Int. J. Nonlinear Anal. Appl. In Press, 1–10 ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/IJNAA.2023.30809.4502

Stock price prediction using data mining algorithms in the Iranian stock market

Mohammad Rehani Nezhad'Alla[a], Saeid Ghane[a,1], Allah Karam Salehi[b]

[a]Department of Management, Masjed-Soleiman Branch, Islamic Azad University, Masjed- Soleiman, Iran

[b]Department of Accounting, Masjed-Soleiman Branch, Islamic Azad University, Masjed-Soleiman, Iran

(Communicated by Madjid Eshaghi Gordji)

Abstract

Uncertainty in the capital market means the difference between the expected values and the values that occur in reality. The design of different analysis and forecasting methods in the capital market is also due to the high value and the need to know prices in the future with more certainty or less uncertainty. To earn profit in the capital market, investors have always sought to find the right share for investment and the right price for buying and selling, and therefore all the forecasting models proposed have always sought to answer three basic questions; What share, in what time frame and at what price should be bought or sold. In this article, we will use the combined method based on LSTM and a neural fuzzy system to predict stock prices in the Iranian market. The results show that the proposed method has an accuracy of over 90% in stock price prediction.

Keywords: stock market, neural fuzzy system, lstm neural network
2020 MSC: 62M45, 68T07

1 Introduction

The tools used in predicting all aspects of the capital market can be divided into three general categories: technical methods, fundamental methods and mathematical methods, including classic time series and regression methods and artificial intelligence methods. Before [1], everyone was only looking to answer the three mentioned questions in the capital market. With the introduction of FAMA's capital market efficiency theory, another big question was faced by investors and analysts; Is it even possible to analyze and predict the future of the stock market? By proposing the capital market efficiency theory, Fama assumed that the market consists of investors who all have access to the same information from the past and present. Political, economic, secret information etc. To believe the Fama theory, it is not even necessary that all investors have access to the same information, it is only necessary that a large number of investors have access to this information in the same and equal way. When the available information is equal, the analysis and forecasting tools are also the same; No investment can earn extraordinary profits. For example, if the first person who has confidential information intends to buy a share; In the time it takes to do this, the sellers get to know this information and sell their shares at a higher price than before. Therefore, there is no opportunity to earn unusual profits. The precise and complete definition of an efficient capital market is a market in which the speed of information transfer is very high. Therefore, FAMA defines three types of efficiency, a poorly efficient market in which the speed of information transmission is low,

[1] Corresponding author
Email addresses: mohamadreyhaninezhad@gmail.com (Mohammad Rehani Nezhad'Alla), dr.saeedghane@gmail.com (Saeed Ghane), salehi_allahkaram@gmail.com (Allah Karam Salehi)

Received: April 2023 *Accepted:* June 2023

but it is still not possible to make unusual profits using past information. An average efficient market where the speed of information transmission is high and it is not possible to make unusual profits using current information. A strong efficient market where the speed of information transfer is very high and even the owners of confidential information do not have the opportunity to earn extraordinary profits. Finally, Fama concludes that price fluctuations in the market are random and cannot be predicted.

After the theory of capital market efficiency was proposed, capital market investors were divided into two groups. The group that believed in the efficiency of the market turned to long-term investments, the reason for this was that this group ruled out the possibility of earning profit through market forecasting, and therefore there was no opportunity to earn unusual profit through short-term buying and selling. Other investors do not believe in this theory and criticize it. This group believed in the prediction of the capital market and therefore used different models for these predictions. In addition to long-term investments, this category also deals with short-term and medium-term investments, and through continuous buying and selling, they try to earn extraordinary profits.

Among the criticisms of the capital market efficiency theory, we can mention the famous rules of the capital market. Among these rules, we can mention the lowering of prices in general at the end of weeks; The reason for this is that during the holidays, unpleasant things may happen that cause a sharp drop in prices at the beginning of the following week. Another rule is the lowering of prices at the end of the year. In general, stocks that have a higher P/E yield are higher. These rules, which have become market customs, are among the criticisms that investors make of the theory of the efficiency of the capital market, which deems it impossible to predict.

