Effect of Multi-Scale Decomposition on Performance of Neural Networks in Short-Term Traffic Flow Prediction

HAICHAO HUANG, JINGYA CHEN, XINTING HUO, YUEFI QIAO, AND LEI MA
College of Civil and Transportation Engineering, Hohai University, Nanjing 210013, China
Corresponding author: Haichao Huang (18852896930@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 52078190, and in part by the Humanity and Social Science Youth Foundation of Ministry of Education of China under Grant 18YJAZH119.

Abstract
Numerous studies employ multi-scale decomposition to improve the prediction performance of neural networks, but the grounds for selecting the decomposition algorithm are not explained, and the effects of decomposition algorithms on other performance of neural networks are also lacking further study. This paper studies the influence of commonly used multi-scale decomposition algorithms including EMD (Empirical Mode Decomposition), EEMD (Ensemble Empirical Mode Decomposition), CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise), VMD (Variational Mode Decomposition), WD (Wavelet Decomposition), and WPD (Wavelet Packet Decomposition) on the performance of Neural Networks. Decomposition algorithms are adopted to decompose traffic flow data into component signals, and then K-means is used to cluster component signals into volatility components, periodic components, and residual components. A Bi-directional LSTM (BiLSTM) neural network is adopted as the standard model for training and forecasting. Finally, three metrics, including prediction performance, robustness, and generalization performance are proposed to evaluate the influence of the multi-scale decomposition algorithm for neural networks comprehensively. By comparing the evaluation results of different hybrid models, this study provides some useful suggestions on proper multi-scale decomposition algorithm selection in short-time traffic flow prediction.

Index Terms
Multi-scale decomposition, short-term traffic flow, K-means, robustness, generalization performance.

I. INTRODUCTION
With the acceleration of urbanization, the car parc is increasing year by year. The resulting traffic congestion and unequal distribution of the right of way seriously affect the safety and efficiency of travel. Accurate short-term traffic prediction is the core problem of ITS (intelligent transportation system), which can provide technical support for the surveillance and the forewarning of traffic flow [1], so as to alleviate congestion. In order to achieve the high-precision prediction of short-term traffic flow, a large number of prediction models have been proposed, which can be mainly divided into three categories: parametric model (statistical theoretical model), non-parametric model (machine learning theoretical model), and hybrid model.

Parametric models such as Kalman filtering [2], [3] is particularly good at dealing with noisy traffic flow prediction, Grey forecasting model [4], [5] can achieve high prediction accuracy even when the number of samples is relatively small, autoregressive integrated moving average (ARIMA) model [6] realize traffic flow prediction by stabilizing data through difference. Furthermore, its improved model seasonal autoregressive integrated moving average (SARIMA) [7] improves the ability to extract traffic flow periodic features by adding seasonal factors and further improves the prediction performance. However, parametric models are often applicable to linear data, and the effect of predicting the complex nonlinear traffic flow is often not satisfactory. Especially when dealing with large-scale data at the same time, [8] the efficiency of parametric models is very low, which cannot meet the timeliness requirements of short-term traffic flow prediction.
Non-parametric models have become the mainstream short-term traffic flow forecasting models due to their strong nonlinear fitting ability and self-learning characteristics. Decision tree models, including Random Forest (RF) [9], [10] and Gradient Boosting Decision Tree (XGBoost) [11], are typical non-parametric models that are not prone to over-fit and have a high anti-interference capacity. Support vector machine (SVM) [12]–[14] is widely used in short-time traffic flow prediction because of its strong nonlinear convergence performance and strict mathematical interpretation. K-Nearest Neighbors (KNN) [15], [16] with a short training period is also a commonly used prediction model. The Artificial Neural Network (ANN) [17], [18] is known for its adjustable structure and powerful generalization performance. As the deep learning theory is proposed, RNN [19] and its improved models GRU [20] and LSTM are widely adopted in short-term traffic flow prediction [1], [21], [22].

Tseleltis et al. [23] uncovered that each prediction model has a specific application range. In predicting short-term traffic flow, the combined model is more accurate and more applicable than a stand-alone model. Tang et al. [24] combined double exponential smoothing and SVR to verify its advantages over a stand-alone model. Saigun et al. [25] combined ARIMA and LSTM to train the linear and nonlinear characteristics of traffic flow respectively. Linjiang et al. [26] proposed a dynamic spatial-temporal feature optimization method for big data based on GBDT.

