Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

A UNIFIED BENCHMARK OF STATISTICAL, MACHINE LEARNING, AND DEEP LEARNING APPROACHES FOR S&P 500 INDEX FORECASTING

Khansa Shakeel[1], Dr. Syed Safdar Hussain[*2], Maryam Khalid[3], Faisal Ghaffar[4], Imad Ali[5], Zoha Saif[6], Muhammad Kashif Majeed[7], Muhammad Daud Abbasi[8]

[1,*2]*Department of Computer Science, Iqra University, Karachi, Pakistan*
[3]*Institute of Geography, University of the Punjab, Lahore, Pakistan*
[4]*Department of Computer System Engineering, University of Engineering and Applied Sciences, Swat, KP, Pakistan*
[5]*Department of Computer Science, University of Shangla, KP, Pakistan*
[6]*Department of Mathematics and Statistics, University of Agriculture, Faisalabad, Pakistan*
[7,8]*Faculty of Engineering Science and Technology, Iqra University, Karachi, Pakistan*

[1]khansa.73573@iqra.edu.pk, [*2]safdar@iqra.edu.pk, [3]maryamkhalid.geog@pu.edu.pk, [4]engr.faisal90@gmail.com, [5]imad.ali@ushangla.edu.pk, [6]saadullahmalhi2@gmail.com, [7]mkashif@iqra.edu.pk, [8]daud.abbasi@iqra.edu.pk

DOI: http://doi.org/10.5281/zenodo.19046837

Keywords
Financial Time Series Forecasting; S&P 500 Index Prediction; Machine Learning Models; Deep Learning Architectures; ARIMA; Logistic Regression; Random Forest; XGBoost; LSTM Networks; Financial Data Analytics.

Article History
Received: 14 January 2026
Accepted: 26 February 2026
Published: 14 March 2026

Corresponding Author: *
Dr. Syed Safdar Hussain

Abstract

Financial time series forecasting remains one of the most challenging problems in quantitative finance due to the highly volatile, noisy, and non-stationary nature of financial markets. Accurate prediction of stock market indices plays a crucial role in investment decision-making, portfolio optimization, and risk management. In recent years, machine learning and deep learning techniques have gained increasing attention for financial forecasting tasks. However, the comparative effectiveness of traditional statistical models, classical machine learning algorithms, and deep learning architectures under a unified experimental framework remains insufficiently explored.

This study presents a comprehensive empirical evaluation of statistical, machine learning, and deep learning approaches for forecasting the S&P 500 stock market index. Using a dataset consisting of 25 years of daily historical data (2000–2024) including Open, High, Low, Close, and Volume (OHLCV) features, we benchmark eight forecasting models across three methodological categories. The evaluated models include ARIMA as a statistical baseline; logistic regression and support vector machines as classical machine learning methods; random forest and XGBoost as ensemble learning approaches; and deep learning architectures including Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and a hybrid CNN–LSTM model.

To ensure a fair comparison, all models are implemented within a unified experimental pipeline incorporating consistent preprocessing, feature normalization, rolling window segmentation, and chronological train–validation–test splitting to prevent data leakage. Model performance is evaluated using two complementary metrics: Root Mean Squared Error (RMSE) for regression forecasting and directional accuracy for classification-based prediction of market movements.

Experimental results reveal that simpler models can outperform more complex architectures in financial time series forecasting under constrained feature spaces. In particular, logistic regression achieved the highest directional accuracy of 81.96%, significantly outperforming several machine learning and deep learning models. Deep learning architectures such as LSTM and CNN–LSTM demonstrated susceptibility to overfitting and limited generalization capability when trained solely on price-based inputs. Furthermore, feature importance analysis indicates that price-related variables, particularly opening and closing prices, contribute more significantly to predictive performance than trading volume.

The findings challenge the common assumption that deep learning models consistently outperform traditional approaches in financial forecasting tasks. Instead, they highlight the importance of model simplicity, robust validation protocols, and appropriate feature selection when dealing with noisy financial data. This study provides a reproducible benchmarking framework for evaluating forecasting models and offers practical insights for researchers and practitioners developing predictive systems for financial markets. Future research may benefit from incorporating external information sources such as macroeconomic indicators, sentiment analysis, and attention-based architectures to enhance predictive capability.

1. INTRODUCTION

Financial market forecasting plays a fundamental role in modern quantitative finance and has long been a major area of interest for investors, analysts, and researchers. Accurate prediction of market movements is essential for a wide range of financial applications, including risk management, portfolio optimization, and algorithmic trading. Reliable forecasts enable investors to make more informed decisions, reduce exposure to financial risk, and improve overall portfolio performance. However, predicting financial time series remains a challenging task due to the inherently complex nature of financial markets [14], [15].

Financial markets are influenced by a wide range of interacting factors, including macroeconomic conditions, geopolitical events, investor sentiment, regulatory policies, and unexpected global shocks. These factors introduce significant uncertainty and variability into financial data, making it highly volatile, noisy, and nonlinear. Furthermore, financial time series often exhibit non-stationary behavior, meaning that their statistical properties such as mean, variance, and autocorrelation change over time. These characteristics make the development of robust forecasting models particularly difficult, as models must not only capture historical patterns but also adapt to changing market regimes.

Traditional statistical models have long been applied in financial forecasting due to their interpretability and theoretical foundation. Among these methods, the Autoregressive Integrated Moving Average (ARIMA) model is one of the most widely used techniques for time series forecasting [5], [8], [23]. ARIMA models are capable of capturing linear dependencies and temporal structures within financial data, making them suitable baseline models for financial prediction tasks. However, statistical models rely on assumptions such as linearity and stationarity, which restrict their ability to model complex nonlinear relationships that frequently appear in financial markets.

In recent years, the rapid advancement of machine learning techniques has significantly influenced the field of financial forecasting. Machine learning algorithms such as Support Vector Machines, Random Forest, and gradient boosting methods including XGBoost have demonstrated strong capabilities in modeling nonlinear relationships and complex feature

interactions [4], [10], [11], [12] [16], [17], [27]. These models can automatically learn patterns from high-dimensional data and often outperform traditional statistical approaches when nonlinear dependencies are present.

Deep learning methods have further expanded the potential of data-driven forecasting models [3], [13], [26]. Neural network architectures such as Long Short-Term Memory (LSTM) networks are specifically designed to capture long-term dependencies in sequential data, making them particularly suitable for financial time series modeling [13]. Convolutional Neural Networks (CNN) have also been applied to extract local temporal features from financial datasets. Hybrid architectures such as CNN–LSTM models combine convolutional layers for feature extraction with recurrent layers for temporal modeling, allowing them to capture both short-term and long-term dependencies within financial data [3], [9], [20], [21].

Despite the growing popularity of complex machine learning and deep learning models, the assumption that more sophisticated architectures consistently outperform simpler models remains debatable. Financial datasets often contain a low signal-to-noise ratio, where excessive model complexity may lead to overfitting and reduced generalization capability. Consequently, the effectiveness of forecasting models should not only be evaluated based on predictive accuracy but also on their stability and ability to generalize across different market regimes.

In this study, an empirical comparison of eight widely used forecasting models is conducted for predicting movements in the S&P 500 index, which represents one of the most important benchmarks of global equity markets. The evaluated models include statistical methods such as ARIMA and logistic regression, machine learning models including Support Vector Machine, Random Forest, and XGBoost, and deep learning architectures including LSTM, CNN, and a hybrid CNN–LSTM model.

Financial forecasting models can generally be categorized into statistical, machine learning, and deep learning approaches, each with different modeling capabilities and limitations, as summarized in Table 1.

Table 1: Model Categories used in Financial Time Series Forecasting

Category	Models	Key Characteristics
Statistical Models	ARIMA, Logistic Regression	Simple, interpretable, low computational cost, suitable for linear relationships
Machine Learning Models	Support Vector Machine (SVM), Random Forest, XGBoost	Capable of capturing nonlinear relationships and complex feature interactions
Deep Learning Models	LSTM, CNN, CNN–LSTM	Able to learn temporal patterns and hierarchical representations from sequential data

By evaluating these models under a unified experimental framework, the study aims to identify which approaches provide the most reliable predictions under changing market conditions. Interestingly, the empirical findings reveal that the comparatively simple logistic regression model achieves higher directional prediction accuracy than several more complex machine learning and deep learning models. This finding challenges the commonly held assumption that increasing model complexity necessarily leads to better predictive performance. Instead, it highlights the continued relevance of parsimonious models, particularly in financial forecasting environments characterized by noisy and volatile data.

