Science World Journal Vol. 20(No 3) 2025 https://dx.doi.org/10.4314/swj.v20i3.37
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

ENHANCING STOCK PRICE PREDICTION USING COMPLETE ENSEMBLE EMPIRICAL MODE DECOMPOSITION AND PRINCIPAL COMPONENT ANALYSIS ALGORITHMS-BASED FEATURE ENHANCEMENT

Binta Mshelbara Hassan, Abdullahi Mohammed, M.A. Bagiwa, *Abdulrazaq Abdulrahim, Ibrahim Hayatu Hassan

Department of Computer Science, Ahmadu Bello University, Zaria

*Corresponding Author Email Address: aabdulrahim@abu.edu.ng

ABSTRACT
Predicting stock prices is a complex task due to the nonlinear, nonstationary, and noisy characteristics of financial time series data. Traditional statistical and economic models often fail to capture the intricate and dynamic behavior of stock markets. To address these challenges, this study proposes three hybrid deep learning architectures that integrate advanced preprocessing and sequence modeling techniques. First, Complete Ensemble Empirical Mode Decomposition (CEEMD) is employed to denoise and decompose the financial time series into Intrinsic Mode Functions (IMFs). Then, Principal Component Analysis (PCA) is applied to extract the most significant features from the IMFs and reduce dimensionality. The transformed data is processed by a Convolutional Neural Network (CNN) to capture local patterns, and subsequently by either a Long Short-Term Memory (LSTM) network or a Bidirectional LSTM (BiLSTM) network to model temporal dependencies. The proposed architectures, including CEEMD-PCA-CNN-LSTM, CEEMD-PCA-CNN-BiLSTM, and CEEMD-PCA-BiLSTM are evaluated using four major stock indices: S&P 500, Dow Jones, DAX, and Nikkei 225. Results show that combining CEEMD and PCA significantly enhances predictive performance. The CEEMD-PCA-CNN-BiLSTM model outperforms others for the Dow Jones and Nikkei 225 datasets, achieving reductions in Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) by up to 93.7%, 92.7%, and 90.2%, respectively. The CEEMD-PCA-BiLSTM model yields the best results for the S&P 500 and DAX indices, with reductions in RMSE, MAE, and MAPE reaching up to 96.8%. These findings demonstrate the effectiveness of combining decomposition, feature selection, and deep learning for robust stock price prediction.

Keywords: Stock Price Prediction; Financial Time Series Forecasting; Complete Ensemble Empirical Mode Decomposition (CEEMD); Principal Component Analysis (PCA); Noise Reduction in Time Series; Bidirectional LSTM (BiLSTM)

INTRODUCTION

The stock market plays a vital role in economic development by facilitating capital flow and enabling investors to trade shares based on various market conditions, including supply and demand dynamics, company performance, macroeconomic indicators, and investor sentiment (Al Qaisi et al., 2016). The price at which a share is currently traded referred to as the stock price is a reflection of these fluctuating factors and is often subject to significant volatility. Stock trading is inherently high-risk and high-reward, attracting investors who seek to capitalize on price movements by accurately forecasting future trends. Predicting stock prices involves analyzing historical data to model and anticipate market behavior. However, the financial time series underlying stock prices are typically nonlinear, nonstationary, and highly noisy characteristics that pose significant challenges to accurate prediction (Chandar et al., 2016; Polanco-Martínez, 2019). Traditional forecasting models based on statistical or econometric methods often fall short in capturing the complex and chaotic nature of stock markets. As a result, modern predictive approaches increasingly rely on artificial intelligence (AI) and machine learning (ML), particularly deep learning (DL) techniques, which have demonstrated strong capabilities in modeling intricate temporal and spatial data dependencies (Selvin et al., 2017).

To enhance the performance of deep learning models in financial prediction, appropriate preprocessing techniques are essential. One such technique is Complete Ensemble Empirical Mode Decomposition (CEEMD), an adaptive signal decomposition method capable of breaking down noisy and nonstationary time series into a set of Intrinsic Mode Functions (IMFs) (Srijiranon et al., 2022). This process helps in isolating meaningful patterns and trends from raw financial data. Following decomposition, Principal Component Analysis (PCA) is employed to reduce the dimensionality of the IMFs while preserving the most informative components. PCA not only streamlines the input data but also accelerates model training and enhances performance by minimizing redundancy and noise (Zhang et al., 2020; Wen et al., 2020).

Deep learning architectures such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks are widely used for feature extraction and sequential modeling. CNNs are effective in capturing local spatial patterns in time series data, while LSTM networks model temporal dependencies. BiLSTM, in particular, offers a bidirectional view of sequential dependencies, enabling the model to consider both past and future contexts, which often leads to superior predictive accuracy.

Given these advancements, this study proposes three novel hybrid deep learning architectures that integrate CEEMD, PCA, CNN, LSTM, and BiLSTM for improved stock price forecasting. The main contributions of the paper are as follows:

1. Development of Hybrid Deep Learning Architectures: We present CEEMD-PCA-CNN-LSTM, CEEMD-PCA-CNN-BiLSTM, and CEEMD-PCA-BiLSTM frameworks

that combine decomposition, dimensionality reduction, and advanced sequence modeling.
2. Enhanced Preprocessing via CEEMD and PCA: The use of CEEMD for noise reduction and PCA for feature selection ensures high-quality input for the deep learning models, thereby improving learning efficiency and accuracy.
3. CNN-Based Feature Extraction: One-dimensional CNNs are used to extract deep spatial features from the reduced feature set, enhancing the model's capacity to learn from complex patterns.
4. Temporal Modeling with BiLSTM: The integration of BiLSTM allows for capturing forward and backward dependencies, improving model generalization and performance over conventional LSTM models.
5. Empirical Validation on Benchmark Datasets: The models are tested on four major stock indices—S&P 500, Dow Jones, DAX, and Nikkei 225—demonstrating significant improvements in prediction accuracy, as evidenced by reductions in Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

These contributions collectively offer a robust framework for accurate and scalable financial time series forecasting, addressing the inherent challenges of noise, nonlinearity, and temporal complexity in stock market data.

