A Residual-Based Hybrid Approach Combining ARMA/ARIMA with Random Forest and GBM for ASELSAN Stock Price Forecasting

Yusuf Emre BAYSAL
Electronics Engineering Department
İstanbul Technical University
İstanbul/Turkey
baysal24@itu.edu.tr

***Abstract*— This study evaluates ARMA, ARIMA, Random Forest, and GBM models, along with their hybrid combinations using residuals, for forecasting ASELSAN stock prices especially focuses on hybrid models and its effects on forecasting future values. Study shows that differenced data significantly improved machine learning model accuracy. While ARMA-based hybrids reduced performance, ARIMA-based hybrids enhanced it. The best results were obtained from Random Forest and GBM models trained on differenced data. These findings highlight the potential of integrating statistical and machine learning methods to improve stock price prediction.**

Keywords—ARMA, ARIMA, Random Forest, GBM, Hybrid Model, Stock Prediction, Hybrid Stock Prediction

I. INTRODUCTION

Accurate stock price prediction is a major challenge in today's volatile financial markets, essential for maximizing profits and mitigating risks. Traditional methods like ARMA and ARIMA effectively capture linear patterns in historical data but are limited by the market's inherent unpredictability, influenced by factors such as market sentiment, political events, and economic indicators.

Machine learning techniques, including Random Forest and Gradient Boosting Machines (GBM), have recently shown superior performance in modeling complex, non-linear relationships in financial data. However, combining statistical and machine learning approaches remains underexplored.

This paper presents four hybrid models integrating ARMA and ARIMA with Random Forest and GBM to enhance stock price prediction. Using MATLAB, we trained these models on Aselsan's stock price data from January 2023 to November 2024, leveraging both statistical robustness and machine learning's predictive power. The contributions of this research are threefold:

- **Development of Hybrid Models**: Combining ARMA/ARIMA with Random Forest and GBM.
- **Implementation in MATLAB**: Providing a comprehensive codebase for model training and prediction.
- **Empirical Validation**: Demonstrating improved accuracy over individual models using Aselsan's historical data.

The remainder of this paper is organized as follows: Section II provides the theoretical background, Section III outlines the methodology, Section IV presents simulations and results, and Section V concludes with contributions and future research suggestions.

II. THEORETICAL BACKGROUND

Accurate stock price prediction leverages both statistical and machine learning (ML) models to capture complex patterns in financial data. This research explores four distinct hybrid models by integrating ARMA, ARIMA with Random Forest (RF) and Gradient Boosting Machines (GBM) to enhance prediction accuracy.

A. Statistical Models

1) ARMA (AutoRegressive Moving Average):

ARMA models combine autoregressive (AR) and moving average (MA) components to capture linear relationships in stationary time-series data, such as historical stock prices. AR captures dependencies on past values, while MA accounts for forecast errors. ARMA is effective for stationary data but struggles with non-stationary series, which can be addressed using differencing or transformations.

2) ARIMA (AutoRegressive Integrated Moving Average):

ARIMA extends ARMA by incorporating differencing to handle non-stationary data, making it suitable for a broader range of financial time series. The "Integrated" component achieves stationarity through differencing. While ARIMA can model trends and seasonality, it primarily captures linear relationships and may not fully address the complexities of stock price movements.

B. Machine Learning Models

1) Random Forest (RF):

Random Forest is an ensemble method that builds multiple decision trees and averages their predictions. It effectively captures non-linear relationships and interactions, enhancing accuracy and controlling overfitting. In financial forecasting, RF robustly identifies diverse stock price patterns.

2) Gradient Boosting Machines (GBM):

GBM sequentially builds trees, where each new tree corrects the errors of the previous ones by optimizing a loss function. This method captures complex patterns with high accuracy and is effective in modeling intricate relationships in stock data. However, GBM can be computationally intensive and sensitive to hyperparameter settings.

C. Hybrid Approaches and Residual Method

Hybrid models combine statistical and ML techniques to leverage their complementary strengths. In this study, four hybrid models are developed:

- ARMA + Random Forest (RF)
- ARMA + Gradient Boosting Machines (GBM)
- ARIMA + Random Forest (RF)
- ARIMA + Gradient Boosting Machines (GBM)

1) Residual Method:

Each hybrid model employs the residual method, where a statistical model (ARMA or ARIMA) first captures the linear components of the stock price. Subsequently, an ML model (RF or GBM) predicts the residuals, addressing the non-linear patterns not captured by the statistical model. This two-step process enhances overall prediction performance by effectively modeling different aspects of the data.

