A flame imaging based online deep learning model for predicting NO_X emissions from an oxy-biomass combustion process

Li Qin, Gang Lu, *Senior Member, IEEE*, Md Moinul Hossain, *Senior Member, IEEE*, Andy Morris, and Yong Yan, *Fellow, IEEE*

***Abstract*—To reduce NO_x (Nitrogen Oxide) emissions from fossil fuel and biomass fired power plants, online prediction of NO_x emissions is important in the plant operation. Data-driven models have been developed to predict NO_x emissions from various combustion processes with good accuracy. However, such models have initially been built based on known combustion conditions, which are historically 'seen'. For new conditions, which are 'unseen', these models usually perform unwell. In this study, an ODL (Online Deep Learning) model is proposed to predict NO_x emissions from an oxy-biomass combustion process for 'seen' and 'unseen' combustion conditions based on source deep learning and condition recognition models. The ODL model is mainly built based on 'unseen' combustion conditions. A new objective function that consists of regression loss and distillation loss is introduced in the ODL model to improve the prediction accuracy. The ODL model is examined using boiler operation data, flame temperature maps and NO_x data obtained under a range of oxy-biomass combustion conditions on an Oxy-fuel Combustion Test Facility. Flame images acquired using a dedicated imaging system are used for computing the temperature distribution of the flame through two-colour pyrometry. The results demonstrate that the proposed model is capable of predicting NO_x emissions under 'seen' and 'unseen' conditions with a mean absolute percentage error of less than 3%, for the 1st, 2nd, and 3rd updates.**

***Index Terms*— NO_x prediction, online deep learning, flame temperature map, condition monitoring, oxy-biomass combustion**

I. Introduction

FOSSIL fuels are being replaced by renewable energy resources to meet the legally binding target of 'net-zero' emission by 2050 under the Paris Agreement. Biomass, regarded as a carbon-neutral and dispatchable renewable fuel, is widely used in new and existing boilers for electric power and thermal energy generation [1]. Combining with the oxy-fuel combustion technology, oxy-biomass combustion has become a promising environmentally-friendly technology which can reduce significantly not only the CO_2 (carbon dioxide) but also NO_x emissions as the recycled NO is reburnt or reduced when it is re-circulated through the combustion system [2]. In the oxy-biomass combustion process, though the use of O_2/flue gas instead of air reduces the NO_x emission, NO_x is still one of the main contributors to the harmful emissions. The NO_x emission is formed mainly from three different sources, i.e., fuel-N, thermal-N, and prompt-N [3], depending upon the fuel type, oxidiser atmosphere and temperature [4]. It is also understood that biomass fuel has a higher conversion of fuel-N to NO_x than coal [5]. Therefore, the control of NO_x emission in such a combustion process is still a concern in the power generation industry. Techniques for the online monitoring and prediction of NO_x emission are thus crucial to meet increasingly stringent environmental regulations.

Techniques developed for monitoring NO_x emission in combustion processes can be categorised into three distinct approaches, i.e., monitoring systems, model-driven prediction, and data-driven machine learning models. The monitoring system measures the NO_x by taking samples in flue gas, and thus a significant time delay exists between the measured and actual NO_x emission values, resulting in inefficiency in NO_x control. The model-driven prediction approaches such as Computational Fluid Dynamics predict the NO_x formation through simulating the physical and chemical reactions associated with the combustion process of given fuels, which is a very useful tool for a theoretical understanding of the combustion phenomenon, but not suitable for real-time processes due to the complexity and computational resources required [6]. The data-driven models are mainly based on historical boiler operation data to predict the NO_x emission through various machine learning techniques, such as ANN (Artificial Neural Network) [7], [8], SVM (Support Vector Machine) [9], [10], ELM (Extreme Learning Machine) [11], and DBN (Deep Brief Network) [12]. Least square support vector regression and deep learning models were also proposed for NO_x prediction based on the flame radical images [13], [14]. The traditional machine learning and deep learning models have provided promising solutions to 'black-box' problems in combustion and many other industrial processes. Although considerable studies [7]-[14] were conducted in the last decade and achieved reasonable success in predicting NO_x in various combustion systems through data-driven modelling, the existing techniques suffer from several drawbacks. For

This work was partly supported by the Biomass and Fossil Fuel Research Alliance (BF2RA Project 30).

Li Qin, Gang Lu, Md Moinul Hossain and Yong Yan are with the School of Engineering, University of Kent, Canterbury, Kent CT2 7NT, UK (e-mail: lq26@ kent.ac.uk, g.lu@kent.ac.uk, m.hossain@kent.ac.uk and y.yan@kent.ac.uk).

