A Novel Hybrid Model Based on CEEMDAN and Bayesian Optimized LSTM for Financial Trend Prediction

Yu Sun[1], Sofianita Mutalib[2*], Liwei Tian[3]
School of Management, Guangdong University of Science and Technology, Dongguan, Guangdong Province, China[1]
School of Computing Sciences, College of Computing, Informatics and Mathematics,
Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia[1, 2]
School of Computing, Guangdong University of Science and Technology, Dongguan, Guangdong Province, China[3]

***Abstract*—Financial time series prediction is inherently complex due to its nonlinear, nonstationary, and highly volatile nature. This study introduces a novel CEEMDAN-BO-LSTM model within a decomposition-optimization-prediction-integration framework to address these challenges. The Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) algorithm decomposes the original series into high-frequency, medium-frequency, low-frequency, and trend components, enabling precise time window selection. Bayesian Optimization (BO) algorithm optimizes the parameters of a dual-layer Long Short-Term Memory (LSTM) network, enhancing prediction accuracy. By integrating predictions from each component, the model generates a comprehensive and reliable forecast. Experiments on 10 representative global stock indices reveal that the proposed model outperforms benchmark approaches across RMSE, MAE, MAPE, and R^2 metrics. The CEEMDAN-BO-LSTM model demonstrates robustness and stability, effectively capturing market fluctuations and long-term trends, even under high volatility.**

***Keywords*—LSTM; Bayesian optimization; CEEMDAN; financial time series; time window selection**

I. Introduction

With the ongoing advancement of global economic integration and the increasing openness of international markets, the influence of stock market fluctuations on the global economy has grown significantly. Accurately capturing stock market trends and forecasting their future movements has thus become a crucial research focus in the financial field. Financial market forecasting methods are primarily categorized into linear and nonlinear models. Early research achieved notable progress in financial time series modeling using linear models [1-3]. However, financial time series typically exhibit high noise, uncertainty, and nonstationary, with dynamic changes in the relationships between variables over time. These limitations often hinder traditional linear models from effectively capturing intricate patterns and sustaining long-term dependencies. Furthermore, the linear models' assumption of data smoothness imposes an additional limitation on their validity.

In order to overcome the limitations of linear models, nonlinear prediction methods have become the mainstream of financial market prediction. Among them, artificial neural networks (ANNs) are widely used due to their ability to process high-dimensional, multimodal and heterogeneous data. For example, Yassin et al. (2017) utilized a stock prediction model based on multilayer perceptron (MLP) to predict the weekly stock price of Apple Inc. and verified its strong prediction performance through one step ahead (OSA) and correlation analysis [4]. Similarly, Zhang et al. (2021) employed a back propagation (BP) neural network to classify and predict stock price patterns, achieving an improved accuracy of 73.29% and providing valuable insights for investors and macroeconomic policies [5]. However, ANN-based models face many challenges, including overfitting, gradient vanishing or exploding, and easy to fall into local optimality, which limit their generalization ability for complex financial data.

With the continuous development and application of deep learning technology, Long Short-Term Memory (LSTM) network has become an important model due to its strong ability to handle noisy, nonlinear and nonstationary data. By introducing advanced gating mechanisms, LSTM effectively overcomes the inherent gradient vanishing and gradient exploding problems of traditional recurrent neural networks (RNNs), making it particularly suitable for processing long sequence data [6, 7]. Its robustness and adaptability have promoted its widespread application in the financial field. Wang et al. (2024) verified that the LSTM model can overcome the limitations of RNN in stock market forecasting and can greatly improve the forecasting performance [8]. These advantages make LSTM a highly promising component in hybrid models for solving complex financial forecasting tasks.

In the field of financial market forecasting, hybrid models have attracted much attention. This type of model combines the strengths of multiple algorithms and offers greater advantages than single models in processing complex and noisy financial data. Compared with single models, hybrid models have stronger generalization capabilities and are more robust, especially in capturing multi-scale features and nonlinear relationships.

In recent years, LSTM-based hybrid models have shown great potential in stock market forecasting due to their ability to capture both long and short-term patterns in financial time series. Table I summarizes recent research based on LSTM

*Corresponding Author.

models, clearly showing the methods and advantages of these hybrid models in stock market forecasting.

TABLE I. SUMMARY OF THE RECENT WORK ON LSTM-BASED STOCK MARKET PREDICTION

Reference	Technique	Dataset	Number of stocks	Metric	Advantages
Mehtarizadeh et al. (2025) [9]	LSTM, Sin-Cosine Algorithm, ARIMA, GARCH	Stock	8	RMSE	Integrates statistical and deep learning models, enhancing the ability to capture the complex dynamics of stock prices.
Baek (2024) [10]	CNN, LSTM, Genetic Algorithm Optimization	Stock index	1	MAPE, MSE, MAE	Effectively improves the accuracy of stock index prediction.
Sang and Li (2024) [11]	Attention Mechanism Variant LSTM	Stock	1	MSE, MAE, R^2	Enhances generalization, prediction accuracy, and convergence ability
Zheng et al. (2024) [12]	Convolutional Neural Network, Bidirectional LSTM, Attention	Stock	1	MSE, MAPE	Demonstrates significantly improved prediction accuracy.
Akşehir et al. (2024) [13]	Two-level decomposition of CEEMDAN, LSTM, Support Vector Regression	Stock index	4	RMSE, MAE, MAPE, R^2	Removes noise from financial time series data, boosting predictive performance.
Gülmez (2023) [14]	Artificial Rabbits Optimization, LSTM	Stock	30	MSE, MAE, MAPE, R^2	Exhibits high generalisability in diverse financial scenarios.
Tian et al. (2022) [15]	LSTM, Bayesian optimization, LightGBM	Diverse financial market datasets	9	RMSE, MAE, F1-score, Accuracy	Provides better approximation and generalisation in stock volatility prediction.

As shown in Table I, most studies on LSTM-based stock market prediction methods employ parameter optimization algorithms. This is mainly because manually adjusting parameters not only increases the computational cost, but may also reduce the prediction accuracy and optimization efficiency. In deep learning networks such as LSTM, hyperparameter selection is crucial to achieve accurate predictions [16]. The performance of LSTM largely depends on the optimal configuration of hyperparameters, such as the number of hidden neurons and the learning rate setting. In addition, since the neural network needs to optimize a large number of parameters, it is prone to falling into local optima during this process and may be at risk of overfitting.

To address these challenges, Bayesian Optimization (BO) provides an efficient solution [17]. BO estimates the objective function through Gaussian Processes (GP) and optimizes it using surrogate modeling. This method can efficiently explore the hyperparameter space, avoid redundant calculations, reduce the risk of falling into local optima, and significantly improve both the effectiveness and training efficiency of the LSTM model, making it more suitable for the complexity of financial time series prediction.

In addition to hyperparameter optimization, data preprocessing also plays a vital role in improving prediction accuracy. High noise is a key characteristic of financial time series and a major challenge for accurate prediction. Signal decomposition methods can transform complex original time series into simpler components, enabling the model to capture potential features more effectively and reduce noise. Among them, variational mode decomposition (VMD) is widely used due to its strong practical performance [18, 19]. However, VMD requires manual selection of the number of modes, introducing subjectivity and uncertainty. In contrast, Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) provides an advanced adaptive data decomposition method that decomposes data into physically meaningful modal components. As a higher-order version of Empirical Mode Decomposition (EMD), CEEMDAN overcomes some of EMD's limitations by introducing adaptive white noise, improving robustness and accuracy in extracting intrinsic features of complex data. CEEMDAN has demonstrated significant theoretical and practical value in economic and financial applications, laying a solid foundation for enhancing model performance [20].

