FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 21, November, 2026, pp 342 – 350
DOI: https://doi.org/10.33003/fjs-2026-1021-6370

Comparative Evaluation of ARIMA, LSTM and Hybrid ARIMA-LSTM Models for Forecasting the Apple Inc. Stock Price

[1]Asabe Sandra Ahmadu, *[2]Nyako Alhaji Baba

[1]Department of Computer Science, Faculty of Computing, Modibbo Adama University Yola
[2]Department of Computer Science, School of Science Technology, Federal Polytechnic Kaltungo, Gombe State, Nigeria

*Corresponding authors' email: abnyako32@gmail.com

Received: 10 Oct. 2026
Accepted: 04 Nov. 2026
Published: 09 Nov. 2026

ABSTRACT
Accurate stock-price forecasting remains challenging due to the complex, nonlinear, and volatile nature of financial time series. This study developed and evaluated a hybrid ARIMA-LSTM model for forecasting Apple Inc. stock prices using 11,107 observations spanning December 1980 to January 2025. The dataset, obtained from Kaggle, used Close price as the target variable and Open, High, Low, Adjusted Close, and Volume as input features. Three forecasting models were evaluated: ARIMA (1,1,1), a stacked LSTM comprising two LSTM layers with 50 units each followed by two Dense layers, and a hybrid ARIMA-LSTM model that combines ARIMA-based linear forecasting with LSTM residual learning. The data were divided chronologically into 80% training and 20% testing sets, and model performance was assessed using MAE, MSE, RMSE, MAPE, and R^2. ARIMA achieved the best overall performance, with an MAE of 1.3328, MSE of 4.3612, RMSE of 2.0884, MAPE of 1.25%, and R^2 of 0.9990. The LSTM recorded an MAE of 3.7500, MSE of 34.3950, RMSE of 5.8650, MAPE of 2.60%, and R^2 of 0.9920. The hybrid model achieved an MAE of 1.3610, MSE of 4.4570, RMSE of 2.1110, MAPE of 1.30%, and R^2 of 0.9990. These results indicate that ARIMA provided more accurate forecasts than the standalone LSTM and hybrid model for the evaluated Apple stock series and test period. However, the findings are limited to a single stock and one chronological train-test split, which restricts their generalisability to other stocks and market conditions. Future research should evaluate additional assets, forecast horizons, and repeated experimental runs and compare the models against a naive persistence benchmark.

Keywords: Apple Stock Price; ARIMA; LSTM; Hybrid; Time Series Forecasting.

INTRODUCTION

The stock market is an essential pillar of modern global economics, influencing a wide range of economic indicators, investment strategies, and decision-making processes. While equity trading was once limited by strict regulations and manual systems (Burkaltseva et al., 2021), today's dynamic and liberalized stock markets play a vital role in determining economic stability, wealth creation, and corporate competitiveness (Loubochkin, 2018). Among the leading companies shaping global equity markets, Apple Inc. (AAPL) has emerged as a central player due to its market capitalization, innovation-driven growth, and wide investor appeal.

Apple stock serves as a benchmark for technology sector performance and is a major component of indices such as the S&P 500 and NASDAQ. Its fluctuations significantly affect investor sentiment, portfolio valuations, and broader market trends (Nenkov, 2024; Wang, 2024). Beyond its role in global markets, Apple's financial performance also influences supply chains, retail distribution, and consumer technology adoption worldwide. During major product launches or earnings announcements, demand for Apple stock surges, and any downturn in its value can create ripple effects across both institutional and retail investors (Wang, 2024).

The volatility of Apple stock highlights the importance of reliable stock price forecasting. Accurate predictions help policymakers, institutional investors, and individuals manage financial risks and make informed decisions. For example, fund managers rely on predictive models to adjust portfolios, while retail investors use forecasts to time their buying and selling decisions. Likewise, businesses connected to Apple's ecosystem from chip manufacturers to app developers benefit from anticipating future stock performance (Gawer & Cusumano, 2014).

Since the adoption of computerized and algorithmic trading in the late 20th century, stock markets have been increasingly influenced by private market forces and investor psychology, leading to heightened unpredictability (Wang, 2024). This unpredictability is especially pronounced in technology stocks like Apple, where innovation cycles, competition, regulatory changes, and global demand shifts frequently trigger price volatility (Gawer & Cusumano, 2014). Consequently, forecasting stock prices has become an essential task in financial planning, portfolio management, and strategic investment.

Traditional statistical models such as Autoregressive Integrated Moving Average (ARIMA), Exponential Smoothing, and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) have been widely used in stock market forecasting but often fall short in capturing the non-linear and complex dynamics of financial time series (Ibraeva *et al.*, 2023; Ajumi & Kaushik, 2017). In response, modern approaches utilizing Artificial Intelligence (AI) and Machine Learning (ML) techniques have gained traction for their ability to process large datasets, identify hidden patterns, and adapt to changing market conditions (Adekoya *et al.*, 2021).

Recent studies have successfully applied Machine Learning models such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), fuzzy inference systems, and deep learning architectures like Long Short-Term Memory

(LSTM) networks and Convolutional Neural Networks (CNN) to improve forecasting accuracy (Odion *et al*., 2025; Lawal et al., 2025; Zhang *et al*., 2024). These models excel in capturing complex temporal dependencies and non-stationary behaviors, which are common in stock price movements (Bandara et al., 2020; Martínez et al., 2022). Furthermore, hybrid models that combine classical statistical methods with machine learning algorithms have demonstrated superior performance in several empirical studies (Zhang, 2003; Khashei *et al*., 2021).

