IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 20, 2020, accepted January 4, 2021, date of publication January 8, 2021, date of current version January 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3050165

A Novel Deep Learning Approach to Predict the Instantaneous NO_x Emissions From Diesel Engine

YANG YU[1,2], YANYAN WANG[1,2], JIAQIANG LI[1,2], MINGLIANG FU[3], ASAD NAEEM SHAH[4], AND CHAO HE[1,2]

[1]School of Machinery and Transportation, Southwest Forestry University, Kunming 650224, China
[2]Key Laboratory of Motor Vehicle Environmental Protection and Safety in Plateau Mountainous Areas of Yunnan Province, Kunming 650224, China
[3]Chinese Research Academy of Environmental Sciences, Beijing 100012, China
[4]Department of Mechanical Engineering, University of Engineering and Technology, Lahore 54000, Pakistan

Corresponding author: Chao He (hehesmile@gmail.com)

This work was supported by the National Natural Science Foundation of China under Grant 51968065, by the Yunnan Provincial Agricultural Joint Special Project under Grant 2017FG001(-010) and Grant 2017FG001(-093), and by the Scientific Research Fund Project of Yunnan Provincial Department of Education under Grant 2020Y0381.

ABSTRACT Accurate and stable prediction of NO_x emissions from diesel vehicles plays a crucial role in the establishment of virtual NO_x sensors and the development and design of diesel engines. This paper presents a method for estimating transient NO_x emissions by complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and a long- and short-term memory neural network (LSTM). First, the CEEMDAN algorithm is used to reduce the non-stationarity and volatility of the transient NO_x emission data to obtain multiple subseries with different frequencies. Secondly, a predictive model is developed for each subsequence using an LSTM neural network. Finally, the results of each subsequence prediction are summed to obtain the final prediction. The proposed model uses NO_x emission data generated by an EU IV diesel bus during real road driving. The results show that (1) The use of CEEMDAN can effectively improve the smoothness of NO_x transient emission data, as well as facilitate more effective extraction of internal characteristics and variations of the raw data. (2) LSTM has better learning and prediction capability for transient changes in NO_x emissions. (3) The results of CEEMDAN-LSTM for RMSE, R^2, MAE and NRMSE are 46.11,0.98, 29.82 and 2.71, respectively, which are better than the other model with improved prediction performance.

INDEX TERMS Diesel engine, NO_x emissions, long short-term memory network, empirical mode decomposition, machine learning.

NOMENCLATURE

NO_x	nitrogen oxide
LSTM	long short-term memory neural network
RES	residual
RMSE	root mean square error
MAE	mean absolute error
PEMS	portable emission measurement system
CFD	computational fluid dynamics
BPNN	back propagation neural network
NSGA-II	non-dominated sequencing genetic algorithm
EGR	exhaust gas recirculation
GPRS	general packet radio service
KNN	k-nearest neighbor
RF	random forest
EEMD	ensemble empirical mode Decomposition
NRMSE	normalized root-mean square Error
ANFIS	adaptive neuro-fuzzy inference system
CEEMDAN	complete ensemble empirical mode decomposition with adaptive noise
IMFs	intrinsic mode functions
EU IV	Euro IV
R2	determination coefficient
MLP	multi-layer perceptron
NO	nitric oxide
OBD	on board diagnostics
ANN	artificial neural network
DNN	deep neural networks
SVM	support vector machines
XGBoost	extreme gradient boosting
CART	categorical regression tree

The associate editor coordinating the review of this manuscript and approving it for publication was Bin Liu.

EMD	empirical mode decomposition
RNN	recurrent neural networks
GBRT	gradient boosted regression tree

I. INTRODUCTION

Diesel engines offer high fuel economy and thermal efficiency and are widely used in heavy vehicles and Non-road machinery. At the same time, the NO_x emissions from diesel engines, due to their higher compression ratio and high combustion temperatures in the cylinder, are much higher than gasoline engines. Excessive NO_x emissions severely affect the environment and human health, which motivates the researchers to conduct critical studies on NO_x emissions from diesel engine and design of the diesel engine. On the other hand, the calibration of NO_x emissions put forward higher requirements, therefore, the study of instantaneous NO_x emissions from diesel engines and the establishment of accurate emission model, plays an important role in the design and development stage of automotive powertrain.

There are significant differences between transient and steady-state operating conditions of the engine [1]. Steady-state data are usually obtained from standard driving cycle tests, using engine chassis-dynamometer platforms to obtain the pollutant emission data, and the results of these tests do not reflect changes in the emissions under real driving conditions, especially when the engine is overloaded from part load to full load. Transient data can be collected by a portable emission measurement system (PEMS) to investigate the emissions from vehicles during actual road operation [2], and such data encompasses the changes in the emissions due to transient changes in the vehicle operating conditions. Although transient data can reflect the changes in vehicle emissions during actual road operation, yet such data are non-stationary and non-linear in nature, making it difficult to develop accurate predictive models.

To study the NO_x emissions, various authors have proposed the models based on physical and chemical relationships such as a previously conducted study, in which Egnell [3] developed a combustion diagnostic method to calculate the nitric oxide (NO) formed during combustion through multi-zone heat release. Provataris *et al.* [4], in an experimental study, investigated the formation of NO_x emissions in the combustion chamber based on the data obtained from the engine operation and combustion analysis. Guardiola *et al.* [5] adopts a semi-empirical model based on heat release rate and the adiabatic flame temperature with NO_x correction considering the combustion mechanism and shows good transient NO_x prediction performance, but the time to complete the prediction is long. Bhave *et al.* [6] implemented a stochastic reactor model, based on a probabilistic density function, to describe the turbulent mixing and further, used a rotational clustering model to predict the range of NO_x emissions. Rosero *et al.* [7] created a grid engine diagram to obtain the engine data from the OBD system and measured the NO_x data by PEMS to reproduce the results in terms of fuel consumption and NO_x emissions. Reitz and Rutland [8] improved the prevailing computational fluid dynamics (CFD) models for multicomponent fuel vaporization and small-flame combustion to accurately predict the trends in NO_x emissions and NO_x-soot balance. Moreover, a number of researchers have also made accurate predictions of NO_x emissions through CFD models [9]–[11]. Although these models are capable of estimating NO_x by calculating the local temperature and chemical composition, however, certain drawbacks related to these methods such as high computational cost, high computational time, increased computational power and high calibration requirements, hinder the application of these models in practice.

In the last few years, machine learning algorithms such as artificial neural network (ANN), support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost), and deep neural network (DNN) have developed rapidly. Compared to physical and chemical relationship models, machine learning can quickly establish non-linear relationships between diesel vehicle operating parameters and emissions, and has been widely used in the engineering field [12]. Wang *et al.* [13] estimated NO_x emissions using a back propagation neural network (BPNN) by mutual information reducing the need for experts in feature selection. Alcan *et al.* [14] used the most significant engine characteristic parameters for nonlinear prediction and reduced the secondary characteristic variables to obtain a more stable model for NO_x emission prediction. Bhowmik *et al.* [15] analyzed the results of error and correlation matrices and the statistical analysis of artificial neural network(ANN) to predict the NO_x emissions with ternary blended fuels with good accuracy. Lotfan *et al.* [16] investigated engine speed, intake temperature, and output power as variable inputs to obtain optimal values for NO_x emissions using artificial neural network and non-dominated sequencing genetic algorithm (NSGA-II). Arsie *et al.* [17] introduced a virtual sensor, developed based on neural networks, for estimating NO_x emissions from diesel locomotives, with good adaptability achieved on diesel engine test bench. Wang *et al.* [18] proposed a method by combining neural networks and nonlinear dynamic systems to replace NO_x sensors, but due to the limitations of artificial neural networks, a room was available for improvement in accuracy. In the previously conducted studies, few scholars have used deep neural networks to develop diesel engine NO_x emission prediction models. Moreover, most of the above-mentioned studies was obtained from the experiments performed on the engine test benches under steady-state load conditions, which cannot be applied to study the real road conditions.

