Research and Comparison of Random Forests and Neural Networks in Shanghai and Shenzhen Financial 20 Index Prediction

Cai-Yu Su
Guangxi Vocational&Technical Institute of Industry,
Nanning, Guangxi, China

Yuan Lei*, Mei-Xia Wang
School of Information and Engineering,
Guangxi University of Foreign Languages,
Nanning,Guangxi,China
*Corresponding author leiyuandata@foxmail.com

2021 World Conference on Computing and Communication Technologies (WCCCT) | 978-1-6654-4110-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/WCCCT52091.2021.00023

***Abstract*—Machine learning has made great achievements in the field of artificial intelligence, especially in the financial industry, which has shown great potential and attracted the attention of the academia and the industry. With the opening of the capital market to the outside world, how to achieve more accurate forecast of the trend of stock index and timely carry out reasonable risk control is one of the important issues concerned by investors. On the basis of literature and theoretical analysis, the public data sets of a well-known securities firm is selected to make predictions of the Shanghai and Shenzhen CSI 20 index based on neural networks and random forest models. The average absolute error (MAE) was then used as a metric to compare the performance of the two machine learning algorithms on the stock index prediction problem. The results show that the model constructed by the random forest algorithm performs better on this problem. Therefore, we believe that making investment decisions based on the results can help investors make effective investment decisions under certain risks.**

Keywords*—random forests, RBF neural network, Ma-chine learning, financial index forecast, stock market

I. Introduction

In the financial stock market, it is difficult to predict the exact value of a single stock because of stock data fluctuates from day to day. In order to easily show the overall stock market trends, many stock markets have defined their own specific financial indices. For example, Shanghai and Shenzhen (CSI) 300 index, Shanghai Stock ExchangeSSE50 index and China Stock Exchange (CSI) 180 Index, etc. These financial indices are general in scope, however, do not reflect trends in specific sectors, such as pharmaceuticals and the Internet. As a result, the major securities firms have launched one or more indexes for the quality stocks in each stock sector. Especially the Internet and pharmaceutical sectors of the stock market have become popular investments due to technological advances and improved living standards. The CSI 20 index in this article was proposed by a well-known securities company. Moreover, it reflects the movement of quality stocks in the Internet and pharmaceutical sectors of the Shanghai Stock Exchange, which has been tested by the market for many years and is loved by industry people. The CSI 20 index is an important index of the securities company.However, due to the particularity of the financial industry, most investors are ignorant of the calculation rules of CSI20. Therefore, investors often need to take great risks when they make investment decisions. How to achieve more accurate stock index forecast and help investors to make effective investment decisions is a high-profile problem.

However, stock indexes are often influenced by many factors, such as complicated factors such as economy and politics. There seems to be no obvious rule that causes stock index fluctuations [1].Moreover, compared with individual stock data, stock index data has a more complex evolution pattern, and also has the characteristics of non-stationary, nonlinear and long-term memory. Traditional models often assume many conditions, and generally can only find the linear relationship between the index and the characteristics. For the nonlinear change law of stock index, the prediction effect of traditional model is not ideal, and its application is very limited [2].

In recent years, with the rapid development of machine learning, data mining technology has been favored by a large number of scholars, gradually showing great advantages in forecasting and classification. A growing number of academics and investors are experimenting with machine learning to predict stock indexes. Such as,Meng et al. (2016) [3]used web crawler techniques to obtain massive textual information from nine well-known financial websites, and then used a random forest algorithm to identify the main variables that affect the returns and yield movements of the SSE Composite Index. Indranil Ghosh et al. (2018) [4] used four advanced machine learning methods, namely adaptive neural fuzzy inference system, dynamic evolutionary neural fuzzy inference system (DENFIS), Jordan neural network, support vector regression and random forest, to establish prediction models to predict future index prices. Chin et al. (2019) [5] compared four prediction models: artificial neural network (ANN), support vector machine (SVM), random forest and naïve-Bayes. Chin computed 10 technical parameters, and expressed them as trend-deterministic data. Jou et al. (2016) [6] used the proposed planar feature representation method and convolutional neural network (CNN) algorithm to process the historical dataset of Taiwan stock index futures. Dadabada et al. (2017) [7] proposed a quantile regression neural network based