Another criticism of investors is that seasonal, cyclical, linear and non-linear trends are clearly identifiable and evident through econometric and time series models. From an empirical point of view, many investors earn profits by using technical and fundamental analysis. In [2], Khetan examines the current trends in the capital market and states that from a practical point of view, this theory is not correct and the theory of capital market efficiency is correct in a perfectly ideal market. Therefore, according to the mentioned criticisms and the practical opportunities perceived in the market by investors, in this article, the assumption of the predictability of the capital market is accepted and we answer the first question in this way.

Now that we believe in the predictability of the market, it is time to answer the three questions raised, what is the share, what is the time frame and what is the price? Many different models have been used to answer this question. The models that predict the target share are generally examined using fundamental analysis and classical forecasting methods such as econometrics and time series. The models that deal with the time range seek to find a specific strategy for trading and use technical analysis and forecasting models or advanced artificial intelligence models. In response to the question of what price, a very large category of models is used. Classical models, fundamental analysis, technical analysis and artificial intelligence algorithms. In this article, we are specifically looking for the answer to the third question, i.e. price forecasting, the most diversity in the field of forecasting is also in this field. In the following, we will Artificial intelligence algorithms, the use of which is growing rapidly among investors, are actually a combination of all the aforementioned forecasting methods with the ability to fit high-order nonlinear curves. These algorithms can work with a large number of variables and find the appropriate relationship between these variables. As mentioned, there are many influencing factors in stock prices, and classic models and old and empirical analyses are not responsive to these factors. For short-term forecasts, classical time series models and technical analysis are generally used, each of which includes a large number of other algorithms that make it difficult to summarize and draw conclusions. Artificial intelligence algorithms have the power to combine all these analyses with each other through optimal weightings and provide a unique and optimal solution.

Among artificial intelligence algorithms, the use of neural networks in the topic of prediction is very high. This is due to the capability of the neural network to work with a large number of variables, a very accurate fitting of the time series, not being affected by outlier data, no limitation for a certain degree of non-linearity, and the flexibility of the network against changes in the model parameters.

According to the mentioned reasons and the results obtained in other research, some of which were mentioned, in this article, we also choose neural networks to predict stock prices. The important point in using neural networks is that these networks initially enter the learning process using 75% of the data (in general). At this stage, due to the high ability of the neural network to fit complex curves, the network may identify the model as more complex than it is. This makes the predicted values closer to the real values and makes a more accurate estimation in the training phase. The point here is that some of the changes and differences between the predicted and actual values are due to the randomness of these changes. There is a random component in all prediction models, but neural networks use more variables to estimate these values, while these values are random and unpredictable. This property, which is called overfitting, is the biggest problem of neural networks.

To eliminate this overfitting, the inputs to the neural network must be reduced. In other words, neural networks should not be allowed to overfit. In this article, data mining is used to reduce the number of variables. Data mining is the science of exploring data to discover knowledge. The science of data mining, by providing different solutions, considers the first pillar of using data to discover knowledge to be the removal of unnecessary data and unnecessary leaves, therefore, in this article, several data mining techniques have been used that will be described later. Also, as mentioned, the models used for short-term forecasts are time series models and technical analysis. Since we intend to forecast the closing price of stocks daily, by combining these two models, we forecast using technical analysis indicators and previous days' prices. briefly explain the types of prediction models and the reason for choosing neural networks as the selected model.

Stock price prediction models can generally be placed in three categories, technical analysis, fundamental analysis and analysis using artificial intelligence algorithms. Technical analysis makes predictions by studying past changes in stock prices and with the assumption that past events will repeat themselves in the future. Chartists or technical experts believe that the only thing that changes the stock price is the amount of demand and supply in the market. In fact, this group believes that stock prices change under the influence of any other fundamental economic factor, these factors only affect supply and demand in the market, and therefore by predicting these values, prices can also be predicted. The two general models used by this category are the matching of patterns and the use of indicators; Patterns are actually past price change trends, which can be used to predict the future based on the belief that they will repeat themselves in the future. Indicators are also mathematical models that predict supply and demand and finally the price by using price indicators such as the opening and closing price and the volume of transactions.