III. Methodology

This section outlines the methods and procedures employed in this research to develop and evaluate hybrid models for stock price prediction. The methodology encompasses data collection, preprocessing, model development, implementation, and training and validation processes.

A. Data Collection

Aselsan's daily opening stock prices from January 2, 2023, to December 6, 2024, were obtained from Investing.com. The dataset was divided into training and testing sets, with the last 10 days reserved as the test set to evaluate the models' predictive performance.

Figure 1 ASELSAN Opening Stock Prices

B. Data Preprocessing

To prepare the data for modeling, the following preprocessing steps were undertaken:

1) Stationarity Check

For ARMA and ARIMA models, it was crucial to ensure the data's stationarity. An Augmented Dickey-Fuller (ADF) test was conducted, yielding a p-value of 0.9297, indicating non-stationarity. Consequently, differencing was applied to achieve stationarity, resulting in a p-value of 0.0010.

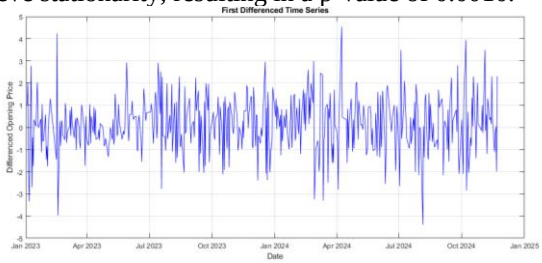

Figure 2 Difference Taken Stock Prices

2) Visualization

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were generated to identify appropriate lag parameters for ARMA and ARIMA models.

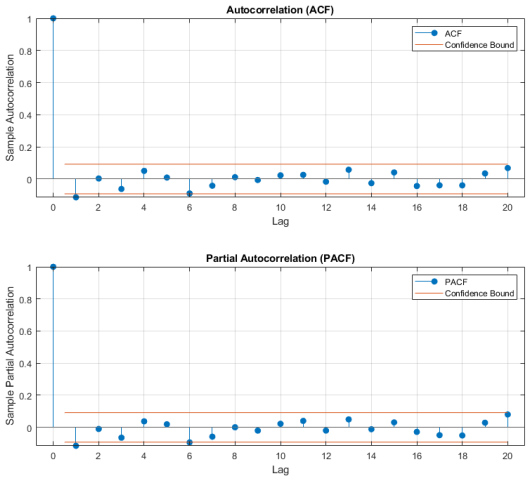

Figure 3 ACF and PACF Graphs of Difference Taken Time Series

3) Train-Test Split

The dataset was divided into a training set (January 2, 2023, to November 22, 2024) and a test set (November 25, 2024, to December 6, 2024). Despite focusing on the first 5 days of predictions for analysis due to error amplification in longer forecasts, the entire 10-day period was retained for comprehensive evaluation.

C. Model Development

Four distinct hybrid models were developed by integrating statistical and machine learning techniques:

- ARMA + Random Forest (RF)
- ARMA + Gradient Boosting Machines (GBM)
- ARIMA + Random Forest (RF)
- ARIMA + Gradient Boosting Machines (GBM)

1) Residual Method

Each hybrid model employed the residual method, which involves two steps:

Step 1: A statistical model (ARMA or ARIMA) is first applied to capture the linear components of the stock price.

Step 2: The residuals (errors) from the statistical model are then predicted using an ML model (RF or GBM) to address non-linear patterns. This approach allows the ML model to focus on the discrepancies left by the statistical model, thereby enhancing overall prediction accuracy.

2) Parameter Selection:

ARMA Model: The Akaike Information Criterion (AIC) test was used to determine the optimal parameters. For the ARMA model, the best fit was found with p=5 and q=4.

ARIMA Model: Similarly, the AIC test identified the optimal parameters for the ARIMA model as p=5, d=1, and q=5.

ML Models: For Random Forest and GBM, hyperparameters such as window size, tree depth, and the number of boosting rounds were optimized using a trial-and-error approach to achieve the best performance.

For Random Forest simulations, Monte Carlo simulations with 100 runs are conducted, since the effect of randomness

algorithm inside the Random Forest method varies the results with high variance.

D. Implementation

The models were implemented using MATLAB, leveraging its robust toolboxes for statistical analysis and machine learning. The MATLAB environment was set up with the Statistics and Machine Learning Toolbox and Econometrics Toolbox, to facilitate model training and evaluation. The code structure was organized into scripts for data preprocessing, model training, prediction, and evaluation to ensure clarity and reproducibility.