Signal processing algorithms can extract features of transient data more efficiently, and have been widely used in conjunction with machine learning in areas such as stock prices, power loads, and wind speed prediction. Cao *et al.* [19] developed a stock market price prediction model using CEEMDAN combined with long- and short-term memory (LSTM) for linear regression analysis of major global stock market indices and outperformed SVM, multilayer

perceptual machine (MLP), and other hybrid models. Li *et al.* [20] combined the improved CEEMDAN with a Multi-Kernel Extreme Learning Machine (MKELM) to optimize the weights and parameters of each kernel in the extreme learning machine using Gray Wolf Optimization (GWO) and showed excellent prediction performance for short-term power load prediction. Lu *et al.* [21] used CEEMDAN pre-processing of raw wind data to make the data smoother, combined with an improved grey box model to predict wind energy, which made accurate predictions of Belgian offshore wind farms and had higher prediction accuracy and stability than six benchmark models. Du *et al.* [22] used CEEMDAN to decompose the raw wind speed data into a finite set of components, and then used multi-objective ant lion optimization to optimize the initial weights between layers and thresholds in the Elman neural network, and the prediction accuracy was much higher than that of the comparative model. CEEMDAN can effectively enhance the stability of transient data. However, the signal processing method has not yet been applied to diesel engine NO_x transient emission prediction. In this research, a model combining CEEMDAN and LSTM for predicting NO_x emissions from diesel engines is proposed to effectively address the non-stationarity and non-linearity of real road emission data. The contributions are as follows:

1) Application of the decomposition algorithm CEEMDAN to the instantaneous prediction of diesel NO_x to reduce the non-stationarity and complexity of the raw data.
2) LSTM algorithm has more stable and accurate performance in time series prediction compared to traditional machine learning algorithm, which can be used for sub-series prediction of CEEMDAN.
3) The predictive performance of the proposed CEEMDAN-LSTM model has been demonstrated in an experimental case, using seven models to verify accuracy and validity.
4) A machine learning algorithm for filling missing values is proposed for time series data.

The paper is structured as follows: section 2 describes the methodology used to build the model. Section III describes the sources of data and how they were processed. Section IV discusses and analyzes the experimental results. Section V summarizes the conclusions.

II. RESEARCH METHODOLOGY

A. COMPLETE ENSEMBLE EMPIRICAL MODE DECOMPOSITION WITH ADAPTIVE NOISE

CEEMDAN is essentially an improvement on the deficiencies of EMD and EEMD, as it addresses the modal overlap defect of EMD [23] and the incomplete white noise cancellation of EEMD additions. CEEMDAN adds adaptive white noise to each decomposition in order to achieve the decomposition of the original sequence with fewer averaging times and a reconstruction error of almost zero.

Suppose the i-th decomposition sequence is denoted as $s^i(n) = s(n) + v^i(n)$, where $s^i(n)$ denotes the original sequence and $v^i(n)$ denotes the white noise added during the i-th (1,2…, I) decomposition. Assuming that the kth modal component by EEMD and CEEMDAN is denoted by $E_k(\cdot)$ and IMF_k, respectively, the specific steps of the CEEMDAN algorithm are as follows;

Step 1: For the time series $s^i(n)$, the I decomposition is performed using the EEMD method to obtain the 1st modal component as shown in Eq. (1).

$$IMF_1(n) = \frac{1}{I}\sum_{i=1}^{I} IMF_1^i(n) \tag{1}$$

Step 2: Calculation of the residual sequence after decomposition by CEEMDAN can be written as Eq. (2):

$$r_1(n) = s(n) - IMF_1(n) \tag{2}$$

Step 3: The ith (1,2,…, I) decomposition of $r_1(n) + \varepsilon_1 E_1\left[v^i(n)\right]$ yields the second modal component as described in Eq. (3).

$$IMF_2(n) = \frac{1}{I}\sum_{i=1}^{I} E_1[r_1(n)] + \varepsilon_1 E_1[v^i(n)] \tag{3}$$

where ε_k denotes the noise amplitude.

Step 4: After, all the $IMF_k(n)$ are obtained, calculate the $k(2, 3\ldots K)$ sequence of residual quantities, and then calculate the $k+1$ modal components according to Step 3, as depicted by Eq. (4) and (5):

$$r_k(n) = r_{k-1}(n) - IMF_k(n) \tag{4}$$

$$IMF_{k+1}(n) = \frac{1}{I}\sum_{i=1}^{I} E_1\{r_k(n) + \varepsilon_k E_k[v^i(n)]\} \tag{5}$$

Step 5: Repeat step 4, when the sequence of residuals is no longer decomposable, i.e., when the extreme point of the residuals is at most 2, stop the decomposition, and obtain the k modal components. The final result of the residuals can be evaluated by Eq. (6):

$$R(n) = s(n) - \sum_{k=1}^{K} IMF_k \tag{6}$$

At this point the sequence is decomposed into k *IMF* and an $R(n)$.

B. LONG SHORT-TERM MEMORY

Long short-term memory can be known as a variety of recurrent neural networks (RNN), which cannot handle long time spans. As the time-series span increases, the learning ability of the RNNs will be lost, resulting in the models, that cannot be trained [24]. Therefore, LSTM adds a state c to the hidden layer h of the original RNN to save the information about the state of the historical time and the current state, as shown in Fig. 1(a). Expanding Fig. 1(a) by time dimension as shown in Fig. 1(b), at time t, the input of LSTM is increased from 1 to 3 from RNN, including the current network input value x_t, the output value of LSTM at the previous moment h_{t-1} and the cell state c_{t-1} at the previous moment, while the output value of LSTM cell includes the current moment h_t and the cell state c_t at the current moment. The state c that saves the history information is updated on the cell state by the

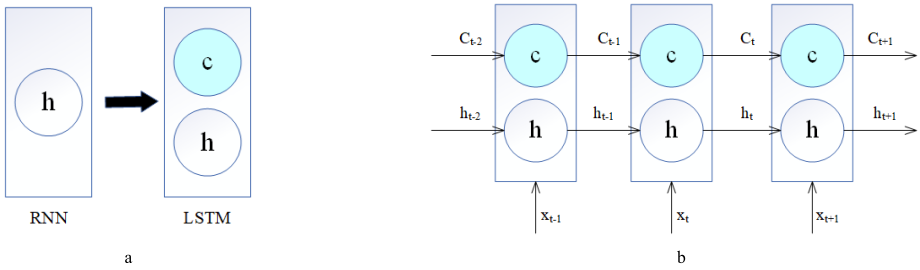


FIGURE 1. RNN to LSTM. (a) RNN change to LSTM. (b)LSTM unfold.