978-1-6654-4110-0/21/$31.00 ©2021 IEEE
DOI 10.1109/WCCCT52091.2021.00023

on particle swarm optimization (PSO) training–PSOQRNN, which used to predict the volatility of financial time series. Bhumika et al. (2018) [8] described artificial neural networks, recurrent neural networks i.e. long-term and short-term memories (LSTM), convolutional neural networks and expanding convolutional networks and compares them in the prediction of closing prices of market indices.Although machine learning model can be used to index relative is various, but the neural network as a supervised learning method, is used to index prediction problem has the advantages of other models do not have stock index prediction must use index history data, which essentially determines the supervised learning in the stock index prediction problems have congenital advantage.Chong et al. (2017) feature data is extracted by unsupervised learning algorithm. The results show that the deep neural network (DNN) can extract additional information from the residual of autoregressive models, but not vice versa. This indicates that the neural network has some advantages over the autoregressive model in predicting problems [9].Araújo et al. (2019) A neural network model is trained with gradient descent method to realize the prediction. Compared with ARIMA, MLP, PELMNN, SVRP and other models, the results show that this model has a better performance in forecasting problems [10]. Although the neural network performs well in this problem, there is no clear way to select the parameters of the neural network, so the use of neural network for prediction is often subject to the selection of parameters [11].As a new and highly flexible machine learning algorithm, random forest is not inferior to neural network in solving prediction problems. Moreover, the stochastic forest does not have to worry about the overfitting problem and has excellent application in the prediction problem [12].Therefore, in this paper uses the random forest model and neural network model in deep learning to fit and forest the CSI 20 index.MAE is calculated from the fitting results of the two models to evaluate their performance in solving the stock index forecasting problem.

We can see that most of the existing research is based on a single deep learning model, and only a few have compared different models. Therefore, the significance of this paper is as follows.

1. We use random forests to predict the CSI 20 index.

2. We use neural network to predict the CSI 20 index.

3. We compare the predicted results of random forest and neural network models, and makes corresponding analysis to draw conclusions.

II. Forecasting Methods

A. Data Standardization Methods

1) Min-max: The Min-Max standardization method is used to normalize the variables and eliminate the dimension among data factors.

The Min-Max formula is as follows.

$y_{ij} = (x_{ij} - x_{\min}) / (x_{\max} - x_{\min}\}$

In the formula, y_{ij} is the normalized value of the jth sample of the i-th variable, x_{ij} is its original value; $x_{\max}$ i represent the maximum values of all samples of the i-th variable and $x_{\min}$ is minimum values.

2) Z-score: The Z-score standardization method is used to normalize the variables and eliminate the dimension among data factors.

The z-score formula is as follows.

$x' = \frac{x-\mu}{\sigma}$

In the formula, μ is the average value and σ is the standard deviation.

B. Random Forests

Show in figure 1,if each ensemble classifier $h_k(\mathbf{x})$ is a decision tree, then the collection is a random forest. We define $h_k(\mathbf{x})$ the parameters of the decision tree for the classifier $\Theta_k = (\theta_{k1}, \theta_{k2}, \ldots, \theta_{kp})$(these parameters include the structure of the tree, the nodes to which variables are split, etc.). We can also write it as $h_k(\mathbf{x}) = h(\mathbf{x} \mid \Theta_k)$.