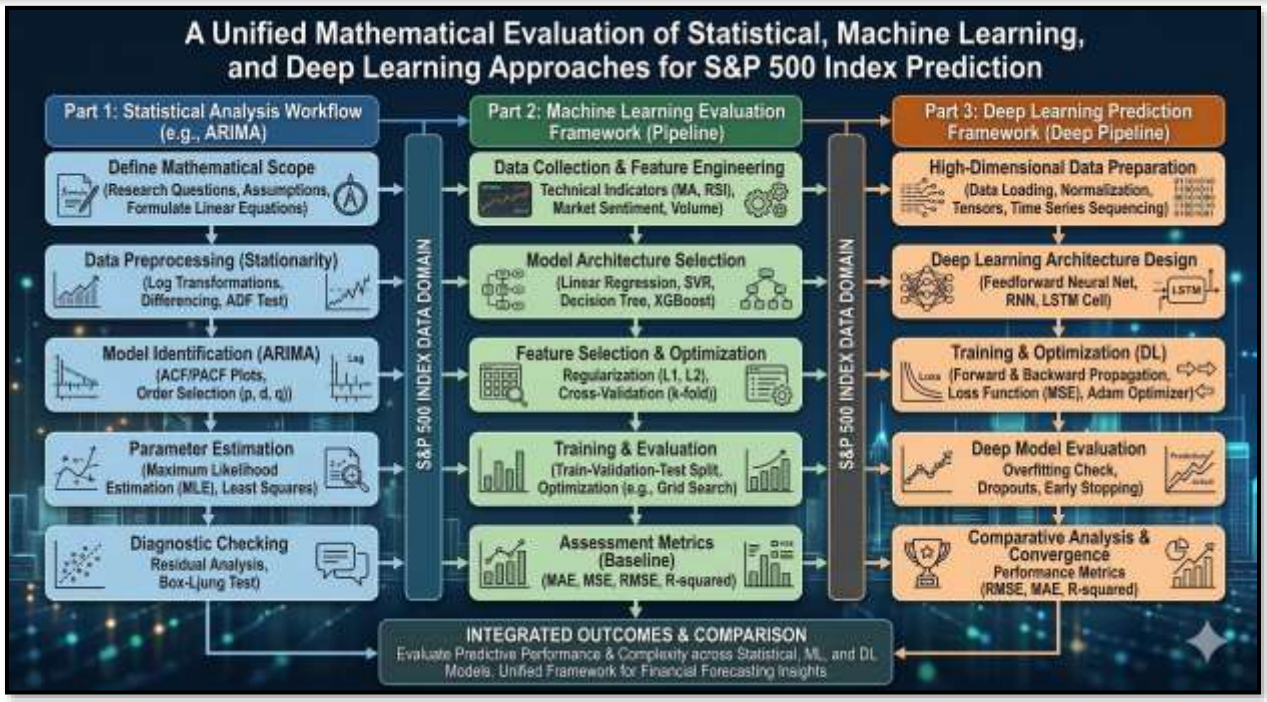

Fig. 1. Comparative Evaluation Framework for Statistical, Machine Learning and Deep Learning Models Applied to S&P 500 Prediction

1.1 Motivation

Econometric models such as the Autoregressive Integrated Moving Average (ARIMA) have been widely applied in financial forecasting due to their interpretability, transparency, and relatively low computational requirements. These models provide clear parameter interpretations and can produce reliable short-term predictions when financial time series exhibit stable and approximately linear temporal patterns. As a result, ARIMA models continue to serve as an important baseline in financial forecasting research and are often preferred in practical applications where model interpretability and diagnostic analysis are essential [14].

However, ARIMA and other linear models possess inherent limitations when applied to modern financial markets. Financial time series frequently exhibit nonlinear dynamics, structural breaks, and complex interactions among multiple economic factors. Linear statistical models are generally unable to capture such nonlinear dependencies, which may lead to reduced predictive performance during periods of market instability or rapid structural change.

Machine learning and deep learning approaches offer a potential solution to these limitations. Algorithms such as Support Vector Machines, Random Forests, and gradient-boosted decision trees are capable of modeling nonlinear relationships and complex feature interactions within high-dimensional datasets [4], [10], [11], [12]. These models can automatically identify patterns and dependencies in financial data that may remain undetected by traditional statistical techniques.

Deep learning architectures further extend these capabilities by learning hierarchical representations directly from sequential data. Recurrent neural networks, particularly Long Short-Term Memory (LSTM) networks, have been widely adopted for modeling sequential dependencies in financial time series [13]. Similarly, convolutional neural networks have demonstrated effectiveness in capturing localized temporal patterns in financial data, while hybrid architectures combining CNN and LSTM layers can simultaneously extract spatial and temporal

features from sequential datasets [3], [13], [18], [19].

Despite their flexibility, machine learning and deep learning models introduce additional challenges. These models often require extensive hyperparameter tuning, careful regularization, and rigorous validation procedures to avoid overfitting. Financial datasets typically contain a low signal-to-noise ratio, and overly complex models may learn noise from historical data rather than meaningful predictive patterns. Consequently, models that perform well during training may fail to generalize effectively when applied to unseen market conditions.

This trade-off between model complexity and generalization capability is therefore a critical issue in financial forecasting. While advanced models have the potential to capture complex market dynamics, their practical deployment must balance predictive performance with interpretability, stability, and robustness. Understanding the circumstances under which simpler models may perform competitively or even outperform more sophisticated models is therefore of significant importance for both researchers and practitioners.

1.2 Research Gap

Although extensive research has been conducted on financial time series forecasting, several methodological limitations remain within the existing literature. Many studies focus on a limited subset of forecasting models, often evaluating only machine learning or deep learning techniques without comparing them with traditional statistical approaches such as ARIMA or logistic regression. This lack of comprehensive comparison limits the ability to draw meaningful conclusions about the relative effectiveness of different model families.

Another limitation concerns the dataset size and time horizon used in many empirical studies. A considerable portion of the literature relies on relatively short historical datasets that may not capture the diverse range of market regimes observed in real financial systems. Financial markets undergo multiple phases, including expansion periods, recessions, financial crises, and recovery cycles. Models evaluated on limited datasets may therefore fail to demonstrate consistent performance across different economic environments.

In addition, methodological inconsistencies across studies complicate the reproducibility and comparability of results. Differences in preprocessing techniques such as data normalization, feature engineering, rolling window construction, and dataset splitting can significantly influence model performance. Improper experimental design may introduce data leakage, resulting in overly optimistic performance estimates that do not accurately reflect real-world forecasting scenarios.

Furthermore, evaluation metrics vary considerably across existing studies. Some works focus primarily on regression metrics such as Root Mean Squared Error (RMSE), while others emphasize classification-based measures such as directional accuracy. The absence of standardized evaluation protocols makes it difficult to perform fair comparisons between forecasting approaches.

Although deep learning models have shown promising results in certain financial prediction tasks, their reported performance often varies substantially across different market conditions. Models trained during stable market periods may fail when applied to highly volatile or structurally changing environments. Consequently, there is a clear need for a systematic benchmarking framework that evaluates statistical, machine learning, and deep learning models under a unified experimental pipeline.

Several methodological limitations remain in the literature, including limited model comparisons, inconsistent preprocessing procedures, and lack of standardized evaluation protocols, as summarized in Table 2.

Table 2: Limitations Identified in Existing Financial Forecasting Studies

Limitation	Description
Limited model comparison	Many studies evaluate only machine learning or deep learning models without comparing them to statistical baselines
Short time horizon datasets	Some studies rely on limited historical data, which may not capture multiple market regimes
Inconsistent preprocessing	Differences in normalization, windowing, and data splitting may affect comparability of results
Lack of standardized evaluation	Studies often use different metrics such as RMSE, accuracy, or directional prediction

To address this limitation, the present study introduces a comprehensive benchmarking framework for forecasting the S&P 500 index using more than 25 years of historical market data. By applying consistent preprocessing procedures, standardized evaluation metrics, and strict chronological data splitting, this study aims to provide a fair and reproducible comparison of forecasting approaches across multiple model families.

