Fuel xxx (xxxx) 121975

Contents lists available at ScienceDirect

Fuel

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

Full Length Article

An ensemble deep learning model for exhaust emissions prediction of heavy oil-fired boiler combustion

Zhezhe Han [a], Jian Li [a], Md. Moinul Hossain [b], Qi Qi [a], Biao Zhang [a], Chuanlong Xu [a]

[a] Key Laboratory of Energy Thermal Conversion and Control of Ministry of Education, School of Energy and Environment, Southeast University, Nanjing 210096, China
[b] School of Engineering, University of Kent, Canterbury, Kent CT2 7NT, UK

ARTICLE INFO

Keywords:
Emission prediction
Flame image
Ensemble method
Stacked denoising autoencoder
Gaussian process regression

ABSTRACT

Accurate and reliable prediction of exhaust emissions is crucial for combustion optimization control and environmental protection. This study proposes a novel ensemble deep learning model for exhaust emissions (NOx and CO_2) prediction. In this ensemble learning model, the stacked denoising autoencoder is established to extract the deep features of flame images. Four forecasting engines include artificial neural network, extreme learning machine, support vector machine and least squares support vector machine are employed for preliminary prediction of NOx and CO_2 emissions based on the extracted image deep features. After that, these preliminary predictions are combined by Gaussian process regression in a nonlinear manner to achieve a final prediction of the emissions. The effectiveness of the proposed ensemble deep learning model is evaluated through 4.2 MW heavy oil-fired boiler flame images. Experimental results suggest that the predictions are achieved from the four forecasting engines are inconsistent, however, an accurate prediction accuracy has been achieved through the proposed model. The proposed ensemble deep learning model not only provides accurate point prediction but also generates satisfactory confidence interval.

Nomenclature

Abbreviations

AE	autoencoder
ANN	artificial neural network
ARD	automatic correlation determination
CDAE	convolutional denoising autoencoder
DAE	denoising autoencoder
EF	empirical formula
ELM	extreme learning machine
FL	fuel load
GPR	Gaussian process regression
LSSVM	least square support vector machine
MAE	mean absolute error
MSE	mean square error
PCA	principal component analysis
RMSE	root mean square error
SAE	stacked autoencoder
SDAE	stacked denoising autoencoder
SVM	support vector machine

Symbols

x_i	the i^{th} original image
x_i	the i^{th} noisy image
z_i	the i^{th} reconstructed image
h_i	deep feature of the i^{th} flame image
y_i	the label of the i^{th} flame image
c	filter size
d	decoder
e	encoder
q	filter number
s	filter stride
f	activation function
H	the labeled image dataset
M	the number of labeled images
R^2	determination coefficient
L_{MSE}	mean square error loss function

Greek letters

$\mathcal{G}$	latent function
$\mathcal{Q}$	Gaussian process function
φ	corruption ratio
ξ	positive slack variable

https://doi.org/10.1016/j.fuel.2021.121975
Received 15 June 2021; Received in revised form 31 August 2021; Accepted 8 September 2021

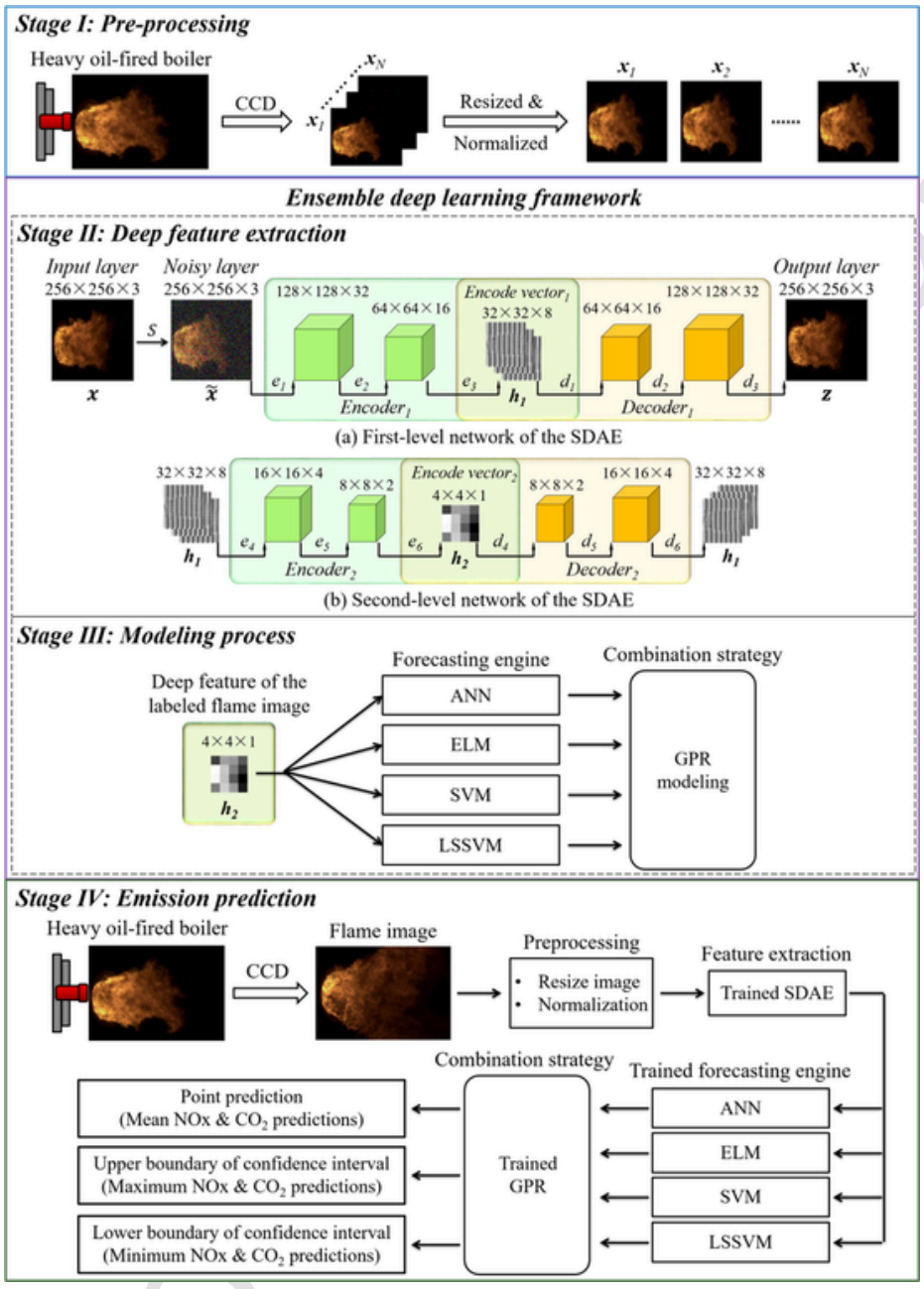

Fig. 1. The architecture for the emission prediction model.

$\mathscr{D}$	the number of testing samples
$\hbar$	hidden neuron
$\mathscr{P}$	forecasting engine
$\mathscr{C}$	regularized constant
σ	width parameter
ℓ	length scale
τ	signal variance

1. Introduction

Heavy oil is widely utilized as fuel for industrial boilers and marine engines due to its low cost, high energy density and abundant availability. However, during its combustion process, several pollutant emissions such as nitrogen oxide (NOx) and carbon dioxide (CO_2) are produced and exhausted. These pollutant emissions cause ozonosphere destruction and global warming, seriously endangering human health and the ecological environment [1]. Therefore, reducing or even eliminating pollutant emissions has become a critical and urgent matter. Many technologies are making efforts to reduce pollutants emissions and greenhouse gas, such as combustion optimization and flue gas treatment. However, regardless of applying emission reduction strategies, accurate and reliable measuring of the emission concentration of pollutants is an indispensable prerequisite.

Various methods have been proposed for emission monitoring, which can be categorized into instrument, simulation and data-driven methods. The instrument method, such as the continuous emission monitoring system (CEMS), is widely used in the boiler system [2]. Nevertheless, the CEMS is difficult to ensure economic and continuous measurement due to high-cost maintenance and frequent offline calibration. The simulation method can offer a clear explanation of the NOx and CO_2 formation process, and the computational fluid dynamics (CFD) analysis is commonly utilized to achieve emission prediction [3,

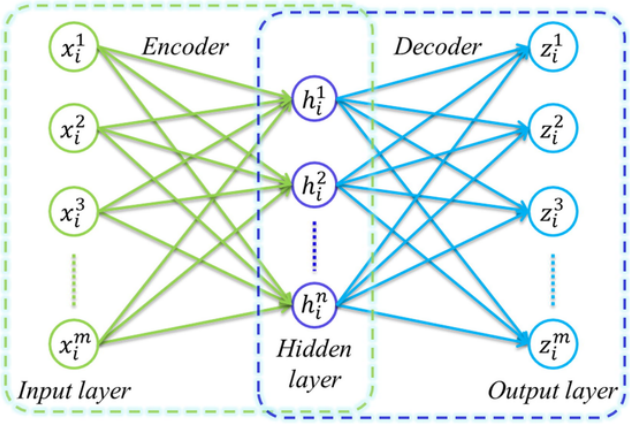

Fig. 2. The architecture of the AE.

4]. For instance, Kang et al. [5] and Park et al. [6] employed the CFD simulations on the combustion process of the heavy oil-fired boiler and achieved the accurate prediction of NOx concentration. However, due to complicated simulation analysis and high computation cost, the simulation method is hard to implement online measurement.

In contrast, the data-driven method is more convenient without exact equations, which may perform better than the standard model [7]. Currently, many artificial intelligence algorithms have been successfully utilized in emission prediction, including artificial neural network (ANN), support vector machine (SVM) and their variations. For example, Krzywanski et al. [8] established an ANN to predict the circulating fluid bed (CFB) boiler emissions using 8 operational variables, including the Ca:S molar ratio, the excess oxygen, flue gas recycle ratio, average riser temperature, average gas velocity in the riser, mean diameter of the coal particles, global O_2 concentration and inlet gas pressure. Zhou et al. [9] utilized the SVM to construct the nonlinear relationship between NOx emissions and 21 operational variables (such as primary air velocities, secondary air velocities, operation load and coal quality, etc.) on a 300 MW coal-fired boiler. Where the SVM provides a better prediction performance than traditional back-propagation and generalized regression neural networks. Tan et al. [10] introduced an extreme learning machine (ELM) to predict NOx emissions of a 660 MW coal-fired boiler using 31 operational variables, i.e., steam pressure, steam flow, steam temperature, rate of primary air and primary airflow rate. Yuan et al. [11] explored a stacked generalization ensemble method for NOx emissions prediction on a 600 MW subcritical variable pressure boiler based on 31 operational variables, i.e., unit load, airflow rate, pulverized coal feeder flow rate and overfire air damper opening percentage. Nevertheless, a major obstacle faced by these data-driven methods is the selection of input variables. As the core of any data-based modeling technique, reasonable input variables directly determine the prediction performance. A thorough review of the relevant literature discovered that the inputs of the prediction model are various, includes 29 types [12], 42 types [13], and even 96 types [14]. It is difficult to determine which operation variables are optimal, and blind selection can easily bring the risk of data redundancy to the prediction model.

