Environmental Research 247 (2024) 118190

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Enhancing vehicular emissions monitoring: A GA-GRU-based soft sensors approach for HDDVs

Luoshu Yang [a], Yunshan Ge [a], Liqun Lyu [a,*], Jianwei Tan [a], Lijun Hao [a], Xin Wang [a], Hang Yin [b], Junfang Wang [b]

[a] School of Mechanical Engineering, Beijing Institute of Technology, Beijing, 100081, China
[b] State Environmental Protection Key Laboratory of Vehicle Emission Control and Simulation, Chinese Research Academy of Environmental Sciences, Beijing, 100012, China

ARTICLE INFO

Keywords:
Soft sensor
Vehicle emissions
Gate recurrent unit
Genetic algorithm

ABSTRACT

Vehicle emissions have a serious impact on urban air quality and public health, so environmental authorities around the world have introduced increasingly stringent emission regulations to reduce vehicle exhaust emissions. Nowadays, PEMS (Portable Emission Measurement System) is the most widely used method to measure on-road NOx (Nitrogen Oxides) and PN (Particle Number) emissions from HDDVs (Heavy-Duty Diesel Vehicles). However, the use of PEMS requires a lot of workforce and resources, making it both costly and time-consuming. This study proposes a neural network based on a combination of GA (Genetic Algorithm) and GRU (Gated Recurrent Unit), which uses CC (Pearson Correlation Coefficient) to determine and simplify OBD (On-board Diagnosis) data. The GA-GRU model is trained under three real driving conditions of HDDVs, divided by vehicle driving parameters, and then embedded as a soft sensor in the OBD system to monitor real-time emissions of NOx and PN within the OBD system. This research addresses the existing research gap in the development of soft sensors specifically designed for NOx and PN emission monitoring. In this study, it is demonstrated that the described soft sensor has excellent R^2 values and outperforms other conventional models. This research highlights the ability of the proposed soft sensor to eliminate outliers accurately and promptly while consistently tracking predictions throughout the vehicle's lifetime. This method is a groundbreaking update to the vehicle's OBD system, permanently adding monitoring data to the vehicle's OBD, thus fundamentally improving the vehicle's self-monitoring capabilities.

1. Introduction

With the yearly increase in vehicle ownership, vehicle emissions have become a hot topic of social concern (MEE, 2022). However, NOx and PN emissions from diesel engines can cause problems such as air pollution, which will cause severe respiratory diseases such as asthma (So et al., 2022; Liu et al., 2023b). Therefore, monitoring vehicle pollutant emissions becomes imperative. At present, the primary method for such monitoring is through PEMS tests. However, this testing approach necessitates the suspension of all ongoing vehicle operations. The testing equipment must be affixed to the vehicle and operated following designated testing protocols. This process entails the coordination of multiple human and material resources and a great number of testing times. Furthermore, for real-time vehicle monitoring, this method lacks the capability to assess pollutant emissions continuously. To reduce the emissions from HDDVs such as NOx and PN, many countries introduced corresponding laws, for example, the Euro-V, Euro-VI, China-V, and China-VI. The European Union limits the PN emissions for diesel cars to 1.2×10^{12} #/kWh. In addition, the China-VI regulation limits the NOx emission of 690 mg/kWh and PN emission of 1.2×10^{12} #/kWh (MEE, 2018). Hence, in order to comply with stringent emission regulations, there is a necessity to devise a novel methodology for conducting emission tests, serving as a viable alternative to the PEMS test. This proposed method should aim to reduce costs without compromising the precision of measurements.

Currently, there is a variety of methods to evaluate vehicle emissions, such as chassis and engine dynamometer testing (Jaworski et al., 2018), road tunnel measurements (Geller et al., 2005), optical remote sensing systems (Xie et al., 2019), plume chasing measurements (Bishop et al., 2010) and PEMS (O'Driscoll et al., 2016), etc. However,

* Corresponding author.
E-mail address: bitlvliqun@163.com (L. Lyu).

https://doi.org/10.1016/j.envres.2024.118190
Received 7 November 2023; Received in revised form 2 January 2024; Accepted 10 January 2024
Available online 17 January 2024

time-consuming, workforce-consuming, and funds-consuming processes are required for each of them, so it is essential to produce a new method that does not need physical sensors to tackle these problems.

The soft sensor is a virtual sensor based on mathematical models and data analysis, which can be utilized in monitoring, measurement, and prediction of many parameters and attributes such as temperature, pressure, and flow rate (Lei and Wang, 2022; Mei et al., 2017; Yan et al., 2017) et al. In the study, the author employed a soft sensor based on an LSTM model to monitor various parameters in wastewater treatment processes. The comprehensive findings indicated that the method successfully predicted key performance indicators of biological treatment in wastewater, achieving automation and reducing the need for installing physical sensors, thereby lowering costs (Yu et al., 2023). When managing intricate changes in industrial processes with traditional instruments and meters, the high level of uncertainty and error associated with these tools makes obtaining accurate measurement results challenging (Wang et al., 2023b; Zhou et al., 2023). It can help to control and optimize the product procedure in time to enhance production efficiency and product quality. Soft sensor usually utilizes machine learning or neural network as a black box to compute the procedure to obtain the optimum outcome, and neural network models are known for their convenience in modeling complex nonlinear systems (Gholipour et al., 2007). In the vehicle field, the soft sensors can be used to monitor parameters such as engine performance, emissions, and fuel efficiency. They help improve driving safety and vehicle performance. By extensive training and validating a substantial volume of data, continuously refining data preprocessing techniques, fine-tuning relevant neural network architectures and parameters, and enhancing the accuracy of soft sensors for predicting pollutants, the integration of soft sensors can streamline vehicle systems by leveraging existing physical sensor data. This eliminates the need for additional sensors, resulting in a simplified sensor layout and overall vehicle system. This approach not only simplifies the vehicle but also reduces costs associated with sensor installation and maintenance (Zhang et al., 2022).

In Liu and Xie's study (Liu and Xie, 2020), a kernel-based method was used to provide an overview of achieving hard-to-measure variable prediction. In (Wang et al., 2015), a comprehensive evaluation of different variable selection methods for PLS (Partial Least Squares)-based soft sensor development is presented, and a new metric is proposed to assess the performance of different variable selection methods, demonstrating the method has strengths and limitations.

As for vehicle emissions, researchers (Tan et al., 2022) proposed LOLIMOT (Local Linear Neuro Fuzzy Model) methods to predict the gasoline engine emissions and the outcome results of R^2 values greater than 0.99. Another literature (Andrade et al., 2022) proposed a soft-sensor solution based on an embedded system designed to retrieve data from vehicles through their OBD-II interface, processing different inputs to provide estimated values of CO_2 emissions over time (Taghavifar et al., 2016). However, little soft sensor approach has been applied to pollutant prediction of HDDVs. Few studies have been able to draw on any structured research into the opinions and attitudes of PN soft sensors due to the randomness of aftertreatment DPF (Diesel Particulate Filter) (Kontses et al., 2020; Zhong et al., 2021).

This study introduces a soft sensor method based on GRU and implements a GA in Python 3.8. Throughout the model-building process, the optimal hyperparameters essential for the model are iteratively determined by integrating the genetic algorithm with response emission data. This approach significantly diminishes the need for manual parameter tuning, enhancing the computational accuracy of the model and concurrently reducing the response time (Yang, 2021). It can utilize the soft sensor embedded in the OBD system as software to test the real-drive emissions. To obtain the optimum performance under different driving conditions, the PEMS data will be selected as the original dataset, which has been divided into three driving conditions (urban, suburb, and highway conditions) and selected as the training dataset. After training, the validation dataset will be used to evaluate the soft sensor's performance and compare it with other models to obtain the metric values of R^2 under different driving conditions.

Moreover, some conventional models were used to predict the dataset and compare the outcome between GA-GRU (Genetic Algorithm with Gated Recurrent Unit) and them (Guo et al., 2024; Liu et al., 2023a). The performance of the described model is better than others. Furthermore, to verify that the sensor can be applied throughout the entire lifecycle of the vehicle, the described soft sensor is used to predict the datasets from different times, and the frequency of these dataset points is also 1 Hz (Lyu et al., 2023). By continuously modifying the model, it is developed into a compatible and reusable soft sensor working with the OBD system. Ensure that the soft sensor model, designed to substitute physical sensors, is seamlessly integrated into the vehicle's OBD system. This integration aims to monitor real-time pollutant emissions during actual driving, effectively minimizing the need for physical sensors. This novel approach to vehicle pollutant testing not only reduces actual production costs and the demand for workforce and resources but also boosts the volume of data output from the vehicle's OBD. Furthermore, it optimizes the post-processing device of the vehicle.

2. Material and methodology

2.1. Model select

2.1.1. Recurrent neural networks and gate recurrent unit

RNN (Recurrent Neural Networks) are a class of networks that can predict future trends, essentially, which are inherently highly correlated with time-series sequences and have been widely addressed in time-series sequences (Geron, 2019). As the pollutants' data undergoes time-series testing at a frequency of 1 Hz, and given that various operating conditions and other factors influence the emissions of HDDVs, these contributing factors to pollutant levels are inherently linked to time-series. Consequently, accurate forecasts of pollutant emissions exhibit a strong correlation with time, so in this work, the RNN network has been used for pollutant prediction. The working principle of RNN is shown in Fig. S1, which is extended by a timeline.

GRU is a variant of RNN aimed at overcoming the long-term dependency problem in traditional RNNs. GRU introduces mechanisms to capture and convey long-term dependency information within sequences more effectively. It incorporates an update gate responsible for determining the extent to which information from previous memory cells should be preserved. The output of this gate, ranging between 0 and 1, signifies the proportion of past information to be retained. Alongside the update gate, GRU also features a reset gate, dictating how much information in prior memory cells is reset to accommodate new input data. The GRU model comprises a memory cell state and a hidden state. Memory cells store historical information, while hidden states encapsulate the current model state. The outputs from the update and reset gates play a pivotal role in fine-tuning the update and transmission dynamics of these two states. The recurrent neural network basis consists of neurons, each of which has two sets of weights, one for the current input $x_{(t)}$ and the other for the feedback output $y_{(t-1)}$ from the previous time step. These weights can be referred to as W_x and W_y. The output vector for the entire loop layer is calculated from **Eq. (1)**.

$$y_{(t)} = \varphi\left(W_x^T x_{(t)} + W_y^T y_{(t-1)} + b\right) \tag{1}$$

Where b is the bias vector; $\varphi(\cdot)$ is the activation function.

The output of the whole recurrent neural network is given by **Eq. (2)**.

$$Y_{(t)} = \varphi\left(W_x X_{(t)} + W_y Y_{(t-1)} + b\right) \tag{2}$$

Where $Y_{(t)}$ is the output for each time step at the moment; $X_{(t)}$ is the inputs at all times; W_x is the input weights for the current time step; W_y is the output weights for the current time step; b is the bias in each time

step.

