The Prediction of Stock Prices

Dingwei Bai*

Information technology, Monash University, Clayton, Australia

*Corresponding author: dbai0008@student.monash.edu

Abstract. The prediction of stock prices has consistently captured the attention of numerous analysts and researchers. Due to the influence of various variables such as economics, politics, investment psychology, and trading techniques on the price trends of stocks, forecasting stock prices inherently presents a challenging problem. In order to accurately predict the changing trends of stock prices, this study proposes a hybrid forecasting model known as ARIMA-SVM. This model is capable of simultaneously accommodating both the linear and nonlinear features of stock price data. Empirical research is conducted using stock price data from four sectors, and a comparison is made between the predictive accuracy of the ARIMA-SVM model, ARIMA model, and SVM model. The results indicate that the predictive accuracy of the ARIMA-SVM model, which integrates two individual models is enhanced.

Keywords: Stock Prices, Deep Learning, Modelling

1 Introduction

The prosperity and stability of financial markets can to a certain extent promote the development of a country's economy. As a vital component of the capital market, the stock market plays an indispensable role in optimizing national resource allocation, raising funds for enterprises, and allowing investors to partake in the dividends of economic growth. The operation of the stock market aligns with that of the macroeconomy, with its cyclical fluctuations reflecting and responding to economic changes, acting as a valuable economic indicator [1].

The analysis and prediction of the stock market have significant implications for nations, businesses, and individuals alike. For a nation, stock market prediction aids in macroeconomic risk management and drives financial reform. At the enterprise level, stock market forecasting offers guidance for fundraising and strategic decision-making. For individuals, the analysis and prediction of the stock market can assist investors in making informed investment decisions, better planning investment portfolios, and achieving financial objectives.

However, due to the dynamic nature of the market, analyzing and predicting stock market trends and price behaviors pose considerable challenges. In fact, the stock market is influenced by numerous interrelated factors such as external political factors, macroeconomic variables, industry-specific variables, company-specific variables, and

C. Chen et al. (eds.), *Proceedings of the 3rd International Conference on Digital Economy and Computer Application (DECA 2023)*, Atlantis Highlights in Computer Sciences 17,
https://doi.org/10.2991/978-94-6463-304-7_77

investor psychological factors, all of which contribute to the volatility of stock prices and the uncertainty of stock research [2].

Currently, stock price prediction techniques can be categorized into three main approaches: (1) Statistical methods. Within time series analysis, several models have been applied to predict trends and prices in the stock market, such as the Autoregressive Integrated Moving Average (ARIMA) model [3] and the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model [4]. However, statistical models typically handle linear prediction models, require variables to follow a statistically normal distribution for better predictive performance, and can exhibit biases in long-term predictions [5-6]. (2) Machine learning methods. These include Random Forest (RF) [7], Support Vector Machine (SVM) [8], and common Artificial Neural Network (ANN) algorithms [9]. While machine learning methods have achieved some success in stock price prediction, their predictive performance and deeper information extraction capabilities are often limited. (3) Deep learning methods. Compared to statistical and machine learning methods, deep learning methods possess more complex learning network structures, such as Recurrent Neural Networks (RNN) [10] and Convolutional Neural Networks (CNN) [11]. Relying on intricate hidden layers, deep learning can more accurately extract hidden information from massive datasets. Given the intricate price features of the stock market, deep learning exhibits strong predictive performance [12]. Despite producing slightly better predictive results, deep learning models suffer from issues of "slow convergence and high computational complexity" and demand substantial training data, thus hindering their scalability and usage.

Practical experience has shown that a single predictive model struggles to meet the demands of stock market forecasting. Therefore, researchers have proposed hybrid models as an approach to enhance predictive accuracy. A hybrid model involves combining two or more different models, exploiting their complementarity to form a unified model. For instance, Zhang et al. (2019) [13] proposed an algorithm that integrates Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) to predict stock price trends and patterns. A plethora of research demonstrates that hybrid models, as opposed to single models, can systematically and comprehensively integrate data from the stock market, thereby improving model fit and increasing the precision of stock prediction studies.

In summary, this study introduces the ARIMA-SVM hybrid model: the ARIMA model accurately addresses linear features within stock market data, while the SVM model handles nonlinear features. The combination of the two enhances the prediction of stock price trends and improves predictive performance, all while avoiding high computational demands. We first elucidate the ARIMA model and the Support Vector Machine model (SVM), then propose the ARIMA-SVM hybrid model based on these two. Finally, empirical research is conducted on stock price data from four sectors. The results indicate that, compared to single models, our proposed ARIMA-SVM model achieves higher predictive accuracy.