Based on the above analysis, this study introduces a novel hybrid model CEEMDAN-BO-LSTM. This model further improves the performance of financial time series prediction by effectively removing the original signal noise, enhancing the long-term and short-term dependency modeling capabilities, and employing an efficient optimization algorithm for hyperparameter tuning. Specific contributions of this study include:

1) Proposal of an innovative hybrid framework: This study constructs a "decomposition-optimization-prediction-integration" framework. The framework uses the CEEMDAN algorithm to perform multi-scale decomposition of stock index time series, automatically adjusts the hyperparameters of LSTM through BO for prediction, and finally integrates the prediction results. This method greatly enhances predictive performance and adaptability, and effectively overcomes the limitations of traditional models in parameter selection and noise reduction.

2) Generalization ability in stock index prediction: The proposed model is experimentally evaluated on 10 representative stock indices worldwide, leading to consistent conclusions. The results demonstrate its effectiveness in capturing the nonlinear characteristics of stock market volatility and achieving significant improvements in forecast accuracy and stability, further highlighting its applicability in financial time series forecasting.

3) Optimized time window segmentation: This study classifies the decomposed components based on energy contribution, dividing them into different frequency components, and assigns a suitable time window to each component, which can be predicted according to the characteristics of different frequencies, further enhancing the prediction performance.

These contributions enable the CEEMDAN-BO-LSTM model to demonstrate significant potential in addressing complex financial market fluctuations, opening up broader application prospects for financial market analysis.

This paper is structured into five sections. The next section introduces the methods used in this study. Section III describes the construction process of the CEEMDAN-BO-LSTM model. Then, Section IV outlines the experimental procedure and analyzes the results. Finally, Section V presents the conclusions of this study, as well as future research directions.

II. METHODS

A. LSTM

The LSTM network is an enhanced variant of the RNN [5]. By incorporating specialized LSTM units, the LSTM network can effectively store and manage both long-term and short-term information. As illustrated in Fig. 1, each LSTM unit comprises three gate mechanisms and two state variables. They are input gate (i_t), forget gate (f_t), and output gate (o_t), as well as cell state (C_t) and hidden state (h_t).

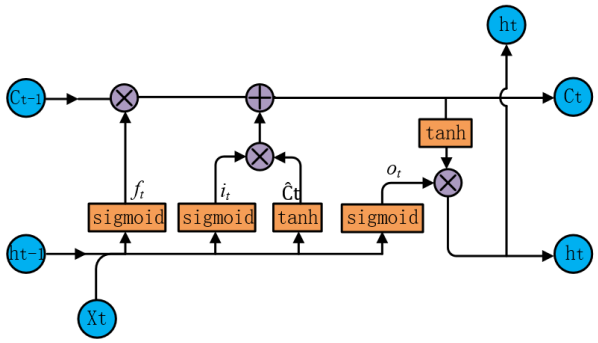

Fig. 1. Structure of LSTM.

According to Fig. 1, these components work collaboratively to regulate the flow of information, allowing LSTM network to maintain and update memory across extended sequences.

The input gate determines the processing of information from the current input data. The mathematical representation is shown in Eq. (1) and Eq. (2).

$$i_t = \sigma(W_i * [h_{t-1}, X_t] + b_i) \quad (1)$$

$$\hat{C}_t = tanh(W_c * [h_{t-1}, X_t] + b_c) \quad (2)$$

The forget gate controls the update of historical data to the memory unit state value. Its mathematical expression is presented in Eq. (3).

$$f_t = \sigma\big(W_f * [h_{t-1}, X_t] + b_f\big) \quad (3)$$

C_t is updated through the combined actions of f_t and i_t. The forget gate regulates the retention of the previous cell state (C_{t-1}), while the input gate introduces new information via the candidate state ($\hat{C}_t$). The update equation is shown in Eq. (4).

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \quad (4)$$

The output gate determines the information to be emitted at the current time step. The detailed calculation procedure is shown in Eq. (5) and in Eq. (6).

$$o_t = \sigma(W_o * [h_{t-1}, X_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

Here, σ (sigmoid function) and tanh (hyperbolic tangent function) serve as activation functions. W and b denote the weight matrices and bias terms, respectively. Through this design, LSTM effectively integrates long-term and short-term information, ensuring stability and accuracy in forecasts. Its flexibility and adaptability make LSTM particularly useful for financial time series forecasting, as it effectively captures the complex nonlinear dynamics of stock market volatility.

B. CEEMDAN

In the field of signal processing, researchers have proposed a variety of advanced methods to analyze the intrinsic characteristics of nonlinear and nonstationary data, including empirical mode decomposition (EMD) and its related techniques. EMD is able to decompose the different frequency components of the original data, revealing the contribution of each frequency component, and providing high-resolution analytical capabilities for nonlinear time series [21].

Although EMD has many advantages, its mode mixing issue limits its accuracy and effectiveness in complex signal decomposition. To address this limitation, researchers proposed Ensemble Empirical Mode Decomposition (EEMD) in 2009. EEMD reduces the impact of mode mixing by adding white noise of different amplitudes to the original signal, performing EMD multiple times, and then averaging the decomposition results. However, since EEMD requires multiple repetitions of the decomposition process, the computational cost is high and may introduce additional computational errors [22].

Based on EEMD, Torres et al. (2011) proposed CEEMDAN [23]. CEEMDAN algorithm further enhances the EEMD by refining the introduction of white noise at each decomposition step. This improvement ensures the consistency of the extracted IMFs, mitigates mode mixing, and reduces reconstruction errors. This method enhances both decomposition accuracy and computational efficiency, making it more reliable for complex and dynamic financial time series analysis.

The following section outlines the detailed algorithmic steps of CEEMDAN:

1) Step 1: Generate noisy data sets. Noisy versions of the original time series $x[n]$ are created using the Eq. (7):

$$x_i[n] = x[n] + \varepsilon_0 w_i[n] \quad (7)$$

Where $w_i[n]$ (i = 1, 2 … I) represent Gaussian white noise that follows a normal distribution, and ε_0 is the standard deviation of the noise. This step introduces controlled noise to reduce mode mixing during the decomposition process.

2) Step 2: Apply the EMD method to each noisy signal to extract the first intrinsic mode function (IMF). The first IMF is

calculated as the average of the IMFs obtained from all noisy realizations as shown in Eq. (8)

$$IMF_1[n] = \frac{1}{I}\sum_{i=1}^{I} IMF_1^i[n] \quad (8)$$

The first residual is then computed as Eq. (9):

$$r_1[n] = x[n] - IMF_1[n] \quad (9)$$

3) Step 3: Iterative decomposition is applied to obtain successive IMFs. For k = 2, ... K, the k-th residue is calculated as shown in Eq. (10). The realizations $r_k[n] + \varepsilon_k EMD_k(w_i[n])$ are decomposed until the first EMD mode is obtained, and the (k+1)-th mode is defined as shown in Eq. (11):

$$r_k[n] = r_{k-1}[n] - IMF_k[n] \quad (10)$$

$$IMF_{k+1}[n] = \frac{1}{I}\sum_{i=1}^{I} EMD_1(r_k[n] + \varepsilon_k EMD_k(w_i[n])) \quad (11)$$

4) Step 4: Proceed to step 3 for the next k. The CEEMDAN algorithm terminates when the residual no longer exceeds the two extreme points, and further decomposition is not possible. The final residual is given by Eq. (12):

$$R[n] = x[n] - \sum_{k=1}^{K} IMF_k \quad (12)$$

This approach effectively resolves mode mixing, ensuring stable and interpretable decomposition results.