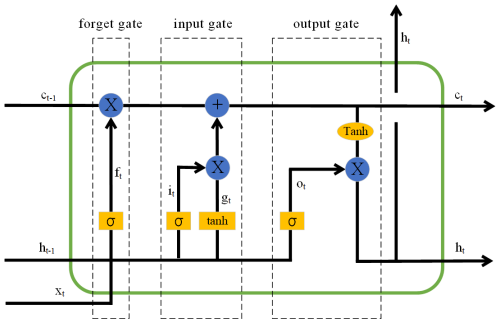


FIGURE 2. Internal unit structure of LSTM.

control operation of three gates (input, forgetting, and output gates) [25]. The internal cell structure of LSTM is shown in Fig. 2.

The LSTM neural network can be calculated by Eq. (7-11) as follows:

$$i_t = \sigma(W_{xi}x_i + W_{hi}h_{i-1} + W_{ci}c_{i-1} + b_i) \quad (7)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (8)$$

$$c_t = f_t h_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (9)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}hh_{t-1} + W_{co}c_t + b_o) \quad (10)$$

$$ht = o_t tanh(c_t) \quad (11)$$

where, x_t is the input; f_t, i_t, o_t are the forgetting gate, input gate and output gate, respectively; c_t is the cell unit state after time t passing through the input gate and forgetting gate; h_t are all output states of LSTM unit; tanh is the activation function of double tangent; δ is sigmoid function. W_{xi}, W_{hi}, W_{ci}, W_{xf}, W_{hf}, W_{cf}, W_{xc}, W_{hc}, W_{xo}, W_{ho}, and W_{co} are the input weight coefficients. b_i, b_f, b_c and b_o is the offset vector.

The LSTM can selectively retain valid information based on the state of the cell unit and the structure of three gates: the input gate, the forgetting gate and the output gate. The input gate handles the input of the current sequence position and determines the input information, whereas, the forgetting gate determines the discarding of historical information based on the activation function. The output gate determines the final output and the retention of information.

C. DEVELOPMENT OF THE CEEMDAN-LSTM NO_X PREDICTION MODEL

The data series of diesel engine NO_x emission concentration is a non-stationary signal, that can be affected by many

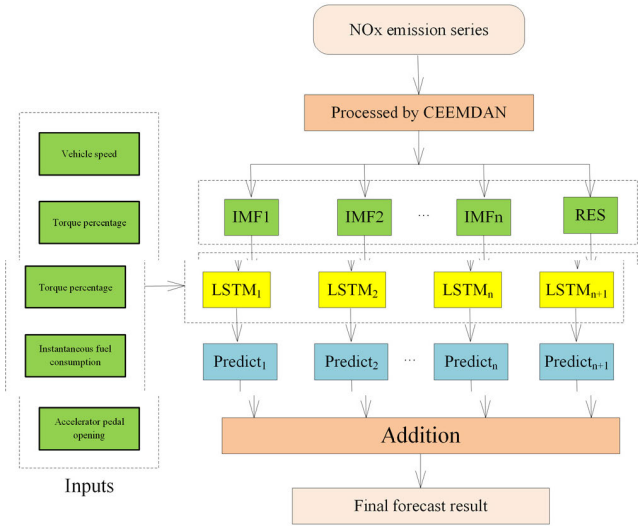


FIGURE 3. Prediction flow chart based on CEEMDAN-LSTM model.

factors. Since CEEMDAN possesses a capability to smoothen the non-stationary time series, so in the current effort CEEMDAN and LSTM were combined to establish a prediction model based on CEEMDAN-LSTM to predict diesel engine NO_x emission concentration. The flowchart for establishing the CEEMDAN-LSTM prediction model is shown in Fig. 3. The major steps are described as follows:

1) CEEMDAN decomposes the NO_x concentration series data by generating multiple IMFs and a RES. At the same time, it decomposes the non-stationary, non-linear NO_x concentration series into data components of different simultaneous frequencies.
2) LSTM modeling is carried out by using the sub-series generated by CEEMDAN. The measurements of the parameters such as engine speed, accelerator pedal opening, instantaneous fuel consumption, percentage of engine load and speed were obtained from the vehicle exhaust monitoring platform and were used as input variables. Whereas, the sub-series decomposed by the CEEMDAN were considered to be output to build n LSTM models. Of the data, 90% were used as training data and 10% were used to validate model accuracy. Adjusting the model parameters not only allows the LSTM to learn more about the variation pattern of

TABLE 1. Basic parameters of diesel engine.

Parameters	Value
Type	6-cylinder, in-line, water-cooled
Air intake system	Turbocharged inter-cooled
Fuel system	Electric control high-pressure common rail system
Cylinder diameter (mm)	124
Stroke (mm)	107
Displacement (L)	6.7
Rated power (kW)	155
Rated Power speed (r/min)	2500
Maximum torque (Nm)	800
Maximum torque-speed (r/min)	1400
Compression ratio	(17.3±0.3):1
Emission standard	EU IV

each series but also improves the accuracy of the NO_x concentration prediction.

3) The predictions obtained from each LSTM model in the previous step are added together to obtain the final results of the CEEMDAN-LSTM model in order to predict the concentrations of NO_x emission.

III. DATA SOURCES AND PROCESSING METHODS

A. DATA SOURCES

In the current study, the data was derived from an online motor vehicle exhaust monitoring platform [26], which was employed to monitor dynamic real-time vehicle emission data using general packet radio service (GPRS) technology. It consisted of NO_x sensor, an OBD data reader and a wireless data transmission unit. The NO_x sensor is a high-precision sensor that can accurately measure NO_x emissions of diesel vehicles per second, the OBD port obtained vehicle and engine information, and the collected data was transmitted to the terminal platform through the wireless data transmission unit. The online monitoring platform collected and managed the data including exhaust gas monitoring parameters, basic vehicle information with strict management and control. The parameters of the diesel engine of a bus were monitored as shown in Table 1. With a sampling interval of 1s, the 10,000 transient emission data were obtained from the online monitoring platform. The selected samples included the data in idling, acceleration, constant speed, deceleration and stopping conditions of the bus.

B. DATA PRE-PROCESSING

Tail gas data pre-processing was employed to eliminate the errors related to missing and abnormal data caused by equipment failure, human error or sensor abnormality during the data collection. Moreover, it helped to make the data more standardized and complete, which is convenient and inevitable for the research and development of diesel engine exhaust emission model.

1) OUTLIER HANDLING

In the recent study, a box plot was used to determine the dispersion of the data distribution, and to identify, whether the data was outlier based on the degree of aggregation, or not [27]. As shown in the box diagram in Fig. 4a, the data distribution was evaluated in terms of the minimum, first quartile (Q1), median (Q2), third quartile (Q3), maximum values of the data set and Q3-Q1=IQR (defined as the quartile distance). The $1.5\times$IQR was used to be the outlier judgment boundary to remove the data less than Q1-$1.5\times$IQR and greater than Q3+$1.5\times$IQR, leaving the null values in the data. Fig.4b shows that the outliers are mainly part of the data out of the actual value range and are concentrated between 2000-3200, which is due to occasional anomalies in the NO_x sensor causing data acquisition errors.

2) MISSING VALUE TREATMENT

Mostly, due to the measurement errors or equipment failures, the missing or incomplete data can negatively affect the analysis of emission data and the results of predictive models. Researchers have proposed various interpolation methods and demonstrated their applications in practice, such as mean interpolation, k-nearest neighbor (KNN) interpolation, hot card interpolation, multiple interpolation and others, however, these interpolation methods only consider a single attribute. A machine learning prediction model, based on extreme gradient boosting (XGBoost) [28], was used to fill in the missing values for the multi-attribute diesel engine emission data. Considering the correlation between the features of the experimental samples, the XGBoost model was generated based on the training of the complete dataset, and the training model was used to predict the filling specialties of the missing data. The method was insensitive to the input data, which can effectively prevent the overfitting of the data. Furthermore, it helped to reduce the influence of undetected anomalies in the data on the filling of missing values, and effectively improve the accuracy of the filled data.

