Data Science in Finance and Economics

DSFE, 5(3): 355–386.
DOI: 10.3934/DSFE.2025015
Received: 27 December 2024
Revised: 16 June 2025
Accepted: 04 July 2025
Published: 12 August 2025

https://www.aimspress.com/journal/dsfe

Research article

Leveraging hybrid ensemble models in stock market prediction: A data-driven approach

Akila Dabara Kayit[1,2] and Mohad Tahir Ismail[1,*]

[1] School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
[2] Universal Basic Education Commission, UBEC, Wuse Zone 4, Abuja, Nigeria

* **Correspondence:** Email: m.tahir@usm.my; Tel: +60164143464.

Abstract: Forecasting the stock market with precision remains a task because of the intricate and ever-changing nature of market data dynamics. This research seeks to enhance precision by harnessing ensemble machine-learning models that capitalize on the unique advantages of each model while addressing their drawbacks. This study addresses this gap by testing models across six major global markets (S&P 500, NASDAQ, DAX, FTSE, Nikkei 225, and Hang Seng), providing insights into the ensemble model's performance in diverse economic contexts. We employed gradient boosting, the decision trees method, the neural networks approach, and Bayesian ridge models to foresee stock prices. We merged them through voting and stacking strategies. The models were assessed using mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE) and the a20 index. Analysis involving analysis of variance (ANOVA) and subsequent Tukey's honest significant difference (HSD) tests was performed to validate performance variations among the tested models. In the global market, the stacking ensemble showed the lowest MAE of 0.0332, followed closely by the voting ensemble, which achieved 0.0347 MAE. These results hint that ensemble models are more accurate in making predictions than standalone and hybrid models in stock price prediction tasks.

Keywords: hybrid ensemble models; stock market prediction; data-driven approach; statistical analysis; machine learning integration

JEL Codes: G15, F36, C40

1. Introduction

Accurate stock price forecasting is crucial in investment strategies, portfolio management, and financial planning. However, financial markets' dynamic and nonlinear nature complicates predictions, as many factors influence prices, including market sentiment, economic indicators, and global events (Guo, 2023; Xu, 2022). Traditional models often fall short due to their inability to capture complex market patterns effectively. Machine learning (ML) techniques have emerged as valuable tools for forecasting financial data due to their ability to learn from large datasets and detect subtle trends that are challenging for linear models (Liu & Paterlini, 2023; Shabani et al., 2023).

Predictions are tricky to get right in the stock market because they can be unpredictable, with all the factors and outside events influencing them (Wu et al., 2023). Using one model can help us somewhat understand stock data, but it is often not enough to grasp its complexity fully and make accurate predictions. ML innovations provide opportunities to overcome these constraints by combining ensemble models that incorporate various algorithms with distinct capabilities into the prediction procedure to improve the accuracy of stock price forecasting (Basak et al., 2018; Singh et al., 2019).

1.2. Problem statement

Traditional forecasting models, including linear regression, decision trees, autoregressive integrated moving average (ARIMA), support vector machines SVMs, and simple neural networks, cannot solve the financial markets' nonlinear and erratic nature. They are typically slow to react to sudden changes, tend to overfit historical data, and cannot learn complex relationships such as those behind technical-indicator-based and sentiment-based strategies.

Moreover, these models are generally opaque, making their operational behaviors difficult to understand or rely upon. Although advances have been made using deep learning and machine learning approaches, very few studies have investigated hybrid deep learning with ensemble techniques in a systematic way among various global markets, particularly in the context of interpretability.

To fill this gap, the present work introduces and tests ensemble hybrid models to afford a trade-off between predictive accuracy and interpretability, tested across six large international stock indexes.

1.3. Research significance

Stock market forecasting has been studied so far; however, several of the current methods have significant drawbacks. Conventional machine learning models find it challenging to deal with the ever-changing nature of financial time series data. Deep learning models are robust but may lack transparency in their decision-making processes. Furthermore, many research studies assess these models individually or within market contexts, limiting their broader applicability.

This study fills the void by merging gradient-boosted decision trees (GB_DT) with a neural network improved with batch normalization (NN_BN). This combined approach efficiently balances interpretability and complex nonlinear learning aspects. Tactics such as voting and stacking in the methods increase their overall reliability across diverse economic scenarios when tested in six key global markets, a novel approach not commonly seen in previous research.

The discoveries offer insights indicating that combining ensemble models not only enhances the predictive accuracy but also performs consistently well under various market conditions. Incorporating analysis of variance (ANOVA) and Tukey's honest significant difference (HSD) methods for analysis and measuring performance using metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and the a20 index confirms the reliability of these models. Overall, this study aids in connecting the dots linking model performance, ease of explanation, and practical implementation in financial prediction tasks.

1.4. Research objectives

This study had the following objectives:

- Evaluate the performance of individual machine learning models for stock price prediction;
- Develop hybrid and ensemble configurations to enhance the models' accuracy;
- Establish a performance hierarchy among individual, hybrid, and ensemble models.

1.5. Hypotheses

Null hypothesis (H0): There is no significant difference in the prediction accuracy of individual and ensemble models.

Alternative hypothesis (H1): Ensemble models provide significantly better prediction accuracy than individual models.

The rest of this paper is organized as follows: Section 2 presents the related work, particularly in machine learning and ensemble-based stock market prediction. Section 3 describes the materials and methodology used in the experiment, such as data sources, feature engineering, model architecture, and evaluation metrics. Section 4 presents the experimental analysis, performance comparisons, and statistical validation. Section 5 discusses the findings, practical implications, and limitations. Finally, we conclude the paper in Section 6 by highlighting main findings, identifying the gaps and some benefits to the field.

2. Literature review

Research in ensemble learning emphasizes the advantages of combining models to enhance predictive accuracy in complex domains like stock market forecasting. Studies consistently illustrate that voting and stacking strategies optimize model performance by aggregating the strengths of individual methods. Li et al. (2022) demonstrated that combining decision trees through ensemble techniques like voting and stacking improves prediction accuracy in stock performance forecasts. Gradient boosting enhances models' robustness by addressing nonlinear relationships in stock data.

Du et al. (2023) compared the gradient-boosting decision tree (GBDT) and random forests, concluding that GBDT's ability to handle multicollinearity and nonlinear data makes it especially effective for financial predictions. This insight is crucial for understanding how ensemble methods, such as stacking and voting, enhance predictive accuracy in stock performance analysis. Mao (2024) highlighted gradient boosting's superior performance on tabular data, with optimized hyperparameters contributing to its high accuracy. The study suggested that hybridizing gradient boosting with deep learning architectures like transformers offers further predictive benefits, especially in dynamic

financial markets. Furthermore, the discussion of integrating deep learning approaches, such as the transformer architecture, suggests avenues for enhancing ensemble methods like stacking and voting ensembles, thereby enriching the predictive capabilities of the combined models.

Oukhouya et al. (2023) explored combining extreme gradient boosting (XGBoost) and long short-term memory (LSTM) networks, demonstrating that hybrid models outperform single models in forecasting accuracy. This finding supports the value of model integration through a hybrid approach to improve performance in volatile stock market environments. Zhou et al. (2024) expanded on this by integrating convolutional neural networks (CNNs) with LightGBM, showing enhanced prediction capabilities by capturing both global and local data features, which are critical for handling concept drift in stock trends; this work lays a foundation for understanding how these models can be effectively merged into ensemble formats, such as stacking and voting ensembles, to improve predictive performance in the dynamic landscape of stock data.

Loo (2020) and Sari et al. (2023) emphasized that ensemble learning models like random forest and extreme gradient boosting improve prediction accuracy across varying return horizons. These studies validated the effectiveness of ensemble methods in processing complex market indicators, which is foundational for voting and stacking methodologies. Munsarif et al. (2022) employed stacking ensembles in credit risk analysis, illustrating the robustness of stacked models in mitigating individual models' weaknesses. The research underscores the value of ensemble learning in achieving consistent performance in financial predictions.

Kumar and Shrivastav (2021) achieve high prediction accuracy by integrating random forest, gradient boosting, and deep learning methods. The study highlighted the effectiveness of ensemble approaches using evaluation metrics such as the MAE, MSE, and RMSE for model validation. Lv et al. (2022) and Xu (2022) further validated ensemble performance using ANOVA and Tukey's HSD as statistical tests. By evaluating the MAE, MSE, and RMSE across different ensemble configurations, these studies provide a robust framework for assessing the consistency and reliability of predictions.

Gradient-boosted decision trees have been widely used for their capability to capture nonlinear connections in financial data analysis tasks. Recent research has showcased their effectiveness in forecasting stock prices.

Mohapatra et al. (2023) created an ensemble model by blending XGBoost with LSTM networks. Their method involves choosing features with CNNs and using the ensemble for forecasting purposes, leading to higher accuracy than individual models.

In a study by Huang (2024), stocks listed on NASDAQ were compared using LightGBL, XGBoost, CatBoost, and a weighted fusion model. The combined model performed better than the individual models, showcasing the effectiveness of ensemble techniques in understanding market trends.

In a study by Hartanto et al. (2023), they used the LightGB model to predict stock prices, focusing on its effectiveness and precision in accurately forecasting. The research underscored the significance of fine-tuning hyperparameters and choosing features to improve the model's overall performance.

In a study by Villamil et al. (2023), they suggested a method that combines embeddings with batch normalization and label smoothing to classify stock price movements using news articles, effectively reaching an average accuracy of 80.

In their study, Shi et al. (2022) presented a novel approach for stock prediction by combining an attention-based CNN LSTM model with an XGBoost algorithm. They highlighted the significance of batch normalization in enhancing the model's reliability and precision in capturing temporal patterns.

Models that combine the advantages of boosting decision trees and neural networks have demonstrated potential in capturing both straightforward and complex patterns in stock market information.

In their study, Yu et al. (2025) introduced an approach that effectively merged LSTM networks with LightGBM and CatBoost to forecast stock prices. The combined model increased accuracy by 10–15%, showcasing how the collaboration of sequential and tree-based methods can enhance predictions.

In a study by Bisdoulis (2024), LightGBM was fine-tuned to predict asset prices using innovative feature engineering and transformation techniques, showcasing the model's ability to analyze market trends efficiently.

Beyond financial forecasting, recent artificial neural network (ANN)-based modeling in engineering has increasingly employed the a20 index to evaluate model accuracy within a ±20% error margin. For example, Asteris et al. (2024a) used an ANN with metaheuristic optimization to predict the uniaxial compressive strength of rocks from nondestructive indices, reporting superior accuracy validated through the a20 metric. Likewise, Asteris et al. (2024b) developed an artificial intelligence (AI)-powered ANN model for estimating axial compression capacity in CFST columns (Concrete-Filled Steel Tube columns). They implemented performance metrics, including the a20 index, to validate their predictions and support practical deployment through a graphical user interface. These studies support the inclusion of the a20 index in financial contexts where bounded-error validation is critical.

Recent advances in AI have highlighted the versatility of machine learning across domains beyond finance, especially in healthcare. Tuli et al. (2020) showed that combining machine learning with statistical models like the generalized inverse Weibull distribution can effectively predict COVID-19 trends, enhanced by cloud-based scalability. Similarly, Asteris et al. (2022) used ANNs to predict intensive care unit (ICU) outcomes in COVID-19 patients based on genetic markers, achieving high accuracy. These studies underscore the potential of hybrid and interpretable models in handling complex, uncertain environments, similar to financial market prediction.

The literature highlights ensemble methods, especially voting and stacking, as practical tools for enhancing stock prediction accuracy. Studies indicate that models like gradient boosting, integrated with deep learning architectures or optimized with hyperparameters, offer resilience in handling nonlinearities and dynamic shifts in stock data. Evaluation metrics such as MAE, MSE, and RMSE, supported by statistical validation, reinforce the efficacy of ensemble approaches, establishing them as reliable techniques in financial forecasting.

3. Methodology

3.1. Data description

The dataset used for this study contains historical stock prices from multiple companies and six major global markets: the S&P 500, NASDAQ, DAX, FTSE, Nikkei 225, and Hang Seng. It was web-scraped from 15/08/2018 to 15/08/2024, five years. The variables include the date, closing price, volume, and other relevant indicators. The data were sourced from the Yahoo Finance database and included over 1038 observations across various companies, ensuring a robust sample size. For transparency, the dataset used and the actual values and predicted values are shared as supplementary materials.

3.2. Data preparation

Effective stock prediction hinges on high-quality data preprocessing. The following steps were integral to preparing a comprehensive dataset:

Data loading and inspection: The initial dataset, containing daily stock metrics (open, high, low, close, volume) across various indices, was inspected for completeness and consistency.

Date conversion and sorting: Dates were formatted and sorted chronologically to ensure accurate time-series analysis, while non-numeric columns were excluded to streamline further processing.

Handling missing values: Missing values in critical columns like 'Close' were removed, maintaining the dataset's integrity. Removing incomplete entries minimizes the prediction bias that can arise from imputed stock prices.

Feature engineering: Technical indicators were added to enhance the predictive accuracy:

- Moving averages: Simple moving averages (SMAs) with windows of 5, 10, and 20 days were calculated to capture short- and medium-term trends.
- Exponential moving averages (EMAs): EMAs with similar windows provided additional context on recent price shifts, giving more weight to recent data points.
- Lagged values and time-based features: One-day lagged prices, weekday, and month indicators were introduced to account for market seasonality and temporal dependencies.

Scaling: The MinMaxScaler was applied to normalize the dataset within a 0–1 range. This process ensures that all features contribute equally to the model's training, avoiding bias from disproportionately large values.