As a consequence, selecting the appropriate input variables is very important for the performance of the emission prediction model. Recently, flame imaging-based model has been gradually attracted more attention and applied in exhaust emission prediction [15,16]. For instance, Li et al. [17] established a least square support vector machine (LSSVM) to analyze the features of flame radical image for NOx emission prediction. Disappointingly, the radical image acquisition carries

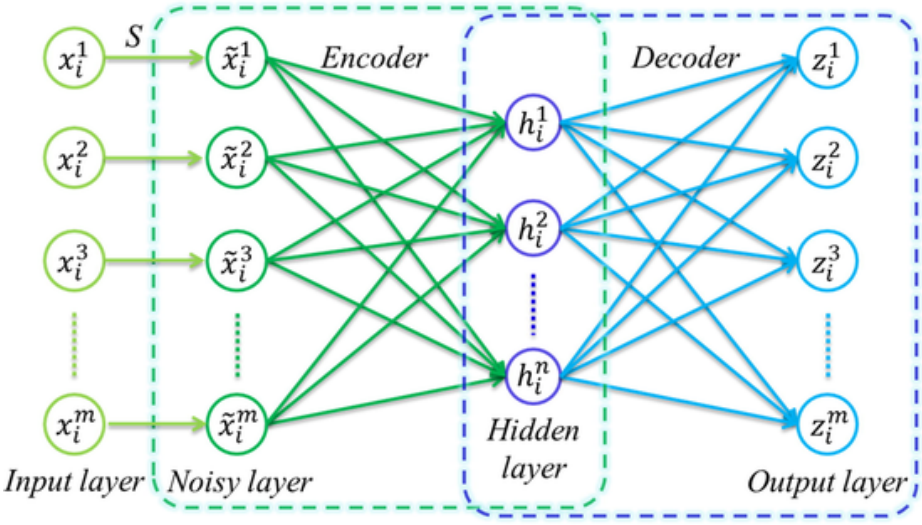

Fig. 3. The architecture of the DAE.

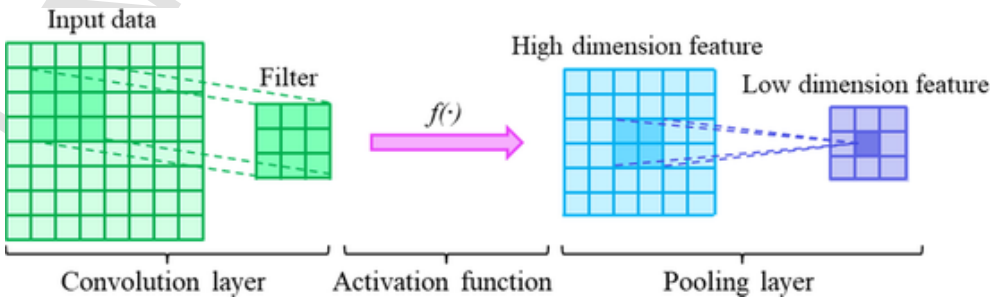

Fig. 4. Diagram of the convolution operation. $f(\bullet)$ represents activation function.

Table 1
The hyper-parameter setting of the SDAE.

Layer	Type	Upsampling	Convolution	Activation	Pooling	Output dimension
Input layer	x	/	/	/	/	256 × 256 × 3
Noisy layer	x	/	/	/	/	256 × 256 × 3
Encoder$_1$	e_1	/	q = 32; c = 3; s = 1	ReLU	c = 2; s = 2	128 × 128 × 32
	e_2	/	q = 16; c = 3; s = 1	ReLU	c = 2; s = 2	64 × 64 × 16
	e_3	/	q = 8; c = 3; s = 1	ReLU	c = 2; s = 2	32 × 32 × 8
Encode vector$_1$	$\boldsymbol{h}_1$	/	/	/	/	32 × 32 × 8
Decoder$_1$	d_1	c = 64	q = 16; c = 3; s = 1	ReLU	/	64 × 64 × 16
	d_2	c = 128	q = 32; c = 3; s = 1	ReLU	/	128 × 128 × 32
	d_3	c = 256	q = 3; c = 3; s = 1	Sigmoid	/	256 × 256 × 3
Output layer	z	/	/	/	/	256 × 256 × 3
Encoder$_2$	e_4	/	q = 4; c = 3; s = 1	ReLU	c = 2; s = 2	16 × 16 × 4
	e_5	/	q = 2; c = 3; s = 1	ReLU	c = 2; s = 2	8 × 8 × 2
	e_6	/	q = 1; c = 3; s = 1	ReLU	c = 2; s = 2	4 × 4 × 1
Encode vector$_2$	$\boldsymbol{h}_2$	/	/	/	/	4 × 4 × 1
Decoder$_2$	d_4	c = 8	q = 2; c = 3; s = 1	ReLU	/	8 × 8 × 2
	d_5	c = 16	q = 4; c = 3; s = 1	ReLU	/	16 × 16 × 4
	d_6	c = 32	q = 8; c = 3; s = 1	ReLU	/	32 × 32 × 8

considerable limitations, requiring an expensive electron multiplying charge coupled device (EMCCD) camera and specific fuel calibration. Comparatively, the red-greenblue (RGB) image is preferred because of its simple color camera and no fuel calibration. Remarkable achievements have been made in emission prediction based on the RGB image. For example, Wang et al. [18] extracted three statistical image features (average temperature, temperature distribution and ignition distance) of the RGB flame image to estimate the NOx emission in the pulverized coal boiler. However, some issues remain to be further considered:

(1) Image feature extraction, i.e., derives significant information from the flame image. Flames under different combustion operation conditions would generate delicate differences, i.e., size, shape, luminosity, etc. These differences should be precisely extracted from flame images for subsequent analysis. But existing feature extraction methods focus on hand-crafted features designed by unique empirical formulas, which may lose the key information. Besides, the tedious formula calculation will seriously cause the response lag of the prediction model. The deep learning methods, such as deep belief network (DBN) [19] and autoencoder (AE) [20], provide a good solution, which can directly mine intrinsic information of the original data through multi-layer nonlinear transformation. For instance, Liu et al. [21] proposed a DBN to extract deep features of flame images for predicting the oxygen (O_2) content of heavy oil combustion. This study demonstrated that the multi-layer DBN-based model learns more representative flame features, and the prediction performance is better than the classical principal component analysis (PCA)-based model. However, the DBN often encounters a thorny problem, that is, high-dimensional color images cause training difficulties [22]. In comparison to the DBN, the AE is a purely unsupervised feature learning model, which can be trained more effectively and easily [23]. Li et al. [24] designed an image feature extractor based on a deep denoising autoencoder to predict NOx emissions in the biomass combustion process. The results suggested that the deep learning-based prediction model outperforms the prediction models based on the hand-crafted feature extraction and basic learning algorithms. However, as the number of the hidden layers, this fully connected deep denoising autoencoder brings an unpredictable growth of the network parameters, resulting in a huge computational burden and inefficient prediction. Thus, an effective image feature extraction method based on the deep learning method still needs to be further explored.

(2) Confidence interval estimation, i.e., quantifying the uncertainty of the prediction result. Most prediction models only offer point prediction values but lack uncertainty information associated with point prediction. It is risky to formulate combustion management strategies purely based on the point prediction results. This requirement can be satisfied by the Gaussian process regression (GPR). Unlike point prediction models, the GPR is a probabilistic model that can provide a probability distribution for each predicted value [25]. However, an issue that cannot be ignored is that the computational complexity and the time-consuming of the GPR will increase the dimension of input variables dramatically. This severely restricts its application in solving the problem of high-dimensional input data.

Aiming at the shortcomings of existing methods, this study proposes an ensemble deep learning model for exhaust emissions prediction. A stacked denoising autoencoder (SDAE) is established to extract robust

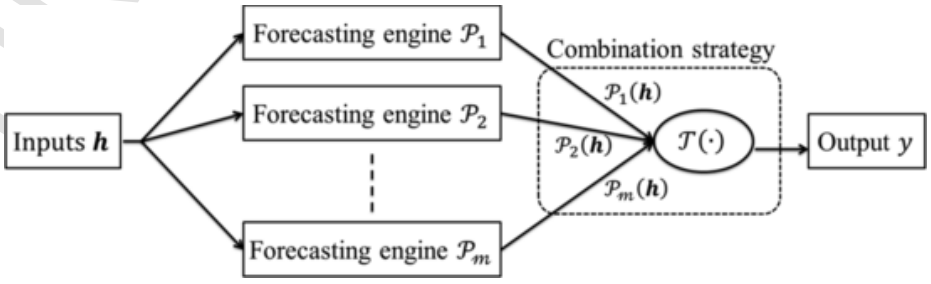

Fig. 5. Combination strategy design of the ensemble learning model.

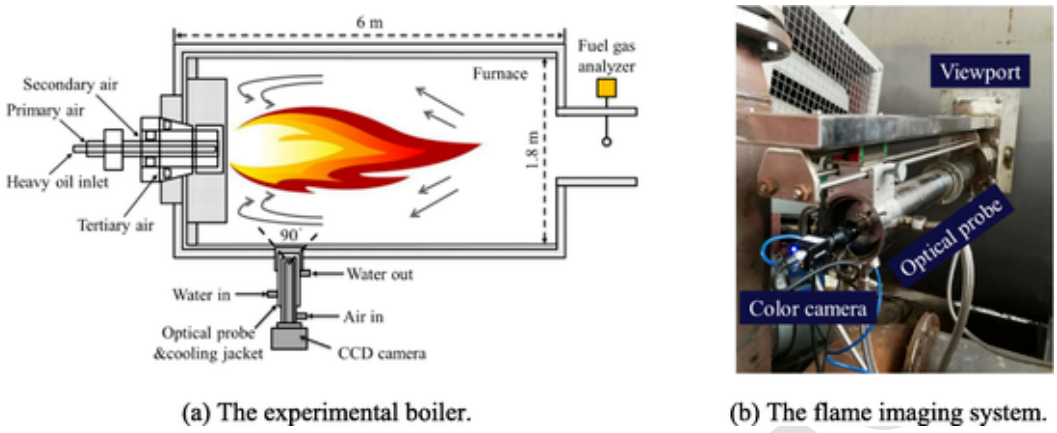

Fig. 6. Schematic diagram of the experimental apparatus.

Table 2
Ultimate analysis of the heavy oil.