The remainder of this paper is organized as follows: Section 2 reviews related works on stock price prediction, highlighting their strengths and limitations. Section 3 presents a detailed description of the proposed methodology, while Section 4 reports and discusses the experimental results. Finally, Section 5 concludes the study and outlines potential directions for future research.

Numerous studies have employed deep learning and hybrid frameworks to improve the accuracy of stock price prediction. Zhang et al. (2023) proposed a CNN-BiLSTM-Attention model that outperforms LSTM, CNN-LSTM, and other variants by integrating bidirectional temporal features and an attention mechanism that effectively weighs the importance of input data. Despite its strong generalization and robustness, the model struggles with prediction accuracy during sharp price spikes, suggesting further enhancement through integration of multi-source heterogeneous data or newer architectures.

Srijiranon et al. (2022) developed a hybrid framework combining FinBERT, PCA, EMD, and LSTM, emphasizing sentiment analysis as a key factor in prediction. The Empirical Mode Decomposition (EMD) technique enhanced performance by decomposing signals into components, allowing tailored LSTM training. While PCA helped minimize prediction errors in initial intrinsic mode functions (IMFs), the model could not utilize sentiment data for decomposed signals, highlighting a limitation that could be mitigated by adaptive hybrid algorithm selection or alternative decomposition strategies like Hybrid Time Series Decomposition (HTD).

Li (2022) focused on feature selection in LSTM models, identifying key features that significantly influence prediction outcomes. The study highlighted the minimal impact of the lowest stock price and suggested further exploration of input dimension optimization, network architecture, and parameter tuning. Rezaei et al. (2021) introduced a CEEMD-CNN-LSTM model that leveraged decomposition and deep learning layers for enhanced predictive performance. While effective, the use of basic CNN and LSTM configurations limited the potential, prompting future work to explore deeper architectures and more sophisticated algorithms. Sunny et al. (2020) compared LSTM with Bi-Directional LSTM (BiLSTM) and found the latter to yield better performance due to its dual-direction learning capability. The BiLSTM achieved lower RMSE values, but results were highly dependent on parameter settings. Further research was proposed to apply the model to broader market categories.

Tsantekidis et al. (2020) presented a CNN-LSTM approach that utilized stationary limit order book features and demonstrated superior stability compared to SVMs, MLPs, and other deep models. Incorporating attention mechanisms was suggested to enhance the network's ability to focus on relevant information while minimizing noise. Zhang et al. (2020) introduced a CEEMD-PCA-LSTM model that effectively handled nonlinear and multi-scale financial time series data. CEEMD aided in data denoising while PCA reduced dimensionality and extracted high-level features. The study recommended incorporating more data types such as technical indicators and macroeconomic variables for improved accuracy and suggested hyperparameter optimization for future work.

Wen et al. (2020) employed a PCA-LSTM model that showed better prediction accuracy than LSTM, CNN, and MLP models. The paper recommended further model and parameter optimization to improve outcomes. Yu and Yan (2020) proposed a PSR-NN-based LSTM network that followed a structured pipeline of data processing, model building, and evaluation. The model demonstrated superior predictive capability compared to other approaches. Yeh et al. (2010) explored Complementary Ensemble Empirical Mode Decomposition (CEEMD) as a data preprocessing method. CEEMD effectively removed noise residues without increasing computational cost, making it a recommended standard over EEMD for noise-reduction tasks in time series analysis.

MATERIALS AND METHODS

This section outlines the sources of the datasets, the data preparation process, and the model training procedures, along with all other materials and methods employed in developing the proposed model.

Data Pre-Processing and Description

The proposed algorithm is trained using historical daily financial time series data from January 4, 2010, to September 27, 2019, obtained from the Yahoo Finance website. The dataset includes the daily closing prices of the S&P 500, Dow Jones, DAX, and Nikkei 225. Table 1 provides a detailed description of the data

Table 1: Dataset analysis

Index	Count	Mean	Min	Max	Standard Deviation
S&P500	2450	1932.946	1022.580	3025.86	567.197
Dow Jones	2450	17343.64	9686.480	27359.16	4949.94
DAX	2473	9444.774	5072.330	13559.599	2350.467
Nikkei 225	2406	15595.31	8160.009	24270.619	4852.428

Complete Ensemble Empirical Mode Decomposition (CEEMD)

The CEEMD frequency decomposition algorithm is a derivative of

the Empirical Mode Decomposition model (Yeh et al., 2010). It has the ability to provide a reconstructed noise-free time series, by decomposing complex signal into a series of intrinsic mode function (IMF) (Purba et al., 2018). Decomposed time series models are more accurate than undecomposed time series models, this is because decomposing time series models CEEMD and EMD algorithms play a significant role in the analysis of nonlinear and non-stationary time series (Rezaei et al., 2021). The CEEMD is an extension method of Ensemble Empirical Mode Decomposition (EEMD) derived from the EMD algorithm, which can separate the high-frequency noise from the raw data by adding the white noise, but the low frequency noise cannot be reduced hence the use of opposite white noise by CEEMD to remove the low frequency noise. This method is an adaptive signal analysis approach based on the signal characteristics of local extrema. It decomposes a time series of Intrinsic Mode Function (IMF) components, and each IMF satisfies the following conditions:

1. Over the entire time range, the number of zero-crossings must be equal or differ by one at most.
2. At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero (the mean value of the upper and lower envelope is zero everywhere).

The steps for CEEMD decomposition are as follows:

1. By adding N pairs of positive and negative Gaussian white noise to the original signal, 2N signal sets are denoted by equation 1:

$$\begin{bmatrix} M_1 \\ M_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} S \\ N \end{bmatrix} \quad (1)$$

Where S is the original signal, N is the Gaussian white noise, M_1 is the sum of the original data with positive noise, and M_2 is the sum of the original data with the negative noise. Similarly, the ensemble IMFs obtained from those negative mixtures contribute to another set of ensemble IMFs with negative residue of added white noises. Thus, the final IMF is the ensemble of both the IMFs with positive and negative noises.

