Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/140109

Analysis of the Stock Market Using the Integration of Statistical and Machine Learning Models

M. Aripov[1,*], *Azhari A. Alhag*[2], *Siham A. Shaddad*[3], *Nadia B. Mohammed Ali*[4], *and Hiba A. A. A. Hussin*[4]

[1] Department of Applied Mathematics and Computer Analysis, Faculty of Mathematics, National University of Uzbekistan, Tashkent, Uzbekistan
[2] Department of Mathematics and Statistics, College of Science, Taif University, Taif, Saudi Arabia
[3] Student Affairs Division, Prince Mohammed Bin Fahd University, Dhahran, Saudi Arabia
[4] Department of Finance, Faculty of Business, Imam Mohammed Ibn Saudi Islamic university, Riyadh, Saudi Arabia

Received: 27 Oct. 2024, Revised: 1 Dec. 2024, Accepted: 13 Dec. 2024
Published online: 1 Jan. 2025

Abstract: Time series prediction is a critical task in various fields, including finance, economics, and field of finance. In this study, we assess the forecasting performance of three distinct models-Artificial Neural Networks (ANN), Autoregressive Integrated Moving Average (ARIMA), and a Hybrid model-using a dataset of Saudi Basic Industries Corporation (SABIC) stock prices, covering the period from January 1, 2016, to August 10, 2024. The models are evaluated based on three widely recognized error metrics: Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Symmetric Mean Absolute Percentage Error (sMAPE). The Hybrid model, which integrates multiple forecasting approaches, consistently outperforms both the ANN and ARIMA models across all three metrics. The results reveal that the Hybrid model provides the most accurate and stable predictions, with significantly lower error values, including a notably lower coefficient of variation (CoefVar) compared to the other models. The ANN model, while effective, exhibits slightly higher variability and error rates, while ARIMA struggles to capture extreme values in the data. Boxplots of actual and predicted values demonstrate that all models successfully capture the general trends in the data without producing substantial outliers. Based on these findings, the Hybrid model is recommended for stock price forecasting, particularly when prediction accuracy, stability, and minimal variability are prioritized.

Keywords: Finance, Artificial Neural Networks, Saudi Basic Industries Corporation, Autoregressive Integrated Moving Average, predictions.

1 Introduction

Saudi Basic Industries Corporation (SABIC) ranks among the largest petrochemical enterprises globally and serves as a fundamental component of the Saudi Arabian economy [1]. Established in 1976, SABIC significantly helps to the advancement of the downstream industries sector and aids in the diversification of the Saudi economy [2]. SABIC operates in over 50 countries, establishing itself as a significant entity in the worldwide petrochemical and chemical industry [3]. The company emphasizes innovation and sustainable development, thereby improving its capacity to address the demands of both global and local markets [4]. SABIC significantly contributes to the Kingdom's Vision 2030 by advancing non-oil businesses and fostering industrial diversification [5].

Hybrid machine learning models are essential in contemporary technology as they integrate the advantages of many algorithms to enhance predictive accuracy and efficiency [6]. Hybrid models amalgamate many strategies, including supervised and unsupervised learning or ensemble methods, to address intricate problems that singular models may find challenging [7]. They are especially significant in sectors such as healthcare, banking, and natural language processing, where precision and adaptability are crucial [8]. Hybrid models can improve performance, diminish bias, and yield more robust predictions in dynamic contexts by utilizing the characteristics of diverse methodologies [9].

* Corresponding author e-mail: hahussin@imamu.edu.sa

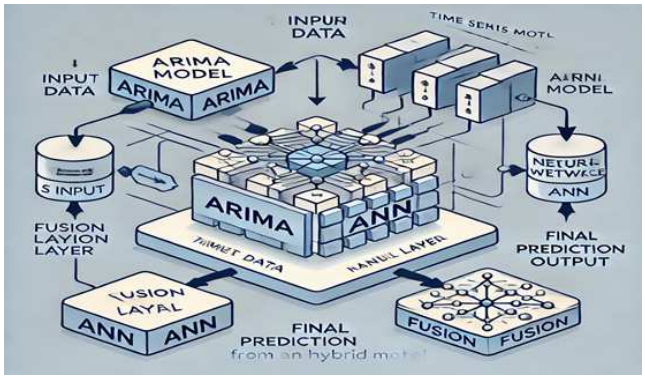

Fig. 1: Hybrid ARIMA-ANN Diagram

1.1 The objective of the study:

This study aims to assess the predictive efficacy of three analytical models—Artificial Neural Networks (ANN), Autoregressive Integrated Moving Average (ARIMA), and a hybrid model—in projecting the stock prices of Saudi Basic Industries Corporation (SABIC). The research examines the correctness of various models by three statistical metrics, intending to determine the most effective model for accuracy, stability, and minimal variance.