However, a recurrent neural network only relies on a single hidden layer for data memory function, which is not enough for processing long sequences, so more memory gates need to be added to the recurrent neural network to enable the neural network to have long-term memory function, which has a higher level of improvement for predicting instantaneous emission in HDDVs.

Furthermore, on the basis of traditional RNN, Chung's paper has indicated that these advanced recurrent units in the LSTM (Long Short-Term Memory) and GRU models are indeed better than more traditional recurrent units, which were mentioned as tanh units (Chung et al., 2014; Hochreiter and Schmidhuber, 1997). Optimized it to have long-term memory ability and have advantages in processing long time series data, especially for HDDVs emission data, and its test results are generally better than RNN.

GRU and LSTM neural networks are variants of RNN. Both of them solve the short-term memory problem of ordinary RNN by setting the gate mechanism to extend the memory time. The gate mechanism effectively alleviates the gradient disappearance and gradient explosion problems during the training process. The gating mechanism effectively alleviates the gradient disappearance and gradient explosion problems during training, simplifies the complexity of tuning, stores the useful features in the memory gate, and removes the useless features in the forgetting gate, which simplifies the complexity of the neural network during training. Compared with the LSTM, GRU combines the forgetting gate and the input gate into one update gate, and the other gate is a reset gate. Without introducing additional internal state conditions and directly introducing linear dependencies between the current and historical states, the GRU network works as follows.

(1) Calculate reset gate and candidate status:

The reset gate r_t is used to control whether the computation of the candidate state $\widetilde{h}_t$ depends on the state $\widetilde{h}_{t-1}$ at the previous moment.

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_\gamma\right) \tag{3}$$

Candidate status at the current moment.

$$\widetilde{h}_t = \tanh(W_r x_t + U_r [\![(r_{t\odot h}]\!]_{t-1}) + b_h) \tag{4}$$

(2) Calculate the update gate and current status:

The update gate z_t is used to control how much information the current state h_t needs to retain from the historical state h_{t-1} and how much new information needs to be accepted from the candidate state $\widetilde{h}_t$.

$$Z_t = \sigma(W_r x_t + U_z h_{t-1} + b_Z) \tag{5}$$

The hidden state h_t is then calculated as follows:

$$h_t = z_{t\odot h_{t-1}} + (\underline{1} - z_t) \odot \widetilde{h}_t \tag{6}$$

When $z_t = 0$ and $r_t = 1$, the hidden state h_t is the computational formula in the simple recurrent neural network, and the GRU network degenerates to a simple RNN.

$$h_t = \tanh(W_h x_t + U_h h_{t-1} + b_h) \tag{7}$$

In addition, the formula for finding the hidden state h_t is organized as follows: there is a linear relationship between the current state h_t and the historical state h_{t-1} , and there is also a nonlinear relationship, which can alleviate the Gradient Vanishing problem to a certain extent.

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tanh(W_h x_t + U_h [\![(r_t \circ h]\!]_{t-1}) + b_h) \tag{8}$$

In order to accurately and quickly calculate rapidly changing instantaneous emissions and monitor them, the model will be used for soft sensors, and this method has the properties of timeliness and robustness, also includes extremely fast computing speed, and can fulfill the need for real driving emission.

2.1.2. Genetic algorithm

GA was designed and proposed with reference to the laws of biological DNA inheritance, simulating a mathematical and computational model of Darwinian biological evolution and genetic mechanisms. Genetic algorithms simulate the changes of biological DNA mathematically and the crossover and mutation of biological processes of chromosomes through computer simulation (Wang et al., 2023a).

The GA is written in Python3.8, and the structure is shown in Fig. 1. First, establish the initial population and then calculate the hyperparameters in the GRU model. If the outcome is dissatisfied, perform the crossover operation and mutation operation. Then, verify the parameters again, following this procedure, and finally output the convergence results as the best parameters in the next training.

2.1.3. GA encoded GRU

Therefore, there are various hyperparameters of the GRU model that could influence the model outcome, such as optimizer learning rate, number of neural network layers, number of hidden layer neurons, batch size, and training generations. In solving multiple-parameter combination problems, finding the optimal combination plays a crucial role in determining the structural parameters of the RNN. The manual tuning is basically based on past training experience, and the hyperparameters with relatively high accuracy are selected and trained. Manual tuning not only has low accuracy but also has strong subjectivity in the selection of parameters, which may be trapped in local minimum. It is an increasingly important issue in neural network. So, in order to avoid the huge error, the GA and GRU models were encoded on Python3.8, the structure of which is shown in Fig. 1. The GA was used to determine the hyperparameters automatically and more accurately.

2.2. Data acquisition/modeling data

2.2.1. Vehicle selection

In this paper, PEMS tests are carried out on long-haul heavy-duty diesel trucks conforming to China VI, which are obtained as our original datasets. The test vehicles adopt the same emission control devices, including EGR (Exhaust Gas Re-circulation), DOC (Diesel Oxidation Catalyst), DPF, SCR (Selected Catalytic Reduction), and ASC (Ammonia Slip Catalyst), which are the mainstream strategies for China-VI HDDVs. Detailed information on the test trucks is shown in Table 1.

To eliminate the effects of different fuels on the emission results and to avoid interfering with the prediction models, all test HDDVs adopted the same batch of commercial #0 diesel. The original datasets were divided into three driving conditions by vehicle speed. The first 20% of the test conditions are urban conditions, the middle 25% are suburb conditions, and the remaining 55% are highway conditions. The data collection frequency is 1Hz as the basis for analysis, and the corresponding data statistics are shown in Table 2 after the basic statistics of the data.

2.2.2. PEMS apparatus

The PEMS apparatus used for the PEMS test is OBS-ONE, manufactured by HORIBA Ltd., which mainly includes the gaseous emission measurement unit, exhaust flow meter unit, OBD, environment monitoring unit, and GPS (Global Positioning System). The PEMS apparatus installation is shown in Fig. S2. The NOx is measured by the CLD (Chemiluminescent Detector) analyzer, and the measurement precision is within ±0.3 % of full scale or ±2.0 % of readings. The exhaust flow rate is measured by a pitot tube, with the measurement precision within ±1.0 % of full scale. At the beginning of each PEMS test, the validation test was conducted on the engine dynamometer to ensure the margins between the PEMS and laboratory apparatus of measurement results met the accuracy requirement. Meanwhile, the PEMS apparatus was set to zero and calibrated before and after each test to ensure the accuracy of

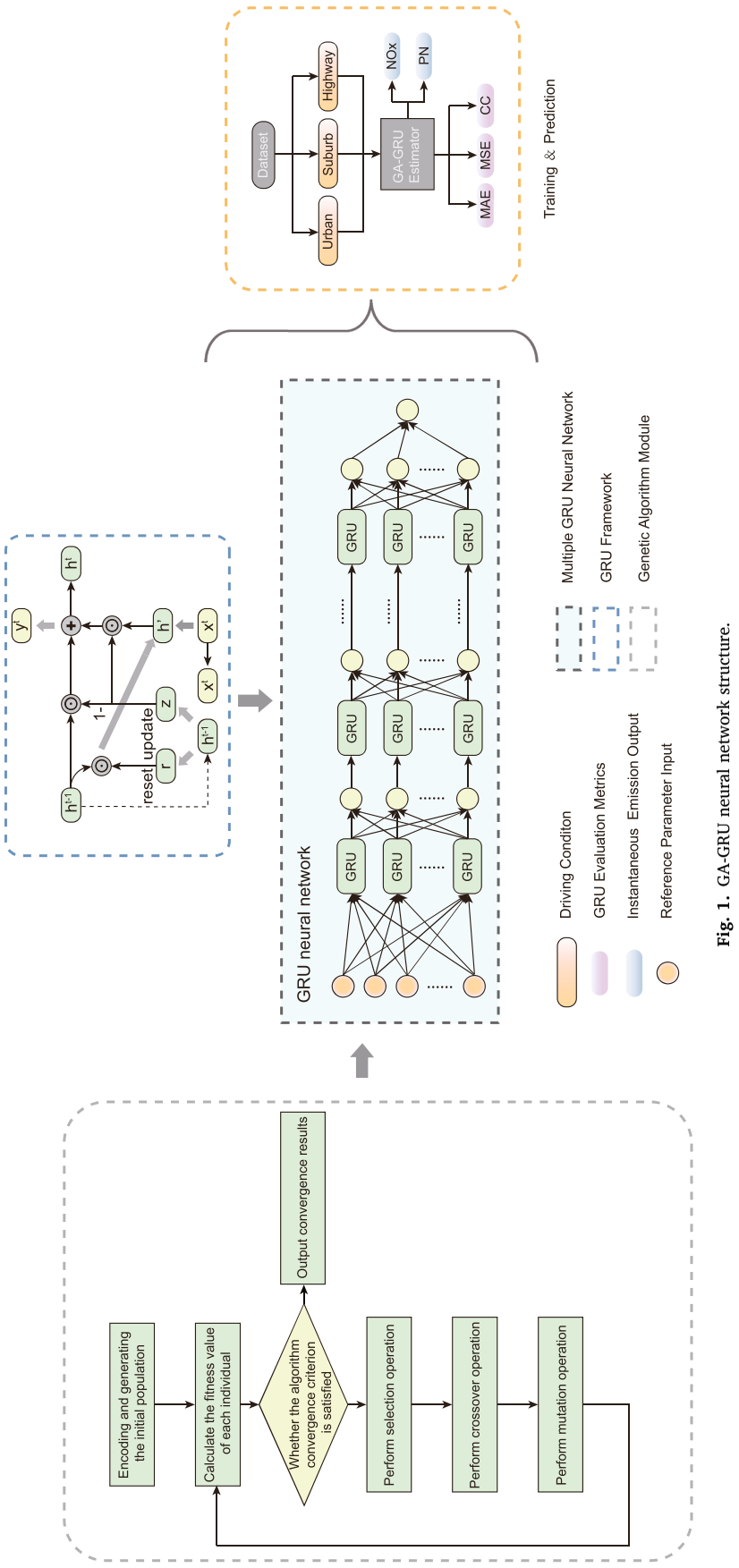

Fig. 1. GA-GRU neural network structure.

Table 1
Specifications of the test trucks.

Parameter	Value
Vehicle Category	Line-haul, N3
Curb weight	8800
Model Year	2021/03
Engine	6-cylinder in-line diesel engine
Rated Power	311 kW (1900 rpm)
Maximum Torque	2100 N m (1000 rpm–1400 rpm)
Capacity	10.45 L
Fuel System	Common rail (180 MPa)
Intake System	Turbocharged inter-cooled
Emission Control	EGR, DOC, DPF, SCR, ASC
Emission Standard	China-VI

measurement results.