C. BO

Bayesian optimization (BO) is a widely used algorithm for optimizing black-box functions with the aim of finding the global optimal solution efficiently. The method achieves optimization by iteratively updating the posterior distribution of the objective function, while the updating process is based on the prior distribution with historical data. Compared to traditional optimization methods, Bayesian optimization excels in handling non-convex problems and avoids falling into local optima [24]. Its objective is to maximize the function f(x) by finding the optimal input x^* within the search space $\chi \in \mathbb{R}$, as shown in Eq. (13):

$$x^* = \underset{x \in \chi}{\operatorname{argmax}} f(x) \quad (13)$$

The essence of Bayesian optimization is rooted in modeling the objective function $f(x)$ as a stochastic process, typically represented by a Gaussian Process (GP) as a surrogate model. The GP provides a probabilistic estimate of $f(x)$ based on historical observations, enabling Bayesian optimization to balance exploration of unexplored regions with exploitation of known high-performing regions. This balance improves optimization efficiency and reduces the risk of missing the global optimum.

Bayesian optimization relies on Bayes' theorem to infer the posterior distribution of f(x) given a historical dataset $D_n = \{(x_i, f(x_i))\}_{i=1}^n$, The posterior distribution is given by Eq. (14):

$$P(f(\mathrm{x})|D_n) = \frac{P\left(D_n \middle| f(\mathrm{x})\right) P(f(\mathrm{x}))}{P(D_n)} \quad (14)$$

Where $P\left(D_n \middle| f(\mathrm{x})\right)$ is the likelihood of the dataset given (x), $P\left(f(\mathrm{x})\right)$ represents the prior distribution of the objective function, and $P(D_n)$ is the evidence.

By employing a Gaussian Process, BO algorithm efficiently estimates $f(\mathrm{x})$ and selects the next sampling point from the posterior distribution $P(f(\mathrm{x})|D_n)$. This iterative process seeks to maximize the objective function while minimizing uncertainty, enabling efficient exploration in high-dimensional spaces.

Bayesian optimization's flexibility and ability to incorporate prior knowledge make it particularly suitable for complex and computationally expensive optimization tasks, such as hyperparameter tuning in machine learning, experimental design, and automated decision-making. Its ability to efficiently handle high-dimensional, noisy, and black-box functions positions it as an effective method for addressing complex optimization challenges.

III. Model Construction

Due to the nonlinear and nonstationary nature of stock market indices, achieving accurate predictions remains a significant challenge. In this context, the combination of effective decomposition techniques and advanced machine learning methods is particularly crucial for handling these complexities. In light of this, this study integrates CEEMDAN temporal decomposition, Bayesian Optimization (BO), and LSTM network to propose a novel stock market forecasting framework: CEEMDAN-BO-LSTM.

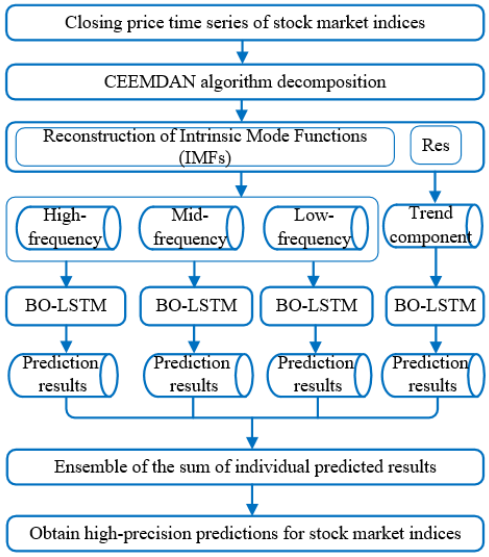

Fig. 2. The workflow of CEEMDAN-BO-LSTM predictive modeling.

As illustrated in Fig. 2, the CEEMDAN-BO-LSTM modeling process addresses the complex, nonlinear, and multi-scale nature of stock market indices through four key steps. In this process, the closing price series of stock market indices serves as input data. The details of these steps are as follows:

1) Step 1: Decomposition. Decompose the closing price time series into multiple IMFs and a residual component using the CEEMDAN algorithm. IMFs capture time series

characteristics at different frequency levels, while the residual represents the overall trend.

2) Step 2: Optimization. The decomposed components are classified according to frequency bands, the time window in the LSTM model is set according to these categories, and the LSTM is then trained using Bayesian optimization.

3) Step 3: Prediction. Optimized LSTM models corresponding to different frequency bands generate predictions for each respective band.

4) Step 4: Integration. Predictions from all components are combined to produce the final forecast for stock market indices. Appropriate evaluation metrics are used to assess the proposed model's reliability and robustness.

IV. Experiment and Result

A. Data Collection and Preprocessing

The historical financial time series dataset used in this study is obtained from Yahoo Finance and covers the daily closing prices of 10 representative global stock indices. The dataset spans 10 years, from October 23, 2014, to October 21, 2024.

Table II presents detailed information on the selected stock indices and their market characteristics. As depicted in Fig. 3, the dynamic trajectories of these indices exhibit typical characteristics of financial time series, including nonlinearity and nonstationary. These characteristics highlight the complexity and challenges associated with forecasting financial trends.

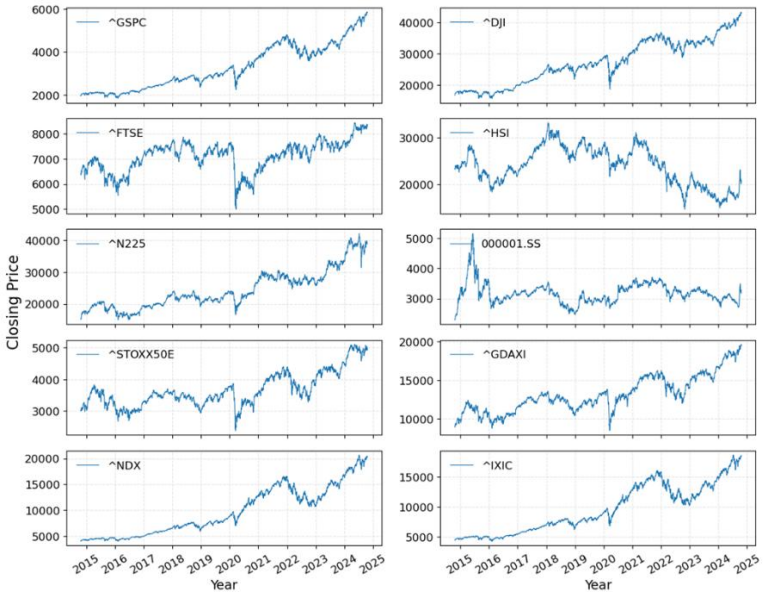

Fig. 3. Historical trends of 10 stock indices.

TABLE II. Major Stock Market Indices and their Characteristics

Index Name	Code	Market Characteristics	Index Name	Code	Market Characteristics
S&P 500 Index	^GSPC	Represents 500 leading U.S. companies across multiple sectors, reflecting the overall U.S. economy.	Shanghai Composite Index	000001.SS	Covers all A-shares on the Shanghai Stock Exchange, serving as a benchmark for China's market.
Dow Jones Industrial Average	^DJI	Consists of blue-chip stocks, reflecting trends in the industrial and financial markets.	Euro Stoxx 50 Index	^STOXX50E	Tracks 50 leading European companies and represents key sectors of the eurozone economy.
FTSE 100 Index	^FTSE	Includes 100 major UK companies, providing insights into the UK and European financial markets.	German DAX Index	^GDAXI	Comprises 40 major companies listed on the Frankfurt Stock Exchange, serving as a key indicator of Germany's economic and corporate performance.
Hang Seng Index	^HSI	Tracks 50 leading Hong Kong companies, highlighting the economic connection between China and Hong Kong.	NASDAQ 100 Index	^NDX	Comprises 100 major U.S. tech and non-financial companies, serving as a key indicator of global tech stocks.
Nikkei 225 Index	^N225	Covers 225 top Japanese companies, reflecting Japan's economic trends and Asia's broader market dynamics.	NASDAQ Composite Index	^IXIC	Includes all companies listed on NASDAQ, capturing global tech stock trends and investor sentiment.

