Stock Price Prediction using Time Series Analysis

Nilay Nisheethkumar Patel
Department of Computer Science
Georgia State University
Atlanta, Georgia, USA
npatel262@student.gsu.edu

Ashwani B Neminimadathil
Department of Computer Science
Georgia State University
Atlanta, Georgia, USA
aneminimadathil@student.gs.edu

Abstract* — Accurate forecasting of stock prices remains a paramount challenge in financial analytics due to the intricate complexities and dynamic nature of the stock market. This paper presents a comprehensive investigation into stock price forecasting specifically for Apple Inc., one of the world's most valuable and influential companies. By leveraging a diverse array of advanced statistical and machine learning techniques, we aim to develop robust predictive models capable of capturing the nuances and seasonal patterns inherent in Apple's stock data [8]. Our methodology employs a multi-pronged approach, evaluating individual models as well as hybrid ensemble strategies. We harness the predictive power of Long Short-Term Memory Networks (LSTM), a state-of-the-art deep learning architecture adept at modeling long-range dependencies in sequential data, making it well-suited for the volatile fluctuations of stock prices. Simultaneously, we explore traditional time series forecasting methods, including the Autoregressive Integrated Moving Average (ARIMA) and its extension, Seasonal ARIMA (SARIMA), which explicitly accounts for seasonal components present in the data. Furthermore, we investigate the applicability of Fuzzy Time Series models, which leverage the principles of fuzzy set theory to handle the inherent uncertainty and imprecision in financial time series data. By incorporating this diverse set of models, we aim to exploit their unique strengths and capture different aspects of the data's complexity. In addition to evaluating these individual models, we explore the potential of ensemble approaches that combine the forecasts from ARIMA, SARIMA, and LSTM models. Through this hybrid strategy, we seek to leverage the complementary strengths of each technique while mitigating their respective weaknesses, potentially enhancing the overall predictive accuracy and robustness of our forecasting framework. This research aims to find the best way to predict Apple's stock price. By doing this, we can give investors, traders, and analysts valuable information to make better decisions in the complex and constantly changing stock market. This will contribute to the ongoing research on how to best predict financial markets.

***Keywords* — Apple Stock Forecasting, LSTM, ARIMA, SARIMA, Fuzzy Time Series, Ensemble models, trends, seasonality, predictive accuracy, financial analytics, time series forecasting.**

I. INTRODUCTION

Accurately forecasting stock prices is a cornerstone of financial analysis, enabling informed decision-making for investors and institutions in the dynamic and complex financial market. As stock market exhibits inherent volatility, non-linearity, and is susceptible to various economic, political, and psychological factors, traditional statistical methods have achieved some success to model and forecast stock prices; however, their limitations in capturing non-linear relationships and complex sequential patterns hinder their effectiveness. This necessitates the exploration of advanced machine learning techniques for improved forecasting accuracy.

Time series analysis (TSA) constitutes a robust statistical framework for deciphering the complexities inherent in temporally ordered data. TSA delves into the dynamic nature of these sequences, uncovering latent patterns, trends, and interrelationships that offer valuable insights across diverse domains.[6].

This paper presents a comprehensive investigation comparing the performance of several prominent forecasting models. We explore the efficacy of Long Short-Term Memory Networks (LSTMs), a powerful deep learning architecture adept at modeling sequential data and uncovering hidden dependencies within time series[1]. Additionally, we examine established statistical methods like Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) for their ability to model trends and seasonality present in financial data[6]. Furthermore, we delve into Fuzzy Time Series models, which leverage fuzzy set theory to incorporate inherent uncertainties prevalent in financial markets[9]. Recognizing the potential benefits of leveraging the strengths of each individual approach, we propose an ensemble learning strategy that combines the predictive power of ARIMA, SARIMA, and LSTMs[4].

Through rigorous empirical evaluation on real-world stock data, we comparatively assess the performance of these models, both individually and in ensemble configurations, using multiple error metrics, including Root Mean Squared Error (RMSE). This comprehensive analysis not only provides insights into the relative efficacy of these techniques but also highlights the most promising approach for capturing the seasonal patterns prevalent in stock market dynamics.