Sample	w(C)%	w(H)%	w(N)%	w(S)%	Calorific value (kJ/g)
Reference value	82.00 ~ 84.50	12.00 ~ 14.00	0.10 ~ 0.50	0.10 ~ 0.40	44.00 ~ 47.00
Measured value	83.64	13.94	0.35	0.22	45.21

features of the flame image. The extracted image features are then fed into four forecasting engines, i.e., ANN, ELM, SVM, and LSSVM, to achieve an initial prediction of the emissions. Afterwards, the GPR combines the initial prediction in a nonlinear way to achieve the final point prediction and the confidence interval estimation. The principal contributions of this work can be summarized as follows: (1) attempted to integrate the flame imaging, deep neural network and probabilistic prediction model for NOx and CO2 emissions prediction; (2) established SDAE model to extract representative features from the flame image automatically. In particular, the denoising coding technique can improve the ability of feature learning; (3) deterministic forecasts made by the ANN, ELM, SVM and LSSVM are transformed into inputs and then combined by the GPR to obtain a probabilistic prediction. Therefore, the ensemble deep learning model fills the gaps in existing linear or simple weighting methods, providing better prediction performance; (4) the proposed prediction method also provides a valuable confidence interval that reflects the possible fluctuation range of the predicted value. The point prediction value is regarded as the mean of the confidence interval distribution.

The remainder of this article is organized as follows. Section 2 details the technical strategy of the emissions prediction based on the ensemble learning model. In Section 3, the experiment is carried out on a heavy oil-fired boiler to construct the flame image dataset. Section 4 investigates the advantages of the ensemble learning model and determines its vital hyper-parameter. The prediction performance of the ensemble learning model is presented in Section 5. Finally, the conclusion and future work from this study are given in Section 6.

2. Methodology

In this study, an ensemble deep learning model based on the flame image for emissions prediction is established. The ensemble learning model [26] can be regarded as a strong learner, which uses multiple weak learners to solve the same problem, and then takes special strategies to make the final decision combined with the opinions of all the weak learners. Fig. 1 presents the overall framework of the proposed emission prediction model, in which the SDAE is used for deep feature extraction of flame images, four forecasting engines include ANN, ELM, SVM and LSSVM are used as the weak learner, and the GPR is taken as the combination strategy model. The implementation process of the ensemble deep learning method is mainly divided into four stages and specified as follows:

Stage I: Pre-processing

A portable image acquisition system is constructed to capture flame images (denoted as $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ in Fig. 1) under different operation conditions of a heavy oil-fired boiler furnace. After that, all flame images are resized to 256 (H) × 256 (V) and normalized into an interval of 0 to 1.

Stage II: Deep feature extraction

A feature learning network (i.e., SDAE) is established and its parameters are randomly initialized. The SDAE is composed of a two-level network that can be trained only using unlabeled flame images. Once the unsupervised training is completed, it can be directly used to extract the representative feature $\boldsymbol{h}_2$ of the flame image.

Stage III: Modeling process

Four forecasting engines (i.e., ANN, ELM, SVM and LSSVM) are established using the extracted deep features and label information of the flame images. Afterwards, the outputs of these four forecasting engines are combined with the label information to train the GPR.

Stage IV: Emission prediction

In this stage, the newly captured flame images without labeling are pre-processed and their representative features are extracted by the trained SDAE and then successively fed into the four forecasting engines to achieve initial emissions prediction. The initial emissions predictions are then subsequently fed into the GPR for final emissions prediction. Especially, the final prediction results consist of point value (i.e., mean NOx & CO_2 predictions) and confidence interval (upper boundary and lower boundary), where the upper boundary represents the maximum NOx & CO_2 predictions, and the lower boundary represents the minimum NOx & CO_2 predictions.

2.1. Deep feature extraction

For any supervised learning network, the network performance depends heavily on the scale and quality of label data. However, it is difficult to prepare a massive high-quality label dataset, requiring precise experimental equipment and significant human effort. An appropriate solution is to apply the unsupervised learning network, which only requires unlabeled data. In this section, a novel unsupervised network SDAE is proposed and its basic components are described below:

Table 3
Overview of the flame image dataset.

Conditions	FL (%)	PA (%)	SA (%)	Total images
C1	20	10	25	1000
C2		20	25	
C3		30	25	
C4		20	20	
C5		20	30	
C6	40	5	35	1000
C7		10	35	
C8		20	35	
C9		30	35	
C10		20	50	
C11	60	10	50	1000
C12		20	50	
C13		30	50	
C14		20	35	
C15		20	65	

Table 4
Descriptive statistics of the NOx concentration.

Emission	FL (%)	Min. (mg/m^3)	Max. (mg/m^3)	Mean (mg/m^3)	St. dev.
NOx	20	274.42	299.76	288.69	7.56
	40	316.25	353.01	334.34	11.98
	60	358.53	374.76	356.39	5.14

2.1.1. Autoencoder

The AE is a typical unsupervised network consisting of three fully connected layers, i.e., input layer, hidden layer and output layer. As depicted in Fig. 2, the data processing of the AE is performed by the encoder and decoder. Suppose an unlabeled dataset with N samples $\{\boldsymbol{x}_i\}_{i=1}^{N}$, the encoder transforms each m-dimensional input sample $\boldsymbol{x}_i$ into a n-dimensional encode vector $\boldsymbol{h}_i$, typically $m>n$. Then, the decoder reconstructs the $\boldsymbol{h}_i$ to the$m$-dimensional output sample z_i.

2.1.2. Denoising autoencoder and convolutional operation

To extract the high-level features of the input sample, the AE usually uses multiple hidden layers. Unfortunately, as the number of hidden layers increases, the fully connected AE suffers from some weaknesses, such as the risk of over-fitting during the training process, and difficulty in training with numerous network parameters. To solve these drawbacks, two improvement strategies are utilized:

(1) Denoising coding technique is introduced into the AE to form the denoising autoencoder (DAE). As shown in Fig. 3, the input samples of the DAE are corrupted by random noises, and then the noise-free samples are reconstructed after encoding and decoding. Different from the AE, the DAE constructs a nonlinear mapping relationship between the reconstructed sample and the noisy sample, rather than the input sample. Because of this, the DAE is constrained during the training process, thus avoiding the simple replication of input information.

In Fig. 3, S represents stochastic mapping, and$\boldsymbol{x}_i$ represents the noisy sample with the dimension of m, calculated by:

$$\boldsymbol{x}_i = \boldsymbol{x}_i + \varphi\varsigma \tag{1}$$

whereφ is the corruption ratio; ς is a normal random distribution within a range of −2.576 to 2.576 with 99% probability. The DAE training is performed by minimizing the reconstruction error (known as loss function) between $\boldsymbol{x}_i$ and z_i. The mean square error (MSE) is often used as the loss function, defined as:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\|z_i - \boldsymbol{x}_i\|^2 \tag{2}$$

(2) Fully connected operation is replaced by convolutional operation. Convolution operation has the characteristics of sparse connectivity and weight sharing, effectively reducing the number of network parameters [27]. As illustrated in Fig. 4, the convolutional operation is composed of a convolutional layer, activation function and a pooling layer. In the convolution layer, the feature map is generated by sliding a specific filter over the entire input data. Then, the generated feature map is activated by the activation function to improve its nonlinear expression ability. The purpose of the pooling layer is to reduce data dimensionality and enhance translation invariance.

2.1.3. Stacked denoising autoencoder

For any deep neural network, the initialization of the weight parameters must be very cautious, otherwise, phenomena such as gradient disappearance or gradient explosion will be easily encountered during the training process. To solve this problem, a pre-train strategy [28] is widely adopted that can splits a multi-hidden layer network into several single-hidden layer networks. In this way, the training failure can be avoided because the weight parameters of each layer will be assigned a specific value instead of random initialization. However, due to the simplicity of single-hidden layer structure and the difficulty of global parameter optimization, this type of network often faces the problems of low training efficiency and poor generalization ability. Hence, this study adopts a compromise scheme and proposes a novel feature extraction network SDAE. It is composed of a two-level network, and each network contains five hidden layers. The overall architecture of the SDAE is presented in Fig. 1(*Stage II*), and the specific hyper-parameter setting is listed in Table 1.

In Table 1: *e*: encoder operation; *d*: decoder operation; *q*: filter number; *c*: filter size; *s*: filter stride.

In the first-level network, the noisy image $\boldsymbol{x}$ is created by adding Gaussian noise to the input flame image $\boldsymbol{x}$. Then, the $\boldsymbol{x}$ is processed by the convolution operation of the encoder e_1, which has 32 filters and each filter with a size of 3 × 3 and a stride of 1. The convolution filter with a size of 3 × 3 and a stride of 1 provides prominent performance of the model as described in [29,30]. Compared with other filter sizes (such as 5 × 5 and 7 × 7), the 3 × 3 convolution kernel used to obtain a feature map with the same dimension has the advantages of deeper network depth and fewer network parameters. The generated convolution feature is activated through the rectified linear unit (ReLU) function [$y(\hbar) = max(0, \hbar)$, $\hbar$ represents hidden neuron]. The ReLU function is widely utilized in deep neural networks due to its advantages of reducing the over-fitting risk. Notably, the ReLU function is replaced by the Sigmoid function [$y(\hbar) = 1/1 + \exp(-\hbar)$] in the d_3 (3th decoder) to ensure that the output range is 0 to 1. The max-pooling [31] is a widely used filter that provides the maximum value of the input data for dimensionality reduction. The pooling filter with a size of 2 × 2 and a strider of 2 can effectively obtain the key information of input data [32].

After the encoder e_1, the flame image $\boldsymbol{x}$ with the dimension of 256 × 256 × 3 is mapped to the feature data with the dimension of 128 × 128 × 32. Through a series of similar encoder operations, the flame image is finally condensed into the encoder vector $\boldsymbol{h}_1$ with the dimension of 32 × 32 × 8.

In the decoder d_1, the encoder vector $\boldsymbol{h}_1$ is dimensionally extended by the upsampling operation, and then processed successively by the convolution and activation. After that, the feature map with the dimension of 64 × 64 × 16 is obtained. Through a series of decoders, the flame image z with the dimension of 256 × 256 × 3 is reconstructed.

Table 5
Descriptive statistics of the CO_2 concentration.

Emission	FL (%)	Min. (%)	Max. (%)	Mean (%)	St. dev.
CO_2	20	6.62	7.52	7.18	0.27
	40	11.51	12.04	11.76	0.16
	60	12.04	13.24	12.54	0.41

In the second-level network, the encoder vector $\boldsymbol{h}_1$ is transformed into the input variable. After similar encoder operations, the encoder vector $\boldsymbol{h}_2$ with the dimension of $4 \times 4 \times 1$ is finally learned. To reduce the data dimension of the flame image, the number of convolution filters in the two-level encoder is gradually decreased, such as 32, 16, 8, 4, 2 and 1. These specific numbers are determined based on the Refs. [20, 27].

2.2. Forecasting engine

Emission prediction is a highly nonlinear regression task. Therefore, various regression models have been applied in this field [11,25]. In this study, four forecasting engines such as ANN, ELM, SVM and LSSVM are selected due to their high fitting accuracy, fast learning speed, strong learning ability and outstanding generalization ability.

