

A Confidence-Calibrated Adaptive Ensemble Voting Framework for Stock Price Direction Prediction

Yuan Lukito*, Nugroho Agus Haryono, Raden Gunawan Santosa
Department of Informatics, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia

Abstract—Predicting stock price direction is crucial for investment decisions, yet most machine learning models optimize classification metrics while ignoring probability calibration. Uncalibrated probabilities often produce overconfident predictions, increasing financial risk. To address this gap, we introduce Confidence-Calibrated Adaptive Ensemble Voting (CAEV-C), a framework combining temperature scaling calibration with dynamic confidence-weighted aggregation. We evaluate CAEV-C using XGBoost, LightGBM, and Bidirectional Long Short-Term Memory (BiLSTM) base learners across seven technology stocks (AAPL, AMZN, GOOGL, META, MSFT, NVDA, and TSLA) under a temporal walk-forward validation scheme. CAEV-C achieves the highest average F1-score of 0.5885, outperforming LightGBM (0.4399), Hard Voting (0.5391), and Soft Voting (0.4451). Furthermore, CAEV-C yields the lowest Expected Calibration Error (ECE) of 0.0347, improving calibration quality over LightGBM (0.2053), Hard Voting (0.2629), and Soft Voting (0.1062). Reliability diagrams confirm that CAEV-C confidence estimates track ideal calibration closely, while Wilcoxon Signed-Rank Tests demonstrate statistically significant ECE improvements over Hard and Soft Voting, as well as significant F1-score gains over Soft Voting ($p < 0.05$). By improving probability reliability while maintaining competitive predictive performance, CAEV-C provides a robust ensemble framework for quantitative stock forecasting.

Keywords—Stock market prediction; ensemble learning; confidence-calibrated ensemble; probability calibration

I. INTRODUCTION

Stock price direction prediction helps investors, financial institutions, and researchers guide trading decisions, portfolio optimization, and risk management. However, forecasting market movements remains difficult because financial time series are noisy, non-stationary, and nonlinear [1]. Macroeconomic conditions, investor sentiment, and unexpected events continuously alter market dynamics, limiting the effectiveness of traditional analytical models. These challenges drive ongoing research into computational methods that handle complex financial time-series data more reliably.

To overcome these limits, researchers frequently turn to machine learning and deep learning algorithms [2]. Gradient boosting models like eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) capture complex nonlinear patterns in financial features [3], whereas recurrent architectures such as Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (BiLSTM) process sequential market dependencies [4].

However, single models perform unevenly across different stocks and market regimes. This inconsistency reduces generalization, leading researchers toward ensemble learning methods that combine diverse models.

Ensemble learning improves prediction robustness by combining outputs from multiple models to reduce individual biases. Standard aggregation techniques include Hard Voting, which takes majority decisions, and Soft Voting, which averages predicted probabilities to incorporate confidence scores [5]. Yet conventional voting mechanisms assume these predicted probabilities are reliable. In practice, machine learning and deep learning models produce uncalibrated probabilities, yielding overconfident or underconfident estimates. Poorly calibrated confidence scores distort ensemble decisions [6] by hiding actual market uncertainty, making probability-weighted voting less reliable.

Stock forecasting studies typically evaluate models using classification metrics such as Accuracy, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC) [1]. While these metrics measure discrimination ability, they do not show whether predicted probabilities reflect actual outcomes. A model can achieve strong classification scores while producing poorly calibrated probabilities, yielding overconfident or underconfident estimates [7]. This gap matters in finance, where trading decisions rely on both predicted direction and prediction confidence. Calibration metrics like Expected Calibration Error (ECE) and Brier Score measure how well probabilities align with real outcomes. However, calibration quality remains rarely evaluated in stock direction research, particularly within ensemble learning frameworks.

Although ensemble learning is widely used in stock market forecasting [5], researchers rarely evaluate the reliability of the confidence estimates used during aggregation. Most studies combine model outputs via Hard or Soft Voting without first calibrating prediction probabilities. Consequently, uncalibrated base predictions carry over into the ensemble output, degrading decision reliability. Most stock market forecasting studies emphasize classification accuracy while overlooking calibration quality. Addressing this gap requires an ensemble framework that combines predictions from multiple models while calibrating probabilities to match real-world outcomes.

To address these limitations, we propose Confidence-Calibrated Adaptive Ensemble Voting (CAEV-C) for stock price direction prediction. The framework integrates predictions from three diverse base models: XGBoost, LightGBM, and BiLSTM. Before ensemble aggregation,

*Corresponding author.

temperature scaling calibrates each model's probability outputs to reduce overconfidence and improve reliability. The framework then uses these calibrated probabilities in a confidence-weighted aggregation scheme, weighting predictions by their confidence levels. By combining calibration with confidence-weighted voting, CAEV-C improves probability reliability while maintaining competitive predictive performance for financial decision-making.

We evaluate CAEV-C using walk-forward validation across seven technology stocks: AAPL, AMZN, GOOGL, META, MSFT, NVDA, and TSLA. CAEV-C achieves the highest average F1-score of 0.5885, outperforming Hard Voting (0.5391) and Soft Voting (0.4451). It also records the lowest Expected Calibration Error (ECE) of 0.0347, improving calibration over Hard Voting (0.2629) and Soft Voting (0.1062). Wilcoxon Signed-Rank Tests confirm statistically significant ECE gains over both baselines and significant F1-score improvements over Soft Voting.

This study offers three main contributions: 1) a confidence-calibrated ensemble voting framework (CAEV-C) for stock price direction prediction, 2) A pipeline combining temperature scaling with confidence-weighted aggregation to calibrate probabilities before decision-making, 3) An empirical evaluation across multiple stocks, calibration metrics, and statistical significance tests to validate predictive performance and reliability.

II. RELATED WORK

A. Stock Price Direction Prediction Using Machine Learning and Deep Learning

Broader access to financial data and faster computing have driven the use of machine learning for predicting stock price directions. Unlike traditional statistical models that rely on strict assumptions about data distributions and linearity, machine learning algorithms can capture complex nonlinear patterns from large datasets [8], [9], [10]. Tree-based methods like Random Forest [11], XGBoost [3], [12], and LightGBM [13], [14] are widely used in stock forecasting for their strong predictive performance, robustness, and ability to handle varied feature types.

Beyond traditional machine learning, deep learning models are widely used for financial time-series forecasting. Recurrent neural networks, specifically LSTM [15], [16], [17], [18] and BiLSTM [19], help capture time-dependent patterns in historical market data. These architectures overcome standard neural networks' difficulty with long sequences, allowing them to track shifting market trends over time. Consequently, deep learning provides a strong alternative for stock direction prediction, particularly when price movements rely on complex sequential patterns.

While machine learning and deep learning models perform well, individual model accuracy varies across different stocks, market conditions, and time horizons. Tree-based models excel at finding feature interactions, whereas deep learning architectures are better at tracking temporal patterns. Because no single model outperforms all others in every scenario, combining predictors through ensemble learning allows models

to complement each other, improving overall forecast stability and generalization.

B. Ensemble Learning for Stock Market Prediction

Ensemble learning is widely used in stock market forecasting to address the limits of individual models [20], [21], [22], [23], [24]. By combining predictions from multiple algorithms, ensemble methods leverage each model's strengths to offset individual weaknesses. Merging these outputs lowers prediction variance and improves overall stability and generalization. Given the high uncertainty and shifting conditions in financial markets, ensemble approaches tend to yield more reliable results than single models across different tasks and environments.

Hard Voting and Soft Voting are two of the most common aggregation techniques in ensemble learning [5]. Hard Voting bases the final prediction on a majority vote across classifiers, which is straightforward to implement and resilient to isolated model errors. Soft Voting instead averages predicted probabilities from each model and selects the class with the highest combined score. By incorporating probability estimates rather than just binary votes, Soft Voting provides more nuanced decisions than simple majority voting. Consequently, financial forecasting research frequently uses both techniques as standard baselines.

Despite their usefulness, traditional voting ensembles have clear drawbacks. Hard Voting relies entirely on class labels and ignores prediction confidence, discarding helpful probability data. Soft Voting assumes that a model's predicted probabilities are accurate and well calibrated. In practice, machine learning and deep learning models often produce overconfident or underconfident estimates [6], which skews aggregated probabilities away from the true likelihood of outcomes. Poor calibration can undermine Soft Voting, pointing to the need for ensemble methods that directly factor in probability reliability when merging predictions.

C. Confidence Calibration and Probability Reliability

Although high accuracy is a primary goal in stock market forecasting, accurate classification alone does not guarantee trustworthy probability estimates. A model can achieve a high Accuracy, F1-score, or ROC-AUC while outputting probabilities that misrepresent the true likelihood of an outcome. For instance, if a model assigns 90% confidence to a set of predictions, a properly calibrated system should be right about 90% of the time. In practice, machine learning and deep learning models often display overconfidence or underconfidence, creating a gap between predicted probabilities and actual outcomes. Because financial decisions rely on both predicted directions and confidence levels, miscalibration can skew risk assessments and lead to poor choices.

Researchers use several calibration techniques to make probabilistic predictions more reliable, including Platt Scaling, Isotonic Regression, and Temperature Scaling [25]. Temperature Scaling is especially popular because it is simple and effective. It introduces a single temperature parameter to adjust output confidence without changing prediction rankings, keeping classification decisions intact while improving

probability estimates. Because it is computationally efficient and performs well in practice, Temperature Scaling is widely used as a post-processing step for machine learning and deep learning models.

Several metrics help assess probability calibration quality [26]. Expected Calibration Error (ECE) measures the gap between predicted confidence and actual accuracy across confidence bins. The Brier Score calculates the mean squared error between predicted probabilities and actual outcomes to evaluate probability accuracy. Researchers also use reliability diagrams to visually compare prediction confidence against empirical accuracy. Together, these tools provide complementary views on whether a model's probabilities are trustworthy for real-world decision-making.