II. LITERATURE REVIEW

Accurately forecasting stock prices remains a challenge, despite the success of traditional methods like ARIMA/SARIMA in capturing trends and seasonality. However, their limitations in handling non-linear relationships and complex patterns hinder effectiveness [6]. LSTMs, a powerful deep learning architecture, demonstrate ability in capturing long-term dependencies and non-linearities in sequential data like stock prices. Studies comparing LSTMs with traditional methods report superior accuracy for LSTMs [10]. Hybrid approaches combining LSTMs with ARIMA show further improvement [12]. Fuzzy Time Series models address this using fuzzy set theory, potentially improving forecasting robustness [8].

While prior research has explored the potential of LSTM, ARIMA/SARIMA, and Fuzzy Time Series models individually for stock price prediction, a comprehensive investigation comparing their performance and the potential benefits of ensemble learning approaches is lacking. This paper addresses this gap by implementing and evaluating these individual models, investigating an ensemble learning strategy that combines their strengths, and conducting a rigorous comparative analysis to identify the most effective approach for accurate stock price forecasting.

III. PROPOSED METHOD

This paper presents a multifaceted methodology for stock price forecasting that leverages both traditional time series analysis techniques and advanced machine learning models.

The initial stage involves meticulous data collection. We gather historical stock price data for Apple Inc., encompassing opening, high, low, closing, adjusted closing prices, and trading volume. To ensure data quality, rigorous preprocessing is undertaken. Techniques such as mean/median imputation address missing values, while outlier detection and removal are conducted. Furthermore, normalization ensures consistency in the data scale. Notably, for ARIMA and SARIMA models, log differencing is applied to achieve stationarity, a prerequisite for these methods to achieve optimal performance. Finally, the preprocessed data is indexed by date for seamless integration into the subsequent modeling pipelines.

Our model selection strategy employs a diverse set of techniques to capture the inherent complexities of stock market dynamics. Long Short-Term Memory Networks (LSTMs) are adept at modeling long-term dependencies and non-linearities prevalent in sequential data. LSTMs, potentially employing a multi-layered structure with a specified number of layers and neurons, are well-suited for capturing the volatile patterns often observed in stock prices.Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) are employed to capture linear relationships and seasonal components within the data, complementing the capabilities of LSTMs [9, 10]. We further explored the application of Fuzzy Time Series models, which harness the principles of fuzzy set theory to manage the inherent uncertainty and imprecision inherent in financial time series data. This multi-model approach aims to exploit the unique strengths of each technique, potentially enhancing the overall predictive accuracy and robustness of our forecasting framework.

The training and evaluation of these models are conducted using a split of the data into training and validation sets. We employ a range of performance metrics, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Root Mean Squared Error (RMSE), to comprehensively assess each model's effectiveness in capturing the underlying patterns within the data.

Moreover, we investigated the potential of ensemble modeling techniques that combine the forecasts from individual models, aiming to leverage their complementary strengths while mitigating their respective weaknesses. This ensemble approach potentially yields more accurate and reliable predictions by integrating the diverse predictive capabilities of the constituent models.

Our methodology culminates in a rigorous comparative analysis of the individual and ensemble models' performance, using the established metrics. This comprehensive evaluation enabled us to identify the most effective model(s) for accurate stock price forecasting, providing valuable insights to support informed decision-making for investors and traders operating in the dynamic financial markets.

IV. DATA DESCRIPTION

A. Data Collection

The dataset utilized in this study comprises historical stock price data for Apple Inc. (AAPL), sourced from Kaggle, a platform renowned for its diverse collection of datasets used in academic and research settings. The data spans a specific period, providing a detailed view of Apple's stock performance over time. This timeframe includes both regular market operations and significant market events, offering a robust foundation for analyzing patterns and forecasting future price movements.

B. Data Features

The dataset includes the following key features that are typical in stock market analysis:

- **Open**: Price of stock at the beginning of trading day.
- **High**: Highest price of stock during the trading day.
- **Low**: Lowest price of stock during the trading day.
- **Close**: Price of stock at the closing of the trading day.
- **Adjusted Close**: Stock's closing price after accounting for dividends, stock splits, and other adjustments.
- **Volume**: No. of shares traded during the trading day.