2 Stock Prediction Models

2.1 Hybrid Model

2.1.1 ARIMA Model

The section headings are in boldface capital and lowercase letters. Second level headings are typed as part of the succeeding paragraph (like the subsection heading of this paragraph). All manuscripts must be in English, also the table and figure texts, otherwise ${}_s w$ we cannot publish your paper. Please keep a second copy of your manuscript in your office.

Stock prices represent a type of time series data, which is a collection of observed data arranged in chronological order, characterized by nonlinearity and non-stationarity. Let $\{Y_t, t \in T\}$ denote stock prices, and $\{y_t, t = 1, 2, \ldots, n\}$ represent the n ordered observed values of stock prices, forming a sequence with a length of n.

The ARIMA (Autoregressive Integrated Moving Average) model is an extension of the ARMA (Autoregressive Moving Average) model. The ARMA(p, q) model combines the AR(p) model and the MA(q) model. The former captures the observed mean reversion effect in financial markets, while the latter explains random disturbance effects. However, the limitation of the ARMA model lies in its inability to handle non-stationary sequences. The ARIMA model builds upon the ARMA model by incorporating differencing to transform non-stationary time series into stationary ones.

In the ARIMA(p, d, q) model, the future value of a variable is assumed to be a linear combination of historical values and historical disturbance terms. The mathematical formula for this model can be expressed as follows:

$$\begin{cases} \mathbf{\Phi}(B)(1-B)^d y_t = \mathbf{\Theta}(B)\epsilon_t \\ \mathbf{\Phi}(B) = 1 - \sum_{i=1}^{p} \phi_i B^i \\ \mathbf{\Theta}(B) = 1 - \sum_{j=1}^{q} \theta_j B^j \\ E(\epsilon_t) = 0, Var(\epsilon_t) = \sigma_\epsilon^2, E(\epsilon_t \epsilon_s) = 0, s \neq t \\ E y_s \epsilon_t = 0, \forall s < t \end{cases} \tag{1}$$

Where:

y_t is the stock price at time t.

ϵ_t is the random disturbance or error term at time t.

ϕ_i and θ_j are coefficients.

B is the lag operator.

$\boldsymbol{\Phi}(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the autoregressive coefficient polynomial.

$\boldsymbol{\Theta}(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ is the moving average coefficient polynomial.

p and q are the orders of autoregressive and moving average terms, respectively.

d is the degree of differencing required to make a non-stationary time series stationary.

p, q and d are all intergers.

The ARIMA model can be simplified into the ARMA(p, q) model, AR(p) model, and MA(q) model, respectively. For instance, the ARIMA(1, 0, 1) model can be represented as follows:

$$y_t = \theta_0 + \phi_1 y_{t-1} + \epsilon_t - \theta_1 \epsilon_{t-1} \tag{2}$$

Indeed, when $p = q = 0$ and $d = 1$, the ARIMA(0, 1, 0) model simplifies to what is known as a random walk model or sometimes referred to as a drunkard's walk model.

2.1.2 SVM model

Support Vector Machine (SVM) was introduced by Vapnik in 1995. It is based on the learning theory of VC dimension, constructing theory and minimal structural risk, to predict sample data [14]. SVM possesses robustness and can utilize kernel functions for non-linear mapping, effectively addressing the "curse of dimensionality". Moreover, unlike some other classification or regression methods that yield local optimal solutions, SVM employs algorithms to obtain the global optimal solution of the objective function. Let (x, y) represent the sample data, where $x = \{x_1, x_2, \ldots, x_i\}$ is the input vector or feature, and $y = \{y_1, y_2, \ldots, y_i\}$ corresponds to the target value or label of x. The input vector x is nonlinearly mapped to a high-dimensional feature space, and the nonlinearity between y and x can be transformed into a linear relationship after mapping. The SVM regression function formula is as follows:

$$f(x) = \omega^T \phi(\mathbf{x}) + b \tag{3}$$

Certainly, the estimated values of coefficients ω and b can be obtained by minimizing a certain function:

$$R(\omega, \zeta, \zeta^*) = \left\{\frac{1}{2}\|\omega\|^2 + C\left(\sum_{i=1}^{N}(\zeta + \zeta^*)\right)\right\} \tag{4}$$