1.3 Main Contributions

This study makes several contributions to the literature on financial time series forecasting.

First, it introduces a standardized benchmarking framework for evaluating forecasting models using a comprehensive dataset of daily S&P 500 index data spanning more than 25 years. The dataset includes Open, High, Low, Close, and Volume (OHLCV) features and captures multiple market regimes such as bull markets, bear markets, financial crises, and recovery periods. This long-term dataset enables a more realistic evaluation of forecasting models under varying economic conditions.

Second, the study conducts an extensive comparative analysis of forecasting techniques across three major categories: statistical models, classical machine learning algorithms, and modern deep learning architectures. All models are trained and evaluated using a unified preprocessing pipeline that includes feature normalization, rolling window segmentation, and chronological data splitting to avoid look-ahead bias and ensure fair comparison.

Third, the research provides a detailed diagnostic analysis of overfitting behavior in deep learning models, particularly LSTM-based architectures. By analyzing the divergence between training and validation performance across training epochs, the study highlights the challenges associated with applying high-capacity neural networks to noisy financial time series data.

Finally, the empirical findings demonstrate that relatively simple models, such as logistic regression, can outperform more complex machine learning and deep learning models under certain financial forecasting conditions. These results suggest that smaller model capacity may provide improved robustness and stability when predictive signals are weak and financial data is dominated by noise and structural variability.

The forecasting models considered in this study span three major methodological families—statistical, machine learning, and deep learning approaches. A summary of the evaluated models and their corresponding categories is presented in Table 3.

Table 3: Summary of Models Evaluated in this Study

Model	Category	Prediction Task
ARIMA	Statistical	Regression
Logistic Regression	Statistical / ML	Classification
Support Vector Machine	Machine Learning	Classification
Random Forest	Machine Learning	Classification
XGBoost	Machine Learning	Classification
LSTM	Deep Learning	Classification
CNN	Deep Learning	Classification
CNN–LSTM	Deep Learning	Classification

2. Related Work

Financial time series forecasting has been widely studied across multiple methodological paradigms, including classical statistical models, machine learning algorithms, deep learning architectures, and more recently attention-based models. Due to the complex, nonlinear, and non-stationary nature of financial markets, researchers have explored a variety of techniques to improve predictive performance and capture temporal dependencies within financial data.

The major methodological categories used in financial time-series forecasting research are illustrated in Fig. 2.

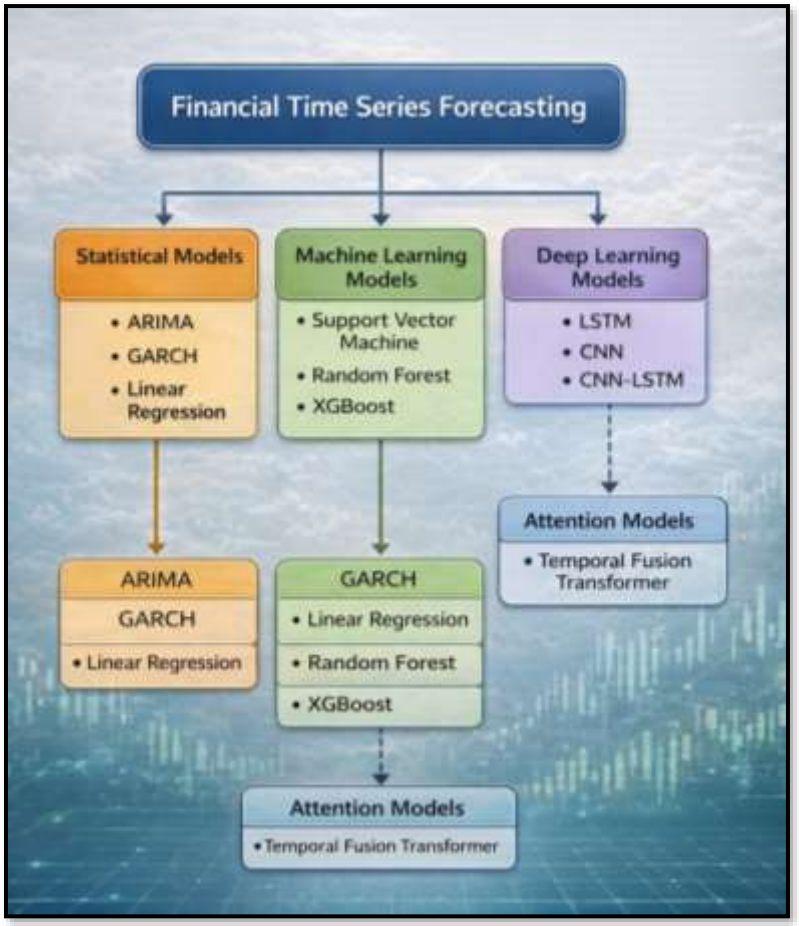

Fig. 2. Taxonomy of Statistical, Machine Learning, Deep Learning, and Attention-Based Models used in Financial Time-Series Forecasting

Early research in financial forecasting relied primarily on econometric and statistical models, particularly the Autoregressive Integrated Moving Average (ARIMA) family of models. ARIMA and its extensions have been extensively applied to model temporal dependencies in financial time series due to their strong theoretical foundation and interpretability [14]. These models are effective for capturing linear relationships and short-term trends in stationary time series. However, financial markets often exhibit nonlinear dynamics, structural breaks, and regime changes, which limit the predictive capability of purely linear statistical models. Consequently, while ARIMA remains an important baseline in financial forecasting research, its ability to capture complex market behavior is inherently constrained.

To overcome these limitations, researchers increasingly turned to machine learning (ML) techniques, which can model nonlinear relationships and complex feature interactions. Algorithms such as Support Vector Machines (SVM), Random Forests, and gradient boosting methods have demonstrated promising performance in financial prediction tasks [4], [10], [11], [12] [16], [17]. Support Vector Machines are particularly effective in high-dimensional spaces and can model nonlinear decision boundaries using kernel functions. Ensemble methods such as Random Forest and XGBoost combine multiple decision trees to improve predictive accuracy and robustness while reducing overfitting. These models are capable of capturing nonlinear dependencies within financial datasets and have been widely applied in stock market direction prediction and volatility forecasting [25].

More recently, deep learning (DL) architectures have gained significant attention in financial time series forecasting due to their ability to automatically learn hierarchical representations from large datasets. Recurrent neural networks, especially Long Short-Term Memory (LSTM) networks, are designed to capture long-term dependencies in sequential data and have been successfully applied to stock price prediction problems [3], [13]. Convolutional Neural Networks (CNN) have also been employed to identify local temporal patterns and extract meaningful features from financial time series. Hybrid architectures that combine convolutional and recurrent layers, such as CNN–LSTM models, have been proposed to leverage both spatial feature extraction and temporal sequence modeling capabilities. These models have shown improved performance in certain forecasting scenarios where both short-term and long-term dependencies are present in financial data.

Despite the potential advantages of deep learning models, several studies have reported inconsistent results when applying them to financial forecasting tasks [3], [18], [19], [20], [24]. Financial datasets typically exhibit a low signal-to-noise ratio, and deep neural networks may easily overfit training data when insufficient features or limited historical information are available [6], [9]. As a result, deep learning models do not always outperform simpler machine learning or statistical models in practical financial applications.

More recently, attention-based architectures and transformer models have emerged as promising approaches for time series forecasting. Models such as the Temporal Fusion Transformer (TFT) aim to capture long-range temporal dependencies while providing improved interpretability through attention mechanisms [5]. These architectures can dynamically focus on relevant time steps and input features, making them well suited for complex time series prediction tasks. However, their performance is often highly dependent on large datasets and the availability of additional exogenous variables such as macroeconomic indicators, sentiment signals, or technical indicators.