2.2.1. Artificial neural network

ANN [33] is a multi-layer network, composed of an input layer, one or more hidden layers, and an output layer. The ANN with a single hidden layer structure has been proved to be sufficient to approximate complicated nonlinear functions [34], so the number of hidden layers is set to 1. Based on this, a three-layer ANN is considered for emissions prediction. The 16-dimensional image feature is fed into the input layer, so the number of input neurons is set to 16. The number of neurons in the hidden layer is set to 8, which is determined via cross-validation in comparison to other values (varies from 1 to 16). The emission (NOx or CO_2) concentration is taken as the output target, so the number of output neurons is set to 1. The ANN consists of two learning processes, i.e., transforming the input image features forward to the output emission concentration and back-propagating the prediction error to tune the weight parameters. These two learning processes operate in an alternating pattern until the prediction error converges.

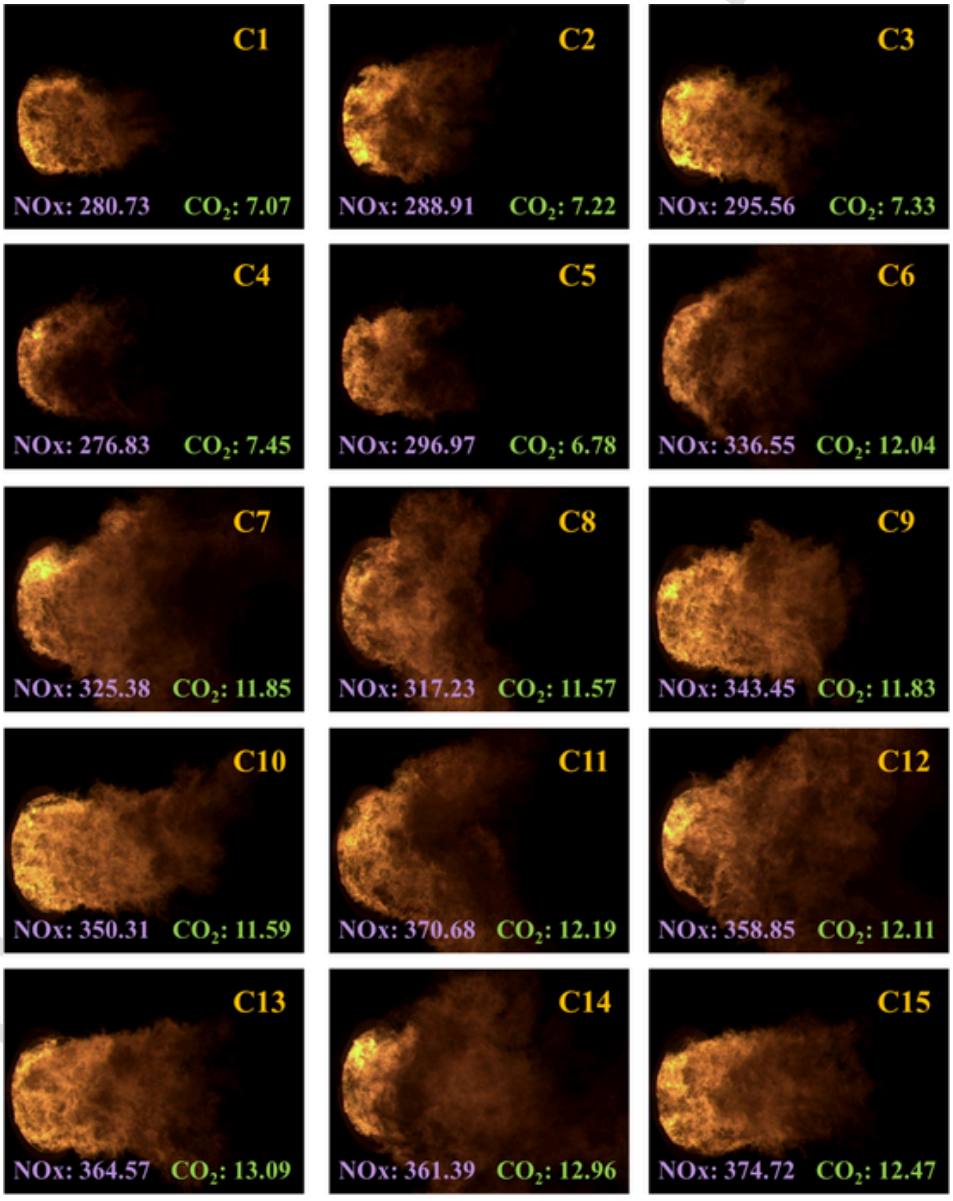

Fig. 7. Example of flame images with NOx concentration (mg/m^3) and CO_2 concentration (%).

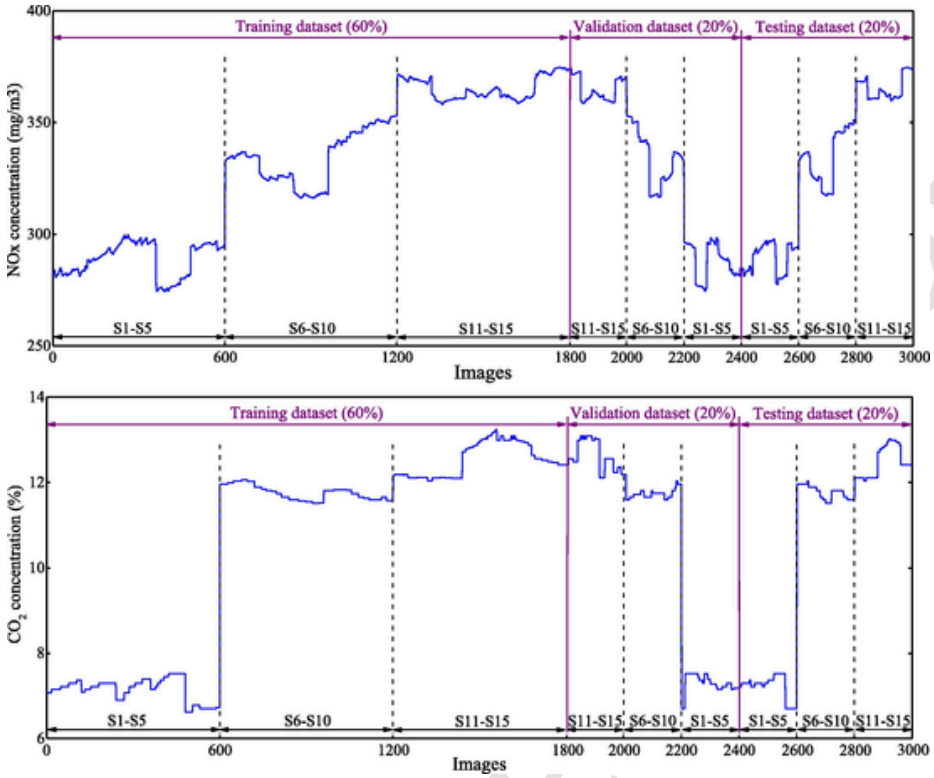

Fig. 8. Labeled images under three different fuel loads: 20%, 40% and 60%.

2.2.2. Extreme learning machine

ELM [35] is a feed-forward network with the advantages of fast learning speed. Similar to the ANN, the ELM is also designed as a single hidden layer structure, which consists of an input layer, a single hidden layer, and an output layer. Suppose a training dataset with M labeled images $\{\boldsymbol{h}_i, y_i | i = 1, 2, \cdots, M\}$, where $\boldsymbol{h}_i$ represents the i^{th} image feature with the dimension of n ($n = 16$); $\boldsymbol{y}_i$represents the corresponding label information (i.e., NOx or CO_2 concentration) with the dimension of k ($k = 1$). The number of neurons in the hidden layer is set to 8, determined via cross-validation compared to other values (varies from 1 to 16). Then, the ELM with g ($g = 8$) hidden neurons performs a zero-error approximation on all training samples, expressed as:

$$\boldsymbol{y}_i^k = \sum_{v=1}^{v=g} \boldsymbol{L}_i^v f\left(\boldsymbol{w}_i^{uv} \boldsymbol{h}_i^u + \boldsymbol{b}_i^v\right) \tag{3}$$

where $\boldsymbol{L}_i^v$ is the output weights from the v^{th} hidden neuron to the k^{th} output neuron; $f(\bullet)$ is the activation function; $\boldsymbol{w}_i^{uv}$ is the input weights from the input u^{th} input neuron and the v^{th} hidden neuron; $\boldsymbol{b}_i^v$ is the bias of the hidden layer. The input weights and biases of the ELM are assigned arbitrarily instead of tuned carefully, while only output weights are determined analytically through generalized inverse operation [36].

2.2.3. Support vector machine

SVM [37] is a statistical learning algorithm based on the principle of structural risk minimization. It performs well for dealing with small samples, high dimensions and local minimum points. Assuming a training dataset $\{H, Y\}$ with M labeled images, $H = \{\boldsymbol{h}_i\}_{i=1}^M$, $Y = \{\boldsymbol{y}_i\}_{i=1}^M$. The main idea of the SVM is to map the input data into a higher dimensional feature space and perform linear regression. In the feature space, the linear regression function can be defined as:

$$\psi(H) = \boldsymbol{w}\phi(H) + b \tag{4}$$

where $\boldsymbol{w}$ is weight vector; b is a bias; $\phi(\bullet)$ represents the nonlinear function mapping input data from the low-dimensional original space to the high-dimensional feature space. Based on the principle of structural risk minimization, the learning process of the SVM can be expressed as seeking a constrained minimum value, defined as:

$$\underset{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*}{Minimize} \frac{1}{2}\|\boldsymbol{w}\|^2 + \mathscr{C}\sum_{i=1}^{i=M}\left(\xi_i + \xi_i^*\right) \tag{5}$$

subject to:

$$\begin{cases} \left(\boldsymbol{w}\phi\left(\boldsymbol{h}_i\right) + b\right) - y_i \le \varepsilon + \xi_i \\ y_i - \left(\boldsymbol{w}\phi\left(\boldsymbol{h}_i\right) + b\right) \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{6}$$

where $\|\bullet\|^2$ is the *L-2* norm; $\mathscr{C}$ is regularized constant;ε is the parameter of theε-intensive loss function; ξ_i and ξ_i^* are positive slack variables, measuring the deviation outside the $[\varepsilon, -\varepsilon]$ region. By introducing the Lagrange equation and the Karush–Kuhn–Tucker (KKT) condition [38], Eq. (4) can be transformed into:

$$\psi(H) = \sum_{i=1}^{i=M}\left(\lambda_i - \lambda_i^*\right) k\left(\boldsymbol{h}_i, \boldsymbol{h}\right) + b \tag{7}$$

where λ_i and λ_i^* are Lagrange multipliers, satisfying $\sum_{i=1}^{i=M}\left(\lambda_i - \lambda_i^*\right) = 0$ $(\lambda_i \ge 0,\ \lambda_i^* \ge 0)$; $k\left(\boldsymbol{h}_i, \boldsymbol{h}\right)$ is the kernel function. Gaussian radial basis functions (RBF) [39] can produce complex decision boundaries with only one unknown parameter, so it is chosen as the kernel function and defined as:

$$k\left(\boldsymbol{h}_i, \boldsymbol{h}\right) = exp\left(-\frac{\left\|\boldsymbol{h}_i - \boldsymbol{h}\right\|^2}{2\sigma^2}\right) \tag{8}$$

