Journal of Knowledge Management Application and Practice
*An International Journal*

http://dx.doi.org/10.18576/jkmap/070102

# Prediction of the Stock Market Using LSTM, ARIMA, and Hybrid of LSTM-ARIMA Models

*Mohanned H. Alharbi*[*]

Finance and Business Sector, Institute of Public Administration, P.O.Box 205 Riyadh 11141, Saudi Arabia.

Received: 23 April. 2024, Revised: 2 May. 2024; Accepted: 14 June. 2024
Published online: 1Jan. 2025.

**Abstract:** This paper will be presented a hybrid ARIMA-LSTM model for forecasting the stock price of Saudi Basic Industries Corporation (SABIC) on the TASI index, highlighting the growing importance of machine learning in financial market predictions. The advantages of the autoregressive integrated moving average (ARIMA) model are its ability to capture linear trends, and the advantages of the Long Short-Term Memory (LSTM) is good at modeling complex nonlinear relationships in time series data, so it is proposed to combine them in a hybrid model. The forecasting process begins with ARIMA, which identifies and addresses the linear components of the data. The residuals generated by ARIMA, which exhibit non-linear and stochastic characteristics, are then passed to the LSTM network to capture intricate patterns. By integrating both models, the ARIMA-LSTM hybrid approach is able to address both the linear and non-linear aspects of the stock data, leading to improved prediction accuracy. The anticipated outcomes are shown that the hybrid model outperforms individual models, which leads to the emphasizing its potential as a powerful forecasting tool. its potential as a powerful forecasting tool. This methodology is especially valuable in financial market analysis, where precise asset performance predictions are crucial for effective portfolio management and investment decisions.

**Keywords:** Forecasting the stock price, Recurrent Neural Network, Hybrid ARIMA LSTM, Machine learning, Artificial neural network.

## 1 Introduction

The growing need for data-driven decision-making in the rapidly evolving financial markets is becoming increasingly clear. While traditional financial models and strategies have been effective in the past, they are no longer sufficient to address the complexities of today's market environment. This paper seeks to improve stock market price prediction by combining the strengths of statistical models and machine learning (ML) techniques, ultimately enhancing portfolio optimization and risk management in financial institutions. The study is structured into five sections: Introduction, Section Two: Literature Review, Section Three: Models Used, Section Four: Results Summary, and Section Five: Conclusion.

The proposed approach demonstrates superior performance, offering both reliability and adaptability when compared to other models. By incorporating a weighted version of the distribution, the third parameter increases its flexibility, enabling it to better represent the diverse characteristics of real-world data compared to its sub-models.

## 2 Literature Review

Data analytics plays a pivotal role in understanding market dynamics, enabling informed decision-making, and refining investment strategies in the stock market [1]. Predicting stock market trends has become increasingly important for formulating effective business strategies. Numerous academic fields, including computer science, statistics, economics, finance, and operations research, have shown a great deal of interest in projecting stock values. According to recent studies, a

[*] Corresponding author E-mail: Harbimh@ipa.edu.sa

key factor in market prediction is the abundance of publicly available data, including information from social media sites and Wikipedia. Because of the complicated interactions between so many elements that affect stock prices, stock market forecasting is a very difficult and complex task. By utilizing a variety of data sources, such as historical stock market performance, Twitter activity, real-time reactions to news events, and internet search trends, the incorporation of Internet of Things (IoT) analytics has greatly increased predictive research [3]. Research has further highlighted the value of financial news in assessing stock market volatility and predicting price movements [4]. Shares of publicly traded corporations are actively exchanged on the stock market, which is a dynamic financial environment. It serves as a crucial indicator of a country's economic health, offering information on business performance and the general business environment.