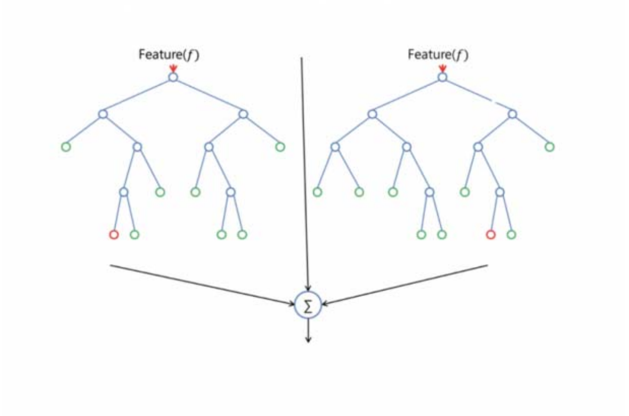


Fig. 1: Random forests model

The question is, how do we choose which features appear in which nodes of the tree k^{th}.In a random forest, factors Θ_k are chosen randomly based on the variables Θ being modeled. A random forest is a family of classifiers $h(\mathbf{x} \mid \Theta_1), \ldots, h(\mathbf{x} \mid \Theta_K)$ based on a classification tree Θ_k with one parameter, chosen randomly from the model random vectors Θ. For the final classification $f(\mathbf{x})$(combined with the classifier $\{h_k(\mathbf{x})\}$), the random forest takes the plural or average of these results as the output of this new data. The formula is as follows $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$.

The process of building a random forest is shown in the figure. Compared with many algorithms, the learning process for random forests is very fast.

C. RBF Neural Networks

An artificial neural network (ANN), or neural network, models the relationship between a set of input and output signals, which is a mathematical or computational model that mimics the structure and function of biological neural network. A neural network consists of a large number of artificial neurons linked together to perform computations. Modern neural network is a kind of nonlinear statistical data modeling

TABLE I: Symbols

Symbol	Annotation
L	Layers of neural network
M_l	The number of neurons in the l layer
$f_l(\cdot)$	Activation function of neurons in the l layer
$\boldsymbol{W}^{(l)} \in \mathbb{R}^{M_l \times M_{l-1}}$	Weight matrix from layer l-1 to layer l
$\boldsymbol{b}^{(l)} \in \mathbb{R}^{M_l}$	Offset from layer l-1 to layer l
$z^{(l)} \in \mathbb{R}^{M_l}$	Net input (net activity value) of neurons in layer l
$\boldsymbol{a}^{(l)} \in \mathbb{R}^{M_l}$	Output (activity value) of neurons in the l layer

tool that is commonly used to model complex relationships between input and output, or to explore data. This paper uses the feed-forward neural network model, in which each neuron belongs to a different layer with no connections within the layer, and all the neurons between the two adjacent layers are connected in pairs. There is no feedback in the whole network, and the signal propagates from the input layer to the output layer, which can be represented by a directed acyclic graph in figure 2.

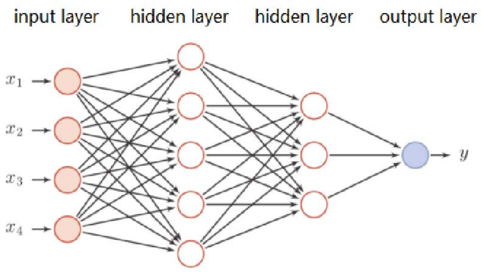


Fig. 2: RBF Neural Networks

The following notation is used to describe the RBF neural network:

The formula of the RBF neural network is as follows. $\boldsymbol{z}^{(l)} = \boldsymbol{W}^{(l)}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}$

$\boldsymbol{a}^{(l)} = f_l\left(\boldsymbol{z}^{(l)}\right)$

The formula of Radial Basis Function is as follows.

$x = a^{(0)} \rightarrow z^{(1)} \rightarrow a^{(1)} \rightarrow z^{(2)} \rightarrow \cdots$

$\rightarrow a^{(L-1)} \rightarrow z^{(L)} \rightarrow a^{(L)} = \phi(x; W, b)$

A neural network can serve as an all-purpose function, which can be used to transform complex features or approximate a complex conditional distribution.

D. Model Scoring Criteria

The average absolute error is used for the score of model predictions, and the formula is as follows.

$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_{\text{actual}} - Y_{\text{predict}}|$

In the formula, N is the number of predictions, and Yactual is the true value, and Ypredict is the predicted value. Moreover, the smaller the average absolute error of the prediction, the more accurate the model is.

