Research on SZSE Component Index Volatility Prediction Based on CEEMDAN-BiLSTM

Xin Yang

School of Digital Economics, Hubei University of Automotive Technology, Shiyan, Hubei, China

Abstract: This study proposes a hybrid framework combining CEEMDAN and BiLSTM to predict high-frequency volatility of the SZSE Component Index. CEEMDAN first decomposes the volatility series into multi-scale components, which are then modeled individually by BiLSTM networks. Using 1-minute intraday data from 2014 to 2024, the proposed model significantly improves forecasting accuracy, achieving an R^2 of 0.9325 and reducing MAPE to 19.045%, substantially outperforming standard BiLSTM (R^2 = 0.5365, MAPE = 33.246%). Results demonstrate that signal decomposition effectively enhances the modeling of multi-scale volatility dynamics..

***Keywords*:** CEEMDAN; Bidirectional LSTM; volatility prediction; SZSE Component Index

1. Introduction

Financial market volatility is essential for risk assessment, portfolio allocation, and derivative pricing. The growth of algorithmic trading has increased demand for accurate intraday volatility forecasts. Realized volatility (RV), derived from high-frequency data, provides a more comprehensive measure than traditional low-frequency proxies[1]. In China, the SZSE Component Index exhibits complex volatility patterns that challenge conventional models[2].

Traditional models like ARCH, GARCH, and ARIMA struggle with high-frequency data due to nonlinearities, non-stationarities, and microstructure noise [3][4]. Machine learning methods, including SVR, Random Forest, and LSTM networks, offer improved capabilities for capturing nonlinear patterns [5], but challenges in handling noise and multi-scale features persist [6].

Hybrid models combining signal decomposition (e.g., EMD) with deep learning have emerged to address these issues [7]. However, existing approaches often lack deep integration between decomposition and prediction.

This paper proposes a CEEMDAN-BiLSTM hybrid framework that integrates signal decomposition with deep learning to enhance volatility prediction. Experiments on 1-minute SZSE Component Index data demonstrate its robustness and accuracy.

2. Methodology

2.1 CEEMDAN Method

Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) improves upon EMD and EEMD by introducing adaptive noise at each decomposition stage, effectively mitigating mode mixing while ensuring near-complete noise cancellation and computational efficiency. The algorithm proceeds as follows:

Let $x(t)$ denote the original volatility series. Define $\omega_m(t)$ as independent white noise realizations with variance ϵ_0. For the initial decomposition stage:

$$IMF_1(t) = \frac{1}{M}\sum_{m=1}^{M} E_1\left(x(t) + \varepsilon_0 \omega_m(t)\right) \tag{1}$$

where $E_1(.)$ extracts the first *IMF* via the EMD algorithm, and M represents the ensemble size. The first residual component is computed as: $r_1(t) = x(t) - IMF_1(t)$ .For subsequent stages indexed by n=2,3,…,N:

$$IMF_n(t) = M^{-1}\sum_{m=1}^{M} E_1\left(r_{n-1}(t) + \varepsilon_{n-1} E_{n-1}(\omega_m(t))\right) \tag{2}$$

The decomposition process terminates when no further IMFs can be extracted according to standard stopping criteria, yielding the final residual component:

$$R(t) = x(t) - \sum_{n=1}^{N} IMF_n(t) \tag{3}$$

This adaptive approach produces IMFs with progressively decreasing frequencies, effectively separating high-frequency noise components from lower-frequency components that capture trend and cyclical patterns in the volatility series.

2.2 Bidirectional LSTM Architecture

BiLSTM addresses this fundamental constraint by employing two parallel LSTM layers operating in opposite directions — one processing the sequence forward and the other processing it backward. The forward layer computes hidden states $\vec{h}_t$ based on past information, while the backward layer computes $\overleftarrow{h}_t$ using future information. The final representation at each time step concatenates both directional outputs:

$$\vec{h}_t = LSTM_{\text{forward}}\left(x_t, \vec{h}_{t-1}\right) \tag{4}$$

$$\overleftarrow{h}_t = LSTM_{\text{backward}}\left(x_t, \vec{h}_{t-1}\right) \tag{5}$$

$$y_t = \sigma\left(W\left[\vec{h}_t, \overleftarrow{h}_t\right] + b\right) \tag{6}$$

where W denotes the weight matrix, b represents the bias term, and σ indicates the activation function. This bidirectional architecture enables the model to capture dependencies that extend in both temporal directions, proving particularly valuable for financial time series where patterns may exhibit complex lead-lag relationships across different time scales.

3. CEEMDAN-BiLSTM Hybrid Framework

The proposed hybrid forecasting framework operates through two sequential stages designed to address the multi-scale nature of financial volatility.

Stage 1: Decomposition into Multi-scale Components

The volatility series is decomposed using CEEMDAN, which mitigates EMD mode mixing through adaptive noise. This separates the series into distinct frequency scales—from high-frequency noise to low-frequency trends—along with a residual term, effectively disentangling heterogeneous volatility components.