2.2.3. Test procedure

All the PEMS tests in this paper were conducted in Tangshan, a port city in northern China. The test route was selected as part of the fixed route where the test vehicle normally transports cargo, and the route complies with the China-VI emission regulations, as shown in Fig. S3. The driving conditions are carried out in the order of urban, suburb, and highway conditions, and the switch of driving conditions shall be determined by the vehicle velocity. It is considered to switch to the suburb and highway conditions when the vehicle velocity exceeds 55 km/h and 75 km/h for the first time, respectively. According to the on-road emission measurement test protocols, the composition of driving conditions is determined by duration time, whose percentages of the urban, suburb, and highway conditions are 20 %, 25 %, and 55 %, respectively. In addition, the altitude difference between the start and the end point of each test was within 30 m, the maximum altitude of the test route was 149 m, and the minimum altitude was 56 m. The test route was determined in a relatively gentle area, and the altitude elevation changes did not exceed 600 m/100 km.

2.2.4. Data selection

To ensure as much as possible that the data from the corresponding model is more easily obtained from the vehicle, this test selects OBD data and corresponding on-road PN and NOx emissions as the original datasets for training and testing. OBD data can be collected through the vehicle's OBD data interface as input data for soft sensors, which is easy to obtain and accurate. Subsequently, soft sensors based on neural networks are embedded in the vehicle OBD, and vehicle emission data is quickly calculated using existing OBD data.

Due to excessive OBD data for HDDVs, to reduce the training and validation time, the CC filter data with strong correlation, and the CC is calculated as Eq. (9).

$$Pearson = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{t=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{9}$$

Where x is the NOx emission parameter and y is the other test parameters.

The CC of each parameter with NOx and PN emission data was calculated by Eq. (9), and the outcome is depicted in Fig. S4. Choose parameters with strong correlation as training parameters. For NOx, thirteen parameters have a strong correlation for training. They are engine coolant temp, vehicle speed, mass air flow rate, ambient air temp, engine oil temp, engine fuel rate, actual engine percent torque, calculated load value, catalyst temp bank sensor, DPF bank delta pressure, DPF bank outlet pressure, DPF bank inlet temp, and DPF bank outlet temp. And for PN, there are eight parameters strongly related to training. They are engine fuel rate, actual engine percent torque, tailpipe ambient absolute humidity, calculated load value, mass air flow rate, catalyst temp bank sensor, DPF bank outlet pressure, and DPF bank outlet temp.

2.2.5. Data processing

After acquiring pollutant emission data through PEMS tests, we employed it as the initial dataset. Utilizing Python 3.8 with Pandas and Numpy libraries, we performed data cleaning to eliminate missing values and outliers. Meanwhile, guided by the Pearson correlation coefficient, the necessary input data was determined. Due to the very large differences in the scale of the test data, the data are too large, making the actual differences between individual data too large for the neural network to learn the data effectively, resulting in poor performance of the neural network learning algorithm, so the initial data need to be normalized. In this paper, min-max normalization is chosen as Eq. (10).

$$y_i = \frac{x_i - \min_{1\le j\le n}\{x_j\}}{\max_{1\le j\le n}\{x_j\} - \min_{1\le j\le n}\{x_j\}} \tag{10}$$

Where: i is the number of parameters; j is the number of samples; x_i is the original parameter of the sample; $\max_{1\le j\le n}\{x_j\}$ is the maximum value of the interval parameter; $\min_{1\le j\le n}\{x_j\}$ is the minimum value of the interval parameter; y_i is the normalized value of the sample.

According to Eq. (10), the input and output parameters are mapped between [0, 1] to establish the normalization of the original data, reduce the numerical differences between the data, improve the learning ability of the neural network, make the neural network training better, and

Table 2
Description of test data statistics.

Driving Conditions	Sample Size	Parameter	Max	Min	Mean	Std
Urban	1748	NOx (ppm)	1128.00	0.01	176.42	265.04
		PN(#)	17123.44	0.00	1680.42	1998.07
		Speed (m/s)	15.82	0.00	8.88	4.55
		Acceleration (m/s^2)	1.13	−1.95	0.00	0.29
		Humidity (%)	56.77	34.49	44.07	7.31
		Temperature (°C)	5.79	1.07	3.23	1.31
Suburbs	2294	NOx (ppm)	1297.00	0.01	398.29	297.89
		PN(#)	16806.04	0.00	2750.47	2357.75
		Speed (m/s)	20.52	0.00	14.34	3.93
		Acceleration (m/s^2)	1.85	−1.71	0.01	0.28
		Humidity (%)	55.13	41.55	48.45	4.16
		Temperature (°C)	4.02	1.09	2.15	0.55
Highway	4979	NOx (ppm)	1612.00	0.01	607.96	504.82
		PN(#)	44531.64	0.00	8787.40	8054.18
		Speed (m/s)	22.79	1.41	19.62	3.32
		Acceleration (m/s^2)	1.48	−2.40	0.00	0.31
		Humidity (%)	50.60	37.30	43.23	2.75
		Temperature (°C)	3.97	1.69	2.72	0.47

improve the accuracy of the subsequent model prediction.

Since the actual test data of PEMS includes three driving conditions, the prediction of the original data is divided into three parts in this paper. The training set and test set are divided according to different driving conditions. The urban condition, suburb condition, and highway condition are divided respectively (Wang et al., 2022b), in which the first 80% of the data of each condition is used as the training set for the training of the neural network model. The last 20% is used as the test set to test the accuracy of the model. The urban, suburb, and highway conditions are analyzed according to the test results. By adopting the mentioned data preprocessing methods, the performance, robustness, and generalization capacity of the GA-GRU model can be enhanced. This ensures that the model is better equipped to accurately capture the fundamental characteristics of vehicle pollutant emission data, leading to more precise prediction results.

2.2.6. Data indicators

MSE is the squared difference of real values, which makes it relatively less sensitive to outliers and helps improve the model's robustness to noise and outliers in the data. MSE intuitively represents the square of the average difference between predicted and actual values. This makes the performance evaluation of the model easier to understand and interpret. For regression tasks, such as predicting continuous values, MSE is a common choice because it accurately quantifies the difference between predicted and true values.

Therefore, the MSE (Mean Square Error) is selected as the loss function M_{mse} in the training and evaluation process of this model (Glorot and Bengio, 2010), meanwhile the RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) are also chosen as the auxiliary metrics.

2.2.7. Contrast models

Additional neural network models and traditional machine learning models were selected for comparative analysis to assess the accuracy of pollutant predictions generated by our model, as outlined below.

(1). Back Propagation

BP (Back Propagation) neural network is a concept proposed by scientists led by Rumelhart and McClelland. It is a multi-layer feedforward neural network trained according to the error backpropagation algorithm and is one of the most widely used neural network models (Wang et al., 2022b).

(2). Supported Vector Machine

The concept of SVM (Supported Vector Machine): Support vector machines were first proposed by Corinna Cortes and Vapnik in 1995 (Cortes and Vapnik, 1995). SVM can perform nonlinear regression through kernel methods and is one of the common kernel learning methods (Zhong-jin, 2008).

(3). Stochastic Gradient Descent

SGD (Stochastic Gradient Descent) is an abbreviation for Stochastic Gradient Descent and is one of the commonly used optimization algorithms in deep learning (Zhang and Zhou, 2023). SGD is a gradient-based optimization algorithm used to update the parameters of deep neural networks. Its basic idea is to randomly select a small batch of samples in each iteration to calculate the gradient of the loss function and update the parameters with the gradient. This randomness makes the algorithm more robust, avoiding getting stuck in local minima, and the training speed is also faster.

(4). Gradient Boosting Decision Tree

GBDT (Gradient Boosting Decision Tree) is an iterative decision tree algorithm in Chinese (Natekin and Knoll, 2013). The algorithm consists of multiple decision trees, and the conclusions of all trees are accumulated to obtain the final answer. It was considered a strong generalization algorithm along with SVM at the beginning of its proposal.

(5). Long Short-Term Memory

LSTM is a type of time recurrent neural network specifically designed to solve the long-term dependency problem of general RNNs. All RNNs have a chain form of repetitive neural network models.

3. Result and discussion

3.1. Optimization of hyperparameters

This section provides further information regarding the hyperparameters of the models. The dataset is partitioned into a training set and a validation set based on different driving conditions, with 20% of the data allocated for urban conditions, 25% for suburb conditions, and 55% for highway conditions. The GA model is employed to acquire the optimal hyperparameters, which can be classified into three different groups based on different driving conditions, as presented in Table 3.

The fitting accuracy of a neural network tends to improve as the number of training generations increases, thanks to the network's flexible training capability. However, it is important to note that the neural network's excessive freedom can lead to overfitting, where the network becomes too closely tailored to the training set. This can result in a simultaneous increase in both training and validation loss. To achieve satisfactory generalization performance of the neural network model, it is imperative to employ regularization techniques that mitigate the risk of overfitting to the training set during the model training phase and identify the optimal number of training iterations for the model.

The utilization of a GA model has the potential to reduce the inefficiencies associated with creating from scratch, including temporal, hardware, and network resource wastage. It can be seen in the table that there are differences in hyperparameters observed between different driving conditions. This phenomenon can be attributed to the utilization of distinct training datasets, which leads to the identification of optimal hyperparameters for each specific set of driving conditions.

The figure assigned to the learning rate for various driving conditions remains constant at 1.00E-02. This is because deviating from this value, either by increasing or decreasing it, may result in the training process failing to converge toward the optimal spots. Based on varying driving conditions, disparities arise within the real driving data, such as urban conditions exhibiting a greater number of traffic lights and junctions. In contrast, highway conditions entail higher driving speeds. These factors contribute to the increased complexity of emission data.

Consequently, the simulation results of the GA model necessitate adaptation to the specific driving conditions. When comparing the conditions of suburbs, it is observed that the batch size for urban and highway conditions is greater, which suggests that when training urban and highway data, it is necessary to raise the batch size in order to

Table 3
Ideal hyperparameters for the GA-GRU model.

Emission	Driving Condition	Learning rate [l_r]	Number of neural network layers [L]	Batch size [b_c]	Number of hidden layer neurons [N]
NOx	Urban	1.00E-02	2	64	32
	Suburb	1.00E-02	3	16	32
	Highway	1.00E-02	4	128	20
PN	Urban	1.00E-02	2	64	32
	Suburb	1.00E-02	3	16	32
	Highway	1.00E-02	4	128	20

minimize the loss of accuracy. The Keras framework offers a callback system to mitigate the issue of overfitting. Within this mechanism, the Early Stopping function plays a crucial role by monitoring the validation loss. It effectively stops the training of the model when the validation loss reaches its minimum, so effectively preventing overfitting.