Andy Morris is with EDF Energy, Central Technical Organisation, Coal and Gas Operations, West Burton A Power Station, Nottinghamshire DN22 9BL, UK (e-mail: Andy.Morris@edfenergy.com).

instance, those models are offline without updates, so these models provide unsatisfactory performance when the new conditions are outside the training conditions. Hence, an online update (i.e., online learning) strategy is required to improve the model performance for new conditions.

To build an online model, sample addition and sample replacement are proposed in [15], [16] to update LSSVM (least squares support vector machine). However, both approaches required access to previous conditions during the update process, which is infeasible if the training data from previous conditions are missing or unavailable. In addition, these approaches require a large storage space of computer system for storing historical datasets. Therefore, how to update the model with data from 'unseen' conditions is becoming challenging and crucial for online NO_x prediction.

In addition, these models rely on a large number of operation data obtained using conventional instruments around the boiler, where the flame characteristics information is missed out. Studies have shown that the flame characteristics are closely related to the combustion stability as well as NO_x emissions [17], [18], so employing the flame data such as temperature maps directly into the model will increase the accuracy of NO_x prediction and improve the efficiency of the model. Finally, most above-mentioned data-driven models focused on predicting NO_x emissions based on boiler operation data from air-coal or biomass combustion processes. There are very few studies on NO_x prediction for oxy-biomass combustion processes. It has also been suggested that a new technical strategy needs to be explored for an effective online deep learning model to predict NO_x emissions.

In recent years, ODL (online deep learning) modelling has been explored to solve multilabel classification problems by updating an initial model (often called a source model) with datasets from 'unseen' conditions (i.e., having not been used to train the source DL or participated in the incremental learning process of ODL model), so that the updated model recognises all the 'seen' and 'unseen' conditions [19]. However, those models often suffer catastrophic forgetting problems, i.e., whilst the updated model recognises the 'unseen' condition, but with a degraded performance to recognise 'seen' conditions (i.e., having been used to train the source DL or participated in the incremental learning process of the ODL model). Various attempts have been made to solve this problem, such as learning without forgetting [19] and learning without memorising [20]. In those models, objective functions which determine how the objective loss is minimised [19], [20] are carefully considered to tackle the catastrophic forgetting problem. It has revealed that, though the catastrophic forgetting problem exists inherently in ODL modelling, it is possible to define an appropriate objective function so that the catastrophic forgetting problem can be minimised.

To address the problems discussed above, the present study focuses on developing a flame imaging-based ODL model for predicting the NO_x emission from an oxy-biomass combustion process. Being inspired by the update strategy of ODL based classification models, an online update strategy for an ODL based regression model with a new objective function that consists of the regression loss and distillation loss. The proposed ODL model is applied to predict NO_x emissions from 'seen' and 'unseen' conditions. Flame temperature maps derived from flame images are used as input datasets of the model. A convolutional neural network (CNN) acts as the source model which is built based on datasets from selected 'seen' operation conditions. The impact of hyper-parameters, such as training epoch, input image size, the number of convolutional layers and batch size, on the model's performance is examined using the grid search technique [21]. The ODL model is updated incrementally using datasets from 'unseen' conditions. A new objective function is proposed to ensure that the ODL model performs well under both the 'seen' and 'unseen' conditions. A stochastic gradient descent method is used to minimise the objective loss by optimising the parameters (e.g., weights and biases) of the model and thus minimise the catastrophic forgetting problem in the updating process [22]. A condition recognition approach based on the cosine similarity is proposed and integrated with the ODL model to recognise the conditions based on boiler operation data. The developed ODL model is evaluated using experimental data (i.e., boiler operation data, flame temperature maps and NO_x) obtained on an Air/Oxy-fuel CTF (Combustion Test Facility) under a range of oxy-biomass combustion conditions. The results obtained by the developed ODL model are analysed and discussed.

II. Methodology

A. Technical strategy

The proposed ODL model is established based on the architecture of a traditional CNN, which is well-known for extracting 'deep' features of input images. Fig. 1 shows the block diagram of the ODL model proposed for NO_x prediction. The flame images obtained under different operation conditions are deployed to calculate the temperature maps based on the two-colour method [23]. The temperature maps are then used as the input dataset of the model. The NO_x emission data are also collected concurrently with the flame image and then used as ground-truth.

The first step is to build a source DL model based on the flame temperature maps and ground-truth NO_x emission data for 'seen' conditions. The 'seen' conditions in this study refer to the historical conditions which are used to train the source DL model or have participated in the incremental learning process of the ODL model.

The second step is to construct the ODL model. Condition recognition is firstly performed to determine whether the input dataset is from a 'seen' or 'unseen' condition. If the dataset is from a 'seen' condition, the existing ODL model is used for NO_x prediction. Otherwise, the dataset will participate in the incremental learning process of the model. During the incremental process, the existing ODL model or source DL model is re-trained to form a new ODL model using the dataset of the 'unseen' condition. An objective function is then deployed to optimise the parameter values of the new ODL model.