Given the strategic importance of Apple Inc. in the global financial system, this study aims to explore and evaluate machine learning techniques for forecasting Apple's stock price. Specifically, the study seeks to develop and evaluate a hybrid ARIMA-LSTM model for accurate and reliable stock price forecasting using historical Apple stock price data from December 1980 to January 2025. The study involves collecting and preprocessing the historical stock price data, identifying the appropriate input features for the ARIMA-LSTM model, developing the hybrid model by integrating the strengths of ARIMA and LSTM, and evaluating its forecasting performance using appropriate performance measures. By utilizing the modern AI-driven forecasting techniques, the study seeks to contribute to more accurate and timely predictions that can assist investors, fund managers, policymakers, and businesses in making well-informed financial and investment decisions.

Review of Related Works

Statistical and Deep Learning Approaches to Financial Time-Series Forecasting

Traditional statistical models and deep learning techniques have been widely applied to financial time-series forecasting, although their predictive performance varies according to the asset class, dataset characteristics, and evaluation procedure. Lawal et al. (2025) comparatively evaluated autoregressive integrated moving average (ARIMA) and long short-term memory (LSTM) models for forecasting the Nigerian Naira/US Dollar exchange rate using Central Bank of Nigeria data covering 2001-2024. Their findings showed that LSTM performed better in terms of mean squared error (MSE), root mean squared error (RMSE), and the coefficient of determination (R^2), whereas ARIMA achieved lower mean absolute error (MAE) and mean absolute percentage error (MAPE). This mixed performance indicates that model superiority depends partly on the evaluation metric and suggests that conclusions drawn from one financial series may not necessarily generalise to another asset class. However, because exchange rates and equity prices may exhibit different temporal characteristics, the findings provide methodological motivation rather than direct evidence of the best model for Apple Inc. stock-price forecasting.

Research on stock-market forecasting has also investigated recurrent neural networks and more complex deep learning architectures. Friday et al. (2025) proposed a CNN-Attention-GRU framework for stock-market trend prediction across multiple international stock indices. Their approach combined convolutional feature extraction, attention mechanisms, and recurrent learning to capture different characteristics of financial data. Gajiwala (2025) investigated transformer-based TimeGPT and Spacetimeformer architectures across different forecasting horizons, while Sajal et al. (2025) compared Transformer, LSTM, GRU, temporal convolutional networks (TCNs), and conventional machine learning models for bank profitability forecasting. These studies illustrate the increasing use of architectures designed to learn complex temporal relationships. Nevertheless, the use of more sophisticated architectures does not, by itself, establish superior performance for every stock-price forecasting task. Comparisons must account for the asset being forecast, prediction horizon, test period, input features, and evaluation metrics. Furthermore, bank profitability forecasting differs from daily stock-price prediction in both the target variable and its temporal structure, making the former a less direct basis for assessing equity-price forecasting models.

Hybrid ARIMA-LSTM and Ensemble Forecasting Approaches

Hybrid forecasting approaches seek to combine the strengths of statistical models and neural networks. Cappello et al. (2025) developed hybrid financial forecasting models integrating ARIMA with neural network architectures, including LSTM and gated recurrent unit (GRU) networks, to represent linear and nonlinear patterns in financial time series. Their findings favoured hybrid approaches over standalone statistical and neural network models in the settings investigated. Similarly, Odion et al. (2025) applied an ARIMA-LSTM hybrid model to USD/NGN exchange-rate forecasting and reported an RMSE of 2.216 and an R^2 of 0.998. Although these results provide evidence that hybrid modelling can be effective for the exchange-rate series examined, the reported RMSE cannot be directly compared with results from another asset or study unless the target scale, data preprocessing, forecast horizon, and evaluation protocol are sufficiently comparable. In particular, the findings do not establish that a hybrid model will outperform standalone ARIMA or LSTM when forecasting Apple Inc. stock prices.

Beyond statistical-neural hybrids, other studies have explored ensemble learning to improve financial prediction. Abir et al. (2025) evaluated LSTM, Random Forest, Gradient Boosting, and Support Vector Machine (SVM) models for currency and stock-market forecasting across BRICS economies. Their results indicated that LSTM performed best for exchange-rate prediction, whereas Gradient Boosting performed best for stock-market prediction. Their ensemble framework sought to improve predictive reliability by combining complementary model capabilities. Malgwi et al. (2026) developed a Hybrid Naïve-LSTM model to forecast the Nigerian Naira–US Dollar exchange rate using daily CBN data from October 2009 to August 2026, incorporating Brent crude oil prices and 49 engineered features. The model was compared with a Naïve persistence model and an ARIMA-LSTM benchmark. Results showed that the Naïve model outperformed the hybrid model, highlighting the strong persistence of exchange-rate movements and the importance of benchmarking complex deep-learning models against simpler baselines. Taken together, these studies suggest that hybridisation and ensembling are useful modelling strategies, but their benefits are empirical rather than universal: a combined model must be evaluated against its individual components on the same test observations to establish whether integration provides a measurable improvement.

Multimodal and Feature-Enriched Financial Forecasting

Another research direction incorporates information beyond historical prices, including financial news, sentiment, technical indicators, and macroeconomic variables. Chauhan et al. (2025) developed the sentiment-aware SenT-In framework, integrating financial news sentiment with market data through CNN and GRU architectures. Carosia et al. (2025) likewise investigated the combination of historical stock prices, technical indicators, and financial news,

reporting improvements in forecasting accuracy and investment performance relative to approaches based solely on historical market information. These studies highlight the potential value of incorporating additional information into financial forecasting systems.