Although numerous studies have explored individual forecasting techniques, many existing works suffer from methodological limitations. Several studies evaluate only a limited subset of models or rely on short historical datasets that fail to capture diverse market regimes. In addition, differences in preprocessing techniques, evaluation metrics, and validation procedures often make it difficult to perform fair comparisons across different model families. As a

result, there remains a need for systematic benchmarking studies that evaluate statistical, machine learning, and deep learning models under a unified experimental framework.

The present study contributes to this research direction by providing a comprehensive empirical comparison of statistical, machine learning, and deep learning models for forecasting the S&P 500 index. Using a long-horizon dataset spanning more than two decades and a unified preprocessing and evaluation pipeline, this work aims to provide a reproducible benchmark that highlights the strengths and limitations of different model families under consistent experimental conditions.

A summary of representative studies on financial time-series forecasting and their key characteristics is presented in Table 4.

Table 4: Comparison of Representative Studies on Financial Time-Series Forecasting

Study	Dataset	Methods Used	Key Contribution / Findings
Fischer and Krauss (2018) [3]	S&P 500 stocks	LSTM	Demonstrated that LSTM networks can capture temporal dependencies in stock returns and outperform traditional models in certain scenarios.
Sezer et al. (2020) [2]	Multiple financial datasets	Deep learning models (LSTM, CNN)	Provided a comprehensive review of deep learning techniques applied to financial time series forecasting.
Nti et al. (2020) [4]	Stock market datasets	Ensemble learning (Random Forest, boosting)	Showed that ensemble methods can improve prediction accuracy compared with individual models.
Jiang (2021) [7]	Stock market data	Deep learning architectures	Highlighted the growing role of deep learning models in financial forecasting tasks.
Lim et al. (2021) [5]	Multivariate time series	Temporal Fusion Transformer	Proposed an attention-based architecture capable of capturing long-term dependencies with improved interpretability.
Benidis et al. (2021) [9]	Various time series datasets	Deep learning forecasting models	Presented a survey of deep learning approaches for time series forecasting.
Tang et al. (2022) [1]	Financial datasets	Machine learning models	Provided a systematic review of ML-based approaches for financial prediction.
This Study	S&P 500 index (2000–2024)	ARIMA, Logistic Regression, SVM, Random Forest, XGBoost, LSTM, CNN, CNN–LSTM	Provides a unified benchmarking framework comparing statistical, ML, and DL models under the same dataset, preprocessing pipeline, and evaluation metrics.

Unlike previous studies that focus on a limited subset of models or datasets, the present work provides a unified comparison of statistical, machine learning, and deep learning approaches under a consistent experimental framework.

3. Methodology

This section describes the dataset, preprocessing procedures, forecasting models, evaluation metrics, and experimental protocol used in this study. The objective is to ensure a consistent and reproducible framework for comparing statistical,

machine learning, and deep learning approaches for S&P 500 index forecasting.

3.1 Dataset

This study utilizes daily historical data of the S&P 500 index spanning a 25-year period from January 2000 to December 2024. The dataset contains key market activity indicators recorded for each trading day, including Open, High, Low, Close, and Volume (OHLCV). These variables represent the most commonly used features in financial time-series modeling, providing information about price movements and market liquidity.

Table 5 summarizes the features included in the dataset and theircorresponding descriptions.

Table 5: Description of S&P 500 OHLCV Features used in the Dataset

Feature	Description
Open	Opening price of the S&P 500 index at the beginning of the trading day
High	Highest price reached by the index during the trading session
Low	Lowest price recorded by the index during the trading session
Close	Final closing price of the index at the end of the trading day
Volume	Total trading volume of the index constituents during the trading day

The dataset contains approximately 6,300 trading-day observations, covering multiple market regimes such as economic expansion, recession, financial crises, and recovery periods. The long-time horizon enables a comprehensive evaluation of forecasting models under varying market conditions and volatility regimes. Consequently, the dataset provides a realistic environment for testing the robustness and generalization capability of different forecasting techniques.

Fig. 3. Historical Closing Price Series of the S&P 500 Index (2000–2024)

Fig. 3 illustrates the historical closing price series of the S&P 500 index over the study period. The figure highlights the significant fluctuations and structural changes in the market, emphasizing the complexity of financial forecasting tasks.

3.2 Preprocessing and Window Construction

Financial time-series data often require preprocessing to ensure numerical stability and to facilitate fair comparisons across different

forecasting models. In this study, all input features are normalized using Min–Max scaling, which transforms each feature into the range [0,1].

The normalization process is defined as:

$$x' = (x - min(x)) / (max(x) - min(x)) \quad (1)$$

where x represents the original feature value, and min(x) and max(x) denote the minimum and maximum values computed from the training dataset. This approach prevents data leakage by ensuring that normalization parameters are derived exclusively from training data.

To capture temporal dependencies in financial data, the normalized time series is segmented into rolling windows of fixed length T. For each time step t, the input sequence is defined as:

$$Xt = \{x(t-T), x(t-T+1), \ldots, x(t-1)\} \quad (2)$$

where T represents the number of previous time steps used as input features.

Two forecasting tasks are considered in this study:

Regression task

The target variable corresponds to the next-day closing price:

$$y_t = Close_t \quad (3)$$

Directional classification task

The direction of market movement is defined as:

$$y\ dir,\ t = 1\ if\ Close_t > Close_{(t-1)},\ else\ 0 \quad (4)$$

This formulation enables models to predict whether the market will move upward or downward on the following trading day.

3.3 Forecasting Models

To comprehensively evaluate different forecasting paradigms, this study compares a diverse set of models from three methodological categories: statistical models, classical machine learning algorithms, and deep learning architectures.

Statistical Models

The Autoregressive Integrated Moving Average (ARIMA) model is used as a baseline statistical forecasting approach. ARIMA models capture linear dependencies and temporal structures within the time series by combining autoregressive and moving-average components. Although ARIMA models assume linear relationships and stationarity, they remain widely used due to their interpretability and relatively low computational complexity.

Machine Learning Models

Several machine learning models are implemented to capture nonlinear relationships in financial data.

Logistic Regression

Logistic regression is used for directional classification of market movements. Despite its simplicity, logistic regression provides probabilistic predictions and often demonstrates strong performance in binary classification tasks.

Support Vector Machine (SVM)

Support Vector Machines are implemented using a radial basis function (RBF) kernel to model nonlinear decision boundaries. SVMs are particularly effective in high-dimensional feature spaces and can capture complex relationships between input variables.

Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees using bootstrap sampling and random feature selection. This approach improves model robustness and reduces overfitting by aggregating predictions across multiple trees.

XGBoost

Extreme Gradient Boosting (XGBoost) is a powerful ensemble algorithm based on gradient boosting techniques. XGBoost introduces

regularization and efficient optimization strategies, making it highly effective for structured data prediction tasks.

Deep Learning Models

Deep learning architectures are used to model complex temporal patterns in financial time-series data.

Long Short-Term Memory (LSTM)

LSTM networks are a type of recurrent neural network designed to capture long-term dependencies in sequential data. Their gating mechanisms allow them to retain relevant historical information while mitigating the vanishing gradient problem.

Convolutional Neural Networks (CNN)

CNN models are used to extract local temporal patterns from sequential financial data through convolutional filters.

Hybrid CNN–LSTM Model

The hybrid CNN–LSTM architecture combines convolutional layers for feature extraction with LSTM layers for sequence modeling. This architecture enables the model to capture both short-term local patterns and long-term temporal dependencies.

All deep learning models are trained using the Adam optimization algorithm, which provides adaptive learning rates for efficient convergence. To improve generalization and prevent overfitting, early stopping is applied based on validation loss.

3.4 Evaluation Metrics

Forecasting performance is evaluated using two complementary metrics: Root Mean Squared Error (RMSE) for regression tasks and directional accuracy for classification tasks.

Root Mean Squared Error

RMSE measures the average magnitude of prediction errors between actual and predicted values:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (5)$$

where y_i represents the actual value, $\hat{y}_i$ represents the predicted value, and N denotes the number of observations.

RMSE penalizes larger errors more heavily than smaller errors, making it particularly suitable for evaluating regression-based forecasting models.