B. Stock Closing Price Sequence Decomposition Based on CEEMDAN

In this study, the S&P 500 index (^GSPC) was selected as a representative example for detailed analysis, while the procedures for the other nine indices were omitted due to space constraints. The ^GSPC dataset consists of 2,515 samples, with the first 80% used for training and the remaining 20% for testing to preserve the temporal structure during model evaluation.

The training set was processed using the CEEMDAN method to extract intrinsic mode functions (IMFs). Following the parameter settings proposed by Torres et al. (2011), the noise standard deviation was set to 0.2 (noise_std=0.2), the number of trials was 500, and a maximum of 2,000 sifting iterations (max_iter=2000) was allowed for each intrinsic mode function (IMF) extraction [23].

Fig. 4 presents the results of the CEEMDAN decomposition, showing six extracted intrinsic mode functions (IMFs) and one residual component (RES). This decomposition effectively separates the original time series into components with distinct frequency characteristics, capturing variations across different scales while preserving the underlying trend. Compared to the original signal, these components are smoother and more regular, enhancing feature representation and providing a robust foundation for subsequent modeling and forecasting.

Table III summarizes the key statistical characteristics of each IMF and the residual component, including their energy contribution rates and Pearson correlation coefficients with the original time series. These metrics reveal the impact of different frequency components on the time series and support their classification into three distinct frequency groups for subsequent modeling and forecasting.

TABLE III. STATISTICAL CHARACTERISTICS OF CEEMDAN-DECOMPOSED ^GSPC COMPONENTS

Component	Sample Size	Mean	Energy Contribution Rate	Correlation Coefficient with the Original Sequence	Average Difference (95% Confidence Interval)
IMF1	2515	0.2533	0.32%	0.0282	[-0.4497, 0.9290]
IMF2	2515	0.2452	0.39%	0.0150	[-0.4715, 1.7823]
IMF3	2515	0.7810	0.71%	0.0733	[-0.2590, 3.5593]
IMF4	2515	1.0539	2.89%	0.0727	[-0.3335, 3.0623]
IMF5	2515	-5.5893	6.99%	0.0435	[-13.4650, 9.6706]
IMF6	2515	-4.9123	88.70%	0.1788	[-13.4649, 9.6706]
RES	2515	3288.74	-	0.9542	[3243.0319, 3322.9385]

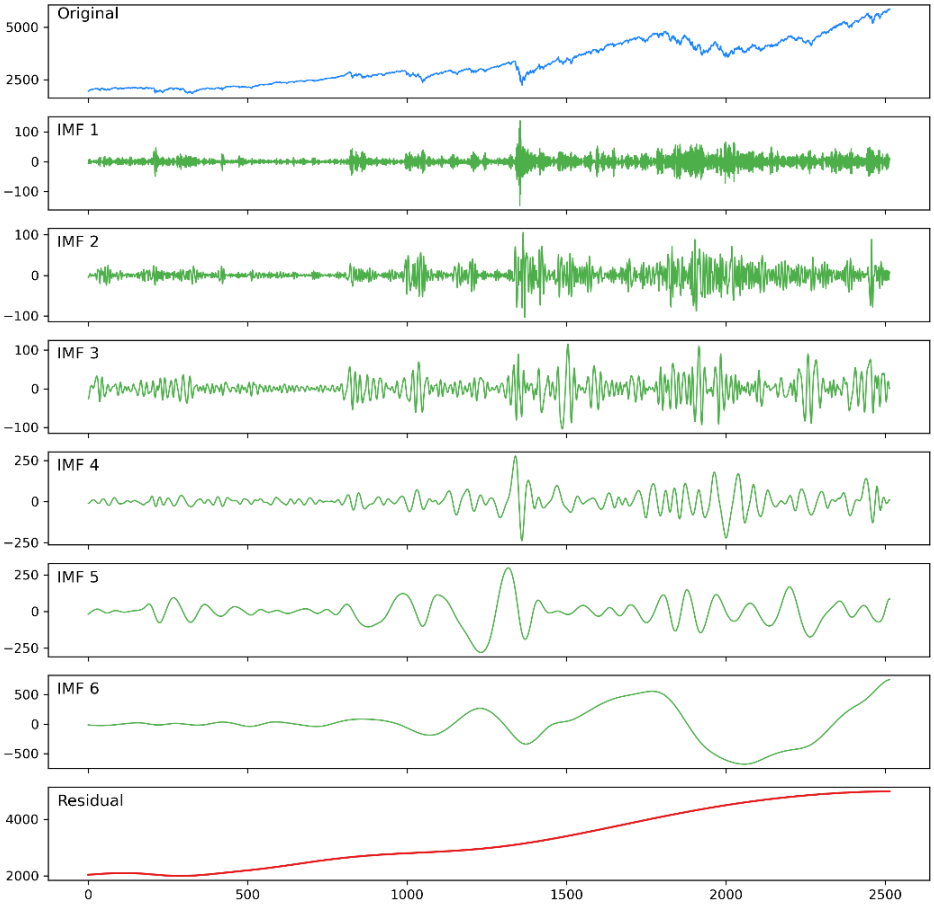

Fig. 4. CEEMDAN decomposition of closing prices for the ^GSPC index.

Based on the statistical analysis in Table III, the IMFs were categorized according to their energy contribution rates and correlation coefficients.

IMF 1 to IMF 3 primarily represent high-frequency components, characterized by low energy contribution rates (0.32%, 0.39%, and 0.71%, respectively) and low Pearson correlation coefficients with the original sequence (0.0282, 0.0150, and 0.0733, respectively). These components primarily capture short-term stochastic fluctuations, which may be influenced by sudden market events, speculative trading, or short-lived policy shifts. IMF 3 exhibits slightly higher explanatory power for short-term volatility.

IMF 4 and IMF 5 correspond to medium-frequency components, with energy contribution rates of 2.89% and 6.99% and correlation coefficients of 0.0727 and 0.0435, respectively. These components reflect medium-term market adjustments and cyclical dynamics, with IMF 5, in particular, being associated with medium-to-long-term fluctuations influenced by economic cycles and technical market corrections.

IMF 6, the dominant low-frequency component, accounts for the highest energy contribution (88.70%) and exhibits a correlation coefficient of 0.1788 with the original sequence. It primarily represents long-term market trends driven by macroeconomic factors and sustained investor sentiment.

The residual component (RES) represents the long-term trend, exhibiting a mean value of 3288.74 and a high correlation coefficient of 0.9542 with the original sequence. This indicates that the RES fully encapsulates the overall market trend.

Applying energy contribution thresholds of 2% for high-frequency components and 60% for low-frequency components effectively distinguishes short-term fluctuations from long-term trends. This separation facilitates predictive modeling by isolating different cyclical patterns in the data. The frequency band allocation of the ^GSPC index is illustrated in Fig. 5.

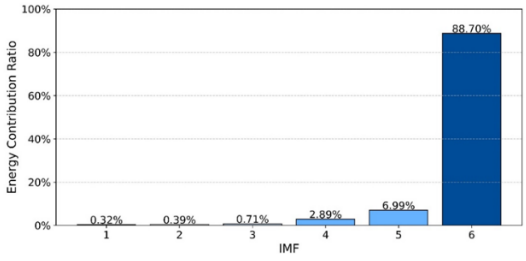

Fig. 5. Frequency band allocation results of IMF components for ^GSPC.