1.2 Study Problem:

Stock price forecasting presents a significant problem owing to the volatility and uncertainty inherent in financial markets. Despite the utilization of numerous statistical and intelligent models in this domain, a knowledge gap persists regarding the efficacy of hybrid models that integrate various methodologies in comparison to conventional models like ANN and ARIMA. The study problem focuses on establishing the most precise and stable model for projecting SABIC stock values, aiding investors and decision-makers in making educated choices.

1.3 Significance of the study:

This work enhances the scientific literature on time prediction by evaluating the performance of several models and assessing the efficacy of the hybrid strategy in financial markets.

The study's findings offer significant advice to investors and portfolio managers regarding the most dependable model for forecasting stock values, hence enhancing investment methods. This study enhances financial forecasting models by examining their strengths and shortcomings, hence facilitating the development of more precise methodologies in the future.

1.4 Literature Review:

The application of machine learning techniques in the financial sector for predictive purposes is a relatively novel subject, steadily increasing in significance [10]. Predicting asset prices and returns continues to be one of the most stimulating tasks for quantitative finance professionals. The substantial rise in data generated and collected in recent years presents an opportunity to utilize machine learning techniques [11]. Forecasting stock prices and predicting market trends are formidable endeavors. Throughout the years, scholars have suggested numerous solutions to these difficulties [12]. Artificial Intelligence (AI) has fundamentally transformed financial forecasting, enhancing risk assessment and decision-making processes [13]. Clustering algorithms are unsupervised learning techniques that discern links and patterns within data. They oversee unlabeled training datasets utilized in classification and decision-making algorithms, particularly within the realm of efficient frontier investment methods [14].

Table 1: Descriptive Statistics (SABIC), and Prediction value

Variable	N	Mean	Coef Var	Minimum	Maximum
SAUDI BASIC INDUSTRIES	2349	99.538	16.57	60.880	139.000
Predicted Value Hybird Model	2349	99.675	15.13	81.120	127.670
Predicted Value ANN Model	2349	99.536	16.57	60.420	139.210
Predicted Value ARIMA Model	2349	99.547	16.57	60.491	139.196

2 Models employed in the study

2.1 ARIMA

It is a statistical model extensively employed for time series forecasting, incorporating three primary components [15].

Autoregressive (AR) component: Utilizes historical values of the series to forecast future values. It is denoted by the parameter p [16]

Integrated (I) component: Transforms the data to achieve stationarity. Denoted by the parameter q [16].

The Moving Average (MA) component utilizes historical forecast errors to enhance forecasts. Characterized by the parameter (p,d,q) [18].

The ARIMA model is represented as ARIMA (p, d, q) [19]:

where

p = The number of lagged observations in the autoregressive model.

d = The frequency of differencing required to attain stationarity in the data.

q = The number of lagged forecast errors in the moving average model

2.2 ANN

It is a computer model derived from the architecture and operation of the human brain. It comprises interconnected neurons (nodes) organized in layers that process input data to discern patterns and provide predictions [20].

Essential Elements of Artificial Neural Networks [21]:

Input Layer: Accepts unprocessed data attributes.

Concealed Layers: Execute calculations and identify patterns.

Output Layer: Generates the ultimate prediction or classification.

Weights and biases: Modified throughout training to enhance model performance.

Activation Functions: Introduce non-linearity, enabling the network to comprehend intricate linkages.

Backpropagation: An algorithm employed to adjust weights according on the discrepancy between real and forecasted values [22].

2.3 Hybird ARIMA- ANN

This method involves the ARIMA model initially analyzing the time series data to identify linear patterns and trends (see Fig 1). The residual errors (the discrepancies between actual and forecasted values) from ARIMA are subsequently input into an Artificial Neural Network (ANN), which captures the nonlinear relationships within the data. The ultimate forecast is derived from the integration of ARIMA and ANN results [23].

