Kameník L, Škvařil J, et al. Comparison of the accuracy and correctness of mortality estimates for intensive care unit patients in internal clinics of the Czech Republic using APACHE II,APACHE IV, SAPS 3 and MPMoIII models[J]. Medicinski Glasnik Official Publication of the Medical Association of Zenica Doboj Canton Bosnia & Herzegovina, 2016, 13(2):82.

[7] Ko M, Shim M, Lee S M, et al. Performance of APACHE IV in Medical Intensive Care Unit Patients:Comparisons with APACHE II, SAPS 3, and MPM0 III[J]. Acute and Critical Care (v.32;2017), 2018, 33(4): 216-221.

[8] Miechowicz J. Prognostic scoring systems for mortality in intensive care units -- the APACHEmodel[J]. Anaesthesiology intensive therapy, 2015,47(1): 46-49.

[9] Salluh J I, Soares M. ICU severity of illness scores: APACHE, SAPS and MPM[J]. Curr Opin CritCare, 2014, 20(5):557-565.

[10] Ferrando-Vivas P, Jones A, Rowan K M, et al. Development and validation of the new ICNARCmodel for prediction of acute hospital mortality in adult critical care[J]. Journal of Critical Care, 2016, 38:335-339.

[11] Hadique S, Culp S, Sangani R G, et al. Derivation and Validation of a Prognostic Model to PredictSix-Month Mortality in an Intensive Care Unit Population[J]. Annals of the American thoracic society, 2017, 14(10): 1556-1561.

[12] L Z, W X, Y W. Prognostic scoring systems for mortality in intensive care units — the APACHEmodel[J]. Research Square, 2020.[183] A. S. Hauser, M. M. Attwood, M. Rask-Andersen, et al. Trends in GPCR drug discovery: new agents, targets and indications[J]. Nature Reviews Drug Discovery, 2017, 16(12): 829-842.

[13] Y. P. Chen, Y. Q. Wang, J. W. Lv, et al. Identification and validation of novel microenvironment-based immune molecular subgroups of head and neck squamous cell carcinoma: implications for immunotherapy[J]. Ann Oncol, 2019, 1(30): 68-75.

[14] B. Li, Y. Cui, D. K. Nambiar, et al. The Immune Subtypes and Landscape of Squamous Cell Carcinoma[J]. Clin Cancer Res, 2019, 12(25): 3528.

[15] D. Bruni, H. K. Angell, J. Galon. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy[J]. Nature Reviews Cancer, 2020, 20(11): 662-680.

[16] Y. Zhang, M. Yu, Y. Jing, et al. Baseline immunity and impact of chemotherapy on immune microenvironment in cervical cancer[J]. Brit J Cancer, 2021, 124(2): 414-424.

[17] R. S. Herbst, J. C. Soria, M. Kowanetz, et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients[J]. Nature, 2014, 515(7528): 563-7.

[18] R. R. Ji, S. D. Chasalow, L. Wang, et al. An immune-active tumor microenvironment favors clinical response to ipilimumab[J]. Cancer Immunology, Immunotherapy, 2012, 61(7): 1019-31.

[19] S. Wang, Z. He, X. Wang, et al. Antigen presentation and tumor immunogenicity in cancer immunotherapy response prediction[J]. eLife, 2019, (8): e49020.

[20] M. O'Hayre, M. S. Degese, J. S. Gutkind. Novel insights into G protein and G protein-coupled receptor signaling in cancer[J]. Current Opinion in Cell Biology, 2014, (27): 126-35.

[21] D. Balli, A. J. Rech, B. Z. Stanger, et al. Immune Cytolytic Activity Stratifies Molecular Subsets of Human Pancreatic Cancer[J]. Clin Cancer Res, 2017, 12(23): 3129.

[22] J. Galon, D. Bruni. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies[J]. Nature Reviews Drug Discovery, 2019, 3(18): 197-218.

[23] D. Sia, Y. Jiao, I. Martinez-Quetglas, et al. Identification of an Immune-specific Class of Hepatocellular Carcinoma, Based on Molecular Features[J]. Gastroenterology, 2017, 153(3): 812-826.

[24] A. Mayakonda, D. C. Lin, Y. Assenov, et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. R Package Version 2.4.05[J]. Genome research, 2018, 11(28): 1747-1756.

[25] F. Martínez-Jiménez, F. Muiños, I. Sentís, et al. A compendium of mutational cancer driver genes[J]. Nature Reviews Cancer, 2020, 20(10): 555-572.