Despite the importance of confidence reliability, few stock forecasting studies evaluate model calibration. Most research focuses on standard classification metrics like Accuracy, F1-score, and ROC-AUC, paying little attention to whether predicted probabilities are trustworthy. Furthermore, while ensemble learning is widely used to boost forecasting accuracy, combining probability calibration directly with ensemble aggregation remains rare. This gap creates an opportunity for ensemble frameworks that incorporate confidence calibration, delivering reliable probability estimates alongside strong predictive performance.

D. Research Gap and Research Positioning

Recent research has advanced stock price direction forecasting using machine learning, deep learning, and ensemble techniques. Machine learning algorithms like XGBoost and LightGBM fit nonlinear patterns in financial data, while deep learning architectures such as LSTM and BiLSTM effectively model temporal dependencies. Ensemble methods, particularly Hard Voting and Soft Voting, combine multiple models to improve prediction stability. However, existing studies focus primarily on standard classification metrics like Accuracy, F1-score, and ROC-AUC, paying far less attention to the reliability of predicted probabilities. Consequently, confidence calibration remains largely underexplored in stock market prediction.

While probability calibration is common in general machine learning, its use in ensemble-based stock prediction remains rare. Most ensemble methods aggregate model outputs without calibrating probabilities before voting, which can pass unreliable confidence estimates into final predictions. This study addresses that issue by introducing the Confidence-Calibrated Adaptive Ensemble Voting (CAEV-C) framework. CAEV-C pairs temperature scaling calibration with confidence-aware ensemble aggregation. Unlike standard voting approaches that focus strictly on accuracy, CAEV-C evaluates both predictive performance and confidence reliability to deliver reliable forecasts for financial decision-making.

CAEV-C combines temperature scaling probability calibration with confidence-aware ensemble aggregation, allowing the system to weigh model predictions alongside their confidence reliability. By evaluating accuracy and calibration quality together through the F1-score, Brier Score, and

Expected Calibration Error (ECE), the framework provides a more reliable ensemble forecasting approach for financial decisions.

III. METHODOLOGY

A. Proposed CAEV-C Framework

This study introduces the Confidence-Calibrated Adaptive Ensemble Voting (CAEV-C) framework for stock price direction prediction. By unifying diverse models, probability calibration, and confidence-aware voting into a single pipeline, the framework improves both forecasting accuracy and probability reliability. Unlike conventional ensembles that combine raw model outputs directly, CAEV-C calibrates each model's probability estimates before voting. This step reduces the impact of overconfident or underconfident predictions, leading to more reliable forecasts.

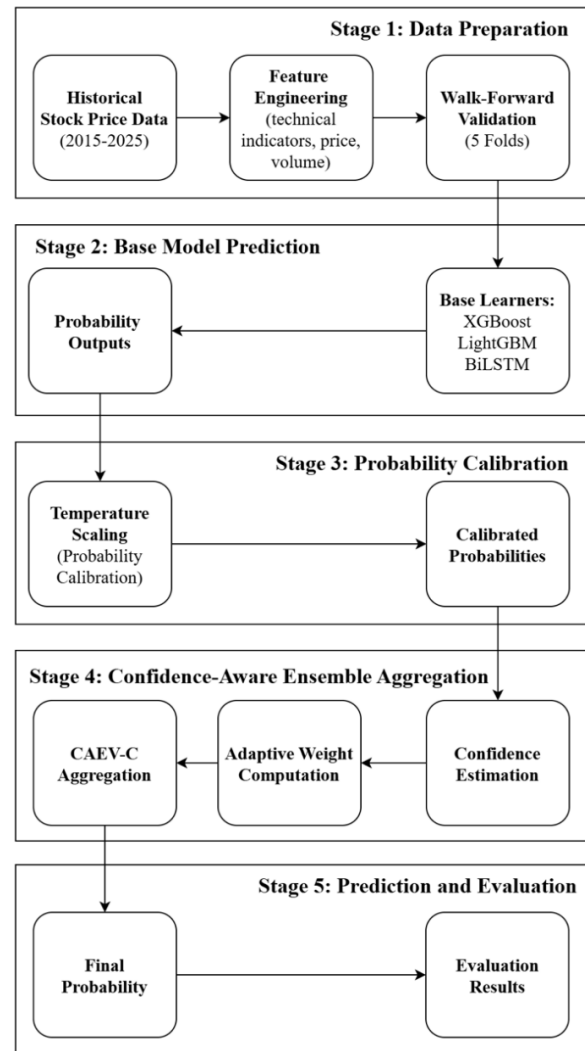


Fig. 1. Overall architecture of the proposed CAEV-C framework consisting of a five-stage workflow.

The framework follows a five-stage workflow as depicted in Fig. 1. First, historical market data is transformed into predictive features and split using walk-forward validation to

preserve temporal order. Second, three distinct base models (XGBoost, LightGBM, and BiLSTM) are trained independently to capture different patterns in financial time series. Third, Temperature Scaling calibrates each model's probability outputs to align predicted confidence with observed outcomes. Fourth, these calibrated probabilities determine prediction confidence levels and calculate adaptive voting weights. Finally, the framework aggregates the calibrated probabilities through confidence-aware voting to produce the final stock direction forecast.

Combining probability calibration with adaptive weighting allows the ensemble to evaluate both predicted class labels and the reliability of their confidence scores. As a result, CAEV-C balances classification accuracy and probability reliability better than standard Hard Voting or Soft Voting.

B. Base Prediction Models

The CAEV-C framework uses three distinct base learners (XGBoost, LightGBM, and BiLSTM) to capture different characteristics of stock market data. Combining diverse models increases prediction diversity, a key factor in ensemble performance. Tree-based algorithms excel at fitting nonlinear relationships in engineered financial features, while recurrent neural networks capture temporal patterns in sequential market data. Merging these complementary approaches provides the ensemble with varied predictive perspectives before calibration and aggregation.

1) *XGBoost*: eXtreme Gradient Boosting (XGBoost) is a gradient-boosted decision tree algorithm that sequentially constructs trees to minimize a specified loss function. Each new tree corrects the residual errors of earlier trees, allowing the model to learn complex nonlinear relationships in data. Because of its built-in regularization and efficient optimization, XGBoost is widely used in stock market prediction and financial forecasting.

For a dataset with N observations and F^* representing the number of features, the XGBoost prediction is given by Eq. (1), where $f_k \in F$, denotes the k -th decision tree, F represents the space of regression trees, and K represents the total number of trees in the ensemble.

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (1)$$

For binary stock direction prediction, the model outputs the estimated probability of an upward price movement, which is then used in the calibration and ensemble aggregation stages.

2) *LightGBM*: Light Gradient Boosting Machine (LightGBM) is a gradient boosting framework built for high speed and scalability without sacrificing predictive accuracy. Unlike traditional tree-based algorithms that grow trees level by level, LightGBM uses a leaf-wise growth strategy, expanding whichever leaf offers the highest information gain. This allows the model to achieve high accuracy with fewer trees.

For a given instance x_i , the LightGBM prediction is expressed as in Eq. (2), where g_k denotes the k -th boosted tree in the LightGBM ensemble.

$$\hat{y}_i = \sum_{k=1}^K g_k(x_i) \quad (2)$$

LightGBM efficiently learns feature interactions, making it well-suited for stock market prediction, where relationships between technical indicators and market variables are often nonlinear and dynamic. As with XGBoost, LightGBM's probability outputs are calibrated before being integrated into the ensemble framework.

3) *Bidirectional Long Short-Term Memory (BiLSTM)*: Bidirectional Long Short-Term Memory (BiLSTM) extends the standard LSTM architecture by processing sequential data in both forward and backward directions. By combining past and future context within a sequence, BiLSTM captures richer temporal representations than unidirectional LSTM networks.

Given an input sequence as expressed in Eq. (3):

$$X = \{x_1, x_2, \dots, x_T\} \quad (3)$$

The forward and backward hidden states are computed as in Eq. (4) and Eq. (5):

$$\vec{h}_t = LSTM_{forward}(x_t, \vec{h}_{t-1}) \quad (4)$$

$$\overleftarrow{h}_t = LSTM_{backward}(x_t, \overleftarrow{h}_{t+1}) \quad (5)$$

The final hidden representation is obtained by concatenating both directions as expressed in Eq. (6), where $[:]$ denotes the concatenation operation.

$$h_t = [\vec{h}_t; \overleftarrow{h}_t] \quad (6)$$

In stock price direction prediction, BiLSTM captures temporal dependencies and sequential patterns that tree-based models may miss. Its output layer estimates the probability of an upward price movement, which is then calibrated and integrated into the CAEV-C aggregation process.

The predicted probabilities from these three base learners serve as the inputs for the calibration stage. To improve confidence reliability before ensemble aggregation, the framework calibrates these outputs using Temperature Scaling, as detailed in the following subsection.

C. Probability Calibration Using Temperature Scaling

Probabilities from machine learning and deep learning models often serve as confidence scores, but high accuracy does not guarantee well-calibrated probabilities. Models frequently yield overconfident probabilities that exceed true event frequencies, or underconfident ones that are overly conservative. Such calibration errors pose risks in stock forecasting because investment decisions depend on predicted market directions alongside their confidence levels. Improving probability reliability is therefore an essential step before probability-based ensemble aggregation.

To address this issue, the framework uses Temperature Scaling, a post-hoc calibration technique that adjusts model confidence without changing prediction rankings. Unlike calibration methods that require model retraining, Temperature Scaling introduces a single scalar parameter T to rescale output logits. This approach is computationally efficient, reduces

prediction overconfidence, and preserves the original model's discriminative ability.