Once the ODL model is built, it is capable of predicting the NO_x emissions for all the 'seen' and 'unseen' conditions. The following sections give a detailed description of the theoretical background of how the source and online DL models are built

based on the 'seen' datasets and updated using the 'unseen' datasets.

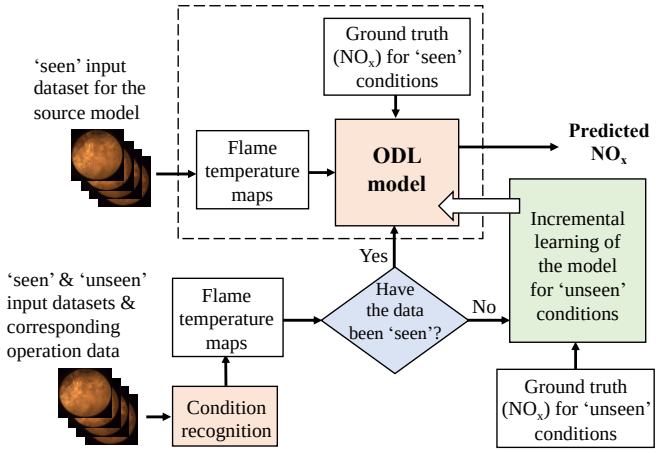

Fig. 1. Block diagram of the proposed ODL model for NO_x prediction.

B. *Source DL model*

1) *The architecture*

The source DL model is basically a conventional CNN model, which is built using the datasets from 'seen' conditions. The model consists of convolutional, pooling, dropout, and fully-connected (FC) layers, as shown in Fig. 2, where N×N is the input image size; m is the batch size; b is the number of convolution channels at the first convolutional layer; a×a is the feature dimension of a single input flame image at the dropout layer; c is the number of convolutional channels at the last convolutional layer; d and g represent the dimension of a single image feature at the first and second FC layers, respectively.

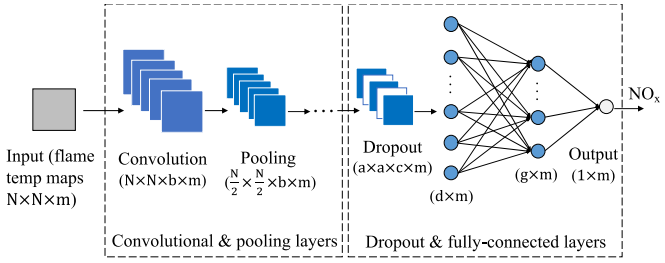

Fig. 2. Architecture of the source DL model.

Convolutional layer- A CNN model has commonly several convolutional layers, each layer extracts deep features from its input and gives a feature map at its output [24]. To ensure that the dimension of the convolution output is the same as that of the convolution input, a 'zero padding' approach is deployed at the convolution input to get the padding output tensor $\mathbf{Z}\epsilon\mathbb{R}^{(Npa)\times(Npa)\times m}$. Convolutional kernels [25] are then used to filter the padding output. The convolution output, $\mathbf{H}_c\epsilon\mathbb{R}^{N\times N\times b\times m}$, can be obtained by,

$$\mathbf{H}_{c(i,j,h,q)}=\sum_{k=1}^{K}\sum_{l=1}^{L}\mathbf{W}_{(k,l,h)}\mathbf{Z}_{(i+K-1,\,j+L-1,q)}+\mathbf{B}_{(i,j,h)}, \quad (1)$$

$$(i,\ j=1,\ 2,\ \cdots,\ N;\ h=1,\ 2,\ 3,\ ..,\ b;\ q=1,\ 2,\ \cdots,\ m)$$

where $\mathbf{W}\epsilon\mathbb{R}^{K\times L\times b}$ and $\mathbf{B}\epsilon\mathbb{R}^{N\times N\times b}$ represent the convolutional kernel weights and biases, respectively; $N_{pa}\times N_{pa}$ are the dimensions of the 1st and 2nd orders of $\mathbf{Z}$, respectively; K and L represent the dimensions of the 1st and 2nd orders of convolutional weights, respectively; *i*, *j*, *h*, and *q* are the subscripts of the 1st, 2nd, 3rd, and 4th orders, respectively; b represents the number of convolutional channels. In each convolutional layer, batch normalisation and activation operations are performed. The batch normalisation reduces the influence of the parameter initialisation and the risk of overfitting during the training process [26], and is expressed as,

$$\mathbf{H}_b=\frac{\mathbf{H}_c-E(\mathbf{H}_c)}{\sqrt{Var(\mathbf{H}_c)+\varepsilon}}, \quad (2)$$

where $\mathbf{H}_b\epsilon\mathbb{R}^{N\times N\times b\times m}$ is the batch normalisation output; ε is a constant (its default value is 10^{-5}) to improve the numerical stability of $\mathbf{H}_b$; $E(\mathbf{H}_c)$ and $Var(\mathbf{H}_c)$ represent the mean and variance of a batch convolutional output, respectively.