Since the traffic flow is easily disturbed by various factors such as weather and traffic detectors during the acquisition process [27], the collected data often contains a lot of noise. At present, numerous studies have shown that the decomposition algorithm can reduce the impact of noise on the prediction model and improve prediction accuracy. One decomposition method is to assume that the traffic flow has a periodic part and can be modeled by trigonometric regression. Then the residual part is obtained by subtracting the periodic part from the original traffic flow to achieve the traffic flow decomposition. Xiaoxue et al. [28] adopted the combination of sine and cosine to model the periodic part, thus decomposing the traffic flow into the periodic part and the residual part. Indicating the hybrid prediction approach is more effective. Reference [22] uses Fourier transform to separate traffic flow into periodic part and volatility part. It was proved that the prediction accuracy of the hybrid model after decomposing is better than that of the traditional model. Zhang et al. [29] introduced variance and assumed that the traffic flow can be decomposed into three parts. Another way is multi-scale decomposition which migrated from the field of signal processing, the decomposition does not require assuming the period T of the function, and is completely determined by the characteristics of the data itself. Ghosh et al. [30] adopted WD to reduce the sample size required for training. Yan et al. [31] introduced Maximal Overlap Discrete Wavelet Transform to decompose time series, and verified that it is superior to the traditional deep learning mechanism. Mehmet et al. [32] concluded that WD has higher accuracy than EMD with correct wavelet type selection. Jiang and Adeli [33] adopted WPD to reduce the noise of traffic data effectively. Cheng et al. [34] proved that WPD can effectively separate historical data. Wang et al. [35] demonstrated hybrid EMD–ARIMA model outperforms the traditional forecasting models in different scenarios. Qiu et al. [36] proposed a hybrid model combining EMD and deep learning method, and proved the superiority by comparing different prediction methods. Tang et al. [37] compared five denoising schemes and proposed that EEMD is superior to other algorithms. Reference [27] also indicated that EEMD can reduce the prediction error. Yanan et al. [38] proposed hybrid models based on deep learning methods and CEEMDAN have great potential in the field of traffic flow prediction. VMD is the state-of-the-art decomposition algorithm, which is seldom employed in the field of short-time traffic flow prediction. However, it is widely utilized in the fields of wind speed [39], runoff [40], chemistry [41], power load [42], and so on.

Currently, a large number of studies show that the multi-scale decomposition algorithm can reduce the influence of noise on the prediction model and improve the accuracy of the prediction. But neither of these studies explain why this decomposition algorithm is chosen over the others, nor does it say when it should be utilized. Reference [37] compares five algorithms, but their purpose is to test the denoising performance and the latest CEEMDAN and VMD are not involved. At the same time, plenty of research concentrates on multi-scale decomposition to improve the prediction accuracy of neural networks, but lacks further research on the other important properties of neural networks, such as robustness and generalization performance.

The rest of the paper is organized as follows. The second part introduces the multi-scale decomposition algorithms briefly. The third part takes the data as an example of empirical analysis. Finally, the fourth part gives the conclusions and suggestions.

II. METHODOLOGY

A. WAVELET DECOMPOSITION AND WAVELET PACKET DECOMPOSITION

Wavelet analysis can decompose signals at multiple scales by translating and scaling the wavelet basis function to obtain signal components at different scales, which overcomes the shortcoming of short-time Fourier transform in single resolution. Wavelet packet decomposition is a more refined signal decomposition method based on orthogonal wavelet analysis. In essence, the wavelet packet decomposition is the further decompose of high-frequency signal obtained by wavelet decomposition. There is neither redundancy nor omission in wavelet packet decomposition, which improves the capability of time-frequency analysis of the signal spectrum. Although wavelet packet decomposition can fully decompose the high-frequency and low-frequency parts of the data, it is generally considered that the high-frequency part of
the data contains a lot of noise. Therefore, the selection of the multi-scale decomposition algorithm differs in different studies [32], [34].

In the multi-scale analysis, if the scale function and the wavelet basis function satisfy the two-scale function equation, it can be set as:

\[
\begin{align*}
    u(t) &= \sum_{k} H(k) \times u(2t - k) \\
    \varphi(t) &= \sum_{k} G(k) \times u(2t - k)
\end{align*}
\]  

(1)

where \( H(k) \) and \( G(k) \) are Gaussian filter coefficients, \( j \) is decomposition scale, \( k \) is the position index and \( t \) is time.

Then Mallat decomposition algorithm can be expressed as:

\[
\begin{align*}
    cD_{j} &= \sum_{k} H(2t - k) \times cD_{j-1} \\
    cA_{j} &= \sum_{k} G(2t - k) \times cD_{j-1}
\end{align*}
\]  

(2)

where \( cA_{j} \) is the low-frequency approximate component coefficient of the signal in the \( j \)-th layer and \( cD_{j} \) is the high-frequency detail component coefficient of the signal in the \( j \)-th layer.

On this basis, the obtained low-frequency component coefficient \( cD_{j} \) is put into Eq. (2), which is the multi-layer wavelet decomposition. If the high-frequency component coefficient \( cA_{j} \) is decomposed into Eq. (2) at the same time, it is the multi-layer wavelet packet decomposition. The decomposition structure is shown in FIGURE 1.