If z represents a classifier's logit before probability transformation, Temperature Scaling computes the calibrated probability as expressed in Eq. (7), where $\hat{p}$ denotes the calibrated probability, z represents the original logit output, T denotes the temperature parameter and $\sigma(\cdot)$ is the sigmoid activation function.

$$\hat{p} = \sigma\left(\frac{z}{T}\right) \quad (7)$$

The temperature parameter controls the sharpness of the probability distribution. At $T = 1$, probability estimates remain unchanged. Setting $T > 1$ softens the distribution, reducing overconfidence, while $T < 1$ sharpens it, increasing prediction confidence.

The framework optimizes an independent temperature parameter for each base learner (XGBoost, LightGBM, and BiLSTM). For each walk-forward fold, the optimal temperature is determined on the validation set by minimizing Negative Log-Likelihood (NLL). Given a validation dataset D_{val} with N samples, the optimization objective is defined as in Eq. (8).

$$T^* = \arg \min_T \left(-\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)] \right) \quad (8)$$

T^* denotes the optimal temperature parameter, y_i represents the ground-truth class label and $\hat{p}_i$ is the calibrated probability of sample i .

Once the optimal temperature parameters are set, the calibrated probabilities serve as inputs to the confidence estimation stage. Reducing calibration errors before ensemble aggregation ensures that the confidence values used during voting accurately reflect each model's true reliability. This calibrated confidence information provides the foundation for the CAEV-C mechanism, detailed in the following subsection.

D. Confidence Estimation

After probability calibration, the framework estimates the confidence of each model prediction. This step forms the foundation of CAEV-C's adaptive aggregation mechanism, quantifying individual prediction reliability before combining outputs. Unlike conventional voting approaches that treat all predictions equally, CAEV-C assigns greater weight to predictions with higher calibrated confidence, allowing the ensemble to adapt to each model's reliability.

Let $\hat{P}_m$ denote the calibrated probability generated by model m , where $\hat{P}_m \in [0, 1]$ and $m \in \{XGBoost, LightGBM, BiLSTM\}$. The confidence score is the maximum probability assigned to either class as expressed in Eq. (9):

$$c_m = \max(\hat{P}_m, 1 - \hat{P}_m) \quad (9)$$

where, c_m denotes the confidence score of m , and $\hat{P}_m$ represents the calibrated probability of the positive class. This yields confidence values within the interval $0.5 \leq c_m \leq 1.0$.

A confidence value near 0.5 reflects high uncertainty because both classes have nearly equal probabilities. Conversely, a value approaching 1.0 shows that the model strongly favours one class over the other. Because confidence scores come from calibrated probabilities rather than raw model outputs, they more accurately reflect the true reliability of individual predictions. This reliability is vital in financial forecasting, where prediction confidence directly influences investment decisions.

The estimated confidence scores are then converted into adaptive voting weights. Models with higher calibrated confidence carry greater weight during ensemble aggregation, while lower-confidence models contribute proportionally less to the final forecast. By incorporating calibrated confidence into the voting process, the framework reduces the impact of unreliable predictions and improves overall ensemble robustness.

The confidence scores from this stage are used in the next subsection to calculate adaptive voting weights and construct the CAEV-C mechanism.

E. Proposed Confidence-Calibrated Adaptive Ensemble Voting (CAEV-C)

This subsection details the Confidence-Calibrated Adaptive Ensemble Voting (CAEV-C) framework. CAEV-C combines calibrated probability estimates from multiple heterogeneous base learners, adjusting each model's contribution according to its calibrated confidence level. Unlike Hard Voting, which considers only class labels, and Soft Voting, which averages probabilities without considering confidence reliability, CAEV-C integrates probability calibration and confidence-aware weighting into a single decision-making process.

The goal of CAEV-C is to improve both predictive accuracy and probability reliability. Giving greater weight to models with higher calibrated confidence reduces the influence of less reliable predictions on the final ensemble output.

1) *Calibrated probability generation*: After training, the three base learners (XGBoost, LightGBM, and BiLSTM) each output a probability estimate for the positive class: $P_{xgboost}, P_{lightgbm}, P_{bilstm}$. These estimates are then calibrated using the Temperature Scaling procedure from the previous subsection, yielding calibrated probabilities: $\hat{P}_{xgboost}, \hat{P}_{lightgbm}, \hat{P}_{bilstm}$. This calibration aligns predicted confidence more closely with observed outcomes before the probabilities enter the ensemble aggregation stage.

2) *Confidence-aware weight computation*: For each calibrated probability, the confidence score is calculated as given in Eq. (9). The confidence scores are then normalized to generate adaptive voting weights as expressed in Eq. (10):

$$w_m = \frac{c_m}{\sum_{j=1}^M c_j} \quad (10)$$

where, w_m denotes the adaptive weight assigned to model m and M represents the total number of base learners. This normalization process ensures that the sum of the weights across all M components sums to 1. Consequently, models

with higher calibrated confidence receive larger voting weights and carry more influence in the final ensemble prediction.

3) *Confidence-calibrated ensemble aggregation*: The ensemble probability is calculated through a weighted combination of calibrated probabilities as expressed in Eq. (11).

$$P_{CAEV-C} = \sum_{m=1}^M w_m \hat{p}_m \quad (11)$$

For the three-model ensemble used in this study, the aggregation can be expressed as Eq. (12), where the sum of the adaptive weight is equal to 1, as expressed in Eq. (12).

$$w_{xgboost} + w_{lightgbm} + w_{bilstm} = 1 \quad (12)$$

This formulation allows the final ensemble prediction to account for both predicted probabilities and their underlying reliability.

4) *Final decision rule*: The final stock price direction is set using a binary classification threshold as expressed in Eq. (13), where $\hat{y} = 1$ indicates an upward stock price movement and $\hat{y} = 0$ indicates a downward stock price movement.

$$\hat{y} = \begin{cases} 1, & P_{CAEV-C} \geq 0.5 \\ 0, & P_{CAEV-C} < 0.5 \end{cases} \quad (13)$$

5) *Algorithm of CAEV-C*: The CAEV-C framework operates through the following steps as expressed in Algorithm 1.

Algorithm 1: CAEV-C

Input:

Historical stock features (X)
Base learners: XGBoost, LightGBM, BiLSTM

Output:

$\hat{y}$

Train each base learner independently

Generate probability predictions: $P_{xgboost}, P_{lightgbm}, P_{bilstm}$

Apply Temperature Scaling to obtain: $\hat{P}_{xgboost}, \hat{P}_{lightgbm}, \hat{P}_{bilstm}$

Compute calibrated confidence scores: $c_m = \max(\hat{p}_m, 1 - \hat{p}_m)$

Normalize confidence scores: $w_m = \frac{c_m}{\sum_{j=1}^M c_j}$

Aggregate calibrated probabilities: $P_{CAEV-C} = \sum_{m=1}^M w_m \hat{p}_m$

Generate final prediction: if $P_{CAEV-C} \geq 0.5$

$\hat{y}_i = 1$

else:

$\hat{y}_i = 0$

Return $\hat{y}_i$

IV. EXPERIMENTAL SETUP

A. Dataset Description

This study uses historical stock market data from seven major NASDAQ technology companies: Apple (AAPL), Amazon (AMZN), Alphabet (GOOGL), Meta Platforms (META), Microsoft (MSFT), NVIDIA (NVDA), and Tesla

(TSLA). These stocks were selected for their heavy trading volume, large market capitalization, and diverse price behaviors, offering a reliable test environment for evaluating model robustness and generalizability. The historical market data were obtained from Yahoo! Finance, as specified in Table I.

TABLE I. HISTORICAL MARKET DATA OBTAINED FROM YAHOO! FINANCE

Parameter	Description/Specification
Ticker symbols	AAPL, AMZN, GOOGL, META, MSFT, NVDA, TSLA
Observation Period	January 1st, 2015 – December 31st, 2025
Data Frequency	Daily
Data Attributes	Date, Open, High, Low, Close

The dataset contains daily trading information for each stock: open, high, low, close, adjusted close, and volume. These variables serve as inputs for feature engineering and forecasting. To ensure experimental consistency, the same preprocessing and feature generation steps were applied to every stock.

The forecasting task is framed as a binary classification problem to predict next-day stock price movements. Following standard practice in financial forecasting, target labels reflect changes in closing prices between consecutive trading days. An upward movement is assigned class 1, while a downward or flat movement is assigned class 0. The target variable is expressed as in Eq. (14), where $Close_t$ and $Close_{t+1}$ represents the closing price at trading day t and $t+1$, respectively.

$$y_t = \begin{cases} 1, & Close_{t+1} > Close_t \\ 0, & otherwise \end{cases} \quad (14)$$

Framing forecasting as a directional prediction task aligns better with real-world trading decisions than predicting exact prices. The resulting labeled dataset is then used for feature engineering, model training, probability calibration, and ensemble evaluation in the CAEV-C framework.

B. Feature Construction

Before model training, feature engineering converts raw stock market data into predictive variables. Because raw prices alone do not fully capture market conditions or investor behavior, this study uses eight technical indicators derived from historical stock data to track price returns, momentum, trend strength, and volatility.

Each of the eight indicators, as presented in Table II, highlights a specific market characteristic. Return measures relative price changes between consecutive trading periods. Momentum tracks the strength and persistence of price trends, while Moving Average Convergence Divergence (MACD) signals trend direction and momentum shifts. Average True Range (ATR) and Bollinger Bands measure market volatility. Finally, the Commodity Channel Index (CCI), Average Directional Index (ADX), and Stochastic Oscillator identify overbought or oversold conditions, trend strength, and momentum behavior.

TABLE II. TECHNICAL INDICATORS USED IN THIS STUDY

Category	Indicator	Parameters
Price Movement	Return	1 day
Momentum	Momentum Stochastic Oscillator	14 days 14 days
Trend and Trend Strength	ADX MACD	14 days 12 days (fast), 26 days (slow), 9 days (signal)
Market Condition	CCI	20 days
Volatility	ATR Bollinger Bands	14 days, 20 days (period), 2 (standard deviation)

Each technical indicator provides complementary information about market behavior. Momentum indicators capture the speed of price movements, trend indicators identify market direction, and volatility indicators measure market uncertainty. Combining these indicators gives both tree-based machine learning models and sequence-based deep learning architectures a solid representation of market activity.