3 Statistical discussion

A time series dataset with 2,349 observations was gathered for the stock price Saudi Basic Industries Corporation (SABIC) between January 1, 2016, and August 10, 2024. The dataset offers summary statistics for three distinct models alongside the actual values of (SABIC)

The table presents the statistical summary of the financial performance of SAUDI BASIC INDUSTRIES and three predictive models: Hybrid Model, Artificial Neural Network (ANN) Model, and Autoregressive Integrated Moving Average (ARIMA) Model. Each of them consists of 2349 data points and the following key statistics are provided: mean, coefficient of variation (CoefVar), minimum, and maximum. All models (Hybrid, ANN, and ARIMA) produce predicted values that are very close to the actual observed value for SABIC, with means ranging from 99.536 to 99.675. The predicted values are almost identical to the actual mean (99.538), indicating a good level of accuracy in the predictions. The Hybrid Model has the lowest coefficient of variation (15.13), which suggests that it has less relative variability in its predictions compared to the other models. The ANN Model and the ARIMA Model both have a CoefVar of 16.57, indicating higher variability in their predictions. The Hybrid Model predicts values within a narrower range (81.120 to 127.670), suggesting that its predicted values are more consistent compared to the ANN Model and ARIMA Model, which have a wider range (60.420–139.210 for ANN and 60.491–139.196 for ARIMA). A narrower range indicates fewer extreme predictions and greater reliability in the output. The Hybrid Model demonstrates the best performance among the three predictive models. It has the lowest variability (CoefVar) and a narrower prediction range, making it the most consistent and reliable in terms of forecasting. While all models are accurate, the Hybrid Model stands out in terms of stability and precision.

The (SABIC) actual data, Predicted Value ANN Model, and Predicted Value hybrid Model have almost identical distributions, with similar, minimum,Q1 , median , Q3 , and maximum values (see Fig. 2-5).