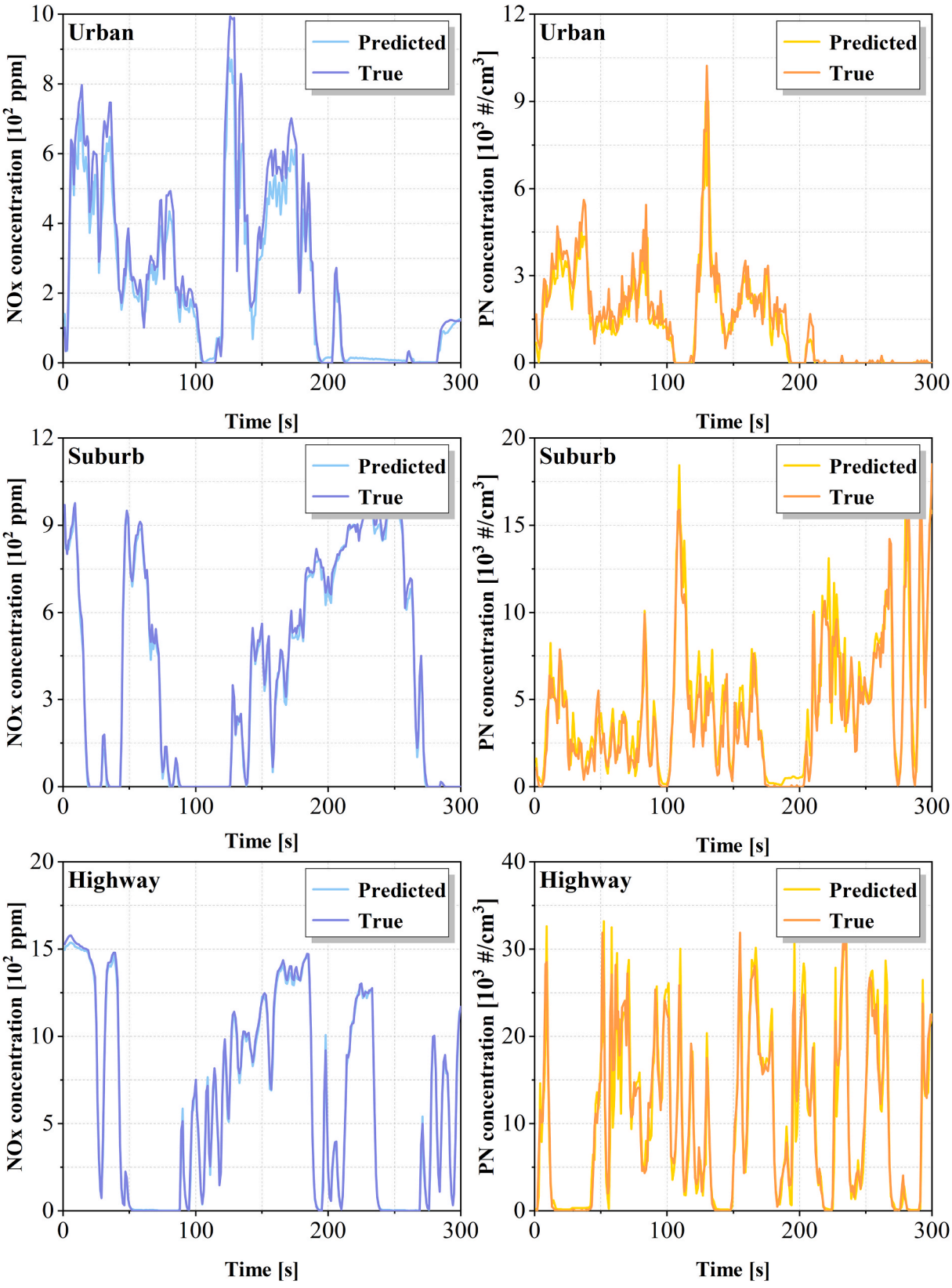

Fig. 2. Comparative analysis of the predicted and actual outcomes.

3.2. Driving condition prediction and comparison

As previously stated, the data was categorized into three distinct classifications based on driving conditions: urban, suburb, and highway conditions. Each cluster is then separated into a training set and a validation set. After training, the system can function as a soft sensor for real-time data monitoring. The comparison between the obtained outcomes and the original data is visually represented in Fig. 2.

There is a noticeable correspondence between the predicted data curve and the real data curve, with minimal divergence observed in the curves representing NOx and PN emissions. Indicates a strong level of agreement between the two datasets, and the two data sets exhibit a significant association.

Additionally, a few differences between the NOx emission prediction data and the PN emission can be observed. It may be concluded that the aftertreatment SCR of NOx is more stable for the aftertreatment DPF of PN. SCR offers a high degree of accuracy in lowering NOx emissions at the right temperature, and it exclusively targets NOx emissions found in exhaust gases. In contrast to SCR, DPF exhibits randomness in PN capture, and the degree of PN dilution in vehicle exhaust is unknown due to its randomness, while the degree of NOx dilution in the SCR is somewhat regular. As a result, the model's predicted outcome for NOx is superior to it for PN.

In the cold start period, the NOx emission is higher than in the warming period. The vehicle's emissions are almost identical to the original emissions since the SCR's temperature does not meet the catalyst temperature of NOx. The catalyst achieves its ideal temperature as the temperature rises, allowing SCR to function and lower its NOx emission to a reasonably low level. With vehicle test times increasing and driving conditions becoming drastic, NOx emissions are severe, so the NOx figure for suburbs and highways is much higher than that for urban areas. Regarding PN emissions, the amount of emissions is correlated with the amount of diesel injected when operating in various conditions, such as suburb and highway conditions. The increase in driving conditions can also lead to changes in fuel injection and incomplete combustion of the engine, leading to an increase in PN emissions.

To demonstrate the GA-GRU model's prediction accuracy and confirm the durability of the soft sensor that was later established. After predicting the validation dataset and training five additional models (discussed previously), the results were obtained to compare with our model, which is displayed in Table 4.

The RMSE values for GA-GRU in relation to NOx predicted emission and real emission are 0.0251, 0.0145, and 0.0192 under different driving conditions, while for PN emission, the corresponding values are 0.0185, 0.0299, and 0.0626 under different driving conditions. Evidently, the performance of GA-GRU models is significantly superior to that of other conventional models. Despite some discrepancies in some driving conditions, the overall level of GA-GRU remains relatively low when compared to LSTM models.

For NOx emission, the R^2 metrics for GA-GRU are 0.9915, 0.09983, and 0.9977 under urban, suburb, and highway conditions. For PN emission, the R^2 metrics for GA-GRU are 0.8933, 0.8899, and 0.9210 under urban, suburb, and highway conditions.

Currently, NOx emissions are the primary focus of neural network applications in emission prediction. According to the previous study (Wei et al., 2022), the NOx emission R^2 for the validation set is 0.84, whereas the R^2 for other vehicles is 0.74. In our study, the metric R^2 outcome is greater.

It is more difficult to predict PN emissions than NOx emissions, so there are few studies on PN prediction. In the previous study, the author conducted RDE (Real Driving Emission) tests on a light petrol vehicle, obtained 10646 data sets, and introduced four ensemble learning algorithms to construct a prediction model for instantaneous PN emissions. The model with the best prediction accuracy has an average R^2 of 0.832 under 10-fold cross-validation (Qiao et al., 2024). Compared to our research, the results of this paper are better than their conclusion. As mentioned above, the prediction of instantaneous PN for HDDVs under real driving conditions based on a neural network, the after-treatment of PN emission, DPF, has uncertain randomness, and due to the incomplete ignition occurring in the diesel engine randomly, not only the parameters of the diesel engine for PN emission have a lot of uncertainty, but the DPF filter rate of PN emission is also unknown. The future study will include more parameters and increase the amount of training to improve the performance of our model.

In terms of PN emission, it is harder to predict PN emission than NOx emission, so there are few studies of PN prediction. In an article, the author conducted RDE testing on a light gasoline vehicle, obtained 10646 sets of data, and introduced four ensemble learning algorithms to construct a prediction model for instantaneous PN emissions. The model that performs best in prediction accuracy has an average R^2 of 0.832 under 10-fold cross-validation (Qiao et al., 2024). Compared with our research, the outcomes of this paper are better than their conclusion. As mentioned previously, the prediction of instantaneous PN for HDDVs under real driving conditions based on neural network the after-treatment of PN emission, DPF, has uncertain randomness, and due to the incomplete ignition occurring in the diesel engine randomly, not only the parameters of the diesel engine for PN emission have plenty of uncertainty, but the DPF filter rate of PN emission is also unknown. The future study will involve more parameters and increase the amount of training to enhance the performance of our model.

Table 5 shows the GA-GRU model's enhanced rate compared with the other five models. Significantly increased accuracy on metrics was observed in the described model compared with others. Furthermore, in different driving conditions, the GA-GRU model also has a perfect performance in terms of prediction accuracy. Even the closest model, LSTM,

Table 4
Predicted performance for NOx and PN.

Model	RMSE			MSE			MAE			R^2		
	Urban	Suburb	Highway	Urban	Suburb	Highway	Urban	Suburb	Highway	Urban	Suburb	Highway
NOx emission												
GA-GRU	0.0251	0.0145	0.0192	0.0006	0.0002	0.0004	0.0179	0.0122	0.0130	0.9915	0.9983	0.9977
LSTM	0.0224	0.0176	0.0243	0.0005	0.0003	0.0006	0.0186	0.0142	0.0209	0.9880	0.9981	0.9904
SVM	0.1036	0.0883	0.1128	0.0107	0.0078	0.0127	0.0885	0.0736	0.0777	0.7208	0.9403	0.8908
BP	0.0736	0.0403	0.0938	0.0054	0.0016	0.0088	0.0564	0.0284	0.0595	0.8109	0.9737	0.9297
GBDT	0.0938	0.0604	0.1022	0.0088	0.0037	0.0104	0.0756	0.0418	0.0674	0.7342	0.9332	0.9093
SGD	0.1310	0.1187	0.1921	0.0172	0.0141	0.0369	0.1172	0.1096	0.1581	0.7628	0.9311	0.7065
PN emission												
GA-GRU	0.0185	0.0299	0.0626	0.0003	0.0009	0.0039	0.0155	0.0207	0.0436	0.8933	0.8899	0.9210
LSTM	0.0178	0.0306	0.0701	0.0003	0.0009	0.0049	0.0112	0.0218	0.0445	0.8696	0.8797	0.8961
SVM	0.0519	0.0839	0.1206	0.0027	0.0070	0.0145	0.0474	0.0627	0.0922	0.6144	0.2280	0.6622
BP	0.0357	0.0634	0.1288	0.0013	0.0040	0.0166	0.0290	0.0467	0.0903	0.4792	0.1687	0.6317
GBDT	0.0400	0.0744	0.1136	0.0016	0.0055	0.0129	0.0294	0.0543	0.0821	0.3529	0.2642	0.6983
SGD	0.0317	0.0599	0.1639	0.0010	0.0036	0.0269	0.0242	0.0483	0.1313	0.6029	0.2100	0.3801

Table 5
Accuracy error between GA-GRU and other models.