All rolling technical indicators were calculated using backward-looking windows over historical price data. Feature values at time t were derived exclusively from observations up to and including trading day t , preventing future market information from contaminating feature construction.

Once feature construction is complete, the feature matrix serves as the input for XGBoost, LightGBM, and BiLSTM. The dataset is then split using a walk-forward validation strategy to preserve the temporal order of financial time-series data and prevent data leakage between training and testing periods.

C. Walk-Forward Validation Strategy

To prevent data leakage and ensure realistic model evaluation, this study uses a walk-forward validation strategy. Unlike random splitting, walk-forward validation preserves the chronological order of financial data by training models on past observations and evaluating them on future data. For each fold, models were trained on historical data, calibrated on a validation set, and evaluated on an unseen test set. After each iteration, the training window moved forward to include more recent observations. Table III presents the configuration of the five walk-forward folds.

TABLE III. THE CONFIGURATION OF THE FIVE WALK-FORWARD FOLDS

Folds	Train (Year)	Validation (Year)	Test (Year)
1.	2015 – 2019	2020	2021
2.	2015 – 2020	2021	2022
3.	2015 – 2021	2022	2023
4.	2015 – 2022	2023	2024
5.	2015 – 2023	2024	2025

To eliminate lookahead bias and prevent data leakage, training, validation, and test sets were partitioned in strict chronological sequence within each walk-forward fold. Feature scaling parameters, technical indicators, base-model training, and temperature calibration were executed strictly on data preceding the corresponding test period, ensuring no future

information influenced any stage of model estimation or calibration.

The validation set was further utilized to determine the optimal temperature parameter for each base learner during the probability calibration process. Performance metrics from all testing partitions were subsequently aggregated to obtain the final evaluation results. This strategy provides a realistic assessment of stock price direction prediction models while avoiding look-ahead bias and preserving temporal consistency.

D. Experimental Environment and Hyperparameter Settings

All experiments were run in Python using Scikit-learn, XGBoost, LightGBM, and PyTorch. Pandas and NumPy handled data preprocessing and feature engineering, while custom scripts managed model evaluation and calibration analysis. Computations ran on Apple Silicon hardware using Metal Performance Shaders (MPS) acceleration for deep learning training. Maintaining a single experimental environment ensures consistency and reproducibility across all models and ensemble configurations.

1) *Base model parameters:* To ensure fair comparison, every model shared the same walk-forward validation protocol, feature set, and evaluation procedures. Hyperparameters as presented in Table IV were selected during preliminary experiments on the validation set and kept fixed across all walk-forward folds. Because this study focuses specifically on probability calibration and confidence-aware ensemble aggregation, extensive hyperparameter tuning was intentionally avoided to prevent confounding performance variations.

TABLE IV. HYPERPARAMETER SETTINGS

Model	Hyperparameter
XGBoost	n_estimators = 100, max_depth = 6, learning_rate (eta) = 0.3, booster = gbtreesubsample = 1, colsample_bytree = 1, objective = 'binary:logistic', random_state = 42
LightGBM	n_estimators = 100, num_leaves = 31, max_depth = -1, boosting_type = 'gbdt' learning_rate = 0.1, objective = binary
BiLSTM	hidden_dim = 64, num_layers = 2 dropout = 0.2, bidirectional = True fully_connected_layer = 2, lookback = 10 epochs = 60, learning_rate = 0.003 batch_size = 32, optimizer = AdamW weight_decay = 0.0001

BiLSTM input sequences were formed using a fixed lookback window of 10 trading days, excluding the boundary samples that lacked a complete historical sequence. Across all walk-forward folds, base models were initialized from scratch and trained independently, ensuring that weights and states never carried between temporal splits. To ensure reproducibility, we trained BiLSTM models with a fixed random seed (42). We applied early stopping based on validation loss with a patience of 20 epochs to prevent overfitting, and computed class weights directly from each fold's training partition to account for class imbalance.

2) *Temperature scaling configuration:* For probability calibration, an independent temperature parameter was optimized for each base learner within every walk-forward fold. The optimal temperature was selected on the validation

set by minimizing negative log-likelihood loss. To prevent extreme calibration behavior, the temperature search space was restricted to $0.01 \leq T \leq 10.00$. The resulting calibrated probabilities were then used by the CAEV-C framework during confidence estimation and adaptive ensemble aggregation.

Temperature parameters were estimated exclusively on the validation split of each walk-forward fold and subsequently applied to calibrate test predictions, ensuring zero test-set exposure during optimization.

These model configurations and calibration steps form the experimental setup. To evaluate CAEV-C, the following subsection introduces the baseline methods used for comparison.

E. Baseline Methods

To evaluate CAEV-C, several baseline models representing different forecasting paradigms were selected for comparison. These include individual machine learning and deep learning models, alongside traditional ensemble approaches. This experimental setup tests whether the confidence-calibrated ensemble mechanism improves upon standalone predictors and standard voting-based ensembles.

The evaluated methods fall into three groups: individual predictive models (XGBoost, LightGBM, and BiLSTM), conventional ensemble methods (Hard Voting and Soft Voting), and the proposed CAEV-C framework. Individual models measure the standalone performance of each algorithm, while conventional ensembles evaluate the impact of combining multiple heterogeneous predictors.

1) *Individual prediction models*: To benchmark classification and calibration performance before ensemble aggregation, three models were evaluated as standalone baselines: XGBoost, LightGBM, and a Bidirectional Long Short-Term Memory (BiLSTM) network. XGBoost models non-linear interactions among technical indicators, LightGBM uses leaf-wise tree growth for efficient gradient boosting, and the BiLSTM captures temporal dependencies within financial time-series data. Evaluating each model independently establishes a baseline for comparison with the subsequent ensemble framework.

2) *Hard voting ensemble*: Hard voting determines the ensemble's prediction by taking a majority vote across individual classifiers. Let $\hat{y}_{xgboost}, \hat{y}_{lightgbm}, \hat{y}_{bilstm}$ denote the binary predictions from the three base learners; the hard voting prediction ($\hat{y}_{hardvoting}$) is calculated using Eq. (15).

$$\hat{y}_{hardvoting} = \text{mode}(\hat{y}_{xgboost}, \hat{y}_{lightgbm}, \hat{y}_{bilstm}) \quad (15)$$

3) *Soft voting ensemble*: Soft voting aggregates predictions by averaging class probabilities across individual models rather than using binary outputs. For three base learners, the aggregated probability is calculated using Eq. (16).

$$P_{softvoting} = \frac{P_{xgboost} + P_{lightgbm} + P_{bilstm}}{3} \quad (16)$$

where, $P_{xgboost}, P_{lightgbm}, P_{bilstm}$ are the predicted probabilities from XGBoost, LightGBM, and BiLSTM, respectively.

4) *Proposed CAEV-C framework*: The Confidence-Calibrated Adaptive Ensemble Voting (CAEV-C) framework extends soft voting by adding probability calibration and confidence-aware weighting. First, Temperature Scaling calibrates each base learner's probability output to improve confidence reliability. Second, adaptive voting weights are dynamically assigned based on these calibrated confidence levels before aggregation.

Combining calibration with confidence-aware weighting addresses key drawbacks of hard and soft voting to balance prediction accuracy with probability reliability. Comparing CAEV-C against these baselines evaluates its performance in predicting stock price direction.

F. Evaluation Metrics and Statistical Analysis

Evaluating CAEV-C requires assessing both classification performance and probability calibration. Classification metrics measure how accurately the model predicts stock price directions, while calibration metrics test whether predicted probabilities reflect actual outcome likelihoods. This dual evaluation is essential in financial forecasting because trading decisions depend on accurate predictions and reliable confidence estimates.

1) *Classification performance metrics*: All models were evaluated using five standard classification metrics: accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC).

2) *Calibration performance metrics*: Because CAEV-C focuses on improving confidence reliability, the evaluation also includes calibration metrics. Probabilistic prediction quality was measured using the Brier Score and Expected Calibration Error (ECE).

The Brier Score measures the mean squared difference between predicted probabilities and actual outcomes as expressed in Eq. (17), where N denotes the total number of observations, $\hat{p}_i$ represents the predicted probability, and y_i denotes the actual class label.

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N (\hat{p}_i - y_i)^2 \quad (17)$$

ECE measures the difference between prediction confidence and observed accuracy. Dividing predicted probabilities into M confidence bins, ECE is defined as given in Eq. (18), where B_m denotes the set of samples belonging to the m -th confidence bin, $\text{acc}(B_m)$ represents the empirical accuracy within bin B_m , and $\text{conf}(B_m)$ represents the average confidence within bin B_m .

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (18)$$

A lower ECE indicates better calibration and closer alignment between predicted confidence and actual outcomes. To complement these numerical metrics, reliability diagrams

visually evaluate calibration by plotting observed accuracy against predicted confidence across multiple intervals.

3) *Statistical significance analysis*: The Wilcoxon signed-rank test was used to determine whether performance differences between ensemble methods are statistically significant. This non-parametric test compares paired results across walk-forward folds without assuming a normal distribution.

The null and alternative hypotheses are: H_0 (there is no statistically significant difference between the compared methods, and H_1 (there is a statistically significant difference between the compared methods).

All tests used a significance level of $\alpha = 0.05$. We compared CAEV-C against Hard Voting and Soft Voting baselines across accuracy, F1-score, ROC-AUC, and ECE to verify whether performance gains reflect consistent improvements across evaluation folds rather than random variation.

V. RESULTS AND DISCUSSION

A. Performance of Base Models

Before evaluating ensemble aggregation, the three baseline models (XGBoost, LightGBM, and BiLSTM) were tested independently across all stocks and walk-forward folds. The evaluation covers both classification and calibration metrics. Table V summarizes the aggregated test results for each model.