III. Results and analysis

A. Data Pre-processing

The data comes from the 2nd Guangxi University Artificial Intelligence Competition, which is a financial data sets of a well-known securities company from April 1, 2011 to December 28, 2019. Among these data, data from 2011 to 2018 is the training set, i.e., the CSI 20 index with pharmaceutical and the Internet; data form 2019 is the test set, which is needed to forecast the corresponding CSI 20 index. We also need to use the data of stocks in the Internet and pharmaceutical sectors and the corresponding CSI 20 index. The large number of stocks in the stock data can lead to overfitting when the model is finally fitted, so we need to select the best 20 stocks for each day according to the CSI 300 index selection method.

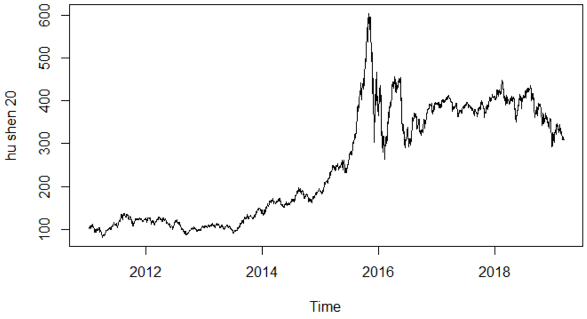


Fig. 3: CSI 20 trend chart

The trend of the CSI 20 Index from April 1, 2011 to December 28, 2019 is shown in the figure 3 above.

We used the data from April 1, 2011 to December 28, 2018 as the training set and the data from 2019 as the test set.

B. Forecast Results of Random Forest

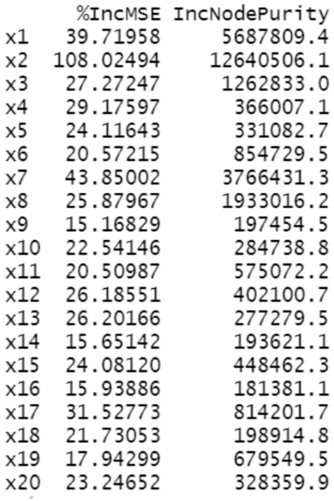
As shown in figure 4, the random forest was used to rank the importance of the characteristics. It can be seen that 20 stocks are very important in random forest and do not require further data filtering.

We use the Z-score standardization to process the data set, and the results of random forest modeling are shown in figure 5. The black line is the actual value of the CSI 20 index, and the red line is the value of the CSI 20 index predicted by random forest.

From figure 5, we can see that the predicted value of the random forest model generally coincides with the curve drawn by the real index value, indicating that the random forest model can effectively predict the CSI 20 index.

C. Forecasting Results of Neural Network

The neural network uses Min-Max standardization to process the data set. In the processed data, the CSI 20 index is used as the dependent variable and the remaining characteristic variables are used as independent variables to view the results of multiple linear regression analysis in this case. The results



```
    %IncMSE IncNodePurity
x1   39.71958     5687809.4
x2  108.02494    12640506.1
x3   27.27247     1262833.0
x4   29.17597      366007.1
x5   24.11643      331082.7
x6   20.57215      854729.5
x7   43.85002     3766431.3
x8   25.87967     1933016.2
x9   15.16829      197454.5
x10  22.54146      284738.8
x11  20.50987      575072.2
x12  26.18551      402100.7
x13  26.20166      277279.5
x14  15.65142      193621.1
x15  24.08120      448462.3
x16  15.93886      181381.1
x17  31.52773      814201.7
x18  21.73053      198914.8
x19  17.94299      679549.5
x20  23.24652      328359.9
```