2. EMD decomposition is performed on the target signal, and each signal obtains a set of IMF components, in which the i^{th} component of the j^{th} IMF is expressed as imf_{ij}.
3. Obtain the results of each IMF after averaging the overall ensemble, which can be formulated by equation 2:

$$imf_j = \frac{1}{2N}\sum_{i=1}^{2N} imf_{ij} \quad (2)$$

Hence, the final decomposition result, x (t), of CEEMD can be denoted by equation 3:

$$x(t) = \sum_{j=1}^{k} imf_j(t) + res \quad (3)$$

Where res is the trend term, res = R(t), representing the trend of original time series S(t), and $imf_j(t)$ is the j^{th} IMF, K is the total number of IMFs of EMD (Yeh et al., 2010)

Principal Component Analysis (PCA)

PCA is applied to reduce the dimensionality of decomposed IMFs, while maintaining the features of the largest variance contribution of the dataset. Feature extraction and reduction is done using the PCA on the IMFs. The PCA with 3 components is picked to be passed into the prediction model. Normalization of the data is important after the decomposition of indices into various frequency spectra; this leads to the adjustment of the data noise effect and implementation of neural networks with high efficiency and speed. By doing so, all features will be transformed into the range [0, 1], meaning that the minimum and maximum value of a feature/variable is going to be between 0 and 1 respectively (Jamal et al., 2014).

$x_i(t)_{scaled}$ denotes normalized data, $x_i(t)$ is ith data of the imfs by equation 4 and equation 5.

$$x_i(t)_{scaled} = \frac{xi(t) - minxi(t)}{xi(t) - minxi(t)} \quad (4)$$

$$x\ \ (t) = \frac{xi(t) - minxi(t)}{maxxi(t) - minxi(t)} \quad (5)$$

Where, $x_i(t)$ is i^{th} data of IMF and $x\ \ (t)$ denotes its normalized data.

The main idea behind normalization is that variables that are measured at different scales do not contribute equally to the model fitting and the model learned function, and might end up creating bias (Botalb et al., 2018).

Convolutional Neural Network (CNN)

The CNN consists of five major components, which are the input layer, convolutional layer, pooling layer, fully connected layer, and output layer. The convolutional layer and the pooling layer are the focus of the model structure as they are capable of extracting features of images. They are mainly used to extract features of the input data and perform dimensionality reduction on the features. The vanishing gradient problem in this study is solved using the ReLu activation function. The CNN model used a one-dimensional convolutional hidden layer of 512 filters of size two and one-dimensional max-pooling of size 2 with a kernel size of 3 on the financial time series data. The output of these layers is transferred to the next stage, which is either the LSTM or BiLSTM to play the role of fully connected layers (Tsantekidis et al., 2020).

Long Short-Term Memory (LSTM) Model

The traditional LSTM is a type of deep recurrent neural network that has the capacity to process sequential data. LSTM network realizes temporal memory function through switch of the gate, and can effectively solve the problem of gradient vanishing and explosion in recurrent neural network. It uses past time series information to predict the next moment's output as it is capable of retaining long time information.

Bi-directional Long Short-term Memory (BiLSTM)

The BiLSTM model is used to learn and predict the extracted feature data (Lu, 2020). The BiLSTM is able to take into account past and future information by connecting the forward LSTM layer and a backward LSTM layer, which facilitates both forward and backward sequence information input thereby making the model more robust. BiLSTM neural network is used to learn the bidirectional serial features from the feature information extracted from the CNN layer, fully exploit the long-term dependent features of the sample data for learning, and finally output the stock price prediction results through the fully linked layer.

Deep Learning Hybrid Prediction Model

The proposed deep learning hybrid prediction model integrates Complementary Ensemble Empirical Mode Decomposition (CEEMD), Principal Component Analysis (PCA), Convolutional Neural Networks (CNN), and Bidirectional Long Short-Term

Science World Journal Vol. 20(No 3) 2025 https://dx.doi.org/10.4314/swj.v20i3.37
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Memory (Bi-LSTM) networks. As illustrated in Figure 1, the framework comprises three to four distinct phases, depending on the specific model configuration:

1. CEEMD is employed to decompose the financial time series data, effectively reducing noise and enhancing signal clarity.
2. PCA is applied to reduce the dimensionality of the decomposed components, extract high-level abstract features, and improve computational efficiency.
3. CNN is utilized to capture complex nonlinear local patterns within the data.
4. LSTM or Bi-LSTM is then used to learn temporal dependencies and generate the final prediction output.

Evaluation Metric

The loss error is used to evaluate the prediction results, which refers to the difference between the actual observed value and the predicted value. To evaluate the effectiveness of the models, the performance of the stock price model is measured using the Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE), represented by equations 6, 7, and 8, respectively

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (6)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \frac{|y_i \quad \hat{y}_i|}{y_i} \times 100 \quad (7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \times 100 \quad (8)$$

Where y_i denotes actual value, $\hat{y}_i$ denotes predicted value and $\overline{y}_i$ denotes the mean of y value. The results are more accurate and reliable when the values of the metrics are lower (Ahmed et al., 2020).