The competitive environment is heavily influenced by the complexity and uncertainty caused by changes in the stages of the business cycle of companies and changes in production processes. These changes make the companies active in the competitive environment to change the traditional production systems and evaluate the new production processes to deliver their products on time and with high quality. To create competitive power in companies, it is necessary to turn their internal capabilities and resources into factors for the success of the organization and create a competitive advantage over other competitors. In other words, the development and cultivation of production capabilities is one of the most important tasks of the production strategy. In fact, the evaluation of competitive production processes by drawing the goals and priorities of companies helps them to increase their competitiveness by choosing appropriate processes and taking possession of various resources including human power, technology, and information technology. Competitive priorities and strategic production decisions are considered to be the most important components of production strategy, and after determining competitive functions from various tools such as total quality management (TQM); Data Envelopment Analysis (DEA) and other production engineering methods try to provide companies with the most effectiveness. Competitive advantage arises when a company has a more coherent understanding of its business cycle based on production capabilities. In other words, the complexity, connection, and growing speed of production developments have created great challenges for competitive organizations, and paying attention to the maturity of the intellectual capital of companies and creating a balance between it and production processes enables companies to be more effective for competition. Intellectual capital can provide more added value for the company by improving production technology, learning, more adaptation of the product to the customer's needs and things like that. This organizational knowledge is broad, which is specific and unique for each company and allows it to continuously adapt itself to the changing conditions and significantly improve the competitive situation of the organization by increasing the level of added value for the key stakeholders of the organization. On the other hand, paying attention to production can speed up industrial growth and development and guide it in a correct and principled direction. Therefore, the existing knowledge and mechanisms in an organization can play an effective role in increasing the competitive advantage of an organization through the improvement of methods and production processes. Since understanding and exploiting vital intangible resources in organizations helps to maintain and gain competitive advantage, successful companies in this field have a higher knowledge-oriented level due to the existence of intellectual capital maturity. The noteworthy point is that the existence of intangible assets is considered as a knowledge-enhancing level, which will be able to strengthen the production functions with greater efficiency to gain more efficiency based on the creation of effective competitive advantages. Intangible assets create the maturity of intellectual capital when the space and form of people's participation are integrated into the service of macro strategies and sub-strategies such as production strategies.

In this research, by examining all the above and the feasibility of combining the methods used for price forecasting, the questions raised will be answered, and for the first time, we will deal with price forecasting by combining lstm method and neural fuzzy network.

2 Theoretical foundations of research

2.1 Application of data mining in the stock market

Data mining plays a pivotal role in addressing diverse challenges across various domains. In the realm of cybersecurity, it contributes to spam detection from extensive datasets through the utilization of evolutionary data mining systems, enhancing the ability to identify and mitigate unwanted messages effectively [3]. Additionally, data mining serves as a fundamental tool for refining recommender systems and optimizing their performance by analyzing user data and patterns. This application substantially improves the accuracy and efficiency of recommendation systems. Furthermore, in the healthcare sector, the diagnosis of atherosclerosis in the coronary arteries is advanced through the application of data mining and machine learning techniques [4]. By extracting valuable insights from medical data, these techniques enhance the precision and predictive capabilities in identifying this cardiovascular condition. These examples underscore the versatile and impactful applications of data mining in addressing multifaceted issues and optimizing decision-making processes across different domains.

Explaining the benefits of investing in international markets and forming a global portfolio of stocks Begum investigated the movement of stock prices in the China-Taiwan ECFA joint stock market using data mining and using syntactic rule techniques and clustering techniques [5]. and classified the stocks in the market and finally concluded that electronic, financial and insurance stocks in the Taiwan market and housing, communication and financial stocks in the Hong Kong market have a strong and promising movement in this market. For this study, they used twenty stock indices for clustering and finding correlation rules. In [5], Begum emphasizes the benefits of investing in international portfolios and states that with the current tendency to work in global markets, the importance of understanding and forecasting currency exchange rates increases. By examining the relationship between the currency exchange rate and Taiwan's international market indices using rules, they have produced rules for generating optimal investment portfolios in the Taiwan market by considering the currency exchange rate, and by using data mining, they have explained the rules to be used in the Markowitz portfolio formation method.

In [6], Hamo divided stocks into two categories, from the point of view of investors, valuable stocks and growing stocks, and by defining the two criteria of book value to market value and stock return, respectively, for valuable stocks and growing stocks, stocks They put the available in the Taiwan market into four categories, which are: valuable, low-value, fast-growing and low-growing. Then, using the time series data mining technique, they investigated how investors behaved with these stocks and found patterns. Among these patterns are: high-value and low-value stocks with overreaction of investors, high-growth and low-growth stocks with normal reaction, and high-value stocks have a sharp drop seasonally.