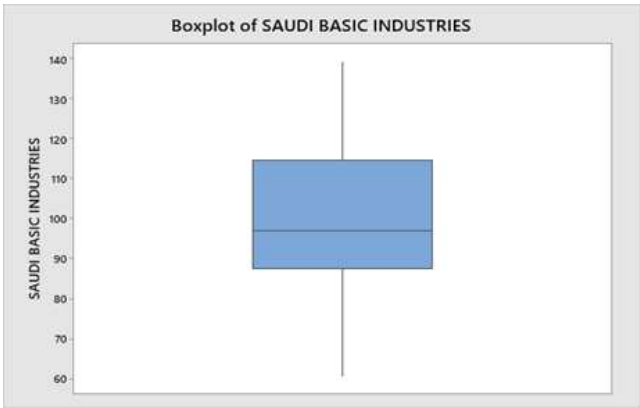

Fig. 2: Boxplot of Actual Data for Saudi Basic Industries

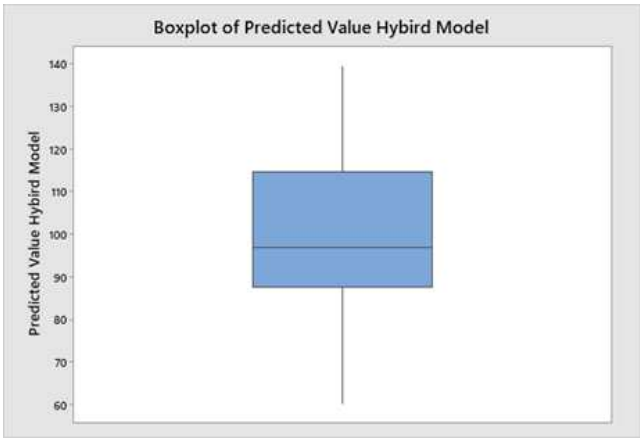

Fig. 3: Boxplot of Predicted Values for Hybrid Model

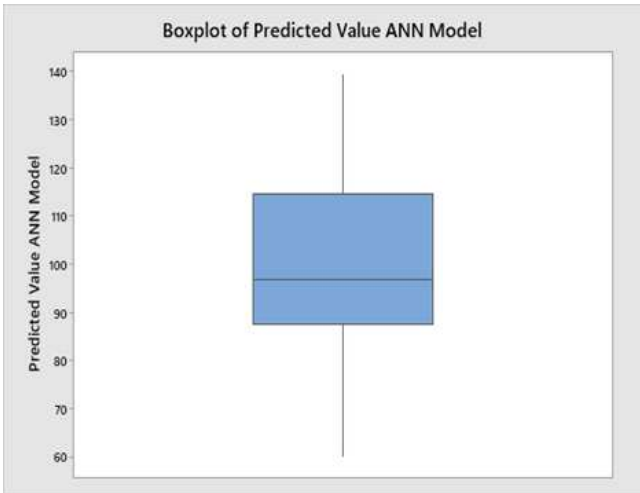

Fig. 4: ANN Model

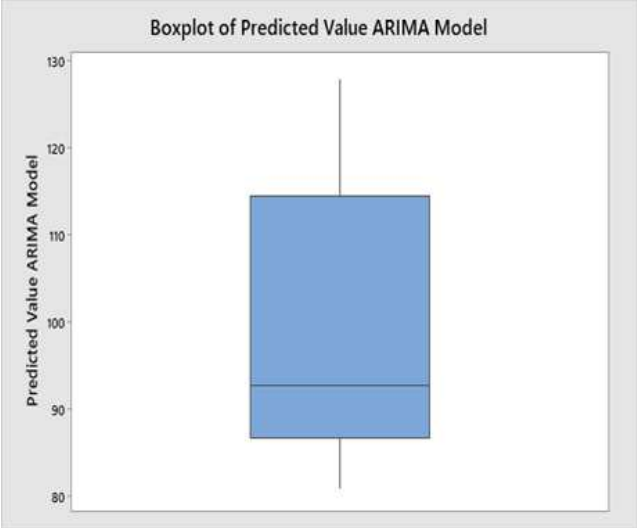

Fig. 5: ARIMA Model

The minimum and maximum values in Fig.1. are contained within the whisker range. This indicates the absence of substantial outliers in the actual data SAUDI BASIC INDUSTRIES. The data points fall within the anticipated range, as delineated by the whiskers (1.5 times the interquartile range above the third quartile and below the first quartile). The data exhibits no substantial outliers. Both the minimum (60.880) and highest (139.000) values fall inside the whisker range. The interquartile range of around 26.95 signifies a substantial dispersion of data around the median, with most values concentrated between 87.650 (Q1) and 114.600 (Q3). The distribution is nearly symmetric, exhibiting a little right skew, as the mean (99.538) marginally exceeds the median (96.920). This boxplot visually depicts the distribution of the real data, illustrating the concentration of values and the lack of extreme outliers. The whiskers of the Hybrid Model resemble those of the Actual Data. The lower whisker extends to 60.491, marginally below the actual data point of 60.880, while the upper whisker reaches 139.196, slightly beyond the real data's maximum of 139.000. The minor discrepancies suggest that the projected values from the Hybrid Model closely align with the actual data's range, exhibiting no significant deviations or outliers (see Fig.2), also the Hybrid Model's minimum value (60.491) and highest value (139.196) reside within the whisker range, signifying the absence of large outliers. The anticipated values in the Hybrid Model remain within the standard range of data and do not display excessive or anomalous values. In the ANN Model (refer to Fig. 3), the whiskers closely approximate those of the actual data. The lower whisker measures 60.420, marginally below

the real data's lower whisker of 60.880, while the higher whisker measures 139.210, slightly beyond the actual data's maximum of 139.000. This minor discrepancy indicates that the ANN Model forecasted values marginally exceeding the actual data range, although still within an acceptable distribution range. The minimum value (60.420) and highest value (139.210) remain within the whisker range, signifying the absence of significant outliers in the ANN Model. The ANN Model demonstrates excellent predicted accuracy relative to the real data, with no substantial discrepancies. The ARIMA Model's lower whisker (81.120) significantly exceeds the actual data's lower whisker (60.880), whilst the upper whisker (127.670) falls short of the real data's upper whisker (139.000). This indicates that the ARIMA Model predictions are confined to a more limited range, omitting some extreme low and high values found in the actual data, the model exhibits no major outliers, as all values reside within the whisker range. (see Fig.4). The dataset utilized for artificial neural network (ANN) analysis is partitioned into two main subsets: a training set and a testing set. The dataset comprises 2,348 cases, with one case omitted from the analysis. The dataset is divided with roughly 68% designated for training and 32% for testing, a standard allocation for machine learning applications, (see table 2).