TABLE V. OVERALL PERFORMANCE OF INDIVIDUAL PREDICTION MODELS

Metric	XGBoost	LightGBM	BiLSTM
Accuracy	0.4992 ± 0.0327	0.5033 ± 0.0380	0.4917 ± 0.0331
Precision	0.5497 ± 0.0911	0.5522 ± 0.0925	0.5294 ± 0.0620
Recall	0.4291 ± 0.2241	0.4350 ± 0.2332	0.4227 ± 0.2391
F1-Score	0.4393 ± 0.1510	0.4399 ± 0.1687	0.4244 ± 0.1720
ROC-AUC	0.5185 ± 0.0297	0.5259 ± 0.0299	0.5097 ± 0.0399
Brier Score	0.3404 ± 0.0449	0.3012 ± 0.0357	0.3377 ± 0.0394
ECE	0.2777 ± 0.0664	0.2053 ± 0.0680	0.2701 ± 0.0668

As shown in Table V, LightGBM achieved the best classification performance among the standalone models, recording the highest accuracy (0.5033), F1-score (0.4399), and ROC-AUC (0.5259). Its leaf-wise tree growth strategy likely helped it capture non-linear interactions among technical indicators better than the other architectures. XGBoost performed similarly to LightGBM with slightly lower metrics, while BiLSTM recorded the lowest F1-score (0.4244).

Although LightGBM led the standalone models, all three produced modest F1-scores ranging from 0.4244 to 0.4399, showing the difficulty a single architecture faces when modeling stock price movement. However, the performance differences between the tree-based and sequence-based models suggest they capture complementary signals from financial time series, supporting the use of a heterogeneous ensemble.

Calibration metrics followed a similar trend. LightGBM earned the lowest Brier score (0.3012) and ECE (0.2053),

outperforming XGBoost (ECE 0.2777) and BiLSTM (ECE 0.2701). Even so, ECE values remained high across all three models, showing a wide gap between predicted confidence and actual outcomes.

These results highlight a key issue with standalone models: reasonable classification accuracy does not guarantee calibrated confidence estimates. This gap justifies combining ensemble learning with probability calibration. The next experiment evaluates whether conventional hard and soft voting ensembles improve prediction accuracy and confidence calibration over individual models.

B. Hard Voting and Soft Voting

To evaluate ensemble learning for stock price direction prediction, we compared Hard Voting and Soft Voting against the individual baseline models. Although both methods combine outputs from heterogeneous base learners, they use different aggregation mechanisms: Hard Voting takes a majority vote of binary predictions, whereas Soft Voting averages predicted probabilities across models. Comparing these strategies shows how aggregation choice affects classification accuracy and probability reliability.

As shown in Table VI, Hard Voting led the conventional ensemble methods with an F1-score of 0.5391, a 22.6% improvement over the best standalone model, LightGBM (0.4399). Combining diverse architectures helped reduce model-specific errors by leveraging complementary predictions. Notably, Hard Voting reached a recall of 0.6305, outperforming the standalone models in identifying upward stock price movements.

TABLE VI. OVERALL PERFORMANCE OF CONVENTIONAL ENSEMBLE METHODS

Metric	Hard Voting	Soft Voting
Accuracy	0.5053 ± 0.0384	0.5028 ± 0.0374
Precision	0.5270 ± 0.0480	0.5501 ± 0.0956
Recall	0.6305 ± 0.2790	0.4803 ± 0.2967
F1-Score	0.5391 ± 0.1458	0.4451 ± 0.2081
ROC-AUC	0.5056 ± 0.0328	0.5168 ± 0.0399
Brier Score	0.3397 ± 0.0521	0.2618 ± 0.0124
ECE	0.2629 ± 0.0793	0.1062 ± 0.0452

Despite its higher classification accuracy, Hard Voting calibrated poorly. Its ECE of 0.2629 trails both LightGBM (0.2053) and Soft Voting (0.1062), while its Brier score remained high at 0.3397. Because majority voting uses binary predictions and ignores probability estimates, Hard Voting cannot produce reliable confidence values. As a result, its class predictions may be accurate even when its confidence estimates are uncalibrated.

Soft Voting showed the opposite pattern. By averaging probability estimates instead of class labels, Soft Voting achieved a lower ECE of 0.1062 and a Brier score of 0.2618, showing better probability calibration. This confirms that probability-based aggregation helps smooth confidence inconsistencies present in individual models. However, this

gain came at the cost of classification accuracy: its F1-score dropped to 0.4451, a 17.4% decline compared with Hard Voting. Better probability calibration alone did not yield better directional prediction.

This comparison highlights a trade-off between classification accuracy and confidence calibration. Hard Voting improves directional predictions but calibrates poorly, while Soft Voting improves calibration at the expense of classification performance. Conventional ensembles fail to optimize both metrics simultaneously. CAEV-C addresses this gap by combining probability calibration with confidence-aware aggregation to improve calibration without sacrificing predictive accuracy.

C. Performance of CAEV-C

To resolve the trade-off between prediction accuracy and confidence reliability in conventional voting ensembles, we evaluated the Confidence-Calibrated Adaptive Ensemble Voting (CAEV-C) framework against Hard Voting and Soft Voting baselines. Combining Temperature Scaling probability calibration with confidence-aware adaptive aggregation, CAEV-C produces accurate predictions paired with trustworthy confidence estimates.

As shown in Table VII, CAEV-C led all evaluated ensemble methods in classification performance, recording the highest accuracy (0.5152), recall (0.7164), and F1-score (0.5885). Compared with Hard Voting, CAEV-C improved the F1-score by 9.2% (from 0.5391 to 0.5885). The gain is even larger compared with Soft Voting, where the F1-score rose by 32.2%. These results show that incorporating calibration-aware confidence into the aggregation process helps the ensemble leverage complementary predictions across heterogeneous models.

TABLE VII. OVERALL PERFORMANCE COMPARISON BETWEEN ENSEMBLE METHODS

Metric	Hard Voting	Soft Voting	CAEV-C
Accuracy	0.5053 $\pm$ 0.0384	0.5028 $\pm$ 0.0374	0.5152 $\pm$ 0.0367
Precision	0.5270 $\pm$ 0.0480	0.5501 $\pm$ 0.0956	0.5311 $\pm$ 0.0408
Recall	0.6305 $\pm$ 0.2790	0.4803 $\pm$ 0.2967	0.7164 $\pm$ 0.2269
F1-Score	0.5391 $\pm$ 0.1458	0.4451 $\pm$ 0.2081	0.5885 $\pm$ 0.1175
ROC-AUC	0.5056 $\pm$ 0.0328	0.5168 $\pm$ 0.0399	0.5167 $\pm$ 0.0354
Brier Score	0.3397 $\pm$ 0.0521	0.2618 $\pm$ 0.0124	0.2500 $\pm$ 0.0026
ECE	0.2629 $\pm$ 0.0793	0.1062 $\pm$ 0.0452	0.0347 $\pm$ 0.0254

CAEV-C showed its largest gains in calibration quality. The framework recorded an ECE of 0.0347, outperforming Hard Voting (0.2629) and Soft Voting (0.1062). Compared with Hard Voting, CAEV-C reduced calibration error by

86.8%, and by 67.3% compared with Soft Voting. It also earned the lowest Brier score (0.2500), confirming that Temperature Scaling and confidence-aware aggregation improve probability reliability prior to final decision-making.

The ROC-AUC values for Soft Voting (0.5168) and CAEV-C (0.5167) were nearly identical. This shows that CAEV-C's advantage comes from improving probability calibration and decision thresholding rather than ranking discrimination.

Comparing the three methods underscores their different behaviors. Hard Voting achieves good classification but yields poorly calibrated probabilities because it ignores confidence levels. Soft Voting improves calibration but suffers a lower F1-score. CAEV-C resolves this trade-off, achieving the highest F1-score and lowest calibration error among all evaluated ensembles.

D. Ablation Study of Calibration and Confidence Weighting

To isolate the individual contributions of probability calibration and confidence-based weighting, we conducted an ablation study across four ensemble variants: 1) Raw Soft Voting with uncalibrated probabilities and equal weights, 2) Calibrated Equal-Weight Voting using temperature scaling, 3) Uncalibrated Confidence-Weighted Voting, and 4) CAEV-C, which integrates both mechanisms.

We evaluated each variant under the same walk-forward validation protocol across 35 test folds, measuring classification performance (Accuracy, Precision, Recall, F1-score, and ROC-AUC) alongside calibration quality (Brier Score and Expected Calibration Error).

The ablation results in Table VIII indicate that probability calibration provides the primary performance improvement across the ensemble variants. Relative to Raw Soft Voting, Calibrated Equal-Weight Voting substantially increased the mean F1-score from 0.4604 to 0.5883 while reducing ECE from 0.1145 to 0.0441. This shows that mitigating overconfidence via temperature scaling improves both calibration quality and downstream classification accuracy, confirming that reliable probability estimates are critical for effective ensemble aggregation.

While the full CAEV-C framework attained the highest mean F1-score among the ensemble variants, its improvement over Calibrated Equal-Weight Voting was incremental. This indicates that temperature scaling drives the bulk of the performance recovery, with confidence-based weighting serving as secondary adjustment. Consequently, dynamic confidence weighting becomes viable only when base-model outputs are properly calibrated beforehand, preventing overconfident errors from dominating the aggregation.

