WORDLE Game Prediction Based on BP Neural Network and Other Models

Yimeng Zhang, Zeqi Zhao*, Jinfan Bai

School of Mathematics and Statistics, Shandong University, Weihai, Shandong, 264209

* Corresponding Author Email: 202100820223@mail.sdu.edu.cn

Abstract. The purpose of this paper is to study the factors that affect the difficulty level of words in word games and the related content of their reported results. This paper provides a detailed analysis of the factors that influence the difficulty level of a word and the variability of reported results in word games. The research highlights the efficacy of using a BP neural network for predicting the distribution of reported results, and the findings have important implications for both game developers and players. After dividing the difficulty of words in word games, the model in this article has achieved an accuracy of over 80% in predicting their difficulty. By providing insight into the factors that influence the difficulty level of a word and the variability of reported results, the paper can help game developers design more engaging and challenging word games, while also assisting players in improving their word game performance.

Keywords: Time series prediction; Bp neural networks; Random forests; K-means clustering.

1. Introduction

Wordle is a popular daily puzzle where players guess a five-letter word in six tries using only real English words. The paper focuses on developing a BP neural network model to predict the future distribution of reported results for a given solution word. While the paper also uses time series and regression analysis to understand daily variations in reported results and the effect of word attributes on scores, BP neural networks are the central focus of the paper as they offer a powerful tool for predicting word game outcomes.

For the research of wordle game, Xiang Zhao's team believes that the paper needs to predict the distribution of the number of attempts in the future difficult mode, so the paper can use neural network to establish the MIMO (Multiple-Input-Multiple-Output) BP neural network model. This model encodes words as input [1]. Snell C, on the other hand, proposes that large language models distill broad knowledge from text corpora. However, they can be inconsistent when it comes to completing user specified tasks. This issue can be addressed by finetuning such models via supervised learning on curated datasets, or via reinforcement learning [2]. Of course, wordle games also have a wide range of applications, for example, the paper can Use Wordle in Counseling Ethics Education [3].

To predict daily variations in reported results, the paper uses time series and regression analysis. By analyzing historical data, the paper can identify patterns and trends and fit a time series model to predict future results. Regression analysis can help us understand the effect of word attributes on Hard Mode scores. However, both models have uncertainty, and there is a range of possible outcomes for predicted values.

To classify solution words by difficulty, the paper uses random forest and K-means clustering algorithms. While feature selection is crucial for identifying relevant features such as word length, frequency of use, and origin, BP neural networks are the primary method for predicting word game outcomes. By classifying words into different difficulty levels, the paper can gain insights into the factors that make a word easy or hard. However, the accuracy of the classification model depends on feature selection and model optimization.

The paper uses a comprehensive data set that includes information on a wide range of word attributes such as length, frequency, and part of speech. This data set has the potential to provide insights into changes in language use and cultural shifts over time, making it useful in fields such as linguistics, sociology, and anthropology. However, limitations exist as the data only includes New

York Times words and has a limited time range, so any analysis must be interpreted in the context of these limitations.

In this section, the paper will introduce the neural network model used in this solution. Neural networks are a type of machine learning algorithm that are modeled after the structure and function of the human brain. They consist of input layers, hidden layers, and output layers, with the input layer receiving data, the hidden layers processing the data, and the output layer generating the final output. Neural networks have been successfully applied in various fields, such as image recognition, natural language processing, and time series forecasting.

For time series forecasting, the data is often cyclical in nature and influenced by seasonal factors. In this solution, the paper assumes that the data follows this pattern. The input data for the Bp neural network is normalized, with each feature being between 0 and 1. This simplifies the calculations and ensures that each feature is given equal weight.

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + a_t \quad (1)$$

In constructing decision trees in random forests, the paper uses the CART algorithm. The K-means clustering algorithm is also used to segment the data into clusters. In this case, the paper assumes that the number of samples in each cluster is more evenly distributed, resulting in better generalization performance for new data.

Overall, these assumptions simplify the model calculations while still producing reliable results.

2. Establishment of BP Neural Network Model

2.1. The ARIMA model

First the paper builds an ARIMA model for Number of reported results forecasting. The modeling process of ARIMA time series forecasting model is as follows.

Step 1: AR model building

If the time series X_t is smooth and there is some dependence between the data before and after $X_t$related to the previous $X_{t-1}, X_{t-2}, \ldots X_{t-p}$, independent of the perturbations (white noise) entering the system at its previous moments, withp memory of order, i.e.