Table 2: Case Processing Summary

		N	Percent
Sample	Training	1598	68.1%
	Testing	750	31.9%
Valid		2348	100.0%
Excluded		1	
Total		2349	

Table 3: Performance Comparison of prediction Models

	MSE	MAPE	sMAPE
ANN	56.40474	0.062826	0.061844
ARIMA	140.4943	0.083988	0.03888
Hybird	5.67934	0.004752	0.005173

This distribution guarantees that the neural network has sufficient data for learning, while also having a reliable selection of data for testing and confirming its predictions. To assess the accuracy of predictions made by different prediction techniques, the provided metrics (MSE, MAPE, and sMAPE) are used. three different models are compared (see table 3) To assess the accuracy of predictions made by different prediction techniques, the provided metrics (MSE, MAPE, and sMAPE) are used. three different models are compared (see table 3) Hybrid Model consistently outperforms both ANN and ARIMA across all three metrics (MSE, MAPE, and sMAPE), offering the best overall accuracy. This highlights the strength of ensemble approaches, which combine multiple models to capitalize on their respective strengths and mitigate their individual weaknesses.

To assess the performance of SABIC stocks and analyse the efficacy of predictive models in forecasting future prices (see Fig. 6), the figure illustrates ARIMA, ANN, and hybrid models.

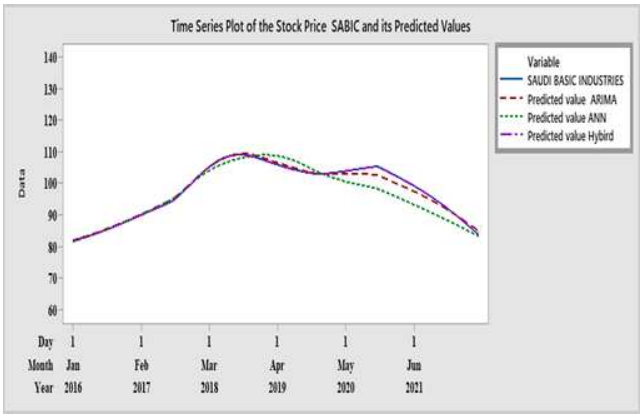

Fig. 6: Time series Plot of the Stock Price SABIC and its Prediction Values

The data demonstrates that ARIMA, ANN, and hybrid models yield accurate prediction. The hybrid and ANN models exhibit more alignment with the genuine values; all models successfully capture the overarching trends in the data.

4 Conclusion

The analysis of the time series dataset for Saudi Basic Industries Corporation (SABIC) stock prices, spanning from January 1, 2016, to August 10, 2024, provides valuable insights into the forecasting capabilities of three distinct predictive models: the Hybrid Model, Artificial Neural Network (ANN), and Autoregressive Integrated Moving Average (ARIMA). All models demonstrated high predictive accuracy, with their predicted values closely aligning with the actual data. The Hybrid Model consistently emerged as the best-performing model, outperforming both the ANN and ARIMA models in terms of Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Symmetric Mean Absolute Percentage Error (sMAPE). This result emphasizes the strength of ensemble methods, which combine multiple models to leverage their strengths and mitigate their weaknesses, resulting in more stable and accurate predictions. In terms of distribution and variability, the Hybrid Model exhibited the least variability, as indicated by its lower coefficient of variation (CoefVar) of 15.13, compared to 16.57 for both ANN and ARIMA. This suggests that the Hybrid Model's predictions are more consistent and reliable. Furthermore, the Hybrid Model's predicted values fall within a narrow range, which indicates that it avoids extreme deviations and provides more stable forecasts