[5]. This research [6] utilized a (LSTM) model to predict the closing stock prices of four leading technology companies. The model's performance was evaluated using the root mean square error (RMSE), which demonstrated that the LSTM effectively captured the stock price trends of the companies with a high degree of accuracy. This study centers on algorithms based on machine learning and deep learning techniques [7]. These algorithms have shown superior accuracy when compared to traditional regression-based models. Artificial recurrent neural networks offer a robust solution for time series forecasting. This paper introduces a price prediction approach using two built-in Python models: LSTM and ARIMA. The data for this study is sourced from the Mulkia Gulf Real Estate dataset, which has been pre-processed and utilized for training deep learning models. After training, model evaluation is performed using several metrics: MAE, MSE, RMSE, validation loss, accuracy, and $R^2$ score [8].

The study [9] also employs LSTMs, mainly using standard parameters and their adjustments, within the Libra framework. The results suggest that, due to data variability and the lack of enhanced hardware or extended time for computation, LSTMs do not outperform the median metrics of the Libra model. The LSTM neural network exhibits remarkable proficiency in collecting temporal relationships in oil rate time series data while considering production restrictions. Additionally, the Particle Swarm Optimization (PSO) technique is utilized to refine the fundamental structure of the LSTM model, hence optimizing its predictive efficacy. [10]. This study [11] This article introduces BiLSTM-MLAM, a sophisticated multi-scale time series forecasting model that utilizes bidirectional (BiLSTM) networks to extract information from both forward and backward temporal directions in time series data. Furthermore, a multi-scale patch segmentation module divides the data into several extended sequences made up of uniform segments, allowing the model to identify patterns across different time scales by dynamically modifying segment lengths.

.Another study [12] This research assesses the predictive efficacy of Long Short-Term Memory (LSTM) and eXtreme Gradient Boosting (XGBoost) models in projecting crude palm oil (CPO) production. The aim is to refine production planning, optimize inventory management, and strengthen CPO sales strategy.

[13] highlights the importance of accurate predictions to address the volatility in palm oil production and evaluates multiple forecasting algorithms. The study uses LSTM, XGBoost, and a hybrid LSTM-XGBoost model to assess backtesting performance, particularly in forecasting stock market trends. The experimental results reveal that the hybrid model significantly outperforms individual XGBoost and LSTM models in trend prediction accuracy across six international stock markets.

Furthermore, a separate study [14] This research utilized four machine learning models—Support Vector Regression (SVR), eXtreme Gradient Boosting (XGBoost), Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM)—to predict MSI 20 prices based on multivariate time series data. The findings indicate that SVR had exceptional predictive performance, with the greatest accuracy of 98.9% with negligible mistakes. Future study may concentrate on improving the efficacy of XGBoost and exploring the viability of alternative models, such as CNN-LSTM, for stock market prediction. In a related study [15], The prediction of photovoltaic (PV) power for the forthcoming 24 hours is executed by integrating a time series forecasting model (LSTM) with a regression model (XGBoost), relying exclusively on direct irradiation data. Numerous climatic elements, such as irradiance, ambient temperature, wind speed, relative humidity, solar position, and dew point, were recognized as critical parameters affecting the variability of photovoltaic output.

## 3 Utilized Models

### *3.1 The ARIMA model*

The ARIMA model, which stands for Autoregressive Integrated Moving Average, is a robust and widely adopted statistical technique used for analyzing and forecasting time series data [16]. It is especially valuable for datasets that exhibit patterns or trends over time, such as financial market data, sales forecasts, and economic indicators. The model is built on three core components, each serving a specific purpose in capturing different aspects of the time series:

**Autoregressive (AR) Component:**

The autoregressive element of the ARIMA model delineates the relationship between the present observation and its preceding values, usually known as lags. It fundamentally illustrates the impact of historical values on the present, which is especially crucial in time series research when observations are intrinsically associated throughout time. The quantity of prior periods included in the model is indicated by p, signifying the order of the autoregressive element. For example, if p = 2, the model employs the two most recent values to predict the present observation. The autoregressive framework assumes that future values can be represented as a linear combination of preceding observations.