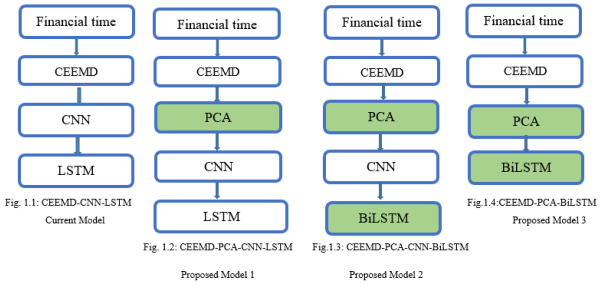


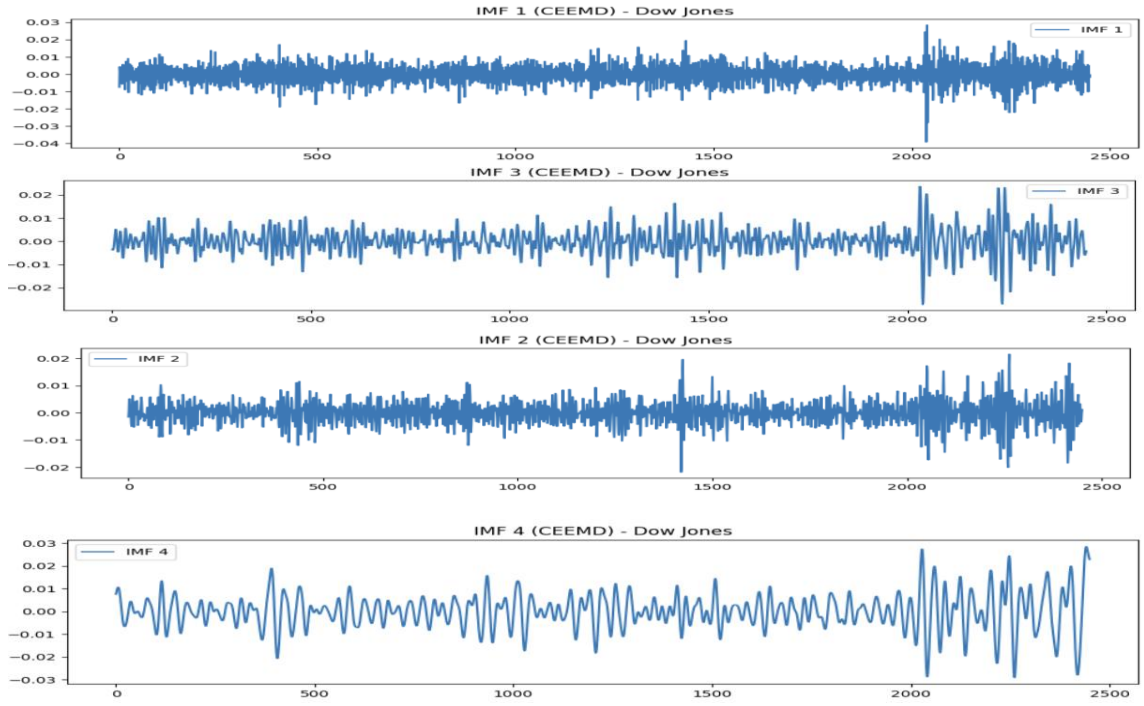
Figure 1 Model framework

RESULTS AND DISCUSSION

This section presents the evaluation of the proposed hybrid algorithms, CEEMD-PCA-CNN-LSTM, CEEMD-PCA-CNN-BiLSTM, and CEEMD-PCA-BiLSTM, and compares their performance against the baseline CEEMD-CNN-LSTM model. The models are assessed using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) to determine their predictive accuracy and overall effectiveness.

Complete Ensemble Empirical Mode Decomposition (CEEMD)

The frequency decomposition algorithms decompose data into various frequency spectra consisting of several IMFs and one residual. The IMFs are arranged from high to low frequency, with the residual component occurring at the end. Training the networks on IMFs with low volatility is easier than IMFs with high volatility (Rezaei et al., 2021). Figure 2 shows a sample decomposed dataset of Dow Jones using the CEEMD algorithms.



Science World Journal Vol. 20(No 3) 2025 https://dx.doi.org/10.4314/swj.v20i3.37
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

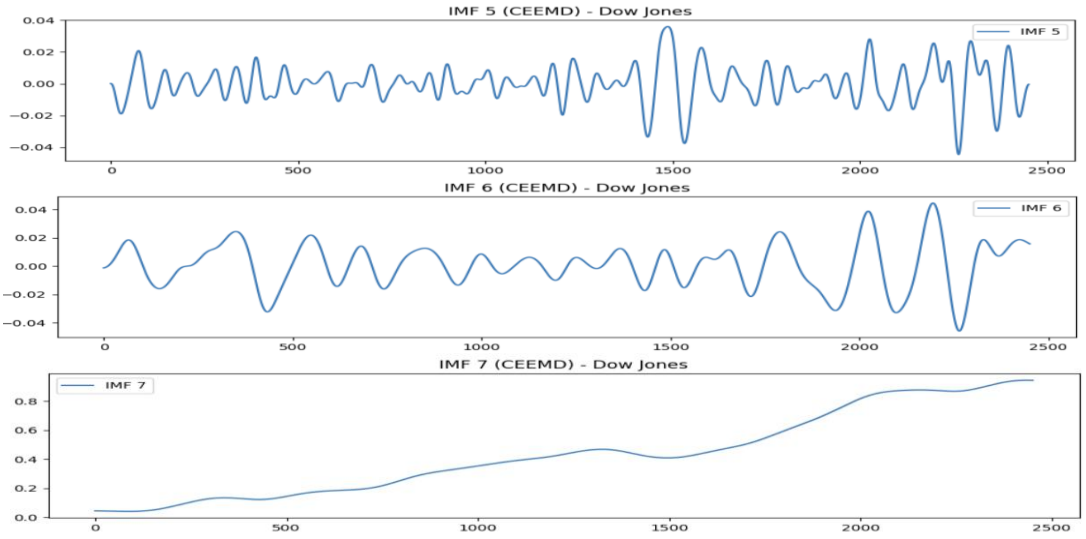


Figure 1: Dow Jones dataset decomposed into IMF components.

The Training Process and Prediction Results
Better prediction is obtained when a sliding window is applied for long-term time series. Data is divided into various intervals for training, with a window of constant length of 4 is applied before the training algorithm is used. The data is split into 80% for the training set and 20% for the testing set.

CNN-LSTM model
The data is passed into the CNN for data pattern extraction with 512 filters, a kernel size of 3, and using the Relu activation function. The model data is first processed through the convolutional layers, followed by the max-pooling layer of size 2 and then the LSTM model with 200 hidden units is used after the pattern extraction and data dynamics to analyse the processed data of two previous layers with a high ability to maintain the time sequences instead of a fully connected CNN layer (Tsantekidis et al., 2020). The number of epochs used for the training is 200 which implies the entire number of times the data is executed for fitting the model (Rezaei et al., 2021).