Model accuracy	LSTM		SVM		BP		GBDT		SGD	
	NOx	PN	NOx	PN	NOx	PN	NOx	PN	NOx	PN
Urban	0.18%	1.36%	14.70%	17.10%	9.60%	26.80%	13.90%	37.10%	12.30%	17.80%
Suburb	0.01%	0.58%	2.90%	49.40%	1.20%	56.50%	3.35%	45.50%	3.40%	51.40%
Highway	0.37%	1.38%	5.50%	15.20%	3.40%	17.20%	4.50%	12.90%	15.80%	37.80%

has a gap compared with the GA-GRU model.

For NOx emissions, the error enhancement rate is around 0.18%–14.70%, 0.01%–3.40%, and 0.37%–15.80% under urban, suburb, and highway conditions, respectively. For PN emissions, it is around 1.36%–37.1%, 0.58%–56.50%, and 1.38%–37.80% under urban, suburb, and highway conditions, respectively.

Compared to the described model, the predictive performance of the BP neural network is inferior, primarily attributed to its fundamental structure. The inherent limitations of the BP neural network may hinder its ability to capture relationships over longer time effectively spans in the time series (Zhang et al., 2023). Notably, the BP neural network is susceptible to being trapped in local minima, particularly under demanding driving conditions that involve frequent starts and stops in suburb driving conditions. In scenarios where intense vehicle driving conditions prevail, the BP neural network may exhibit relatively poor performance in time series processing tasks, underscoring the superior capabilities of the GA-GRU model in such contexts.

The strength of SVM lies in its potential to excel in sequence prediction tasks involving smaller datasets and linear relationships (Pourhosseini et al., 2023; Wang et al., 2022a). However, when confronted with large-scale datasets and nonlinear relationships, the performance of SVM may encounter limitations. Given that pollutant emission prediction data primarily comprises time series information, SVM struggles to adapt to such extensive datasets, leading to suboptimal results when compared to the GA-GRU model.

The GA-GRU model's performance and assessment metrics are comparable to those of the LSTM model. The close performance and evaluation metric alignment between the GA-GRU and LSTM models can be attributed to their shared architecture and sequence modeling capabilities (Islam and Hossain, 2021). Meanwhile, genetic algorithms possess the capability to optimize both the structure and parameters of neural networks (Udaybhanu and Mahendra Reddy, 2024), allowing them to adapt to the specific requirements of tasks. This adaptability makes GA-GRU more flexible in capturing intricate patterns within pollutant emission data, consequently enhancing its predictive performance. Furthermore, the described model is faster and has a faster response time for training or prediction. On the on-vehicle soft sensor, it is critical to provide the appropriate result promptly.

3.3. Model interpretation

Taylor diagrams were created to represent the intuitive performance of several neural network models to assess the models from various aspects. Taylor diagram was first proposed by Karl E. Taylor in 2001 (Hooshmand et al., 2005; Zhou et al., 2021). It concentrates on three main metrics in an equation: CC, RMSE, and standard deviation, shown in **Eq. (14)**.

$$M_{rmse}{}^2 = \sigma_f^2 + \sigma_r^2 - 2\sigma_f\sigma_r R \tag{14}$$

Where M_{rmse} is the root mean square error. σ_f and σ_r are deviations of predicted data and real data. R is the correlation coefficient.

Three indicators are plotted in a single chart in a Taylor diagram. Better model performance was shown by a larger angle between the model point and the Y-axis, which also showed a significant correlation between the predicted data and the real data. The model has a high quality that can capture vehicle NOx and PN emission data and predict its trend and variation. The closer the model point is to the center of the RMSE circle, the smaller the RMSE is (Xu and Han, 2020). The standard deviation scale line on the Taylor plot shows the ratio of the model standard deviation to the observed standard deviation. A reduced standard deviation model, on the other hand, denotes excellent stabilization and suggests that certain extreme values cannot exist when predicting the instantaneous emission. With the use of the Taylor chart, one may evaluate the performance of several models and determine which one performs best in terms of uncertainty, correlation, and standard deviation.

Fig. 3 illustrates the performance of various models compared under different driving conditions and indications. The GA-GRU model performs the same for NOx emissions in urban conditions under different driving conditions but exhibits a little variation in suburb conditions. The SGD model performs the poorest, with the BP, GBDT, and SVM models doing marginally worse. Moreover, the GA-GRU model point is nearer to the RMSE circle center in highway conditions than the LSTM model. The GA-GRU performs well for PN emissions in urban, suburb, and highway conditions as well. But when compared to the conclusions on NOx and PN emissions, the performance of PN emission is marginally worse than that of NOx emission, and in urban conditions, both NOx and PN emission are marginally worse than NOx emission, which is in line with the previous conclusion.

The PEMS test requires high-quality data to predict results accurately. Due to the principles of neural networks, outliers will significantly affect the output. In this study, the long-term memory GA-GRU is employed to make predictions, and the earlier results of the windowing method will affect the later results. Finding a way to reduce the outliers in the prediction data is therefore crucial. The violin plot is used in this section to illustrate the number of outliers in the performance of the different models, as shown in Fig. 4.

In the violin diagram, the black block indicates the overall distribution of predicted values, and the white points are the medium number. Upward and downward extension lines illustrate the maximum and minimum prediction values, which can be called outliers because these value points have a huge deviation from our original data. The body of the violin diagram indicates the distribution of prediction values. The more concentrated the data is near the median, the fewer so-called outliers are in the predicted values, indicating that the model performs better in data prediction.

Obviously, there are a few outliers under the GA-GRU and LSTM models, both in the three driving conditions for NOx emissions. However, significantly decreased outliers were observed in GA-GRU and LSTM compared with others. For PN emission, the outcomes of different models are the same as those for NOx prediction, and the performance of the GA-GRU and LSTM models is also better than other models. However, there is little discrepancy between NOx emission and PN emission. Some discrepancies are indicated in the plots for PN emissions between different driving conditions, such as that suburb emissions are higher than urban emissions and highway emissions are higher than suburb emissions.

The following could be the causes of this phenomenon: There is randomness in the DPF processing process for HDDVs. The engine performance of the car affects how much pollution it emits when it is moving at a low pace, like in an urban setting. The engine's combustion performance is greater in urban settings because of the engine's slower speed and lower working power. The speed of the vehicle grows along

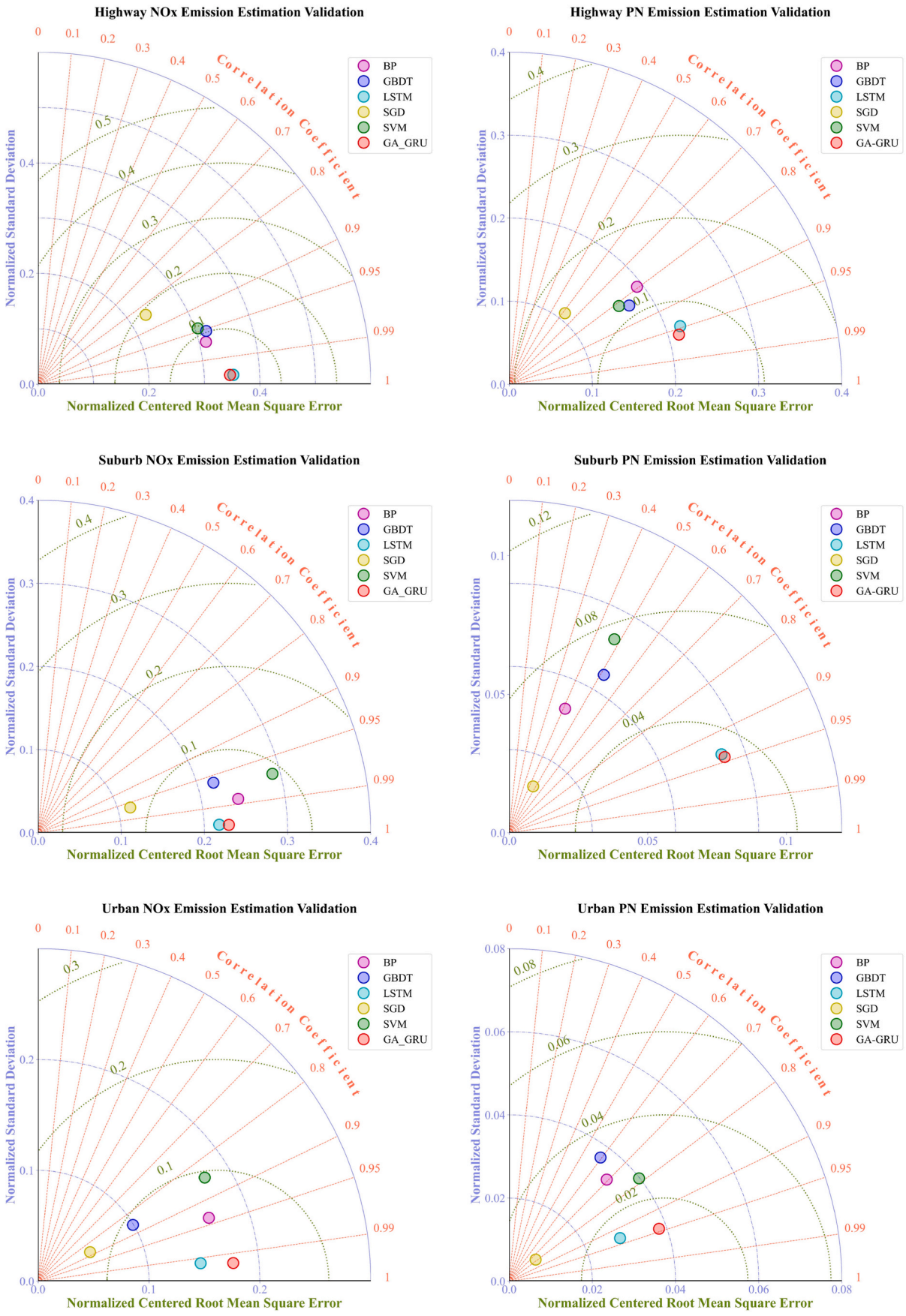

Fig. 3. Performance of different models by Taylor diagram.