$$s.t. \begin{cases} y_i - \omega^T \phi(x_i) - b \le \epsilon + \zeta_i \\ -y_i + \omega^T \phi(x_i) + b \le \epsilon + \zeta_i^* \\ \zeta_i \, \zeta_i^* \ge 0, i = 1,2, \ldots, N \end{cases} \tag{5}$$

the expression includes the Euclidean norm term and the empirical risk term. The Euclidean norm is used to control the complexity of the model, while the empirical risk term constrains the error obtained during training. By introducing Lagrange multipliers β and β^*, and maximizing the dual function of the equation, the expression is transformed as follows:

$$f(x, \beta, \beta^*) = \sum_{i=1}^{N}(\beta_i - \beta_i^*) K(x_i, x) + b \tag{6}$$

Here, $K(x_i, x_j)$ is referred to as the kernel function. The kernel function is a crucial component of Support Vector Machines (SVM) that allows the SVM to operate in a high-dimensional feature space without explicitly computing the transformations. The kernel function computes the dot product of two feature vectors x_i and x_j , which is equivalent to $K(x_i, x_j) = \phi(x_i)\phi(x_j)$, where ϕ represents the mapping of the feature vectors to a higher-dimensional space.

Commonly used kernel functions include Linear Kernel, Polynomial Kernel and Radial Basis Function (RBF) Kernel (also known as Gaussian Kernel). In the context of using SVM for predicting and analyzing the stock market, the Radial Basis Function (RBF) kernel, also known as the Gaussian kernel, often yields the best results. Its formula is as follows:

$$K(x_i, x_j) = \frac{exp\left(-\|x_i - x_j\|^2\right)}{2\sigma^2} \tag{7}$$

2.1.3 ARIMA-SVM model

The direction and trends of stock prices are intricate, often requiring a combination of models to extract meaningful insights from the data. Thus, a hybrid approach that encompasses both linear and non-linear modeling is imperative. The ARIMA model is adept at capturing linear features within the data, while the SVM model excels in handling non-linear features. The ARIMA-SVM hybrid model amalgamates the strengths of these two models and mitigates their weaknesses, allowing for a more comprehensive extraction of data information and yielding more accurate predictions. The mathematical formula for the hybrid model can be expressed as follows:

$$Y_t = L_t + N_t \tag{8}$$

Where:

Y_t represents the time series data.

L_t corresponds to the linear component of the hybrid model.

N_t corresponds to the non-linear component of the hybrid model.

In this research, a weighted averaging method is used for combination. We assign weights to the linear component (ARIMA model) and the non-linear component (SVM model) of the hybrid model. Hence, the final model in this study is as follows:

$$Y_t = (W_{ARIMA} \times Y_{ARIMA} + W_{SVM} \times Y_{SVM}) \times \frac{1}{2} \tag{9}$$

Where:

$\mathrm{W_{ARIMA}}$ and W_{SVM} are the weights assigned to the ARIMA and SVM models, respectively.

Y_{ARIMA} is the prediction from the ARIMA model.

Y_{SVM} is the prediction from the SVM model.

Empirical Study

To assess the predictive performance of the proposed method in this study, the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are employed as objective metrics. Let N represent the number of test data samples, $x_k(i)$ and

$x_k^{pre}(i)$ denote the true value and the model-predicted value of the ith stock price to be forecasted. The formulas for calculating RMSE and MAE are as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_k(i) - x_k(i)^{pre}| \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[x_k(i) - x_k(i)^{pre}]^2} \tag{11}$$

3 Data and Methodology

We selected historical stock price data from the Shenzhen Stock Exchange's New Energy sector for Enjie Energy (SZ#002812) and Ganfeng Lithium (SZ#002460). Figures 1 and 2 depict the time series plots of the stock prices for Enjie Energy from September 14, 2016, to February 2, 2023, and for Ganfeng Lithium from February 18, 2013, to February 2, 2023, respectively.