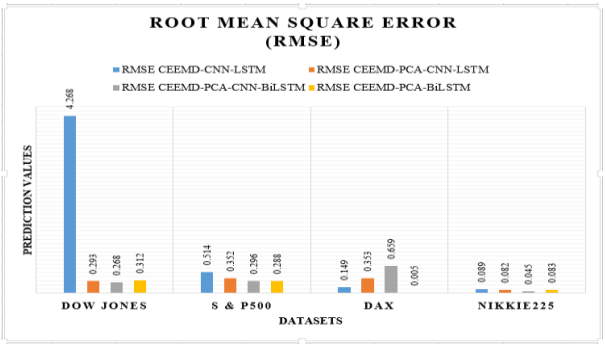
CNN-BiLSTM model
The CNN layer, extracts the serial features from the data and the BiLSTM model fully exploits the long-term dependent features of the sample data for learning and finally output the stock price prediction results through the fully linked layer (Zhang et al., 2023). The BiLSTM is used to learn the bidirectional serial features from the feature information extracted from the data.

Proposed Algorithm Results
The proposed model is a hybrid algorithm consisting of frequency decomposition, feature extraction and deep learning models (CNN-LSTM, CNN-BiLSTM and BiLSTM) which has been previously explained in detail in the preceding sections. The time series data is decomposed into various frequency spectra using CEEMD algorithm and the analysed IMFs as well as the residual is passed into the PCA to extract the relevant features in the IMF. The IMFs with 3 components are passed into the CNN to extract the non-linear local features of the data and finally into the LSTM or BiLSTM model to predict the sample data. We compare the result of the proposed model with that of Rezaei et al (2021) that used that used CEEMD-CNN-LSTM and CEEMD-PCA-CNN-LSTM based on evaluation metrics such as RMSE, MAE, MAPE as shown in the table 2, 3 and 4, and figure 3, 4 and 5. However, our proposed model achieved highest results.

The RMSE results using four datasets
This section examines the RMSE results of the proposed models using four distinct datasets. The performance of these models is compared with the findings reported by Rezaei et al. (2021) to assess improvements in predictive accuracy. The results are represented in table 2 and figure 3.

Table 2: The RMSE results of the four models using four datasets

Datasets	CEEMD-CNN-LSTM	CEEMD-PCA-CNN-LSTM	CEEMD-PCA-CNN-BiLSTM	CEEMD-PCA-BiLSTM
Dow Jones	*4.26793*	*0.29296*	***0.26837***	*0.31208*
S&P500	*0.51412*	*0.35192*	*0.29645*	***0.28820***
DAX	*0.14918*	*0.35261*	*0.65918*	***0.00482***
Nikkie 225	*0.08931*	*0.82136*	***0.04461***	*0.08260*



Science World Journal Vol. 20(No 3) 2025
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University
https://dx.doi.org/10.4314/swj.v20i3.37

Figure 3: RMSE results for the CEEMD-CNN-LSTM, CEEMD-PCA-CNN-LSTM, CEEMD-PCA-CNN-BiLSTM and CEEMD-PCA-BiLSTM models.

The MAE results using four datasets

This section presents the Mean Absolute Error (MAE) results for the proposed models evaluated across four different datasets. The performance of these models is compared to the results reported by Rezaei et al. (2021) to assess the relative effectiveness of the proposed approach. Table 3 and Figure 4 represent the results in this section.

Table 3: The MAE results of the four models using four datasets

Datasets	CEEMD-CNN-LSTM	CEEMD-PCA-CNN-LSTM	CEEMD-PCA-CNN-BiLSTM	CEEMD-PCA-BiLSTM
Dow Jones	2.57886	0.20479	**0.18932**	0.21572
S&P500	0.27971	0.24836	0.20984	**0.20292**
DAX	0.08137	0.20554	0.38878	**0.00339**
Nikkie225	0.06128	0.05937	**0.031437**	0.056235

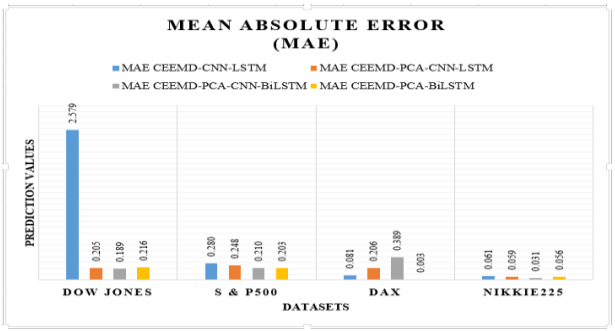


Figure 4: MAE results for the CEEMD-CNN-LSTM, CEEMD-PCA-CNN-LSTM, CEEMD-PCA-CNN-BiLSTM, and CEEMD-PCA-BiLSTM models.

The MAPE results using four datasets

This section presents the Mean Absolute Percentage Error (MAPE) results for the proposed models, evaluated across four distinct datasets. The performance of these models is compared with the results from Rezaei et al. (2021) to gauge their relative performances. By analysing the MAPE values, we aim to provide a comprehensive understanding of the models' prediction performance and assess their effectiveness in comparison to previous research as shown in table 4 and figure 5.