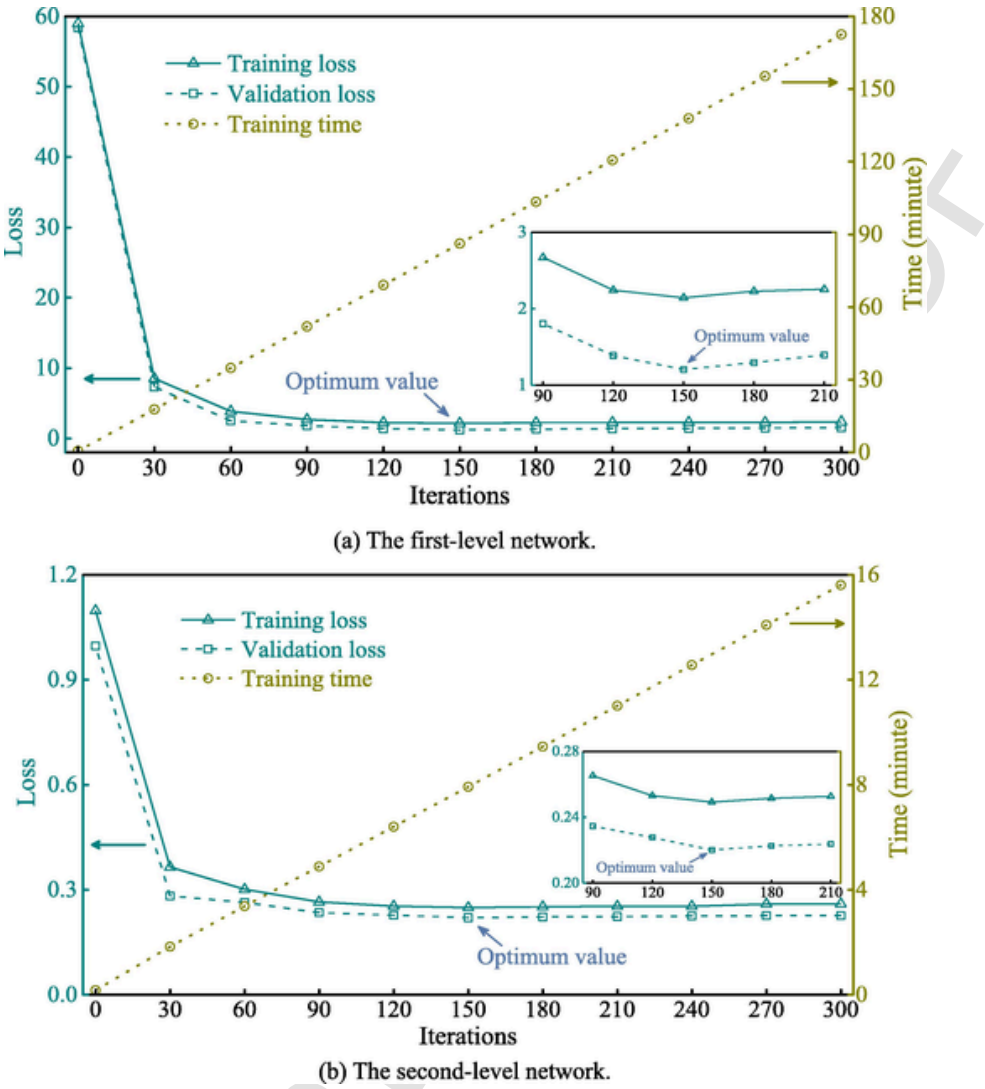

Fig. 9. Training progress of the SDAE under different iterations.

where σ is the width parameter of the Gaussian kernel. To construct the SVM with superior performance, the proper values of the model coefficients $\mathscr{C}$, ε and σ are necessary. In this study, the grid search, and cross-validation methods [11] are implemented to determine the optimal coefficients, resulting in $\mathscr{C} = 10$, $\varepsilon = 3.56$, and $\sigma = 0.25$.

2.2.4. Least square support vector machine

LSSVM [40] is a modified version of the SVM, which converts inequality constraints into equality constraints. Compared with the SVM, the LSSVM has advantages in good generalization and low computation complexity. For the given training dataset $\{\boldsymbol{h}_i, y_i | i = 1, 2, \cdots, M\}$, the LSSVM approximates the expected output through the following multivariate function:

$$y_i = \boldsymbol{w}^T \phi(\boldsymbol{h}_i) + b \tag{9}$$

where $\phi(\bullet)$ denotes nonlinear mapping function; coefficients $\boldsymbol{w}$ and b are estimated by solving the following optimization problem:

$$\underset{\boldsymbol{w},b,e}{Minimize} J(\boldsymbol{w}, e) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{\mathscr{C}}{2}\sum_{i=1}^{i=M} e_i^2 \tag{10}$$

subject to:

$$\boldsymbol{w}^T \phi(\boldsymbol{h}_i) + b = y_i - e_i \tag{11}$$

where $J(\boldsymbol{w}, \boldsymbol{e})$ is the structural risk; e_i is the i^{th} error variable. Like the SVM, the RBF function is also used as the kernel function of the LSSVM model. For the construction of the LSSVM model with prominent performance, the appropriate regularized constant $\mathscr{C}$ and RBF kernel parameter σ should be selected. These two hyper-parameters are finally set to 20 and 0.38 by the trial-and-error.

2.3. Combination strategy by Gaussian process regression

Since the single forecasting engine is limited by its architecture, hyper-parameters, and training data quality, it is difficult to achieve satisfactory prediction accuracy [11]. Alternatively, the ensemble learning model incorporates the advantages of each forecasting engine by adopting a special combination strategy, thus significantly improving the prediction performance [41]. The combination strategy design of the ensemble learning model is shown in Fig. 5.

In Fig. 5, $\mathscr{P}_1, \mathscr{P}_2, \cdots, \mathscr{P}_m$ represent forecasting engine (also known as base model); $\boldsymbol{h}$ represents inputs, i.e., image feature; $\mathscr{P}_m(\boldsymbol{h})$ represents the output of the forecasting engine; $\mathscr{T}(\bullet)$ represents the combination strategy model (also known as meta-model); y represents the output, i.e., NOx or CO_2 concentration. Generally, the combination strategy can

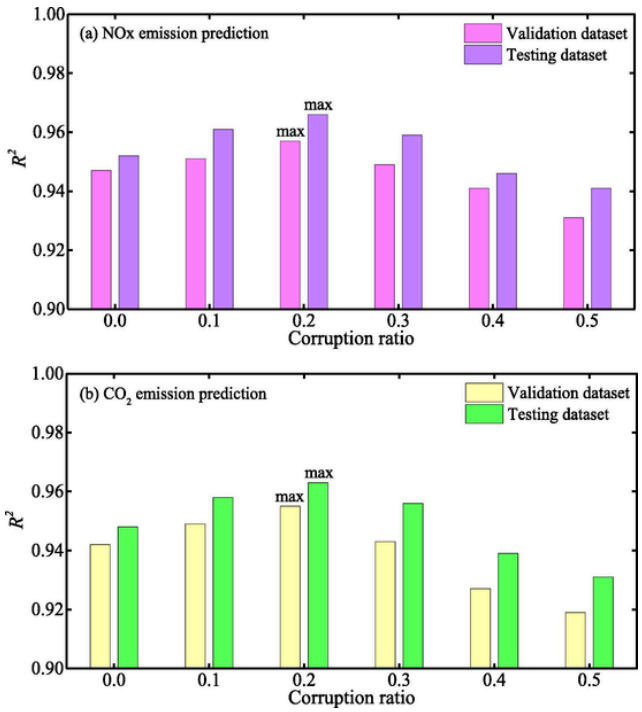

Fig. 10. Prediction performance of the ensemble learning model under different corruption ratios. The top subplot shows the prediction accuracy of NOx emission, and the bottom subplot shows the prediction accuracy of CO_2 emission.

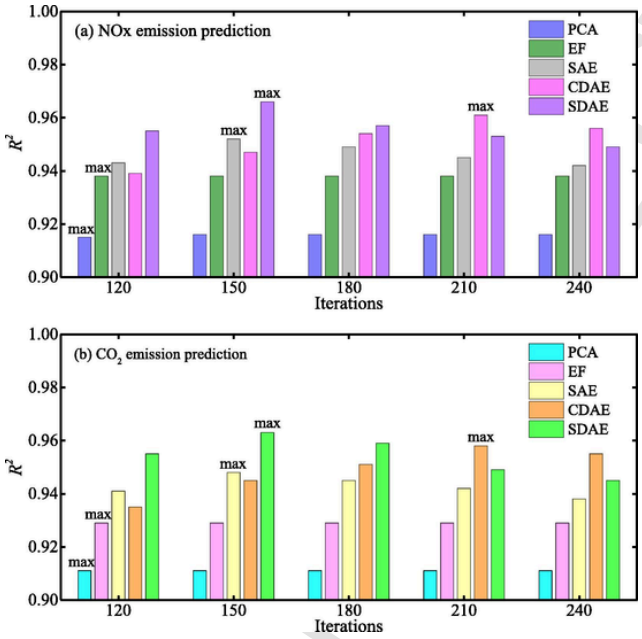

Fig. 11. Prediction performance under different feature extraction networks and iterations.

be divided into voting, weighted averaging and stacked generalization. The voting strategy is often used in classification tasks to determine the target category of inputs by counting the maximum number of votes for all labels. Different from the linear strategy (i.e., based on weighted averaging), the nonlinear strategy (i.e., based on a stacked generalization) can effectively eliminate the correlation between individual forecasting engines and thus improve the prediction accuracy. Because of this, the stacked generalization strategy is selected to construct the ensemble

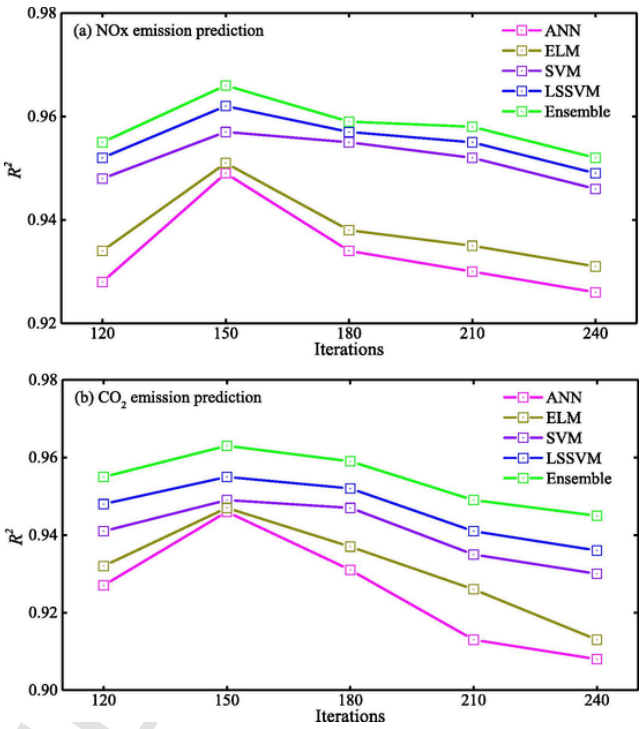

Fig. 12. Performance of the NOx and CO_2 emissions prediction under different feature analysis models.

learning model, and the GPR is used as a strategy model due to its unique ability to quantify prediction uncertainty.