In this study, the activation function, i.e., ReLU (Rectified Linear Unit [27]), is used along with the convolution operation to extract the nonlinear features of the convolution input, which is expressed as,

$$\mathbf{H}_{r(i,j,k,q)}=\mathrm{ReLU}(\mathbf{H}_{b(i,j,k,q)})=\begin{cases}0, & \text{if } \mathbf{H}_{b(i,j,k,q)}<0\\ \mathbf{H}_{b(i,j,k,q)}, & \text{if } \mathbf{H}_{b(i,j,k,q)}\geq 0\end{cases}, \quad (3)$$

where $\mathbf{H}_r\epsilon\mathbb{R}^{N\times N\times b\times m}$ represents the extracted output feature after the ReLU operation.

Pooling layer- The pooling layer, usually after a convolutional layer, is to reduce the dimension of the feature map [28]. The widely used max-pooling approach is employed to downsample the input of the pooling layer using a max-filter [29]. The filter size of 2×2 is chosen to reduce the dimension of $\mathbf{H}_r$ to half at each pooling layer.

Dropout layer- Dropout is an operation that omits the hidden neurons of input features at a dropout layer randomly with an appropriate probability during the training process to reduce the risk of 'overfitting', and increase the generalisation ability and prediction accuracy of the model [30]. In comparison with other machine learning models, such as ANN and SVM, a DL model has generally more layers, often resulting in increased complexity of the model and a high risk of 'overfitting' in the training process. A dropout can be deployed to prevent the 'overfitting' after either the conventional, pooling or FC layers. In this study, the dropout was performed before the first FC layer.

Fully-connected (FC) layer- An FC layer is composed of many neurons, each neuron is connected to all the neurons in the adjacent layers [24]. Two FC layers are used in the source DL and ODL models in this study. The output of the first FC layer can be expressed as,

$$\mathbf{H}_{f1}=\mathrm{ReLU}(\mathbf{W}_{f1}\mathbf{H}_h+\mathbf{B}_{f1}), \quad (4)$$

where $\mathbf{H}_{f1}\epsilon\mathbb{R}^{g\times m}$ represents the output of the first FC layer; $\mathbf{W}_{f1}\epsilon\mathbb{R}^{g\times d}$ and $\mathbf{B}_{f1}\epsilon\mathbb{R}^{g\times m}$ represent the weights and biases of the first FC layer, respectively. The output tensor of the dropout layer is reshaped to the input $\mathbf{H}_h\epsilon\mathbb{R}^{d\times m}$ of the first FC layer. $\mathbf{H}_{f1}$

is also the input of the second FC layer which operates without the ReLU function, i.e.,

$$\mathbf{Y} = \mathbf{W}_{\text{f2}}\mathbf{H}_{\text{f1}} + \mathbf{B}_{\text{f2}}, \tag{5}$$

where $\mathbf{Y}$ is the model output in a batch; $\mathbf{W}_{\text{f2}} \in \mathbb{R}^{1\times g}$ and $\mathbf{B}_{\text{f2}} \in \mathbb{R}^{1\times m}$ are the weights and biases of the second FC layer, respectively.

2) Objective function

In a regression model, the objective function defines the difference between the estimated value and the ground-truth value of the model. In this study, the objective function of the source DL model is expressed as,

$$J_0 = \frac{1}{\text{m}}\left\|\mathbf{Y}_0^o - \mathbf{Y}_0^s\right\|_2^2, \tag{6}$$

where $\mathbf{Y}_0^o \in \mathbb{R}^{1\times m}$ is the model's output vector for a batch; $\mathbf{Y}_0^s \in \mathbb{R}^{1\times m}$ is the scaled ground-truth (NO_x) vector. To increase the convergence speed, the ground-truth data are scaled as follows,

$$\mathbf{Y}_0^s = \frac{\mathbf{Y}_0 - \mathbf{Y}_{\min}}{\mathbf{Y}_{\max} - \mathbf{Y}_{\min}}, \tag{7}$$

where $\mathbf{Y}_0 \in \mathbb{R}^{1\times m}$ is the ground-truth vector; $\mathbf{Y}_{\min}$ and $\mathbf{Y}_{\max}$ represent the minimum and the maximum of the ground-truth values, respectively.

To minimise J_0, parameter values (e.g., weights and biases) of the model are optimised using the stochastic gradient descent method [22].