In order to enable the proposed model to effectively capture the unique fluctuation patterns and dynamic characteristics of each frequency band, a specific time window is assigned to the three different frequency components in this study. The time windows for the high-frequency, medium-frequency, and low-frequency components are set to 5 days, 10 days, and 20 days, respectively. The residual component, which primarily represents the long-term trend, is assigned a time window of 30 days. This configuration allows the model to fully learn and represent the overall long-term market movements and trend characteristics.

By adjusting the time windows based on the frequency characteristics of the components, the model is better equipped to analyze and predict market behavior across different time horizons.

C. Analysis of Experimental Results

Since the process of constructing the prediction model for the S&P 500 index (^GSPC) is identical to that of other stock indices, it is not discussed in detail here due to space limitations. To compare the effectiveness of the CEEMDAN decomposition method, the original closing price sequence was also decomposed using the EMD method for comparison. Taking the ^GSPC index as an example, the EMD method decomposed the sequence into 8 IMFs and RES, resulting in a total of 9 components. The specific decomposition results are presented in Fig. 6. The original ^GSPC closing price sequence is displayed at the top, followed by the IMF components arranged from high to low frequency, with the trend component at the bottom. The construction of the prediction model based on the EMD decomposition follows the same methodology as the CEEMDAN-BO-LSTM approach proposed in this study.

In order to further verify the effectiveness of the decomposition technique, the Bayesian optimized LSTM (BO-LSTM) model and the traditional LSTM model are involved in the comparative analysis in this study. The LSTM models constructed in this study all adopt a double-layer structure, and the specific parameter configurations are shown in Table IV.

TABLE IV. OPTIMIZED HYPERPARAMETERS AND THEIR SEARCH RANGES

Hyperparameters	Search Range
lstm_units1	[50, 200]
lstm_units2	[50, 150]
learning_rate	[0.0001, 0.01]
batch_size	[16, 64]

After completing the above preparations, the ^GSPC index was first trained and then predicted. To comprehensively evaluate the prediction performance, root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) are utilized. The optimal parameter configuration for the double-layer LSTM model is determined through Bayesian optimization. The final settings include a batch size of 31, a learning rate of 0.005, 115 units in the first LSTM layer, and 79 units in the second LSTM layer. Table V summarizes the predictive performance comparison for the ^GSPC index across different models.

According to the prediction results shown in Table V, it can be observed that the CEEMDAN-BO-LSTM model outperforms all other models. The RMSE value of this model is 32.6788, the MAE value is 27.9756, the MAPE value is 0.5910%, and the R^2 value is 0.9969, which strongly demonstrates its superior predictive accuracy and robust fitting capability. In contrast, the RMSE values of the CEEMDAN-LSTM and EMD-BO-LSTM models are 41.4031 and 40.9361, respectively, performing better than the BO-LSTM and traditional LSTM models but still falling short compared to the CEEMDAN-BO-LSTM model.

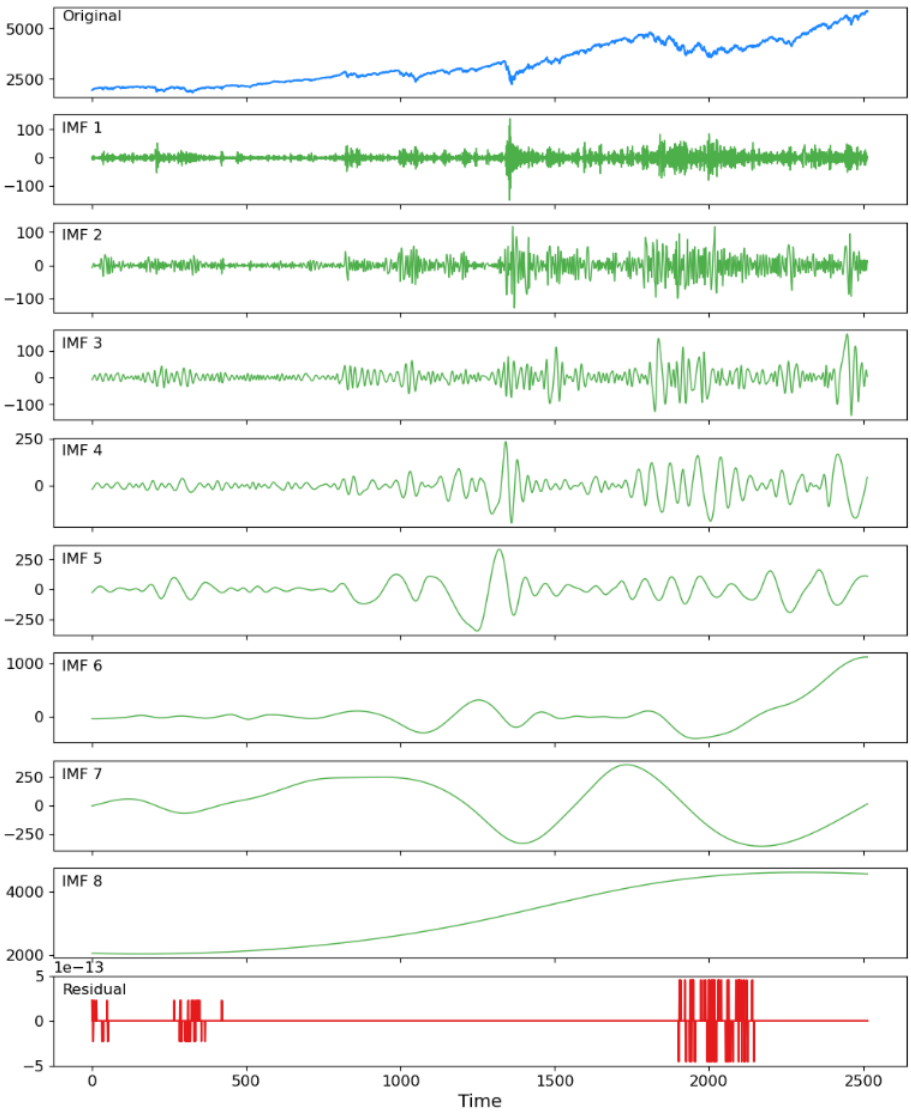

Fig. 6. EMD Decomposition of Closing Prices for the ^GSPC Index.

TABLE V. PERFORMANCE COMPARISON OF DIFFERENT LSTM-BASED MODELS FOR ^GSPC PREDICTION

Metric \ Model	CEEMDAN-BO-LSTM	CEEMDAN-LSTM	EMD-BO-LSTM	EMD-LSTM	BO-LSTM	LSTM
RMSE	32.6788	41.4031	40.9361	156.1560	56.6349	75.2508
MAE	27.9756	32.6506	29.1307	102.8529	45.1560	60.4651
MAPE	0.5910%	0.6729%	0.5991%	1.9808%	0.9487%	1.2603%
R^2	0.9969	0.9951	0.9951	0.9290	0.9909	0.9840

Compared to Bayesian-optimized models, non-optimized LSTM-based models exhibit weaker predictive performance and limited capability in capturing market trends. For example, the EMD-LSTM model produces large prediction errors, but its predictive performance improves significantly after Bayesian optimization of core parameters.

Next, to further visualize and compare the predictive performance of different models, Fig. 7 illustrates the final prediction results, Fig. 8 presents the error distribution, and Fig. 9 depicts the model fitting performance.