Using a corporate classifier, Hamo categorized and predicted prices through a large amount of input data such as technical analysis data and market indicators. The problem that the company's cluster algorithm has is that it cannot work with a lot of data, so they designed a genetic-based cluster to predict buy and sell signals. The obtained results indicate the high power of the proposed algorithm and its ability to compete with many other classification methods such as statistical classification.

2.2 Application of neural network and data mining in stock price prediction

Rastogi showed that the accuracy of neural networks varies mostly in the range of 70-80% [7]. This amount is better than statistical methods by 5-20% on average. In this article, the author concludes that if expert systems are used in neural networks, the average accuracy of neural networks will be higher than the mentioned amount.

In [8], Kansal explains that artificial neural networks can identify stocks that potentially have significant potential for growth in terms of fundamental indicators; be used They also showed that neural network architecture outperforms similar classical and artificial intelligence algorithms that do not use neurons. In this research, rules for conducting transactions with a threshold limit of 50 and above were found by neural networks, which were able to achieve 100% of the goals. From other results of the research, it can be mentioned that the researchers came to the conclusion that by continuing to use neural networks over time, the predictive power of this network increases.

In [9], Rajguru by studying and reviewing nineteen articles concluded that neural networks have a high power in stock forecasting and have performed better than all their alternatives. In [10], Nuaimi has investigated the application of data mining in many areas of decision-making and has acknowledged the high power of data mining in improving results. The examined areas are: medical sciences regarding disease diagnosis, sports areas regarding the selection of players, game strategy, predicting financial bankruptcy of banks, improving product demand forecasting, optimizing classification, choosing the right product, comparing the classification of retailers and manufacturers, predicting the production plans of factories, Deciding on the yes/no rules of continuous works, creating consistent maps for shareholders and investors

3 Suggested method in price forecasting

3.1 Normalization

After extraction, for all data sets, if there are NaN values, they are replaced with the average of the corresponding feature. Since a neural network should receive data that are not relatively large and heterogeneous, we used the Min-Max normalization method to bring all features to the range of 0 to 1.

$$X = \frac{x - min(x)}{max(x) - min(x)}$$

3.2 Time series data

A time series is a sequence of time-dependent values. LSTM to predict this sequence, unlike the autoencoder version, works with supervision. In this way, the data should be divided into input and output. In addition, LSTM is excellent compared to classical statistical linear models because it can make multi-input prediction problems easier. In our model, LSTM uses past data to predict the next 30 days. We have determined this number of past days in the form of window size with the value of 32. Therefore, the input data set of our tensor is a matrix with dimensions of 35x5. It means that we have 5 features with 32 lines. The output data takes into account not only the window size but also the forecast period, which in our case is 30 days. The output data set starts from row 35 and is made of pieces of length 30. The prediction interval determines the output size of the LSTM network [11, 12].

3.3 Division into training and test data

This step is one of the most important steps, especially regarding the price of the refinery stock index. We divided our dataset into two parts, training and testing. Both training and testing sets have input and output characteristics.

3.4 LSTM implementation

The main feature of feedforward networks is that they do not retain anything, and this is due to their one-way nature. In fact, every moment only the current input is sent to the hidden layer and the previous input is not stored in the memory. Now, considering that we are dealing with time series and need to maintain the previous prices, we need a neural network with memory.

Recurrent neural networks that have a loop to preserve past information are a suitable option. However, because traditional neural networks have the problem of maintaining long-term dependencies and vanishing gradients we have used recurrent networks (LSTM type).

LSTM transmits time information with state cells and gates. The state cell is the same horizontal line that is embedded from Ct-1 to Ct to maintain dependencies, and the output of LSTM is adjusted by the state of the cells, as a result, it is suitable for prediction based on the input history and not only the last input. On the other hand, when more time passes forward, we need to forget very old entries. LSTM by the forgetting gate can decide what information to keep and what information to forget, which will be presented in the definition of gates.