XGBoost is an integrated learning algorithm, that works on the basic idea to integrate several weak classifiers together to form a strong classifier with a categorical regression tree (CART) that is capable of splitting the category and continuous variables [29]. XGBoost can automatically utilize multiple threads of the CPU for parallel operations, while improving the algorithm with increased accuracy.

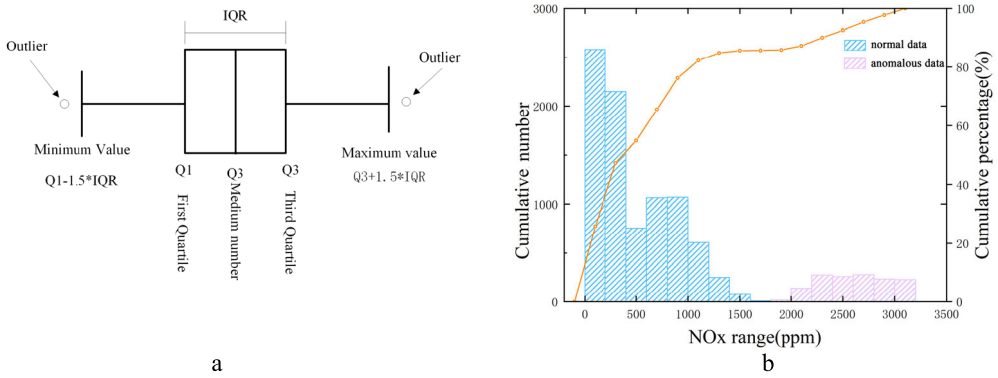


FIGURE 4. Outlier judgment charts. (a) Box diagram to judge the distribution of outliers. (b) Outlier frequency statistics.

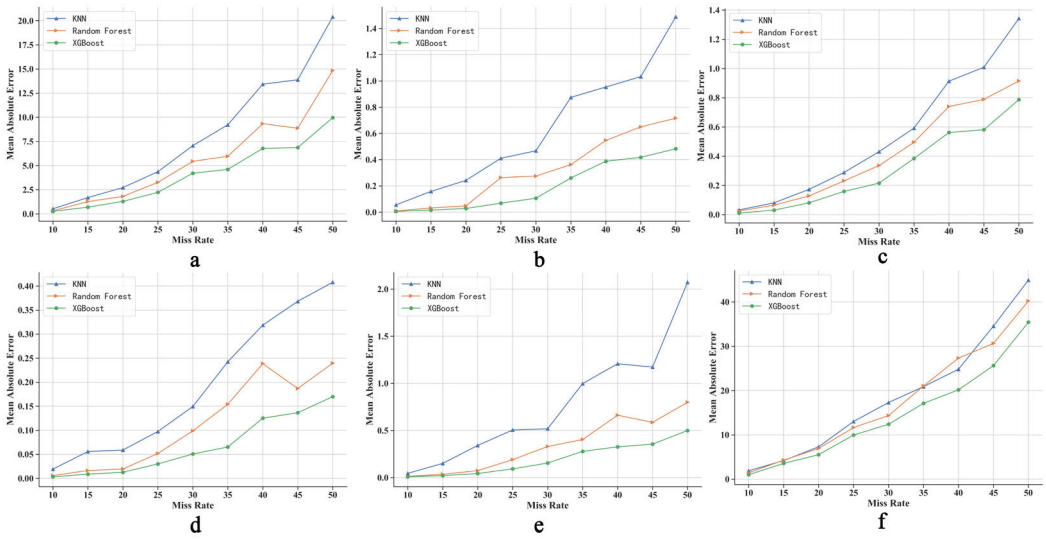


FIGURE 5. Comparison of filling results with different missing ratios. (a)engine speed. (b)gas pedal percentage. (c)speed. (d) instantaneous fuel consumption. (e) torque percentage. (f) NO_x concentration.

The process of using XGBoost regression learning to fill the missing values idea is the same as machine learning to fill the predicted values to the missing values. The process is as follows.

Step 1: Count the number of missing values for each characteristic variable in the missing data set A.

Step 2: (2) To avoid that the feature variables entered into the XGBoost model are missing values, make a copy of the raw data B and prepopulate the missing parts using the mean value of each feature.

Step 3: Dataset A is used as the output of the XGBoost model in descending order of missing values, and the remaining features of Dataset B are used as input variables. The complete part of data set A is used as the training set and the missing part as the prediction set. The XGBoost model is trained to use the trained model to predict missing values and complete a feature fill.

Step 4: Fill the completed feature data to replace the corresponding feature in dataset B. Repeat step (3) until all feature replacements are complete.

Due to a strong correlation between various features of the data used, a regression model was built through the complete data set to consider the size of the missing values with the perspective of attribute correlation to construct more effective data information and enhance the credibility of the information and stability of the prediction model.

To verify the accuracy of XGBoost filling, 500 complete datasets were selected from the experimental data in the current paper, and 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50% of the data were randomly excluded to form different proportional missing value datasets. The KNN fill and random forest (RF) fill were used to compare the test results with XGBoost fill. The mean absolute percentage (MAE) was selected as the evaluation criterion, and the result pair is shown in Fig. 5.

From Fig. 5, it can be found that the accuracy of the XGBoost filling is higher than KNN filling and random forest filling, and it can also achieve better filling in the case of higher missing rate. KNN filling is more sensitive to missing rate such as with an increase in missing rate, its

TABLE 2. Pearson coefficient analysis of correlations between variables.

Pearson coefficient	Vehicle speed	Engine speed	Torque percentage	Instantaneous fuel consumption	Accelerator pedal opening	NO_x concentration
Vehicle speed	1.00	0.64	0.08	0.24	0.37	0.18
Engine speed	0.64	1.00	0.27	0.49	0.58	0.25
Torque percentage	0.08	0.27	1.00	0.94	0.90	0.58
Instantaneous fuel consumption	0.24	0.49	0.94	1.00	0.96	0.62
Accelerator pedal opening	0.37	0.58	0.90	0.96	1.00	0.61
NO_x concentration	0.18	0.25	0.58	0.62	0.61	1.00

accuracy decreases. Random forest and XGBoost belong to the same integrated learning, the difference is that the random forest uses bagging algorithm and XGBoost uses boosting algorithm. In bagging algorithm-based learning process, each training set is independent of each other such as the previous learning results do not have an impact on the subsequent results whereas, in boosting algorithm each learning depends on the previous learning results, therefore XGBoost is more suitable for populating missing data in continuous time series.

In this approach, data features with non-missing values were used to predict the missing data for each feature, experiments were conducted at missing rates from 10% to 50%, and by comparing with KNN filling and random forest filling, XGBoost was able to achieve better filling results for continuous-type time series at 50% data missing rate.