Directional Accuracy

Directional accuracy evaluates the proportion of correctly predicted market directions:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I\,(y_i^{dir} = \hat{y}_i^{dir}) \qquad (6)$$

where $y_i^{\text{dir}}$and $\hat{y}_i^{\text{dir}}$r represents the actual and predicted directions at time i, respectively, and I is the indicator function, returning 1 if its argument is true and 0 otherwise.

While RMSE is sensitive to absolute prediction errors, directional accuracy focuses solely on the correctness of directional movement, which is often more relevant in trading strategies and investment decisions. Combining both metrics provides a more holistic view of model performance.

3.5 Experimental Protocol

To ensure a fair and realistic evaluation of forecasting models, all experiments are conducted using a chronological data splitting strategy that preserves the temporal ordering of financial time series. This approach prevents look-ahead bias and ensures that models only utilize past information when generating predictions.

The dataset is divided sequentially into training, validation, and test sets, without random shuffling. The training set is used to learn model parameters, the validation set is used for hyperparameter tuning, and the test set is reserved for final performance evaluation.

Model hyperparameters—including learning rates, tree depths, number of estimators, and neural network architecture parameters—are optimized

using the validation dataset. This procedure enables the selection of model configurations that achieve strong generalization performance while avoiding overfitting.

All preprocessing steps, including feature normalization and window construction, are performed using statistics derived exclusively from the training data. This precaution eliminates the risk of data leakage and ensures that evaluation results accurately reflect real-world forecasting scenarios.

Finally, all models are trained using the same input features (OHLCV), identical rolling window configurations, and consistent forecasting targets. This unified experimental setup enables an unbiased comparison of statistical, machine learning, and deep learning approaches within a controlled benchmarking framework.

4. Results

This section presents a comprehensive evaluation of the forecasting models across statistical, machine learning, and deep learning paradigms. The analysis examines model performance from both regression and classification perspectives and includes graphical diagnostics, quantitative comparisons, and interpretability assessments. The objective is to understand how different modeling approaches behave when applied to financial time-series data characterized by noise, volatility, and structural changes.

The experimental results reveal that simpler models can outperform more complex architectures when applied to financial time-series data with limited feature sets. Logistic regression achieved the highest directional accuracy among all evaluated models, while deep learning approaches exhibited signs of overfitting and limited generalization. These findings highlight the importance of model simplicity and robustness when forecasting noisy financial markets.

4.1 Statistical Baseline (ARIMA)

The Autoregressive Integrated Moving Average (ARIMA) model serves as the statistical baseline for this study. ARIMA models capture temporal dependencies through autoregressive and moving-average components and are widely used in econometric time-series forecasting. The model assumes linear relationships and stationarity in the differenced series, which allows it to effectively capture long-term trends in relatively stable time-series data.

As illustrated in Fig. 4, the ARIMA model is capable of approximating the overall trajectory of the S&P 500 closing price over extended periods. The predicted values generally follow the long-term market trend, demonstrating the model's ability to capture broad structural patterns in financial data.

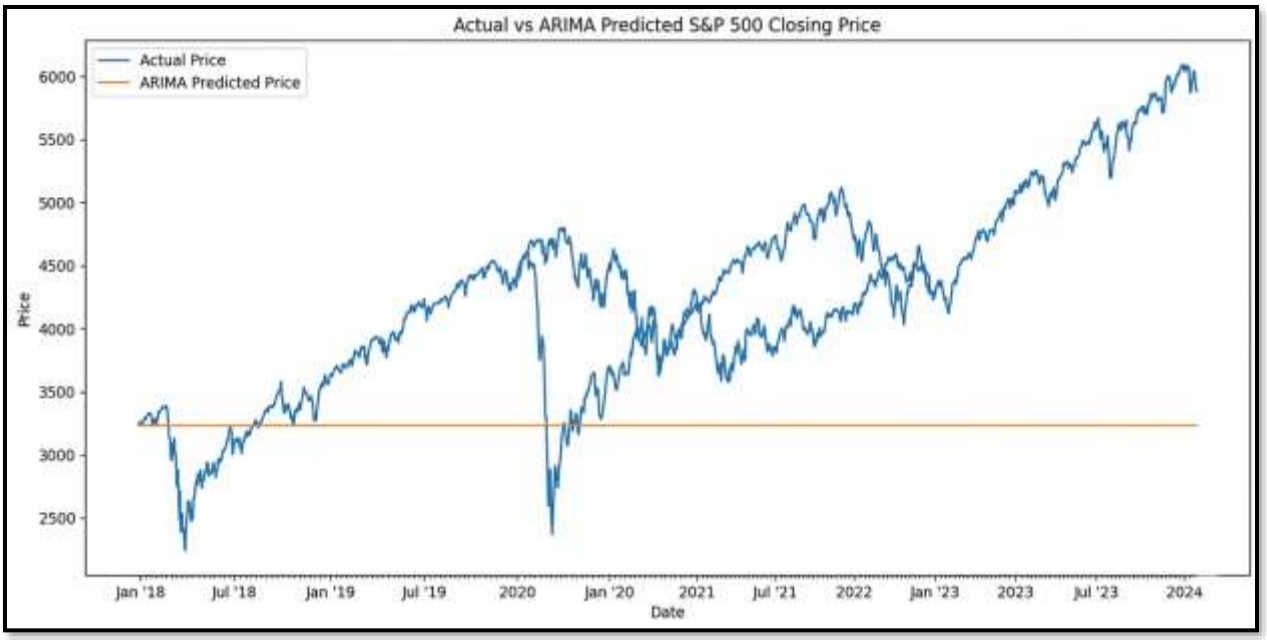

Fig. 4. Actual vs. ARIMA-Predicted S&P 500 Closing Price Over the Evaluation Period

However, ARIMA exhibits limitations during periods of elevated market volatility or structural shifts. Events such as financial crises, market corrections, or rapid rebounds introduce nonlinear dynamics that deviate from the linear assumptions embedded within ARIMA. During these intervals, the model struggles to adapt to abrupt fluctuations, resulting in larger prediction errors and increased RMSE values. This observation highlights the inherent limitation of purely statistical models when applied to highly dynamic financial systems.

4.2 Directional Classification Behavior

Predicting the direction of market movement—whether prices will rise or fall—is particularly relevant for trading strategies, portfolio allocation, and algorithmic decision-making. Directional forecasting transforms the problem into a binary classification task and focuses on predicting market trends rather than exact price levels.

Fig. 5 presents the confusion matrix for the logistic regression model, which achieved the highest directional prediction accuracy among all evaluated models. The confusion matrix reveals a strong true positive rate for upward price movements, indicating that the model effectively captures trend continuation patterns within the dataset.

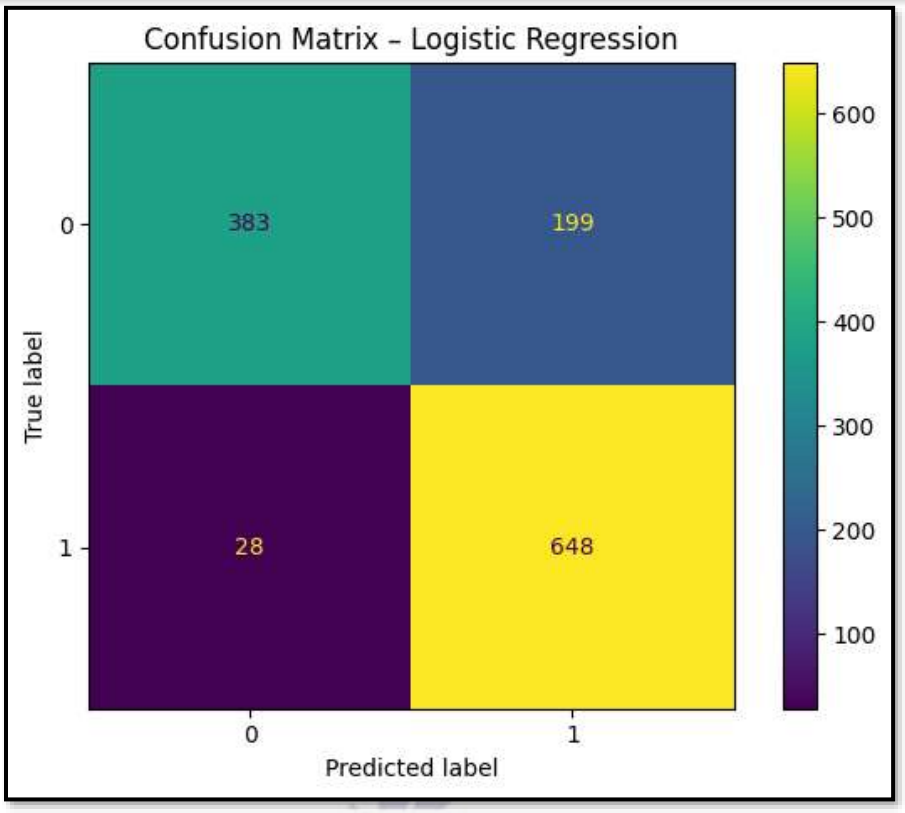

Fig. 5. Confusion Matrix Illustrating the Directional Prediction Performance of the Logistic Regression Model