C. ODL model

The ODL model is expected to be updated incrementally only using the dataset from an 'unseen' condition, as shown in Fig. 3. The purpose of the incremental learning process is to find the optimal parameter values of the ODL model and update the model for NO_x prediction of both 'seen' and 'unseen' conditions. This can be done by two steps. Firstly, the output of the $(i\text{-}1)^{th}$ ODL model is obtained based on the input dataset (i.e., flame temperature maps) of the i^{th} 'unseen' condition. Secondly, the i^{th} ODL model is formed by re-training the $(i\text{-}1)^{th}$ ODL model using the dataset of the i^{th} 'unseen' condition and the output of the $(i\text{-}1)^{th}$ ODL model. If i=1, the 0^{th} ODL model is the source DL model.

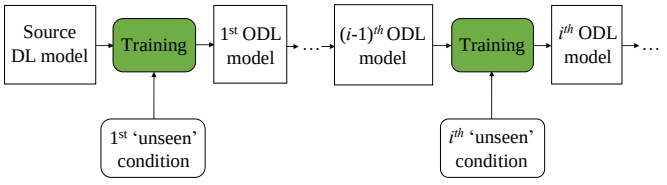

Fig. 3. Incremental learning process.

It is also worth noting that, in the proposed ODL model, all the 'unseen' data will become the 'seen' data once they are used to update the model. They are equally important and should be treated fairly based on their contributions to the updating process of the model. This is done by proposing a dedicated objective function where a loss balance weight is introduced to 'balance' the distillation loss (relating to all the 'seen' and previously 'unseen' conditions) and the regression loss (relating to the 'unseen' conditions only). A detailed description of the proposed objective function is given in the following section.

1) ODL's objective function

When building an ODL model, say the i^{th} ODL model, we intend to re-train the $(i\text{-}1)^{th}$ ODL model based on the dataset only from the i^{th} 'unseen' condition. Generally, the training purpose is to minimise the regression loss, i.e., the difference between the output of the i^{th} ODL model and its corresponding ground truth data, represented commonly as a mean squared error, so that the i^{th} ODL model is functional for the i^{th} 'unseen' condition. To reduce the forgetting problems and ensure that the i^{th} ODL model works for 'seen' conditions, a distillation loss which is the difference (again, represented as a mean square error) between the outputs of the $(i\text{-}1)^{th}$ and i^{th} ODL models is introduced in the objective function. A small distillation loss indicates that the parameter values are very similar between the i^{th} and $(i\text{-}1)^{th}$ ODL models, which means that the i^{th} ODL model can predict the NO_x for 'seen' conditions as the $(i\text{-}1)^{th}$ ODL model does. Therefore, it is important to minimise the regression and the distillation losses of the i^{th} ODL model. To achieve this goal, the following objective function is proposed,

$$\begin{aligned} &J_i = \eta_i L_d\left(\mathbf{Y}_{i-1}^o, \mathbf{Y}_i^o\right) + (1-\eta_i) L_r\left(\mathbf{Y}_i^s,\ \mathbf{Y}_i^o\right), \\ &\left(0 \le \eta_i \le 1,\ i=1, 2, \cdots, \text{n}\right) \end{aligned} \tag{8}$$

where J_i is the loss which carries the information of distillation loss L_d and regression loss L_r; η_i is the loss balance weight for the i^{th} ODL model; $\mathbf{Y}_{i\text{-}1}^o \in \mathbb{R}^{1\times m}$ is the output of the $(i\text{-}1)^{th}$ ODL model based on the input data of i^{th} 'unseen' condition; $\mathbf{Y}_i^o \in \mathbb{R}^{1\times m}$ is the output of the i^{th} ODL model; $\mathbf{Y}_i^s \in \mathbb{R}^{1\times m}$ is the scaled ground-truth (i.e., NO_x) vector from the i^{th} 'unseen' condition; n is the number of ODL model updates.

The regression loss, L_r, the distillation loss, L_d, and the output of the i^{th} ODL model, $\mathbf{Y}_i^o$, in (8) are defined as follows,

$$L_r\left(\mathbf{Y}_i^s,\ \mathbf{Y}_i^o\right) = \frac{1}{\text{m}}\left\|\mathbf{Y}_i^s - \mathbf{Y}_i^o\right\|_2^2, \tag{9}$$

$$L_d\left(\mathbf{Y}_{i-1}^o, \mathbf{Y}_i^o\right) = \frac{1}{\text{m}}\left\|\mathbf{Y}_{i-1}^o - \mathbf{Y}_i^o\right\|_2^2, \tag{10}$$

$$\begin{aligned} \mathbf{Y}_i^o = f_i(&\mathbf{X}_i, \mathbf{W}_{1(i)}, \mathbf{W}_{2(i)}, \ldots, \mathbf{W}_{\text{n}(i)},\ \mathbf{B}_{1(i)},\ \mathbf{B}_{2(i)},\ \\ &\ldots,\ \mathbf{B}_{\text{n}(i)}, \mathbf{W}_{\text{f1}(i)}, \mathbf{W}_{\text{f2}(i)},\ \mathbf{B}_{\text{f1}(i)},\ \mathbf{B}_{\text{f2}(i)}) \end{aligned} \tag{11}$$