A brief overview of the GPR algorithm is as follows. The GPR [42] is a non-parametric probability model based on Bayesian theory, which is suitable for dealing with small sample learning problems. Based on the training dataset $\{H, Y\}$, the nonlinear relationship between the input and output established by the GPR can be defined as:

$$y_i = \varphi_i(\boldsymbol{h}_i) + \rho_i \tag{12}$$

where $\varphi_i(\bullet)$ is the latent function, let $\varphi_i = \varphi_i(\bullet)$; $\rho_i \ \mathcal{N}\left(0, \sigma_M^2\right)$ is Gaussian noise with a mean value of zero and a variance of σ_M^2. For the input dataset H, the Gaussian process function $\mathcal{Q}(H)$ is expressed as a collection of the latent functions that obey a multivariate Gaussian distribution, given by:

$$\mathcal{Q}(H) \ \mathcal{N}(\mu(H), K(H,H)) \tag{13}$$

where $\mathcal{Q} = \left[\varphi_1, \varphi_2, \cdots, \varphi_N\right]^T$; $\mu(H)$ is the mean function, usually taken to zero; $K(H,H)$ is kernel function matrix. The GPR performance is highly dependent on $K(H,H)$, indicating that careful consideration should be taken to select the kernel function. A variety of kernel functions are available for the GPR and they are categorized into: isotropic (ISO) and automatic correlation determination (ARD) [43]. In practice, the traditional ISO function often suffers dimensional failures for high-dimensional inputs [44]. Whereas, the ARD function not only maps the high-dimensional inputs but also distinguishes the importance of different inputs and thereby deletes irrelevant input information. Above these reasons, the ARD function is selected as the kernel function of GPR in this study, specified as:

$$k\left(\boldsymbol{h}_u, \boldsymbol{h}_v\right) = \tau^2 exp\left(-\frac{\left\|\boldsymbol{h}_u - \boldsymbol{h}_v\right\|^2}{2\ell^2}\right) \tag{14}$$

where $\boldsymbol{h}_u$ and $\boldsymbol{h}_v$ are any two samples in the dataset H; τ denotes the signal variance; ℓ stands for the length scale.

Table 6
Performance evaluation of the ensemble learning model for NOx emission prediction.

Dataset	Model	RMSE (mg/m^3)	MAE (mg/m^3)	R^2	Calculation time (ms/f)
Validation dataset	ANN	7.88	5.08	0.94	38.63
	ELM	7.71	6.06	0.94	**38.58**
	SVM	7.58	5.82	0.95	38.61
	LSSVM	7.17	5.49	0.95	38.59
	GPR	7.09	5.43	0.95	44.04
	Ensemble	**6.79**	**5.07**	**0.96**	38.78
Testing dataset	ANN	7.31	5.41	0.94	38.63
	ELM	7.10	5.61	0.95	**38.58**
	SVM	6.71	5.05	0.95	38.61
	LSSVM	6.28	4.83	0.96	38.59
	GPR	6.13	4.74	0.96	44.04
	Ensemble	**5.96**	**4.71**	**0.97**	38.78

The best values are shown in bold.

3. Experiments on a 4.2 MW heavy oil-fired boiler

3.1. Experimental setup

Experimental investigations were conducted on a 4.2 MW heavy oil-fired boiler. The main structure of the boiler furnace and the implementation of the image acquisition system are presented in Fig. 6. The furnace has a height of 6 m and a cross-section of 1.8 m × 1.8 m. During the boiler system operation, the heavy oil is accelerated by the high-speed rotating cup and then collided with the reverse primary air to realize rapid atomization. The secondary air and tertiary air provide sufficient oxidizers to guarantee combustion completion. Note that the quality of the heavy oil remains stable throughout the experiments, and its ultimate properties are summarized in Table 2.

The flame images are captured by a CCD camera (HIKVISION Ltd, MV-CA003-50GC) with a pixel size of 640(H) × 480(V). The camera is mounted on the end of the optical probe, protected by an air–water cooling jacket. A 90° objective lens is assembled at the front of the optical probe that can fully observe the flame root area. A gas analyzer (Testo Ltd, testo330-1 LL) is installed at the outlet of the furnace to continuously measure the NOx and CO_2 concentrations in the untreated flue gas.

3.2. Data collection and preparation

In the experiments, the heavy oil-fired boiler operates under three different fuel loads (FLs) of 20%, 40% and 60%. Under each FL, five combustion conditions are created according to different primary air (PA) and secondary air (SA). In all experiments, the tertiary air is always kept at a 20% opening. Table 3 depicts the flame image dataset obtained from 15 different combustion conditions. For each combustion condition, 1000 RGB flame images are collected.

The descriptive statistics of NOx and CO_2 concentrations are illustrated in Tables 4 and 5. It can be observed that the emission concentrations varied significantly with the change of the FL. All acquired flame images and the measured emission concentrations are recorded on the computer. The sampling frequency of the camera is 4f/s (frame per second), while the sampling interval of the gas analyzer is 1 s. That means every four flame images correspond to one concentration value. Fig. 7 displays typical flame images from 15 different combustion conditions and their corresponding emission concentrations.

To eliminate the influence of different image sizes and accelerate the convergence of the prediction model, all flame images are resized to 256(H) × 256(V) and normalized to 0 to 1 using the min–max scale [32]. For the full dataset listed in Table 3, 3000 images (i.e., 200 images/condition) are randomly selected and manually labeled, while the remaining 12,000 images remain unlabeled. These unlabeled images are used for unsupervised training of the feature extraction network SDAE. The labeled images are further divided into three datasets, which are employed for training (60%), validation (20%) and testing (20%) the emission prediction model. The structure of the labeled dataset is shown in Fig. 8, where the NOx concentration and CO_2 concentration are non-stationarity (fluctuate sharply).

Note that, all the calculations are implemented in Anaconda software with Python programming language, and run in a computing system with Intel i7-8700 K CPU, 64 GB RAM and GeForce GTX 1080 Ti GPU.

3.3. Performance evaluation

To evaluate the performance of the prediction model, three evaluation metrics include root mean square error (RMSE), mean absolute error (MAE) and determination coefficient (R^2) are utilized. These metrics measure the deviation between the observed value and the predicted value, described as:

$$RMSE = \sqrt{\frac{1}{\mathscr{D}}\sum_{i=1}^{i=\mathscr{D}}\left(\hat{P}_i - P_i\right)^2} \tag{15}$$

$$MAE = \frac{1}{\mathscr{D}}\sum_{i=1}^{i=\mathscr{D}}\left|\hat{P}_i - P_i\right| \tag{16}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{\mathscr{D}}\left(P_i - \hat{P}_i\right)^2}{\sum_{i=1}^{\mathscr{D}}\left(P_i - \bar{P}_i\right)^2} \tag{17}$$

where $\mathscr{D}$ represents the number of testing samples; P_i is the observed value; $\hat{P}_i$ is the predicted value; $\bar{P}_i$ is the average of all measurements in the testing dataset. RMSE reflects the overall deviation between the measured and predicted values. MAE reveals the similarity between the measured and predicted values. R^2 considers the related degree between the observed and predicted values. A lower value of RMSE and MAE demonstrates better prediction performance, and a higher value of R^2 indicates higher accuracy of the prediction.

4. Model establishment

4.1. Model training

The proposed ensemble deep learning model consists of two independent training processes, i.e., unsupervised training of the feature extraction model and supervised training of the feature analysis model. During the SDAE training, all weight parameters are initialized by Gaussian distribution and then updated iteratively by the stochastic gradient descent algorithm [31]. To further improve the robustness of the SDAE, Gaussian noise with a corruption ratio of 0.2 ($\varphi = 0.2$) is added into the training dataset (i.e., 12,000 unlabeled images). This corruption ratio is determined by cross-validation with other ratios (e.g., 0.1, 0.3, 0.4, and 0.5), which helps to maximize the feature learning ability [refer to Fig. 10]. Fig. 9 records the training and validation progress of the SDAE with a learning rate of 0.01.

It can be seen that the training time of the two-level network increases linearly with the number of iterations, and 300 iterations consume 172.5 min (the first-level network) and 15.6 min (the second-level network), respectively. In the first-level network, the training loss and validation loss decrease rapidly in the first 120 iterations and reach the minimum value at the 150th iteration. With the further increase of iteration, these losses show a slight upward trend. Similar training experience also appears in the second-level network, that is, training loss and validation loss converge rapidly at first, but increase slightly after the 150th iteration. This indicates that the two-level network has expe-

Table 7
Performance evaluation of the ensemble learning model for CO_2 emission prediction.

Dataset	Model	RMSE (%)	MAE (%)	R^2	Calculation time (ms/f)
Validation dataset	ANN	0.59	0.45	0.93	38.62
	ELM	0.58	0.45	0.94	**38.57**
	SVM	0.57	0.44	0.94	38.61
	LSSVM	0.55	0.38	0.95	38.59
	GPR	0.53	0.36	0.95	44.04
	Ensemble	**0.51**	**0.34**	**0.96**	38.76
Testing dataset	ANN	0.55	0.42	0.93	38.62
	ELM	0.54	0.41	0.94	**38.57**
	SVM	0.53	0.40	0.94	38.61
	LSSVM	0.49	0.34	0.95	38.59
	GPR	0.46	0.35	0.95	44.04
	Ensemble	**0.45**	**0.33**	**0.96**	38.76

The best values are shown in bold.

rienced under-fitting and over-fitting states in the training process, which causes network performance degradation. Therefore, too little or too much training is not the best choice. Considering the computation cost and preventing the occurrence of over-fitting, the optimal iteration of the two-level network is set to 150.

The trained SDAE can be directly used to extract deep features of the original flame image. After that, based on the deep features and label information of the training dataset (1800 labeled images), four forecasting engines (ANN, ELM, SVM and LSSVM) are trained, respectively. Finally, the forecasts of the four forecasting engines are combined with the label information to train the GPR. After the ensemble learning model training, the dataset outside of the training dataset (such as the validation dataset and testing dataset) is used to verify its feasibility in practical application.