**Integrated (I) Component:** The integrated element of the ARIMA model tackles the issue of non-stationarity in time series data. Non-stationary data displays trends or changing patterns over time, complicating the forecasting process. To convert the data into a stationary series, differencing is utilized, whereby each observation is substituted with the difference between its current and preceding value. This method effectively eradicates trends, seasonality, and other non-stationary attributes. The differencing order, indicated as d, signifies the frequency of differencing required to attain stationarity. Upon achieving stationarity, the data becomes more conducive to modeling, hence improving the precision of future forecasts.

**Moving Average (MA) Component:** The moving average part of ARIMA models the relationship between an observation and the residual errors of previous forecasts. Instead of using past observations directly, this component focuses on the forecast errors from previous periods. These errors represent the difference between the predicted and actual values and contain valuable information that can be used to adjust future predictions. The order **q** represents the number of past forecast errors that will be considered in the model. By incorporating past forecast errors, the model can correct for inaccuracies in earlier predictions and improve future forecasts.

The combination of autoregressive (AR), integrated (I), and moving average (MA) components creates a strong foundation for time series data modeling. The ARIMA model is denoted as ARIMA(p, d, q), with p signifying the order of the autoregressive component, d representing the number of differencing operations necessary for stationarity, and q indicating the order of the moving average component. The adaptability of ARIMA allows it to handle a wide array of time series features, from basic linear trends to complex patterns, rendering it an effective forecasting instrument.

The key advantage of ARIMA is its simplicity and effectiveness in modeling data without requiring external predictors. However, it assumes that the underlying data is linear and stationary, which can be a limitation for datasets with non-linear trends or seasonality. To address this, variations of ARIMA, such as SARIMA (Seasonal ARIMA) and ARIMAX (ARIMA with exogenous variables), have been developed to handle seasonal patterns and incorporate external predictors, respectively.

In practical applications, the ARIMA model is often used in conjunction with diagnostic tools like ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots to identify the appropriate values for **p**, **d**, and **q**. These tools help in selecting the optimal configuration for the model, ensuring that it captures the underlying patterns in the data and provides accurate forecasts.

Overall, ARIMA remains a cornerstone technique in time series forecasting, especially for financial and economic data where historical values play a significant role in predicting future trends.

## *3.2 LSTM*

The LSTM networks provide a sophisticated architecture inside Recurrent Neural Networks (RNNs), specifically designed to model and predict time series data, sequential patterns, and datasets with temporal relationships. [17]. In contrast to standard RNNs, LSTMs possess distinctive memory cells that enable them to capture long-term dependencies, thereby resolving the primary challenge that conventional RNNs encounter when learning from remote time steps. [18]. This capability makes LSTMs particularly well-suited for tasks involving sequential data, such as time series forecasting, natural language processing (NLP), and other applications where temporal relationships are key [19].

LSTMs have revolutionized the way sequence modeling problems are approached. Their ability to preserve information over extended periods enables deep learning models to not only retain knowledge but also effectively leverage it over long sequences, which is essential for many complex tasks that involve temporal dependencies. As a result, LSTMs have become

a powerful tool for a wide range of applications, from speech recognition to stock price prediction, where understanding the context of past events is crucial for making accurate forecasts [20].

### *3.3 A Hybrid ARIMA-LSTM Model*

The hybrid ARIMA-LSTM model combines the complementing advantages of Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks, providing a more resilient and accurate framework for time series forecasting. ARIMA, recognized for its efficacy in identifying linear patterns and trends, constitutes the model's foundation by addressing the stationary and linear elements of the time series. It methodically identifies relationships between observations and their prior values, ensuring that differencing is utilized when needed to eradicate trends and non-stationary behavior, thus improving overall forecast accuracy.

On the other hand, LSTM, a type of recurrent neural network (RNN), is specifically designed to address nonlinear dependencies and long-term temporal relationships in data. Unlike traditional RNNs, LSTMs use specialized gating mechanisms that allow the network to "remember" information over extended periods, making them particularly effective for sequential data with complex patterns and varying trends over time.