IV. RESULTS AND DISCUSSION

To evaluate the performance of the model, the root-mean square error (RMSE), the coefficient of determination (R^2), mean absolute error (MAE) and normalized root-mean square error (NRMSE) were evaluated using the Eq. (12-15) as follows;

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{12}$$

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \tag{13}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{14}$$

$$\mathrm{NRMSE} = \frac{RMSE}{y_{max} - y_{min}} * 100 \tag{15}$$

Here, y represents the experimental value, $\hat{y}$ shows the predicted value of the model, y_{max} denotes the maximum NO_x concentration, y_{min} represents the minimum NO_x concentration, and n reflects the number of samples in a test set.

Furthermore, the proposed model was compared with the four basic models including random forest [30], support vector regression [31], XGBoost [28] and LSTM [32].

A. FEATURE VARIABLE SELECTION

The selection of input parameters is critical for the model to obtain accurate predictions. Pearson's correlation coefficient is widely used to measure the degree of correlation between two variables, and its value is between -1 and 1, the closer it is to 1, the higher the positive correlation between the two variables, and the closer it is to -1, the higher the negative correlation between the two variables. In this paper, the correlations between variables and NO_x emissions are analyzed according to Pearson's phase relationship, and the correlation coefficients between selected variables and NO_x are greater than 0.15. The Pearson coefficients are shown in Table 2. As mentioned before, 90% of the data was treated for training purpose and 10% of the data was considered as test model results. The subsequences obtained from the decomposition of the CEEMDAN algorithm were input to the LSTM neural network for training. The input variable parameters were 5 whereas the output result was 1. The grid search was performed by the number of hidden layers, the number of neurons, the number of iterations and the batch size. The final determination of the number of hidden layers is 2, the first hidden layer is 12 neurons, the second hidden layer is 20 neurons, the number of iterations is 400, and the batch size is 8 can obtain the best results. The learning rate is adjusted by the optimizer 'Adam', which can effectively solve the problems of vanishing learning rate, slow convergence or large fluctuations of the loss function.

B. ANALYSIS OF CEEMDAN RESULTS

The CEEMDAN algorithm was used to decompose the NO_x concentration series data due to the strong non-linear and non-stationary characteristics of NO_x emission concentrations of diesel vehicles. The added white noise was 0.4, the average magnitude of the white noise was 10,000, and the decomposition yielded 14 IMF values and 1 RES, as shown in Fig. 5. It was observed that with the increase of decomposition times, the stability and regularity of the subsequences were obviously enhanced. The fluctuation of IMF1 to IMF6 was violent and the change of frequency was high with a certain degree of random disorder. This reflected the uncertainty and randomness of NO_x concentration change in diesel engine. The change in the IMF values from IMF7 to IMF9 was relatively moderate and more stable overall, but still there were points where the magnitude of change is large and the volatility cycle is unstable. The range from IMF10 to

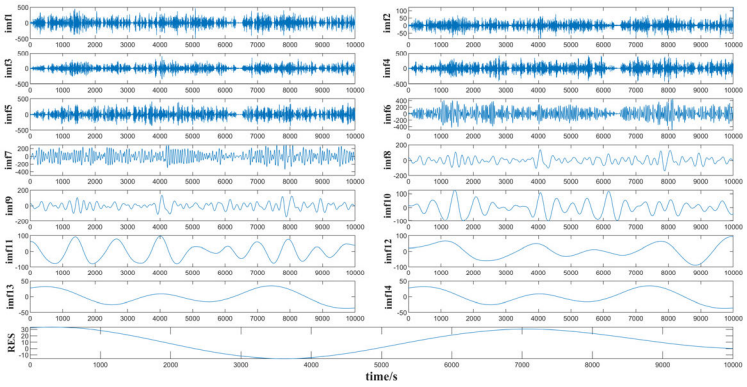


FIGURE 6. Sequential decomposition results of NO_X concentrations based on CEEMDAN decomposition.

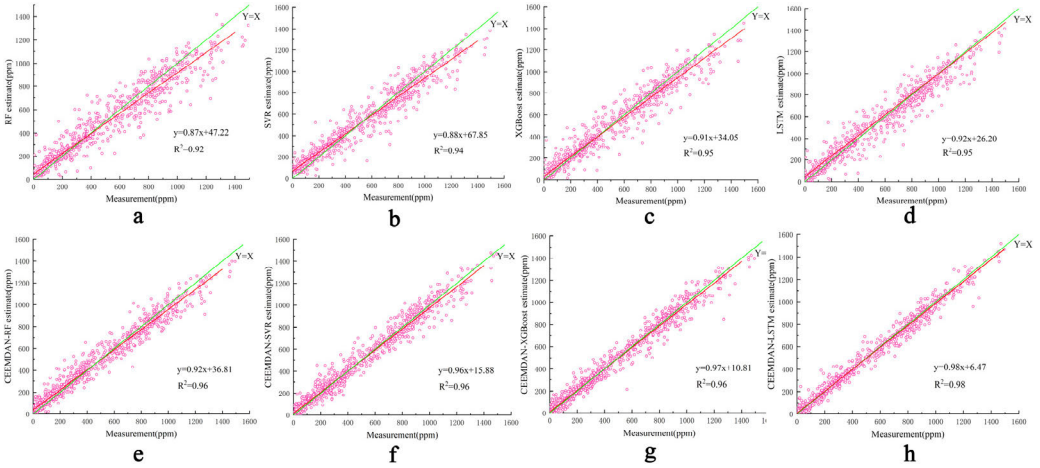


FIGURE 7. Regression analysis of predicted results and measurements from different models. (a) RF. (b) SVR. (c) XGBoost. (d) LSTM. (e) CEEMDAN-RF. (f) CEEMDAN-SVR. (g) CEEMDAN-XGBoost. (h) CEEMDAN-LSTM.

RES was observed to be changing in a flat manner which can be obviously understood as per periodic rule.

The CEEMDAN algorithm was used to decompose diesel NO_X concentrations into subseries of different feature scales. The IMF reflected, not only the local features of NO_X concentrations at different frequencies, but also the hidden feature changes that would not be identified by direct observation of the data. The decomposed data was more regular than the original data series. Although there was still varying degree of non-stationarity between these components, yet their interactions were isolated because they had different characteristic scales. This isolation can reduce the difficulties that non-stationary data bring to the prediction, thus reducing the complexity and noise of the NO_X transient emission data and increasing the accuracy of the LSTM neural network prediction.

C. PREDICTIVE RESULTS DISCUSSION ANALYSIS

Fig 7 shows the prediction results of the eight model-test sets, with horizontal and vertical coordinates corresponding to the test and model prediction values, respectively. The linear function of the test and predicted values were obtained by linear fitting. It can be witnessed that, the closer the slope of the linear function to 1, the better the prediction performance. The linear function slopes of RF, SVR, XGBoost, LSTM, CEEMDAN-RF, CEEMDAN-SVR, CEEMDAN-XGBoost and CEEMDAN-LSTM were 0.87, 0.88, 0.91, 0.92, 0.92,0.96,0.97 and 0.98, respectively, with the CEEMDAN-LSTM model having the largest slope, indicating the highest prediction accuracy. At the same time, it can be noticed that the prediction results are uniformly distributed on both sides of Y=X, indicating that the prediction error was Gaussian, which means that the prediction was relatively ideal. Compared to the four base models, the CEEMDAN decomposed model predictions are closer to Y=X, and the four base model predictions are more discrete.