E. Training and Validation

1) Training Process:

Statistical Models: ARMA and ARIMA models were trained on stationary and non-stationary datasets, respectively. After training, these models generated predictions for the test set.

ML Models: Random Forest and GBM models were independently trained on both stationary and non-stationary datasets.

Iterative Prediction: For Random Forest, an iterative approach was used, making one-day-ahead predictions and appending them to the dataset for subsequent forecasts. This process was repeated over 10 days, with detailed analysis of the first 5 days due to compounding errors in longer forecasts. In contrast, GBM generated a 10-day bulk forecast directly after training on the dataset, providing a comprehensive assessment of the model's performance over the test period.

Validation Metrics: Mean Squared Error (MSE) was used to evaluate the prediction accuracy of each model by comparing the predicted values against the actual stock prices. The MSE provided a quantitative measure of the models' performance, facilitating a comparative analysis between individual and hybrid models.

IV. Simulations and Results

Throughout the study, the following simulations were conducted and are discussed in the results section.

Raw values represent the company's actual opening stock prices in the stock market, while differenced values indicate the changes between consecutive prices.

- ARMA (differenced values)
- ARIMA (raw values)
- GBM (differenced values)
- GBM (raw values)
- Random Forest (differenced values)
- Random Forest (raw values)
- ARMA + Random Forest (differenced values)
- ARMA + GBM (differenced values)
- ARIMA + Random Forest (raw values)
- ARIMA + GBM (raw values)

Although the test set consists of the last 10 days of data, comparisons were made specifically for the first 5 days following the training dataset to facilitate accurate evaluation. This approach was chosen because prediction errors tend to amplify over longer forecasting horizons.

For instance, when the Hybrid (ARIMA + GBM) model was trained using the differenced values from the training set, it generated predictions for both the 5-day and 10-day periods.

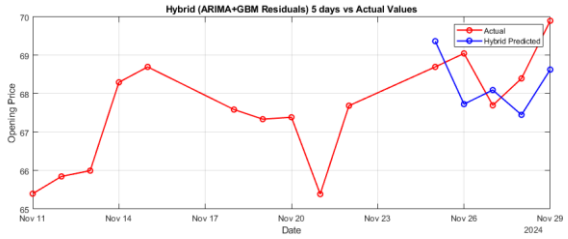

Figure 4 Hybrid (Arima + GBM) 5 days Model Prediction

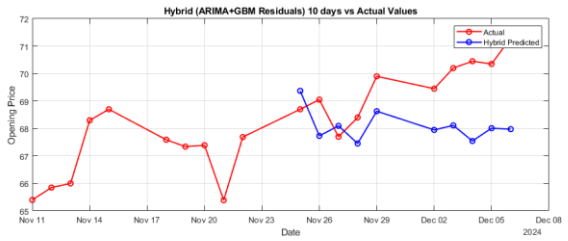

Figure 5 Hybrid (Arima + GBM) 10 days Model Prediction

The table below presents the 5-day and 10-day results for the 10 different model trainings examined in this study.

Model	MSE 5 Days	MSE 10 Days
ARMA (differenced values)	1.1278	2.3722
ARIMA (raw values)	1.3508	3.2725
GBM (differenced values)	0.6287	0.7532
GBM (raw values)	2.0368	4.8891
Random Forest (differenced values)	1.0075*	2.6454*
Random Forest (raw values)	1.5583*	5.3740*
ARMA + Random Forest (differenced values)	1.2389	3.2398
ARMA + GBM (differenced values)	2.1520	6.1314
ARIMA + Random Forest (raw values)	1.1095	3.5692
ARIMA + GBM (raw values)	0.9710	3.6428

*Table 1: 5 and 10 Days MSE Results, *Monte Carlo averages with 100 runs,*

As evident from the table above, the performance of all models in making long-term predictions progressively decreases. The remainder of the study will focus solely on the 5-day forecast graphs, MSE results, and their interpretations. Additionally, the 5-day MSE prediction values presented in this table serve as a summary of the entire study.

A. Model Studies and Model Parameters

1) ARMA Model

For the ARMA model, the procedures outlined in Section III. Methodology was followed. The data was made stationary, and using the AIC test, the optimal ARMA parameters were determined to be (5, 4), after which the model was trained. The results were visualized, and the MSE was calculated as 1.1278.