By combining these two approaches, the hybrid ARIMA-LSTM model benefits from the linear modelling capabilities of ARIMA and the nonlinear pattern recognition of LSTM. This integration allows the model to effectively capture both short-term fluctuations and long-term trends, which are common in real-world time series data, especially in domains such as financial markets, energy forecasting, and climate prediction. As a result, the hybrid model is highly adept at producing more precise and resilient forecasts, even when dealing with intricate, volatile, or noisy data [21-23].

## 4 Results Summary

Figure 1 demonstrates that SABIC experienced notable price volatility, with a downward trend throughout the previous year. A sharp decline is particularly evident during the COVID-19 pandemic, followed by a gradual recovery over time.

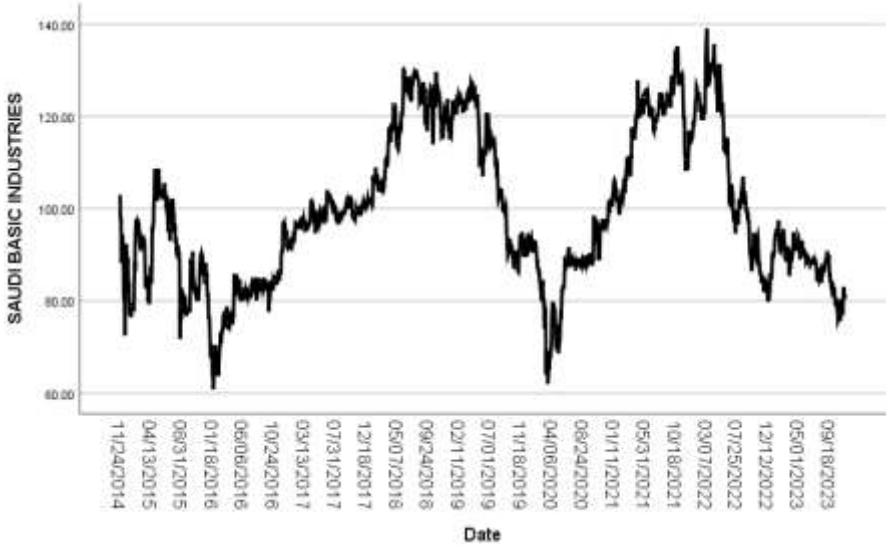


**Fig. 1.**Time series data for SABIC from 24 November 2014 to 24 November 2023.

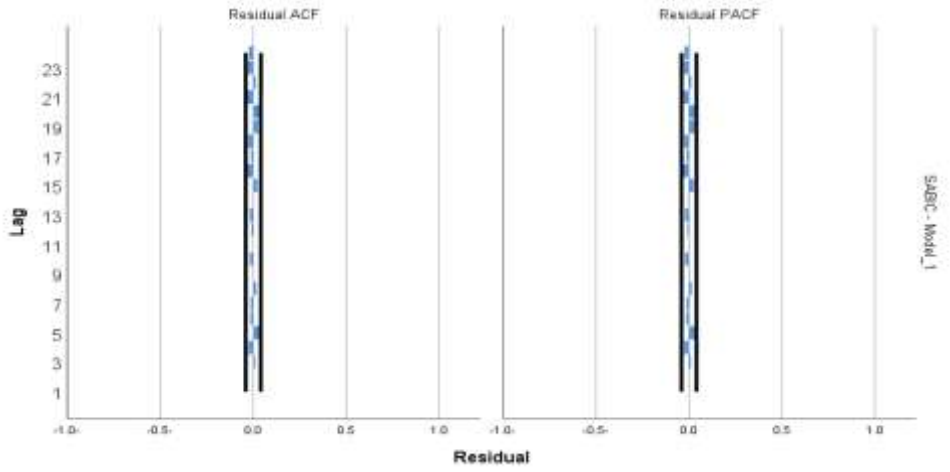


**Fig. 2:** The Residual of ACF and PACF for the SABIC ARIMA (1,1,0) Model_1**.**

Figure 2 displays the Residual Autocorrelation Function (ACF) for the SABIC Model_1, which shows the correlation between residuals at different lags along with their associated standard errors. The ACF values range between -0.035 and 0.041, indicating weak correlations across various time lags, with the highest positive autocorrelation observed at lag 5 and the most significant negative autocorrelation at lag 16. The standard error remains consistent across all lags at 0.021, suggesting uniform error estimation. Notably, any autocorrelation values beyond ±1.96 times the standard error (±0.04116) may be considered significant, indicating potential patterns in the residuals that could require further examination.