4.2. Performance analysis

To optimize the performance of the ensemble learning model, a series of comparative studies are carried out involving corruption ratios, feature extraction networks and feature analysis models. The detailed description of each evaluation is illustrated in the following sections.

4.2.1. Effect of different corruption ratios

In general, the prediction model with good robustness is more preferred, which is conducive to the application in a harsh environment. In this study, Gaussian noise with different corruption ratios is added to the training image to improve the robustness of the SDAE and make it capable of obtaining essential feature information from the noisy images. Because of this, Gaussian noise with a special corruption ratio φ [determined through Eq. (1)] is added to the training dataset. Therefore, it is of great significance to study the influence of different corruption ratios on the accuracy of the emission prediction. Six different corruption ratios from 0 to 0.5 with a step size of 0.1 are considered, and the obtained prediction results are shown in Fig. 10.

It is evident that the prediction accuracy (R^2) of the ensemble learning model presents a trend of first increasing and then decreasing with the increase of the corruption ratio. The maximum prediction accuracies are achieved when $\varphi = 0.2$, that is, 0.96 (validation dataset) and 0.97 (testing dataset) for NOx prediction accuracy, 0.96 (validation dataset) and 0.96 (testing dataset) for CO_2 prediction accuracy. The above results suggest that a proper corruption ratio can effectively improve the feature learning ability of the SDAE, while excessive noise interference will seriously reduce the data quality, eventually resulting in poor prediction performance. As a consequence, to ensure the high-precision performance and strong robustness of the ensemble learning model, the corruption ratio is set to 0.2.

4.2.2. Performance of different feature extraction networks

To demonstrate the superiority of the SDAE, a comprehensive comparison is carried out with other typical feature extraction networks, including PCA, empirical formula (EF), stacked autoencoder (SAE) and convolutional denoising autoencoder (CDAE). The PCA [45] is a linear data dimension reduction method that can linearly map the sample data from high-dimensional space to low-dimensional space, and eliminate redundant information. The EF is a statistical feature extraction method based on pre-defined formulas or empirical parameters, requiring prior expert knowledge. Here, sixteen statistical features were considered, referring to geometric parameters and brightness parameters [46]. The structure of the SAE is similar to the SDAE, except that the denoising coding technique is not considered [i.e., $\varphi = 0$, refer to Eq. (1)]. The structure of the CDAE is the combination of the two-level network of the SDAE, including ten hidden layers. Fig. 11 illustrates the prediction accuracy of the ensemble learning model based on different feature extraction networks under different iterations.

Among these feature extraction networks, PCA and EF are two kinds of non-learning methods, so their testing accuracies always remain a fixed value. It can be seen that the prediction accuracy is worst for the PCA feature and R^2 values are 0.92 and 0.91 for NOx and CO_2 emissions prediction, respectively. This indicates that the shallow PCA method cannot mine the intrinsic information in the flame image effectively. Following the PCA, the performance of the EF-based model is also disappointing, with the NOx prediction accuracy of 0.95 and the CO_2 prediction accuracy of 0.65. In contrast, the performance of the learning method is better. The testing accuracy of the CDAE-based model is relatively higher than that of the SAE-based model, although requiring more training cost (i.e., 210th iteration). As expected, the model based on SDAE features provides the best prediction performance and reaches its maximum at the 150th iteration. Two important conclusions can be drawn based on the above results; (1) the SDAE is better than the SAE, confirming that the introduced denoising coding technique is beneficial for the feature learning ability; (2) the SDAE is better than the CDAE, suggesting that the newly designed two-level structure network is easier to be trained. In addition, the computational burden is reduced, that is, the SDAE training requires 94.13 min (i.e., 150th iteration) and the CDAE training requires 122.65 min (i.e., 210th iteration) to achieve better performance. In addition, it is also noticed that for both NOx and CO_2 emissions, the prediction accuracy of the SDAE first increases and then decreases with the increase of iterations. This phenomenon is consistent with the result in Fig. 9, that is, the SDAE has experienced two states of under-fitting and over-fitting successively, which have a certain impact on the prediction effect.

4.2.3. Performance of different feature analysis models

After the feature extraction network is determined, the feature analysis model still needs to be carefully selected. It is well known that a suitable feature analysis model is promising to maximize prediction accuracy. In this study, four forecasting engines and the ensemble learning model are compared. All feature analysis models are established based on the same SDAE features, and the obtained prediction results are summarized in Fig. 12.

It can be observed that in all cases, the prediction accuracy of the ensemble learning model (represented by the green line in the figure) is always in the highest position, which is higher than that of other forecasting engines. In addition, with the increase of iterations, the prediction accuracy of all feature analysis models first increases and then decreases gradually, and reaches the maximum at the 150th iteration. In fact, this trend mainly depends on the accuracy of the SDAE features, which further confirms that representative features play a decisive role in prediction performance.

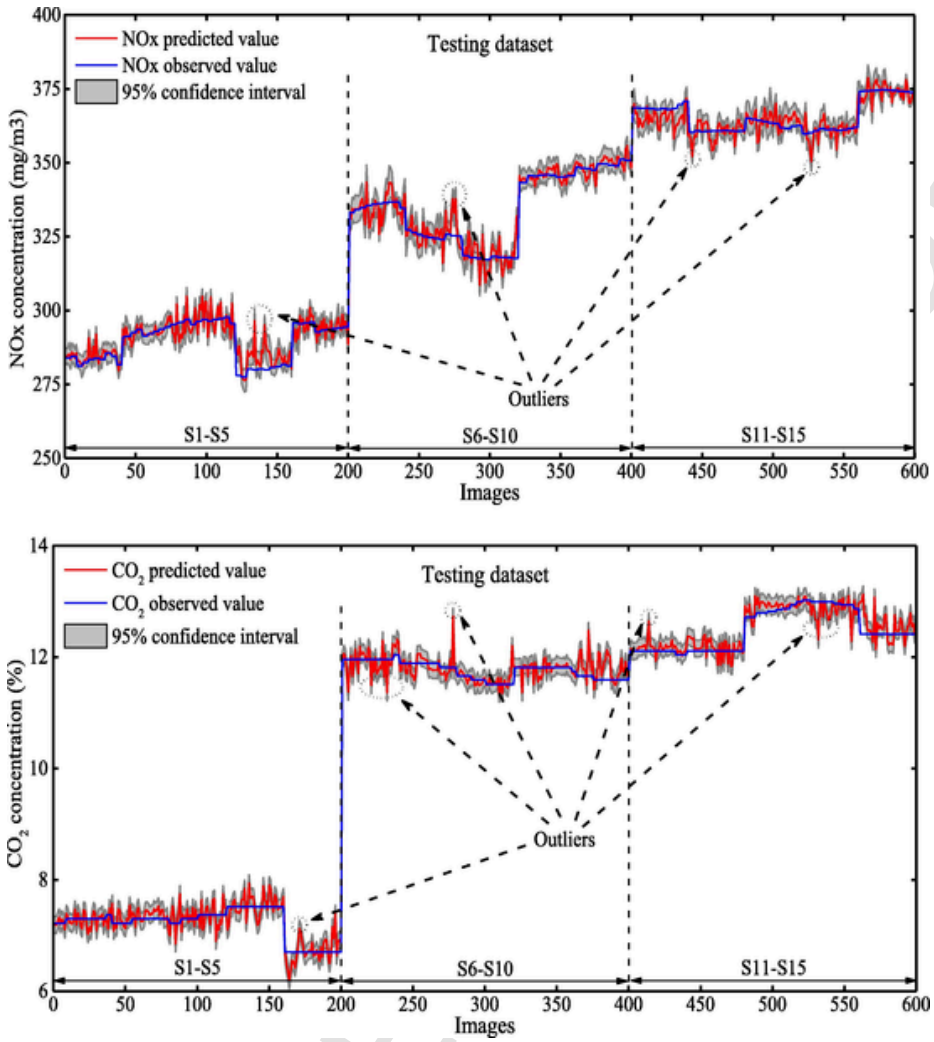

Fig. 13. NOx and CO_2 emissions prediction based on ensemble learning model.

5. Model evaluation and discussion

5.1. Point prediction

In the application process, the deep feature of the new flame image is extracted by the SDAE and then sent into four forecasting engines and the GPR for the NOx and CO_2 emissions estimation. Here, the validation dataset (20%) and testing dataset (20%) are used for presenting the performance of the emission prediction model. The accuracy of NOx and CO_2 prediction of the individual forecasting engine and the ensemble learning model based on the SDAE feature are listed in Tables 6 and 7, respectively. It can be seen, the ensemble learning model based on the SDAE feature has achieved better performance. This demonstrates that the ensemble learning model based on the SDAE feature has strong generalization ability, and thus the extracted image features can be applied to different emissions predictions.

In contrast to the individual forecasting engine, the ensemble learning model provides better prediction accuracy with the lowest RMSE and MAE and the highest R^2. For the NOx emission prediction (as shown in Table 6), it can be seen that: (1) for the validation dataset, RMSE, MAE and R^2 achieved 6.79 mg/m^3, 5.07 mg/m^3 and 0.96, respectively; (2) for the testing dataset, RMSE, MAE and R^2 achieved 5.96 mg/m^3, 4.71 mg/m^3 and 0.97, respectively. For the CO_2 emission prediction (as shown in Table 7), it can be seen that: (1) for the validation dataset, RMSE, MAE and R^2 achieved 0.51%, 0.34% and 0.96, respectively; (2) for the testing dataset, RMSE, MAE and R^2 achieved 0.45%, 0.33% and 0.96, respectively. Therefore, it is suggested that the ensemble learning model improves the prediction performance by integrating the four forecasting engines.

Along with this, due to the slightly complicated structure of the ensemble learning model, more calculation time is required: 38.78 ms/f (milliseconds per frame) for the NOx prediction and 38.76 ms/f for the CO_2 prediction. Even so, all flame images collected by the CCD camera with a frame rate of 4f/s can be processed in time. This means the ensemble learning model can meet the requirements of real-time emission prediction. Moreover, it can be found that the forecasting engine ELM has the shortest calculation time, which indicates that the ELM has a faster response speed due to its unique structure. Relatively, the calculation time of the forecasting engine GPR is the longest, even surpassing the ensemble learning model. This proves that the computational complexity of the GPR is seriously affected by the dimension of input data, that is, higher-dimensional input will require more computation costs.

5.2. Confidence interval prediction

Rather than merely point prediction, the ensemble learning model can also provide confidence interval information. Fig. 13 depicts the point prediction and corresponding 95% confidence interval estimation through the ensemble learning model. This demonstrated that although there are few negligible outliers, most of the predicted values can well fit the actual observed values. These outliers may be attributed to the fluctuation of the combustion flame or the disturbance of the image ac-

quisition system, which can be ignored. In addition, it is also noted that nearly all actual observations are within the confidence interval, which indicates that the prediction result has high credibility.