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + a_t \quad (2)$$

Let B^k be the k-step lag operator, i.e., $B^k X_t = X_{t-k}$ then the autoregressive model can be expressed as $X_t = \phi_1 B X_{t-1} + \phi_2 B^2 X_{t-2} + \ldots + \phi_p B^p X_{t-p} + a_t$, let $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$, the model can be abbreviated as

$$\phi(B) X_t = a_t. \quad (3)$$

Step 2: MA model building

If the time series X_t(t=1, 2...) is smooth, it is not related to the previous $X_{t-1}, X_{t-2}, \ldots X_{t-p}$ and is related to the perturbation (white noise) of its previous moment into the system, with q memory of order, i.e.

$$X_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \quad (4)$$

Introducing the lag operator and letting$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$, the moving average model can be abbreviated as$X_t = \theta(B) a_t$.

Step 3: ARMA model building

A stochastic process consisting of two parts [4], autoregressive and moving average, is called an autoregressive moving average process. The specific principle is as follows, if the time seriesX_t is (t=1, 2...) smooth, and is also related to the previous$X_{t-1}, X_{t-2}, \ldots X_{t-p}$, then this system is an autoregressive moving average system, constituting$\phi(B) X_t = \theta(B) a_t$, modeled as follows.

$$\theta(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p) = (1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q)a_t \tag{5}$$

Step 4: ARIMA modeling

When the time series X_t does not satisfy the smoothness, the paper usually uses the differencing [5] technique to make the series smooth and then apply the ARMA model.

Parameterd represents the order of the difference. Here is the formula for the difference (Δ is the difference operator):

First-order differential

$$\Delta X_t = X_t - X_{t-1} \tag{6}$$

(2) Second-order differential

$$\Delta^{(2)} X_t = \Delta(X_t - X_{t-1}) = X_t - 2X_{t-1} + X_{t-2} = (1-\Delta)^2 X_t \tag{7}$$

(3) Order Difference

$$\Delta^{(d)} X_t = (1-\Delta)^d X_t \tag{8}$$

The order difference is divided into a smooth series if the initial value series is known. If the initial value of $X_1, X_2, \cdots, X_d$ is known, X_t can be obtained by the following equation.

$$W_t = \nabla^d X_t, t = d+1, \cdots, n, \tag{9}$$

When $d=1$ when X_t the expression of is as follows.

$$X_t = X_1 + \sum_{j=1}^{t-1} W_{j+1} = X_k + \sum_{j=1}^{t-k} W_{j+k}, t > k \geq 1 \tag{10}$$

When $d=2$ when X_t the expression of is as follows.

$$W_t = \nabla^d X_t, t = d+1, \cdots, n, X_t = X_2 + (t-2)(X_2 - X_1) + \sum_{j=1}^{t-2}(t-j-1)\, W_{j+2} = X_k + (t-k)(X_k - X_{k-1}) + \sum_{j=1}^{t-k}(t-j-k+1)W_{j+k}, t > k \geq 2 \tag{11}$$

After differential processing to apply the ARMA model to the time series, the paper uses the AIC minimum constant order criterion to determine the ARIMA model p, q values.

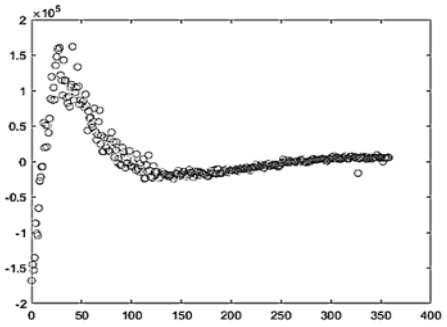


Figure 1. Residual scatter plot.

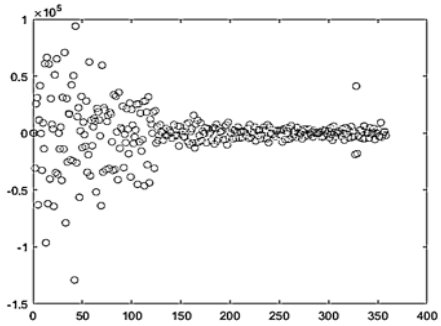


Figure 2. Second-order difference plot.