Figure 2 also presents the Residual Partial Autocorrelation Function (PACF) for the SABIC Model_1, showing the partial autocorrelations alongside their standard errors for lags 1 to 24. The PACF values, like the ACF, range from -0.035 to 0.041, with notable peaks at lag 5 (0.041), the highest positive partial autocorrelation, and at lag 16 (-0.035), the most significant negative partial autocorrelation. The standard error remains constant at 0.021 across all lags, ensuring uniform error evaluation. For partial autocorrelations to be considered significant, they must exceed the threshold of ±1.96 times the standard error (±0.04116). Most PACF values fall within this range, suggesting that the residuals are predominantly free from significant partial autocorrelations, further supporting that no major patterns are left unexplained in the model.

The PACF indicates that the model's residuals do not exhibit significant dependence across the lags. This suggests that the model has successfully captured most of the underlying dynamics, leaving only minimal unexplained autocorrelation in the residuals.

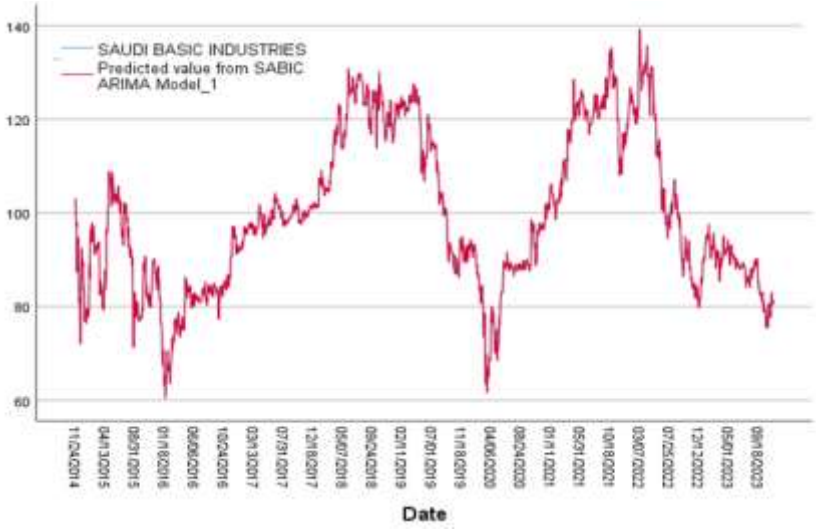


**Fig. 3.** The predicted, and the real value for SABIC.

Figure 3 demonstrates the alignment between the ARIMA (1,1,0) model and the actual data, highlighting the accuracy of the chosen modeling approach. The minimal error reinforces the model's effectiveness in capturing the underlying patterns.