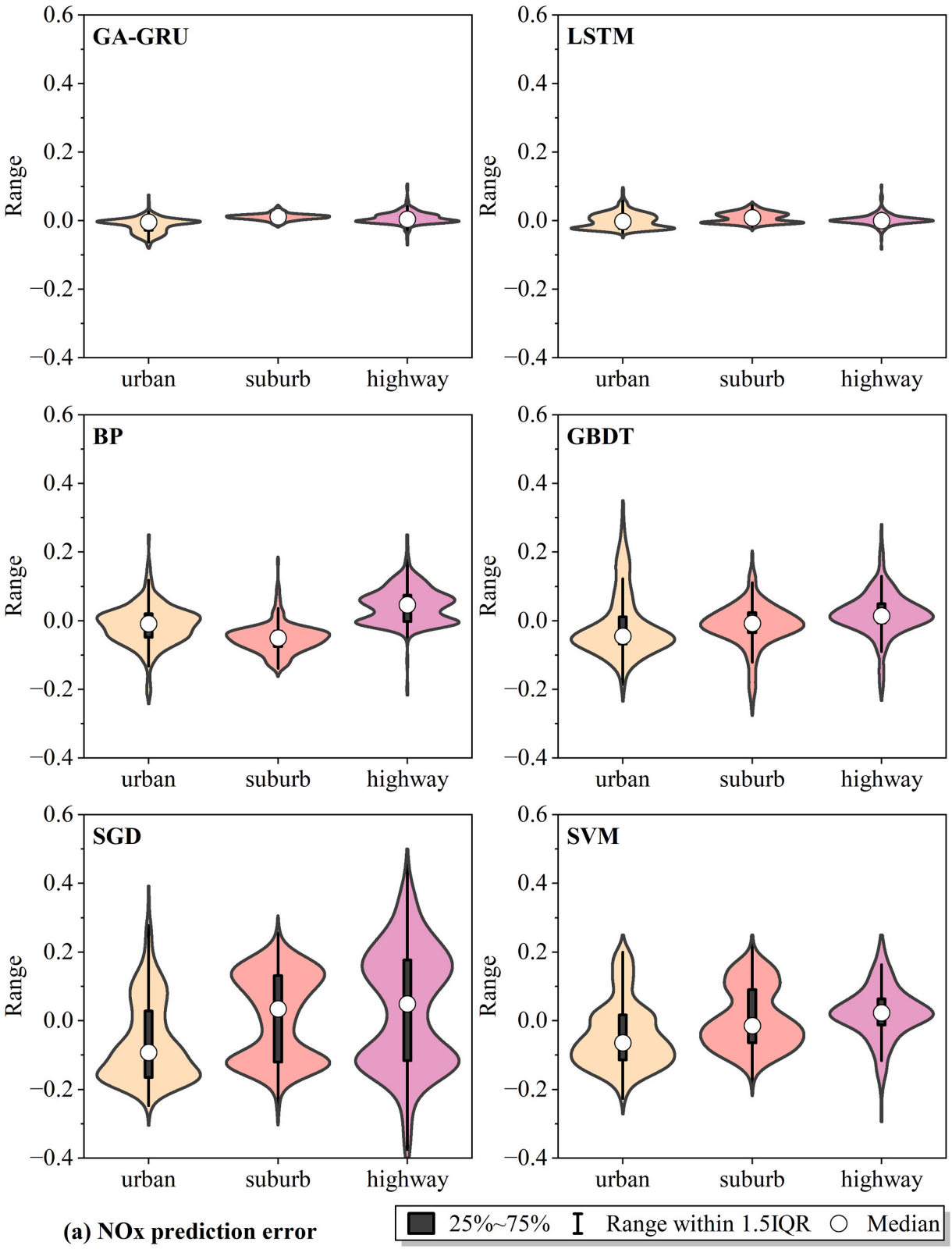

Fig. 4. Estimation error under different models and driving conditions. (a) the NOx prediction; (b) the PN prediction.

with its power and engine speed while it is operating in a drastic condition. Air inertia causes the fuel intake rate to increase at this point, which raises the engine's air flow rate and causes incomplete combustion. This raises the amount of PN that the engine produces, which raises the quantity of PN that enters the DPF after-treatment. As the driving conditions worsen, the increase in factors impacting the capture effectiveness of PN through DPF results in a higher deviation in the outliers provided by the soft sensor for PN emission detection. Compared to the

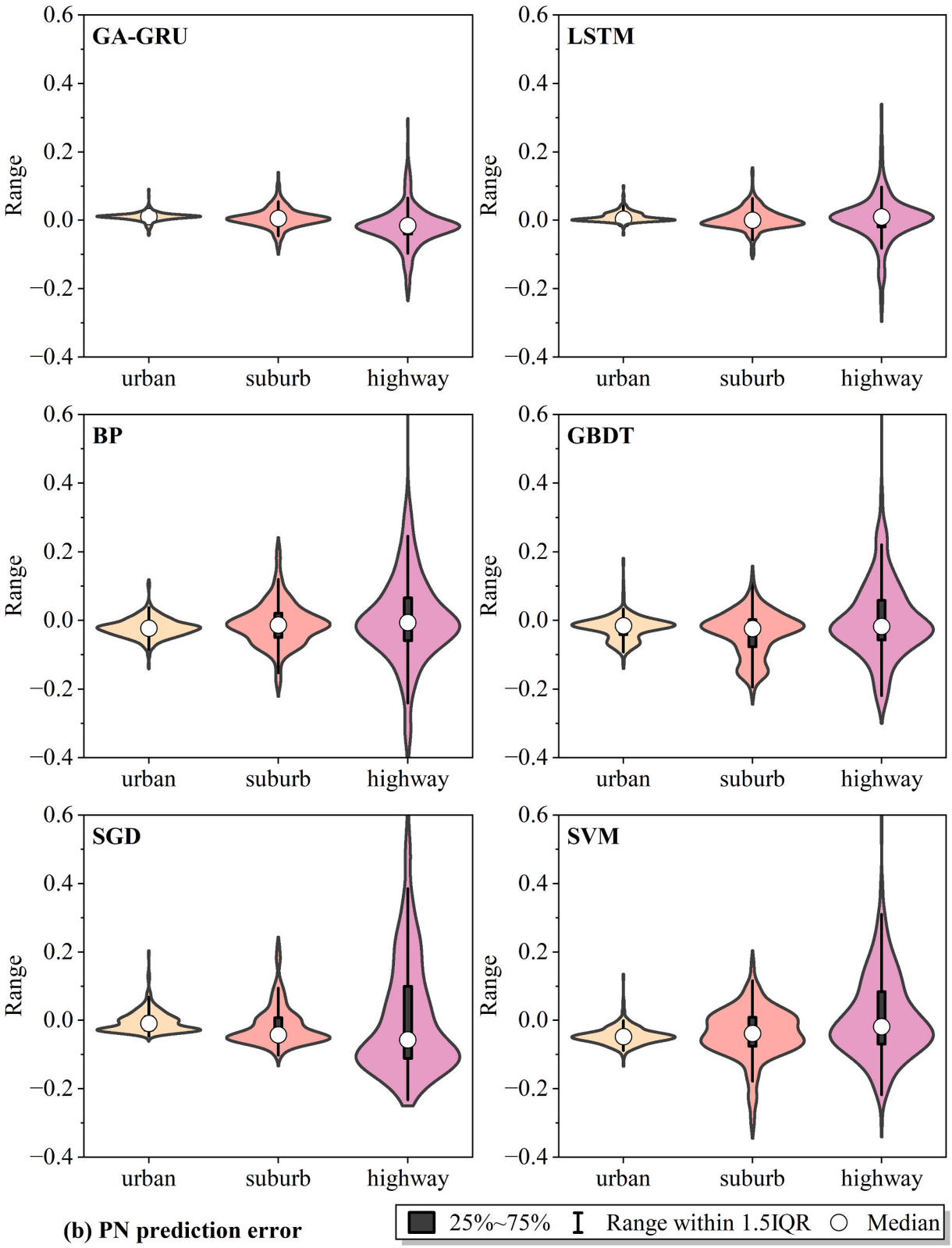

Fig. 4. (*continued*).

LSTM model, the proposed model can calculate faster and respond faster in a shorter time, making it suitable for integration into an OBD system.

3.4. Vehicle tracking

To track vehicle pollutant emission data throughout the entire lifecycle and then integrate soft sensors into the OBD system, in addition to the training set data, we also validated the model using a validation set tested in winter. However, the vehicle's entire life ought to be tracked, so additional datasets were created and tested over the summer of the following year.

Then the described model was used to predict the instantaneous

emissions of NOx and PN, and the results are shown in Fig. 5. The R^2 values for NOx emissions in suburb and highway conditions are somewhat higher than in urban conditions. The reason is that in urban conditions, cars must start their engines during what are known as "cool start stages," when the temperature is relatively low. It is these cold start phases that cause NOx emissions to increase significantly. Additionally, the after-treatment SCR's temperature is comparatively low at the cold start stage. Since the catalyst's ideal temperature has not yet been attained, the effectiveness of SCR is random and in an unpredictable state. However, under different conditions, the NOx projected metric R^2 performance is 0.9669, 0.9817, and 0.9877, respectively. When PN emission training results are compared to the original datasets, the R^2 values are a little better than the original results, which are 0.9089, 0.9260, and 0.9070 under different conditions, respectively. The ambient temperature can be determined as the cause. The relatively low temperatures seen throughout the winter will cause incomplete combustion in diesel engines. Furthermore, these variables may provide some uncertainty in neural network predictions. As a result, the predicted performance of PN emissions in the winter is somewhat more difficult than in the summer due to the influence of numerous unknown characteristics on the prediction's outcome.