Fig 8 presents a detailed comparison of experimental values and predictions for the RF, SVR, XGBoost, LSTM, CEEMDAN-RF, CEEMDAN-SVR, CEEMDAN-XGBoost and CEEMDAN-LSTM models. All of the eight models were able to predict the trends of NO_X emissions with reasonable accuracy, with the largest error occurring during the transition period, such as when engine conditions changed rapidly. However, it was found from the endpoints of the Fig. 8a

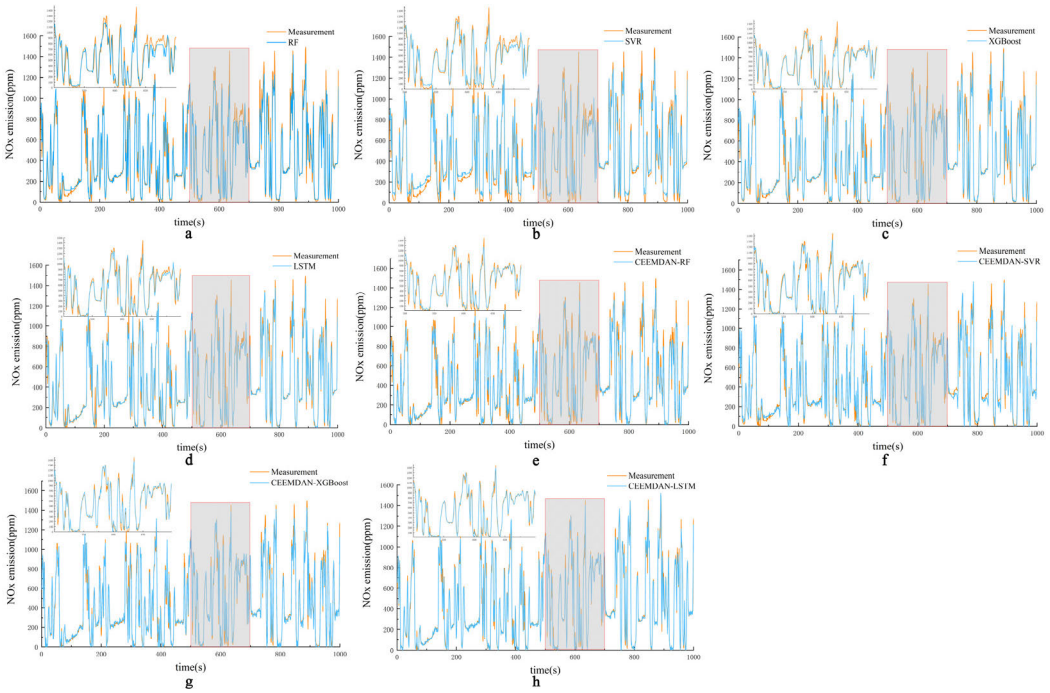


FIGURE 8. Comparative analysis of the predictions and measurements of different models. (a) RF. (b)SVR. (c) XGBoost. (d) LSTM. (e) CEEMDAN-RF. (f) CEEMDAN-SVR. (g) CEEMDAN-XGBoost. (h) CEEMDAN-LSTM.

to Fig. 8d and the data-change intensive section that all the four underlying models suffer from partial aggregation prediction distortion and inability to accurately predict the high-frequency components. From the plot, it can be observed that the base model has more prediction deviation points under high-load and high-speed conditions due to large fluctuations in engine operation, resulting in larger or smaller predictions than the experimental values. Observing Fig. 8e to Fig. 8h, the four basic models are optimized by the CEEMDAN algorithm, and the model fit is significantly improved, with a 30.50% reduction in RMSE for RF, a 24.86% reduction for SVR, a 27.26 reduction for XGBoost, and a 40.66% reduction for LSTM. One of the major advantages of using the CEEMDAN algorithm is that the predicted values can better fit the experimental values at various extremes and points of drastic changes, which is not true in the case of other base models. Moreover, it has a substantial prediction effect on NO_X emissions, which also shows that CEEMDAN algorithm is useful to improve the prediction accuracy of transient time series. Meanwhile, CEEMDAN-LSTM boosts the most and has the best predictive performance.

To make an accurate quantitative comparison between all the models, the parameters such as RMSE, R^2, MAE, and NRMSE were evaluated for each model and presented in Table 3. The CEEMDAN algorithm processed models that were superior to the four underlying models, with CEEMDAN-LSTM being the most effective. Due to the wide range of NO_X concentrations in the test data (0–1701ppm), the RMSE values were relatively large, but all the predicted values yielded good results. Also, in terms of prediction accuracy, The CEEMDAN decomposition optimized model is significantly improved over the base model. CEEMDAN-RF improves R^2, MAE and NRMSE by 4.38%, 23.32% and 30.50%, respectively, compared to RF. CEEMDAN-SVR improves R^2, MAE and NRMSE by 3.19%, 27.98% and 24.86%, respectively, compared to SVR. CEEMDAN-XGBoost improves R^2, MAE and NRMSE by 3.19%, 27.98% and 24.86%, respectively, compared to SVR. CEEMDAN-XGBoost vs. XGBoost improved R^2, MAE, and NRMSE by 2.11%, 18.69%, and 27.26%, respectively, compared to LSTM. CEEMDAN-LSTM improved R^2, MAE, and NRMSE by 3.16%, 23.32%, and 37.69%, respectively, compared to LSTM.

The results of the model stability and error analysis are shown in Figure 9, and it is observed that the base models all have some errors of more than 100 ppm, indicating that the prediction values deviate from the experimental values. After using CEEMDAN decomposition, the prediction errors of the four basic models are significantly reduced, among which CEEMDAN-RF is relatively poorer, but also significantly improved than the RF basic model. CEEMDAN-LSTM has the best prediction effect, only a very small part of the error exceeds 100 ppm, indicating that CEEMDAN-LSTM model has better prediction in transient prediction. Accuracy and stability. Table 4 shows the results of some previously conducted efforts related to the prediction of NO_X transient emissions from diesel vehicles.

Comparing the quantities enlisted in Table 4, it can be observed that the CEEMDAN-LSTM model has higher prediction accuracy in real road prediction, and the combined CEEMDAN-LSTM model exhibits improved prediction performance as compared to machine learning models (GBRT,

TABLE 3. Statistical results of prediction results.

Evaluation indicators	Model							
	RF	SVR	XGBoost	LSTM	CEEMDAN-RF	CEEMDAN-SVR	CEEMDAN-XGBoost	CEEMDAN-LSTM
RMES/ppm	98.28	84.73	80.16	77.71	68. 31	63. 67	58. 31	46.11
R^2	0.92	0.94	0.95	0.95	0. 96	0. 97	0. 97	0.98
MAE/ppm	62.66	63.69	49.23	47.86	48. 05	45. 87	40. 03	29.82
NRMSE/%	5.78	4.98	4.71	4.57	4.02	3.74	3.43	2.71

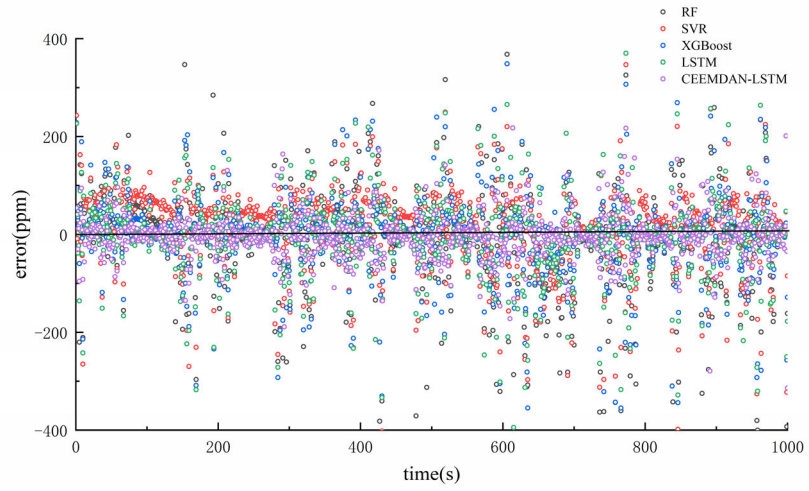


FIGURE 9. Error plot of NO_x concentration prediction results.