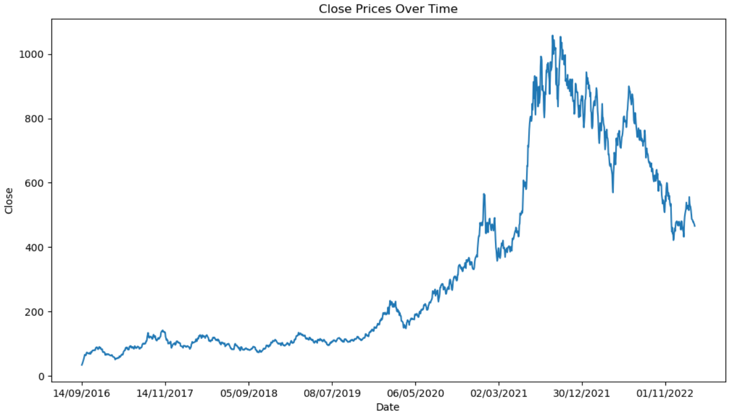

Fig. 1. Time series plot of Enjie Energy's stock prices

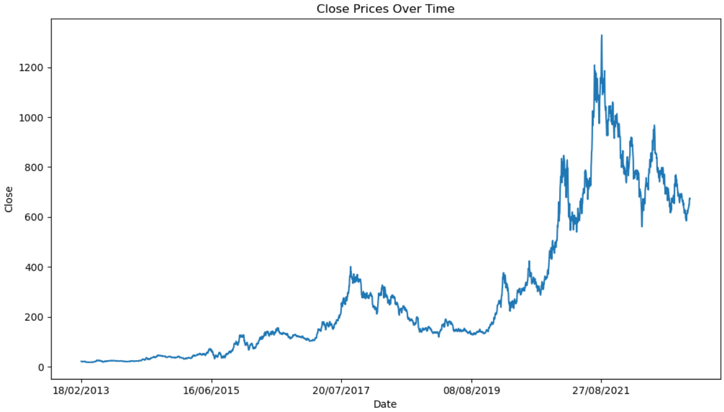

Fig. 2. Time series plot of Ganfeng Lithium's stock prices

From the visualizations, we observe that the stock price of Enjie Energy gradually increased from 2016, reaching a peak in 2021. While there was a minor increase afterward, the overall trend appears to be downward. On the other hand, the stock price of

Ganfeng Lithium showed a slight peak in 2017, followed by a continuous rise and a peak in 2021, followed by a general decline.

Before conducting data mining, we shifted the stock price data by 2 days. This adjustment establishes a correspondence between inputs and outputs for the model, aiding in defining appropriate independent and dependent variables for analysis.

In addition, we imported stock price data from three other sectors using the Quandl library. These sectors include:

- Google's stock price data from the U.S. Technology sector.
- Bank of America's stock price data from the U.S. Financial sector.
- Exxon Mobil's stock price data from the U.S. Energy sector.
- By incorporating data from these diverse sectors, we aim to enhance the robustness and breadth of our analysis.

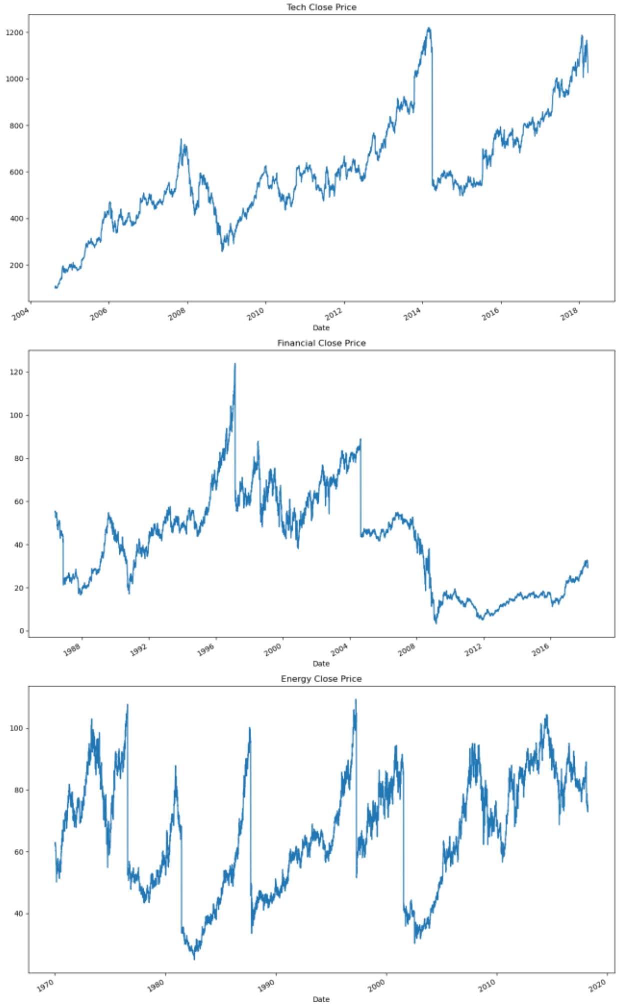

Fig. 3. Time Series Plots of Stock Prices in the Technology, Financial, and Energy Sectors.