Its reconstruction can be expressed as:

\[
s(t) = 2 \left[ \sum_{k} h(t - 2k) cA_{j+1} + \sum_{k} g(t - 2k) cD_{j+1} \right]
\]  

(3)

where \( s(t) \) is the original signal, \( h \) and \( g \) are wavelet reconstruction coefficients, and their relationship with Gaussian filter coefficients can be expressed as:

\[
\begin{align*}
    H(k) &= h(-k) \\
    G(k) &= g(-k) \\
    g(k) &= (-1)^{1-k} h(1-k)
\end{align*}
\]  

(4)

\section*{B. EMPIRICAL MODE DECOMPOSITION AND IMPROVED ALGORITHM}

As an adaptive signal time-frequency processing method, EMD [43] stabilizes the signal to generate a series of data sequences with local characteristics of different time scales. Each sequence is an intrinsic mode function (IMF). Although EMD has an excellent processing effect on nonlinear and non-stationary signals, the inherent defects of mode mixing limit its wide application. However, it is not important whether the modes are mixing or not in short-term traffic flow prediction. EEMD [44] solves the mode mixing by additional white noise. On this basis, CEEMDAN [45] adds adaptive white noise to each decomposition to improve the integrity of EEMD and reduce the reconstruction error. The basic principles of the three algorithms are the same but differ in noise, then the latest CEEMDAN is introduced as an example.

1) Determine all local maximum and minimum points of the original signal \( s(t) \). Cubic spline interpolation is employed to fit all the local maximum and minimum points. The upper and lower envelopes formed by cubic spline difference are denoted as \( u_0(t) \) and \( l_0(t) \) respectively.

2) The mean value of the upper and lower envelope \( m_0(t) = (u_0(t) + l_0(t))/2 \). The first IMF component can be expressed as

\[
IMF_1(t) = s(t) - m_0(t)
\]  

(5)

3) Steps (1) and (2) are the rationale for EMD. The main difference between the three algorithms is that EMD has no additional white noise for the \( s(t) \), while EEMD has additional Gaussian white noise \( w_1(t) \) with zero mean value and homogeneity of variance, and CEEMDAN adds noise coefficient \( \varepsilon_0 \) on this basis. The original signal can be expressed as \( s_1(t) = s(t) + \varepsilon_0 w_1(t) \). Decomposition \( s_1(t) \) by EMD to obtain the first component which was defined as Intrinsic Mode Function(IMF).

\[
\overline{IMF_1}(t) = \frac{1}{l} \sum_{i=1}^{l} IMF_1(t)
\]  

(6)

4) Calculate the first residual signal:

\[
r_1(t) = s(t) - \overline{IMF_1}(t)
\]  

(7)

5) The operator \( E(\cdot) \) is defined as the \( i \)-th modal component generated by the EMD, and the \( IMF_i \) is \( i \)-th modal component generated by CEEMDAN likewise. Decomposition \( r_1(t) + \varepsilon_1 E_1(w_1(t)) \) and obtained the second \( \overline{IMF_2}(t) \).

\[
\overline{IMF_2}(t) = \frac{1}{l} \sum_{i=1}^{l} E_1(r_1(t) + \varepsilon_1 E_1(w_1(t)))
\]  

(8)

6) Repeat steps (4) and (5) until the acquired residual signal can no longer be decomposed. The judgment criterion is that the number of extremal points of the residual signal should not exceed two, and the final residual signal is:

\[
R_m(t) = s(t) - \sum_{j=1}^{n} \overline{IMF_j}(t)
\]  

(9)
where \( m \) is the total modulus of CEEMDAN, and the reconstruction of the original signal \( s(t) \) can be expressed as:

\[
s(t) = \sum_{j=1}^{m} IMF_j(t) + R_m(t)
\]

\[ (10) \]

**C. VARIATIONAL MODE DECOMPOSITION**

VMD [46] is a state-of-the-art time-frequency signal decomposition algorithm with the character of adaptive and completely non-recursive. By constructing a variational model, the modal function \( u_k(t) \) and the corresponding center frequency \( \omega_k \) are alternately and iteratively updated to find the optimal solution of the constrained variational model. Finally, all the components are obtained in this process. VMD overcame the endpoint effect and model mixing of EMD and has a solid mathematics theory foundation. the algorithm can be represented as:

1. Perform Hilbert transform for each \( u_k(t) \) to obtain unilateral spectrum:

\[
\left[ \delta(t) - \frac{j}{\pi t} \right] \ast u_k(t)
\]

\[ (11) \]

where \( \delta(t) \) is the Dirac distribution function and \( \ast \) is convolution operation.

2. The signal of each mode is multiplied by the estimated center frequency \( e^{-j\omega_k t} \), then the spectrum is shifted to the baseband:

\[
\left\{ \left[ \delta(t) + \frac{j}{\pi t} \right] \ast u_k(t) \right\} \cdot e^{-j\omega_k t}
\]

\[ (12) \]

3. Gaussian smoothing estimation is carried out for L2 regularization of demodulation signal gradient to obtain the bandwidth of each mode function, then the constrained variational model can be expressed as:

\[
\sum_{k} \left\{ \sin \left[ \sum_{k} \partial_t \left[ \left[ \delta(t) + \frac{j}{\pi t} \right] \ast u_k(t) \right] \cdot e^{-j\omega_k t} \right] \right\}^2
\]

\[ (13) \]

where \( \partial_t \) solving for the partial derivative and \( k \) is the number of components.