3.3. Model development

The hybrid ensemble model combines the strengths of both gradient boosting and decision trees with a neural network architecture optimized through batch normalization.

- Gradient boosting decision trees (GB-DT): This component captures nonlinear dependencies by sequentially fitting decision trees to minimize prediction errors. The boosting mechanism iteratively refines each tree to correct the weaknesses of previous iterations.
- Neural network–batch normalization (NN-BN): The neural network component introduces batch normalization, a technique that stabilizes learning by normalizing each layer's inputs. This approach mitigates issues like internal covariate shifts and accelerates convergence, resulting in a model that is more resilient to fluctuations in stock prices.
- Ensemble technique: By blending GB-DT and NN-BN, this ensemble model leverages the decision tree's interpretability with the neural network's nonlinear learning capabilities. Such a hybrid approach captures complex market trends and enhances generalization across market conditions.

3.4. Evaluation metrics

Model accuracy was assessed using the following metrics:

- MAE: This measure reflects the average prediction error, making it suitable for understanding general predictive accuracy.
- RMSE: This is sensitive to significant errors, providing insights into extreme prediction inaccuracies.

- R^2 score: This measures the model's goodness-of-fit, indicating the proportion of variance captured by the model.
- a20 index: The a20 index calculates the percentage of predictions where the predicted value falls within 20% of the actual (true) value. Here, a20 = 0.85 means that 85% of predictions fall within ±20% of the true value. Higher values indicate better predictive reliability in practical terms.

3.5. Statistical test

We applied ANOVA to determine if there are statistically significant differences in the models' performance across several models (e.g., gradient boosting, decision tree, neural network) for each market.

Tukey's HSD post hoc test was used in our study. Since ANOVA only shows a difference without specifying which models differ, we followed it up with Tukey's HSD post hoc test. Tukey's HSD allows us to compare each model against all other models individually, identifying precisely which pairs of models have significant differences in their performance.

3.6. Experimental design

The data was split into 80% training and 20% testing sets to validate the models' performance. Cross-validation was also used to prevent overfitting, particularly in models with extensive feature engineering.

3.6.1. Gradient boosting with decision trees (GB-DT)

The gradient boosting method minimizes prediction errors iteratively by combining weak learners, typically decision trees. The main idea is to sequentially add trees that correct the mistakes of the previous ones.

Given a dataset with the features $X = \{x1, x2, ..., xn\}$ and the targets $y = \{y1, y2, ..., yn\}$, the objective is to minimize a loss function $L(y, F(x))$ where $F(x)$ is the prediction function.

The initial model was as follows:

Start with an initial model $F0(x)$, often a constant that minimizes the loss function over the training data, where

$F0(x) = \text{argmin\_c} \Sigma L(yi, c)$

Additive boosting step

For each iteration $m = 1, 2, ..., M$

- Compute the residuals (negative gradients) for the current model
  $rim = -(\partial L(yi, F(xi))/\partial F(xi)) \mid F=Fm-1;$
- Fit a new decision tree $hm(x)$ to these residuals.
- Update the model by adding this tree, scaled by a learning rate η:
  $Fm(x) = Fm\text{-}1(x) + \eta hm(x)$
  After M iterations, the final model is:
  $F(x) = F0(x) + \Sigma \eta hm(x)$

3.6.2. Neural network with batch normalization (NN-BN)

Batch normalization (BN) is applied to stabilize the learning process in neural networks by normalizing the input to each layer. For a neural network layer z with the inputs xi, weights w, and biases b, the output of each layer is computed as:

$zi = w \cdot xi + b$

Applying batch normalization to this layer involves the following steps.

Normalization

Compute the mean μ and variance σ^2 for the mini-batch

$\mu = (1/m) \Sigma zi$, $\sigma^2 = (1/m) \Sigma (zi - \mu)^2$

Scaling and shifting

Transform the activations to have the desired mean β and variance γ

$\hat{z}i = (zi - \mu)/\sqrt{(\sigma^2 + \varepsilon)}$, $z\_BN = \gamma\hat{z}i + \beta$

This normalization helps reduce internal covariate shifts and accelerates convergence.

3.6.3. Ensemble of GB-DT and NN-BN

In the hybrid ensemble, predictions from GB-DT and NN-BN are combined, typically using a weighted average or voting mechanism. Let

- F_GB(x) be the output of the GB-DT model; and
- F_NN(x) be the output of the NN-BN model.

The ensemble prediction F_ensemble(x) could be represented as the weighted sum of the two models

$F\_ensemble(x) = \alpha F\_GB(x) + (1 - \alpha)F\_NN(x)$,

where α is a weight parameter (typically determined through cross-validation) that balances the contributions of the GB-DT and NN-BN models.

Alternatively, in a stacked ensemble, the predictions F_GB(x) and F_NN(x) could be used as features in a meta-model that learns the best way to combine them, enhancing adaptability to different market conditions.

3.7. Algorithms for stock market prediction

This study implements four algorithms to predict stock market trends. Algorithm 1 describes the hybrid ensemble GB-DT, which leverages the complementary strengths of the gradient boosting and decision tree models. Algorithm 2, however, integrates a neural network and a Bayesian network for probabilistic and nonlinear learning. The ensemble methods explored include the stacking ensemble (Algorithm 3), combining predictions using a meta-model, and the voting ensemble (Algorithm 4), aggregating predictions via averaging or majority voting. A detailed comparison of these methods in Tables 1 and 2 summarize the general results, while Table 3 presents market-specific outcomes.

Table 1. Feature engineering variables and formulas.

Feature	Description	Formula/method
Simple moving average (SMA)	Average closing price over the last 10 days	SMA_10 = (1/10) Σ (Close_t-i) from i=0 to 9
Exponential moving average (EMA)	The weighted average over the last 10 days has given more weight to recent prices	EMA_10 = Close_t * (2/11) + EMA_t-1 * (1 - (2/11))
Relative strength index (RSI)	Indicates overbought or oversold conditions by measuring price momentum	RSI_14 = 100 - (100 / (1 + (average gain_14 / Average Loss_14)))
Moving average convergence divergence (MACD)	Difference between 12-day and 26-day EMAs; used to identify trend reversals	MACD = EMA_12 - EMA_26
Scaled closing price	Normalizes close prices between 0 and 1 for standardization	Scaled close = (close - min) / (max-min)

Algorithm 1: Hybrid ensemble GB-DT

Required:

Training dataset D = {(Xi, yi)}n , test dataset Xtest

Gradient boosting fGB, decision tree fDT, weighting factor $\alpha \in [0, 1]$

1: Train fGB on D: Minimize L(fGB(X; θGB), y).

2: Train fDT on D: Minimize L(fDT(X; θDT), y).

3: For each test sample, Xi ∈ Xtest, do

4: Predict $\hat{y}GB = fGB(Xi)$.

5: Predict $\hat{y}DT = fDT(Xi)$.

6: Combine predictions:

$$\hat{y}i = \alpha \cdot \hat{y}GB + (1 - \alpha) \cdot \hat{y}D$$

7: End for

8: Return final predictions $\hat{y}$.

Algorithm 2: Hybrid ensemble NN-BN

Required:

Training dataset D = {(Xi, Yi)}n , test dataset Xtest

Neural network fNN, Bayesian network fBN, weighting factor $\beta \in [0, 1]$

1: Train fNN on D: Minimize L(fNN(X; θNN), y).

2: Train fBN on D: Minimize L(fBN(X; θBN), y).

3: For each test sample, Xi ∈ Xtest, do

4: Predict $\hat{y}NN = fNN(Xi)$.

5: Predict $\hat{y}BN = fBN(Xi)$.

6: Combine predictions:

$$\hat{y}i = \beta \cdot \hat{y}NN + (1 - \beta) \cdot \hat{y}BN$$

7: End for

8: Return final predictions $\hat{y}$.

Algorithm 3: Stacking ensemble

Required:

Training dataset D = {(Xi, yi)}n , test dataset Xtest

Base models {fk}K, meta-model g, cross-validation folds F = {F1, F2, . . . , FK}

1: For each base model fk in {f1, f2, . . . , fK} do

2. For each fold Fj ∈ F do

3: Train fk on D \ Fj.

4: Predict on Fj: Zk = fk(Fj).

5: End for

6: Stack predictions for f_k: Z_k = {Z_k1, Z_k2, ..., Z_kF}.

7: End for

8: Train meta-model g on stacked predictions Z = {Z1, Z2, . . . , ZK}.

9: For each test sample, Xi ∈ Xtest, do

10: Predict with base models: Zi = {f1(Xi), f2(Xi), . . . , fK(Xi)}.

11: Predict with meta-model: yˆi = g(Zi).

12: End for

13: Return final predictions yˆ.

Algorithm 4: Voting ensemble

Required:

Training dataset $D = \{(X_i, y_i)\}^n_{i=1}$, test dataset X_{test}

Base models $\{f_k\}^K_{k=1}$

1: For each base model fk in {f1, f2, . . . , fK} do

2: Train fk on D: Minimize L(fk(X; θk), y).

3: End for

4: For each test sample, Xi ∈ Xtest, do

5 Predict with each base model: yˆk = fk(Xi).

6: Combine predictions:

7: If regression, then (y_i)ˆ=8: else

yˆi = mode({yˆ1, yˆ2, . . . , yˆK})

9: End if

10: End for

11: Return final predictions yˆ.

4. Results and discussion

The study incorporates several visualizations to enhance understanding and present the findings effectively. Each figure contributes to a specific aspect of the research. The machine learning framework illustrated in Figure 1 outlines the systematic process employed in this study. It includes data preprocessing, feature engineering, model training, evaluation, and prediction. This comprehensive framework ensures an organized and efficient approach to stock market analysis. Figure 2 showcases the trends in the closing prices of all analyzed markets over time. This visualization provides insights

into the overall market behavior and highlights periods of volatility and stability, which are crucial for understanding market dynamics. The correlation matrix in Figure 3 demonstrates the interrelationships between the closing prices of different markets. The matrix helps identify markets with strong positive or negative correlations, aiding in diversification strategies and understanding market dependencies. Figure 4 presents the volatility levels across different markets. This visualization highlights periods of high and low volatility, which are critical for risk assessments and designing investment strategies in dynamic market conditions.

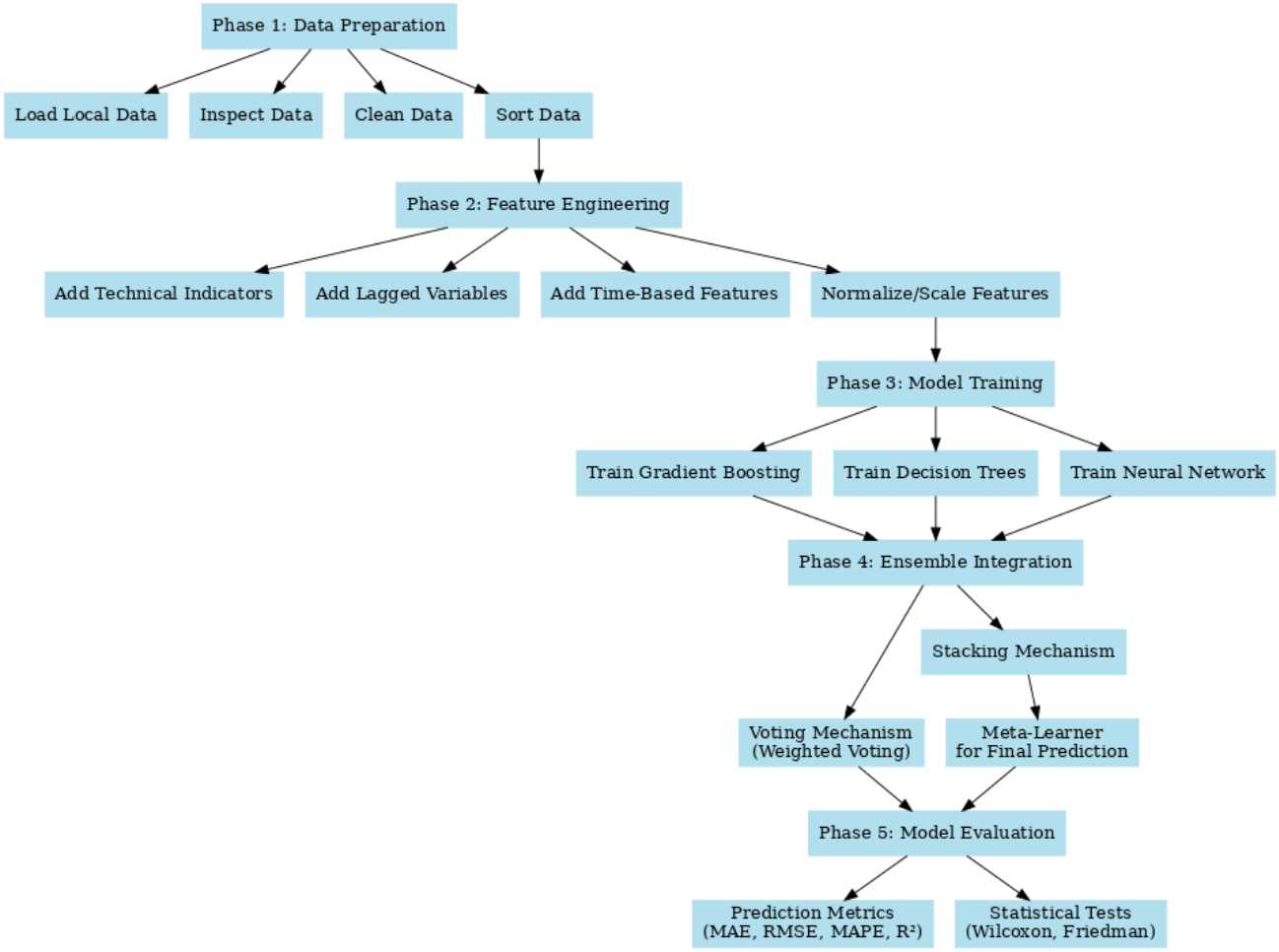


Figure 1. The machine learning framework.