Table 4: The MAPE results of the four models using four datasets

Datasets	CEEMD-CNN-LSTM	CEEMD-PCA-CNN-LSTM	CEEMD-PCA-CNN-BiLSTM	CEEMD-PCA-BiLSTM
Dow Jones	431.79872	44.76181	**41.93632**	46.06914
S&P500	52.20201	54.4211	44.36883	**42.19995**
DAX	16.18961	29.46627	53.58397	**1.466112**
Nikkie225	14.21728	15.13042	**8.67288**	12.32794

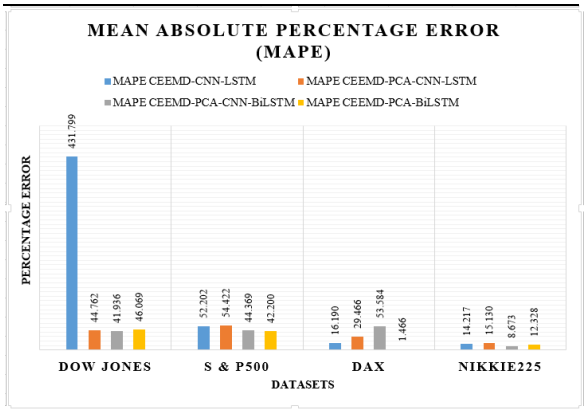


Figure 5: MAPE results for the CEEMD-CNN-LSTM, CEEMD-PCA-CNN-LSTM, CEEMD-PCA-CNN-BiLSTM and CEEMD-PCA-BiLSTM models.

Comparing the Performance of CEEMD-PCA-CNN-LSTM with CEEMD-CNN-LSTM model.

The results in Figure 6 show that introducing the PCA to the CEEMD-CNN-LSTM shows a significant reduction of 93.136%, 92.059% and 89.634% for the RMSE, MAE and MAPE, respectively, for the Dow Jones dataset. A reduction of 31.549% and 11.206% for the RMSE and MAE, respectively, for the S&P500, however, the MAPE increased with the PCA to 4.253%. On the other hand, an increase of 136.365%, 152.593% and 82.007 for the RMSE, MAE, and MAPE for the DAX datasets with the PCA. A reduction of 8.036 and 3.132% for the RMSE and MAE, respectively for Nikkie225 and an increase of 6.423 for the MAPE. The results show that the PCA did not improve the performance of the DAX dataset on the RMSE, MAE and MAPE. Also, the MAPE of S&P500 and Nikkie225 was not improved with the PCA. The performance of the model is improved with the PCA model on the Dow Jones for the RMSE, MAE and MAPE. The S&P500 and Nikkie225 datasets improved the RMSE and MAE with the PCA.

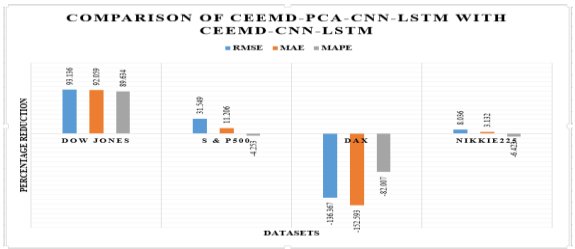


Figure 6: Comparing the CEEMD-PCA-CNN-LSTM with CEEMD-CNN-LSTM models on the RMSE, MAE and MAPE.

Science World Journal Vol. 20(No 3) 2025 https://dx.doi.org/10.4314/swj.v20i3.37
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Comparing the Performance of CEEMD-PCA-CNN-BiLSTM with CEEMD-CNN-LSTM model.

From the results in Figure 7, the performance of CEEMD-CNN-LSTM model is compared with the performance of CEEMD-PCA-CNN-BiLSTM model, and the latter shows a significant reduction of 93.712%, 92.659% and 90.288% for the RMSE, MAE and MAPE, respectively for the Dow Jones dataset. A reduction of 42.338%, 24.978% and 15.006% for the RMSE, MAE and MAPE, respectively for the S&P500. On the other hand, an increase of 341.870%, 377.774% and 230.977% for the RMSE, MAE and MAPE for the DAX datasets. A reduction of 50.048%, 48.704% and 38.998% for the RMSE, MAE, and MAPE, respectively for Nikkie225. The results show that, the PCA and BiLSTM did not improve the performance of the DAX dataset on the RMSE, MAE and MAPE. However, the performance of the model is improved with the PCA and BiLSTM models on the Dow Jones, S&P500 and Nikkie225 datasets.

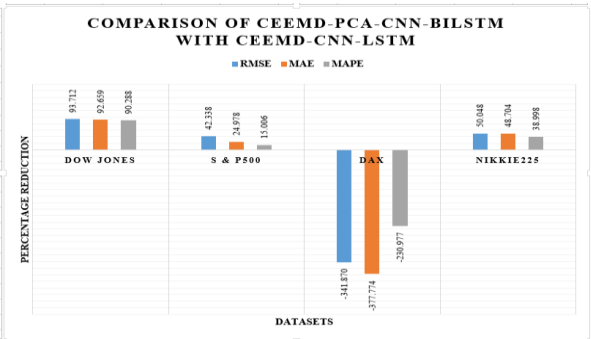


Figure 7: Comparing the CEEMD-PCA-CNN-BiLSTM with CEEMD-CNN-LSTM models on the RMSE, MAE and MAPE.

Comparing the Performance of CEEMD-PCA-BiLSTM with CEEMD-CNN-LSTM model.

Figure 8 shows the performance of CEEMD-CNN-LSTM model compared with the performance of CEEMD-PCA-BiLSTM model, and the latter shows a significant reduction of 92.688%, 91.635% and 89.331% for the RMSE, MAE, and MAPE, respectively, for the Dow Jones dataset. A reduction of 43.942%, 27.453% and 19.160% for the RMSE, MAE, and MAPE for the S&P500, respectively. Also, a reduction of 96.771%, 95.831% and 90.944% for the RMSE, MAE, and MAPE for the DAX datasets. A reduction of 7.513%, 8.239% and 13.289% for the RMSE, MAE, and MAPE, respectively, for Nikkie225. The results show that, the CEEMD-PCA-BiLSTM model performed well on all datasets when compared to the CEEMD-CNN-LSTM model.