TABLE 4. Comparison of the results of the NO_x transient prediction paper.

Methodological model	Range of experimental values	RMSE	R^2	NRMSE (%)
gradient boosted regression tree (GBRT) [33]	0—0.5 g/s	0.073 g/s	0.80	14.60
engine maps [7]	0—0.24 g/s	0.022 g/s	0.74	9.17
multi-layer perceptron (MLP) [34]	0—400 ppm	N/A	N/A	5.11
artificial neural network (ANN) [35]	0—2.5 g/s	0.098 g/s	0.95	3.92
control-oriented semi-empirical model [36]	0—1400 ppm	44 ppm	0.96	3.14
adaptive Neuro-Fuzzy Inference System (ANFIS) [37]	0—0.1 g/s	0.012 g/s	0.73	12.10
adaptive Support Vector Machine [38]	0—1700 ppm	185.52 ppm	0.40	10.91
CEEMDAN-LSTM (This research)	0—1701ppm	46.11 ppm	0.98	2.71

MLP) and artificial neural networks. This is due to the fact that CEEMDAN takes the original high-frequency-variation time series and decomposes it into multiple subseries in different frequency domains to reduce the effects of sudden and multiple changes in NO_x concentration. It provides better prediction of non-stationary, non-linear NO_x concentration data for diesel vehicles, operating under multiple working conditions and complex environmental surroundings.

V. CONCLUSION

In order to improve the NO_x transient emission prediction performance of diesel vehicles, a method for estimating diesel vehicle transient NO_x emissions based on the combination of signal processing CEEMDAN algorithm and LSTM neural network is proposed. CEEMDAN decomposes the NO_x emission transient data with the engine in an unsteady state into more regular and smooth subseries, and uses LSTM to predict each subseries. The proposed model is compared with seven other models and the following conclusions can be drawn.

1) Among the models covered in the paper, the LSTM model using deep learning algorithms has better prediction performance. The LSTM neural network has the best prediction performance among the four base models and the four hybrid models, which indicates that LSTM has a better learning ability for complex time series.

2) The model using signal processing has higher prediction accuracy than the base model. CEEMDAN-RF, CEEMDAN-SVR, CEEMDAN-XGBoost and CEEMDAN-LSTM outperform RF, SVR, XGBoost and LSTM, respectively, indicating that the decomposition method is an effective way to improve NO_x transient prediction for diesel vehicles.
3) The CEEMDAN-LSTM model improved the RMSE compared to RF, SVR, XGBoost, LSTM, CEEMDAN-RF, CEEMDAN-SVR, and CEEMDAN-XGBoost by 53.08%, 45.58%, 42.48%, 40.66%, 32.50%, 27.58%, and 20.92%, respectively. It demonstrates that the CEEMDAN-LSTM proposed in this paper has good applicability and reliability in the field of NO_x transient prediction in diesel engines.

Overall, the proposed hybrid model can provide more reliable and accurate prediction results for instantaneous NO_x emissions from diesel vehicles, thus providing a model basis for replacing physical NO_x sensors. On the other hand, accurate NO_x instantaneous emission prediction for diesel vehicles can reduce the pollution emission level and improve the development and design efficiency of diesel engines.

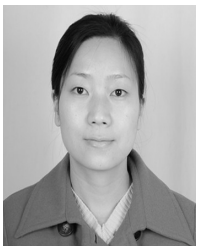
Although the model has a higher predictive performance than the general basic model, such as temperature, atmospheric pressure and humidity also have an impact on emissions. Therefore, the effects of seasonal variation, altitude change, etc. on emission prediction will be considered in future studies and incorporated into the NO_x transient prediction.