However, multimodal forecasting differs from the present study in its input requirements and modelling objectives. The present research focuses on forecasting Apple stock prices using historical market variables rather than integrating news sentiment or other external information sources. Consequently, these studies provide useful context for future extensions but are not direct benchmarks for the performance of the ARIMA, LSTM, and hybrid ARIMA-LSTM models evaluated here. A meaningful comparison of such approaches would require consistent target definitions, test periods, forecast horizons, and evaluation metrics, as well as access to comparable input data.

Research Gap

The reviewed literature demonstrates that statistical, deep learning, hybrid, ensemble, and multimodal approaches have all been investigated for financial forecasting. However, findings across exchange-rate prediction, stock-market forecasting, and other financial prediction tasks are not directly interchangeable because the target variables, data characteristics, input features, and evaluation protocols differ. Studies by Lawal et al. (2025), Odion et al. (2025), and Cappello et al. (2025) provide methodological motivation for investigating ARIMA-LSTM integration, but their findings alone do not establish that this combination is superior for Apple stock-price forecasting. Likewise, the fact that a study focuses on Apple Inc. or uses a long historical dataset is not, by itself, sufficient to establish methodological novelty.

Accordingly, the contribution of the present study is a dataset-specific empirical evaluation of ARIMA (1,1,1), a stacked LSTM, and a hybrid ARIMA-LSTM model using Apple Inc. stock-price observations spanning December 1980 to January 2025.

MATERIALS AND METHODS

This section describes the dataset, data preprocessing and the step by step of the process flow of the various methods and methodology proposed in this research. Figure 1 gives a clear diagrammatic description of the proposed model.

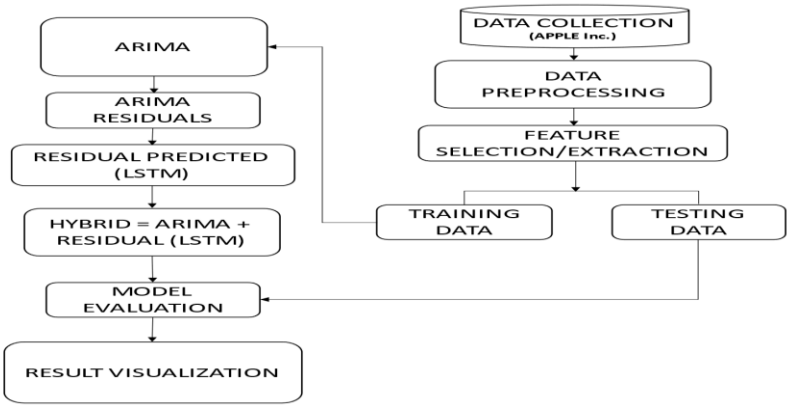


Figure 1: Framework of the Proposed Model

Data Collection

The study evaluates a hybrid ARIMA-LSTM model for forecasting Apple Inc.'s stock prices using historical data from December 1980 to January 2025. The dataset, obtained from (Kaggle, 2025), contains 11,107 observations with seven variables: Date, Open, High, Low, Close, Adjusted Close, and Trading Volume. The data provides a basis for examining long-term price trends, fluctuations, and market patterns in Apple's stock. Table 1 presents an overview of the dataset.

Table 1: Overview of The Dataset

Variable	Description
Date	Date
Open	The price at market open.
High	The highest price for that day.
Low	The lowest price for that day.
Close	The price at market close, adjusted for splits.
adj_close	The closing price after adjustments for all applicable splits and dividend distributions.
Volume	The number of shares traded on that day.

Data Processing

The dataset was first checked for quality and completeness, and no missing or null values were found, eliminating the need for data imputation. The Date variable was then converted from string format to datetime using the mm/dd/YY format to support chronological ordering, temporal analysis, feature extraction, and visualization.

Feature Selection

The Close price was selected as the target variable because it represents the final daily trading price and is widely used in financial forecasting. The Open, High, Low, Adj_Close, and Volume variables were used as input features to capture daily price movements, trading activity, and market dynamics that may influence the closing price.

Data Splitting

The dataset was divided into training and testing sets, with the training set used to learn patterns in the data and train the models. 20% of the data was used for testing and 80% for training, ensuring a balanced approach to model training and evaluation.

Forecasting Models

In this study, three forecasting approaches were developed and evaluated: ARIMA, LSTM, and a hybrid ARIMA-LSTM model. The individual models were first developed to establish their forecasting capabilities, after which ARIMA and LSTM were combined to exploit their complementary strengths. ARIMA is effective in capturing linear temporal patterns, while LSTM can learn complex nonlinear and long-term dependencies in time-series data.

Statistical analysis and ARIMA

Before developing the ARIMA model, the stationarity of the closing-price series was examined using the Augmented Dickey-Fuller (ADF) test. The original series produced an ADF statistic of 4.4379 and a p-value of 1.0, indicating non-stationarity since the p-value exceeded 0.05. Therefore, first-order differencing was applied. The ADF test on the differenced series produced a p-value of 6.0381×10^{-28}, confirming stationarity and making the series suitable for ARIMA modeling.