4. Considering the influence of Gaussian noise second penalty factor \( \alpha \) and Lagrangian multiplier \( \lambda \) are introduced to ensure the reconstruction accuracy of the original signal \( s(t) \), and the constrained variational formula is transformed into the unconstrained variational formula:

\[
L(\{u_k\}, \{\omega_k\}, \lambda)
\]

\[
= \alpha \sum_{k} \left\| \partial_t \left[ \left[ \delta(t) + \frac{j}{\pi t} \right] \ast u_k(t) \right] \cdot e^{-j\omega_k t} \right\|^2 \]

\[ (14) \]

\[
+ \left\| s(t) - \sum_{k} u_k(t) \right\|^2 \]

\[ (15) \]

\[
+ \lambda \left\| s(t) - \sum_{k} u_k(t) \right\|^2
\]

\[ (16) \]

\[
\sum_{k} \left\| u_k^{n+1} - u_k^n \right\|^2 = \varepsilon
\]

\[ (17) \]

**D. BI-DIRECTIONAL LONG SHORT-TERM MEMORY**

LSTM is representative of the recurrent neural network. The gating mechanism is introduced to realize the function of spatiotemporal memory and alleviate the problem of gradient disappearance and explosion effectively. The introduced gating mechanism like input gate, forgetting gate, and output gate can control when to forget historical information, which makes LSTM have an excellent effect in solving nonlinear time series problem and is widely applied in short-term traffic flow prediction. Its process can be expressed as:

\[
i_t = \sigma \left( W_{xi} X_t + W_{hi} H_{t-1} + W_{ci} C_{t-1} + b_i \right)
\]

\[ (19) \]

\[
f_t = \sigma \left( W_{xf} X_t + W_{hf} H_{t-1} + W_{cf} C_{t-1} + b_f \right)
\]

\[ (20) \]

\[
c_t = f_t c_{t-1} + i_t \tanh \left( W_{xc} X_t + W_{hc} H_{t-1} + b_c \right)
\]

\[ (21) \]

\[
o_t = \sigma \left( W_{xo} X_t + W_{ho} H_{t-1} + W_{co} C_{t-1} + b_o \right)
\]

\[ (22) \]

\[
H_t = o_t \tanh(C_t)
\]

\[ (23) \]

where \( i_t, f_t, o_t, c_t \) are the output of input gates \( i \), forget gates \( f \), input gates \( o \) and cell state \( c \) at time \( t \) respectively, \( W_{xi}, W_{xf}, W_{xi}, W_{xc} \) are the bias of input \( X_t \) and input gates \( i \), forget gates \( f \), input gates \( o \) and cell state \( c \) respectively; \( W_{hi}, W_{hf}, W_{ho}, W_{hc} \) are the bias of hidden layer output \( H_t \) and corresponding gated units; \( W_{ci}, W_{cf}, W_{co} \) are the bias of cell state output \( C_t \) and corresponding gated units; \( b_i, b_f, b_o, b_c \) is the bias vector; \( \sigma \) is the activation function.

The principle of BiLSTM is the same as that of ordinary LSTM. It integrates a forward and a backward LSTM to
realize bidirectional trained at the same time. Then the correlation between traffic flow is further considered which further improves the global and integrity of feature extraction. The BiLSTM structure is shown in FIGURE 2.

### III. EXPERIMENT

#### A. DATA DESCRIPTIONS AND PREPROCESSING

1) **DATA**

The experimental data are collected from the well-known short-term traffic flow dataset of Caltrans Performance Measurement System (PeMS), range from 2017-05-01 00:00:00 to 2017-06-30 23:59:00. The traffic flow data are aggregated every 5 minutes, so there are 17 568 samples totally after the data cleaning. All the datasets are split into 80% training dataset (14 055 samples) and 20% test dataset (3512 samples).

2) **DECOMPOSITION PARAMETERS**

Reference [37] indicated that the method with Daubechies type achieves higher accuracy on short-term traffic flow with periodicity. In this case, we adopted the Db4 wavelet for decomposition by testing, while WD carried out a 4-layer decomposition to obtain four high-frequency components and a low-frequency component. WPD carried out a 3-layer decomposition to obtain four high-frequency components same as WD. At the same time, the high-frequency components containing noise are further decomposed by WPD to obtain four low-frequency components.

EMD adaptively decomposes traffic flow data to obtain 14 components without preset parameters; The three parameters of EEMD including noise standard deviation, number of realizations, and the maximum number of shifting iterations are set as [0.2 100 1000] respectively, and we obtained 15 components after decomposition. CEEMDAN adopts the same decomposition parameters as EEMD and obtained 15 components ultimately.