6. Conclusions

In this study, a novel ensemble deep learning model is proposed for combustion emission prediction by integrating SDAE, ANN, ELM, SVM, LSSVM and GPR models. The effectiveness of the model is evaluated through the flame images captured under different combustion conditions on a 4.2 MW heavy oil-fired boiler. The main conclusions are as follows:

(1) The established SDAE can accurately extract the representative features of the flame image, and its training efficiency and prediction accuracy have made significant progress, outperforms other networks such as PCA, EF, SAE and CDAE.
(2) By comparing different forecasting engines, the ensemble learning model provides the lowest RMSE and MAE while the highest R^2 regardless of NOx or CO_2 emissions prediction. Among them, the prediction accuracy of NOx and CO_2 emissions reach to $R^2 = 0.97$ and $R^2 = 0.96$, respectively.
(3) The prediction time of the ensemble learning model for the NOx and CO_2 emissions is 38.78 ms/f and 38.76 ms/f, respectively. Although the prediction time consumed by the ensemble learning model is slightly higher than that of the single forecasting engine, it can still meet the requirements of real-time monitoring;
(4) In addition to point prediction, the ensemble learning model can also generate the confidence interval and quantitatively characterize the prediction uncertainty. In practical testing, the actual observed values are covered in the confidence interval, confirming the reliability of the prediction model.

The proposed prediction method provides a promising monitoring tool for NOx and CO_2 emissions during the combustion process. Future work will strengthen the practical applications of the prediction model, such as coal combustion and biomass co-firing combustion.

CRediT authorship contribution statement

Zhezhe Han: Conceptualization, Methodology, Software, Data curation, Validation, Writing – original draft. **Jian Li:** Validation, Formal analysis, Investigation. **Md. Moinul Hossain:** Validation, Formal analysis, Writing – review & editing, Supervision. **Qi Qi:** Resources, Writing – review & editing. **Biao Zhang:** Software, Visualization, Data curation. **Chuanlong Xu:** Supervision, Resources, Writing – review & editing, Project administration, Funding acquisition.

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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant number 51976038], the Natural Science Foundation of Jiangsu Province, China for Young Scholars [grant number BK20190366], and the China Scholarship Council [grant number 202006090164].

References

[1] Yang T, Ma K, Lv Y, Bai Y. Real-time dynamic prediction model of NOx emission of coal-fired boilers under variable load conditions. Fuel 2020;274:117811.
[2] Nazari S, Shahhoseini O, Sohrabi A, Davari S, Paydar R, et al. Experimental determination and analysis of CO2, SO2 and NOx emission factors in Iran's thermal power plants. Energy 2010;35(7):2992–8.
[3] Zhou H, Meng S. Numerical prediction of swirl burner geometry effects on NOx emission and combustion instability in heavy oil-fired boiler. Appl Therm Eng 2019;159:113843.
[4] Schluckner C, Gaber C, Landfahrer M, Demuth M, Hochenauer C. Fast and accurate CFD-model for NOx emission prediction during oxy-fuel combustion of natural gas using detailed chemical kinetics. Fuel 2020;264:116841.
[5] Kang M, Jeong H, Farid M, Hwang J. Effect of staged combustion on low NOx emission from an industrial-scale fuel oil combustor in South Korea. Fuel 2017;210:282–9.
[6] Park J, Park S, Ryu C, Baek S, Kim Y, et al. CFD analysis on bioliquid co-firing with heavy fuel oil in a 400 MWe power plant with a wall-firing boiler. Appl Therm Eng 2017;124:1247–56.
[7] Lapeyre C, Misdariis A, Cazard N, Veynante D, Poinsot T. Training convolutional neural networks to estimate turbulent sub-grid scale reaction rates. Combust Flame 2019;203:255–64.
[8] Krzywanski J, Nowak W. Artificial intelligence treatment of SO_2 emissions from CFBC in air and oxygen-enriched conditions. J Energy Eng 2015;142:4015017.
[9] Zhou H, Zhao J, Zheng L, Wang C, Cen K. Modeling NOx emissions from coal-fired utility boilers using support vector regression with ant colony optimization. Eng Appl Artif Intell 2012;25(1):147–58.
[10] Tan P, Xia J, Zhang C, Fang Q, Chen G. Modeling and reduction of NOx emissions for a 700 MW coal-fired boiler with the advanced machine learning method. Energy 2016;94:672–9.
[11] Yuan Z, Meng L, Gu X, Bai Y, Cui H, et al. Prediction of NOx emissions for coal-fired power plants with stacked-generalization ensemble method. Fuel 2021;289:119748.
[12] Li Q, Yao G. Improved coal combustion optimization model based on load balance and coal qualities. Energy 2017;132:204–12.
[13] Liukkonen M, Heikkinen M, Hiltunen T, Hälikkä E, Kuivalainen R, et al. Artificial neural networks for analysis of process states in fluidized bed combustion. Energy 2011;36(1):339–47.
[14] Tuttle J, Vesel R, Alagarsamy S, Blackburn L, Powell K. Sustainable NOx emission reduction at a coal-fired power station through the use of online neural network modeling and particle swarm optimization. Control Eng Pract 2019;93:104167.
[15] Golgiyaz S, Talu M, Onat C. Artificial neural network regression model to predict flue gas temperature and emissions with the spectral norm of flame image. Fuel 2019;255(1):115827.
[16] González-Cencerrado A, Gil A, Peña B. Characterization of PF flames under different swirl conditions based on visualization systems. Fuel 2013;113:798–809.
[17] Li N, Lu G, Li X, Yan Y. Prediction of pollutant emissions of biomass flames through digital imaging, contourlet transform, and support vector regression modeling. IEEE Trans Instrum Meas 2015;64(9):2409–16.
[18] Wang F, Wang X, Ma Z, Yan J, Chi Y, et al. The research on the estimation for the NOx emissive concentration of the pulverized coal boiler by the flame image processing technique. Fuel 2002;81(16):2113–20.
[19] Liu Y, Yang C, Cao Z, Yao Y. Ensemble deep kernel learning with application to quality prediction in industrial polymerization processes. Chemometrics Intell Lab Syst 2018;174:15–21.
[20] Qiu T, Liu M, Zhou G, Wang L, Gao K. An unsupervised classification method for flame image of pulverized coal combustion based on convolutional auto-encoder and hidden markov model. Energies 2019;12(13):2585.
[21] Liu Y, Fan Y, Chen J. Flame images for oxygen content prediction of combustion systems using DBN. Energy Fuels 2017;31(8):8776–83.
[22] Lyu Y, Chen J, Song Z. Image-based process monitoring using deep learning framework. Chemometrics Intell Lab Syst 2019;189:8–17.
[23] Shao H, Jiang H, Lin Y, Li X. A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders. Mech Syst Sig Process 2018;102:278–97.
[24] Li N, Lu G, Li X, Yan Y. Prediction of NOx emissions from a biomass fired combustion process based on flame radical imaging and deep learning techniques. Combust Sci Technol 2015;188:233–46.
[25] Chen J, Chan L, Cheng Y. Gaussian process regression based optimal design of combustion systems using flame images. Appl Energy 2013;111:153–60.
[26] Lv Y, Liu J, Yang T, Zeng D. A novel least squares support vector machine ensemble model for NOx emission prediction of a coal-fired boiler. Energy 2013;55(15):319–29.
[27] Wan K, Hartl S, Vervisch L, Domingo P, Barlow R, et al. Combustion regime identification from machine learning trained by Raman/Rayleigh line measurements. Combust Flame 2020;219:268–74.
[28] Hinton G, Salakhutdinov R. Reducing the dimensionality of data with neural networks. Science 2006;313(5786):504–7.
[29] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 770–8.
[30] Han Z, Huang Y, Li J, Zhang B, Hossain M, et al. A hybrid deep neural network based prediction of 300MW coal-fired boiler combustion operation condition. Sci China Technol Sci 2021;64.
[31] Han Z, Hossain M, Wang Y, Li J, Xu C. Combustion stability monitoring through flame imaging and stacked sparse autoencoder based deep neural network. Appl

Energy 2019;259:114159.
[32] Han Z, Li J, Zhang B, Hossain M, Xu C. Prediction of combustion state through a semi-supervised learning model and flame imaging. Fuel 2020;289(4):119745.
[33] Ren T, Modest M, Fateev A, Sutton G, Zhao W, et al. Machine learning applied to retrieval of temperature and concentration distributions from infrared emission measurements. Appl Energy 2019;252(15):113448.
[34] Yilmaz I. Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. Environ Earth Sci 2010;61(4):821–36.
[35] Huang G, Zhu Q, Siew C. Extreme learning machine: theory and applications. Neurocomputing 2006;70(1–3):489–501.
[36] Akyol K. Comparing of deep neural networks and extreme learning machines based on growing and pruning approach. Expert Syst Appl 2020;140:112875.
[37] Zhou H, Tang Q, Yang L, Yan Y, Lu G, et al. Support vector machine based online coal identification through advanced flame monitoring. Fuel 2014;117:944–51.
[38] Tuttle J, Blackburn L, Andersson K, Powell K. A systematic comparison of machine learning methods for modeling of dynamic processes applied to combustion emission rate modeling. Appl Energy 2021;292:116886.
[39] Laref R, Losson E, Sava A, Siadat M. On the optimization of the support vector machine regression hyperparameters setting for gas sensors array applications. Chemometrics Intell Lab Syst 2018;184:22–7.
[40] Zhai Y, Ding X, Jin X, Zhao L. Adaptive LSSVM based iterative prediction method for NOx concentration prediction in coal-fired power plant considering system delay. Appl Soft Comput 2020;89:106070.
[41] Zhang C, Wei H, Zhao X, Liu T, Zhang K. A Gaussian process regression based hybrid approach for short-term wind speed prediction. Energy Convers Manage 2016;126:1084–92.
[42] Rasmussen C, Williams C. Gaussian processes for machine learning. MIT Press; 2006.
[43] Liu K, Hu X, Wei Z, Li Y, Jiang Y. Modified Gaussian process regression models for cyclic capacity prediction of lithium-ion batteries. IEEE Trans Transp Electrif 2020;5(4):1225–36.
[44] Liu K, Li Y, Hu X, Lucu M, Widanage W. Gaussian process regression with automatic relevance determination kernel for calendar aging prediction of lithium-ion batteries. IEEE Trans Ind Inf 2020;16(6):3767–77.
[45] Yang G, Wang Y, Li X. Prediction of the NOx emissions from thermal power plant using long-short term memory neural network. Energy 2020;192:116597.
[46] Bai X, Lu G, Hossain M, Szuhánszki J, Daood S, et al. Multi-mode combustion process monitoring on a pulverised fuel combustion test facility based on flame imaging and random weight network techniques. Fuel 2017;202:656–64.