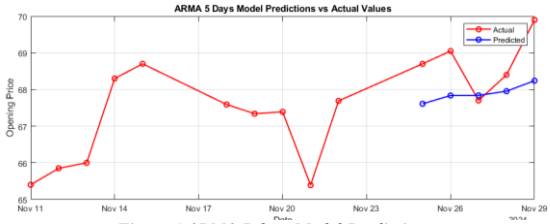

Figure 6 ARMA 5 days Model Prediction

2) *ARIMA Model*

For the ARIMA model, the procedures outlined in Section III. Methodology were followed. The differencing parameter d was set to 1, and using the AIC test, the optimal ARIMA parameters were determined to be (5, 1, 5). Subsequently, the model was trained, the results were visualized, and the MSE was calculated to be 1.3508.

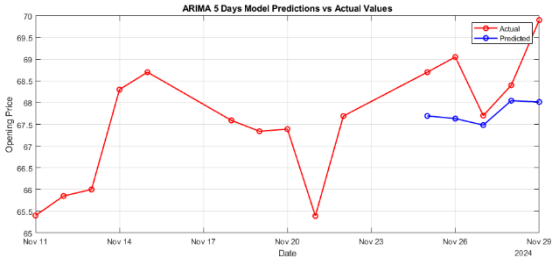

Figure 7 ARIMA 5 days Model Prediction

3) *Random Forest Model with Differenced Values*

For the Random Forest model, the methodology outlined in Section III was implemented. The data was made stationary and the model was trained iteratively to generate daily predictions. A window size of 60 and 50 trees were selected by evaluating the MSE for each parameter combination against the test dataset to identify the optimal settings. Due to the inherent randomness in the Random Forest algorithm, simulation runs can produce highly varied results. In the simulation presented in the figure below, the model achieved an MSE of 0.6039 while Monte Carlo average with 100 runs is 1.0075.

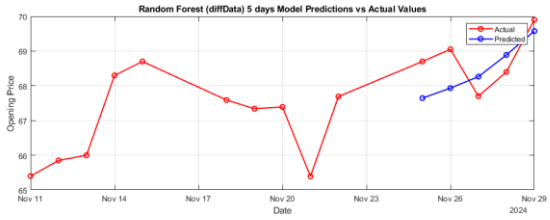

Figure 8 Random Forest 5 days Model Prediction with Differenced Values

4) *Random Forest Model with Raw Values*

For the Random Forest model trained on raw stock prices, the methodology outlined in Section III was implemented to forecast future stock prices. A window size of 45 and 10 trees were selected by evaluating the MSE with same technique explained above. Again, due to the inherent randomness in RF, simulation runs can produce varied results. In the simulation presented in the figure below, the model achieved an MSE of 0.7002, while Monte Carlo average with 100 runs is 1.5583, which is worse than differenced values simulation runs.

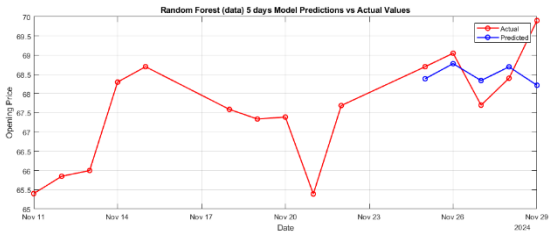

Figure 9 Random Forest 5 days Model Prediction with Raw Values

5) *GBM Model with Differenced Values*

For the GBM model with differenced stock prices, the methodology outlined in Section III was implemented. The data was made stationary and future stock prices were forecasted. A window size of 15 and 50 boosting stages were selected by evaluating the MSE for each parameter. In the simulation presented in the figure below, the model achieved an MSE of 0.6287. The MSE was very close to that of the Random Forest model using differenced values, while the prediction graph demonstrated far better fit compared to Random Forest based on visual evaluation.

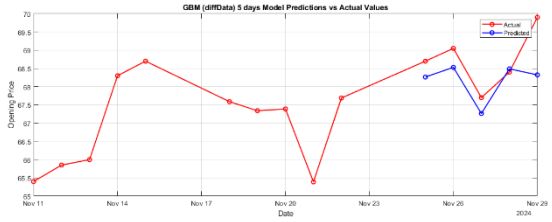

Figure 10 GBM 5 days Model Prediction with Differenced Values

6) *GBM Model with Raw Values*

For the GBM model trained on raw stock prices, the methodology outlined in Section III was implemented. A window size of 29 and 135 boosting stages were selected by evaluating the MSE. In the simulation presented in the figure below, the model achieved an MSE of 2.0368. This MSE is considerably higher compared to the results obtained using differenced values. A similar trend was observed with the Random Forest model, indicating that using differenced values for model training significantly enhances prediction accuracy.