3) **TRAINING PARAMETERS**

BiLSTM neural network is adopted to test. The number of iterations is set as 100, and the initial learning rate is set as 0.005. To prevent the learning rate from being too large, the model oscillates back and forth. The learning rate is dynamically adjusted by the attenuation method, and the learning rate decreases by 50% for every 25 iterations. The dropout is set as 0.2 to avoid the overfitting problem. The loss function is the root mean square error (RMSE). The time-step is set as 3 and output layer neurons is 1, namely, we adopted the former 5, 10, 15 minutes of traffic flow to predict the present traffic flow. Considering the efficiency of the models, the number of hidden layer neurons is 64.

#### B. PREDICTION PERFORMANCE EVALUATING

The prediction performance is the most important metric to evaluate the short-term traffic flow prediction model, which directly determines the quality of the model. The predic-
tion performance is usually evaluated by the error between the predicted value and the measured value. Three metrics are adopted to evaluate the prediction performance comprehensively. Average absolute percentage error (MAPE) reflects the relative deviation between the predicted value and the actual value, which can directly measure the prediction results. It is widely utilized to evaluate the quality of prediction models, but it cannot directly reflect the difference between the predicted value and measured value. Root mean square error (RMSE) can directly reflect the absolute difference between the predicted value and measured value and be sensitive to the extremely large or extremely small error, it can effectively make up for the deficiency of MAPE. The determination coefficient ($R^2$) reflects the degree of similarity between the predicted value and measured value. These metrics are calculated by the following formula:

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|$$

(25)

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2}$$

(26)

$$R^2 = 1 - \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2}$$

(27)

where $\hat{y}_t$ is the predicted value, while $y_t$ is the measured value.

**TABLE 2.** Comparison of prediction performance of the multi-scale decomposition hybrid model.

| Models                  | RMSE     | MAPE    | $R^2$  |
|-------------------------|----------|---------|--------|
| BLSTM                   | 14.7490  | 0.0886  | 0.9540 |
| EMD-BLSTM               | 13.2723  | 0.0765  | 0.9627 |
| EEMD-BLSTM              | 14.1258  | 0.0840  | 0.9578 |
| CEEMDAN-BLSTM           | 12.2554  | 0.0710  | 0.9682 |
| VMD-BLSTM               | 10.5369  | 0.0635  | 0.9766 |
| WD-BLSTM                | 6.5376   | 0.0362  | 0.9910 |
| WPD-BLSTM               | 10.7510  | 0.0606  | 0.9756 |

**FIGURE 3.** The clustering results of K-means (a) EMD (b) EEMD (c) CEEMDAN (d) VMD (e) WD (f) WPD.
For the convenience of comparison, the errors in FIGURE 4 have been processed into relative weights. The final predicted values of the three components after reconstruction are listed in TABLE 2, and the combination model with the best effect is marked in bold. Firstly, in the part of three components, for the volatility component, WP has the lowest error and the best fitting effect; EMD and CEEMDAN show similar performance; EEMD is less effective; VMD and WPD have the largest error. For the periodic component and residual component, the decomposition algorithm based on wavelet function is obviously not competent, WD and WPD cannot effectively separate the two components from the traffic flow, resulting in poor K-means clustering results, and much greater error than EMD, EEMD, etc. Relatively speaking, VMD also performs poorly, slightly better than WD and WPD, but not as well as EMD. Especially, the effect of WPD is pretty poor in trend component prediction. Secondly, after the reconstruction of the three components, the prediction error of the volatility component almost determines the total error. Finally, from TABLE 2 we notice that decomposition algorithms can significantly improve the prediction of neural networks range from 0.6232 to 8.2114 in RMSE. WD has the best performance and outperforms EMD and EEMD which is the same as the conclusion in reference [32] and [37] respectively. Surprisingly, the newly proposed VMD shows excellent performance, close to that of WPD and second only to WD, CEEMDAN has a mediocre performance, EMD and EEMD had the worst effects. We ranked decomposition algorithms according to their predictive errors, from best to worst: WD > VMD = WPD > CEEMDAN > EEMD > EMD.

Multi-step prediction of traffic flow on several time horizons allows for a wider range of applications to apply the predictions. To test the influence of multi-scale decomposition on the multi-step prediction of the model, We tested the variation of prediction accuracy for four horizons of 5 minutes, 10 minutes, 15 minutes, and 20 minutes respectively. The RMSE and MAPE are given in FIGURE 5. It can be found that the longer horizon did not deteriorate the prediction accuracy significantly by introducing multi-scale decomposition. Besides, VMD performed well in multi-step prediction, Even when the prediction horizon reaches 20 min (4-step ahead), the error is not much bigger than that of 1-step horizon ahead prediction. It implies that VMD is a relatively stable method for multi-step prediction.