With the vehicles driving, there will be some extreme figures during our test, and these figures can be classified as outliers, which can be seen in Fig. 5 in the high-emission areas. The failure or inaccuracy of the sensor may lead to the occurrence of abnormal values. The soft sensors' performance can be modified by removing these figures in our future

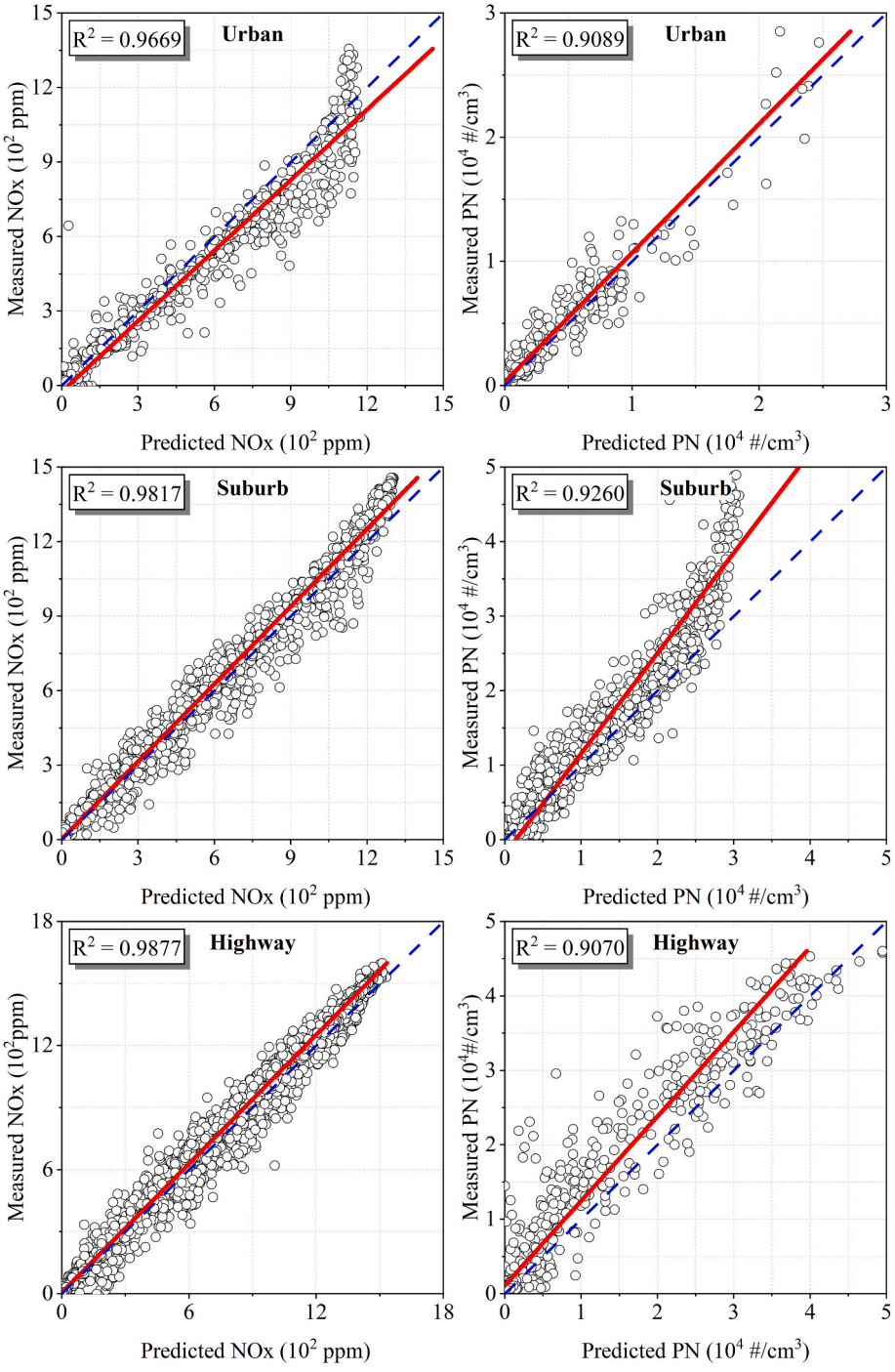

Fig. 5. Predicted outcomes at different times to analyze vehicle tracking.

studies.

In summary, compared with other conventional machine learning and neural network methods such as SVM, BP, GBDT, etc. In terms of predicting vehicle NOx and PN emissions, the GA-GRU model combines the characteristics of a genetic algorithm to optimize the model structure and parameters efficiently and automatically while fully utilizing the temporal modeling ability of GRU. Firstly, the GA-GRU model performs well in prediction accuracy, with lower performance indicators such as RMSE and R^2, allowing for more accurate capture of emissions measurements. Secondly, the GA-GRU model can better adapt to emission data from different vehicles and environmental conditions and has stronger generalization ability. Compared with traditional BP neural networks, GA-GRU overcomes the weakness of BP in dealing with gradient vanishing problems, improves training stability, and improves model performance. Finally, compared to the complex LSTM model, the GA-GRU model performs better in terms of computational efficiency, achieving faster inference times and providing better interpretable results. Therefore, the GA-GRU model has the advantage of comprehensive performance in predicting vehicle NOx and PN emissions, combining accuracy, generalization ability, computational efficiency, and interpretability, making it an excellent soft sensor model for predicting HDDVs' emissions.

4. Conclusion

This study presents a GA-GRU model as a soft sensor integrated into the OBD system whose objective is to predict NOx and PN emissions with the purpose of monitoring pollutant emissions and ensuring compliance with relative regulations. The selection of training datasets focused on OBD data, which was chosen due to its ease of acquisition and excellent accuracy in representing genuine working conditions. The primary benefits of the model outlined above can be summarized as follows: The application of a genetic algorithm is proposed to optimize the hyperparameters of the GRU model, aiming to achieve optimal performance while minimizing training time. The R^2 values for NOx emission in urban, suburb, and highway conditions are 0.9915, 0.9983, and 0.9977, respectively. These values have improved the accuracy by 0.18%–14.70%, 0.01%–3.40%, and 0.37%–15.80%, respectively. The R^2 values for PN emissions are 0.8933, 0.8899, and 0.9210 for urban, suburb, and highway conditions, respectively. These values indicate an improvement in accuracy ranging from 1.36% to 37.10%, 0.58%–56.50%, and 1.38%–37.80%, respectively. For both NOx and PN emissions, the metrics R^2 and RMSE demonstrate superior performance compared to the conventional models.

The soft sensor approach proposed in this study demonstrates a high level of reusability that is compatible with many types of vehicles and different driving conditions. At different times of testing, the performance of the R^2 metric for urban, suburb, and highway conditions are 0.9669, 0.9817, and 0.9877, respectively, for NOx emission. As for PN emission, the correspondent metrics R2 are 0.9089, 0.9260, and 0.9070, respectively, under different driving conditions. So, before utilizing the soft sensor to adapt to various driving conditions, it is only necessary to prepare the original datasets and establish a straightforward training process.

Subsequently, the trained model was integrated into the OBD system to facilitate its future utilization. The soft sensor is capable of monitoring the levels of NOx and PN emissions through the OBD system. It detects potential flaws and errors, collects emission data, and generates reports. Meanwhile, it provides alerts to the driver in appropriate scenarios, thereby ensuring that the vehicle emissions consistently comply with regulations' restrictions.

The described soft sensor also plays a crucial role in the control of vehicle emissions and the environment protection. For NOx emission, it can be applied on SCR to control the urea injection. In terms of PN, it can optimize the structure of DPF. The technique offers precise and up-to-date measurements of emissions, contributing to the mitigation of pollutant discharges, enhancement of fuel efficiency, cost reduction in maintenance, and the promotion of sustainability by minimizing negative environmental consequences.

In addition, the utilization of soft sensors not only offers to reduce the number of physical sensors required and minimize human labor, but it also has a further advantage of reducing the frequency of required RDE tests, hence contributing to a reduction in overall vehicle production costs. In future research, as emission regulations become even more stringent and new components of pollutant gases are introduced, additional pollutant gases will be incorporated into the soft sensors. It will further enhance the accuracy of pollutant prediction and integrate it into the OBD system.

CRediT authorship contribution statement

Luoshu Yang: Conceptualization, Methodology, Validation. **Yunshan Ge:** Funding acquisition, Resources, Validation. **Liqun Lyu:** Supervision, Visualization, Writing – review & editing. **Jianwei Tan:** Investigation, Resources. **Lijun Hao:** Investigation, Validation. **Xin Wang:** Writing – review & editing, Funding acquisition. **Hang Yin:** Data curation, Validation. **Junfang Wang:** Data curation, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

Funding: This work was supported by the National Key Research & Development Project of China (2022YFC3701802), the National Natural Science Foundation of China (52272342), and the Research and Demonstration of Key Technologies for Vehicle and Non-road Machinery Real Driving Test on Plateau (2019-GX-A6).

Abbreviations

Abbreviation	*Full name*
ASC	Ammonia Slip Catalyst
BP	Back Propagation
CLD	Chemiluminescent Detector
DOC	Diesel Oxidation Catalyst
DPF	Diesel Particulate Filter
EGR	Exhaust Gas Re-circulation
GRU	Gated Recurrent Unit
GA	Genetic Algorithm
GPS	Global Positioning System
GBDT	Gradient Boosting Decision Tree
HDDV	Heavy-Duty Diesel Vehicle
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MSE	Mean Square Error
NOx	Nitrogen Oxides
OBD	On-board Diagnosis
PLS	Partial Least Squares
PN	Particle Number
CC	Pearson Correlation Coefficient
PEMS	Portable Emission Measurement System
RDE	Real Driving Emission
RNN	Recurrent Neural Networks
RMSE	Root Mean Square Error

SCR	Selected Catalytic Reduction
SGD	Stochastic Gradient Descent
SVM	Supported Vector Machine

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2024.118190.

References

Andrade, P., Silva, I., Silva, M., Flores, T., Cassiano, J., Costa, D.G., 2022. A TinyML soft-sensor approach for low-cost detection and monitoring of vehicular emissions. Sensors 22. https://doi.org/10.3390/s22103838.

Bishop, G.A., Peddle, A.M., Stedman, D.H., Zhan, T., 2010. On-road emission measurements of reactive nitrogen compounds from three California cities. Environ. Sci. Technol. 44, 3616–3620. https://doi.org/10.1021/es903722p.

Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. https://arxiv.org/abs/1412.3555.

Cortes, C., Vapnik, V., 1995. SUPPORT-VECTOR networks. Mach. Learn. 20, 273–297. https://doi.org/10.1023/a:1022627411411.

Geller, V.D., Sardar, S.B., Phuleria, H., Fine, P.N., Sioutas, C., 2005. Measurements of particle number and mass concentrations and size distributions in a tunnel environment. Environ. Sci. Technol. 39, 8653–8663. https://doi.org/10.1021/es050360s.

Geron, A., 2019. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, second ed. O'Reilly Media.

Gholipour, A., Lucas, C., Araabi, B.N., Mirmomeni, M., Shafiee, M., 2007. Extracting the main patterns of natural time series for long-term neurofuzzy prediction. Neural Comput. Appl. 16, 383–393. https://doi.org/10.1007/s00521-006-0062-x.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. J. Mach. Learn. Res. 9, 249–256.

Guo, J., Chen, C., Wen, H., Cai, G., Liu, Y., 2024. Prediction model of goaf coal temperature based on PSO-GRU deep neural network. Case Stud. Therm. Eng. 53, 103813 https://doi.org/10.1016/j.csite.2023.103813.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hooshmand, A., Shamshiri, S., Alisafaee, M., Lotfi-Kamran, P., Naderi, M., Navabi, Z., Alizadeh, B., 2005. Ieee, binary taylor diagrams: an efficient implementation of taylor expansion diagrams. In: IEEE International Symposium on Circuits and Systems (ISCAS), pp. 424–427. Kobe, JAPAN.