Forget gate: $f_t = \sigma(WfS_{t-1} + W_fS_t)$

Input gate: $i_t = \sigma(W_iS_{t-1} + W_iS_t)$

Output gate: $o_t = \sigma(W_oS_{t-1} + W_oS_t)$ (1)

$C = tanh(W_cS_{t-1} + W_cX_t)$ (2) Intermediate state cell $c_t = (i_t * C_t) + (f_t * c_{t-1})$ (3) state cell (memory for next input) $h_t = o_t * tanh(c_t)$

(4) Calculate the new state

Each gate has a different set of weights. In the last equation, the input gate and the middle cell state are added by the old cell state and the forgotten gate. Then the output of this operation is used to calculate the new state. The advantage of LSTM over traditional neural networks is that instead of using only one tanh layer, it uses 4 layers of neural network that interact with each other.

3.4.1 Setting hyperparameters

Optimizer: Many neural networks use random gradient descent to avoid getting stuck in a local minimum. Other optimizers such as Adagrad, RMSProp also work well, but in our research, Adam had better results (all these optimizers are included in the Cross package).

Cost function: We evaluated the performance of the model with the MeanAbsoluteError function.

$$MAE = \sum_{i=1}^{n} \frac{|Y(pred)i - Y(i)|}{n}$$

Activator function: The most popular activator functions are the sigmoid function, Tanh and ReLu. We used ReLu to predict the price of Bitcoin.

Dropout rate: This rate determines how many percent of network nodes will randomly have zero input to increase the learning speed of the network.

The number of hidden layer neurons: we considered the number of 100 neurons. Although it was tried with other values this value was optimal for our needs.

Number of rounds: We train the network in 100 rounds. According to the available hardware and overfit with higher rounds, a good result was obtained with the same number.

Batch Size: We tried different values and chose 32 as the window size.

Network architecture: We have used Cross Sequential API and it was designed as a sub-network:

- One LSTM layer
- A Droopout layer
- A dense layer
- An Activation layer

4 Simulation

The price prediction of the refinery stock index by simulation in Matela software with the help of an improved neural fuzzy system with DE, GA, PSO and LSTM algorithms only and finally the weighted average combination of ANFIS PSO with LSTM has been done as follows.

In Figures 1 to 4, the upper figure shows the predicted value and the actual value, the lower left figure shows the prediction error value in terms of the number of samples, and the lower right figure shows the histogram of the prediction error, which is comparing Figures 3 to 5, it can be concluded that the ANFIS PSO algorithm has better performance than other algorithms.

5 Weighted average method

In this method, the values of ANFIS PSO methods are obtained with the help of the RMSE value of the training stage, and the higher the accuracy, the more weight it has in the average value. According to the table below, in the optimal ANFIS PSO method, the average error value is low, but the error variance is high, and also in the LSTM method, the average error value is high and the error variance is low, so the combination of ANFIS PSO and LSTM is used to reduce the average error and error variance to an acceptable level. We use

$$OWM = \frac{w_A * \mu_A + w_L * \mu_L}{\mu_A + \mu_L} \quad (5)$$

w_A: ANFIS PSO accuracy in the train phase μ_A:

value predicted by ANFIS PSO w_L: accuracy of

LSTM in the train phase μ_L: value predicted by

LSTM

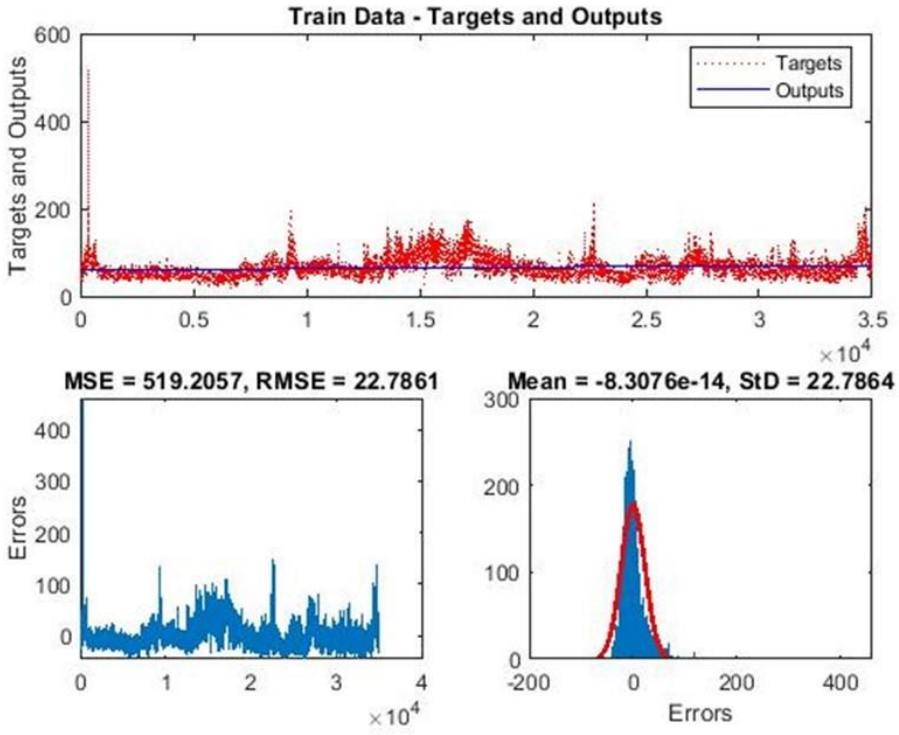

Figure 1: Load prediction by neural fuzzy system and DE optimization algorithm

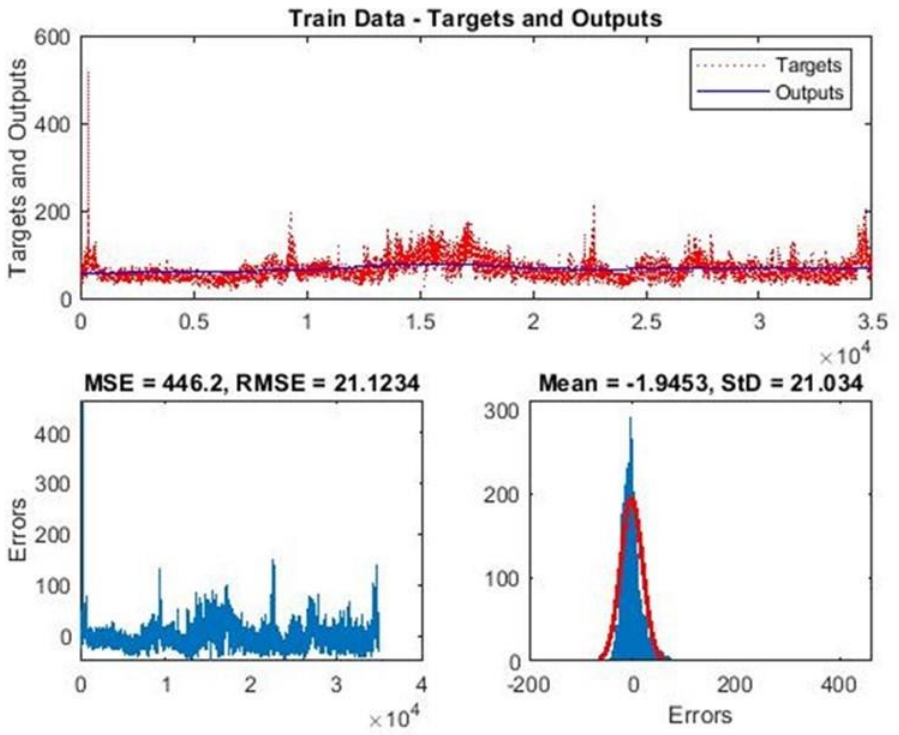

Figure 2: Load prediction by neural fuzzy system and GA optimization algorithm

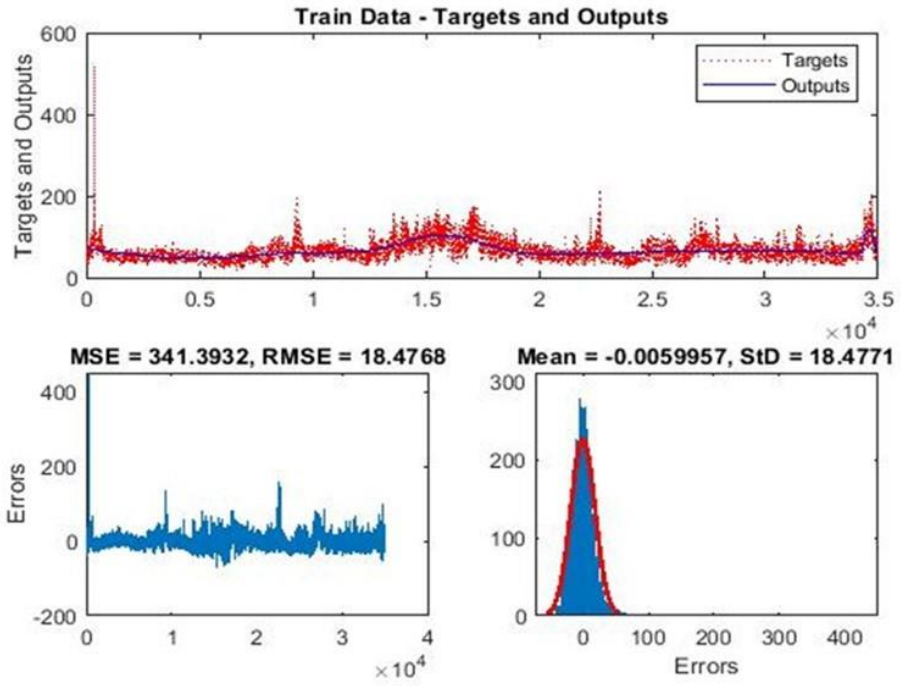

Figure 3: Load prediction using deep learning

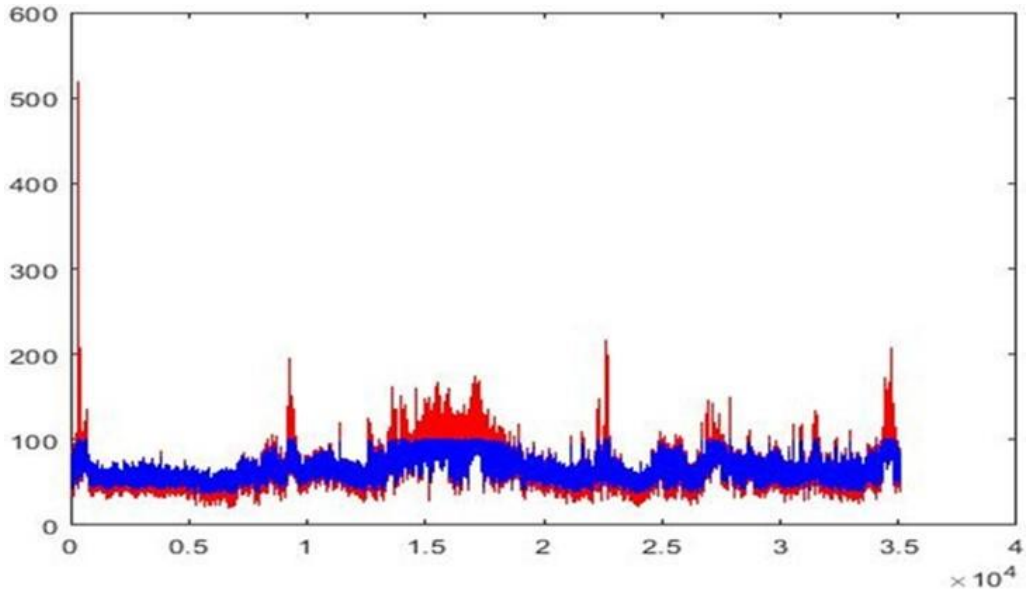

Figure 4: Load prediction using Weighted average method

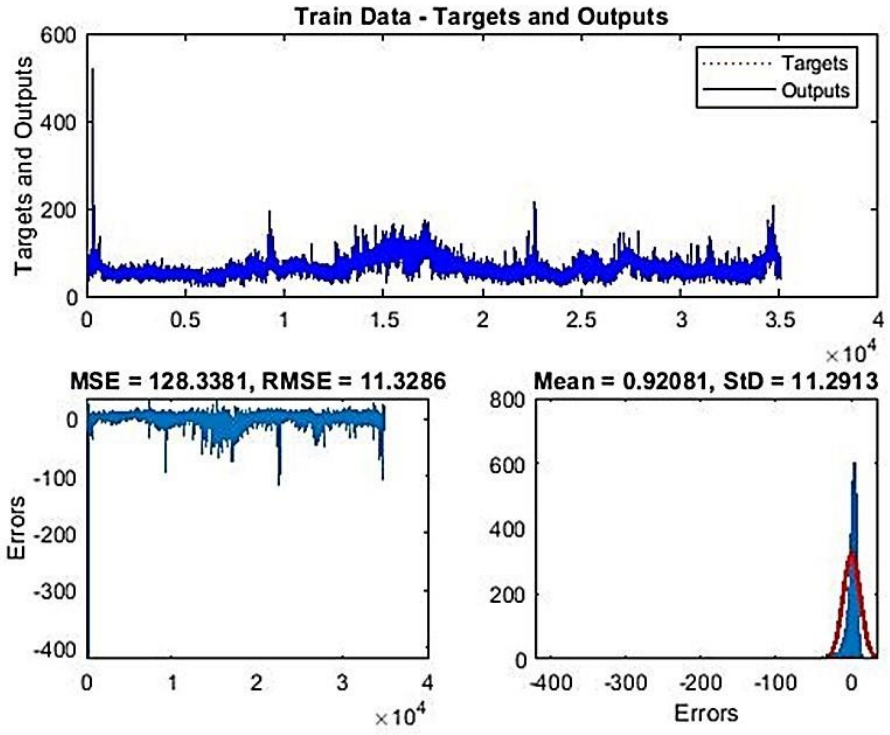

Figure 5: Comparison of the error of different methods of price prediction

6 Conclusion

In this article, firstly, the stock index data of the refinery industry is pre-processed and then using fuzzy neural system (ANFIS) with optimization by DE algorithm, fuzzy neural system (ANFIS) with optimization by GA algorithm, fuzzy neural system (ANFIS) With optimization by PSO algorithm, we will forecast the short-term price, according to the MSE and RMSE error criteria according to the table, we find that the method of load forecasting by combining ANFIS with PSO has less error than other methods, so it is the best method for combining The information is selected by