The Ljung-Box test was also conducted to examine temporal dependence. At lag 10, the test produced a Q-statistic of 215.77 with a p-value of 8.20×10^{-41}. However, the series to which this statistic applies must be verified and explicitly identified. In the revised analysis, the test should be reported separately for the original closing-price series, the differenced series, and the fitted ARIMA residuals, where appropriate. The Ljung-Box test on the original series assesses serial dependence; the test on model residuals helps determine whether significant autocorrelation remains unexplained by the fitted model.

The ARIMA model was selected to represent the linear temporal structure of the closing-price series. First-order differencing (d=1) was used to address non-stationarity. The autoregressive order (p) and moving-average order (q) should not be selected solely from the autocorrelation function (ACF) and partial autocorrelation function (PACF). To strengthen model selection, candidate ARIMA specifications should be compared using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), with model adequacy assessed through residual diagnostics. The ARIMA (1,1,1) specification was retained if this comparison supports it. The candidate orders, corresponding information criteria, and residual diagnostic results should be reported in the revised results.

The chronological data partition was used to prevent information from the test period from influencing model selection. The training data were used to estimate candidate models, while a separate validation period within the pre-test data should be used to guide model selection where feasible. The final test period should remain untouched until model selection and configuration are complete. The model configuration is summarized in Table 2.

Table 2: ARIMA Model Configuration

Parameter	Value	Description
p	1	One autoregressive lag
d	1	First-order differencing
q	1	One moving-average lag
Model	ARIMA(1,1,1)	Final forecasting configuration
Selection criteria	AIC and BIC	Comparison of candidate specifications
Diagnostic test	Ljung-Box	Assessment of residual autocorrelation

The forecasting evaluation should explicitly define the forecast horizon. If the intended task is one-trading-day-ahead prediction, each forecast should estimate the next trading day's closing price using only information available at the forecast origin. The ARIMA parameters should be estimated using the training data and kept fixed during the test period, unless a rolling re-estimation procedure is explicitly implemented and reported. For rolling one-step-ahead evaluation, the model state may be updated using newly observed historical values without re-estimating the model parameters at every step. The same forecast-origin convention must be used when comparing ARIMA, LSTM, the hybrid model, and the baseline. The residual Q-Q plot in Figure 2 is used to assess deviations from normality. Deviations at the tails suggest heavy-tailed residual behaviour or potential outliers. Residual normality is not, by itself, a requirement for point forecasting, but remaining autocorrelation and changing residual variance should be acknowledged because they may affect forecast uncertainty and model adequacy.

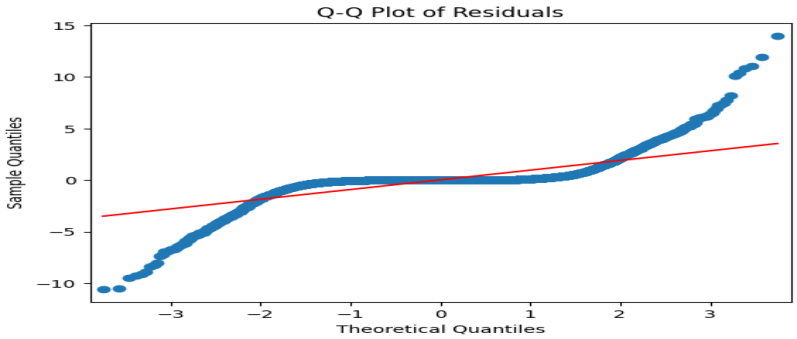


Figure 2: QQ Plot of Residuals

LSTM Model

The LSTM model was developed using TensorFlow/Keras to capture temporal dependencies in historical Apple stock-price data. It employed a 60-time-step look-back window and comprised two stacked LSTM layers with 50 units each, followed by a dense layer of 25 neurons and a single-neuron output layer. The model was trained using the Adam optimizer, mean squared error (MSE) loss, and a batch size of 32. The dataset was divided chronologically into 80% training data and 20% testing data, with a validation subset separated from the training period where implemented. Scaling parameters were fitted exclusively to the training data and applied consistently to validation and test data, while predictions were transformed back to the original price scale before evaluation.

To prevent target leakage, input features were restricted to information available at the forecast origin, with contemporaneous Open, High, Low, and Adjusted Close values lagged or excluded where necessary. The model's input dimensionality and trainable parameters were verified against the Keras model summary. For reproducibility, the training procedure should document the random seed, maximum epochs, validation strategy, and stopping criterion. Where feasible, multiple random seeds and validation-based early stopping should be used to improve the reliability of performance estimates. The same training and validation principles should be applied to the hybrid residual LSTM to ensure a fair comparison between the models. The configuration is summarized in Table 3.

Table 3: LSTM Model Configuration

Layer	Configuration
LSTM 1	50 units, return sequences
LSTM 2	50 units
Dense 1	25 neurons
Output	1 neuron
Lookback	60 time steps
Optimizer	Adam
Loss	MSE
Batch size	32
Epochs	100

Hybrid ARIMA-LSTM Model

To combine the strengths of the statistical and deep learning approaches, a hybrid ARIMA-LSTM model was developed. The model first uses ARIMA (1,1,1) to capture the linear component of the stock-price series, while LSTM is subsequently used to learn the nonlinear patterns contained in the ARIMA residuals. This approach is based on the assumption that a financial time series can be represented as a combination of linear and nonlinear components:

$$x_t = L_t + N_t \quad (1)$$

Where, x_t represents the observed value, L_t is the linear component estimated by ARIMA, and N_t represents the nonlinear component learned by LSTM from the ARIMA residuals. The residual is calculated as $r_t = x_t - L_t$.