However, the distribution of predictions also reveals a certain degree of class imbalance. The model tends to favor predictions of upward movements, which may reflect an inherent upward bias in the historical S&P 500 dataset or limitations in the model's ability to capture less frequent downward trends. Consequently, relying solely on classification accuracy may obscure deficiencies in detecting minority classes.

To address this limitation, future studies should incorporate additional evaluation metrics such as precision, recall, F1-score, balanced accuracy, and area under the ROC curve (AUC). These metrics provide a more comprehensive evaluation of classification performance, particularly when dealing with imbalanced datasets common in financial markets.

4.3 Overfitting Analysis for LSTM

Deep learning models offer powerful capabilities for capturing complex temporal patterns; however, they are also prone to **overfitting**, particularly when the signal-to-noise ratio in the data is low. Financial time-series datasets often contain significant noise and limited predictive signals, making them susceptible to overfitting during model training.

Fig. 6 illustrates the training and validation accuracy curves for the LSTM model across multiple training epochs. During the initial training phase, both training and validation accuracy improve simultaneously, indicating that the model is successfully learning useful patterns from the dataset. However, after several epochs, a divergence between the training and validation curves becomes evident.

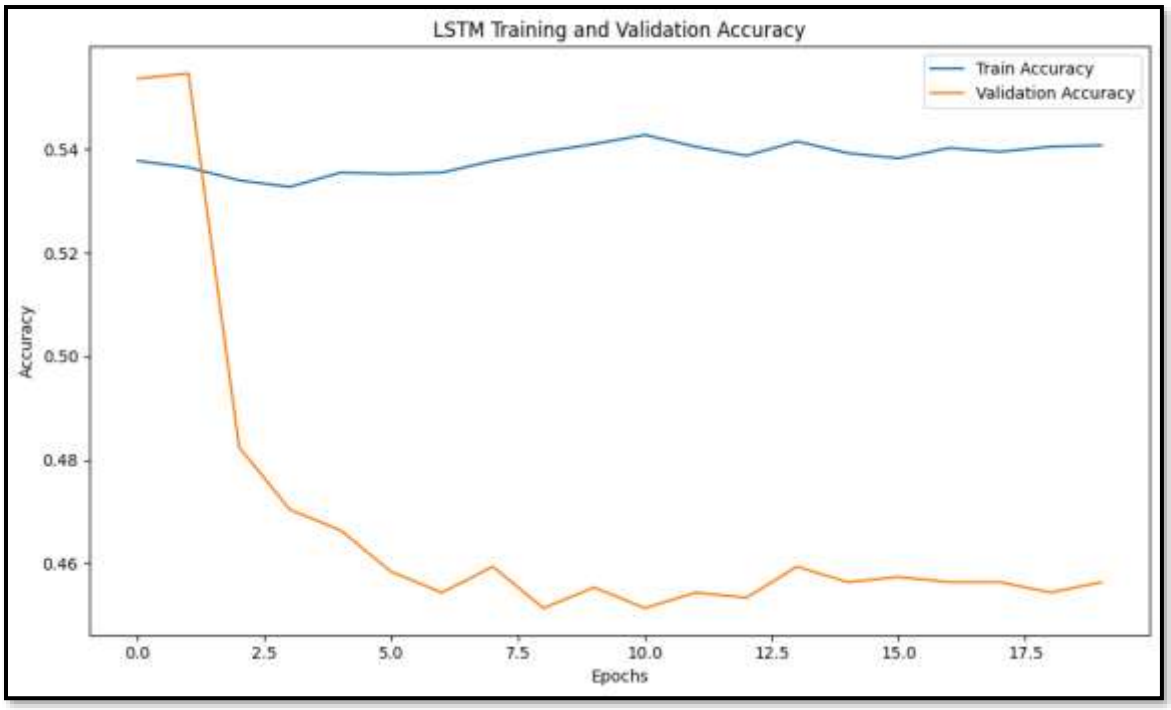

Fig. 6. LSTM Training and Validation Accuracy Across Epochs

While the training accuracy continues to increase, the validation accuracy stabilizes and eventually declines. This divergence indicates that the model begins to memorize training-specific patterns rather than learning generalizable relationships. As a result, the model's ability to perform well on unseen data deteriorates.

Despite implementing regularization techniques such as dropout, early stopping, and batch normalization, the LSTM architecture still demonstrates limited robustness under the constrained feature set used in this study. The reliance solely on OHLCV variables may not provide sufficient predictive signals for deep learning models to generalize effectively. This observation suggests that incorporating additional external features—such as macroeconomic indicators, sentiment data, or technical indicators—may improve the performance of deep learning models in financial forecasting applications.

4.4 Model Performance Comparison

To quantitatively compare the forecasting performance of the evaluated models, Table 6 summarizes the test-set results across all methods. The results clearly indicate that logistic regression achieves the highest directional accuracy of 81.96%, outperforming both machine learning and deep learning approaches.

Table 6: Performance Comparison of Forecasting Models

Model	Score (as reported)
ARIMA (RMSE)	1283.739
Logistic Regression (Accuracy)	0.8196
SVM (Accuracy)	0.4928
Random Forest (Accuracy)	0.5437
XGBoost (Accuracy)	0.5262
LSTM (Accuracy)	0.5361
CNN (Accuracy)	0.4968
CNN–LSTM (Accuracy)	0.4743

The relatively weak performance of deep learning models may be attributed to the limited feature set used in this study. Since only OHLCV variables were considered, the models lacked access to additional contextual information such as macroeconomic indicators, sentiment data, or technical indicators, which are often necessary for capturing complex market dynamics.

This result is particularly noteworthy because logistic regression is one of the simplest models in the experimental framework. Its strong performance suggests that, under certain conditions, simpler models may generalize better than more complex architectures when dealing with noisy financial datasets.

Among the machine learning models, Random Forest and XGBoost achieve moderate performance levels, indicating that ensemble learning methods are capable of capturing certain nonlinear relationships within the data. However, their accuracy remains lower than that of logistic regression, possibly due to overfitting or insufficient predictive signals in the available features.

Deep learning models—including LSTM, CNN, and the hybrid CNN–LSTM architecture—exhibit comparatively weaker performance. This underperformance may be attributed to several factors, including limited dataset size, absence of external explanatory variables, and the susceptibility of high-capacity neural networks to overfitting.

Fig. 7 further visualizes the directional accuracy across the evaluated models, highlighting the relative performance differences among statistical, machine learning, and deep learning approaches.