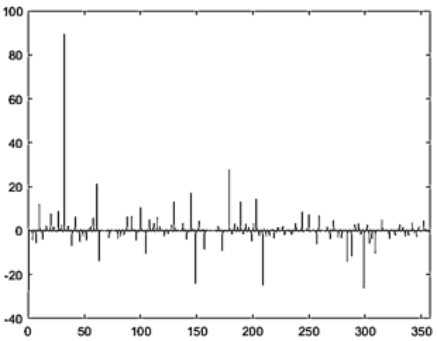


Figure 3. Partial autocorrelation function.

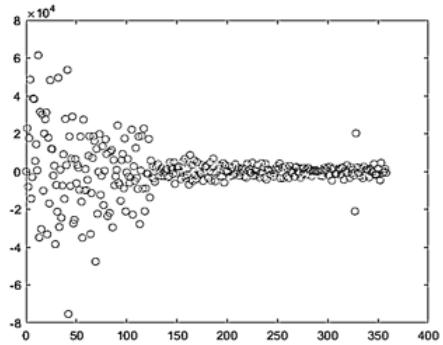
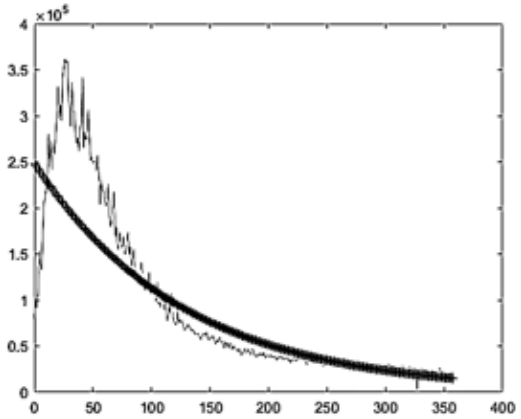


Figure 4. Plot of first-order difference.



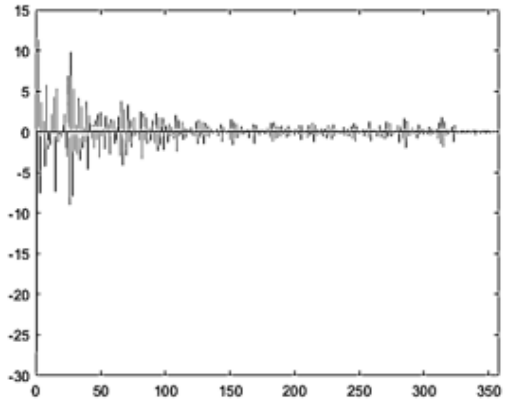


Figure 5. Exponential regression prediction. **Figure 6.** Autocorrelation function.

Solving the ARIMA Model, getting as in Figure 1, 2, 3, 4, 5, 6, the results showed a seasonal pattern with increasing popularity over time. No significant relationship found between word attributes and scores reported, but limited attributes considered may have influenced the result. The model predicts reported results, but no clear relationship found between word attributes and scores reported.

2.2. Build the BP neural network model

The paper uses a neural network to predict reported results for solved words in the future, trained with historical data. It can be implemented with a feed-forward network taking a vector input of words and dates, and outputting a probability distribution of outcomes (1-6, X).

BP neural networks are multilayer feedforward networks trained with error backpropagation, inspired by neurobiology, and widely used in fields such as biomedicine. Based on the continuous refinement and improvement of neural network methods by previous generations, the neural network [6] approach has been widely used in biomedicine with state-of-the-art performance. Their applications include classification, prediction, and modeling, but their randomness means that each operation's results vary. If the training model is saved, subsequent data can be directly uploaded for classification, schematically as Figure 7 [7].