Stage 2: Component-wise BiLSTM Modeling

Each IMF and the residual term are independently modeled using separate BiLSTM networks. This component-wise approach accommodates the distinct temporal dependency structures of different frequency bands. For each component $c \in \{\text{IMF1},\ldots,\text{IMFN},\text{R}\}$

$$\hat{c}_t = BiLSTM_c\left(c_{t-p}, c_{t-p+1}, \ldots, c_{t-1}; \theta_c\right) \tag{7}$$

where p denotes the lookback window length and θ_c represents component-specific parameters

4. Empirical Analysis

4.1 Data Preprocessing and Descriptive Statistical Analysis

The empirical analysis utilizes 1-minute intraday price data for the SZSE Component Index spanning January 2, 2014 to December 31, 2024, comprising 2,676 trading days after excluding holidays and partial trading sessions. Realized volatility for trading day t is computed using standard methodology:

$$r_{t,j} = \ln\left(\frac{P_{t,j+1}}{P_{t,j}}\right), RV(t) = \sum_{j=1}^{n} r_{t,j}^2 \tag{8}$$

Where $P_{t,j}$ corresponds to the index value recorded at the j-th minute within trading day t, and J = 240 denotes the total number of 1-minute intervals comprising a full trading session.

Statistical diagnostics, including KPSS stationarity test, Ljung-Box autocorrelation test, and Jarque-Bera normality test, confirm that the realized volatility series of the SZSE Component Index exhibits non-stationarity, significant serial

dependence, and non-normal distribution. These characteristics preclude direct forecasting of the raw series, necessitating CEEMDAN decomposition for multi-scale feature extraction. The test results are summarized in Table 1.

Table 1. Volatility Series Statistical Test Results for SZSE Component Index

Index	KPSS	L-B (15)	J-B
SZSE Component Index	23.81^{***}	22381^{***}	13846^{***}

Note: *** denotes statistical significance at the 1% level.

4.2 Comparative Model Performance Evaluation

The predictive capability of the model is assessed using four widely adopted evaluation metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), the Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE). These metrics provide various perspectives on the model's forecasting accuracy.

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}, MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{9}$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right|, R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \tag{10}$$

Here y_i refers to the actual value of the i-th data point, $\hat{y}_i$ represents the predicted value for the same point, and $\overline{y}$ is the mean of the observed data.

To evaluate the proposed CEEMDAN-BiLSTM (C-BiLSTM) model, comparative experiments are conducted against six benchmarks: LSTM, GRU, BiLSTM, and their decomposition-based variants C-LSTM and C-GRU. Results are presented in Table 2.

Table 2. SZSE Prediction Performance Comparison Across Models

SZSE Component Index	R^2	RMSE(10^{-5})	MAE(10^{-5})	MAPE
GRU	0.5142	6.2234	2.7987	45.2345
LSTM	0.5256	6.1543	2.6934	42.1123
BiLSTM	0.5365	6.0876	2.4876	33.2456
C-GRU	0.9089	2.7034	1.4234	20.756
C-LSTM	0.9067	2.7389	1.5765	23.189
C-BiLSTM	0.9325	2.3245	1.2876	19.045

Table 2 reveals three key findings. First, models without decomposition perform poorly ($R^2 \approx 0.51$–0.54, MAPE > 33%). Second, CEEMDAN decomposition substantially improves performance, with decomposition-based models achieving R^2 values of 0.9067–0.9325—a 70–75% improvement. Third, among these, C-BiLSTM outperforms C-GRU and C-LSTM, reducing RMSE by up to 15.1% and MAPE by up to 4.14 percentage points, confirming the value of bidirectional processing.

5. Conclusion

This paper proposes a CEEMDAN-BiLSTM hybrid framework for SZSE Component Index volatility prediction. Using 11 years of 1-minute data, empirical results show: (1) CEEMDAN decomposition improves performance by isolating multi-scale volatility components; (2) C-BiLSTM outperforms C-LSTM and C-GRU, achieving $R^2 = 0.9325$ and MAPE = 19.045%, demonstrating the value of bidirectional processing. These findings have practical implications for risk management, offering a robust foundation for financial decision-making.

References

[1] Zhang, Y., Peng, Y. & Song, Y. Realized Volatility Forecasting for Stocks and Futures Indices with Rolling CEEMDAN and Machine Learning Models. Comput Econ 66, 1215–1268 (2025).

[2] Zeng, Q., Zhang, J., & Zhong, J. (2024). China's futures market volatility and sectoral stock market volatility predic-

tion. Energy Economics, 132, 107429.
[3] Agnolucci, P. (2009). Volatility in crude oil futures: A comparison of the predictive ability of GARCH and implied volatility models. Energy Economics, 31(2), 316-321.
[4] Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Systems With Applications, 103, 25-37.
[5] Beniwal, M., Singh, A., & Kumar, N. (2023). Forecasting long-term stock prices of global indices: A forward-validating Genetic Algorithm optimization approach for Support Vector Regression. Applied Soft Computing, 145, 110566.
[6] Shen, J., & Shafiq, M. (2025). STL-ELM: A computationally efficient hybrid approach for predicting high volatility stock market. Scientific African, 28, e02590.
[7] Li, H., Mei, Y., Hao, X., & Chen, Z. (2024). Out-of-sample equity premium predictability: An EMD-denoising based model. Pacific-Basin Finance Journal, 88, 102536.