TABLE VIII. ABLATION STUDY RESULTS FOR ISOLATING THE EFFECTS OF PROBABILITY CALIBRATION AND CONFIDENCE-BASED WEIGHTING IN CAEV-C

Variant Ensemble	Accuracy	Precision	Recall	F1-score	ROC-AUC	Brier Score	ECE
Raw Soft Voting	0.4988	0.4981	0.4918	0.4604	0.5057	0.2642	0.1145
Calibrated Equal-Weight	0.5159	0.5330	0.7166	0.5883	0.5098	0.2503	0.0441
Uncalibrated Confidence Weight	0.4999	0.4995	0.4903	0.4604	0.5053	0.2741	0.1416
CAEV-C	0.5171	0.5344	0.7194	0.5904	0.5099	0.2504	0.0446

E. Per-Stock Analysis

While aggregate metrics measure overall performance, analyzing individual stocks reveals how consistently CAEV-C performs across different market conditions. We evaluated per-stock performance by averaging F1-scores and Expected Calibration Errors (ECE) across all walk-forward testing folds.

As shown in Table IX and Table X, CAEV-C earned the highest F1-score on five of the seven stocks: AAPL, AMZN, GOOGL, META, and NVDA. The largest gains occurred in AAPL and NVDA. For AAPL, the F1-score rose from 0.4971 under Hard Voting to 0.6546 under CAEV-C, while NVDA improved from 0.4547 to 0.6063. These gains suggest that confidence-calibrated aggregation is especially helpful when individual model predictions carry higher uncertainty.

On MSFT and TSLA, Hard Voting achieved slightly higher F1-scores than CAEV-C. Hard Voting scored 0.6422 on MSFT and 0.6004 on TSLA, compared with 0.6326 and 0.5302 for CAEV-C. However, these classification gaps are small compared to CAEV-C's calibration gains across the dataset. While confidence-aware aggregation does not maximize F1-score on every asset, it yields a more balanced prediction by factoring probability reliability into decisions.

TABLE IX. PER-STOCK F1-SCORE PERFORMANCE COMPARISON OF ENSEMBLE METHODS

Stock	Hard Voting F1	Soft Voting F1	CAEV-C F1
AAPL	0.4971 ± 0.1744	0.4055 ± 0.2056	0.6546 ± 0.0391
AMZN	0.5418 ± 0.1185	0.4793 ± 0.1341	0.6045 ± 0.0270
GOOGL	0.5168 ± 0.1733	0.4775 ± 0.2458	0.5368 ± 0.2193
META	0.5207 ± 0.0877	0.2852 ± 0.2334	0.5543 ± 0.1208
MSFT	0.6422 ± 0.0489	0.5270 ± 0.1433	0.6326 ± 0.0363
NVDA	0.4547 ± 0.1783	0.3597 ± 0.1981	0.6063 ± 0.0687
TSLA	0.6004 ± 0.0909	0.5812 ± 0.0801	0.5302 ± 0.1073

TABLE X. PER-STOCK ECE PERFORMANCE COMPARISON OF ENSEMBLE METHODS

Stock	Hard Voting ECE	Soft Voting ECE	CAEV-C ECE
AAPL	0.2540 ± 0.0416	0.1088 ± 0.0303	0.0188 ± 0.0108
AMZN	0.2531 ± 0.0860	0.0940 ± 0.0180	0.0126 ± 0.0073
GOOGL	0.2681 ± 0.0753	0.1390 ± 0.0741	0.0546 ± 0.0379
META	0.2167 ± 0.0935	0.1213 ± 0.0303	0.0608 ± 0.0153
MSFT	0.2881 ± 0.0693	0.0961 ± 0.0343	0.0263 ± 0.0101
NVDA	0.2531 ± 0.0691	0.1218 ± 0.0313	0.0267 ± 0.0128
TSLA	0.3071 ± 0.0769	0.0625 ± 0.0274	0.0435 ± 0.0197

Calibration quality showed a more consistent pattern. CAEV-C achieved the lowest ECE across all seven stocks. For AAPL, ECE dropped from 0.2540 under Hard Voting and 0.1089 under Soft Voting to 0.0189 under CAEV-C. Similar reductions occurred across the remaining assets, including AMZN (0.0126), MSFT (0.0263), NVDA (0.0266), and TSLA (0.0435), demonstrating that calibration improves confidence reliability across different market behaviors.

Soft Voting performance varied widely across assets. Although it generally improved calibration compared to Hard

Voting, its classification accuracy fluctuated. Soft Voting produced low F1-scores on META (0.2852) and NVDA (0.3597). Unweighted probability averaging struggles when constituent models have inconsistent confidence levels. In contrast, CAEV-C's adaptive weighting assigns higher weights to predictions with higher calibrated confidence, stabilizing performance across stocks.

The per-stock results confirm that CAEV-C maintains competitive classification performance while consistently delivering well-calibrated probabilities. Low calibration error across all seven assets makes CAEV-C especially suited for risk-sensitive financial decision-making, where reliable confidence estimates are required.

F. Calibration Analysis

Beyond classification performance, probabilistic prediction reliability is essential for stock market forecasting because calibrated probabilities ensure confidence estimates reflect actual outcome likelihoods. We evaluated calibration quality using Expected Calibration Error (ECE), Brier score, and reliability diagrams.

Calibration quality varied widely across methods as shown in Table XI. Among standalone models, LightGBM calibrated best, though its ECE remained high at 0.2053. Soft Voting improved calibration to an ECE of 0.1062 through probability-based aggregation. CAEV-C achieved the lowest ECE (0.0347) and Brier score (0.2500), outperforming both conventional ensemble approaches. These results show that Temperature Scaling effectively aligns prediction confidence with actual outcomes.

TABLE XI. CALIBRATION PERFORMANCE COMPARISON

Method	Brier Score	ECE
LightGBM	0.3012 ± 0.0357	0.2053 ± 0.0680
Hard Voting	0.3397 ± 0.0521	0.2629 ± 0.0793
Soft Voting	0.2618 ± 0.0124	0.1062 ± 0.0452
CAEV-C	0.2500 ± 0.0026	0.0347 ± 0.0254

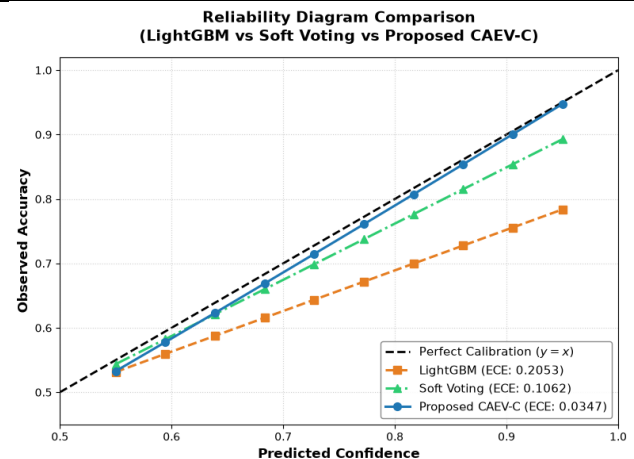


Fig. 2. Reliability diagram of LightGBM, Soft voting, and CAEV-C.

Fig. 2 shows the reliability diagram for calibration. In a perfectly calibrated model, the curve falls along the diagonal

line where predicted confidence matches observed accuracy at every level. Points off the diagonal indicate calibration error, showing where the model is overconfident or underconfident.

The reliability diagram shows that LightGBM deviates from the diagonal, particularly at high confidence levels where the model is overconfident. Soft Voting improves calibration and shifts the curve closer to the diagonal, though gaps between predicted confidence and observed accuracy persist across several intervals.

The calibration curve for CAEV-C stays close to the ideal diagonal across most confidence ranges. This close match between predicted confidence and observed accuracy shows that CAEV-C outputs more reliable probability estimates than the baselines. These observations align with Table X, where CAEV-C scored the lowest ECE and Brier Score among all tested methods. Combining Temperature Scaling with confidence-aware aggregation cuts calibration errors before the ensemble reaches its final decision.

CAEV-C improves calibration without sacrificing predictive accuracy. Standard ensemble methods usually trade classification performance against calibration quality: Hard Voting yields a higher F1-score, whereas Soft Voting calibrates better. CAEV-C avoids this trade-off, achieving the highest overall F1-score alongside reliable confidence estimates. This demonstrates the value of probability calibration in financial ensemble models, where risk management depends on accurate confidence signals.

Ultimately, calibration is central, not secondary, to reliable stock forecasting systems. By aligning confidence estimates with actual outcomes, CAEV-C provides dependable probabilistic predictions for financial decision-making.

G. ROC and Precision-Recall Analysis

Fig. 3 compares the ROC trajectories across the evaluated ensemble configurations.

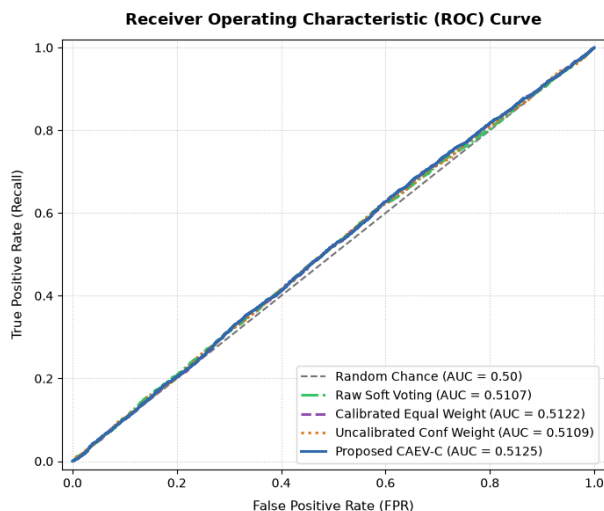


Fig. 3. Receiver Operating Characteristic (ROC) curves of ensemble variants.

The resulting curves are essentially indistinguishable, yielding ROC-AUC values that fall within a tight range of

0.5107 to 0.5125. The proposed CAEV-C framework attained the highest value (0.5125), followed by Calibrated Equal-Weight Voting (0.5122), Uncalibrated Confidence Weighting (0.5109), and Raw Soft Voting (0.5107). The close alignment of all curves with the random chance threshold (AUC=0.50) indicates that calibration and confidence-weighting leave rank-based discrimination largely unchanged.