The hybrid modeling process began by fitting the ARIMA (1,1,1) model to the training data and obtaining its predictions. The residuals between the actual and ARIMA-predicted values were then calculated and normalized using MinMax scaling. The normalized residuals were transformed into sequences using a 60-time-step lookback window and used to train the LSTM network. The LSTM retained the same 50-50-25-1 architecture but was trained for 50 epochs using the Adam optimizer and MSE loss.

During testing, ARIMA first generated the base forecasts, after which the trained LSTM predicted the nonlinear residual component. The ARIMA and LSTM outputs were then combined to produce the final hybrid forecast. The major configurations of the hybrid model are presented in Table 4.

Table 4: Hybrid Model Configuration

Component	Configuration
ARIMA	(1,1,1)
ARIMA role	Linear/base forecast
LSTM	50-50-25-1
LSTM role	Nonlinear residual correction
Lookback window	60 time steps
Optimizer	Adam
Loss function	MSE
Batch size	32
Residual LSTM epochs	50

The performance of the ARIMA, LSTM, and hybrid ARIMA-LSTM models was evaluated using MAE, MSE, RMSE, MAPE, and R^2. These were estimated using Equations (2), (3), (4), and (5) respectively, with *n* representing the number of observations.

MAE: The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.

$$\text{MAE}_{(y_i, \hat{y}_i)} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

Where; y_i is the actual value, $\hat{y}_i$ is the predicted value and n is the number of observations.

MSE: Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.

$$MSE_{(yi,\ yi)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2 \quad (3)$$

RMSE: is the square root of Mean Squared error. It measures the standard deviation of residuals.

$$RMSE_{(yi,\ yi)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2} \quad (4)$$

Where RMSE is just the square root of MSE.

R-s: The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model.

$$R^2_{(yi,\ yi)} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

Software and Computational Environment

The experiments were conducted using Google Colaboratory (Google Colab), a cloud-based Python development environment that provides access to computational resources for data analysis and machine learning. Python was used for data preprocessing, exploratory data analysis, model development, training, forecasting, and performance evaluation. The ARIMA model was implemented using the Statsmodels library, while the LSTM and hybrid ARIMA–LSTM models were developed using TensorFlow and Keras. Additional libraries, including NumPy, Pandas, Scikit-learn, and Matplotlib, were used for numerical computation, data manipulation, preprocessing, evaluation, and visualization. Google Colab provided an accessible and reproducible environment for implementing and evaluating the forecasting models.

RESULTS AND DISCUSSION

Results

The results of the study are presented based on the historical behaviour of Apple Inc.'s stock price, the diagnostic assessment of the ARIMA model, the forecasting performance of the three models, and their comparative evaluation. The dataset comprised 11,107 daily observations covering December 1980 to January 2025. The models considered were ARIMA(1,1,1), LSTM, and the hybrid ARIMA-LSTM model.

Historical Stock Price Behaviour

Figure 3 presents the historical movement of Apple Inc.'s closing stock price from December 1980 to January 2025. The series shows a clear long-term upward movement, although the trend is accompanied by periods of increased fluctuations. The earlier portion of the series records relatively low price levels, while substantially higher values are observed in the later years. The increasing magnitude of price movements over time indicates that the series contains both persistent trends and changing levels of variability, making it appropriate for evaluating models capable of capturing linear and nonlinear temporal patterns.



Figure 3: Historical Apple Stock Price (1980-2025)

ARIMA Model Diagnostics

The stationarity analysis revealed that the original Apple stock-price series was non-stationary, as indicated by an ADF test statistic of 4.4379 and a p-value of 1.0000. After applying first-order differencing, the ADF test produced a p-value of 6.0381×10^{-28}, confirming that the differenced series was stationary. This finding justified the use of one order of differencing and supported the selection of d = 1 for the ARIMA model.

The Ljung-Box test at lag 10 produced a Q-statistic of 215.77 with a p-value of 8.20×10^{-41}, indicating significant autocorrelation in the series. The ACF and PACF plots further supported the inclusion of one autoregressive term and one moving-average term. Based on these diagnostic results, ARIMA (1,1,1) was selected for the forecasting analysis.

To maintain the integrity of the forecasting evaluation, the ARIMA model was fitted using the training portion of the chronologically ordered dataset, comprising 8,886 observations, while the remaining 2,221 observations were reserved for testing. The model configuration and diagnostic outputs were verified against the training data to ensure that the test observations were not used during model fitting.

The estimated ARIMA coefficients from the verified training-only fit were examined for statistical significance. The original reported AR (1) and MA (1) coefficients were 0.5552 and -0.5734, respectively, with p-values below 0.001. The originally reported information criteria were AIC = 30,165.629, BIC = 30,187.575, and HQIC = 30,173.019.

Residual diagnostics were conducted to assess whether the fitted model adequately captured the temporal dependence in the training data. The originally reported Ljung-Box test at lag 1 yielded a Q-statistic of 0.50 and a p-value of 0.48, indicating insufficient evidence of residual autocorrelation at that lag. However, the Jarque-Bera test indicated departures from residual normality, while the heteroskedasticity test was significant, suggesting non-constant residual variance. These findings indicate that the model may have captured important linear temporal dependencies, but the residual distribution and variance characteristics remain limitations. The diagnostic results were therefore interpreted alongside the model's forecasting performance on the held-out test set.