REFERENCES

[1] Z. Gao, J. C. Conklin, C. S. Daw, and V. K. Chakravarthy, "A proposed methodology for estimating transient engine-out temperature and emissions from steady-state maps," *Int. J. Engine Res.*, vol. 11, no. 2, pp. 137–151, Apr. 2010.
[2] J. D. K. Bishop, M. E. J. Stettler, N. Molden, and A. M. Boies, "Engine maps of fuel use and emissions from transient driving cycles," *Appl. Energy*, vol. 183, pp. 202–217, Dec. 2016.
[3] R. Egnell, "Combustion diagnostics by means of multizone heat release analysis and NO calculation," *SAE Trans.*, vol. 107, pp. 691–710, May 1998.
[4] S. A. Provataris, N. S. Savva, T. D. Chountalas, and D. T. Hountalas, "Prediction of NO_x emissions for high speed DI diesel engines using a semi-empirical, two-zone model," *Energy Convers. Manage.*, vol. 153, pp. 659–670, Dec. 2017.
[5] C. Guardiola, J. J. López, J. Martín, and D. García-Sarmiento, "Semiempirical in-cylinder pressure based model for NO_x prediction oriented to control applications," *Appl. Thermal Eng.*, vol. 31, pp. 3275–3286, Jun. 2011.
[6] A. Bhave, M. Kraft, L. Montorsi, and F. Mauss, "Modelling a dual-fuelled multi-cylinder HCCI engine using a PDF based engine cycle simulator," SAE Tech. Paper 2004-01-0561, 2004, pp. 3275–3286, vol. 31.
[7] F. Rosero, N. Fonseca, J.-M. López, and J. Casanova, "Real-world fuel efficiency and emissions from an urban diesel bus engine under transient operating conditions," *Appl. Energy*, vol. 261, Mar. 2020, Art. no. 114442.
[8] R. Reitz, "Development and testing of diesel engine CFD models," *Prog. Energy Combustion Sci.*, vol. 21, no. 2, pp. 173–196, 1995.
[9] R. Mobasheri, Z. Peng, and S. M. Mirsalim, "Analysis the effect of advanced injection strategies on engine performance and pollutant emissions in a heavy duty DI-diesel engine by CFD modeling," *Int. J. Heat Fluid Flow*, vol. 33, no. 1, pp. 59–69, Feb. 2012.
[10] B. S. Dahifale and A. S. Patil, "Diesel engine performance improvement for constant speed application using CFD," in *Proc. ASME Int. Mech. Eng. Congr. Expo*. New York, NY, USA: American Society of Mechanical Engineers, 2017, pp. 1–11.
[11] D. Yoo, D. Kim, W. Jung, N. Kim, and D. Lee, "Optimization of diesel combustion system for reducing PM to meet tier4-final emission regulation without diesel particulate filter," SAE Tech. Paper 2013-01-2538, 2013.
[12] İ. Ceylan, O. Erkaymaz, E. Gedik, and A. E. Gürel, "The prediction of photovoltaic module temperature with artificial neural networks," *Case Stud. Thermal Eng.*, vol. 3, pp. 11–20, Jul. 2014.
[13] G. Wang, O. I. Awad, S. Liu, S. Shuai, and Z. Wang, "NO_x emissions prediction based on mutual information and back propagation neural network using correlation quantitative analysis," *Energy*, vol. 198, May 2020, Art. no. 117286.
[14] G. Alcan, M. Unel, V. Aran, M. Yilmaz, C. Gurel, and K. Koprubasi, "Diesel engine NO_x emission modeling using a new experiment design and reduced set of regressors," *IFAC-PapersOnLine*, vol. 51, no. 15, pp. 168–173, 2018.
[15] S. Bhowmik, A. Paul, R. Panua, S. K. Ghosh, and D. Debroy, "Performance-exhaust emission prediction of diesosenol fueled diesel engine: An ANN coupled MORSM based optimization," *Energy*, vol. 153, pp. 212–222, Jun. 2018.
[16] S. Lotfan, R. A. Ghiasi, M. Fallah, and M. H. Sadeghi, "ANN-based modeling and reducing dual-fuel engine's challenging emissions by multi-objective evolutionary algorithm NSGA-II," *Appl. Energy*, vol. 175, pp. 91–99, Aug. 2016.
[17] I. Arsie, A. Cricchio, C. M. De, F. Lazzarini, C. Pianese, and M. Sorrentino, "Neural network models for virtual sensing of NO_x emissions in automotive diesel engines with least square-based adaptation," *Control Eng. Pract.*, vol. 61, pp. 11–20, Apr. 2017.
[18] Y.-Y. Wang, Y. He, and S. Rajagopalan, "Design of engine-out virtual NO_x sensor using neural networks and dynamic system identification," *SAE Int. J. Engines*, vol. 4, no. 1, pp. 837–849, Apr. 2011.
[19] J. Cao, Z. Li, and J. Li, "Financial time series forecasting model based on CEEMDAN and LSTM," *Phys. A, Stat. Mech. Appl.*, vol. 519, pp. 127–139, Apr. 2019.
[20] T. Li, Z. Qian, and T. He, "Short-term load forecasting with improved CEEMDAN and GWO-based multiple kernel ELM," *Complexity*, vol. 2020, Feb. 2020, Art. no. 1209547.
[21] H. Lu, X. Ma, K. Huang, and M. Azimi, "Prediction of offshore wind farm power using a novel two-stage model combining kernel-based nonlinear extension of the arps decline model with a multi-objective grey wolf optimizer," *Renew. Sustain. Energy Rev.*, vol. 127, Jul. 2020, Art. no. 109856.
[22] P. Du, J. Wang, Z. Guo, and W. Yang, "Research and application of a novel hybrid forecasting system based on multi-objective optimization for wind speed forecasting," *Energy Convers. Manage.*, vol. 150, pp. 90–107, Oct. 2017.
[23] C. Lv, J. Zhao, C. Wu, T. Guo, and H. Chen, "Optimization of the end effect of Hilbert-Huang transform (HHT)," *Chin. J. Mech. Eng.*, vol. 30, no. 3, pp. 732–745, May 2017.
[24] F. Zhao, Z. Ruan, Z. Yue, D. L. S. Hung, S. Som, and M. Xu, "Time-sequenced flow field prediction in an optical spark-ignition direct-injection engine using bidirectional recurrent neural network (bi-RNN) with long short-term memory," *Appl. Thermal Eng.*, vol. 173, Jun. 2020, Art. no. 115253.
[25] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. C, Emerg. Technol.*, vol. 54, pp. 187–197, May 2015.
[26] W. Di, K. Yu, and L. Ze-Rui, "Design of vehicle exhaust remote monitoring data center platform," *J. Atmos. Environ. Opt.*, vol. 11, p. 453, Nov. 2016.
[27] M. J. Baxter, "Detecting multivariate outliers in artefact compositional data," *Archaeometry*, vol. 41, no. 2, pp. 321–338, Aug. 1999.
[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[29] M. Wang, J. Yu, and Z. Ji, "Credit fraud risk detection based on XGBoost-LR hybrid model," in *Proc. Int. Conf. Electron. Bus.*, Guilin, China, vol. 2, no. 6, 2018, pp. 336–343.
[30] A. Liaw, M. Wiener, and A. Liaw, "Classification and regression with random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[31] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004.
[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[33] Y. Pan, S. Chen, F. Qiao, S. V. Ukkusuri, and K. Tang, "Estimation of real-driving emissions for buses fueled with liquefied natural gas based on gradient boosted regression trees," *Sci. Total Environ.*, vol. 660, pp. 741–750, Apr. 2019.

[34] Q. Zhang, A. Pennycott, R. Burke, S. Akehurst, and C. Brace, "Predicting the nitrogen oxides emissions of a diesel engine using neural networks," SAE Tech. Paper 2015-01-1626, 2015, doi: 10.4271/2015-01-1626.
[35] R. Jaikumar, S. M. S. Nagendra, and R. Sivanandan, "Modeling of real time exhaust emissions of passenger cars under heterogeneous traffic conditions," *Atmos. Pollut. Res.*, vol. 8, no. 1, pp. 80–88, Jan. 2017.
[36] R. Finesso, G. Hardy, C. Maino, O. Marello, and E. Spessa, "A new control-oriented semi-empirical approach to predict engine-out NO_x emissions in a euro VI 3.0 l diesel engine," *Energies*, vol. 10, no. 12, p. 1978, Nov. 2017.
[37] K. Gopalakrishnan, A. Mudgal, and S. Hallmark, "Neuro-fuzzy approach to predictive modeling of emissions from biodiesel powered transit buses," *Transport*, vol. 26, no. 4, pp. 334–352, Jan. 2012.
[38] M. Ghanbari, G. Najafi, B. Ghobadian, R. Mamat, M. M. Noor, and A. Moosavian, "Support vector machine to predict diesel engine performance and emission parameters fueled with nano-particles additive to diesel fuel," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 100, Dec. 2015, Art. no. 012069.

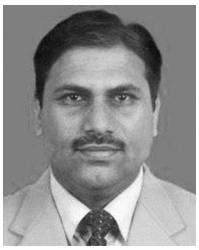
YANG YU received the bachelor's degree from Panzhihua University, in June 2018. He is currently pursuing the master's degree with Southwest Forestry University. His current research interests include machine learning, motor vehicle emissions, and deep learning.



YANYAN WANG received the master's degree from Kunming University of Science and Technology, in 2018. She is currently an Experimenter with Southwest Forestry University. Her main research interest includes calibration of electronically controlled common rail diesel engine.

JIAQIANG LI received the Ph.D. degree from the Beijing Institute of Technology, in 2017. He is currently a Lecturer with Southwest Forestry University. His main research interests include vehicle emission characteristics, mechanism and matching of solid SCR, and generation mechanism and control strategy of NOx emission.

MINGLIANG FU received the Ph.D. degree from Beijing Institute of Technology. He is currently an Associate Professor with the Chinese Research Academy of Environmental Sciences. His main research interests include motor vehicle emissions, remote sensing monitoring, ship emissions, and air pollution control.



ASAD NAEEM SHAH received the Ph.D. degree from the Beijing Institute of Technology. He is currently Professor with the University of Engineering and Technology. His interests include diesel engine combustion control, diesel fuel emission control, and air emission pollution control.

CHAO HE received the Ph.D. degree from the Beijing Institute of Technology, in 2008. He is currently the Dean of the School of Machinery and Transportation, Southwest Forestry University. He has published more than 53 articles. His main research interests include intelligent control of internal combustion engine, energy saving and new energy vehicles, and car crash safety.