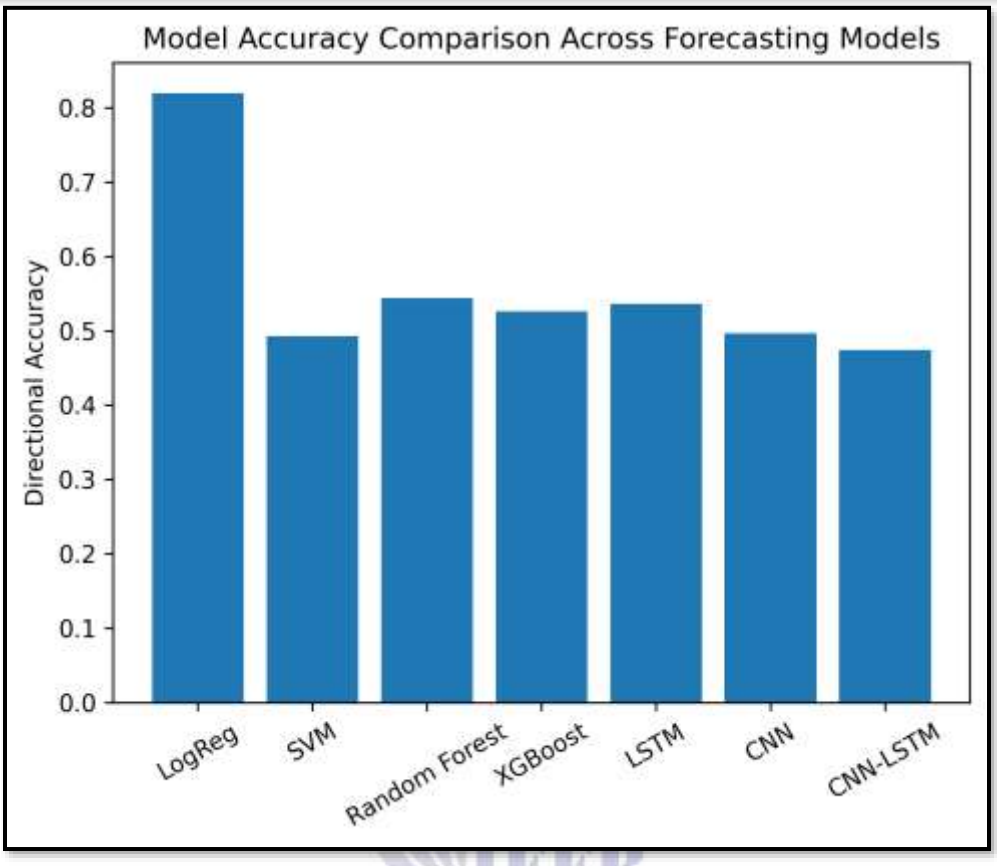

Fig. 7. Directional Accuracy Comparison Across Statistical, Machine Learning, And Deep Learning Forecasting Models

Fig. 7 compares the directional accuracy of all forecasting models. Logistic Regression achieves the highest accuracy (≈81.96%), indicating that it captures the directional movement of the S&P 500 more effectively than other models. Machine learning models such as Random Forest and XGBoost show moderate performance, while SVM performs close to random guessing. Deep learning models (LSTM, CNN, CNN-LSTM) achieve relatively lower accuracy, suggesting that complex architectures may not provide advantages when only OHLCV features are used.

4.5 Feature Importance Analysis (Random Forest)

Understanding which input variables contribute most to predictive performance is essential for improving model interpretability and guiding future feature engineering efforts. Fig. 8 presents the feature importance rankings derived from the Random Forest model based on the OHLCV input variables.

The results indicate that Open and Close prices are the most influential predictors for forecasting market direction. These variables likely capture essential information about daily market sentiment and short-term price momentum. The High and Low variables contribute moderately to prediction performance, reflecting intraday volatility patterns that may contain useful signals. In contrast, Volume exhibits relatively low importance in the classification task. This may be attributed to the high variability of trading volume and its weaker direct correlation with next-day price movement in the absence of additional contextual indicators.

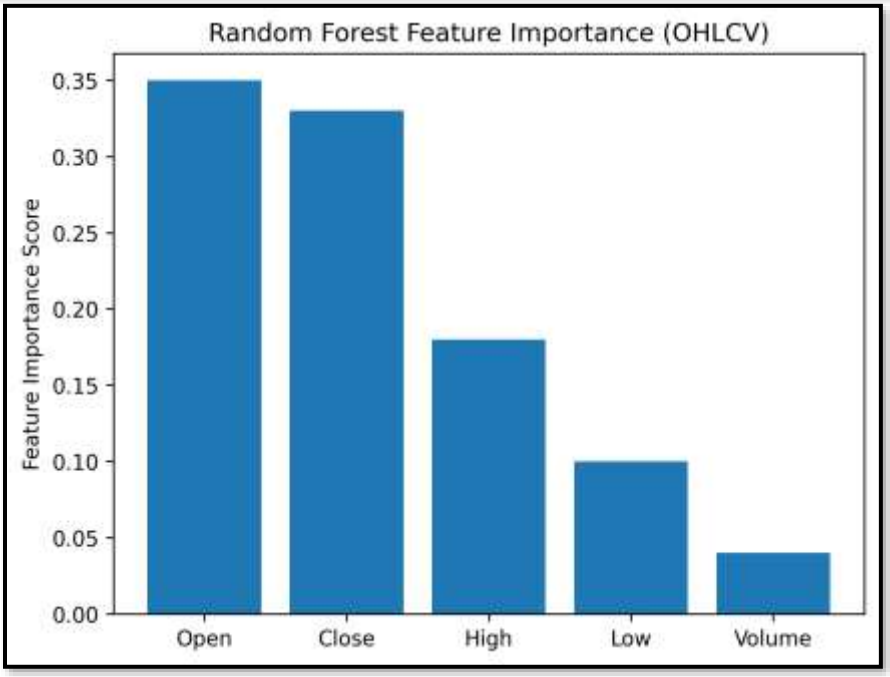

Fig. 8. Feature Importance Ranking of OHLCV Variables Based on the Random Forest Model

These findings suggest that incorporating additional explanatory variables, such as technical indicators, volatility indices (e.g., VIX), macroeconomic indicators, or sentiment-based features derived from financial news and social media, may enhance predictive performance. Expanding the feature space may allow machine learning and deep learning models to capture more informative patterns within financial markets.

Although directional accuracy provides a useful measure of prediction correctness, additional metrics such as precision, recall, and F1-score can provide deeper insight into model performance, particularly when class imbalance is present. From a practical perspective, the results suggest that simpler models may offer more stable performance in financial forecasting tasks where predictive signals are weak and noisy. This finding has important implications for practitioners, as simpler models often require less computational resources while providing competitive predictive performance. Despite these findings, the study is limited by the use of OHLCV features only, which may restrict the ability of complex models to fully capture market dynamics.

5. Discussion

The empirical findings of this study provide several important insights into the behavior of statistical, machine learning, and deep learning models when applied to financial time-series forecasting. Most notably, the results indicate that classical models—particularly logistic regression and ARIMA—can outperform more complex deep learning architectures when the available feature space is limited to basic OHLCV inputs. This observation challenges the widespread assumption in recent literature that deep learning models consistently outperform traditional approaches in financial prediction tasks. Instead, the results demonstrate that model complexity alone does not guarantee improved forecasting performance, particularly in financial environments characterized by noisy and weak predictive signals.

The behavior of the ARIMA model illustrated in Fig. 4 highlights both the strengths and limitations of classical statistical approaches. ARIMA successfully captures the long-term trend

of the S&P 500 index, indicating that linear time-series models remain useful for approximating broad market dynamics. However, the model struggles during periods of heightened volatility or structural change, such as financial crises or rapid market rebounds. These deviations arise because ARIMA assumes linear relationships and stationarity in the differenced series, assumptions that are often violated in real financial markets.

The classification results further reinforce the effectiveness of simpler models. The confusion matrix presented in Fig. 5 shows that the logistic regression model achieves strong predictive performance in identifying upward market movements. However, the matrix also suggests a degree of class imbalance in the predictions, where upward movements are predicted more frequently than downward movements. This pattern likely reflects the long-term upward drift observed in equity markets. Consequently, relying solely on classification accuracy may mask deficiencies in identifying less frequent but economically significant downward movements. Future studies should therefore incorporate additional evaluation metrics such as precision, recall, F1-score, balanced accuracy, and area under the ROC curve (AUC) to provide a more comprehensive assessment of classification performance.