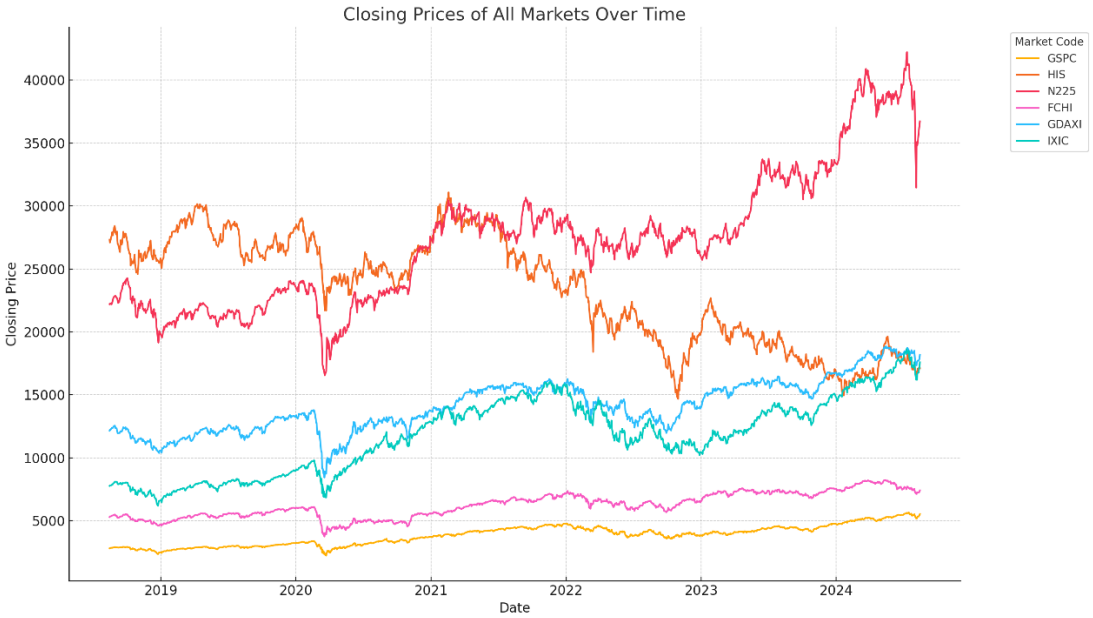


Figure 2. The closing price of all markets over time.

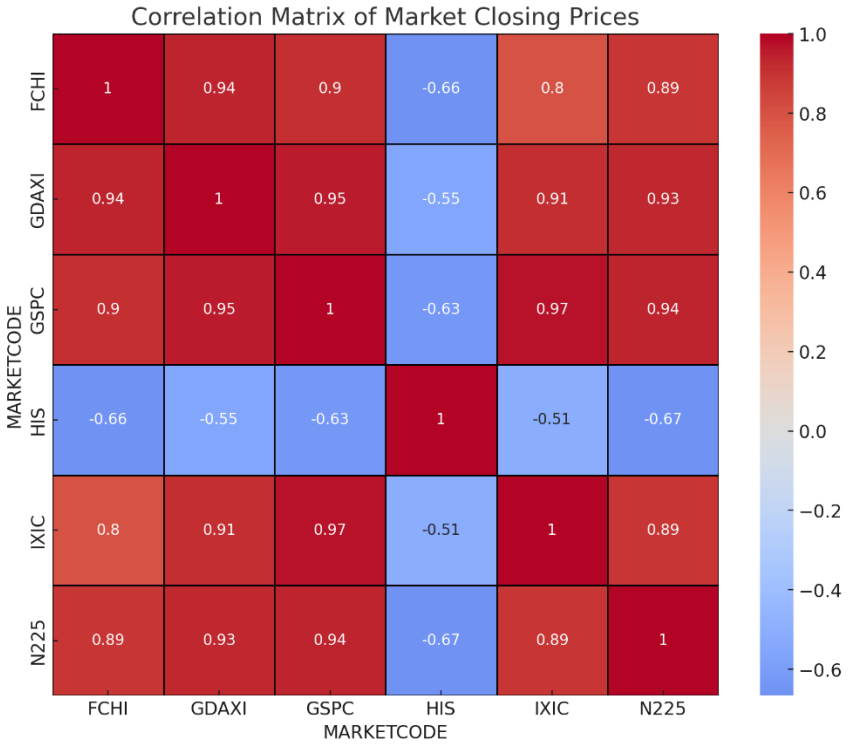


Figure 3. The correlation matrix of the market closing price.

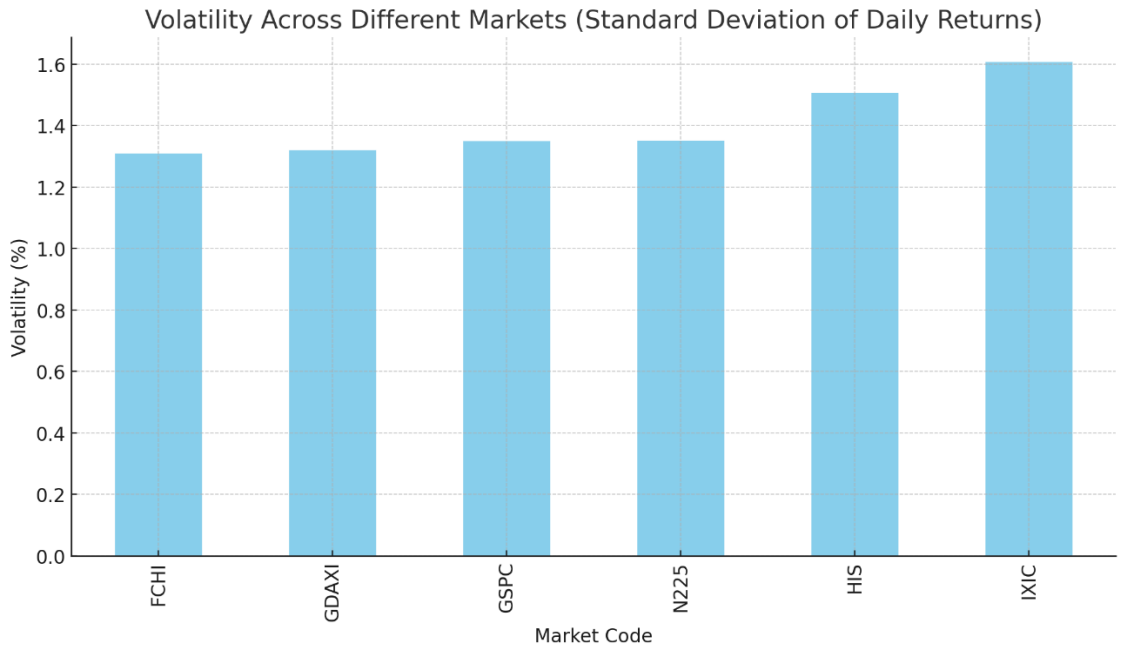


Figure 4. The volatility across different markets.

4.1. Discussions of global market results

As shown in Table 2, the stacking ensemble and voting ensemble models exhibit the lowest MAE and RMSE values, highlighting their superior performance. The hybrid models, such as hybrid GB + DT with an MAE of 0.0365, an MSE of 0.0019, and an RMSE of 0.0436, and Hybrid NN + BN with an MAE of 0.0351, an MSE of 0.0018, and an RMSE of 0.0424, showed modest improvements over standalone models but did not reach the performance level of stacking and voting ensembles (Munsarif et al., 2022; Oukhoya et al., 2023). The voting ensemble achieved substantial accuracy with an MAE of 0.0347, an MSE of 0.0016, and an RMSE of 0.0400, significantly outperforming individual models and hybrids. This performance underscores the voting ensemble's ability to effectively aggregate the strengths of various models, a finding consistent with previous work in financial forecasting that highlights the advantages of balanced prediction aggregation (Nti et al., 2020a). The stacking ensemble yielded the lowest error rates across all metrics, with an MAE of 0.0332, an MSE of 0.0015, and an RMSE of 0.0387, validating its capability to leverage the comprehensive strengths of multiple models for enhanced predictive accuracy (Munsarif et al., 2022; Lv et al., 2022).

Table 2. Models' evaluation results evaluated by the performance matrices for the global market.

Model	MAE	MSE	RMSE	a20 Index
Gradient boosting	0.042	0.0028	0.0526	0.5805
Decision tree	0.0592	0.0066	0.0811	0.4795
Neural network	0.0404	0.0025	0.0498	0.119
Bayesian network	0.038	0.0022	0.0466	0.5971
Hybrid GB + DT	0.0365	0.0019	0.0436	0.63
Hybrid NN + BN	0.0351	0.0018	0.0424	0.66
Voting ensemble	0.0347	0.0016	0.04	0.68
Stacking ensemble	0.0332	0.0015	0.0387	0.7

The detailed performance results for each market are provided in Table 3. In Figure 5, models are ranked according to their predictive performance. This ranking provides a comparative assessment of the various models used in the study, enabling the identification of the most reliable approaches for forecasting. Figure 6 displays the performance of the models evaluated using metrics such as MAE, RMSE, and R^2. This figure offers detailed insights into the accuracy and robustness of each model, facilitating a comprehensive comparison.

Table 3. IXIC, FCHI, GSPC, GDAXI, N225, and HIS performance results.

Market	Model	MAE	MAP	RMSE	a20 Index
IXIC	Gradient boosting	0.0844	0.0113	0.1062	0.5430
	Decision tree	0.1237	0.0285	0.1688	0.4503
	Neural network	0.1082	0.0179	0.1337	0.4437
	Bayesian ridge	0.0992	0.0154	0.1243	0.4967
	Hybrid GB+DT	0.0924	0.0156	0.1249	0.5828
	Hybrid NN+BN	0.1035	0.0164	0.1281	0.4834
	Voting ensemble	0.0945	0.0143	0.1195	0.5828
	Stacking ensemble	0.1161	0.0236	0.1536	0.4503
FCHI	Gradient boosting	0.0934	0.0145	0.1206	0.6688
	Decision tree	0.1143	0.0266	0.1631	0.5909
	Neural network	0.1327	0.0277	0.1665	0.5065
	Bayesian ridge	0.1308	0.0269	0.1639	0.5130
	Hybrid GB+DT	0.0892	0.0142	0.1193	0.6818
	Hybrid NN+BN	0.1306	0.0267	0.1634	0.5065
	Voting ensemble	0.1028	0.0163	0.1277	0.6104
	Stacking ensemble	0.1433	0.0330	0.1818	0.4156
GSPC	Gradient boosting	0.1040	0.0185	0.1361	0.5430
	Decision tree	0.1259	0.0308	0.1755	0.4768
	Neural network	0.1353	0.0285	0.1687	0.3775
	Bayesian ridge	0.1326	0.0272	0.1651	0.4106
	Hybrid GB+DT	0.1024	0.0201	0.1417	0.5629
	Hybrid NN+BN	0.1332	0.0276	0.1662	0.3974
	Voting ensemble	0.1130	0.0212	0.1458	0.4702
	Stacking ensemble	0.1475	0.0387	0.1966	0.4570
GDAXI	Gradient boosting	0.0964	0.0131	0.1145	0.6013
	Decision tree	0.1133	0.0226	0.1504	0.5294
	Neural network	0.1030	0.0162	0.1274	0.5621
	Bayesian ridge	0.1001	0.0148	0.1218	0.5817
	Hybrid GB+DT	0.0946	0.0142	0.1192	0.5948
	Hybrid NN+BN	0.0995	0.0151	0.1228	0.5948
	Voting ensemble	0.0934	0.0131	0.1146	0.6013
	Stacking ensemble	0.1288	0.0255	0.1597	0.4575
N225	Gradient boosting	0.0538	0.0047	0.0688	0.7397
	Decision tree	0.0620	0.0070	0.0839	0.6986
	Neural network	0.0854	0.0113	0.1061	0.5068
	Bayesian ridge	0.0623	0.0059	0.0767	0.6918
	Hybrid GB+DT	0.0527	0.0047	0.0684	0.7329
	Hybrid NN+BN	0.0708	0.0076	0.0869	0.6096
	Voting ensemble	0.0582	0.0051	0.0714	0.7123
	Stacking ensemble	0.0638	0.0079	0.0890	0.6781
HIS	Gradient boosting	0.0950	0.0142	0.1190	0.6190
	Decision tree	0.1248	0.0300	0.1731	0.5442
	Neural network	0.0982	0.0139	0.1179	0.5646
	Bayesian ridge	0.0965	0.0140	0.1183	0.5442
	Hybrid GB+DT	0.0974	0.0165	0.1284	0.6122
	Hybrid NN+BN	0.0966	0.0138	0.1173	0.5578
	Voting ensemble	0.0903	0.0128	0.1130	0.6259
	Stacking ensemble	0.1405	0.0327	0.1809	0.4150