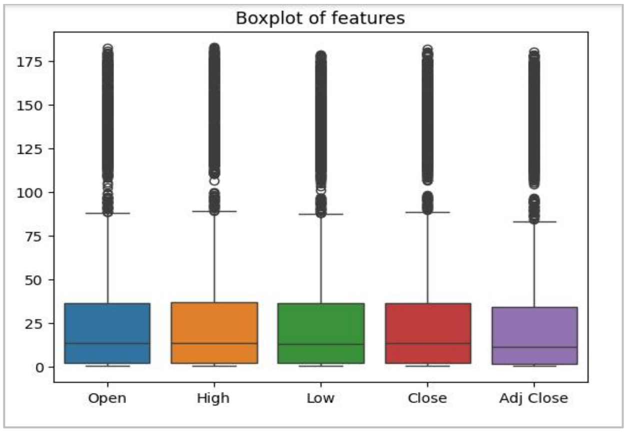
Leveraging these foundational features, predictive models are deployed to forecast the opening price for the subsequent trading day, integrating a holistic perspective of market dynamics for informed decision-making processes.

The dataset underwent meticulous preprocessing steps to ensure data integrity crucial for accurate time series forecasting:

- **Cleaning**: The data was first cleaned to remove any inconsistencies or format issues. This included ensuring all trading days were accounted for and that no duplicate entries existed.
- **Handling Missing Values**: Any missing data points were identified and imputed. Forward filling was used

to handle any gaps in the data, preserving the continuity of the stock's price movement.

- **Normalization**: To aid in the convergence of machine learning algorithms and ensure that features contribute equally to the analysis, the data was normalized. Price data was scaled using the MinMaxScaler, transforming each value to a range between 0 and 1.
- **Outlier Detection and Handling**: The dataset was analyzed for outliers, which were then handled either by capping (winsorizing) or transforming the data to reduce the impact of extreme values.



V. Model Implementation

A. Technical Setup

The computational analysis for this study was conducted using a robust technical setup designed to handle large datasets and complex computations efficiently:

- **Software**: The analysis was performed primarily in Python (version 3.8), utilizing powerful libraries such as Pandas for data manipulation, NumPy for numerical computations, and Matplotlib for data visualization.
- **Machine Learning Frameworks**: TensorFlow and Keras were employed to design and train the LSTM model, while statsmodels was used for ARIMA and SARIMA models. The scikit-fuzzy library provided tools for implementing Fuzzy Time Series.
- **Development Environment**: Jupyter Notebooks served as the primary development environment, offering an interactive coding interface that facilitated data exploration and iterative testing of models.

B. Algorithm Details

I. **Long Short-Term Memory Networks (LSTM):**
The LSTM model used consists of two layers with 50 units each, incorporating dropout layers to prevent overfitting. The model was compiled using the Adam optimizer, with a learning rate of 0.001, and trained to minimize mean squared error (MSE) as the loss function.

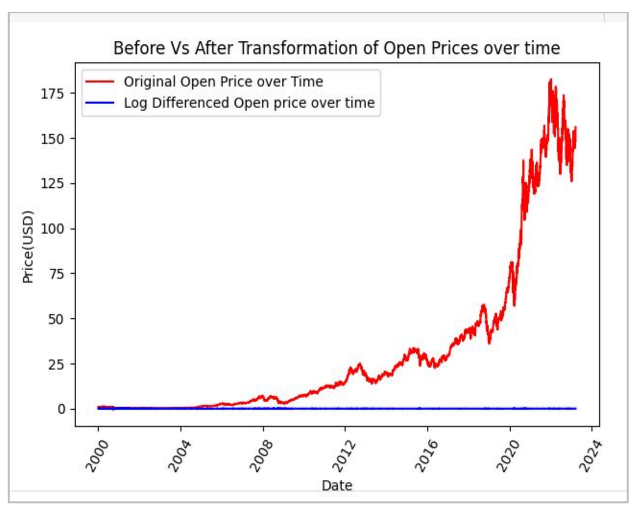
II. **ARIMA/SARIMA**:
Optimal hyperparameter selection is crucial for accurate forecasting with ARIMA (p, d, q) and SARIMA (P, D, Q, s) models. This work employs a two-pronged approach. First, a meticulous grid search evaluates various (p, d, q) and (P, D, Q, s) combinations, minimizing the Akaike Information Criterion (AIC) to identify the best fit. Second, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots inform lag selection for model parameters. This combined approach ensures data-driven hyperparameter tuning for accurate ARIMA/SARIMA model fitting.