weighted average method with deep learning network, then load prediction is done by LSTM alone to compare with the combination of the neural fuzzy system (ANFIS) optimized by PSO algorithm and LSTM deep learning. Finally, the simulation results from the combination With the help of weighted average, the two mentioned methods show that combining weighted information with higher accuracy than other simulated methods can predict the load and has less error than other methods.

References

[1] N. Sran and N. Kaur, *Comparative analysis of existing load balancing techniques in cloud computing*, Int. J. Engin. Sci. Inv. **2** (2013), no 1.

[2] A. Khetan, V. Bhushan, and S. Chand Gupta, *A novel survey on load balancing in cloud computing*, Int. J. Engin. Res. Technol. **3** (2013), no. 2.

[3] Y. Ranjith Kumar, M. Madhu Priya, and K. Shahu Chatrapati, *Effective distributed dynamic load balancing for the clouds*, Int. J. Engin. Res. Technol. **2**(2013), no. 2.

[4] S. Asadzadeh and M.A. Shayegan, *Diagnosis of atherosclerosis of the coronary arteries of the heart with data mining and machine learning techniques*, Trans. Machine Intell. **6** (2023), no. 4, 198–207.

[5] S. Begum and C.S.R. Prashanth, *Review of load balancing in cloud computing*, Int. J. Comput. Sci. Iss. **10** (2013), no. 1, 343.