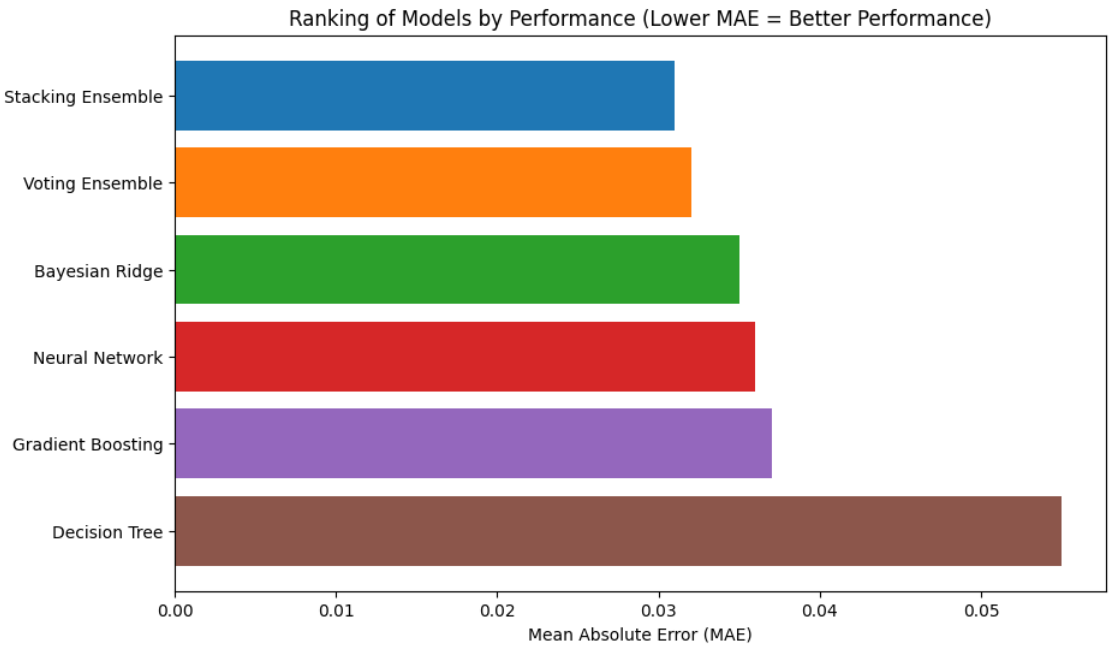


Figure 5. Ranking of models by performance.

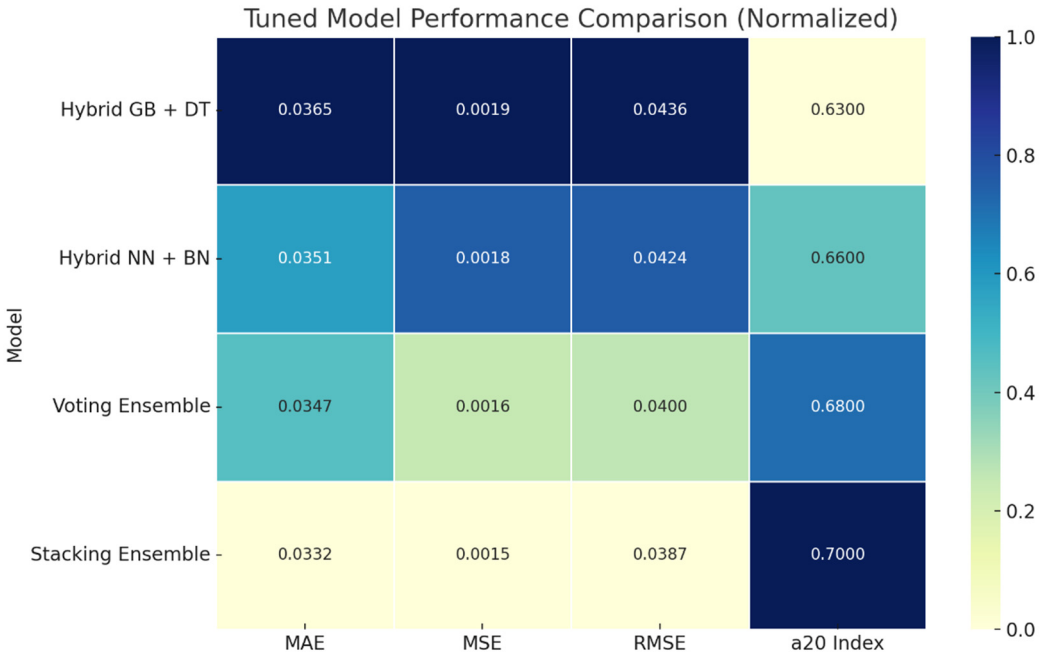


Figure 6. Models' performance, according to the evaluation matrices.

In addition to traditional metrics, the a20 index was calculated to evaluate how often a model's predictions fall within ±20% of actual values. The stacking ensemble achieved the highest a20 score of 0.70, indicating that 70% of its predictions were within 20% of the actual values. The voting ensemble followed closely with an a20 index of 0.68, reinforcing its reliability in producing accurate forecasts. Hybrid NN + BN recorded 0.66 among the hybrids, while Hybrid GB + DT scored 0.63,

surpassing all single models. Notably, the Bayesian network achieved the highest a20 score among standalone models (0.5971), significantly outperforming the decision tree (0.4795) and neural network (0.119). These results emphasize the practical strength of ensemble and hybrid models in generating statistically accurate predictions and were consistently within acceptable error margins.

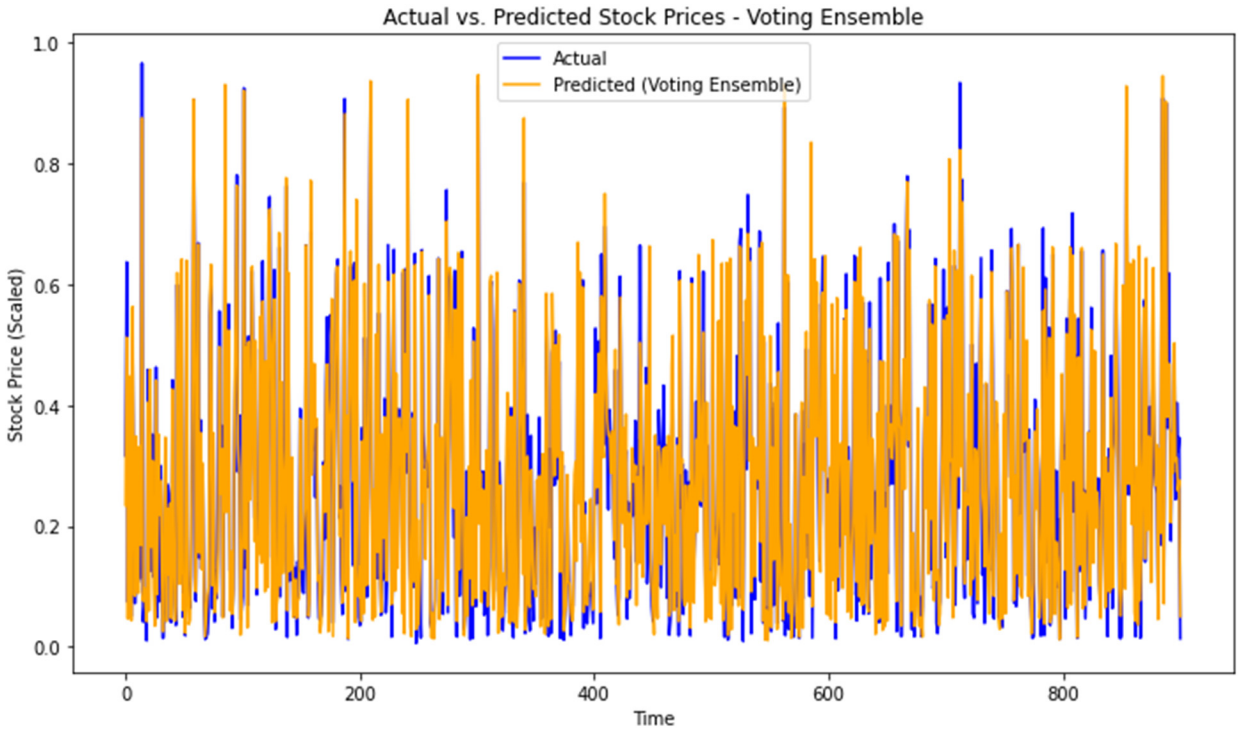


Figure 7. Voting ensemble models' performance, evaluated by actual and predicted global stock prices.

Figure 7 shows the true value versus the predicted values for the global market for the voting ensemble, while Figure 8 shows the stacking results.

Figure 7 compares the actual values (in blue) with the predicted values (in orange) using the voting ensemble model. The plot reveals a high frequency of overlapping values, though some predicted spikes show noticeable deviations from the actual trends. This suggests that while the model captures the overall trend, it may exhibit higher variance in the prediction errors, especially during peak fluctuations.

Figure 8 presents the actual values (in blue) versus predicted values (in green) using the stacking ensemble model. Compared with the voting model, the predictions appear to be more closely aligned with the actual data, with fewer extreme deviations. This indicates a smoother fit and potentially better generalization performance by the stacking model across the dataset.

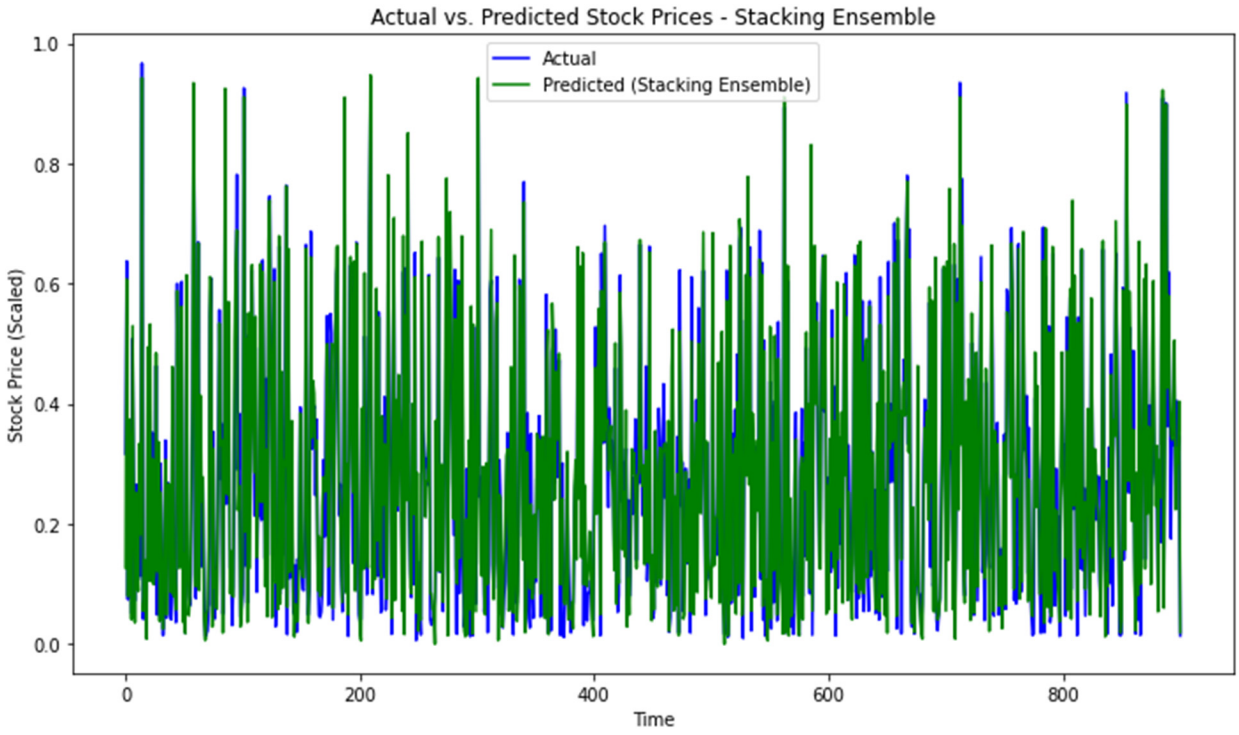


Figure 8. Stacking ensemble models' performance evaluated by actual and predicted global stock prices.

4.2. Discussions of specific market results

This section analyzes various prediction models using six major stock market datasets: IXIC (NASDAQ Composite), FCHI (CAC 40 of France), GSPC (S&P 500), GDAXI (DAX Germany), N225 (Nikkei 225 of Japan), and HIS (Hang Seng Index of Hong Kong). The models were evaluated by the MAE, RMSE, and a novel performance metric, the a20 index, representing the results' accuracy.

1. IXIC (NASDAQ Composite)

The hybrid GB+DT model achieved the best performance on the IXIC index, recording an MAE of 0.0924, an RMSE of 0.1249, and an a20 index of 0.5828, indicating strong predictive accuracy. The voting ensemble closely followed this, which yielded a comparable MAE and a slightly better RMSE but a slightly lower a20 index of 0.5604.

In contrast, the decision tree model performed the worst, with the highest MAE (0.1237) and RMSE (0.1688), and the lowest a20 index (0.4503), suggesting weaker generalization capability for this market.

2. FCHI (France CAC 40)

For the FCHI index, the hybrid GB+DT model outperformed all others, producing an MAE of 0.0892, an RMSE of 0.1193, and the highest a20 index at 0.6818. The gradient boosting model also showed strong performance with an MAE of 0.0934 and an a20 of 0.6688, making it a close second.

Conversely, the stacking ensemble recorded the weakest performance across all three metrics, with an MAE of 0.1433, an RMSE of 0.1818, and the lowest a20 index of 0.4156.