Fig. 4: Random forest feature ranking

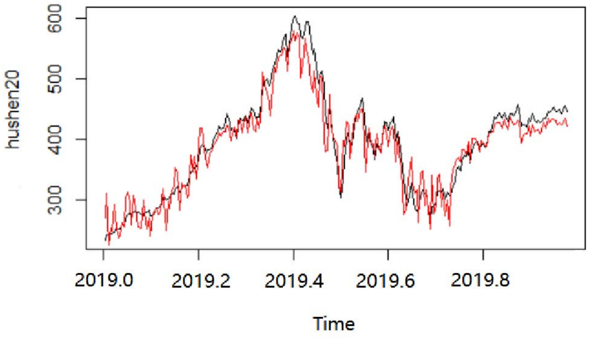


Fig. 5: Random forest results

are shown in figure 6 and it can be seen that the model passes the F-test to determine the coefficient R-squared to 0.8153.

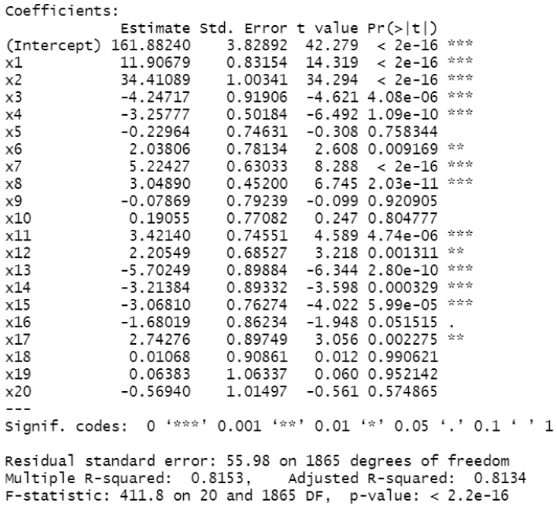
There are two hidden layers in the network, and the number of neurons in each layer is 15 and 5. The network structure is shown in figure 8.

The results of neural network model are shown in figure 8. In the figure, the black line is the actual value of the CSI 20 index and the red line is the value the value predicted by neural network. In the figure, we can see that only a small part of the curve drawn by the neural network model overlaps with the real value of the index, while most of the other curves have a large gap with the real value. Thus, this shows that the neural network model can predict the CSI 20 index, but with a large error.

D. Comparison of Random Forest and Neural Network

$MAE = \frac{1}{N}\sum_{i=1}^{N} |Y_{\text{actual}} - Y_{\text{predict}}|$

The scoring formula is shown above, and we need to calculate the error between the predicted value and the true