Forecasting Results

The forecasting results for ARIMA, LSTM, and the hybrid ARIMA-LSTM model are presented in Figure 4 and Table 5. The comparative prediction plot shows the relationship between the observed Apple stock prices and the values predicted by the three models during the testing period.

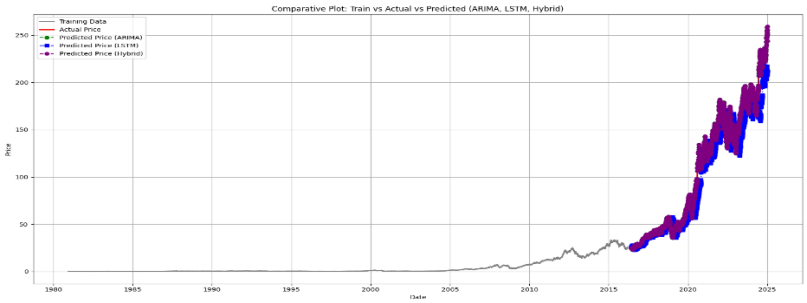


Figure 4: Actual and Predicted Apple Stock Prices During the Testing Period

Given the substantial difference between the training and testing price ranges, the preprocessing and scaling procedures were considered important factors in interpreting the forecasting results. The training data covered an earlier period when Apple's nominal stock prices were considerably lower than those observed during the later testing period. Consequently, performance on the test set needed to be assessed with particular attention to the models' ability to generalise beyond the price levels represented in the training data. For the LSTM-based models, the scaling procedure and inverse transformation of predictions were verified to ensure that the test observations were not used to fit the scaler and that predictions were evaluated in the original price units.

Figure 4 compares the observed closing prices with the predictions produced by the three models. The models exhibited different levels of agreement with the observed series. ARIMA produced the lowest aggregate forecast errors in the initial evaluation, whereas the standalone LSTM showed larger deviations. The hybrid ARIMA-LSTM model achieved performance close to that of ARIMA. However, the interpretation of these differences depends on the validity of the preprocessing procedure, the alignment between forecast dates and actual observations, and the models' performance under the chronological test design.

Table 5 presents selected observations and their corresponding predictions. These examples illustrate the differences between the models at individual points in the testing period. The reported prediction dates and values should be verified against the original forecast arrays before drawing conclusions about which model tracked the observed prices most closely. The final comparison should be based on correctly aligned predictions and actual closing prices, evaluated using the same test observations.

Table 5: Overview of Actual Vs Predicted of the Proposed Models

DATE	ACTUAL	LSTM Predicted	ARIMA Predicted	Hybrid Predicted
6/2/2016	24.43	25.43914	24.9532	24.59785
6/3/2016	24.48	25.24649	25.83342	24.42128
6/6/2016	24.6575	25.27377	25.25247	24.47635
6/7/2016	24.7575	25.29519	23.6214	24.65847
6/8/2016	24.735	25.56714	25.03785	24.76254
12/27/2024	255.59	211.7529	226.2985	259.0277
12/30/2024	252.2	211.9963	226.7255	255.6105
12/31/2024	250.42	211.1532	225.7109	252.2182
1/2/2025	243.85	212.0801	226.7667	250.4496
1/3/2025	243.86	208.0456	221.7565	243.9638

Table 5 provides selected examples of actual and predicted prices. For instance, on 27 December 2024, the actual closing price was 255.59, while the LSTM, ARIMA, and hybrid models predicted 211.75, 226.30, and 259.03, respectively. Similarly, on 30 December 2024, the actual price was 252.20 compared with predictions of 212.00, 226.73, and 255.61 from the LSTM, ARIMA, and hybrid models, respectively. On 3 January 2025, the actual closing price was 243.86, while the corresponding predictions were 208.05, 221.76, and 243.96. These examples indicate that the hybrid model was particularly effective in tracking the observed prices toward the end of the testing period.

Evaluation Metrics

The quantitative performance of the ARIMA, LSTM, and hybrid ARIMA-LSTM models was assessed using MAE, MSE, RMSE, MAPE, and R^2. These metrics quantify prediction errors and the degree of agreement between the predicted and observed closing prices.

The initial evaluation showed that ARIMA achieved an MAE of 1.3328, MSE of 4.3612, RMSE of 2.0884, MAPE of 1.25%, and R^2 of 0.9990. The standalone LSTM recorded an MAE of 3.7500, MSE of 34.3950, RMSE of 5.8650, MAPE of 2.60%, and R^2 of 0.9920. The hybrid ARIMA-LSTM model recorded an MAE of 1.3610, MSE of 4.4570, RMSE of 2.1110, MAPE of 1.30%, and R^2 of 0.9990.

Table 6: Evaluation Metrics

ALGORITHM	MAE	MSE	RMSE	MAPE	R2
ARIMA	1.3328	4.3612	2.0884	1.25%	0.9990
LSTM	3.7500	34.395	5.8650	2.60%	0.9920
HYBRID	1.3610	4.4570	2.1110	1.3%	0.9990

These values represent the initial evaluation and should be confirmed after verifying the training-only scaling procedure, prediction-date alignment, and the treatment of the changing price level. In particular, the high R^2 values should be interpreted cautiously because stock-price levels can exhibit strong temporal trends, and a high level-based R^2 does not necessarily establish accurate forecasting of price changes.