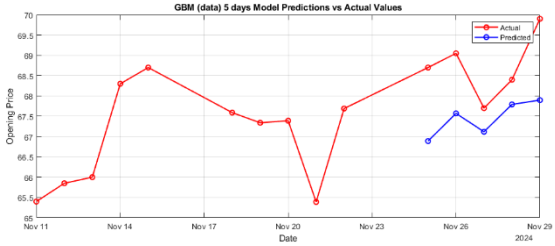

Figure 11 GBM 5 days Model Prediction with Raw Values

7) *Hybrid Model with ARMA + Random Forest*

For the hybrid models, this study combined the ARMA and Random Forest models using the residuals method as explained in Section III. The data was made stationary and future stock prices were forecasted. Utilizing the AIC test, the optimal ARMA parameters were determined to be (5, 4). A window size of 60 and 100 trees were selected. In the simulation presented in the figure below, the hybrid model achieved an MSE of 1.2389. This result is slightly worse than the performance of the ARMA model alone, which is

considered as bad since the aim of the hybrid model is to increase the performance.

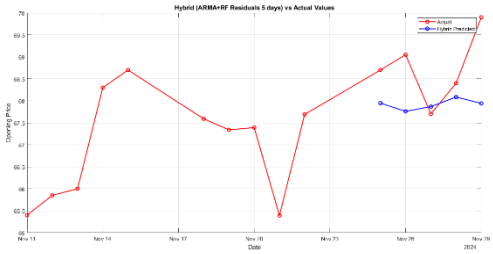

Figure 12 ARMA + Random Forest 5 days Model Prediction

8) Hybrid Model with ARMA + GBM

By combining ARMA and GBM, the residuals method employed. The data was made stationary by applying differencing to accommodate the ARMA model, and future stock prices were forecasted. Utilizing the AIC test, the optimal ARMA parameters were determined to be (5, 4). A window size of 200 and 55 boosting stages were selected by evaluating the MSE. Simulation presented in the figure below, the hybrid model achieved an MSE of 2.1520. This result is worse than the performance of the ARMA model alone and the GBM model trained on differenced values alone.

However, the trend captured in the graph below successfully follows the ups and downs of the stock prices, indicating that the model effectively tracked the fluctuations. Despite this, the overall results were negatively impacted by offset differences. For future work, incorporating offset estimation into a more optimized system could lead to significantly better prediction outcomes.

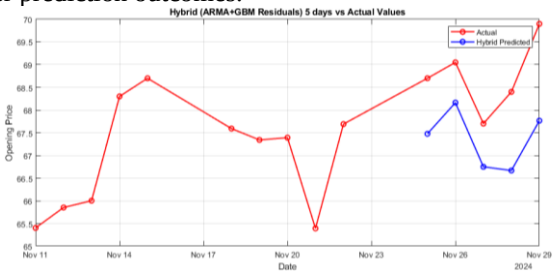

Figure 13 ARMA + GBM 5 days Model Prediction

9) Hybrid Model with ARIMA + Random Forest

By combining ARIMA and Random Forest, the residuals method was employed to forecast future stock prices using raw stock data. The differencing parameter d was set to 1, and the optimal ARIMA parameters were determined to be (5, 1, 5) based on the AIC test. A window size of 60 and 100 trees were selected by evaluating the MSE. In the simulation presented in the figure below, the hybrid model achieved an MSE of 1.1095. This result slightly outperforms the performance of the ARIMA model alone and the Random Forest model trained on raw values alone. The results indicate that the residuals method slightly improved the model's efficiency.

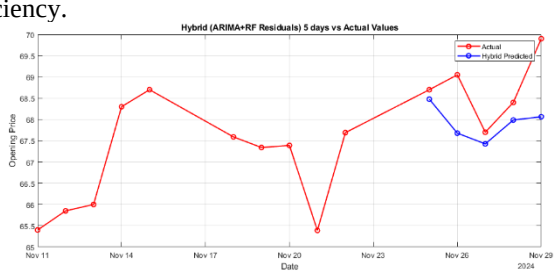

Figure 14 ARIMA + Random Forest 5 days Model Prediction

10) Hybrid Model with ARIMA + GBM

Hybrid model with ARIMA and GBM, the residuals method was employed to forecast future prices using raw stock data. The differencing parameter d was set to 1, and the optimal ARIMA parameters were determined to be (5, 1, 5) using the AIC test. A window size of 200 and 55 boosting stages were selected by evaluating the MSE. In the simulation presented in the figure below, the hybrid model achieved an MSE of 0.9710.