where f_i represents the i^{th} ODL model; $\mathbf{X}_i \in \mathbb{R}^{a\times a\times m}$ is a tensor representing the temperature maps from the i^{th} 'unseen' condition; $\mathbf{W}_{1(i)}$, $\mathbf{W}_{2(i)}$, …, $\mathbf{W}_{\text{n}(i)}$, $\mathbf{B}_{1(i)}$, $\mathbf{B}_{2(i)}$, …, $\mathbf{B}_{\text{n}(i)}$, $\mathbf{W}_{\text{f1}(i)}$, $\mathbf{W}_{\text{f2}(i)}$, $\mathbf{B}_{\text{f1}(i)}$, $\mathbf{B}_{\text{f2}(i)}$ represent the parameters of the i^{th} ODL model.

To minimise J_i, all the parameter values of the i^{th} ODL model are optimised using the stochastic gradient descent method. As can be seen from (8)-(10), the minimum objective loss, $\min(J_i)$, happens when $\mathbf{Y}_i^o = \eta_i \mathbf{Y}_{i\text{-}1}^o + (1\text{-}\eta_i)\mathbf{Y}_i^s$, i.e.,

$$\min(J_i) = \frac{(2\eta_i^2 - 2\eta_i + 1)}{m} \left\| \mathbf{Y}_i^s - \mathbf{Y}_{i-1}^o \right\|_2^2 . \quad (12)$$

It is clear that the minimum objective loss is related to η_i, $\mathbf{Y}_i^s$ and $\mathbf{Y}_{i\text{-}1}^o$. The output $\mathbf{Y}_{i\text{-}1}^o$ of the $(i\text{-}1)^{th}$ ODL model generally resides in the domain of the scaled ground-truth of the 'seen' conditions. When $\mathbf{Y}_i^s$ is close to $\mathbf{Y}_{i\text{-}1}^o$, the parameter values of i^{th} ODL model are the same as that of the $(i\text{-}1)^{th}$ ODL model, and thus the i^{th} ODL model can predict the NO_x for the previously 'seen' and i^{th} 'unseen' conditions. In practice, however, $\mathbf{Y}_i^s$ can be very different from $\mathbf{Y}_{i\text{-}1}^o$, resulting in an unacceptable objective loss. One way to solve this problem is to scale the ground-truth data of the i^{th} 'unseen' condition to a new domain which is close to that of $\mathbf{Y}_{i\text{-}1}^o$, i.e.,

$$\mathbf{Y}_i^s = \frac{\mathbf{Y}_i - \mathbf{Y}_{\min} + D_i}{\mathbf{Y}_{\max} - \mathbf{Y}_{\min}} , \quad (13)$$

and,

$$D_i = \bar{\mathbf{Y}}_{\text{old}} - \bar{\mathbf{Y}}_{\text{new}(i)} , \quad (14)$$

where $\mathbf{Y}_i$ represents the ground-truth data of the i^{th} 'unseen' condition; $\mathbf{Y}_{\min}$ and $\mathbf{Y}_{\max}$ represent the minimum and maximum ground-truth values of the 'seen' and 'unseen' conditions, respectively; D_i represents the difference between the average ground-truth of the 'seen' conditions, $\bar{\mathbf{Y}}_{\text{old}}$, and that of the i^{th} 'unseen' condition, $\bar{\mathbf{Y}}_{\text{new}}$.

When testing input $\mathbf{X}$ is fed into the i^{th} ODL model, the output $\mathbf{Y}$ of which can be expressed as,

$$\mathbf{Y} = f_i(\mathbf{X}, \mathbf{W}_{1(i)}, \mathbf{W}_{2(i)}, ..., \mathbf{W}_{n(i)}, \mathbf{B}_{1(i)}, \mathbf{B}_{2(i)}, ..., \mathbf{B}_{n(i)}, \mathbf{W}_{f1(i)}, \mathbf{W}_{f2(i)}, \mathbf{B}_{f1(i)}, \mathbf{B}_{f2(i)}) . \quad (15)$$

Hence, the predicted NO_x, $\mathbf{Y}_{\text{pre}}$, is calculated by,

$$\mathbf{Y}_{\text{pre}} = \mathbf{Y}(\mathbf{Y}_{\max} - \mathbf{Y}_{\min}) - D_i + \mathbf{Y}_{\min} . \quad (16)$$

As D_i changes with the i^{th} condition, it's crucial to determine D_i through a condition recognition approach to obtain $\mathbf{Y}_{\text{pre}}$.