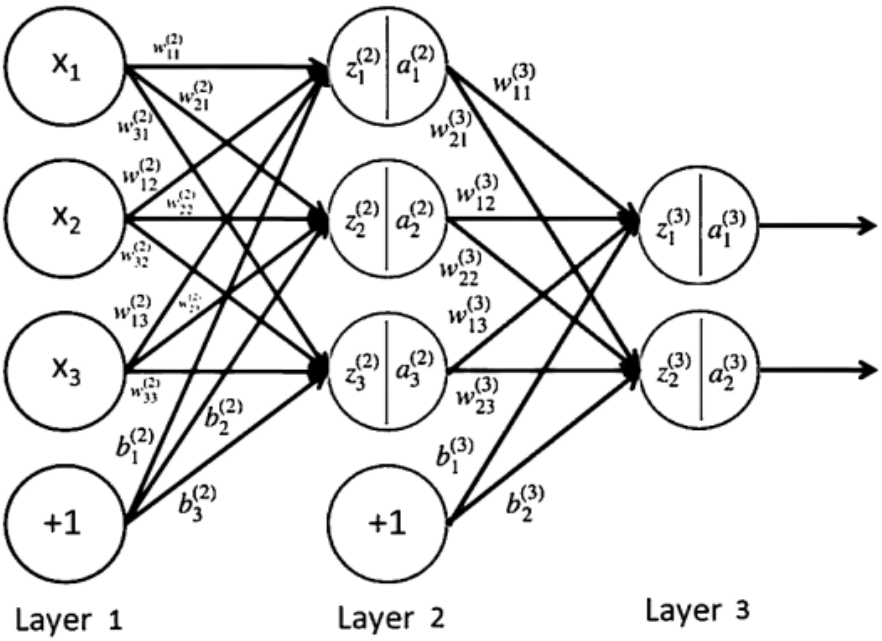


Figure 7. Schematic diagram of the structure of bp neural network.

As can be seen from the figure [8], a neural network is a kind of network which is divided into an input layer, a hidden layer and an output layer. The input layer means that after receiving an input, the result can be output directly after the calculation of the parameters w, b, and is responsible for receiving the input data; the hidden layer is located between the input and output layers and acts as a connection, which is not visible to the outside; and the output layer, which has access to the neural network output data. In this case, there is no connection between neurons in the same layer; each connection has a weight value.

According to the above schematic diagram, the paper first construct and draw the structure of the neural network [9] as follows, which is basically $kx + b$ in the form of a neural network, wherex_1, and x_2 denotes each node, andk_1, and k_2 denotes the respective weights, 1 is the default bias node that exists, and b is the parameter value, and $g(z)$ is the activation function, and a is the output value, and the following equation is obtained.

$$\begin{gathered} g(b + x_1 k_1 + x_2 + k_2) = z \\ g(z) = a \end{gathered} \tag{12}$$

The paper then preprocessed the initial data set using standard normalization to reduce the adverse effects of individual anomalous data on the model prediction results, and the preprocessing equation is shown below.

$$x' = \frac{x-\mu}{\delta} \tag{13}$$

μ Is the mean of all sample data, δ is the standard deviation of all sample data.

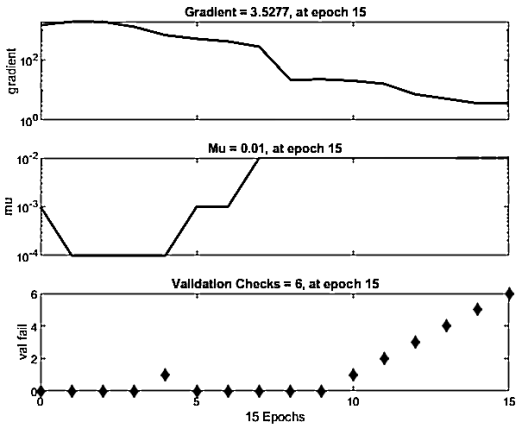


Figure 8. bp gradient diagram.

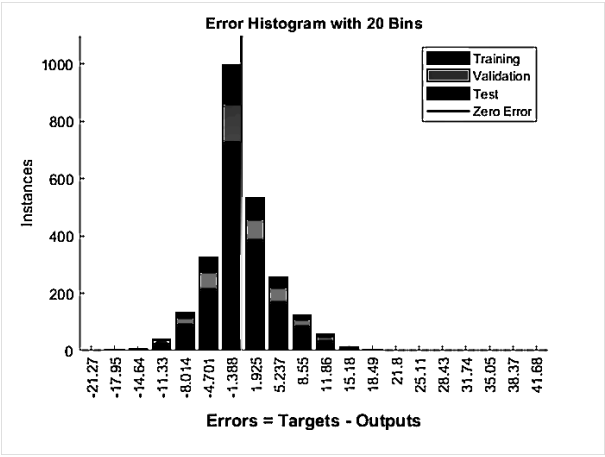


Figure 9. Bp neural network training diagram.