Firstly, we established SVM models for both stock datasets using Gaussian kernel function. The input comprised both the unshifted and shifted stock data. We trained the SVM models on the training dataset to subsequently predict the test dataset. Subsequently, we conducted stationarity tests on both stock datasets. The optimal values of p and q were chosen by analyzing the ACF and PACF plots along with the BIC criterion. The data were fitted using ARIMA models, and white noise tests were performed on the resulting residuals. Experimental results indicated that an ARIMA(3,1,5) model was established for Enjie Energy, ARIMA(6,1,4) for Ganfeng Lithium, ARIMA(0,1,1) for Google, and ARIMA(1,1,2) for Bank of America and Exxon Mobil. The p-values for the residuals of Enjie Energy, Ganfeng Lithium, Google, Bank of America, and Exxon Mobil exceeded 0.05, confirming the models' validity. Lastly, we combined the two individual models using weighted averaging to create the ARIMA-SVM hybrid model.

4 Results

Table 1 and Table 2 present the prediction results of the three models for Enjie Energy and Ganfeng Lithium, respectively, on a 10% training dataset. Table 3 showcases the prediction performance of the three models across the three sectors with a forecasting horizon of 10 days. Figures 4 and 5 visualize the prediction outcomes of the three models for Enjie Energy and Ganfeng Lithium, respectively, on a 10% training dataset.

Table 1. Prediction Results for Enjie Energy's Three Models

Model	RMSE	MAE
SVM	29.48	0.04
ARIMA	16.82	0.04
ARIMA-SVM	16.82	0.02

Table 2. Prediction Results for Ganfeng Lithium's Three Models

Model	RMSE	MAE
SVM	38.03	0.04
ARIMA	21.20	0.04
ARIMA-SVM	21.20	0.02

Table 3. Prediction Results for Three Sectors and Three Models

Data	Model	RMSE	MAE
Google	SVM	34.63	0.03
	ARIMA	26.35	0.02
	ARIMA-SVM	26.34	0.02
Bank of America	SVM	1.04	0.89
	ARIMA	0.79	0.02
	ARIMA-SVM	0.79	0.68
Exxon Mobil	SVM	1.12	0.01