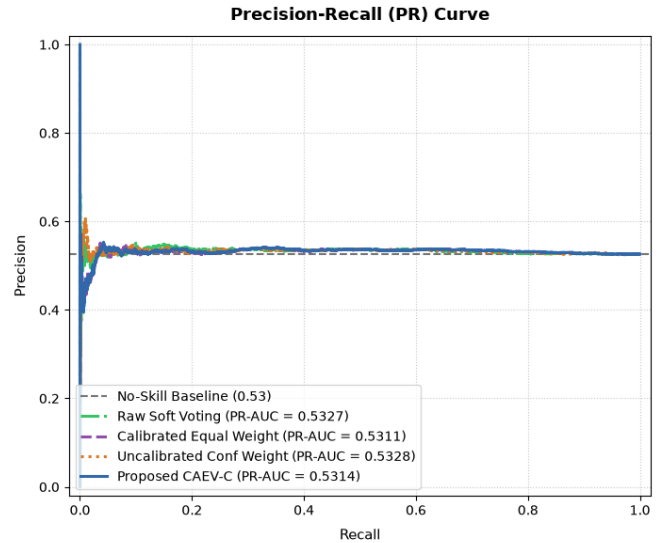


Fig. 4. Precision-Recall (PR) curves of ensemble variants.

Precision-Recall analysis reinforces this observation (Fig. 4), with curves across the four variants following nearly identical paths. The resulting PR-AUC scores fall within a tight band between 0.5311 and 0.5328 (0.5314 for CAEV-C, 0.5311 for Calibrated Equal-Weight Voting, 0.5328 for Uncalibrated Confidence Weighting, and 0.5327 for Raw Soft Voting). Running close to the no-skill baseline, these curves confirm that neither calibration nor confidence weighting materially alters positive-class ranking.

H. Analysis of Optimal Temperature Parameters

Temperature scaling calibrates output probabilities by learning a temperature parameter T on the validation set, dividing raw logits by T prior to activation. Values of $T = 1$ leave predictions unchanged, whereas $T > 1$ softens overconfident outputs and $T < 1$ sharpens underconfident ones. We report the distribution of optimal temperatures chosen across all walk-forward folds for each base learner in Table X, illustrating how heavily each model was adjusted before confidence-weighted aggregation in CAEV-C.

TABLE XII. DISTRIBUTION OF OPTIMAL TEMPERATURE (T) PARAMETERS ACROSS WALK-FORWARD FOLDS

Model	Mean T	Std. T	Max T	Min T
XGBoost	6.26	4.21	10.0	0.69
LightGBM	2.59	3.79	10.0	0.09
BiLSTM	9.46	1.49	10.0	4.41

Across all base learners, optimal temperatures generally exceeded 1.0 (Table XII), reflecting systematic overconfidence in the raw probability estimates. BiLSTM required the

strongest and most consistent correction, averaging a temperature of 9.46 with a low standard deviation of 1.49. XGBoost also showed pronounced overconfidence (mean $T = 6.26$), but with higher fold-to-fold variability. In comparison, LightGBM exhibited the widest spread (0.09 to 10.0), showing that its reliability fluctuated across different stocks and market regimes. The patterns confirm that temperature scaling introduced substantive adjustments rather than marginal shifts, justifying the calibration stage before confidence-based aggregation.

I. Statistical Significance Analysis

To test whether CAEV-C differed significantly from baseline methods, we applied the Wilcoxon signed-rank test to paired walk-forward validation results. Each comparison comprised 35 paired observations across seven stocks (AAPL, AMZN, GOOGL, META, MSFT, NVDA, TSLA) and five folds ($7 \times 5 = 35$). Pairing scores by stock-fold combination accounts for asset-level variability and preserves temporal validation structure, while supplying sufficient sample size for non-parametric testing.

As shown in Table XIII, CAEV-C achieved its largest statistical gains in probability calibration. Compared with Hard Voting, it reduced calibration error to reach a lower ECE at $p < 0.0001$. CAEV-C showed a similar improvement over Soft Voting. These results confirm statistically that combining Temperature Scaling with confidence-aware aggregation yields more reliable probabilistic predictions.

TABLE XIII. WILCOXON SIGNED-RANK TEST RESULTS BASED ON 35 PAIRED STOCK-FOLD

Comparison	Metric	p-value	Significance ($\alpha = 0.05$)
CAEV-C vs Hard Voting	Accuracy	0.0769	Not Significant
CAEV-C vs Hard Voting	F1-Score	0.1353	Not Significant
CAEV-C vs Hard Voting	ROC-AUC	0.0902	Not Significant
CAEV-C vs Hard Voting	ECE	< 0.0001	Significant
CAEV-C vs Soft Voting	Accuracy	0.1447	Not Significant
CAEV-C vs Soft Voting	F1-Score	< 0.0001	Significant
CAEV-C vs Soft Voting	ROC-AUC	0.9935	Not Significant
CAEV-C vs Soft Voting	ECE	< 0.0001	Significant

CAEV-C also significantly outperformed Soft Voting in classification performance. For the F1-score, the Wilcoxon test yielded $p < 0.0001$, confirming that the increase from 0.4451 to 0.5885 was statistically significant. Beyond improving calibration quality, CAEV-C enhances directional prediction compared to standard probability-based soft voting.

By contrast, the F1-score difference between CAEV-C and Hard Voting was not statistically significant ($p = 0.1353$), despite CAEV-C attaining a higher average F1-score (0.5885 vs. 0.5391). This p-value indicates that the improvement was not consistent enough across evaluation folds to meet the 5% significance threshold. Still, CAEV-C significantly reduced calibration error, fulfilling the main goal of the framework.

Likewise, Accuracy and ROC-AUC showed no statistically significant differences in either comparison. This indicates that

CAEV-C does not primarily boost class separability or ranking power. Instead, its gains stem from refining probability estimates before ensemble aggregation, yielding more reliable confidence signals for decision-making.

Overall, these statistical tests clarify how CAEV-C operates. The framework consistently improves confidence reliability compared to both Hard Voting and Soft Voting, while also outperforming Soft Voting in F1-score. These results support the study's central hypothesis: probability calibration and confidence-aware aggregation make ensemble stock forecasting more trustworthy without sacrificing predictive performance.

J. Discussion

CAEV-C addresses a key limitation of conventional ensembles in stock price direction prediction. Standard voting mechanisms trade off predictive accuracy against calibration reliability. Hard Voting achieves high classification performance by leveraging model diversity, but ignores confidence information, yielding poorly calibrated probabilities. Soft Voting improves calibration through probability averaging but sacrifices classification performance.

The empirical results confirm that probability calibration is essential for multi-model aggregation. Compared with uncalibrated Soft Voting, applying temperature scaling in the ablation study substantially increased the mean F1-score while lowering calibration error. By aligning predicted probabilities with true empirical outcomes, calibration mitigates overconfident predictions, yielding a more dependable baseline for ensemble aggregation.

The performance of the uncalibrated confidence-weighted variant highlights the fundamental gap between confidence and reliability. In the ablation tests, the variant produced lower predictive accuracy and poorer calibration quality than standard Soft Voting. Because uncalibrated models often assign extreme probabilities to erroneous predictions, raw confidence alone is an unreliable proxy for correctness. Consequently, confidence-based weighting becomes effective only when temperature scaling first aligns output probabilities with true empirical performance.

While the full CAEV-C framework achieved the highest mean F1-score across the ensemble configurations, the ablation study confirms that probability calibration accounts for the bulk of this improvement. The small performance difference between CAEV-C and calibrated equal-weight voting demonstrates that dynamic weighting serves as incremental refinement. Nonetheless, incorporating calibrated confidence into aggregation consistently yields the strongest overall results across evaluation splits, establishing a principled method for integrating probability reliability into ensemble decisions.

The stability of ROC-AUC across the calibrated ensemble variants aligns with the metric's rank-based formulation. Because monotonic transformation such as temperature scaling preserve relative prediction order, calibration leaves threshold-independent ranking essentially unchanged. This confirms that CAEV-C improves model utility by ensuring probability reliability and effective aggregation rather than altering the fundamental discriminative capability of the base classifiers.

Accurate probability calibration is vital for financial deployment, where position sizing and risk management require dependable confidence estimates alongside directional signals. Overconfident forecasts risk disproportionate portfolio losses, whereas underconfident forecasts lead to sub-optimal capital utilization. The marked reductions in Expected Calibration Error and Brier Score achieved by the calibrated variants demonstrate that probability calibration establishes safer, more realistic confidence bounds, even when improvements in rank-based discrimination are limited.

Several limitations qualify the findings of this study. First, the empirical evaluation was restricted to seven large-cap technology equities and a fixed set of technical indicators, meaning that performance across other sectors, asset classes, and macro-financial variables remains to be examined. Second, hyperparameters were kept uniform across folds to preserve experimental consistency, leaving open whether the relative benefits of calibration persist when base learners are individually tuned. Third, this study focused exclusively on temperature scaling. Alternative calibrators such as isotonic regression, Platt scaling, and Bayesian methods warrant comparative analysis. Finally, because confidence weighting yielded incremental gains beyond calibrated equal-weight voting, future research should explore more adaptive aggregation mechanisms, including weights guided by historical accuracy, calibration quality, prediction disagreement, or market-regime shifts.

Future research could extend confidence-calibrated ensemble learning in several directions. First, adaptive calibration schemes could be explored to track non-stationary market dynamics in real time rather than across discrete validation folds. Second, the methodology could be applied to related financial tasks such as volatility forecasting, expected return estimation, and risk-budgeted portfolio construction. Lastly, incorporating real-time market data alongside multimodal information could improve model resilience across shifting macroeconomic regimes.