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 161.88240    3.82892  42.279  < 2e-16 ***
x1           11.90679    0.83154  14.319  < 2e-16 ***
x2           34.41089    1.00341  34.294  < 2e-16 ***
x3           -4.24717    0.91906  -4.621 4.08e-06 ***
x4           -3.25777    0.50184  -6.492 1.09e-10 ***
x5           -0.22964    0.74631  -0.308 0.758344    
x6            2.03806    0.78134   2.608 0.009169 ** 
x7            5.22427    0.63033   8.288  < 2e-16 ***
x8            3.04890    0.45200   6.745 2.03e-11 ***
x9           -0.07869    0.79239  -0.099 0.920905    
x10           0.19055    0.77082   0.247 0.804777    
x11           3.42140    0.74551   4.589 4.74e-06 ***
x12           2.20549    0.68527   3.218 0.001311 ** 
x13          -5.70249    0.89884  -6.344 2.80e-10 ***
x14          -3.21384    0.89332  -3.598 0.000329 ***
x15          -3.06810    0.76274  -4.022 5.99e-05 ***
x16          -1.68019    0.86234  -1.948 0.051515 .  
x17           2.74276    0.89749   3.056 0.002275 ** 
x18           0.01068    0.90861   0.012 0.990621    
x19           0.06383    1.06337   0.060 0.952142    
x20          -0.56940    1.01497  -0.561 0.574865    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.98 on 1865 degrees of freedom
Multiple R-squared:  0.8153,	Adjusted R-squared:  0.8134 
F-statistic: 411.8 on 20 and 1865 DF,  p-value: < 2.2e-16
```

Fig. 6: Neural regression

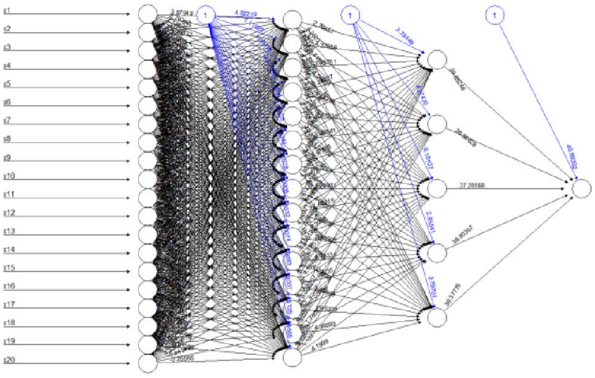


Fig. 7: Practical model of neural network

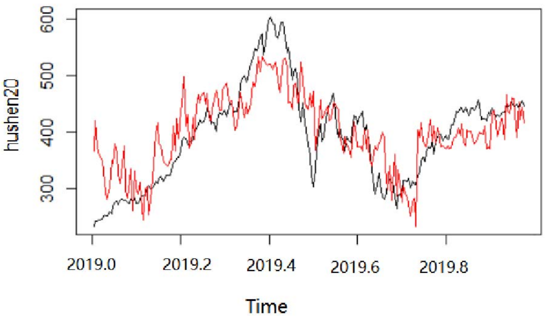


Fig. 8: Neural network results

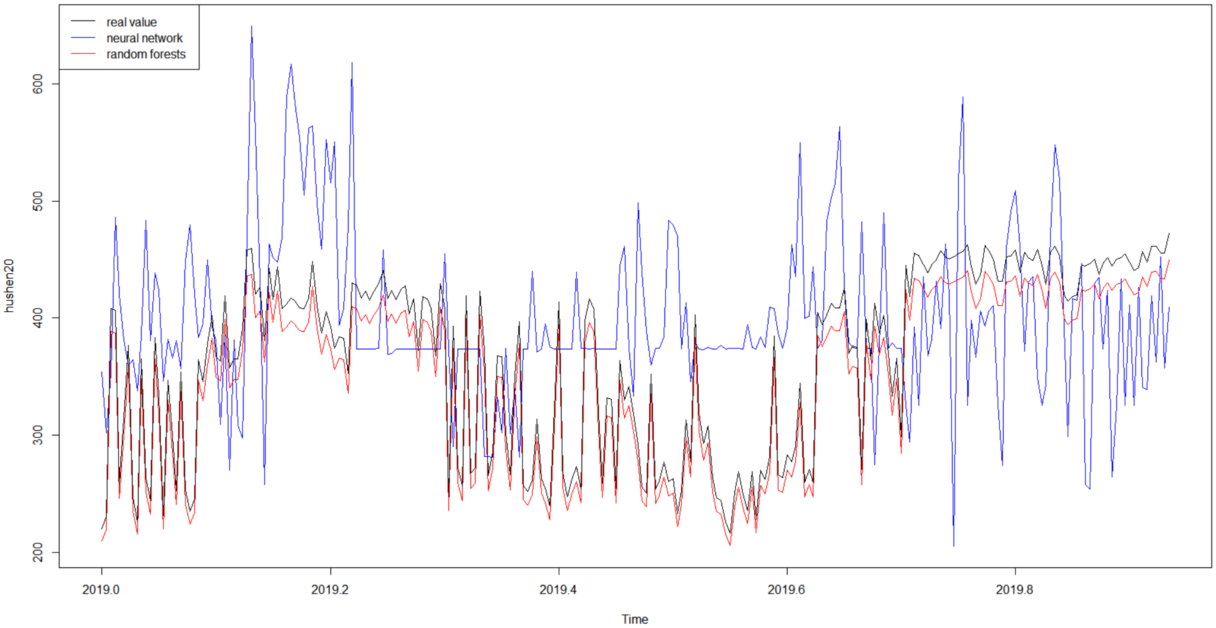


Fig. 9: Comparison of random forests and neural networks

TABLE II: MAE results

	random_forest	neural_network
MAE	31.23	50.35

value, which is absolute. Therefore, as shown in the following figure 9, the absolute values of the errors between the random forest and the neural network and the true values are shown. However, the average absolute error of the random forest is smaller than that of the neural network. Random forest reduces the number of operations and improves the prediction accuracy compared to neural networks, and the algorithm is insensitive to multicollinearity and stable for missing data and unbalanced data, which can be well adapted to datasets of up to several thousand explanatory variables.

As shown in the following table 2, calculating the MAE by calculating the absolute value of the error and summing it up also yields the same result.

IV. Conclusions

This paper is based on the prediction of CSI 20 index by neural network and random forest model, and uses its average absolute error as an indicator to analyze the difference between two machine learning algorithms on financial problems. We draw following conclusions: (1) Random forest has good performance in the prediction of CSI 20 index, and its results have high prediction accuracy. Compared to neural network models, random forest is more suitable for handling data sets of several thousand variables. (2) Financial markets can have huge ups and downs, and the simple random forest and neural network models are trained from past data and they have large errors for sudden ups and downs. In other words, none of them can handle black swan events.