D. Condition recognition

During the experiments (refer to Section III), flame images and corresponding boiler operation data are recorded simultaneously for each test condition. Three different operation data, i.e., O_2 concentration, primary to total flow, and secondary-to-tertiary flow split (refer to Table I in Section III), are used for the condition recognition through the cosine similarity approach [31]. These operation data are noted as (x, y, z), where x, y, z represent the O_2 concentration, primary to total flow, and secondary-to-tertiary flow split, respectively. The cosine similarity between the operation data of condition A_j (i.e., a new condition which needs to be examined) and previously 'seen' conditions B_i (i.e., the conditions which are used to train or re-train the source DL or ODL model) are calculated by,

$$S = \frac{\mathbf{A}_j \cdot \mathbf{B}_i}{\|\mathbf{A}_j\| \|\mathbf{B}_i\|} = \frac{x_j x_i + y_j y_i + z_j z_i}{\sqrt{x_j^2 + y_j^2 + z_j^2} \sqrt{x_i^2 + y_i^2 + z_i^2}} , \quad (17)$$

where $\mathbf{A}_j=(x_j, y_j, z_j)$ and $\mathbf{B}_i=(x_i, y_i, z_i)$ are the vectors representing operation data from the j^{th} test and i^{th} test, respectively. If S=1 and $\|\mathbf{A}_j\|=\|\mathbf{B}_i\|$, then $i=j$, i.e., condition $\mathbf{A}_j$ is recognised as the i^{th} condition. If the criterion is not met, $\mathbf{A}_j$ is recognised as an 'unseen' condition.

III. Collection Of Experimental Data

A. Experimental setup

Experimental data were obtained from oxy-biomass flame tests carried out on the UKCCSRC PACT 250 kW_{th} Air/Oxy-fuel CTF [32]. The CTF has a single burner firing down in a cylindrical combustion chamber with an inner diameter of 0.9 m. The burner is a purpose-designed Low-NO_x biomass-fired burner with a primary annulus for conveying pulverised biomass fuels. The rest of the oxidiser is introduced through swirled secondary and tertiary annuli to ensure the completion of the combustion. The burner also has an internal flow splitting system to control the secondary to tertiary flow ratio. A flame imaging system [23] was used for flame image acquisition. Fig. 4 is the block diagram of the CTF and the flame imaging system, whilst Fig. 5 shows the site installation of the system. The flame imaging system mainly consists of an optical probe with a water jacket, a digital camera, and a personal computer with system application software. The probe is a rigid endoscope with a 90° field of view and protected by a water-air cooled jacket. The camera is an industrial RGB CMOS camera with an image resolution of 1216 (H) × 1936 (V) pixels and a frame rate of up to 900 frames per second with reduced resolution. The probe was installed at the viewport on a top section of the furnace where the burner quarl and primary reaction zone of the flame was fully visualised.

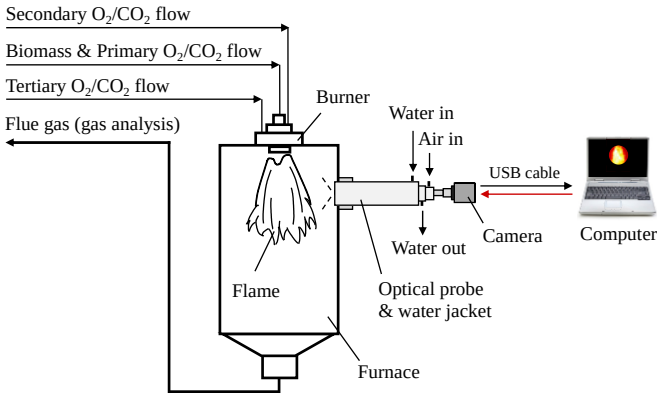

Fig. 4. Block diagram of the CTF and the flame imaging system.

B. Test conditions

A total of ten test conditions were conducted using a pulverised whitewood as biomass fuel under different oxy-combustion conditions with a fixed furnace load of 150 kW_{th}. The recycled flue gas was simulated by using CO_2. The secondary and tertiary flows (CO_2+O_2) were used as rig variables to obtain different percentages of O_2, primary to total flow, and secondary to tertiary flow ratio (split). Table I

summarises the test conditions and corresponding operation data.

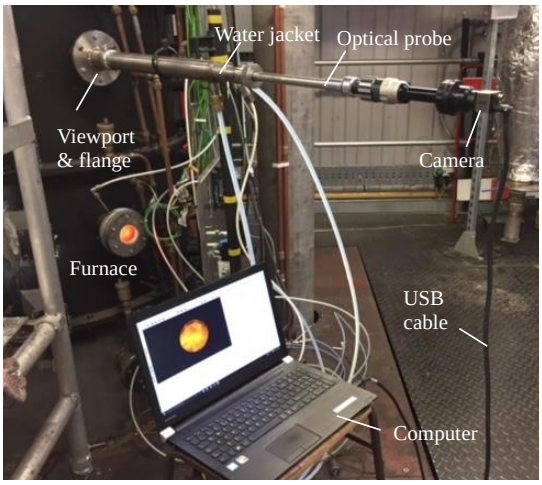

Fig. 5. Site installation of the flame imaging system.