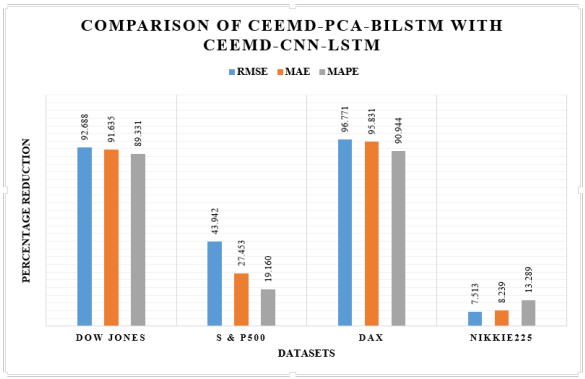


Figure 8: Comparing the CEEMD-PCA-BiLSTM with CEEMD-CNN-LSTM models on the RMSE, MAE, and MAPE.

Exploring the relationship between the Mean, Standard Deviation, and the Results obtained

The most competitive result is seen in CEEMD-PCA-CNN-BiLSTM, which performs better on Dow Jones and Nikkie225. However, the CEEMD-PCA-BiLSTM performed better on S&P500 and DAX. In examining the quality of the dataset, Figure 9 shows that the two datasets that performed best on the CEEMD-PCA-CNN-BiLSTM, which are the Dow Jones and the Nikkie225 have a higher mean and standard deviation. The datasets that performed best on the CEEMD-PCA-BiLSTM, which are the S&P500 and DAX have a lower mean and standard deviation. From the analysed result, the presence of the CNN seems to make a significant contribution when the mean and standard deviation is high (Dow Jones and the Nikkie225). However, the model will perform well without the CNN when the mean and standard deviation is low (S&P500 and DAX). The above results show that it is important to study the nature and the characteristics of datasets in order to ascertain the qualities of datasets that make them to perform well on some models and not so well on others.

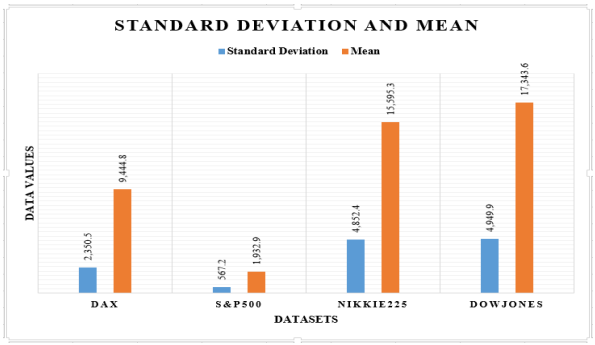


Figure 9: The mean and standard deviation of DAX, S&P 500, Nikkie225 and Dow Jones datasets.

DISCUSSION

The empirical findings of this study demonstrate that integrating CEEMD, PCA, and deep learning models yields substantial improvements in stock price prediction compared to baseline architectures. Specifically, the incorporation of PCA alongside CEEMD consistently enhanced predictive performance by reducing redundancy in the decomposed IMFs and emphasizing the most informative features. This confirms the importance of effective preprocessing in handling the nonlinear and noisy characteristics of financial time series, a challenge often highlighted in prior works

(Zhang et al., 2020; Rezaei et al., 2021).

A notable outcome is the superior performance of BiLSTM-based models across multiple datasets, with the CEEMD-PCA-CNN-BiLSTM excelling on the Dow Jones and Nikkei 225 indices, while the CEEMD-PCA-BiLSTM achieved the best accuracy on the S&P 500 and DAX indices. This aligns with earlier evidence that BiLSTM, by considering both past and future temporal dependencies, often outperforms unidirectional LSTM in capturing the dynamics of volatile stock markets (Sunny et al., 2020; Zhang et al., 2023). Furthermore, the observed dataset-specific behavior—where models with CNN layers perform better on high-variance datasets (Dow Jones, Nikkei 225), while models without CNN yield superior results on low-variance datasets (S&P 500, DAX)—suggests that the statistical properties of financial indices should guide model selection. This insight contributes to the methodological literature by highlighting the role of mean and variance in influencing model suitability, an area underexplored in existing research.

When compared with related studies, our hybrid frameworks achieve notable error reductions. For instance, Rezaei et al. (2021) reported effective results with CEEMD-CNN-LSTM, but our CEEMD-PCA-CNN-BiLSTM improved RMSE, MAE, and MAPE by over 90% on some indices. Similarly, Zhang et al. (2023) demonstrated the potential of CNN-BiLSTM-Attention models, yet their framework struggled during sharp market fluctuations. Our results suggest that decomposition and feature selection steps (CEEMD and PCA) help mitigate these weaknesses by stabilizing the input signals before temporal modeling. Thus, the current work provides complementary evidence that preprocessing is just as critical as network architecture in financial prediction tasks.

From a practical standpoint, these findings have implications for financial analysts, algorithmic traders, and portfolio managers. More reliable stock price forecasts can enhance trading strategies, improve risk management, and support decision-making under uncertainty. For example, the reduced forecasting error rates observed in this study could translate into more precise entry and exit points for traders, thereby increasing profitability and reducing exposure to losses. Moreover, the relatively low computational overhead introduced by PCA suggests that the proposed methods are scalable and adaptable for real-time applications.

Despite these contributions, certain limitations must be acknowledged. First, the models rely exclusively on historical closing prices, without incorporating external variables such as trading volume, macroeconomic indicators, or investor sentiment, which have been shown to influence stock dynamics (Srijiranon et al., 2022). Second, the evaluation is limited to four major indices, which, while diverse, may not fully capture the heterogeneity of emerging or less liquid markets. Third, although CEEMD effectively addresses noise, it is still computationally intensive, potentially limiting its applicability in high-frequency trading contexts. Future research could explore lighter decomposition methods or approximate CEEMD variants to balance accuracy and speed.

Going forward, several extensions are worth considering. Integrating sentiment analysis from financial news or social media could enrich the feature space, as demonstrated in FinBERT-based studies. Attention mechanisms could also be added to further enhance feature weighting and improve performance during abrupt price fluctuations. Additionally, hyperparameter optimization using metaheuristic algorithms may further refine the predictive capabilities of the hybrid models. Finally, testing the frameworks on cryptocurrency or commodity markets, which exhibit even higher volatility, would help establish the generalizability of the approach.