Figure 8 shows the gradients, which are generally decreasing; and shows that the neural network uses a nonlinear least squares strategy with a larger value of the damping factor, which means that the algorithm converges better. Figure 8 also shows the generalization test, where the training set is used to test the error of the validation set for each generation of training completion. A value of 0 on the vertical axis of the plot indicates a decreasing error during the training process. The default is to terminate training when the error does not decrease six times in a row. In figure above, the error of the validation set does not decrease several times in a row starting from the 15th generation of training, which indicates that the training is sufficient and may fall into overfitting if not stopped in time.

Figure 9 in the horizontal coordinates indicate the median of the error interval, the vertical coordinates indicate the number of samples located in the error interval, the perfect situation is the error 0, only one column in the middle 0, the normal situation is the most samples close to 0, the larger the sample the smaller the absolute value of the normal situation, combined with the figure the paper see that the data are in the normal situation.

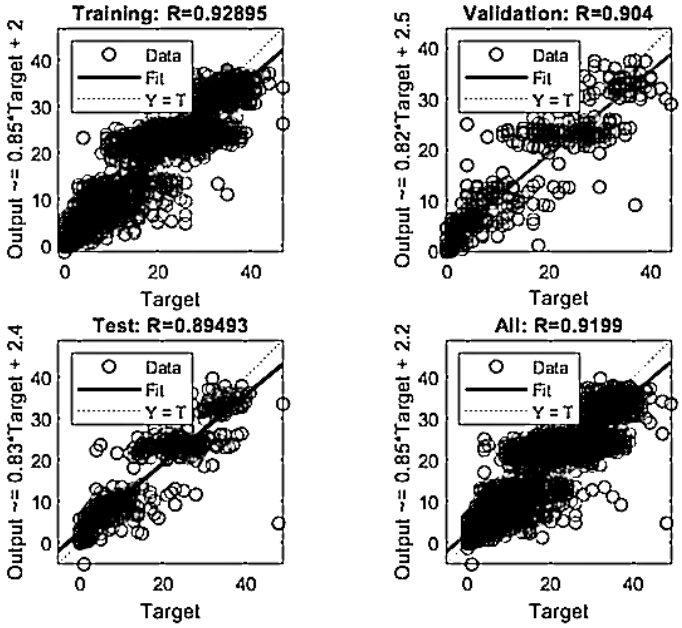


Figure 10. Regression of target with respect to output.

Analyzing Figure 10, the paper found that the R-values of the four sets of regression data are 0.92895, 0.904, 0.89493, and 0.9199, respectively, with values very close to 1, indicating that these results are better.

3. Difficulty prediction results of models

Random Forest (RFF) [10] is an ensemble learning algorithm that combines multiple decision trees for high-accuracy predictions. It uses variable randomization and a minimal set of variables to identify active networks and paths in datasets. For the data that have been preprocessed, we can obtain the test set as X, the category as C, and the number of decision tree models as T, whereby the relevant model can be obtained as

$$H(X) = arg\ \ max\left\{\sum_{i=1}^{T} I(h_i(X) = y)\right\} \tag{14}$$

Where $h_t(X)$ is the output of the firsti the output of the decision tree model, theI is the indicator function, and the function value is 1 when each indicator is meaningful, otherwise it is 0.

K-means (K-means clustering algorithm) is a method of dividing a cluster partition, a data set with N tuples or records is split into K subgroups, each subgroup is a cluster, K<N, each subgroup satisfies.

1. Each grouping contains at least one data record.
2. Each data record belongs to and only belongs to one group.

Algorithm steps.

Step 1: Standardize data and select 4 initialized samples as clustering centers.

Step 2: Assign macroeconomic indicator samples to the closest cluster center.

Step 3: For each class aj, recalculate their cluster centers (the center of mass of all samples belonging to that class).

$$a_j = \frac{1}{c_i}\sum x \tag{15}$$

Repeat steps 2 and 3 until the termination condition is reached.

The data were imported, solved in python, and the ROC curves, clustering effects, as Figure 11 and Figure 12, and predicted difficulty of the word "EERIE" were obtained.