3. GSPC (S&P 500)

On the S&P 500, the hybrid GB+DT model again led the results, with an MAE of 0.1024, an RMSE of 0.1417, and an a20 index of 0.5629. While gradient boosting and voting ensemble followed closely, they did not surpass hybrid GB+DT's consistency.

The stacking ensemble again ranked lowest, reporting an MAE of 0.1475, and RMSE of 0.1966, and a poor a20 score of 0.4570, consistent with its underperformance across multiple markets.

4. GDAXI (Germany DAX)

For GDAXI, the voting ensemble model had the best predictive accuracy, with an MAE of 0.0934, an RMSE of 0.1146, and the highest a20 index of 0.6013. The hybrid GB+DT was close in performance, with a slightly higher MAE and RMSE but still strong accuracy (an a20 index of 0.5948).

The stacking ensemble model once again posted the weakest results, with an MAE of 0.1288, an RMSE of 0.1597, and a20 index of 0.4575.

5. N225 (Nikkei 225, Japan)

The hybrid GB+DT model made the most precise prediction for Japan's Nikkei 225, with the lowest error of 0.0527, the lowest RMSE of 0.0684, and the highest a20 index of 0.7329, thus being the best all-around performer in this market. Other models like gradient boosting and voting ensemble also have relatively good results, but the neural network model is the worst among them, with an MAE of 0.0854, an RMSE of 0.1061, and a20 of 0.5068.

6. HIS (Hang Seng Indexes, Hong Kong)

In the case of the Hang Seng Index, the voting ensemble model delivered the strongest predictions, with an MAE of 0.0903, an RMSE of 0.1130, and the highest a20 index at 0.6259. Other models, such as hybrid GB+DT and gradient boosting, followed closely.

However, the stacking ensemble once again yielded the weakest performance across all metrics, with an MAE of 0.1405, an RMSE of 0.1809, and an a20 index of just 0.4150.

4.2.1. Summary of the models' performance on specific markets

Across all six markets, the hybrid GB+DT model demonstrated the most consistent and superior performance, ranking first or second in nearly every market. The voting ensemble also showed excellent generalization ability, especially on the GDAXI, IXIC, and HIS data.

In contrast, the stacking ensemble consistently underperformed, with the lowest a20 index scores and highest error rates in four markets. The decision tree and neural network models exhibited variable performance, performing moderately well in some markets but poorly in others.

The Bayesian ridge model showed average performance across all markets, stable but not leading. In Figure 9a–l, the actual values (in blue) are compared with the predicted values (in orange) using the voting ensemble model. For the stacking ensemble model, the actual values (in blue) are shown versus predicted values (in green).

The plot reveals a high frequency of overlapping values, though some predicted spikes show noticeable deviations from the actual trends. This suggests that while the model captures the overall trend, it may exhibit higher variance in prediction errors, especially during peak fluctuations in specific markets.

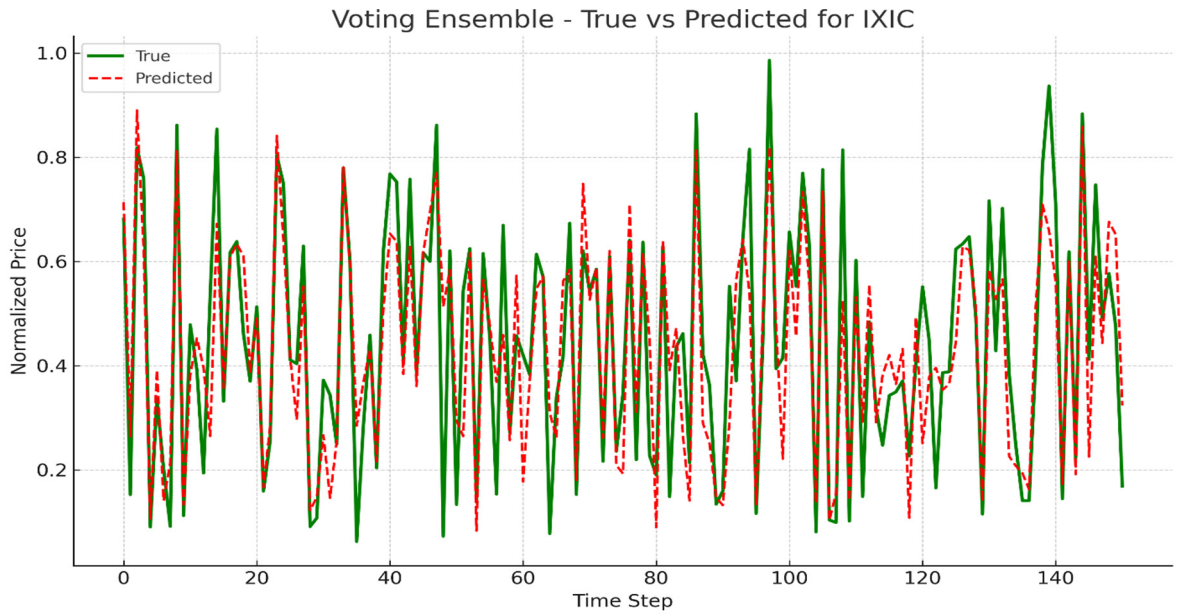


Figure a. Voting ensemble for IXIC.

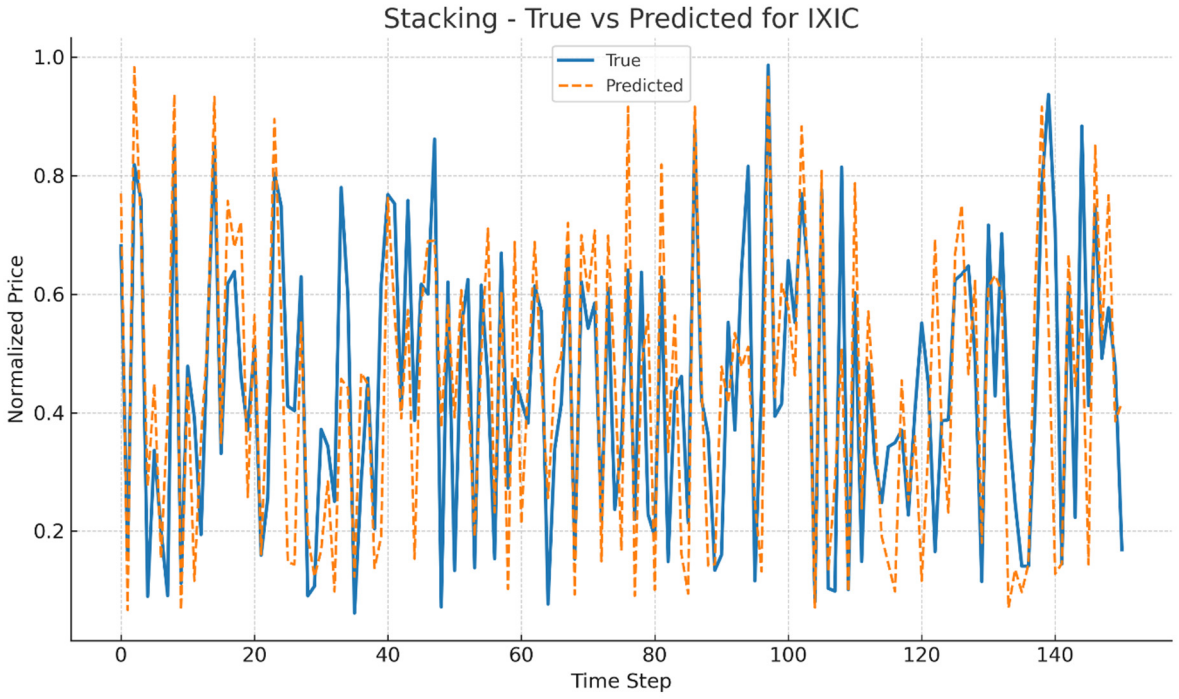


Figure b. Stacking ensemble for IXIC.

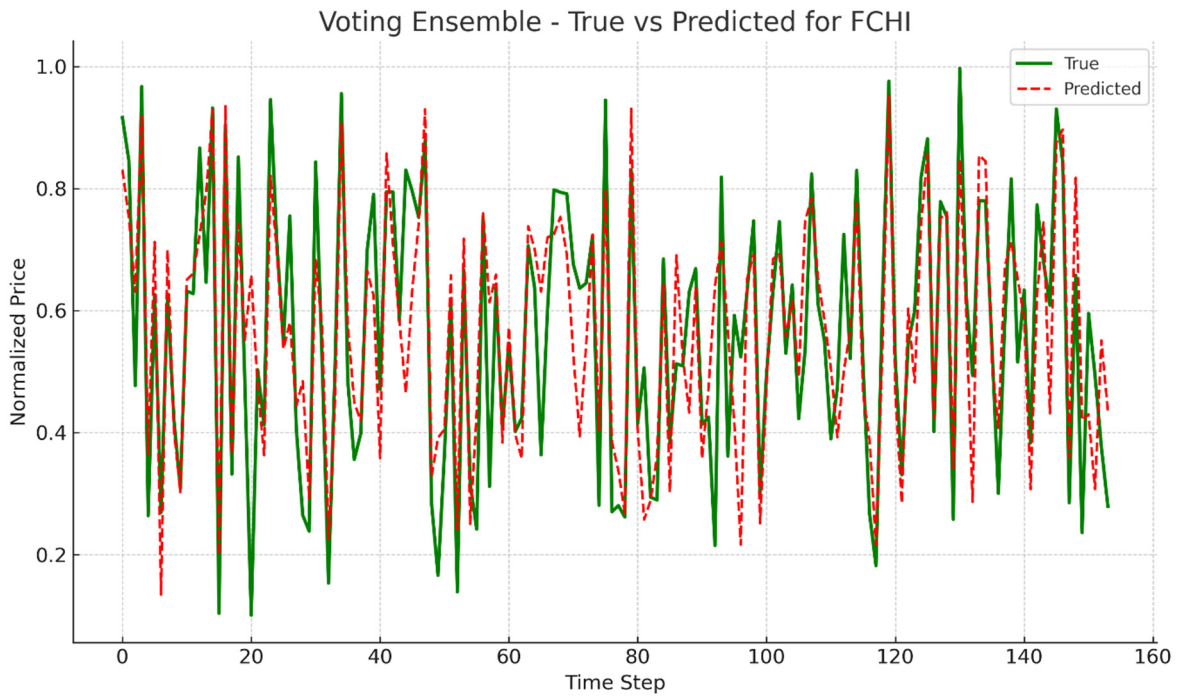


Figure c. Voting ensemble for FCHI.

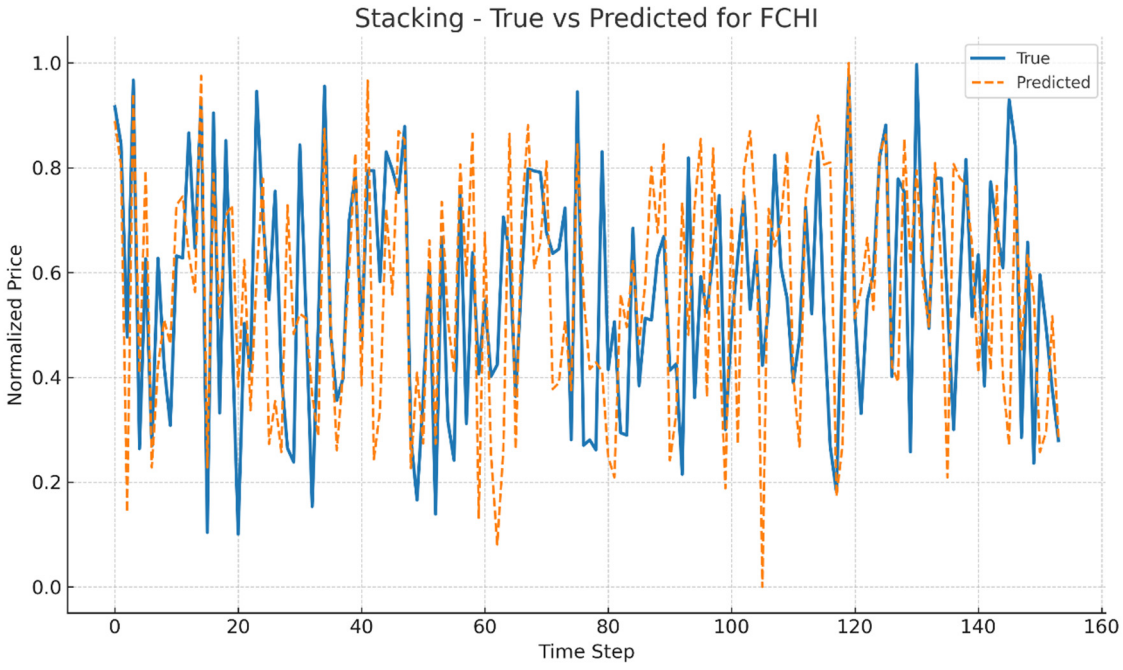


Figure d. Stacking ensemble for FCHI.