Baseline Model: To verify the improvement of neural network performance by multi-scale decomposition, commonly used prediction models are introduced for comparison, including Long short-term memory(LSTM), Random Forest(RF), and Particle swarm optimization least squares support vector machine(PSO-LSSVM).

FIGURE 6 had been processed the same as FIGURE 4, and we can demonstrate the following conclusions from FIGURE 6. Firstly, the prediction performance of the general LSTM neural network is not different(no more than 1%
MAPE) from that of traditional machine learning like RF and LSSVM. Secondly, LSTM with bi-directional training can significantly improve the prediction accuracy of short-term traffic flow, RMSE reduced by 7.2494, and MAPE reduced by 4.67%. $R^2$ is up 5.63%. Finally, the combination of multi-scale decomposition and neural networks can further reduce the prediction error. Taking the WD-BiLSTM hybrid model as an example, compared with LSTM, its RMSE and MAPE have been reduced by 15.4617 and 9.91% respectively. $R^2$ increased by 9.33%. The effect is remarkable.

### C. ROBUSTNESS EVALUATING

The robustness of the neural network refers to the model’s ability to maintain good performance under certain disturbances, while in the field of short-term traffic flow, the robustness is shown as the ability to resist data noise. In order to evaluate the influence of multi-scale decomposition on the robustness of neural networks, two evaluation methods are designed. One is to simulate the random disturbance of the detector. For this purpose, Gaussian white noise with different signal-noise ratio (SNR) is added. The other is to simulate the random failure of the detector. Therefore, different proportions of traffic flow data are set to 0 randomly.

FIGURE 7 (a) and (b) show similar results that when white Gaussian noise with SNR from $-5$ dB to $-25$ dB is added, the overall error of all decomposition algorithms almost does not increase. Since the small amount of added noise is equivalent to data enhancement, the regularization effect is played. When white noise with SNR of $-5$ dB is added to all prediction models, the errors are reduced to a certain extent. EEMD had the most significant response effect to regularization, followed by CEEMDAN and WPD. The regularization had no obvious effect on WD, which had the best prediction performance originally. At the same time, the prediction performance of the model without the decomposition algorithm is not improved by adding noise, which shows that multi-scale decomposition can improve the sensitivity of neural networks to regularization. When the SNR reaches $-35$ dB and $-45$ dB, the prediction accuracy of the models fall off a cliff and lose the prediction ability. Noise disturbance has the most serious effect on EMD, CEEMDAN, and WD algorithms.

FIGURE 7 (c) and (d) show that for data missing, the early error of all decomposition algorithms increases linearly and then flattens out. It can be noted that the multi-scale decomposition is quite sensitive to data missing. When the missing rate is 5%, the prediction error of half of the combined models is higher than that of the BiLSTM. With the further increase of miss rate, the prediction error of all combination models is gradually higher than that of BiLSTM. When the miss rate reaches 25%, the BiLSTM neural network is the prediction model with the least prediction error. In particular, WD with the best prediction effect is the most sensitive to data missing. When the missing rate reaches 5%, the error doubles by 3 times, and when the missing rate finally reaches 25%, it is
The model with the largest error. However, EEMD performs well in the noise resistance test but has a poor performance in data missing resistance, and its error growth rate is second only to WD. The data missing resistance performance of EMD is still not good, and the error is close to that of WPD. VMD and CEEMDAN show good performance in this test, which is superior to other decomposition algorithms.

D. GENERALIZATION PERFORMANCE EVALUATING

At present, in the study of short-term traffic flow prediction, the data of the former part of the same section is taken as the training set and the latter part as the test set, so as to evaluate the generalization performance of the neural network, and the generalization performance is equivalent to the prediction accuracy. However, under the circumstance of big data, the short-term traffic flows in the same section will show a similar pattern. The trained neural network may perform extremely high accuracy in this section, but it achieves a poor effect in other sections, that is, overfitting occurs. Therefore, the generalization performance of neural networks is not comprehensive enough. Our goal is to obtain a prediction model for the road network. Therefore, 3 of the remaining 23 sections are selected randomly for prediction by using the trained neural network model to comprehensively evaluate the generalization performance of the neural network and the influence of the decomposition algorithm.

FIGURE 8 show that when the models are adopted to predict different sections, the prediction error is significantly increased, and the degree of augmenting is different, which indicates that it is not comprehensive to evaluate the generalization performance of the neural network only by training and simulation of traffic flow data in the same section. At the same time, the multi-scale decomposition greatly improved the generalization performance of the neural networks. When the traditional BiLSTM is applied to other sections, the prediction error increased greatly and
different sections fluctuated greatly. No matter which multi-scale decomposition algorithm is utilized to optimize the neural network, the error fluctuation decreased obviously. WD maintained a good generalization performance in general, and the average MAPE is the minimum, but the prediction error of different sections fluctuated greatly. On the other hand, VMD showed excellent generalization performance and stability. Average MAPE is second only to WD, but the stability is far better than WD. Besides, the stability of CEEMDAN and EEMD is not much different, and the prediction error of CEEMDAN is slightly smaller than that of EEMD. The generalization performance of EMD and WPD is also unsatisfactory, and they are not suitable for road network prediction. VMD and WD are the optimal decomposition algorithms for road network traffic flow prediction.