Conclusion

The major concept suggested by the algorithm was to create a collaboration between CEEMD, PCA, CNN, LSTM, BiLSTM and CNN-BiLSTM models by combining them together which could process the data, reduce the effect of noise and extract deep features in time sequences yielding to a better prediction accuracy. The CEEMD-PCA-CNN-BiLSTM and the CEEMD-PCA-BiLSTM yielded competitive results on the datasets and also outperformed the CEEMD-CNN-LSTM and the CEEMD-PCA-CNN-LSTM on all datasets across all metrics of evaluation. From the datasets obtained, it can be seen that the mean and standard deviation of S&P500 and DAX is low in the two datasets that performed well on the CEEMD-PCA-BiLSTM. However, it is seen that the mean and standard deviation of the datasets is high on the two datasets that performed well on the CEEMD-PCA-CNN-BiLSTM which is the Dow Jones and Nikkie225. The best results obtained from the CEEMD-PCA-BiLSTM has further shown that not all datasets will require the use of the CNN model. The CEEMD-PCA-BiLSTM model performed well on the datasets with a low mean and standard deviation and the CEEMD-PCA-CNN-BiLSTM model shows that the CNN is relevant on datasets with a high mean and a high standard deviation. However, this will require future works to further prove the influence of mean and standard deviation in choosing a model for stock price data prediction. This study goes to show the importance of exploratory analysis and understanding the quality and nature of datasets before making a choice of a prediction model.

REFERENCES

Al Qaisi, F., Tahtamouni, A., & Al-Qudah, M. (2016). Factors affecting the market stock price-The case of the insurance companies listed in Amman Stock Exchange. International Journal of Business and Social Science, 7(10), 81-90. International Journal of Business and Social Science, 7(10), 81–90.

Botalb, A., Moinuddin, M., Al-Saggaf, U. M., & Ali, S. S. A. (2018). Contrasting Convolutional Neural Network (CNN) with Multi-Layer Perceptron (MLP) for Big Data Analysis. International Conference on Intelligent and Advanced System, ICIAS 2018, February 2019, 1–5. https://doi.org/10.1109/ICIAS.2018.8540626

Chandar, S. K., Sumathi, M., & Sivanandam, S. N. (2016). Prediction of Stock Market Price using Hybrid of Wavelet Transform and Artificial Neural Network. Indian Journal of Science and Technology, 9(8), 0974-5645 9(February). https://doi.org/10.17485/ijst/2016/v9i8/87905

Jamal, P., Ali, M., Faraj, R. H., Ali, P. J. M., & Faraj, R. H. (2014). 1-6 Data Normalization and Standardization: A Technical Report. Machine Learning Technical Reports, 1(1), 1–6.

Li, D. (2022). Feature Selection Based on Stock Prediction Model. Journal of Physics: Conference Series, 2386(1), 4–9. https://doi.org/10.1088/1742-6596/2386/1/012021

Lu, W. (2020). A CNN-BiLSTM-AM method for stock price prediction. Neural Computing and Applications, 0123456789. https://doi.org/10.1007/s00521-020-05532-z

Polanco-Martínez, J. M. (2019). Dynamic relationship analysis

between NAFTA stock markets using nonlinear, nonparametric, non-stationary methods. Nonlinear Dynamics, 97(1), 369–389. https://doi.org/10.1007/s11071-019-04974-y

Purba, H., Musu, J. T., Diria, S. A., Permono, W., Sadjati, O., Sopandi, I., & Ruzi, F. (2018). Completed Ensemble Empirical Mode Decomposition: A Robust Signal Processing Tool to Identify Sequence Strata. IOP Conference Series: Earth and Environmental Science, 132(1). https://doi.org/10.1088/1755-1315/132/1/012033

Rezaei, H., Faaljou, H., & Mansourfar, G. (2021). Stock price prediction using deep learning and frequency decomposition. Expert Systems with Applications, 169, 114332. https://doi.org/10.1016/j.eswa.2020.114332

Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. 2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017, 2017-January (January 2018), 1643–1647. https://doi.org/10.1109/ICACCI.2017.8126078

Srijiranon, K., Lertratanakham, Y., & Tanantong, T. (2022a). A Hybrid Framework Using PCA, EMD and LSTM Methods for Stock Market Price Prediction with Sentiment Analysis. Applied Sciences (Switzerland), 12(21). https://doi.org/10.3390/app122110823

Srijiranon, K., Lertratanakham, Y., & Tanantong, T. (2022b). A Hybrid Framework Using PCA, EMD and LSTM Methods for Stock Market Price Prediction with Sentiment Analysis. Applied Sciences (Switzerland), 12(21). https://doi.org/10.3390/app122110823

Sunny, A. I., Mohd, M., Maswood, S., & Alharbi, A. G. (2020). Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model. 87–92.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2020). Using Deep Learning for price prediction by exploiting stationary limit order book features. Applied Soft Computing Journal, 93, 106401. https://doi.org/10.1016/j.asoc.2020.106401

Wen, Y., Lin, P., & Nie, X. (2020). Research of stock price prediction based on PCA-LSTM model. IOP Conference Series: Materials Science and Engineering, 790(1). https://doi.org/10.1088/1757-899X/790/1/012109

Yeh, J. R., Shieh, J. S., & Huang, N. E. (2010). Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method. Advances in Adaptive Data Analysis, 2(2), 135–156. https://doi.org/10.1142/S1793536910000422

Yu, P., & Yan, X. (2020). Stock price prediction based on deep neural networks. Neural Computing and Applications, 32(6), 1609–1628. https://doi.org/10.1007/s00521-019-04212-x

Zhang, J., Ye, ., & Lai, Y. (2023). Stock Price Prediction Using CNN-BiLSTM-Attention Model. Mathematics, 11(9), 1–18. https://doi.org/10.3390/math11091985

Zhang, Y., Yan B., & Aasma, M. (2020). A novel deep learning framework: Prediction and analysis of financial time series using CEEMD and LSTM. Expert Systems with Applications, 159, 113609. https://doi.org/10.1016/j.eswa.2020.113609