III. **Fuzzy Time Series**:
The Fuzzy Time Series model was configured with membership functions defined based on the data's distribution. The model utilized fuzzy relational matrices to establish the relationships between successive data points, capturing the inherent fuzziness in the stock market data.

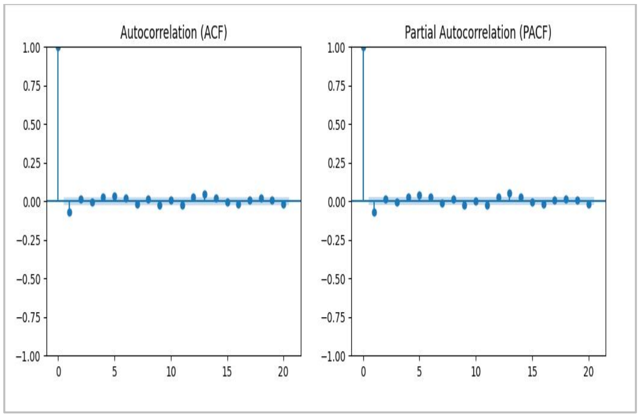
C. Training Process

The dataset was divided into training, validation, and test sets to ensure a robust evaluation of each model:

- **Data Splitting**: 80% of the data was allocated to the training set and remaining 20% to the test set. This split was designed to provide a substantial training dataset while still allowing for thorough testing and validation of model predictions.
- **Augmented Dickey-Fuller (ADF) Tests**: To ensure the effectiveness of ARIMA and SARIMA models, we employed ADF test for robust stationarity analysis. This statistical test evaluates the null hypothesis of a unit root (non-stationary) against the alternative of stationarity. By analyzing the ADF test's p-value, we determined if the data requires differencing to achieve stationarity for optimal model performance.
- **Log Differencing:** This transformation aims to achieve stationarity by capturing linear trends within the data, making it suitable for optimal performance of the ARIMA/SARIMA models.



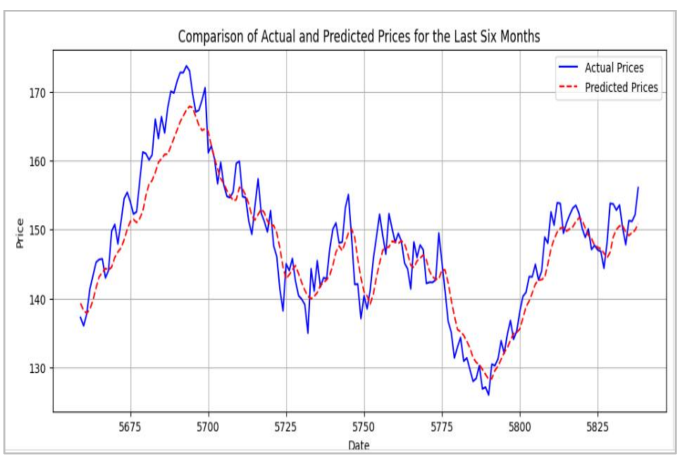
- **Hyperparemeter tuning:** Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots inform lag selection, crucial for precise hyperparameter tuning in ARIMA/SARIMA models, complementing grid search optimization for accurate forecasting.



- **Model Evaluation**: Model performance was assessed using the Root Mean Squared Error (RMSE) on the validation set during the training process to monitor overfitting and underfitting.

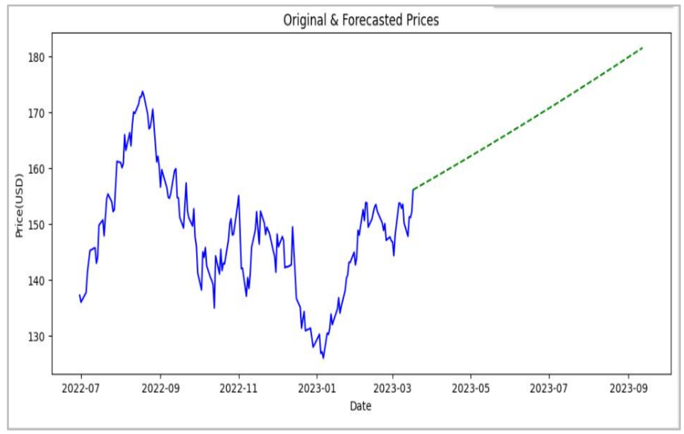
VI. RESULTS

Forecast using LSTM:
Our model exhibits promising forecasting performance. The predicted values indicate a potential upward trend, with a Root Mean Squared Error (RMSE) of as low as 0.03. This low RMSE suggests a high degree of accuracy in capturing the underlying patterns within the stock price data.