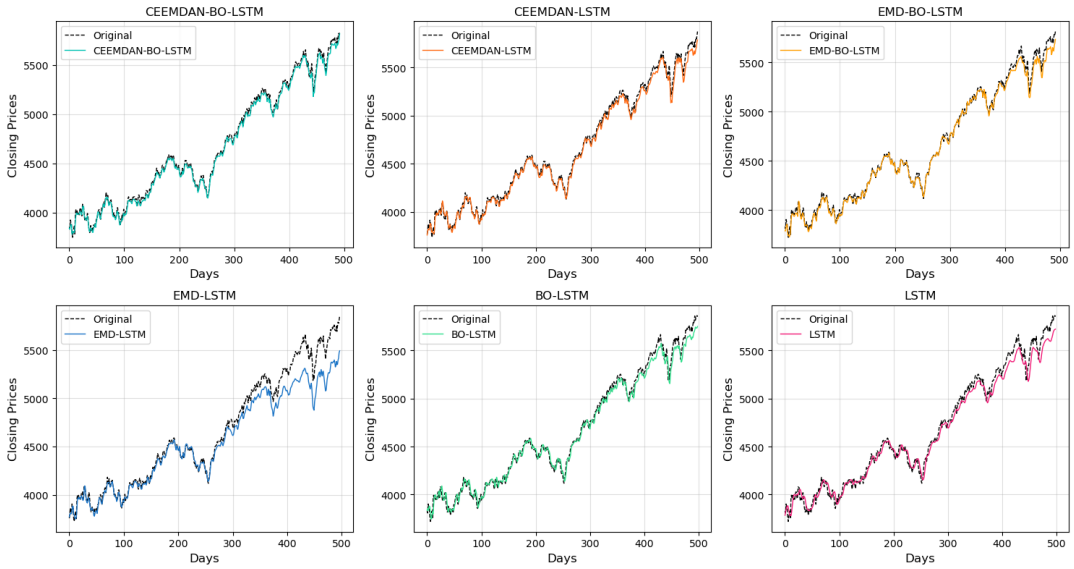

Fig. 7. Individual forecasts of different models on the ^GSPC index.

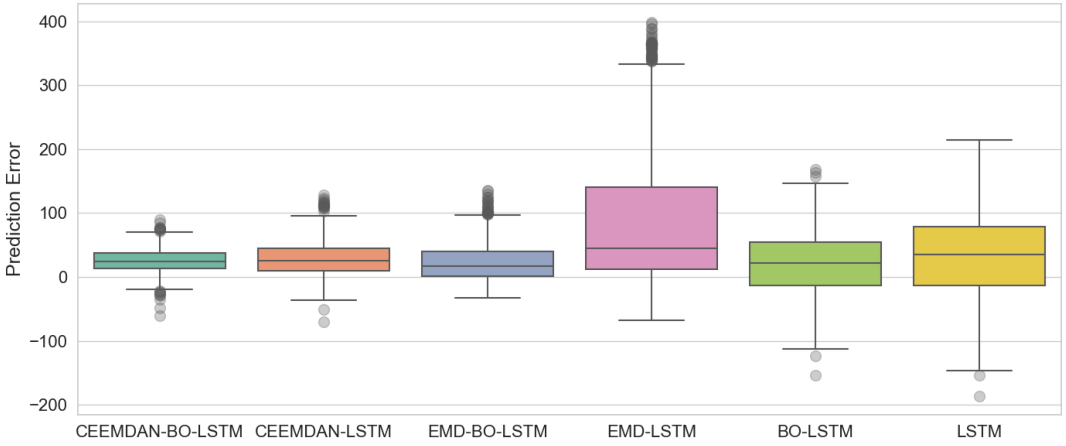

Fig. 8. Comparison of prediction errors for the ^GSPC index across different models.

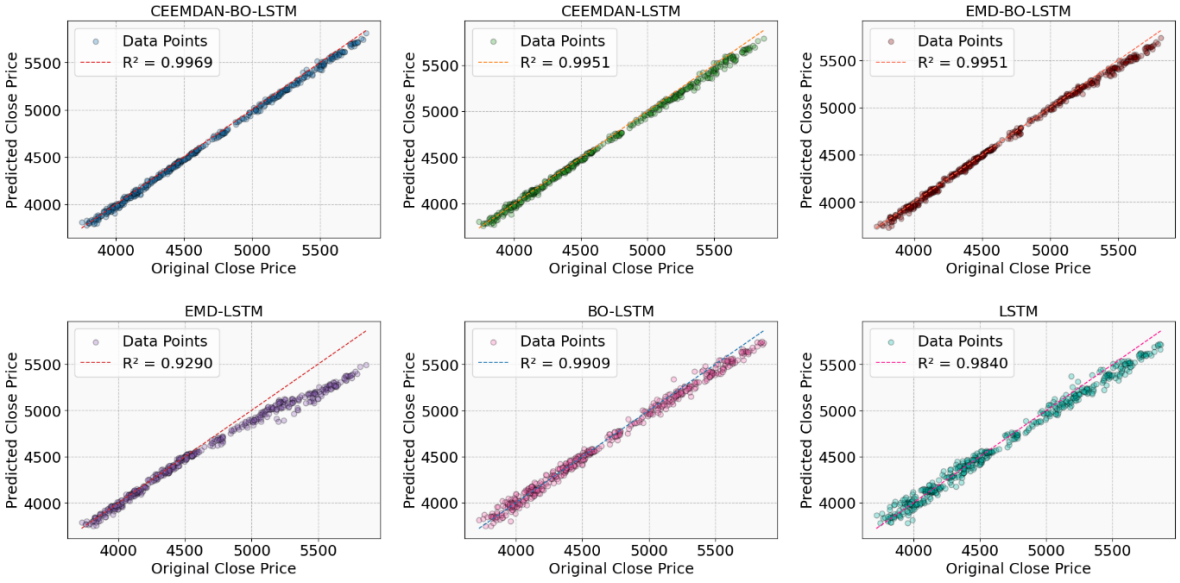

Fig. 9. Fitting performance of different models for the ^GSPC index predictions.

As illustrated in Fig. 7, it is evident that the CEEMDAN-BO-LSTM model exhibits the best fitting performance among all models, with its prediction curve closely aligning with the actual values. The model is effective in capturing the dynamic trends in different frequency bands, particularly around trend shifts and extreme points, demonstrating strong adaptability to complex market dynamics. In contrast, while the CEEMDAN-LSTM and EMD-BO-LSTM models also fit the overall trend well, they exhibit slight phase shifts during periods of severe market volatility, especially around local peaks and troughs. On the other hand, the fitting performance of the EMD-LSTM, BO-LSTM, and traditional LSTM models is relatively weak. Among them, both the EMD-LSTM model and the traditional LSTM model exhibit poor fitting performance in the second half of the prediction period. However, their predictive accuracy improves significantly after Bayesian optimization.

As shown in the box plots of error distributions in Fig. 8, the CEEMDAN-BO-LSTM model exhibits a narrower interquartile range (IQR) and shorter whiskers, indicating lower prediction uncertainty and greater stability. In contrast, the CEEMDAN-LSTM and EMD-BO-LSTM models have slightly wider error distributions, though their overall volatility remains moderate. On the other hand, the error box plots of the EMD-LSTM, BO-LSTM, and traditional LSTM models are significantly larger, reflecting more volatile prediction errors and lower stability.

Furthermore, the superiority of the CEEMDAN-BO-LSTM model is further demonstrated in the fitting performance plot shown in Fig. 9. The predicted points align closely with the ideal fitting line, achieving a high R^2 value of 0.9969. In contrast, the prediction points of the EMD-LSTM model and the traditional LSTM model are more dispersed, with R^2 values of only 0.9290 and 0.9840, respectively, reflecting their limitations in capturing short-term fluctuations and long-term trends. Overall, the CEEMDAN-BO-LSTM model significantly outperforms the other models in terms of fitting accuracy, prediction error, and stability.

To further validate the robustness and reliability of the proposed model, the same methodologies are applied to predict nine additional stock indices. The prediction results are summarized in Table VI.