Islam, M.S., Hossain, E., 2021. Foreign exchange currency rate prediction using a GRU-LSTM hybrid network. Soft Computing Letters 3, 100009. https://doi.org/10.1016/j.socl.2020.100009.

Jaworski, A., Kuszewski, H., Ustrzycki, A., Balawender, K., Lejda, K., Wos, P., 2018. Analysis of the repeatability of the exhaust pollutants emission research results for cold and hot starts under controlled driving cycle conditions. Environ. Sci. Pollut. Control Ser. 25, 17862–17877. https://doi.org/10.1007/s11356-018-1983-5.

Kontses, A., Ntziachristos, L., Zardini, A.A., Papadopoulos, G., Giechaskiel, B., 2020. Particulate emissions from L-Category vehicles towards Euro 5. Environ. Res. 182, 109071 https://doi.org/10.1016/j.envres.2019.109071.

Lei, Q., Wang, H., 2022. Noise-tolerant Co-trained semisupervised soft sensor model for industrial process. IEEE Sensor. J. 22, 19411–19423. https://doi.org/10.1109/jsen.2022.3201706.

Liu, J., Qiu, T., Penuelas, J., Sardans, J., Tan, W., Wei, X., Cui, Y., Cui, Q., Wu, C., Liu, L., Zhou, B., He, H., Fang, L., 2023a. Crop residue return sustains global soil ecological stoichiometry balance. Global Change Biol. 29, 2203–2226. https://doi.org/10.1111/gcb.16584.

Liu, J., Tetzlaff, D., Goldhammer, T., Wu, S., Soulsby, C., 2023b. Quantifying changes and trends of NO3 concentrations and concentration-discharge relationships in a complex, heavily managed, drought-sensitive river system. J. Hydrol. 622 https://doi.org/10.1016/j.jhydrol.2023.129750.

Liu, Y., Xie, M., 2020. Rebooting data-driven soft-sensors in process industries: a review of kernel methods. J. Process Control 89, 58–73. https://doi.org/10.1016/j.jprocont.2020.03.012.

Lyu, L., Ji, Z., Yin, H., Wang, J., Yang, W., Tan, J., Hao, L., Wang, X., Wang, H., Ge, Y., Wang, Y., 2023. NOx emission deterioration in modern heavy-duty diesel vehicles based on long-term real driving measurements. Environ. Res. 232, 116396 https://doi.org/10.1016/j.envres.2023.116396.

MEE, (Ministry of Ecology and Environment, P.R.C.), ASMR, (State Administration for Market Regulation, P.R.C.), 2018. Limits and Measurementmethods for Emissions Fromdiesel Fuelled Heavy-Dutyvehicles. CHINA VI), (17691-2018.

MEE, (Ministry of Ecology and Environment, P.R.C, 2022. China Mobile SourceEnvironmental ManagementAnnual Report, 2022.

Mei, C., Su, Y., Liu, G., Ding, Y., Liao, Z., 2017. Dynamic soft sensor development based on Gaussian mixture regression for fermentation processes. Chin. J. Chem. Eng. 25, 116–122. https://doi.org/10.1016/j.cjche.2016.07.005.

Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorob. 7 https://doi.org/10.3389/fnbot.2013.00021.

O'Driscoll, R., ApSimon, H.M., Oxley, T., Molden, N., Stettler, M.E.J., Thiyagarajah, A., 2016. A Portable Emissions Measurement System (PEMS) study of NOx and primary NO2 emissions from Euro 6 diesel passenger cars and comparison with COPERT emission factors. Atmos. Environ. 145, 81–91. https://doi.org/10.1016/j.atmosenv.2016.09.021.

Pourhosseini, F.A., Ebrahimi, K., Omid, M.H., 2023. Prediction of total dissolved solids, based on optimization of new hybrid SVM models. Eng. Appl. Artif. Intell. 126 https://doi.org/10.1016/j.engappai.2023.106780.

Qiao, P., Ni, J., Huang, R., Cheng, Z., 2024. Prediction of instantaneous particle number for light-duty gasoline vehicles under real driving conditions based on ensemble learning. J. Clean. Prod. 434, 139859 https://doi.org/10.1016/j.jclepro.2023.139859.

So, R.A., Andersen, Z.J., Chen, J., Stafoggia, M., de Hoogh, K., Katsouyanni, K., Vienneau, D., Rodopoulou, S., Samoli, E., Lim, Y.H., Amini, H., Cole-Hunter, T., Shahri, S.M.T., Maric, M., Bergmann, M., Liu, S., Azam, S., Loft, S., Westendorp, R.G. J., Mortensen, L.H., Bauwelinck, M., Klompmaker, J.O., Atkinson, R., Janssen, N.A. H., Oftedal, B., Renzi, M., Forastiere, F., Strak, M., Thygesen, L.C., Brunekreef, B., Hoek, G., Mehta, A.J., 2022. Long-term exposure to air pollution and mortality in a Danish nationwide administrative cohort study: beyond mortality from cardiopulmonary disease and lung cancer. Environ. Int. 164 https://doi.org/10.1016/j.envint.2022.107241.

Taghavifar, H., Taghavifar, H., Mardani, A., Mohebbi, A., Khalilarya, S., Jafarmadar, S., 2016. Appraisal of artificial neural networks to the emission analysis and prediction of CO 2 , soot, and NOx of n-heptane fueled engine. J. Clean. Prod. 1729–1739.

Tan, Q., Han, X., Zheng, M., Tjong, J., 2022. Neural network soft sensors for gasoline engine exhaust emission estimation. JOURNAL OF ENERGY RESOURCES TECHNOLOGY-TRANSACTIONS OF THE ASME 144. https://doi.org/10.1115/1.4052793.

Udaybhanu, G., Mahendra Reddy, V., 2024. A hybrid GA-ANN and correlation approach to developing a laminar burning velocity prediction model for isooctane/blends-air mixtures. Fuel 360, 130594. https://doi.org/10.1016/j.fuel.2023.130594.

Wang, H., Ji, C., Shi, C., Ge, Y., Meng, H., Yang, J., Chang, K., Wang, S., 2022a. Comparison and evaluation of advanced machine learning methods for performance and emissions prediction of a gasoline Wankel rotary engine. Energy 248, 123611. https://doi.org/10.1016/j.energy.2022.123611.

Wang, H., Ji, C., Shi, C., Yang, J., Wang, S., Ge, Y., Chang, K., Meng, H., Wang, X., 2023a. Multi-objective optimization of a hydrogen-fueled Wankel rotary engine based on machine learning and genetic algorithm. Energy 263, 125961. https://doi.org/10.1016/j.energy.2022.125961.

Wang, J., Wang, D., Zhang, F., Yoo, C., Liu, H., 2023b. Soft sensor for predicting indoor PM2.5 concentration in subway with adaptive boosting deep learning model. J. Hazard Mater. 465, 133074. https://doi.org/10.1016/j.jhazmat.2023.133074, 133074.

Wang, Z., Yang, Y., Huang, J., Wang, Y., Wei, L., Qin, W., 2022b. Numerical study of back-propagation suppression and intake loss in an air-breathing pulse detonation engine. Aero. Sci. Technol. 126 https://doi.org/10.1016/j.ast.2022.107566.

Wang, Z.X., He, Q.P., Wang, J., 2015. Comparison of variable selection methods for PLS-based soft sensor modeling. J. Process Control 26, 56–72. https://doi.org/10.1016/j.jprocont.2015.01.003.

Wei, N., Zhang, Q., Zhang, Y., Jin, J., Chang, J., Yang, Z., Ma, C., Jia, Z., Ren, C., Wu, L., Peng, J., Mao, H., 2022. Super-learner model realizes the transient prediction of CO_2 and NOx of diesel trucks: model development, evaluation and interpretation. Environ. Int. 158 https://doi.org/10.1016/j.envint.2021.106977.

Xie, H., Zhang, Y., He, Y., You, K., Fan, B., Yu, D., Li, M., 2019. Automatic and fast recognition of on-road high-emitting vehicles using an optical remote sensing system. Sensors 19. https://doi.org/10.3390/s19163540.

Xu, Z., Han, Y., 2020. Short communication comments on 'DISO: a rethink of Taylor diagram'. Int. J. Climatol. 40, 2506–2510. https://doi.org/10.1002/joc.6359.

Yan, W., Tang, D., Lin, Y., 2017. A data-driven soft sensor modeling method based on deep learning and its application. IEEE Trans. Ind. Electron. 64, 4237–4245. https://doi.org/10.1109/tie.2016.2622668.

Yang, X.-S., 2021. Chapter 6 - genetic algorithms. In: Yang, X.-S. (Ed.), Nature-Inspired Optimization Algorithms, second ed. Academic Press, pp. 91–100.

Yu, B., Pooi, C.K., Tan, K.M., Huang, S., Shi, X., Ng, H.Y., 2023. A novel long short-term memory artificial neural network (LSTM)-based soft-sensor to monitor and forecast wastewater treatment performance. J. Water Proc. Eng. 54 https://doi.org/10.1016/j.jwpe.2023.104041.

Zhang, F., Jin, T.Q., Xue, Z.G., Zhang, Y.H., 2022. Recent progress in three-dimensional flexible physical sensors. Int. J. Soc. Netw. Min. 13, 17–41. https://doi.org/10.1080/19475411.2022.2047827.

Zhang, L., Zhang, J., Ren, P., Ding, L., Hao, W., An, C., Xu, A., 2023. Analysis of energy consumption prediction for office buildings based on GA-BP and BP algorithm. Case Stud. Therm. Eng. 50 https://doi.org/10.1016/j.csite.2023.103445.

Zhang, Z., Zhou, S., 2023. Adaptive proximal SGD based on new estimating sequences for sparser ERM. Inf. Sci. 638, 118965 https://doi.org/10.1016/j.ins.2023.118965.

Zhong-jin, Y., 2008. Kernel-based support vector machines. Computer Engineering and Application 44 (1–6), 24.

Zhong, C., Gong, J.K., Wang, S.L., Tan, J.Q., Liu, J.A., Zhu, Y., Jia, G.H., 2021. NO2 catalytic formation, consumption, and efflux in various types of diesel particulate filter. Environ. Sci. Pollut. Control Ser. 28, 20034–20044. https://doi.org/10.1007/s11356-020-11870-1.

Zhou, J.-y., Yang, C.-h., Wang, X.-l., Cao, S.-y., 2023. A soft sensor modeling framework embedded with domain knowledge based on spatio-temporal deep LSTM for process industry. Eng. Appl. Artif. Intell. 126 https://doi.org/10.1016/j.engappai.2023.106847.

Zhou, Q., Chen, D., Hu, Z., Chen, X., 2021. Decompositions of Taylor diagram and DISO performance criteria. Int. J. Climatol. 41, 5726–5732. https://doi.org/10.1002/joc.7149.