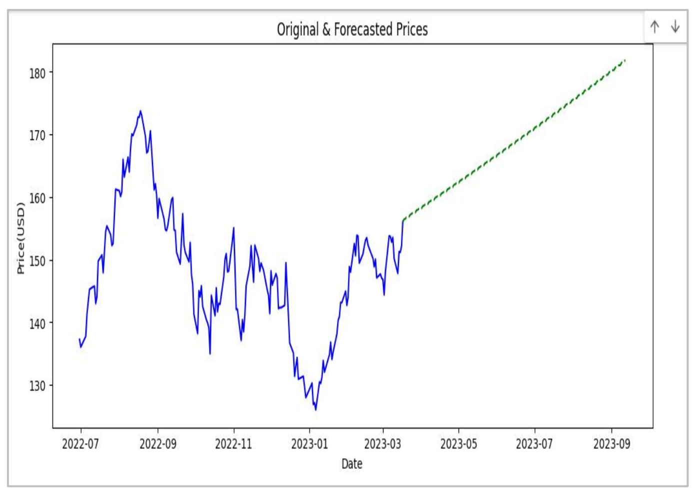


Forecast using ARIMA:
A comprehensive evaluation using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) was conducted to identify the optimal ARIMA model configuration. The ARIMA(1,0,1) model emerged as the frontrunner, exhibiting the lowest AIC and BIC values. This indicates a superior trade-off between model complexity and goodness-of-fit compared to other candidate models. Notably, all models achieved a Root Mean Squared Error (RMSE) of 0.03.

	Model	AIC	BIC	RMSE
0	ARIMA(1, 1, 1)	-20524.23	-20504.89	0.03
1	ARIMA(1, 0, 1)	-20537.41	-20511.61	0.03
2	ARIMA(0, 1, 1)	-20500.15	-20487.26	0.03
3	ARIMA(1, 1, 0)	-18598.74	-18585.85	0.03
4	ARIMA(2, 1, 1)	-20524.07	-20498.28	0.03

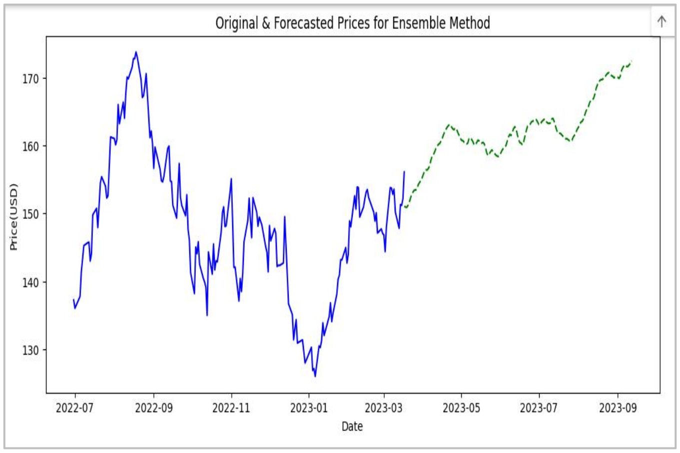


Forecast using SARIMA:
Configurations with a seasonal order of 5 consistently yielded lower AIC and BIC values compared to those with a seasonal order of 21. This suggests that incorporating a seasonality component with a period of 5 units (e.g., 5 days, weeks, etc.) improves model performance for the given time series data.