### E. COMPREHENSIVE COMPARATIVE EVALUATION

For the convenience of comparison, the performance of each model under different tests is ranked and scored. Models that performed well will achieve higher scores. The prediction performance and anti-missing performance are comprehensively evaluated by RMSE and MAPE. In the part of anti-noise performance, the regularization effect is taken as the main standard, and the error from $-35$dB to $-45$dB is used for auxiliary evaluation. In the part of generalization performance, the error fluctuation of the different sections is the main factor, supplemented by the average MAPE.

From the observation of results in FIGURE 9, several interesting findings can be summarized as follows:

1. In most cases, the multi-scale decomposition algorithm can effectively improve the performance of the neural network. Although these algorithms are quite sensitive to missing data. However, when forecasting traffic flow in real situations, interpolation or machine learning is often used to fill in the missing data, so the impact of missing data is relatively light.

2. VMD and WD are excellent multi-scale decomposition with the best comprehensive performance. VMD has shown stable performance in all tests and can deal with various complex traffic flow conditions with wide applicability. Although WD is extremely sensitive to data loss, if the original data is properly preprocessed, it will be the most superior decomposition algorithm.

3. The performance of CEEMDAN and WPD is relatively close in general, and they can be substituted for each other. If the prediction accuracy and anti-noise performance have higher requirements, WPD can be employed; If we focus on the generalization performance of the model, CEEMDAN will be a better choice.

4. The overall performance of EMD and EEMD is mediocre. EMD, in particular, has relatively poor performance in almost all tests, and we can always find a superior alternative algorithm for it. So we do not recommend it to be employed in priority under any circumstances. While performing poorly in other tests, EEMD has unmatched noise resistance and can be considered when raw traffic flow data contains a lot of noise and cannot be effectively separated.

### IV. CONCLUSION

This study assessed the impact of the multi-scale decomposition algorithm for the neural network model based on the combination of BiLSTM and the decomposition algorithm, using the traffic flow data collected from PEMS. We utilize K-means clustering to unify input parameters and achieve dimensionality reduction. Three metrics of prediction performance, robustness, and generalization performance are proposed to comprehensively evaluate the influence of the multi-scale decomposition algorithm. From the observation of the results from this study, several interesting conclusions can be drawn. Firstly, the multi-scale decomposition algorithm is quite effective in improving the performance of the neural network in all aspects. Secondly, VMD is the most widely applicable algorithm, WD is excellent after data preprocessing, and EEMD has unmatched noise resistance. Finally, CEEMDAN can be used to pursue prediction accuracy and anti-noise performance, and WPD can be used to pursue generalization performance, but in general, they are inferior to VMD. EMD is not recommended in any cases.

Comparing the performance of different combination models, this study provides some useful suggestions on proper multi-scale decomposition algorithm section in short-time traffic flow prediction. However, in the current study, we only consider historical traffic flow but do not involve the response of decomposition algorithms to spatial factors such as weather and holidays, which will be the focus of our further research.

### REFERENCES

[1] Q. Tang, M. Yang, and Y. Yang, “ST-LSTM: A deep learning approach combined spatio-temporal features for short-term forecast in rail transit,” J. Adv. Transp., vol. 2019, pp. 1–8, Feb. 2019.

[2] J. W. Gao, Z. W. Leng, B. Zhang, X. Liu, and G. Q. Cai, “The application of adaptive Kalman filter in traffic flow forecasting,” Adv. Mater. Res., vol. 680, pp. 495–500, Apr. 2013.

[3] L. Cai, Z. Zhang, J. Yang, Y. Yu, T. Zhou, and J. Qin, “A noise-immune Kalman filter for short-term traffic flow forecasting,” Phys. A, Stat. Mech. Appl., vol. 536, Dec. 2019, Art. no. 122601.

[4] Z.-X. Wang and P. Hao, “An improved grey multivariable model for predicting industrial energy consumption in China,” Appl. Math. Model., vol. 40, nos. 11–12, pp. 5745–5758, Jun. 2016.
[5] B. Zeng and C. Li, “Improved multi-variable grey forecasting model with a dynamic background-value coefficient and its application,” *Comput. Ind. Eng.*, vol. 118, pp. 278–290, Apr. 2018.

[6] J. Wang, W. Deng, and Y. Gao, “New Bayesian combination method for short-term traffic flow forecasting,” *Transp. Res. C. Emerg. Technol.*, vol. 43, pp. 79–94, Jan. 2014.