TABLE VI. COMPARISON OF PREDICTION RESULTS ACROSS VARIOUS MODELS FOR NINE ADDITIONAL STOCK INDICES

Model / Metric	^DJI				^FTSE				^HSI			
	RMSE	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***
CEEMDAN-BO-LSTM	142.0854	112.3485	0.31%	0.9977	23.7777	18.5461	0.24%	0.9946	147.3649	111.4311	0.61%	0.9918
CEEMDAN -LSTM	259.5091	223.6002	0.61%	0.9923	42.1390	33.4897	0.43%	0.9834	200.1531	154.0437	0.86%	0.9850
EMD-BO-LSTM	231.8059	188.7823	0.54%	0.9937	25.2575	19.2848	0.25%	0.9939	158.3405	120.0416	0.66%	0.9906
EMD-LSTM	219.2336	180.4671	0.50%	0.9945	39.5812	31.2914	0.40%	0.9854	326.7038	277.6644	1.55%	0.9602
BO-LSTM	487.2037	385.5664	1.03%	0.9739	98.7783	81.2571	1.03%	0.9088	332.9701	244.7510	1.34%	0.9589
LSTM	450.2276	365.7604	1.00%	0.9777	98.6217	75.9350	0.96%	0.9091	403.8065	308.4712	1.68%	0.9395
Model / Metric	**^N225**				**000001.SS**				**^STOXX50E**			
	RMSE	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***
CEEMDAN-BO-LSTM	285.8826	184.8111	0.55%	0.9960	18.9642	14.1334	0.46%	0.9845	34.3053	27.6026	0.60%	0.9913
CEEMDAN-LSTM	408.1841	281.7839	0.81%	0.9918	39.8539	32.4014	1.07%	0.9315	51.6794	42.9811	0.93%	0.9810
EMD-BO-LSTM	408.4615	294.1020	0.84%	0.9918	24.8386	19.3618	0.64%	0.9735	39.7511	34.2836	0.75%	0.9886
EMD-LSTM	541.3611	437.5387	1.28%	0.9856	34.3530	27.1668	0.90%	0.9491	66.2227	57.1415	1.25%	0.9692
BO-LSTM	1096.3721	844.6950	2.34%	0.9412	31.8605	21.2103	0.69%	0.9563	68.8182	54.2253	1.17%	0.9661
LSTM	1471.2706	1087.955	2.97%	0.8941	40.7106	29.9967	0.96%	0.9286	115.5603	97.0145	2.09%	0.9044
Model / Metric	**^GDAXI**				**^NDX**				**^IXIC**			
	RMSE	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***	***RMSE***	***MAE***	***MAPE***	***R^2***
CEEMDAN-BO-LSTM	65.9233	51.0301	0.31%	0.9980	124.4471	99.8485	0.68%	0.9980	84.5104	66.0946	0.48%	0.9987
CEEMDAN-LSTM	172.7391	151.907	0.90%	0.9866	132.9840	105.0143	0.71%	0.9978	123.8861	101.1588	0.73%	0.9972
EMD-BO-LSTM	203.5323	168.399	0.99%	0.9814	196.8166	152.8102	0.95%	0.9950	140.0372	97.1818	0.65%	0.9963
EMD-LSTM	298.8572	243.266	1.41%	0.9603	293.6994	216.5402	1.30%	0.9927	199.6780	158.5990	1.09%	0.9926
BO-LSTM	227.5541	174.661	1.03%	0.9770	231.6684	185.2062	1.18%	0.9933	200.0174	154.5512	1.12%	0.9927
LSTM	172.7391	151.907	0.90%	0.9866	271.6015	217.8445	1.42%	0.9907	321.7919	259.8644	1.77%	0.9812

As presented in Table VI, the CEEMDAN-BO-LSTM model consistently outperforms the other models across the nine stock indices. It achieves the best results for all indices. For example, on the ^DJI, ^FTSE, and ^NDX indices, their RMSE values are 142.0854, 23.7777, and 124.4471, respectively, which are significantly lower than those of other models. Meanwhile, their MAPE values are 0.31%, 0.24%, and 0.68%, respectively, and they have the lowest error rates compared to other models. This indicates that the model consistently maintains high forecasting accuracy and low error under different market conditions.

Although the EMD-BO-LSTM model fails to achieve the highest overall accuracy, it performs well on specific stock indices (such as 000001.SS, ^FTSE, and ^HSI) with lower MAE and RMSE values than other models. This suggests that the application of the LSTM model combining EMD decomposition with Bayesian optimization in stock index prediction is effective and improves the prediction performance to some extent. Nevertheless, its performance remains inferior to that of the CEEMDAN-BO-LSTM model proposed in this paper. The results further indicate that EMD still has some limitations in terms of effectiveness and stability when compared with the CEEMDAN algorithm.

Furthermore, the superiority of the CEEMDAN-BO-LSTM model is further demonstrated in the fitting performance plot shown in Fig. 9. The predicted points align closely with the ideal fitting line, achieving a high R^2 value of 0.9969. In contrast, the prediction points of the EMD-LSTM model and the traditional LSTM model are more dispersed, with R^2 values of only 0.9290 and 0.9840, respectively, reflecting their limitations in capturing short-term fluctuations and long-term trends. Overall, the CEEMDAN-BO-LSTM model significantly outperforms the other models in terms of fitting accuracy, prediction error, and stability.

To further validate the robustness and reliability of the proposed model, the same methodologies are applied to predict nine additional stock indices. The prediction results are summarized in Table VI.

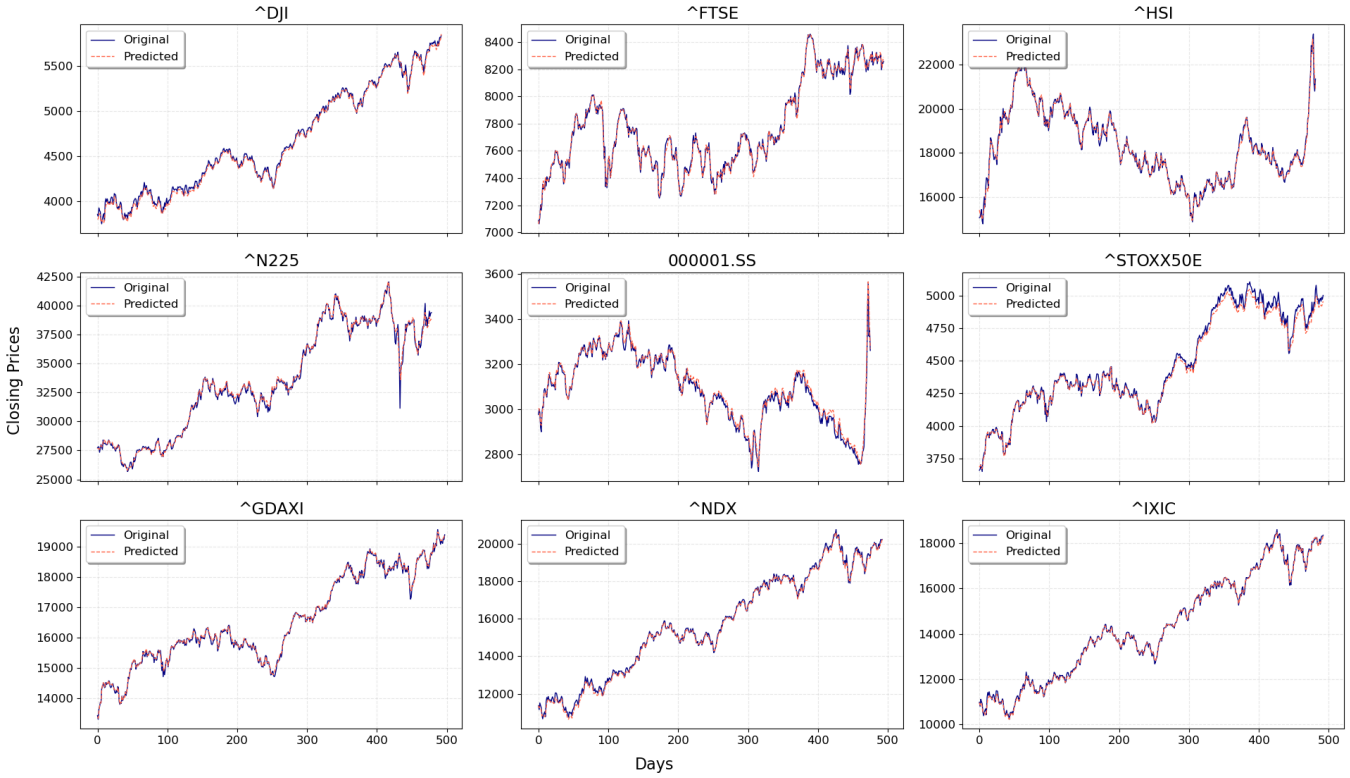

Fig. 10. Fitting performance of different models.