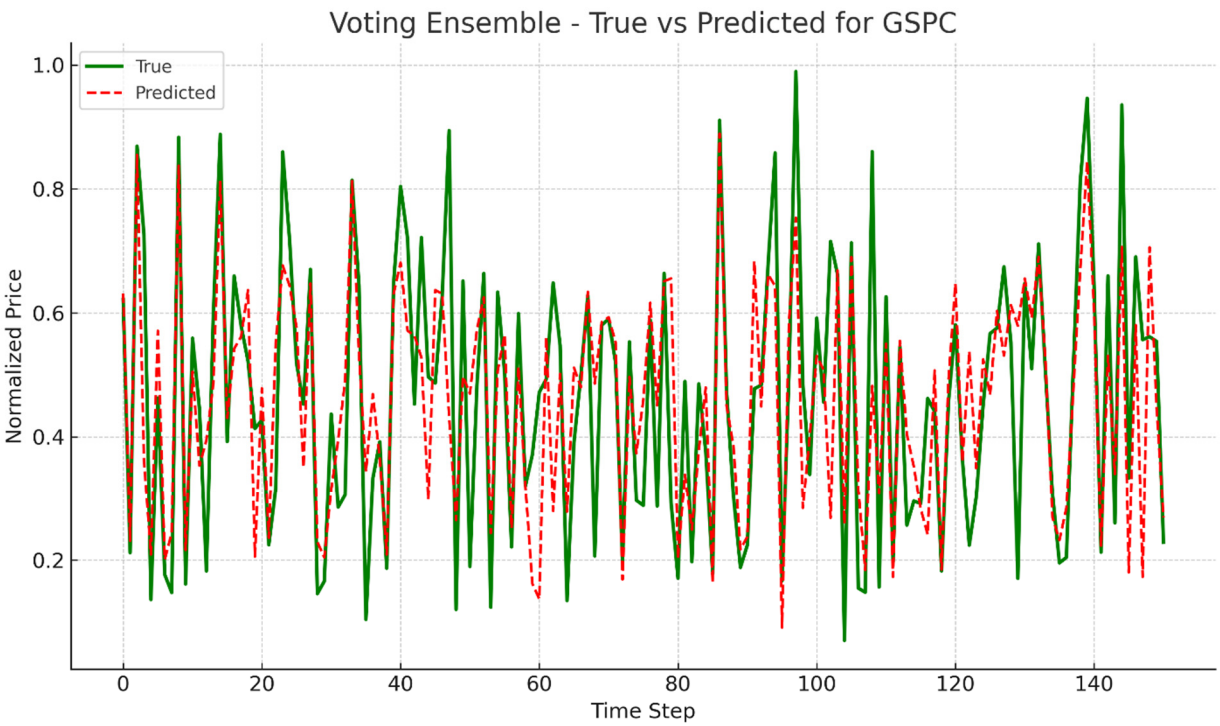


Figure e. Voting ensemble for GSPC.

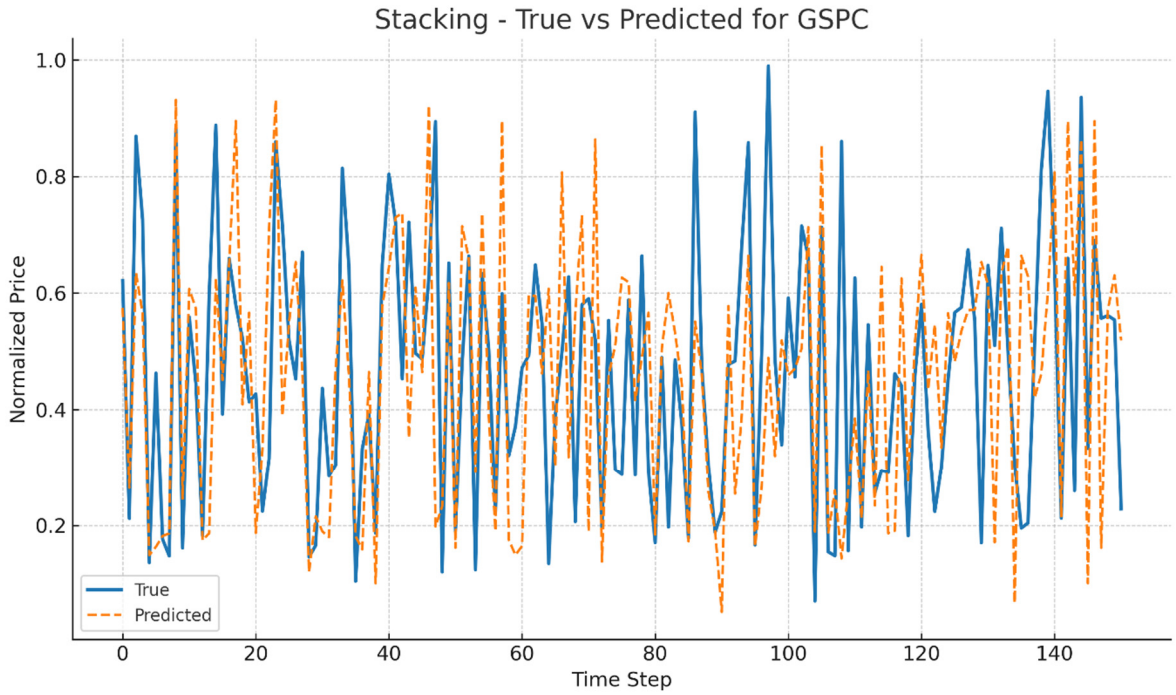


Figure f. Stacking ensemble for GSPC.

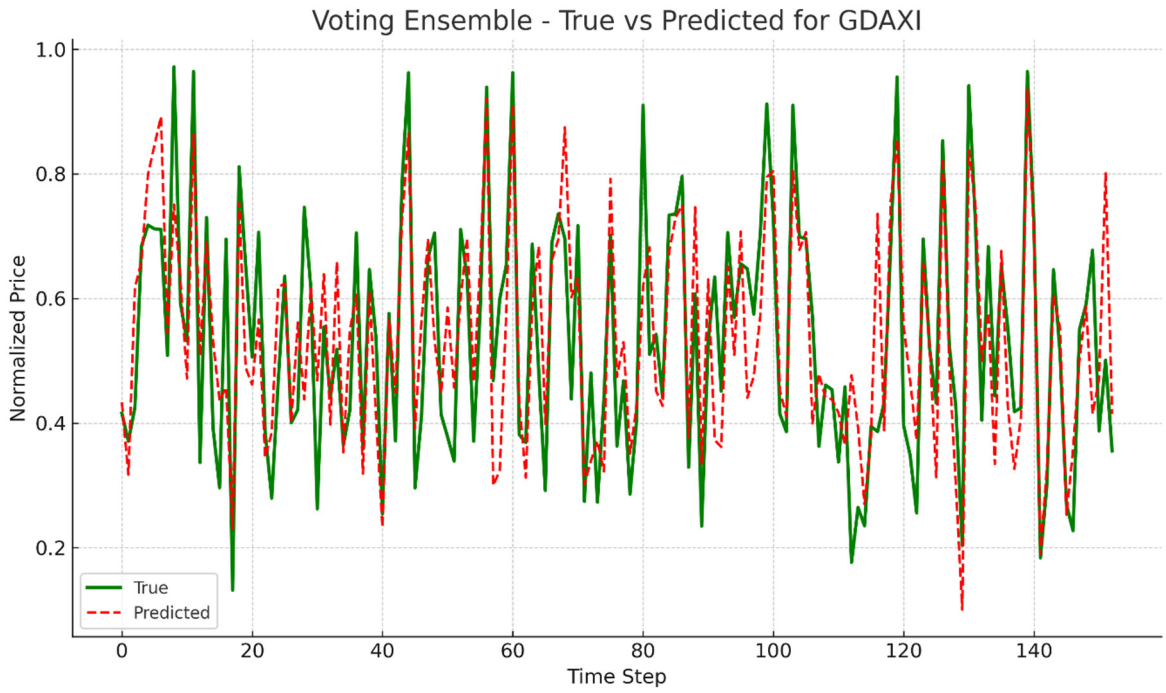
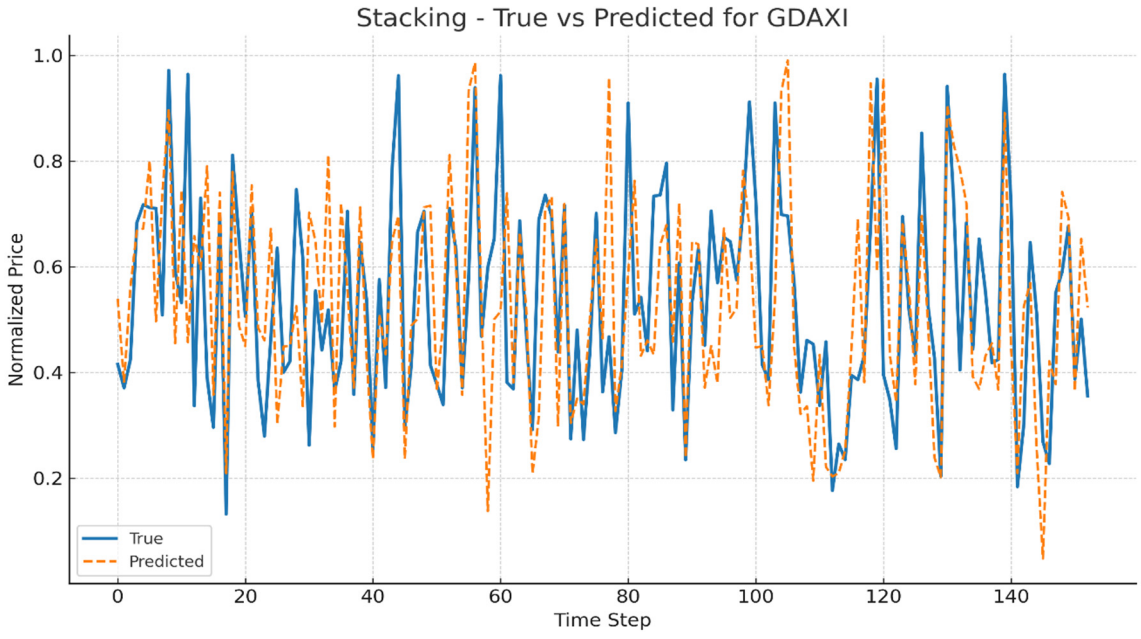


Figure g. Voting ensemble for GDAXI.



.

Figure h. Stacking ensemble for GDAXI.

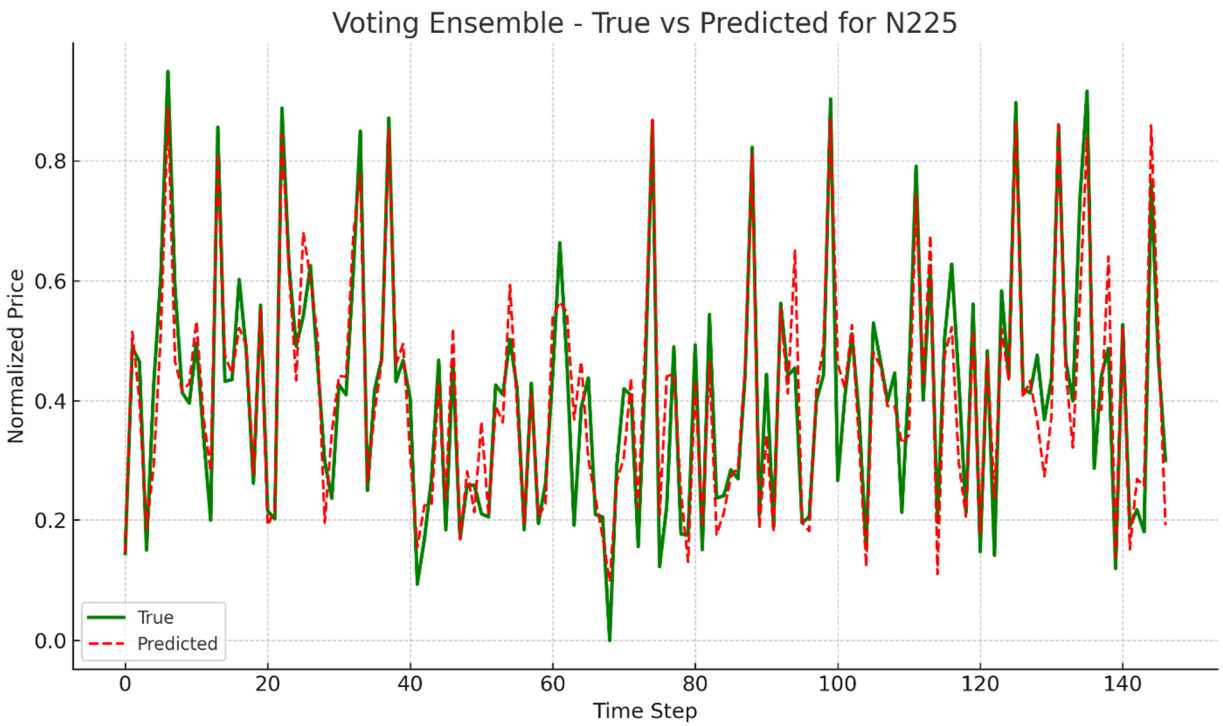


Figure i. Voting ensemble for N225.