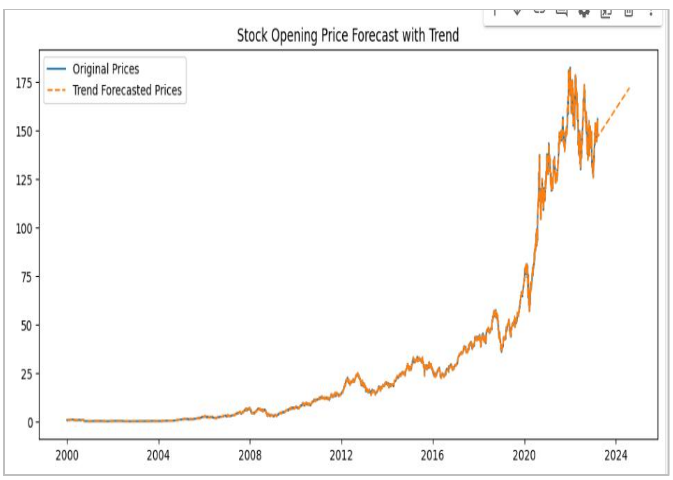


Forecast using Ensemble Method:
The ensemble model, combining LSTM, ARIMA, and SARIMA forecasts, predicted an upward trend for future values. However, the Root Mean Squared Error (RMSE) of 16 indicates a relatively high level of prediction error compared to the individual models discussed previously. While the

upward trend aligns with the model's prediction, the high RMSE suggests a need for further investigation into potential limitations of the ensemble approach for this specific dataset.



Forecast using Fuzzy Time Series:
The Fuzzy Time Series model predicted a significant upward trend in future stock prices. However, the Root Mean Squared Error (RMSE) of 55 indicates a substantial level of prediction error compared to other models evaluated in this study. While the model captured the direction of potential growth, the high RMSE suggests limitations in its ability to precisely quantify the magnitude of future price movements.



VII. Limitations

While the study provided valuable insights, it encountered limitations such as the scope of data, which was confined to one stock ticker over a specific period. Additionally, model biases and the exclusion of external economic factors could have influenced the forecasting outcomes. The complexity introduced by integrating multiple models also raised computational demands and model interpretability issues.

VIII. Conclusion

This research highlighted the capabilities and limitations of employing LSTM, ARIMA, SARIMA, and Fuzzy Time Series models individually and in a hybrid setup for forecasting stock prices. SARIMA's robust performance across all metrics indicates its suitability for scenarios with strong seasonal patterns.

IX. Acknowledgment

We express our sincere gratitude to Dr. Yanqing Zhang for his invaluable guidance throughout the course of this project. Dr. Zhang's expertise in fuzzy logic and his deep understanding of complex system analysis have been instrumental in shaping our approach to this research. His mentorship was not only vital in our understanding and application of fuzzy logic concepts but also inspired us to explore innovative solutions in the field of financial forecasting.

X. References

[1] S. Selvin, R. Vinayakumar, E.A. Gopalakrishnan, V.K. Menon, K.P. Soman, "Stock price prediction using LSTM, RNN and CNN-sliding window model," International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1643-1647, 2017.

[2] Greff, K., Srivastava, R. K., Koutník, J., Steppner, B., & Schmidhuber, J. (2017). LSTM: A search space odyssey. IEEE transactions on neural networks and learning machines, 28(10), 2215-2229.

[3] E. Chong, C. Han, F.C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," Expert Systems with Applications, vol. 83, pp. 187-205, 2017.

[4] T. Tran, D. Pham, "Ensemble forecasting for stock market prediction," International Conference on Knowledge and Systems Engineering (KSE), pp. 292-297, 2017.

[5] G.P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," Neurocomputing, vol. 50, pp. 159-175, 2003.

[6] R. Adhikari, R.K. Agrawal, "An introductory study on time series modeling and forecasting," LAP Lambert Academic Publishing, 2013.

[7] J.V. Hansen, R.D. Nelson, "Data mining of time series using stanced time delay embedding," Technical Report, 1997.

[8] S. Shen, H. Jiang, T. Zhang, "Stock market forecasting using machine learning algorithms," Technical Report, Department of Electrical Engineering, Stanford University, 2012.

[9] I.S.K. Engin, V.V. Sivakumar, V.G. Kerone, "Fuzzy time series forecasting model based on fuzzy logic and artificial neural networks," IEEE International Conference on Fuzzy Systems, pp. 236-241, 2004.

[10] T.H.K. Yu, K.H. Huarng, "A neural network-based fuzzy time series model to improve forecasting," Expert Systems with Applications, vol. 37, no. 4, pp. 3366-3372, 2010.

[11] S.M. Chen, J.R. Hwang, "Temperature prediction using fuzzy time series," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 30, no. 2, pp. 263-275, 2000.

[12] Shafi, M. I., Farooq, U., & Raza, M. S. (2020). A comparative study of ARIMA, LSTM and ensemble machine learning models for stock price prediction. International Journal of Pattern Recognition and Artificial Intelligence, 34(01), 2050002.