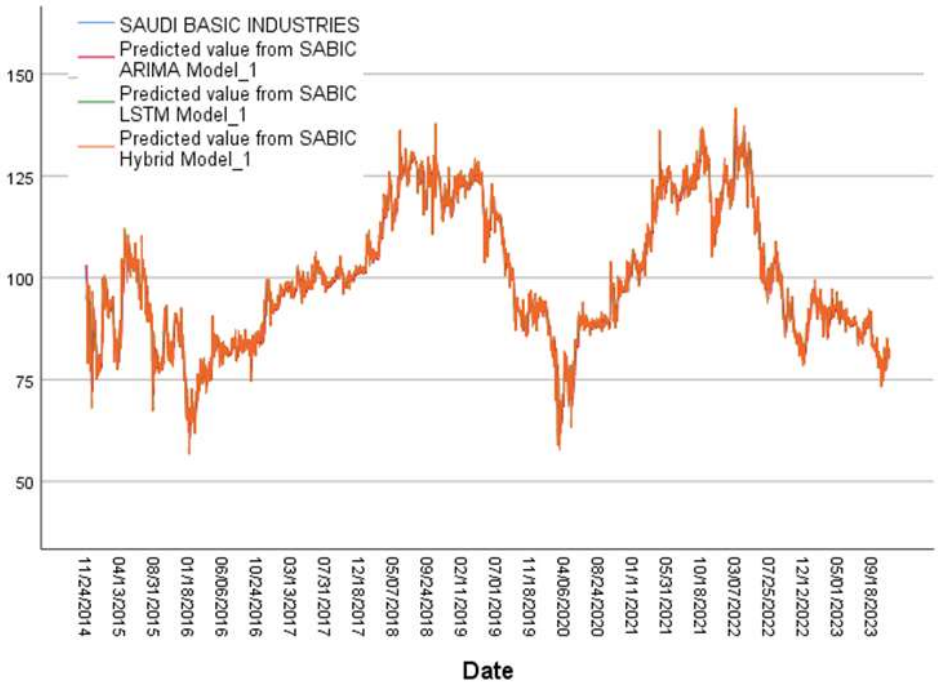


**Fig. 4.** Comparison of Real SABIC Data with LSTM, the hybrid and ARIMA Model Predictions (24 November 2014 to 24 November 2023)

Figure 4 presents a comparison between the actual data and the prediction accuracy of the LSTM, hybrid, and ARIMA models for SABIC data covering the period from November 24, 2014, to November 24, 2023. The results indicate that while all models closely track the actual data, the hybrid model's predictions are more in sync with the actual data line, showcasing its slightly better predictive performance.

**Table 1. The Criteria to Evaluate Models**

| Models | MAPE | MASE | SMAPE |
|---|---|---|---|
| Hybrid | 0.00088869 | 0.00088869 | 0.00773635 |
| LSTM | 0.01002203 | 0.01002203 | 0.01959017 |
| ARIMA | 0.01170023 | 0.02081381 | 0.02024742 |

The Hybrid model outperforms both LSTM and ARIMA across all evaluation metrics, delivering exceptional accuracy with minimal prediction errors. This superiority can be attributed to its ability to integrate the strengths of multiple modeling approaches, effectively capturing both linear and nonlinear patterns. While the LSTM model performs better than ARIMA, its error rates remain higher than those of the Hybrid model, indicating its limitations in fully capturing the dataset's complexity. On the other hand, ARIMA, although useful for simpler time-series data, shows the highest error rates, reflecting its struggle to accommodate the intricate dynamics of the dataset.

## 5 Conclusions

This study underscores the value of the three models—ARIMA, LSTM, and the hybrid model—in analyzing complex time-series data and forecasting SABIC's performance over recent years. The findings indicate that the hybrid model outperforms the others in terms of accuracy, with predictions that closely align with actual data. This is further supported by key statistical performance metrics such as MAPE, MSE, and SMAPE. While the ARIMA model proved effective in capturing the underlying data dynamics, evidenced by the absence of significant autocorrelation in the residuals, the hybrid model's

integration of both ARIMA and LSTM features significantly boosted its predictive power. This made the hybrid model particularly well-suited for handling non-linear and complex fluctuations. The study also highlights the models' ability to capture major events, such as the COVID-19 pandemic, and their capacity to model the subsequent recovery trajectory. These results offer valuable insights for decision-makers, emphasizing the importance of utilizing advanced forecasting models to improve financial performance predictions and enhance strategic planning accuracy.

Looking ahead, future research could explore the integration of other machine learning techniques, such as eXtreme Gradient Boosting (XGBoost), to further refine the predictive capabilities of stock price forecasting models. Additionally, expanding the scope of the model to include external variables, such as global market trends or geopolitical factors, could enhance its robustness and provide a more comprehensive approach to financial forecasting.

## References

[1] Y. K. Gupta and N. Sharma, N. Propositional aspect between apache spark and hadoop map-reduce for stock market data. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) IEEE., 479-483 (2020).