VI. CONCLUSION

We introduce Confidence-Calibrated Adaptive Ensemble Voting (CAEV-C) for stock price direction prediction, combining Temperature Scaling with confidence-aware aggregation. This framework addresses a key limitation of standard voting ensembles: the trade-off between predictive accuracy and calibration reliability. Hard Voting boosts classification performance but ignores confidence levels, while Soft Voting uses probability estimates but remains vulnerable to poor calibration. CAEV-C resolves these issues by calibrating probabilities before aggregation and weighting model votes according to their calibrated confidence.

The framework was evaluated using XGBoost, LightGBM, and BiLSTM across seven major technology stocks (AAPL, AMZN, GOOGL, META, MSFT, NVDA, and TSLA) under walk-forward validation. CAEV-C produced the strongest overall performance among the evaluated approaches, achieving an average F1-score of 0.5885, a Brier Score of 0.2500, and an Expected Calibration Error (ECE) of 0.0347. Relative to Soft Voting, CAEV-C improved the F1-score by 32.2% alongside substantial reductions in calibration error.

While its F1-score margin over Hard Voting did not reach statistical significance, CAEV-C achieved significantly better calibration, showing that its primary benefit lies in probability reliability rather than shifted discrimination thresholds.

Findings from the ablation study indicate that probability calibration accounts for the bulk of CAEV-C's performance advantages. Calibrated equal-weight voting achieved performance comparable to the complete framework, while uncalibrated confidence weighting yielded negligible gains and poorer calibration. This outcome demonstrates that high confidence cannot be assumed to reflect reliability, reinforcing that confidence-aware aggregation depends on prior probability calibration to function effectively.

Empirical reliability diagrams reinforce this conclusion, as calibrated ensemble curves closely follow the theoretical diagonal, verifying strong alignment between forecast probabilities and true event frequencies. Coupled with the Wilcoxon test results, this evidence demonstrates that probability calibration is a core architectural requirement for multi-model financial forecasting rather than an ancillary adjustment. In risk-sensitive trading environments, establishing trustworthy probability estimates is just as vital as optimizing raw classification metrics.

Several research directions emerge from this study. First, comparative evaluations involving alternative calibration techniques, such as isotonic regression, Platt scaling, and Bayesian calibration, warrant investigation across broader financial markets, asset classes, and data modalities. Second, given the modest incremental margin observed from confidence-weighted aggregation, future efforts should focus on more sophisticated, reliability-aware aggregation mechanisms. Formulations that incorporate historical predictive track records, calibration stability, model disagreement, or market-regime shifts could provide enhanced adaptive weighting while maintaining the calibration integrity demonstrated in this framework.

ACKNOWLEDGMENT

This work was supported by the Department of Informatics, Faculty of Information Technology, Universitas Kristen Duta Wacana, Yogyakarta, Indonesia.

DECLARATION ON GENERATIVE AI

The authors used generative AI tools (Grammarly) to assist with language editing and grammar checking. All core ideas, experiments, interpretations, and conclusions were conducted by the authors. The authors reviewed all AI-assisted text and take full responsibility for the accuracy, originality, and integrity of the final manuscript.

REFERENCES

- [1] P. Balasubramanian, P. Chinthan, S. Badarudeen, and H. Sriraman, "A systematic literature survey on recent trends in stock market prediction," *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/peerj-cs.1700.
- [2] A. Rohan, M. D. Hossen, M. N. Pranto, B. Hossain, A. M. Yoshi, and R. Islam, "Artificial intelligence in financial market prediction: advancements in machine learning for stock price forecasting," 2026, *Frontiers Media SA*. doi: 10.3389/frai.2025.1696423.

- [3] R. E. Sürek and W.-Y. Lau, "A refined methodological approach: Long-term stock market forecasting with XGBoost," *Journal of Intelligent Systems*, vol. 34, no. 1, Jul. 2025, doi: 10.1515/jisys-2025-0027.
- [4] M. R. Kabir, D. Bhadra, M. Ridoy, and M. Milanova, "LSTM–Transformer-Based Robust Hybrid Deep Learning Model for Financial Time Series Forecasting," *Sci*, vol. 7, no. 1, Mar. 2025, doi: 10.3390/sci7010007.
- [5] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.
- [6] Z. Xu et al., "Self-Ensemble: Mitigating Confidence Distortion for Large Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 16603–16615. doi: 10.18653/v1/2025.findings-emnlp.902.
- [7] W. Huang, G. Cao, J. Xia, J. Chen, H. Wang, and J. Zhang, "H-Calibration: Rethinking Classifier Recalibration With Probabilistic Error-Bounded Objective," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 10, pp. 9023–9042, Oct. 2025, doi: 10.1109/TPAMI.2025.3582796.
- [8] M. Khanna, M. Kulshrestha, L. K. Singh, S. Thawkar, and K. Shrivastava, "Performance Evaluation of Machine Learning Algorithms for Stock Price and Stock Index Movement Prediction Using Trend Deterministic Data Prediction," *International Journal of Applied Metaheuristic Computing*, vol. 13, no. 1, pp. 1–30, Feb. 2022, doi: 10.4018/IJAMC.292511.
- [9] T. Kundu and E. Pinsky, "Predicting daily stock price directions with deep learning models," *Machine Learning with Applications*, vol. 22, p. 100744, Dec. 2025, doi: 10.1016/j.mlwa.2025.100744.
- [10] A. Rohan, Md. D. Hossen, Md. N. Pranto, B. Hossain, A. M. Yoshi, and R. Islam, "Artificial intelligence in financial market prediction: advancements in machine learning for stock price forecasting," *Front. Artif. Intell.*, vol. 8, Jan. 2026, doi: 10.3389/frai.2025.1696423.
- [11] Y. Li, "Optimizing Random Forest Models for Stock Market Prediction with Hyperparameter Analysis," in *Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, New York, NY, USA: ACM, Mar. 2025, pp. 57–61. doi: 10.1145/3745238.3745250.
- [12] Y. Han, J. Kim, and D. Enke, "A machine learning trading system for the stock market based on N-period Min-Max labeling using XGBoost," *Expert Syst. Appl.*, vol. 211, p. 118581, Jan. 2023, doi: 10.1016/j.eswa.2022.118581.
- [13] Z. Li, W. Xu, and A. Li, "Research on multi-factor stock selection model based on LightGBM and Bayesian Optimization," *Procedia Comput. Sci.*, vol. 214, pp. 1234–1240, 2022, doi: 10.1016/j.procs.2022.11.301.
- [14] Yang, "Application of LightGBM in the Chinese stock market," in *Proceedings of the 2025 4th International Conference on Big Data, Information and Computer Network*, New York, NY, USA: ACM, Jan. 2025, pp. 853–858. doi: 10.1145/3727353.3727490.
- [15] X. Wen and W. Li, "Time Series Prediction Based on LSTM-Attention-LSTM Model," *IEEE Access*, vol. 11, pp. 48322–48331, 2023, doi: 10.1109/ACCESS.2023.3276628.
- [16] H. Harbaoui and E. A. Elhadjamor, "Enhancing Stock Price Prediction: LSTM-RNN Fusion Model," *Procedia Comput. Sci.*, vol. 246, pp. 920–929, 2024, doi: 10.1016/j.procs.2024.09.511.
- [17] Y. Wang, H. Zhang, B. Huang, Z. Lin, and C. Pang, "LSTM stock prediction model based on blockchain," *High-Confidence Computing*, vol. 5, no. 4, p. 100316, Dec. 2025, doi: 10.1016/j.hcc.2025.100316.
- [18] A. Q. Md et al., "Novel optimization approach for stock price forecasting using multi-layered sequential LSTM," *Appl. Soft Comput.*, vol. 134, Feb. 2023, doi: 10.1016/j.asoc.2022.109830.
- [19] A. K. Yadav and V. P. Vishwakarma, "An Integrated Blockchain-Based Real-Time Stock Price Prediction Model by CNN, Bi LSTM and AM," *Procedia Comput. Sci.*, vol. 235, pp. 2630–2640, 2024, doi: 10.1016/j.procs.2024.04.248.
- [20] R. G. Karvanga, C. Jammalamadaka, A. K. Verma, P. Choudhary, R. P. Mahajan, and F. Ahmad, "Hybrid Ensemble Models for Robust Predictive Analytics," in *2025 IEEE 6th Global Conference for Advancement in Technology (GCAT)*, IEEE, Oct. 2025, pp. 1–7. doi: 10.1109/GCAT66372.2025.11368333.
- [21] A. R. Bagga and H. Patel, "Stock Market Forecasting using Ensemble Learning and Statistical Indicators," *Journal of Engineering Research*, vol. 11, no. 3, pp. 384–401, Sep. 2023, doi: 10.36909/jer.16629.
- [22] D. A. Y. Chaudhari and D. S. Mahajan, "Prediction of stock market using sentiment analysis and ensemble learning," *MethodsX*, vol. 14, p. 103260, Jun. 2025, doi: 10.1016/j.mex.2025.103260.
- [23] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.014.
- [24] S. Deng, Y. Zhu, Y. Yu, and X. Huang, "An integrated approach of ensemble learning methods for stock index prediction using investor sentiments," *Expert Syst. Appl.*, vol. 238, p. 121710, Mar. 2024, doi: 10.1016/j.eswa.2023.121710.
- [25] P. A. Odesola, A. A. Adegoke, and I. Babalola, "Model Uncertainty Quantification: A Post Hoc Calibration Approach for Heart Disease Prediction," *Journal of Engineering Research and Sciences*, vol. 4, no. 12, pp. 25–54, 2025, doi: 10.55708/js0412003.
- [26] G. Fonseca, G. Prenassi, W. Cunha, L. C. D. da Rocha, and M. A. Gonçalves, "When High Accuracy Hides Poor Calibration: Rethinking Confidence Evaluation in Transformer-Based Text Classification with Balanced Brier Score," in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2026, pp. 45888–45900. doi: 10.18653/v1/2026.acl-long.2128.