References

[1] E. N. Žalışkan. The impact of strategic human resource management on organizational performance. Journal of Naval Sciences and Engineering, 6(2), 100-116, (2010).
[2] T. Alabdullatif. SABIC Green Logistics Systems & Profitability: To explore chemical industries green logistics and contribution to profitability with a particular case of SABIC, (2017).
[3] R., Maglad, R., Shaheen, & M., Samkari, . Impact of initial public offering on the financial performance of petrochemical industry in Saudi Arabia. In Proceedings of the International Conference on Industrial Engineering and Operations Management (Vol. 18, pp. 592-603), (2019).
[4] S. Silvestre, Bruno, and Diana Mihaela Ţîrcă. Innovations for sustainable development: Moving toward a sustainable future. Journal of cleaner production, (208), 325-332, (2019).
[5] L. F., Alqublan, The adoption of technologies in The Kingdom of Saudi Arabia's Sovereign Wealth Fund in propelling its attainment of Vision 2030 goals, (2021).
[6] K. E. , Bassey. Hybrid renewable energy systems modeling. Engineering Science & Technology Journal, 4(6), 571-588, (2023).
[7] A., Al Mamun, M., Sohel, N., Mohammad, M. S. H., Sunny, D. R., Dipta, & E.Hossain. (A comprehensive review of the load forecasting techniques using single and hybrid predictive models. IEEE access, 8, 134911-134939, 2020).
[8] Kiasari, M., Ghaffari, M., & Aly, H. H. (2024). A comprehensive review of the current status of smart grid technologies for renewable energies integration and future trends: The role of machine learning and energy storage systems. Energies, 17(16), 4128.
[9] M. R., Pulicharla . Hybrid Quantum-Classical Machine Learning Models: Powering the Future of AI. Journal of Science & Technology, 4(1), 40-65, (2023).
[10] T., Zema, Kozina, A., Sulich, A., RŎmer, I., & M., Schieck. Deep learning and forecasting in practice: an alternative costs case. Procedia Computer Science, 207, 2958-2967, (2022).
[11] P., Ndikum, Machine learning algorithms for financial asset price forecasting, arXiv preprint arXiv:2004.01504, (2020).
[12] Z., Hu, Y., Zhao, M., Khushi. A survey of forex and stock price prediction using deep learning. Applied System Innovation, 4(1), 9, (2021).
[13] A. A., Mumammad. Enhancing Financial Forecasting Accuracy Through AIDriven Predictive Analytics Models, RESEARCH AND ENGINEERING JOURNALS,4(12), (2021)
[14] P., Eslamieh, M., Shajari, A., Nickabadi,. User2vec: A novel representation for the information of the social networks for stock market prediction using convolutional and recurrent neural networks. Mathematics, 11(13), 2950, (2023).
[15] S., Akhter, K. U., Eibek, S., Islam, A. R. M. T., Islam, , R., Chu, S., Shuanghe, Predicting spatiotemporal changes of channel morphology in the reach of Teesta River, Bangladesh using GIS and ARIMA modeling. Quaternary International, 513, 80-94, (2019).
[16] A. A., A., DarJain, M., Malhotra, A. R.,Farooqi, O., Albalawi,. Time Series analysis with ARIMA for historical stock data and future projections. Soft Computing, 1-12, (2024)
[17] G. Vijayalakshmi, K. Pushpanjali, and A. Mohan Babu. A comparison of ARIMA & NNAR models for production of rice in the state of Andhra Pradesh. Int J Stat Appl Math, vol. 8, no. 3, pp. 251–257, (2023).
[18] M. L. Ayala and D. L. L. Polestico. Modeling COVID-19 cases using NB-INGARCH and ARIMA models: A case study in Iligan City, Philippines. Procedia Comput. Sci., vol. 234, pp. 262–269, (2024).

[19] T. Umairah, N. Imro'ah, and N. M. Huda, Arima model verification with outlier Factors using control chart, BAREKENG J. Math. Its Appl., (18), (1), pp. 0579–0588, (2024).
[20] M. Melina, Sukono, H. Napitupulu, and N. Mohamed. Modeling of Machine Learning-Based Extreme Value Theory in Stock Investment Risk Prediction: A Systematic Literature Review. Big Data, pp. 1–20, (2024).
[21] C., Twumasi and J. Twumasi. Machine learning algorithms for forecasting and backcasting blood demand data with missing values and outliers: A study of Tema General Hospital of Ghana. Int. J. Forecast., 38(3), pp. 1258–1277, (2022).
[22] P. Więcek, D, Kubek. The impact time series selected characteristics on the fuel demand forecasting effectiveness based on autoregressive models and Markov chains. Energies, 17(16), 4163, (2024).
[23] A., Atesongun, M. A., Gulsen. Hybrid Forecasting Structure Based on Arima and Artificial Neural Network Models. Applied Sciences, 14(16), 7122, (2024).