Subject to confirmation through the corrected evaluation, ARIMA achieved the lowest reported errors, followed closely by the hybrid model, while the standalone LSTM produced larger errors. The results do not demonstrate that the hybrid model outperformed ARIMA. Instead, they suggest that the additional residual-learning component did not yield a measurable improvement under the initial experimental conditions. The final comparison should use predictions generated without information leakage and should distinguish price-level accuracy from the ability to forecast changes in stock prices.

Discussion

The initial results indicate that ARIMA achieved lower forecasting errors than the standalone LSTM and hybrid ARIMA-LSTM models on the evaluated Apple stock-price series. ARIMA recorded an MAE of 1.3328 and an RMSE of 2.0884, compared with 3.7500 and 5.8650, respectively, for LSTM. The hybrid model recorded an MAE of 1.3610 and an RMSE of 2.1110. However, these results must be interpreted cautiously because the selected predictions in Table 5 were inconsistent with the reported aggregate metrics. The prediction arrays, date alignment, and evaluation metrics must therefore be verified before firm conclusions can be drawn about the relative performance of the models.

The LSTM's higher reported errors should be examined in relation to the substantial change in Apple's nominal closing-price range between the training and testing periods. A model trained on historical price levels may struggle when test observations extend beyond the range represented during training, particularly when the target is scaled using parameters estimated from the training data. Although fitting the scaler exclusively to the training data is essential to prevent data leakage, it does not guarantee effective extrapolation. The scaling method, target transformation, and inverse-transformation procedure should therefore be documented and verified. Modelling log returns or differenced prices could provide a useful complementary experiment because these representations focus on price changes rather than directly learning the evolving nominal price level. Any resulting forecasts must be converted appropriately to price levels if the primary objective remains closing-price prediction.

The hybrid ARIMA-LSTM model produced initial errors close to those of ARIMA but did not achieve lower aggregate errors. Thus, the initial results do not demonstrate that residual-based nonlinear correction improved forecasting accuracy. This finding applies only to the evaluated configuration and period and should not be generalised to all hybrid forecasting approaches. A valid comparison requires correctly aligned predictions, consistent evaluation dates, and the same forecast horizon across all models.

The residual diagnostics reported for ARIMA indicated heteroskedasticity and high kurtosis, suggesting time-varying residual variance and heavy-tailed behaviour. These findings warrant further examination of the residuals from all three models. In particular, the LSTM and hybrid residuals should be assessed using residual plots, autocorrelation functions, and Ljung-Box tests to determine whether systematic temporal dependence remains. Residual variance should also be examined over time, and error measures should be calculated across distinct test periods or volatility regimes. These analyses would help establish whether the models' errors increase during periods of rapid price movement or elevated volatility. A GARCH-based extension could be investigated if modelling conditional volatility is an objective, but its benefits would need to be demonstrated empirically.

The results should also be compared with the forecasting errors reported in the studies reviewed in Section 2. Such comparisons must account for differences in datasets, observation periods, forecast horizons, target transformations, and evaluation protocols. Numerical error values from different studies are not directly comparable when these experimental conditions differ. Accordingly, the literature comparison should identify the datasets and forecasting settings used by previous studies and discuss whether their reported findings are consistent with, or differ from, the results obtained in this study.

Several limitations should be acknowledged. First, the analysis focuses on a single stock, Apple Inc., so the findings may not generalise to other equities or financial markets. Second, the chronological holdout evaluation represents performance over one test period and may not capture performance under different market conditions. Third, the substantial change in nominal price levels and the need to verify prediction alignment and preprocessing limit the conclusions that can be drawn from the initial metrics. Finally, without additional residual analysis, period-specific error assessment, and comparison against a naïve random-walk baseline, the evidence is insufficient to establish consistent superiority over simpler forecasting methods.

CONCLUSION

This study developed and evaluated ARIMA (1,1,1), LSTM, and hybrid ARIMA-LSTM models for forecasting Apple Inc. stock prices using historical data spanning more than four decades. Initial results indicated that ARIMA achieved the lowest reported errors, while the hybrid model produced comparable but slightly higher errors, and standalone LSTM recorded the highest errors. However, inconsistencies between selected predictions and aggregate evaluation metrics must be resolved before definitive conclusions can be drawn. The study is limited by its focus on a single stock, reliance on one chronological train-test split, and the need to verify forecast-date alignment, scaling procedures, and potential target leakage. Future research should employ walk-forward evaluation and a naïve random-walk baseline, compare price-level forecasting with log-return and differenced-price approaches, and investigate residual dependence and volatility-related errors to determine whether extensions such as ARIMA-GARCH or improved residual-based hybrid models offer measurable performance gains.

DATA AVAILABILITY

The data used in this study are publicly available on Kaggle. The code developed and used in the study is available from the corresponding author upon reasonable request.

REFERENCES

Abir, S. I., Saimon, S. I., Saha, T., Sarwer, M., Hasan, M., Sultana, N., & Islam, I. (2025). Comparative analysis of currency exchange and stock markets in BRICS using machine learning to forecast optimal trends for data-driven decision making. *Journal of Economics, Finance, and Accounting Studies*, *7*(1), 26–48. https://doi.org/10.32996/jefas.2025.7.1.3

Adekoya, A. F., Nti, I. K., & Weyori, B. A. (2021). Long short-term memory network for predicting exchange rate of the Ghanaian cedi. *FinTech*, *1*(1), 25–43. https://doi.org/10.3390/fintech1010002

Ajumi, O., & Kaushik, A. (2017). Exchange rates prediction via deep learning and machine learning: A literature survey on currency forecasting. *International Journal of Science and Research*, *7*(12), 1252–1262. https://www.academia.edu/download/81531908/ART20193849.pdf

Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, *140*, Article 112896. https://doi.org/10.1016/j.eswa.2019.112896

Burkaltseva, D., Niyazbekova, S., Borsch, L., Jallal, M. A. K., Apatova, N., Nurpeisova, A., … Zhansagimova, A. (2022). Methodological foundations of the risk of the stock markets of developed and developing countries in the conditions of the crisis. *Journal of Risk and Financial Management*, *15*(1), Article 3. https://doi.org/10.3390/jrfm15010003

Cappello, C., Congedi, A., De Iaco, S., & Mariella, L. (2025). Traditional prediction techniques and machine learning approaches for financial time series analysis. *Mathematics*, *13*(3), Article 537. https://doi.org/10.3390/math13030537

Carosia, A. E. D. O., da Silva, A. E. A., & Coelho, G. P. (2025). Predicting the Brazilian stock market with sentiment analysis, technical indicators and stock prices: A deep learning approach. *Computational Economics*, *65*(4), 2351–2378. https://doi.org/10.1007/s10614-024-10636-y

Chauhan, J. K., Ahmed, T., & Sinha, A. (2025). A novel deep learning model for stock market prediction using a sentiment analysis system from authoritative financial website's data. *Connection Science*, *37*(1), Article 2455070. https://doi.org/10.1080/09540091.2025.2455070

Friday, I. K., Pati, S. P., Mishra, D., Mallick, P. K., & Kumar, S. (2025). CAGTRADE: Predicting stock market price movement with a CNN-attention-GRU model. *Asia-Pacific Financial Markets*, *32*(2), 583–608. https://doi.org/10.1007/s10690-024-09463-w

Gajiwala, C. (2025). Stock market analysis using deep learning. *Journal of Computer Science and Technology Studies*, *7*(2), 559–566. https://doi.org/10.32996/jcsts.2025.7.2.59

Gawer, A., & Cusumano, M. A. (2014). Industry platforms and ecosystem innovation. *Journal of Product Innovation Management*, *31*(3), 417-433. https://doi.org/10.1111/jpim.12105

Ibraeva, Z. E., Bektemyssova, G., & Ahmad, A. R. (2023). Fuzzy model for time series forecasting. *Scientific Journal of Astana IT University*, *2023*, 93-102. https://doi.org/10.37943/13eotu7482

Kaggle (2025). Apple Stock Market Historical Data (1980-2025) https://www.kaggle.com/datasets/umerhaddii/apple-stock-data-2025

Khashei, M., Bijari, M., & Hejazi, S. R. (2012). Combining seasonal ARIMA models with computational intelligence techniques for time series forecasting. *Soft Computing*, *16*(6), 1091-1105. https://doi.org/10.1007/s00500-012-0805-9

Lawal, M. A. M., Abdulrauf, A., Delmut, R. D., & Tohomdet, L. K. Evaluating the Effectiveness of ARIMA and LSTM Models in Predicting USD/NGN Trends. *Machine learning*, *13*, 14. https://iccait.com.ng/wp-content/uploads/2026/05/12_Mohammed_Lawal_et-al.pdf

Loubochkin, M. M. (2018). Economic growth and company stock pricing in the global stock market. Вестник Санкт-Петербургского университета. Экономика, 34(3), 396. https://doi.org/10.21638/spbu05.2018.303

Malgwi, Y. M., Abba, S. A., & Pam, N. G. (2026). A Persistence-Driven Naïve-LSTM Residual Hybrid Model for Forecasting the United States Dollar-Nigerian Naira Exchange Rate. *FUDMA Journal of Sciences*, *10*(18), 162-170.

Martínez, F., Frías, M. P., Pirez-Godoy, M. D., & Rivera, A. J. (2022). Time series forecasting by generalized regression neural networks trained with multiple series. *IEEE Access*, *10*, 3275–3283. https://doi.org/10.1109/ACCESS.2022.3140377

Nenkov, D. (2024). "The magnificent seven" technology stocks and their impact on the S&P 500: A review 4 years later. *Finance, Accounting and Business Analysis*, *6*(2), 180–195. https://www.ceeol.com/search/article-detail?id=1293489

Odion, P. O., Lawal, M. M., & Abdulrauf, A. (2025). A comparative analysis of an enhanced hybrid model for predicting Dollar against Naira exchange rate using deep learning and statistical methods. *Journal of Computing Theories and Applications*, *2*(4), 511–522. https://doi.org/10.62411/jcta.12513

Sajal, A., Chy, M. S. K., Jamee, S. S., Uddin, M. N., Khan, M. S., Gharami, A. K., & Ahmed, M. (2025). Forecasting bank profitability using deep learning and macroeconomic indicators: A comparative model study. *International Interdisciplinary Business Economics Advancement Journal*, *6*(6), 8–20. https://doi.org/10.55640/business/volume06issue06-02

Wang, M. (2024). Do new products affect a company's share price? An event study of Apple *SSRN*. http://dx.doi.org/10.2139/ssrn.5567219

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175. 10.1016/S0925-2312(01)00702-0

Zhang, W., Li, J., & Li, Y. (2023, December). A novel loss function considering the distance between forecasting and historical values in financial time series forecasting models. In *Proceedings of the 5th Management Science Informatization and Economic Innovation Development Conference* (pp. 8-10). https://doi.org/10.4108/eai.8-12-2023.2344823

©2026 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.