This paper uses random forest and neural network to divide the CSI 20 index predicted data into two stages from variable input and numerical fit to provide technical reference for further improving the prediction performance. At the same time, the random forest model provides a new idea for rough and range forecasting of financial markets. More complex models can be used to complete more accurate forecasts in the future.

References

[1] M. Wu and Y. Wang, “Risk analysis of world major stock index before and after the 2008 financial crisis – based on garch-var approach,” *International Journal of Financial Research*, vol. 9, no. 2, 2018.

[2] C. W. J. Granger, “Long memory relationships and the aggregation of dynamic models,” *Journal of Econometrics*, vol. 14, no. 2, pp. 227–238, 2006.

[3] M. Xuejing, Y. Yafei, and Z. Xinquan, “The research on financial news and stock market investment strategy——based on text mining for financial website,” *Review of Investment Studies*, 2016.

[4] I. Ghosh, M. K. Sanyal, and R. K. Jana, “Fractal inspection and machine learning-based predictive modelling framework for financial markets,” *Arabian Journal for ence and Engineering*, vol. 43, no. 8, pp. 4273–4287, 2018.

[5] C. S. Huang and Y. S. Liu, “Machine learning on stock price movement forecast: The sample of the taiwan stock exchange,” *International Journal of Economics and Financial Issues*, vol. 9, 2019.

[6] J. F. Chen, W. L. Chen, C. P. Huang, S. H. Huang, and A. P. Chen, “Financial time-series data analysis using deep convolutional neural networks,” in *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, 2016.

[7] D. Pradeepkumar and V. Ravi, “Forecasting financial time series volatility using particle swarm optimization trained quantile regression neural network,” *Applied Soft Computing*, 2017.

[8] B. Gupta, M. Singhania, and A. Aggarwal, “Financial time-series forecasting : from neural networks to dilated convolutions,” *Post-Print*, 2018.

[9] E. Chong, C. Han, and F. C. Park, “Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies,” *Expert Systems with Applications*, vol. 83, pp. 187–205, 2017.

[10] R. d. A. Araújo, N. Nedjah, A. L. Oliveira, and R. d. L. Silvio, “A deep increasing–decreasing-linear neural network for financial time series prediction,” *Neurocomputing*, vol. 347, pp. 59–81, 2019.

[11] M. M. Mijwel, “Artificial neural networks advantages and disadvantages,” *Retrieved from LinkedIn: https://www. linkedin. com/pulse/artificial-neuralnet works-advantages-disadvantages-maad-m-mijwel*, 2018.

[12] I. K. Nti, A. F. Adekoya, and B. A. Weyori, “Random forest based feature selection of macroeconomic variables for stock market prediction,” *American Journal of Applied Ences*, vol. 16, no. 7, pp. 200–212, 2019.