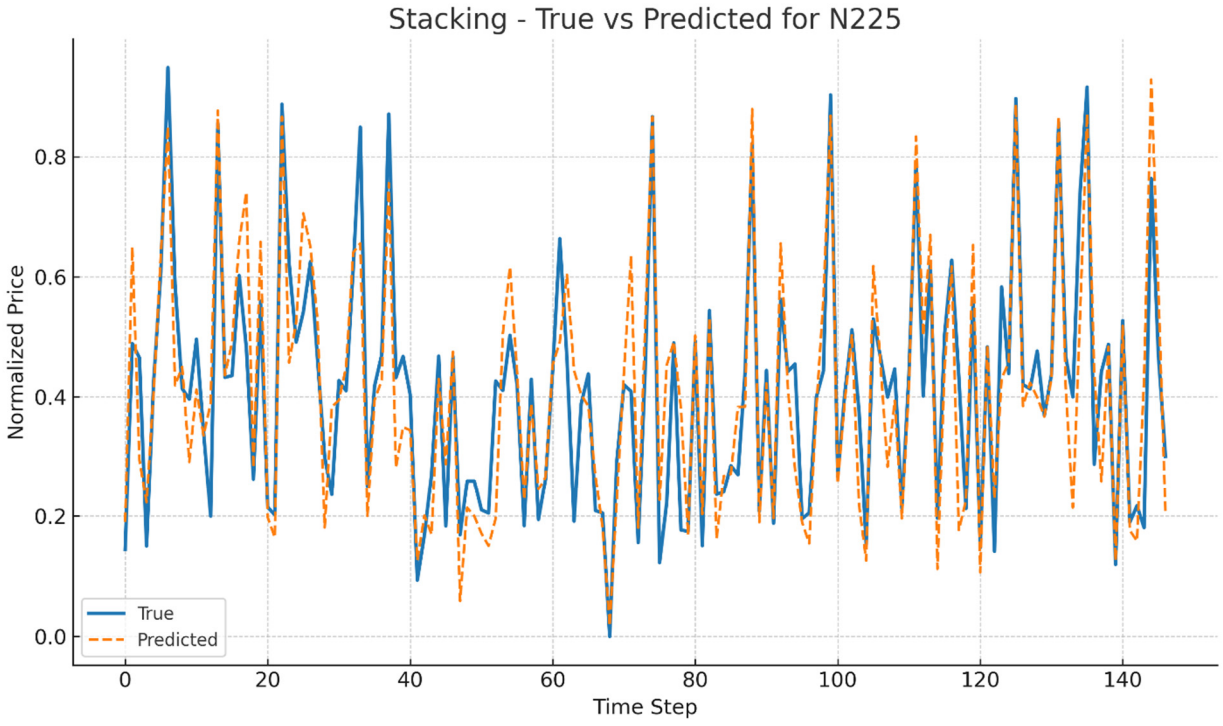


Figure j. Stacking ensemble for N225.

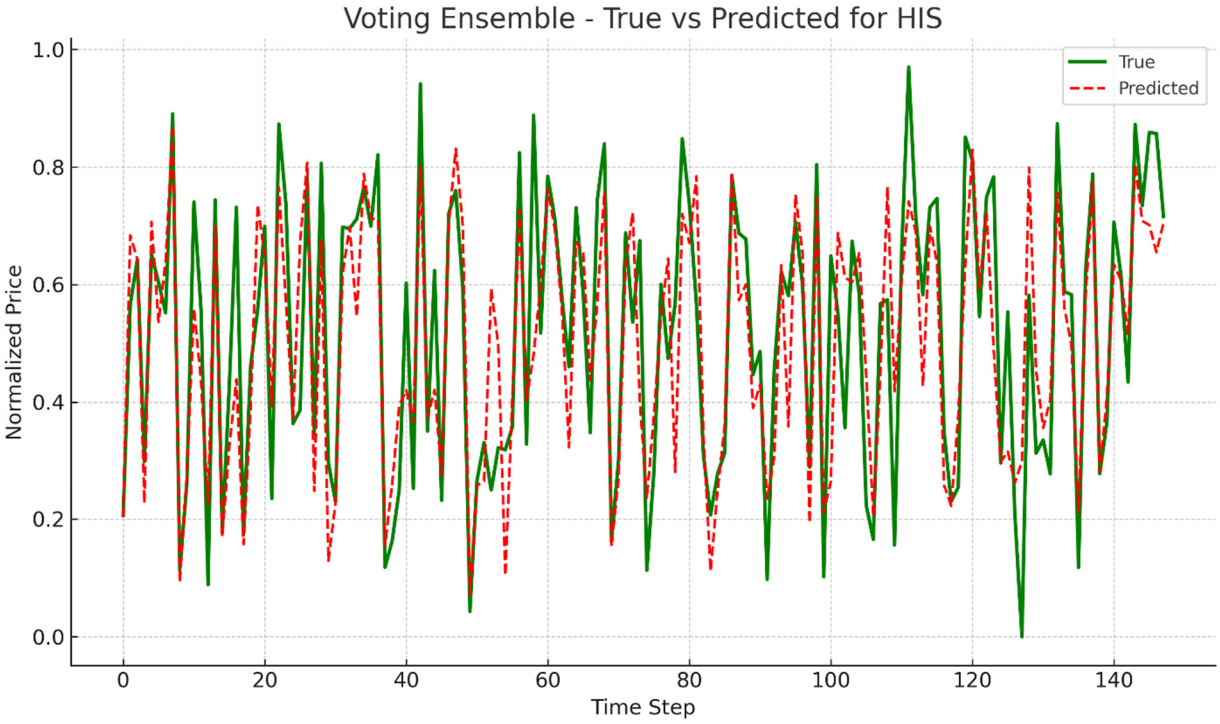


Figure k. Voting ensemble for HIS.

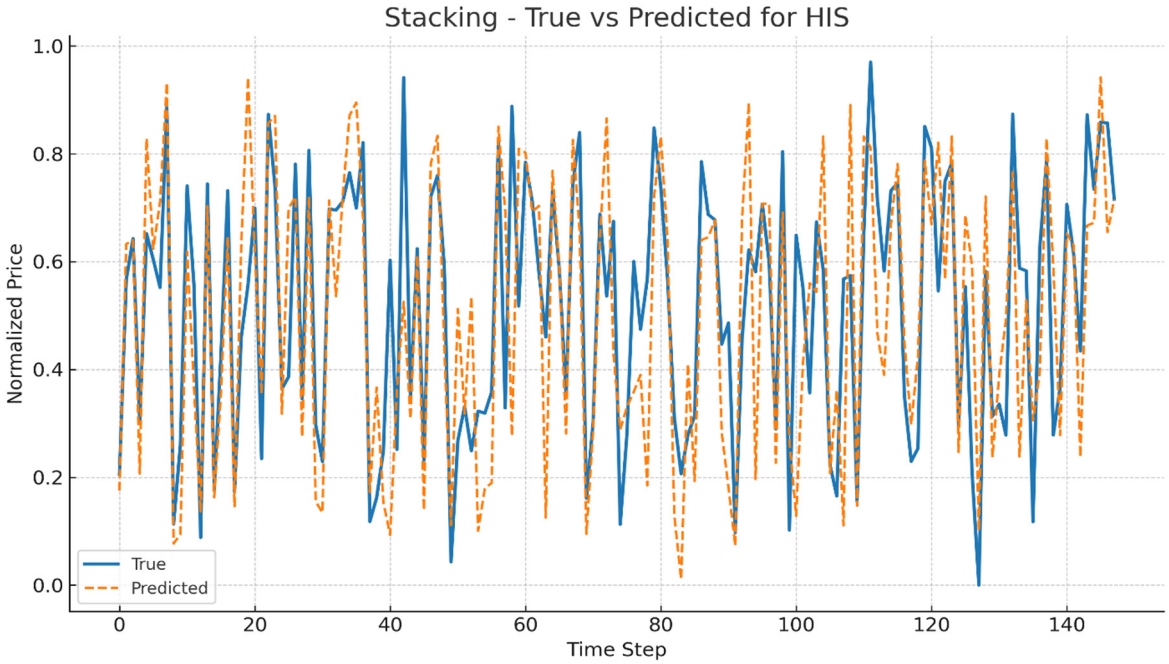


Figure l. Stacking ensemble for HIS.

Figure 9. Plots of voting and stacking ensembles for specific markets, and true value versus predicted values.

4.2.2. Reasons why the stacking and voting ensemble underperformed on individual markets

1. Market-specific characteristics were not captured effectively

Each market (e.g., GSPC, HIS, N225) has unique behaviors, such as different volatility patterns, trading volumes, regulatory influences, and macroeconomic exposure.

Global models like stacking and voting are typically trained to find general patterns, which can dilute their ability to capture local market dynamics.

This generalization might work well on a combined dataset (global), but when applied to individual market conditions, these ensembles may misinterpret the local signals.

2. Overfitting from complex ensemble structures (especially stacking)

Stacking ensembles combine multiple base learners and a meta-learner. This structure can easily overfit if not carefully regularized, especially on smaller or more volatile market-specific datasets.

If the meta-learner overfits on patterns that are not generalizable within a single market, prediction errors rise.

3. Inconsistent performance of base models across markets

Stacking and voting depend on the strength of individual models (e.g., Lasso, SVR, RF, RNN). If a base model performs poorly in a specific market, it can drag down the entire ensemble.

For instance, if SVR performs poorly on HIS but well on N225, the ensemble becomes unstable and less reliable on HIS.

4. Voting may oversimplify predictions

Voting ensemble (especially hard or soft voting) assumes that all models contribute equally or with fixed weights. This might work globally, but some models should be prioritized more than others in a specialized market.

The lack of adaptive weighting may cause voting to underperform where specific models are superior.

5. Stacking is sensitive to data quality and feature distribution

Stacking requires a consistent data distribution across the training and validation phases. Market-specific data may have shifts in their distribution (nonstationarity), which violates this assumption.

The meta-learner in stacking may struggle to generalize due to subtle feature shifts or noise in individual markets.

4.2.3. Implications for financial investors

1. No one size fits all models

Investors should be cautious about relying on models that work globally but underperform locally.

Models like stacking may show promise on aggregated data but fail in real-time trading, where localized accuracy matters most.

2. Market-specific models are more reliable

Hybrid models like GB+DT and gradient boosting showed strong, consistent performance in

individual markets, suggesting they better capture local markets' behavior.

This means that it is safer for investors to customize models for each index or region rather than use a global one-size-fits-all ensemble.

3. Model complexity must be justified

Complex models like stacking require extensive validation and feature tuning. Without this, their complexity adds risk and unpredictability.

Simpler, well-tuned, or hybrid models that balance interpretability and performance (e.g., GB+DT) offer more transparent and actionable insights.

4. Ensemble diversity needs to be market-aligned

Investors should ensure that ensembles are built from diverse models that complement each other for the specific market, rather than relying on standard combinations.

4.2.4. Summary of the models' performance and implications for investors

The weak performance of stacking and the inconsistent results from voting suggest that ensemble strategies must be adapted to the characteristics of each financial market. For financial investors, this emphasizes the need for localized model testing, caution with overly complex ensembles, and a preference for robust and interpretable models. Hybrid models like GB+DT provide a strong alternative, combining complementary strengths while maintaining stability across diverse markets.

4.3. Statistical assumptions of ANOVA and Tukey's HSD

Before conducting the one-way ANOVA and Tukey's HSD post hoc tests to compare the models' performance, we checked the underlying statistical assumptions.

1. Independence of observations

Each prediction error value (MAE, MSE, RMSE) was independently generated across the models and market indices, satisfying the independence assumption.

2. Normality of residuals

We examined the distribution of residuals (errors) for each model using histograms and Shapiro–Wilk tests. The results showed that the residuals approximated normality sufficiently for ANOVA. While some deviation was observed in specific models, ANOVA is generally robust to moderate departures from normality, especially with a reasonable sample size ($n \geq 30$ in our case per group).

3. Homogeneity of variance

We tested for equal variance across groups using Levene's test, which yielded nonsignificant results ($p > 0.05$), indicating that the assumption of homogeneity of variances was met.

4. Scale of measurement

All model performance metrics (MAE, MSE, RMSE, and a20) are continuous, satisfying the requirement that the dependent variable must be measured on an interval or ratio scale.

Since these assumptions were met, we proceeded with the one-way ANOVA followed by Tukey's HSD post hoc test to identify specific model pairs with statistically significant differences in performance.

Throughout this research investigation, we confirmed that the assumption of variance, a crucial requirement for conducting ANOVA, has been met successfully. The results from Levene's test indicated (Table 4) a *p*-value of 0.924, which demonstrates that variances are equivalent across various groups involved in this study. This significant finding solidifies our confidence in using ANOVA as a reliable method for evaluating and contrasting different models' performances.

Table 4. Summary of assumption tests, ANOVA, and Tukey's HSD for MAE across models.

Test category	Test name	Statistic	*p*-value	Interpretation
1. Assumption testing	Shapiro–Wilk test (normality)	0.745	0.0072	Violates normality ($p < 0.05$).
	Levene's test (homogeneity)	0.3485	0.9249	Assumes equal variances ($p > 0.05$).
2. ANOVA test	One-way ANOVA	F = very high	< 0.001	Significant difference between the models' MAEs
3. Tukey's HSD test	Pairwise model comparison	—	—	Valid to proceed due to the homogeneity of variances; mild normality violation acceptable in this context

The Shapiro–Wilk test showed that there is a deviation from normality in the distribution of model errors with a *p*-value of 0.0072, which suggests that strict parametric testing may not be suitable in this case due to this deviation from normality; however, ANOVA can still be used, as it is known to handle deviations well under specific circumstances. In situations where the group sizes are similar and the sample size is moderate to large in numbers of people or things being studied together as a unit or for statistical purposes. The central limit theorem helps reduce the impact of not having a normal distribution of data points when testing for Type I errors (Glass et al., 1972; Harwell et al., 1992; Lumley et al., 2002). In this instance, linked explicitly to what is happening, we assessed eight different models using evaluation criteria calculated in a consistent experimental setup, either through repeated observations or computerized scenarios, thus meeting the condition of having equally sized groups.