[6] A. Hamo and A. Saeed, *Towards a reference model for surveying a load balancing*, Int. J. Comput. Sci. Network Secur.**13** (2013), no. 2, 42.

[7] D. Rastogi, A. Bansal, and N. Hasteer, *A techniques of load balancing in cloud computing: A Survey*, 7th Int. Conf. Comput. Sci. Engineering (CSE), April, 2013.

[8] N.J. Kansal and I. Chana, *Cloud load balancing techniques: A step towards green computing*, Int. J. Comput. Sci. Iss. **9** (2012), no. 1.

[9] A.A. Rajguru and S.S. Apte, *A comparative performance analysis of load balancing algorithms in distributed system using qualitative parameters*, Int. J. Recent Technol. Engin. **1** (2012), no. 3, 175–179.

[10] K. Al Nuaimi, N. Mohamed, M. Al Nuaimi, and J. Al-Jaroodi, *A survey of load balancing in cloud computing: Challenges and algorithms*, IEEE Second Symp. Network Cloud Comput. Appl., 2012.

[11] H. Ehsani Chimeh and M. Karami, *Spam detection from big data based on evolutionary data mining systems*, Trans. Machine Intell. **1** (2018), no. 1, 1–9.

[12] H. Avini, Z. Mirzaei Zavard Jani, and A. Avini, *Improved recommender systems using data mining*, Trans. Machine Intell. **5** (2022), no. 2, 97–114.