[7] X. Zhang, Y. Pang, M. Cui, L. Stallones, and H. Xiang, “Forecasting mortality of road traffic injuries in China using seasonal autoregressive integrated moving average model,” *Ann. Epidemiol.*, vol. 25, no. 2, pp. 101–106, Feb. 2015.

[8] S. V. Kumar, “Traffic flow prediction using Kalman filtering technique,” *Procedia Eng.*, vol. 187, pp. 582–587, 2017.

[9] J. Ou, J. Xia, Y. Li, W. Fan, and W. Rao, “Short-term traffic flow forecasting for urban roads using data-driven feature selection strategy and bias-corrected random forests,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2645, no. 1, pp. 157–167, Jan. 2017.

[10] S. Yassin and Pooja, “Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach,” *Social Netw. Appl. Sci.*, vol. 2, no. 9, pp. 1–3, Sep. 2020.

[11] S. Yang, J. Wu, Y. Du, Y. He, and X. Chen, “Ensemble learning for short-term traffic prediction based on gradient boosting machine,” *J. Sensors*, vol. 2017, pp. 1–15, May 2017.

[12] H. Su, L. Zhang, and S. Yu, “Short-term traffic flow prediction based on incremental support vector regression,” in *Proc. 3rd Int. Conf. Natural Comput. (ICNC)*, Hukou, Hainan, 2007, pp. 640–645.

[13] X. Wang, K. An, L. Tang, and X. Chen, “Short-term prediction of freeway exiting volume based on SVM and KNN,” *Int. J. Transp. Sci. Technol.*, vol. 4, no. 3, pp. 337–352, Sep. 2015.

[14] Y. Xu, D. W. Hu, and B. Su, “Short-term traffic flow prediction based on optimised support vector regression,” *Int. J. Appl. Decis. Sci.*, vol. 10, no. 4, p. 305, 2017.

[15] L. Lin, Y. Li, and A. Sadek, “A k nearest neighbor based local linear wavelet neural network model for on-line short-term traffic volume prediction,” *Procedia-Social Behav. Sci.*, vol. 96, pp. 2066–2077, Nov. 2013.

[16] P. Dell’Acqua, F. Bellotti, R. Berta, and A. De Gloria, “Time-aware multivariate nearest neighbor regression methods for traffic flow prediction,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3393–3402, Dec. 2015, doi: 10.1109/TITS.2015.2453116.

[17] W. Zheng, D.-H. Lee, and Q. Shi, “Short-term freeway traffic flow prediction: Bayesian combined neural network approach,” *J. Transp. Eng.*, vol. 132, no. 2, pp. 114–121, Feb. 2006.

[18] S. Sun, R. Huang, and Y. Gao, “Network-scale traffic modeling and forecasting with graphical lasso and neural networks,” *J. Transp. Eng.*, vol. 138, no. 11, pp. 1358–1367, Nov. 2012.

[19] H. Wigmø, “A model of a neural network with recurrent inhibition,” *Kybernetik*, vol. 16, no. 10, pp. 113–112, 1974.

[20] R. Fu, Z. Zhang, and L. Li, “Using LSTM and GRU neural network methods for traffic flow prediction,” in *Proc. 31st Youth Academic Annu. Conf. Chin. Assoc. Autom. (YAC)*, Nov. 2016, pp. 324–328.

[21] B. Yang, S. Sun, J. Li, X. Lin, and Y. Tian, “Traffic flow prediction using LSTM with feature enhancement,” *Neurocomputing*, vol. 332, pp. 320–327, Mar. 2019.

[22] L. Chen, L. Zheng, J. Yang, D. Xia, and W. Liu, “Short-term traffic flow prediction: From the perspective of traffic flow decomposition,” *Neurocomputing*, vol. 413, pp. 444–456, Nov. 2020.

[23] D. I. Tsiontis, E. I. Vlahogianni, and M. G. Karlaftis, “Improving short-term traffic forecasts: To combine models or not to combine?” *IET Intell. Transp. Syst.*, vol. 9, no. 2, pp. 193–201, Mar. 2015.

[24] J. Tang, H. Wang, Y. Wang, X. Liu, and F. Liu, “Hybrid prediction approach based on weekly similarities of traffic flow for different temporal scales,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2443, no. 1, pp. 21–31, Jan. 2014.

[25] S. Lu, Q. Zhang, G. Chen, and D. Seng, “A combined method for short-term traffic flow prediction based on recurrent neural network,” *Alexandria Eng. J.*, vol. 60, no. 1, pp. 87–94, Feb. 2021.

[26] L. Zheng, J. Yang, L. Chen, D. Sun, and W. Liu, “Dynamic spatial-temporal feature optimization with ERI big data for short-term traffic flow prediction,” *Neurocomputing*, vol. 412, pp. 339–350, Oct. 2020.

[27] Y. Peng and W. Xiang, “Short-term traffic volume prediction using GA-BP based on wavelet denoising and phase space reconstruction,” *Phys. A, Stat. Mech. Appl.*, vol. 549, Jul. 2020, Art. no. 123913.