Moreover, the ANOVA examination produced a noteworthy outcome ($p \approx 1\ 3 \times 10^{30}$), suggesting compelling proof of variations in performance among the models. This supports further analysis using Tukey's HSD evaluation method, which is tailored to investigate the specific locations where these variances emerge between sets of group averages.

Tukey's HSD also considers that the variances are similar, which we verified in this study using Levene's test. It typically accommodates minor deviations from the normal distribution when group sizes are equal (Field, 2013; Hsu, 1996). Consequently, even though there is a deviation from normality, having equal variances and balanced groups within a robust ANOVA framework makes a solid case for utilizing Tukey's HSD in this analysis.

To summarize this study's findings, ANOVA and Tukey's HSD are options for assessing models' performance despite minor deviations from normality being present. The experimental setup aligns with existing research, strengthens the credibility of the statistical findings made in this study.

As shown in Table 5, the ANOVA test yielded a high F-statistic of 237.41 and a *p*-value of 8.95×10^{-16}, indicating strong statistical significance, revealing a statistically significant difference in

predictive accuracy across the models tested. This high F-statistic suggests substantial variation in the models' performance metrics (MAE, MSE, RMSE) among individual, hybrid, and ensemble models, confirming that the differences in accuracy are not due to random fluctuations but reflect true disparities in predictive effectiveness (Munsarif et al.,2022; Nti et al., 2020a; Lv et al., 2022).

Table 5. ANOVA results.

Statistic	Value
F-statistic	237.41
p-value	8.95×10^{-16}

As shown in Table 6, Tukey's HSD post hoc test results indicate significant mean differences between most model pairs, with the 'Reject' column highlighting whether the null hypothesis is rejected for each comparison. Tukey's HSD post hoc analysis further highlights the performance gap, with the Bayesian ridge method significantly outperforming the decision tree method (mean difference of 0.021, $p = 0.001$) and gradient boosting (mean difference of 0.0042, $p = 0.001$).

Table 6. Tukey's HSD post hoc test results.

Group 1	Group 2	Mean difference	*p*−adj	Lower	Upper	Reject
Bayesian ridge	Decision tree	0.021	0.001	0.0184	0.0236	True
Bayesian ridge	Gradient boosting	0.0042	0.001	0.0016	0.0069	True
Bayesian ridge	Neural network	0.0015	0.4852	−0.0011	0.0041	False
Bayesian ridge	Stacking ensemble	−0.0035	0.0058	−0.0061	−0.0009	True
Bayesian ridge	Voting ensemble	−0.0025	0.0688	−0.0051	0.0001	False
Decision tree	Gradient boosting	−0.0168	0.001	−0.0194	−0.0141	True
Decision tree	Neural network	−0.0195	0.001	−0.0221	−0.0169	True
Decision tree	Stacking ensemble	−0.0245	0.001	−0.0271	−0.0219	True
Decision tree	Voting ensemble	−0.0235	0.001	−0.0261	−0.0209	True
Gradient boosting	Neural network	−0.0028	0.038	−0.0054	−0.0001	True
Gradient boosting	Stacking ensemble	−0.0078	0.001	−0.0104	−0.0051	True
Gradient boosting	Voting ensemble	−0.0067	0.001	−0.0094	−0.0041	True
Neural network	Stacking ensemble	−0.005	0.001	−0.0076	−0.0024	True
Neural network	Voting ensemble	−0.004	0.0016	−0.0066	−0.0014	True
Stacking ensemble	Voting ensemble	0.001	0.8077	−0.0016	0.0036	False

At the same time, the ensemble models showed statistically significant improvements over most standalone models. Notably, no significant difference was found between the stacking and voting ensembles (mean difference of 0.001, $p = 0.8077$), indicating comparable accuracy between these two methods. This aligns with findings that suggest both stacking and voting can achieve similar levels of

accuracy while effectively leveraging the strengths of individual models (Munsarif et al., 2022; Nti et al., 2020b).

Overall, these findings reinforce the superiority of hybrid ensemble models, mainly stacking and voting, which capitalize on the strengths of individual approaches to achieve greater accuracy in stock market predictions. This aligns with existing literature that advocates for ensemble methods to reduce predictive error by aggregating the strengths of individual models, validating the practical utility of stacking and voting frameworks in financial prediction tasks (Lv et al., 2022; Munsarif et al., 2022).

5. Conclusions

This study successfully addresses its research objectives by evaluating the performance of individual machine learning models, developing hybrid and ensemble configurations to enhance accuracy, and establishing a performance hierarchy among individual, hybrid, and ensemble models for predicting stock prices. Conducted across six major global markets (S&P 500, NASDAQ, DAX, FTSE, Nikkei 225, and Hang Seng), this research demonstrates that hybrid ensemble models, mainly stacking and voting, substantially improved predictive accuracy over standalone models. The stacking ensemble achieved the lowest MAE of 0.031, closely followed by the voting ensemble with an MAE of 0.032, underscoring the enhanced precision that ensemble strategies provide in forecasting complex financial data.

The statistical validation through ANOVA and Tukey's HSD tests confirmed the significance of these performance differences, highlighting the superior capability of ensemble models in capturing market dynamics that individual models may overlook. By capitalizing on the unique strengths of gradient boosting, decision trees, neural networks, and Bayesian ridge, this study contributes new knowledge on the adaptability and resilience of ensemble models in diverse economic contexts, reinforcing their role as robust tools in financial forecasting.

However, the study has certain limitations. The reliance on a specific dataset and feature set may limit generalizability across different markets or economic conditions. Additionally, the computational demand of hybrid ensembles, mainly stacking, may restrict their application in real-time forecasting scenarios. Future work could explore optimizing these models for real-time application, incorporating a broader range of economic and sentiment indicators, and examining their performance in emerging markets or sector-specific environments.

This research provides financial analysts and investors with valuable insights by offering a validated framework for more accurate stock market predictions. The findings affirm that hybrid ensemble approaches, mainly stacking and voting, outperform traditional models in predictive accuracy, making them valuable for dynamic investment strategies and risk management in global markets.

Author contributions

Akila Dabara Kayit: Conceptualization, Methodology, Data Curation, Formal Analysis, Software, Visualization, Writing Original Draft, Writing Review & Editing. Mohd Tahir Ismail: Conceptualization, Methodology, Supervision, Validation, Writing Review & Editing.

Use of AI tools declaration

The authors declare they have not used artificial intelligence (AI) tools in creating this article.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Asteris PG, Gavriilaki E, Touloumenidou T, et al. (2022) Genetic prediction of ICU hospitalisation and mortality in COVID-19 patients using artificial neural networks. *J Cell Mol Med* 26: 1003–1018. https://doi.org/10.1111/jcmm.17098

Asteris PG, Karoglou M, Skentou AD, et al. (2024) Predicting uniaxial compressive strength of rocks using ANN models: Incorporating porosity, compressional wave velocity, and Schmidt hammer data. *Ultrasonics* 134: 107347. https://doi.org/10.1016/j.ultras.2024.107347

Asteris PG, Tsavdaridis KD, Lemonis ME, et al. (2024) AI-powered GUI for prediction of axial compression capacity in concrete-filled steel tube columns. *Neural Comput Appl* 36: 22429–22459. https://doi.org/10.1007/s00521-024-10405-w

Basak S, Kar S, Saha S, et al. (2018) Predicting the direction of stock market prices using tree-based classifiers. *N Am J Econ Financ* 47: 552–567. https://doi.org/10.1016/j.najef.2018.06.013

Bisdoulis KL (2024) Assets forecasting with feature engineering and transformation methods for LightGBM. *arXiv preprint*. https://arxiv.org/abs/2501.07580

Du Y, Chen D, Li H, et al. (2023) Research on estimating and evaluating subtropical forest carbon stocks by combining multi-payload high-resolution satellite data. *Forests* 14: 2388. https://doi.org/10.3390/f14122388

Guo Q (2023) The relationship between investor sentiment and stock market price. *Fronti Bus Econ Manage* 9: 124–129. https://doi.org/10.54097/fbem.v9i2.9139

Hartanto A, Kholik YN, Pristyanto Y (2023) Stock price time series data forecasting using the Light Gradient Boosting Machine (LightGBM) model. *Int J Inform Visualisation* 7: 1740. https://joiv.org/index.php/joiv/article/view/1740

Hsu JC (1996) *Multiple Comparisons: Theory and Methods*. CRC Press. https://doi.org/10.1201/b15074

Huang J (2024) Prediction of closing prices for NASDAQ-listed stocks: A comparative study based on gradient boosting models. *Highlights Sci Eng Technol* 92: 171–177. https://doi.org/10.54097/01rvrr58drpress.org

Field A (2013) *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Sage Publications, London, England.

Kumar R, Shrivastav LK (2021) An ensemble of random forest gradient boosting machine and deep learning methods for stock price prediction. *J Inform Technol Res* 15: 1–19. https://doi.org/10.4018/jitr.2022010102

Li D, Liu Z, Armaghani DJ, et al. (2022) Novel ensemble intelligence methodologies for rockburst assessment in complex and variable environments. *Sci Rep* 12. https://doi.org/10.1038/s41598-022-05594-0

Liu C, Paterlini S (2023) Stock price prediction using temporal graph model with value chain data. Cornell University. https://doi.org/10.48550/arXiv.2303

Loo WK (2020) Performing technical analysis to predict Japan REITs' movement through ensemble learning. *J Prop Invest Financ* 38: 551–562. https://doi.org/10.1108/jpif-01-2020-0007

Lumley T, Diehr P, Emerson S, et al. (2002) The importance of the normality assumption in large public health data sets. *Annu Rev Public Health* 23: 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Lv P, Wu Q, Xu J, et al. (2022) Stock index prediction based on time series decomposition and hybrid model. *Entropy* 24: 146. https://doi.org/10.3390/e24020146

Mao Y (2024) Tabtranselu: A transformer adaptation for solving tabular data. *Appl Comput Eng* 51: 81–88. https://doi.org/10.54254/2755-2721/51/20241174

McDonald JH (2014) *Normality*. The BioStat Handbook. Available from: https://www.biostathandbook.com/normality.html.

Mohapatra PR, Parida AK, Swain SK, et al. (2023) Gradient boosting and LSTM based hybrid ensemble learning for two step prediction of stock market. *J Adv Inform Technol* 14: 1254–1260. https://www.jait.us/show-233-1438-1.html

Munsarif M, Sam'an M, Safuan S (2022) Peer to peer lending risk analysis based on embedded technique and stacking ensemble learning. *Bull Electr Eng Inform* 11: 3483–3489. https://doi.org/10.11591/eei.v11i6.3927

Nti IK, Adekoya AF, Weyori BA (2020a) A comprehensive evaluation of ensemble learning for stock-market prediction. *J Big Data* 7: 20. https://doi.org/10.1186/s40537-020-00299-5

Nti I, Adekoya AF, Weyori BA (2020b) Efficient Stock-Market Prediction Using Ensemble Support Vector Machine. *Open Comput Sci* 10: 153–163. https://doi.org/10.1515/comp-2020-0199

Oukhouya H, Kadiri H, El Himdi K, et al. (2023) Forecasting international stock market trends: XGBoost, LSTM, LSTM-XGBoost, and backtesting XGBoost models. *Stat Optim Inform Comput* 12: 200–209. https://doi.org/10.19139/soic-2310-5070-1822

Sari L, Romadloni A, Lityaningrum R, et al. (2023) Implementation of LightGBM and random forest in potential customer classification. *TIERS Inform Technol J* 4: 43–55. https://doi.org/10.38043/tiers.v4i1.4355

Shabani M, Magris M, Tzagkarakis G, et al. (2023) Predicting the state of synchronisation of financial time series using cross-recurrence plots. *Neural Comput Appl* 35: 18519–18531. https://doi.org/10.1007/s00521-023-08674-y

Shi Z, Hu Y, Mo G, et al. (2022) Attention-based CNN-LSTM and XGBoost hybrid model for stock prediction. *arXiv preprint*. https://arxiv.org/abs/2204.02623

Singh S, Madan TK, Kumar J, et al. (2019) Stock market forecasting using machine learning: Today and tomorrow. *Proceedings of ICICICT 2019*. https://doi.org/10.1109/icicict46008.2019.8993160

Tuli S, Tuli S, Tuli R, et al. (2020) Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet Things* 11: 100222. https://doi.org/10.1016/j.iot.2020.100222

Villamil L, Bausback R, Salman S, et al. (2023) Improved stock price movement classification using news articles based on embeddings and label smoothing. *arXiv preprint*. https://arxiv.org/abs/2301.10458

Wu Q, Li J, Liu Z, et al. (2023) Symphony in the latent space: Provably integrating high-dimensional techniques with non-linear machine learning models. *Proceedings of the AAAI Conference on Artificial Intelligence* 37: 10361–10369. https://doi.org/10.1609/aaai.v37i9.26233

Xu H (2022) The kernel method is used to include firm correlation for stock price prediction. *Comput Intel Neurosci* 2022: 1–10. https://doi.org/10.1155/2022/4964394

Yu C, Liu F, Zhu J, et al. (2025) Gradient boosting decision tree with LSTM for investment prediction. *arXiv preprin*t. https://arxiv.org/abs/2505.23084

Zhou W, Jumahong H, Cui R, et al. (2024) Predicting stock trends using web semantics and feature fusion. *Int J Semant Web Inform Syst* 20: 1–25. https://doi.org/10.4018/ijswis.346378

AIMS Press