This result outperforms both the ARIMA model alone and the GBM model trained on differenced values alone. The findings indicate that the residuals method effectively enhanced the model's efficiency, demonstrating improved prediction accuracy through the integration of statistical and machine learning techniques.

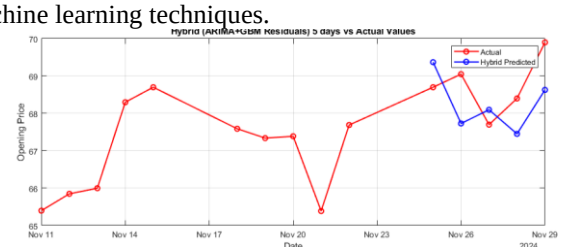

Figure 15 ARIMA + GBM 5 days Model Prediction

V. Conclusion

This study investigated the effectiveness of ARMA, ARIMA, Random Forest, and Gradient Boosting Machines (GBM) models, along with their various hybrid combinations using the residuals method, in forecasting future stock prices of Aselsan. Table 1 provides a comprehensive summary of the 10 different model trainings, showcasing the Mean Squared Error (MSE) for 5-day and 10-day predictions. Detailed results, including graphical representations, are presented in Section IV.

A. Key Findings:

1) Individual Models:

ARMA and ARIMA Models: When used independently, ARMA and ARIMA models produced closely similar results, indicating their comparable effectiveness in capturing linear patterns in stock price data.

Random Forest and GBM Models: Similarly, Random Forest and GBM models yielded comparable outcomes when used alone. Notably, the performance of these machine learning models significantly improved when trained on differenced values rather than raw stock prices. Visual trend analysis revealed that the GBM model provided more rawistic trend tracking compared to Random Forest.

2) Hybrid Models:

ARMA-Based Hybrids: Combining ARMA with Random Forest and GBM resulted in decreased performance,

which contradicts the objectives of enhancing prediction accuracy through hybridization.

ARIMA-Based Hybrids: In contrast, integrating ARIMA with Random Forest and GBM improved the models' performance compared to using ARIMA alone. The hybrid models effectively captured stock price trends, as evidenced by the graphical results.

Overall Performance: The best predictive performance was achieved by Random Forest and GBM models trained on differenced values alone, outperforming both their individual and hybrid counterparts.

B. Future Work:

The machine learning models in this study were solely trained and tested on Aselsan's stock price data. In raw-world scenarios, where future data is unknown and test datasets are unavailable, parameter optimization becomes challenging. To address this limitation, future research should focus on optimizing machine learning models with more comprehensive and diverse datasets. Additionally, incorporating various financial indicators such as trading volume, Moving Average Convergence Divergence (MACD), and Relative Strength Index (RSI) could enhance the accuracy of machine learning models by capturing a broader range of market dynamics.

Exploring deep learning models may also offer superior performance compared to traditional machine learning approaches due to their ability to model complex, non-linear relationships in data. Furthermore, it is essential to tailor prediction models to specific stocks and focus on short-term forecasts. Models should be continuously updated with new market data to maintain their relevance and accuracy, rather than being developed as generic stock prediction tools.

In conclusion, while hybrid models using the residuals method showed mixed results, this study highlights the potential of machine learning models, particularly Random Forest and GBM, when trained on appropriately preprocessed data. Future enhancements in data integration and model optimization are expected to further improve prediction accuracy and reliability in stock price forecasting.

REFERENCES

[1] A. Yadav, V. Kumar, S. Singh and A. K. Mishra, "A Survey on Stock Price Prediction using Machine Learning Techniques", *2023 Winter Summit on Smart Computing and Networks (WiSSCoN*), pp. 1-12, 2023, March.

[2] Huanze Tang, Stock Prices Prediction Based on ARMA Model, 2021 *International Conference on Computer, Blockchain and Financial Development* (CBFD), DOI: 10.1109/CBFD52659.2021.00046.

[3] Guo, D.-H., Tu, T.-H., & Zhang, Z.-D. (2018). Stock Price Correlation Coefficient Prediction with ARIMA-LSTM Hybrid Model. Retrieved from https://arxiv.org/abs/1808.01560

[4] T. Manojlovic and I. Stajduhar, "Predicting stock market trends using random forests: A sample of the Zagreb stock exchange", *2015 38th International Convention on Information and Communication Technology Electronics and MChoi, H.K., 2018. Stock price correlation coefficient prediction with*