As shown in Fig. 10, the CEEMDAN-BO-LSTM model demonstrates strong predictive capability for the closing prices of the nine stock indices, with the predicted values closely aligning with the actual values. However, in the later part of the ^N225 test set, a sharp decline in stock price occurs, and the model underestimates the stock price decline at this point, although the model has incorporated relevant information.

Similarly, in the latter part of the ^STOXX50E test set, the closing price exhibits significant fluctuations, and the model's predictions during this period also show some deviations. These deviations suggest that extreme market volatility poses a significant challenge to predictive modeling.

Nevertheless, overall, the proposed model has demonstrated strong predictive performance, effectively adapting to complex market dynamics and providing accurate predictions in most scenarios.

V. Conclusion

In this study, a novel CEEMDAN-BO-LSTM stock market prediction model is proposed. The model follows a

decomposition-optimization-prediction-integration framework to forecast ten representative stock indices worldwide. It demonstrates excellent performance across all evaluation metrics, with an R^2 value close to 1, indicating outstanding predictive accuracy.

The experimental results show that the CEEMDAN decomposition method can improve the accuracy and stability of the decomposition, which significantly enhances the ability of the model proposed in this study to capture stock market dynamics. Meanwhile, through Bayesian optimization, the hyperparameters are precisely tuned, which not only mitigates the overfitting problem but also enhances the model's training efficiency. In addition, the model leverages the memory capability of the double-layer LSTM network for predictions, and the sliding window is optimally configured based on different frequency bands, further improving predictive accuracy.

Despite the high predictive accuracy and stability of the CEEMDAN-BO-LSTM model, certain limitations remain. Firstly, the CEEMDAN decomposition process increases computational complexity. Although the frequency band setting categorizes the data into low-frequency, medium-frequency, high-frequency, and trend components, each component requires separate model training, leading to prolonged training time. Additionally, the model's predictive performance is highly dependent on the optimization of LSTM hyperparameters, making the appropriate selection of hyperparameter search ranges a critical issue.

Future research will integrate ensemble learning techniques with deep learning models, combining the high predictive accuracy of ensemble methods with the memory capability of deep learning networks. Adaptive learning mechanisms will also be explored to improve the effectiveness of model integration and adaptation. Furthermore, external factors such as social sentiment and policy changes will be considered. Text analysis will be incorporated to enable the financial time series model to learn from socio-economic developments, enhancing the effectiveness and reliability of predictions.

ACKNOWLEDGMENT

This research is funded by Guangdong Province Key Discipline Research Capacity Enhancement Project (No. 2022ZDJS146); Key Natural Science Project of Guangdong University of Science and Technology (No. GKY-2023KYZDK-16).

REFERENCES

[1] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, "Stock price prediction using the ARIMA model," in Proceedings of the 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, IEEE, pp. 106–112, 2014.

[2] H. Herwartz, "Stock return prediction under GARCH—An empirical assessment," International Journal of Forecasting, vol. 33, no. 3, pp. 569–580, 2017.

[3] A. L. S. Maia, F. de A. T. de Carvalho, "Holt's exponential smoothing and neural network models for forecasting interval-valued time series," Int. J. Forecasting, vol. 27, no. 3, pp. 740–759, 2011.

[4] I. M. Yassin, M. F. A. Khalid, S. H. Herman, et al., "Multi-layer perceptron (MLP)-based nonlinear auto-regressive with exogenous inputs (NARX) stock forecasting model," International Journal of Advanced Science Engineering and Information Technology, vol. 7, no. 3, pp. 1098–1103, 2017.

[5] D. Zhang and S. Lou, "The application research of neural network and BP algorithm in stock price pattern classification and prediction," Future Generation Computer Systems, vol. 115, pp. 872–879, 2021.

[6] S. Hochreiter, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] K. Pawar, R. S. Jalem, and V. Tiwari, "Stock market price prediction using LSTM RNN," in Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018, Springer Singapore, pp. 493–503, 2019.

[8] J. Wang, S. Hong, Y. Dong, et al., "Predicting stock market trends using LSTM networks: overcoming RNN limitations for improved financial forecasting," Journal of Computer Science and Software Applications, vol. 4, no. 3, pp. 1–7, 2024.

[9] H. Mehtarizadeh, N. Mansouri, B. M. H. Zade, and M. M. Hosseini, "Stock price prediction with SCA-LSTM network and Statistical model ARIMA-GARCH," The Journal of Supercomputing, vol. 81, no. 2, p. 366, 2025.

[10] H. Baek, "A CNN-LSTM stock prediction model based on genetic algorithm optimization," Asia-Pacific Financial Markets, vol. 31, no. 2, pp. 205–220, 2024.

[11] S. Sang and L. Li, "A novel variant of LSTM stock prediction method incorporating attention mechanism," Mathematics, vol. 12, no. 7, p. 945, 2024.

[12] H. Zheng, J. Wu, R. Song, L. Guo, and Z. Xu, "Predicting financial enterprise stocks and economic data trends using machine learning time series analysis," Appl. Comput. Eng., vol. 87, pp. 26–32, 2024.

[13] Z. D. Akşehir and E. Kılıç, "A new denoising approach based on mode decomposition applied to the stock market time series: 2LE-CEEMDAN," PeerJ Computer Science, vol. 10, p. e1852, 2024.

[14] B. Gülmez, "Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm," Expert Systems with Applications, vol. 227, p. 120346, 2023.

[15] L. Tian, L. Feng, L. Yang, and Y. Guo, "Stock price prediction based on LSTM and LightGBM hybrid model," The Journal of Supercomputing, vol. 78, no. 9, pp. 11768–11793, 2022.

[16] Y. He and K. F. Tsang, "Universities power energy management: A novel hybrid model based on iCEEMDAN and Bayesian optimized LSTM," Energy Reports, vol. 7, pp. 6473–6488, 2021.

[17] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in Advances in Neural Information Processing Systems, vol. 25, pp. 2951–2959, 2012.

[18] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," IEEE Transactions on Signal Processing, vol. 62, no. 3, pp. 531–544, 2013.

[19] H. Niu, K. Xu, and W. Wang, "A hybrid stock price index forecasting model based on variational mode decomposition and LSTM network," Applied Intelligence, vol. 50, pp. 4296–4309, 2020.

[20] K. Su, C. Zheng, and X. Yu, "Portfolio allocation with CEEMDAN denoising algorithm," Soft Computing, vol. 27, no. 21, pp. 15955–15970, 2023.

[21] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, vol. 454, pp. 903–995, 1998.

[22] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," Advances in Adaptive Data Analysis, vol. 1, no. 1, pp. 1–41, 2009.

[23] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 4144–4147, 2011.

[24] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, "Taking the human out of the loop: A review of Bayesian optimization," Proceedings of the IEEE, vol. 104, no. 1, pp. 148–175, 2015