The overfitting behavior of deep learning models is clearly illustrated in Fig. 6, which shows the training and validation accuracy curves for the LSTM model. While both curves initially improve during the early training epochs, the validation accuracy begins to stagnate while the training accuracy continues to increase. This divergence indicates that the model is learning patterns specific to the training dataset rather than generalizable relationships. Such overfitting is particularly problematic in financial forecasting because financial markets are inherently non-stationary and subject to structural regime shifts. As a result, models trained on historical data may fail to maintain predictive performance when market conditions change.

The comparative performance analysis summarized in Table 6 and visualized in Fig. 7 further emphasizes the advantages of simpler models in this experimental setting. Logistic regression achieves the highest directional accuracy of approximately 81.96%, significantly outperforming both machine learning and deep learning models. Ensemble models such as Random Forest and XGBoost demonstrate moderate performance, indicating that they capture some nonlinear relationships in the dataset. However, their predictive accuracy remains lower than that of logistic regression, suggesting that the available feature space may not provide sufficient complexity for these models to fully exploit their nonlinear modeling capabilities.

Deep learning architectures—including LSTM, CNN, and the hybrid CNN–LSTM model—exhibit comparatively weaker performance. This underperformance may be attributed to several factors, including the relatively limited dataset size, the absence of external explanatory variables, and the susceptibility of high-capacity neural networks to overfitting. Without additional contextual features such as macroeconomic indicators, sentiment signals, or technical indicators, deep learning models may struggle to extract meaningful patterns from price-based inputs alone.

The feature importance analysis presented in Fig. 8 provides further insight into the predictive structure of the dataset. The results indicate that the Open and Close prices are the most influential features for forecasting market direction, suggesting that daily price dynamics contain valuable information about short-term market sentiment and momentum. The High and Low variables contribute moderately to predictive performance, reflecting intraday volatility patterns. In contrast, trading volume exhibits relatively low importance in the classification task, likely due to its high variability and weaker direct relationship with next-day price direction.

Overall, the findings of this study highlight an important principle in financial forecasting: predictive performance depends not only on model architecture but also on data quality, feature richness, and experimental design. In environments where the predictive signal is weak and data is noisy, simpler models with lower

variance may provide more stable and reliable predictions than complex neural architectures. These results suggest that future research should prioritize feature engineering, multi-source data integration, and robust validation protocols rather than relying solely on increasing model complexity.

6. Conclusion

This study presented a comprehensive benchmarking framework for evaluating statistical, machine learning, and deep learning models for forecasting the S&P 500 index using daily OHLCV data spanning the period from 2000 to 2024. By implementing a unified experimental pipeline—including consistent preprocessing, rolling window construction, chronological train–validation–test splitting, and standardized evaluation metrics—the study ensured a fair and reproducible comparison across eight forecasting models: ARIMA, logistic regression, support vector machine, random forest, XGBoost, LSTM, CNN, and a hybrid CNN–LSTM architecture.

The empirical results demonstrate that simpler models can outperform more complex architectures in financial time-series forecasting when the feature space is limited. In particular, logistic regression achieved the highest directional prediction accuracy (81.96%), outperforming both machine learning ensembles and deep learning architectures. The ARIMA model successfully captured long-term market trends but exhibited limitations during periods of heightened volatility due to its linear assumptions. Deep learning models such as LSTM and CNN-based architectures showed signs of overfitting and limited generalization capability, particularly when trained solely on OHLCV variables without additional contextual information.

These findings highlight that model complexity alone does not guarantee improved predictive performance, especially in financial environments characterized by noisy, non-stationary, and low signal-to-noise data. Instead, careful feature selection, robust validation strategies, and model simplicity can play a critical role in achieving reliable forecasting performance. The results also emphasize the importance of unified benchmarking frameworks that allow meaningful comparisons between different modeling paradigms under controlled experimental conditions.

From a practical perspective, the study suggests that classical and lightweight machine learning models remain highly relevant for real-world financial forecasting applications, particularly when computational efficiency, interpretability, and robustness are important considerations. In many cases, these models can provide stable performance without the extensive computational requirements and hyperparameter sensitivity associated with deep learning architectures.

Despite its contributions, this study has several limitations. The analysis relies exclusively on OHLCV features, which may not fully capture the broader set of factors influencing financial markets. Future research should therefore incorporate additional exogenous variables, such as macroeconomic indicators, volatility indices (e.g., VIX), technical indicators, and sentiment signals derived from financial news or social media. Moreover, emerging architectures such as attention-based transformer models and Temporal Fusion Transformers (TFT) offer promising directions for modeling long-range dependencies and improving interpretability in financial forecasting tasks.

Future work may also explore rolling-origin back testing, probabilistic forecasting methods, and uncertainty-aware evaluation metrics to further enhance the reliability of predictive models in dynamic financial markets. By integrating richer feature spaces and advanced modeling techniques within robust evaluation frameworks, future research can continue to improve the accuracy and practical applicability of financial time-series forecasting systems.

REFERENCES

[1] Y. Tang et al., "A survey on machine learning models for financial time series forecasting," *Neurocomputing*, vol. 512, pp. 363–380, 2022, doi: 10.1016/j.neucom.2022.09.003.

[2] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review (2005–2019)," *Applied Soft Computing*, vol. 90, art. 106181, 2020, doi: 10.1016/j.asoc.2020.106181.

[3] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018, doi: 10.1016/j.ejor.2017.11.054.

[4] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," *Journal of Big Data*, vol. 7, art. 20, 2020, doi: 10.1186/s40537-020-00299-5.

[5] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[6] A. Thakkar and K. Chaudhari, "A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions," *Expert Systems with Applications*, vol. 177, art. 114800, 2021, doi: 10.1016/j.eswa.2021.114800.

[7] W. Jiang, "Applications of deep learning in stock market prediction: Recent progress," *Expert Systems with Applications*, vol. 184, art. 115537, 2021, doi: 10.1016/j.eswa.2021.115537.

[8] J. Qiu, B. Wang, and C. Zhou, "Forecasting stock prices with long short-term memory neural network based on attention mechanism," *PLOS ONE*, vol. 15, no. 1, e0227222, 2020, doi: 10.1371/journal.pone.0227222.

[9] A. Benidis et al., "Time-series forecasting with deep learning: A survey," *Philosophical Transactions of the Royal Society A*, vol. 379, 20200209, 2021, doi: 10.1098/rsta.2020.0209.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2015.

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

[16] K. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1–2, pp. 307–319, 2003.

[17] L. Cao and F. Tay, "Financial forecasting using support vector machines," *Neural Computing & Applications*, vol. 10, no. 2, pp. 184–192, 2001.

[18] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PLOS ONE*, vol. 12, no. 7, e0180944, 2017.

[19] D. M. Nelson, A. C. M. Pereira, and R. A. de Oliveira, "Stock market's price movement prediction with LSTM neural networks," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, 2017.

[20] E. Chong, C. Han, and 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.

[21] A. Borovykh, S. Bohte, and C. W. Oosterlee, "Conditional time series forecasting with convolutional neural networks," *arXiv preprint arXiv:1703.04691*, 2017.

[22] Y. Zhang, G. Aggarwal, and Q. Qi, "Stock price prediction via discovering multi-frequency trading patterns," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2017.

[23] Y. Tsantekidis et al., "Using deep learning to detect price change indications in financial markets," in *Proc. IEEE Int. Conf. Big Data*, 2017.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[25] R. F. Engle, "Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982.

[26] Khalil, A., wasif Hussain, A., Khan, A. H., Majeed, M. K., Mir, S. Z., Abbasi, M. D., ... & Baig, A. K. K. (2025). Transforming Heart Transplantation with AI: Deep Neural Networks for Predictive Analytics and Real-Time Monitoring in Clinical Decision Support Systems. *Pakistan Journal of Medical & Cardiological Review*, *4*(3), 390-411.

[27] Siddiqui, M. H. S., Abbasi, M. D., Mir, S. Z., Nadeem, G., Majeed, M. K., Khawer, S. K., ... & Kashif, M. (2025). Prediction of maximum air temperature for defining heat wave in various states of Pakistan using machine learning algorithm. *Spectrum of Engineering Sciences*, 942-960.