[2] B.Aasi, , et al. Stock price prediction using a multivariate multistep LSTM: a sentiment and public engagement analysis model. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-8). (2021).

[3] S. Siami-Namini, Set al. The performance of LSTM and BiLSTM in forecasting time series. In 2019 IEEE International conference on big data (Big Data) (pp. 3285-3292). (2019).

[4] A., Gheondea- Eladi, Patient decision aids: a content analysis based on a decision tree structure. BMC medical. A., Gheondea- Eladi, Patient decision aids: a content analysis based on a decision tree structure. BMC medical informatics and decision making., **19**, 1-15 (2019).

[5] B. Gülmez., Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm. Expert Systems with Applications., **227**, 120346. (2023).

[6] Z. Li, et al. Stock market analysis and prediction using LSTM: A case study on technology stocks. Innovations in Applied Engineering and Technology., 1-6 (2023).

[7] S. Siami-Namini, N.,Tavakoli, & A. S. Namin, The performance of LSTM and BiLSTM in forecasting time series. In 2019 IEEE International conference on big (2019).

[8] Albeladi, K., Zafar, B., & Mueen, A. (2023). Time Series Forecasting using LSTM and ARIMA. International Journal of Advanced Computer Science and Applications., **14(1)**, 313-320.

[9] R. Prater, et al. Generalized Performance of LSTM in Time-Series Forecasting. Applied Artificial Intelligence, 38(1), 2377510. (2024).

[10] X., ong, et al Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. Journal of Petroleum Science and Engineering., **186**, 106682, (2020).

[11] Y. Fan, et al. BiLSTM-MLAM: A Multi-Scale Time Series Prediction Model for Sensor Data Based on Bi-LSTM and Local Attention Mechanisms. Sensors, 24(12), 3962. (2024).

[12] K. Aqbar, and R.A. Supomo. Performance analysis of lstm and xgboost models optimization in forecasting crude palm oil (cpo) production at palm oil mill (pom). International Journal of Computer Applications., **975**, 8887 (2023).

[13] H. Oukhouya, et al. Forecasting International Stock Market Trends: XGBoost, LSTM, LSTM-XGBoost, and Backtesting XGBoost Models. Statistics, Optimization & Information Computing., **12(1)**, 200-209 (2024).

[14] H. Oukhouya, and K. El Himdi. Comparing machine learning methods—svr, xgboost, lstm, and mlp—for forecasting the moroccan stock market. In Computer Sciences & Mathematics Forum (Vol. 7, No. 1, p. 39). (2023).

[15] K. B. A Didavi, et al. LSTM and XGBoost Models for 24-hour Ahead Forecast of PV Power from Direct Irradiation. Renewable Energy Research and Applications., **5(2)**, 229-241 (2024).

[16] A. L. Schaffer, T. A. Dobbins & S. A. Pearson. Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. BMC medical research methodology., **21**, 1-12 (2021).

[17] L.; Benos, et al Machine Learning in Agriculture: A Comprehensive Updated Review. Sensors., **21**, 3758 (2021).

[.18] G., Sonkavde, et al. Forecasting stock market prices using machine learning and deep learning models: a systematic review, performance analysis and discussion of implications. International Journal of Financial Studies., **11(3)**, 94 (2023).

[19] D. Radojičićet al. A comparative study of the neural network models for the stock market data classification−A multicriteria optimization approach. Expert Systems with Applications., **238**, 122287 (2024).

[.20] R. Zhang, LSTM-based Stock Prediction Modeling and Analysis. In 2022 7th International Conference on Financial Innovation and Economic Development (pp. 2537-2542). (2022).

[21] B. Charbuty, & A. Abdulazeez, Classification based on decision tree algorithm for machine l earning. Journal of Applied Science and Technology Trends., **2(01)**, 20-28 (2021).

[22] M., Christofi, et al. Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. The International Journal of Human Resource Management., 33(6), 1237-1266 (2022).

[23] M. M Taye, Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future. Directions. Computers., **12(5)**, 91(2023).