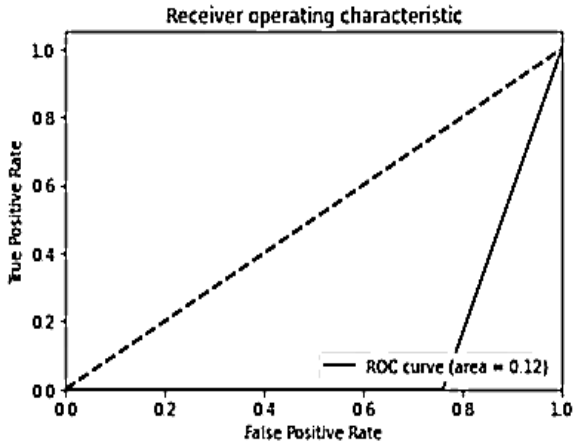


Figure 11. ROC curve.

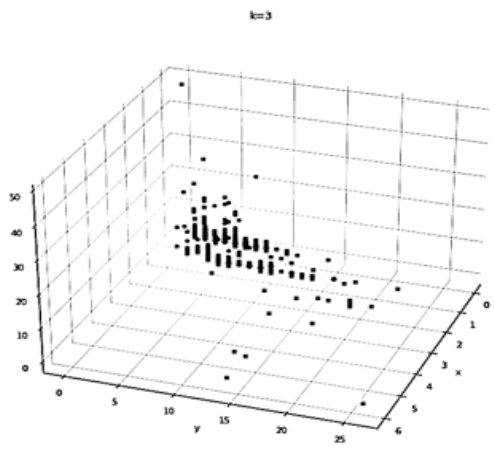


Figure 12. Clustering effect.

The model classifies "EERIE" as medium difficulty, with a higher accuracy (80%) using the random forest algorithm than the k-means clustering algorithm (70%). It effectively predicts solution word difficulty and attributes contributing to difficulty. Future work can incorporate more features and explore other ML algorithms for further improvement. This model has potential for personalized and challenging word puzzle game experiences.

4. Conclusion

The model combines various machine learning techniques to analyze word score reports and predict their outcomes. BP neural networks are the primary method for predicting word game outcomes, while time series and regression analysis are used to understand daily variations in reported results and the effect of word attributes on scores. Random forest and K-means clustering algorithms are utilized to classify solution words by difficulty. The model accurately predicts the number of results and score distribution for a given word. However, limitations include potential inaccuracies due to data quality and the model's sensitivity to changes in data. Future research could focus on developing more interpretable models that provide a clearer understanding of the underlying patterns and validating the model with additional data sets. Additionally, the model could be extended to incorporate additional features and data sources, such as text analysis and demographic information.

References

[1] Yang Q, Fang Y, Zheng Y. Word Data Research and Prediction Based on Wordle Game [J]. Academic Journal of Computing & Information Science, 2023, 6(4).

[2] Snell C, Kostrikov I, Su Y, et al. Offline rl for natural language generation with implicit language q learning [J]. arXiv preprint arXiv:2206.11871, 2022.

[3] Min Z. A study of the number of Wordle users and experience predictions [J]. Academic Journal of Mathematical Sciences, 2023, 4(2).

[4] Guo Chunling. A random forest algorithm-based method for monitoring the operation status of English automatic translation equipment [J]. Automation and Instrumentation, 2023, (01): 178-183.

[5] Liu X, Wang XIN, Wu ZHENG, Dai ZHAOXIN, Sun ZHE. Research on the location of elderly facilities based on random forest algorithm and supply-demand relationship [J]. Survey and Mapping Engineering, 2023, 32(01): 49-55.

[6] Wang Junli. Exploration of a random forest algorithm-based fraud identification method for corporate financial statements [J]. Investment and Entrepreneurship, 2022, 33(24): 58-60.

[7] Xu D.F., Shi B.C., Xu L.Q., Ye Xueqiang, Wang L.N... SOH time series prediction for automotive lithium-ion batteries based on NARX neural network [J]. Automotive Engine, 2022, (06): 71-75.

[8] Yu Jianfeng, He Yunliang, Wu Huahua, Wei Snapdragon, Chen Bo, Zhong Zhenyuan. A random forest algorithm-based distribution network outage research and evaluation scheme design [J]. Microcomputer Applications, 2022, 38(12): 76-79.

[9] Ren Qiang, Shi Ruihao. A neural network-based time series prediction method for slow air leakage of tires [J]. Automotive Electrical, 2022, (12): 26-28+32.

[10] Li Yuefang, Wang Xibo, GAO Yanfei, Ma Feiyan, Lv Hang, Liu Guangqi. Tire tread groove recognition based on BP neural network [J]. Journal of Shandong Institute of Transportation, 2023, 31(01): 1-6.