Data	Model	RMSE	MAE
	ARIMA	0.88	0.01
	ARIMA-SVM	0.87	0.01

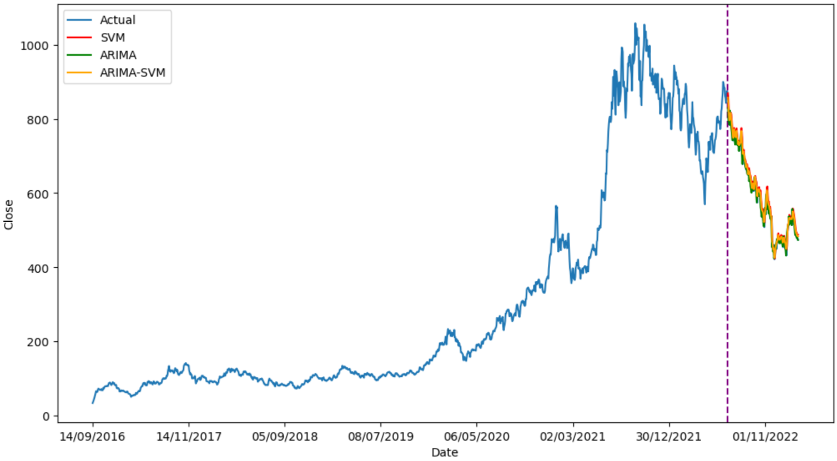

Fig. 4. Visualization of Prediction Results for Enjie Energy's Three Models

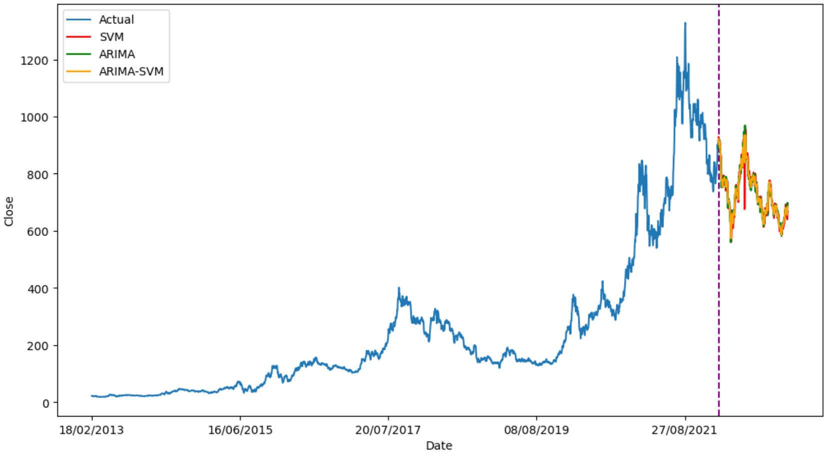

Fig. 5. Visualization of Prediction Results for Ganfeng Lithium's Three Models

By comparing the Mean Absolute Error and Root Mean Squared Error, it is evident that the predictive performance of the single ARIMA model is superior to that of the single SVM model, as it captures the stock price trend changes more effectively. Overall, the ARIMA-SVM hybrid model demonstrates an improvement in predictive accuracy compared to the individual models.

5 Conclusion

The trends and tendencies of stock prices have always been a prominent research area within the financial investment market, playing a crucial role in the realm of finance. However, stock price fluctuations are influenced by numerous factors, including economic indicators, political events, and corporate actions like mergers or splits, as well as unforeseeable occurrences. Consequently, accurately predicting stock prices is a highly challenging task, yet one that holds significant value. With the continuous improvement of statistical and machine learning techniques, an increasing number of researchers have proposed the concept of hybrid models. By combining two or more models and leveraging their respective strengths and weaknesses, predictive accuracy can be enhanced.

This study established an ARIMA-SVM hybrid model capable of simultaneously handling both linear and nonlinear features within stock price data. Empirical research was conducted on stock price data from four sectors, comparing the predictive outcomes with those of the two individual models. The results showcased that the hybrid model outperforms the single models. The fusion of ARIMA and SVM advantages in the hybrid model improves the accuracy of stock price prediction, thereby providing valuable insights for relevant investors.

However, further exploration is warranted to determine how to select optimal parameters for the hybrid model. In theory, the fusion of two individual models should naturally yield a better-performing hybrid model. However, reality does not always align with theory. In future research, the focus should expand to address parameter selection and weighting for hybrid models, aiming to further enhance their predictive effectiveness.

References

1. Guo, Zhou, Chen. Economy barometer analysis of China Stock Market:A dynamic analysis based on the thermal optimal path method [J]. Journal of Management Sciences in China, 2012，(15)：1-9.
2. Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. Expert systems with applications, 67, 126-139.
3. Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS 16th international conference on computer modelling and simulation (pp. 106-112). IEEE.
4. Awartani, B. M., & Corradi, V. (2005). Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries. International Journal of forecasting, 21(1), 167-183.
5. Pan, Y. (2023). Stock Price Forecast of Chinese Leading Liquor Stocks Using the Simple Moving Average. BCP Business & Management, 36, 16–24.
6. Cheng, C. H., & Yang, J. H. (2018). Fuzzy time-series model based on rough set rule induction for forecasting stock price. Neurocomputing, 302, 33-45.
7. Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.

8. Yang, Y., & Yang, Z. (2005). Financial Time Series Forecasting Based on Support Vector Machines. Systems Engineering - Theory & Practice, 14(2), 176-181.
9. Wanjawa, B. W., & Muchemi, L. (2014). ANN model to predict stock prices at stock exchange markets. arXiv preprint arXiv:1502.06434.
10. Zhu, Y. (2020, October). Stock price prediction using the RNN model. In Journal of Physics: Conference Series (Vol. 1650, No. 3, p. 032103). IOP Publishing.
11. Mehtab, S., & Sen, J. (2020, November). Stock price prediction using CNN and LSTM-based deep learning models. In 2020 International Conference on Decision Aid Sciences and Application (DASA) (pp. 447-453). IEEE.
12. Deng, F., & Wang, H. (2018). Application of LSTM Neural Network in Stock Price Trend Prediction—A Study Based on Individual Stock Data from the U.S. and Hong Kong Stock Markets. Financial Economics, (14), 96-98.
13. Zhang, P., Liu, H., Pei, D., & Wang, X. (2019). Stock Price Prediction Based on SVM-KNN. Journal of Statistics and Applications, 8(6), 859-871.
14. Vapnik V. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995.