TABLE I

SUMMARY OF TEST CONDITIONS

Test	O_2 (vol %)	Primary to total flow (%)	Secondary to tertiary flow split	Mean NO_x in flue gas (ppm)	NO_x STD (ppm)	Used as 'seen' or 'unseen' dataset
1	21	24	4	73	1	Seen
2	23	24	4	83	2	Seen
3	25	24	4	79	2	Seen
4	27	24	4	78	2	Unseen
5	27	21	4	67	2	Unseen
6	27	27	4	81	2	Seen
7	27	30	4	94	2	Seen
8	27	24	1	84	4	Unseen
9	27	24	2	73	3	Seen
10	27	24	3	75	2	Seen

For each test condition, a total of 150-170 flame images were captured by the imaging system over about four minutes when the condition was stable. NO_x data were also taken concurrently by a gas analyser from the flue gas (Table I), which were then used as the reference NO_x (ground-truth) for the model construction. Fig. 6 illustrates the typical flame images and corresponding temperature maps under different test conditions. The flame temperature distributions were computed based on the two-colour method [23]. It should be mentioned that flame images from some tests were taken under different camera settings to avoid possible image saturation. Previous studies reveal that the two-colour method-based flame imaging system appears to be linear [23]. This means that the camera settings, such as the exposure time, iris, and objective distance, have no or a very limited impact on the temperature measurement. In addition, it is found that the temperature of the furnace wall (shown as the light blue in the flame temperature maps in Fig. 6), was consistent throughout the tests, and thus should have no or limited influence on the accuracy of the proposed model. It is for this reason, the temperature maps of the complete field of view were used to form the input datasets.

The flame and NO_x emission data from Tests 1-3, 6, 7, 9 and 10 were used to build the source DL model whilst those from Tests 4, 5 and 8 (Table I) were used to train, validate and test the ODL model. Each dataset was randomly split as a training dataset (75%), a validation dataset (12.5%), and a testing dataset (12.5%). The Python language was deployed for programming. The source DL and ODL models were trained, validated, and tested on a personal computer with an intel i9-10900K processor, 32GB RAM and 8GB GPU.

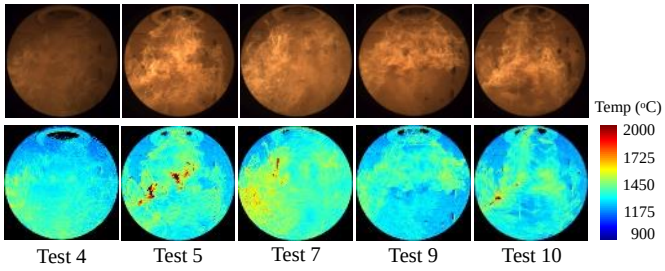

Fig. 6. Example flame images and corresponding temperature maps under different test conditions.

IV. MODEL CONSTRUCTION

A. Evaluation criteria

MAPE (Mean Absolute Percentage Error), σ (Standard Deviation of Absolute Percentage Error) and APE_{max} (Maximum Absolute Percentage Error), are deployed to evaluate the model performance, and they are defined as,

$$\mathrm{MAPE}=\frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\left|\frac{\bar{\mathbf{Y}}_{\mathrm{pre}(i)}-\bar{\mathbf{Y}}_{\mathrm{true}(i)}}{\bar{\mathbf{Y}}_{\mathrm{true}(i)}}\right|\times 100\%, \tag{18}$$

$$\sigma=\sqrt{\frac{\sum_{j=1}^{\mathrm{n}}\left(\left|\frac{\bar{\mathbf{Y}}_{\mathrm{pre}(i)}-\bar{\mathbf{Y}}_{\mathrm{true}(i)}}{\bar{\mathbf{Y}}_{\mathrm{true}(i)}}\right|-\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left|\frac{\bar{\mathbf{Y}}_{\mathrm{pre}(i)}-\bar{\mathbf{Y}}_{\mathrm{true}(i)}}{\bar{\mathbf{Y}}_{\mathrm{true}(i)}}\right|\right)^2}{\mathrm{n}-1}}\times 100\%, \tag{19}$$

$$\mathrm{APE}_{\max}=\max\left(\left|\frac{\bar{\mathbf{Y}}_{\mathrm{pre}(1)}-\bar{\mathbf{Y}}_{\mathrm{true}(1)}}{\bar{\mathbf{Y}}_{\mathrm{true}(1)}}\right|,\left|\frac{\bar{\mathbf{Y}}_{\mathrm{pre}(2)}-\bar{\mathbf{Y}}_{\mathrm{true}(2)}}{\bar{\mathbf{Y}}_{\mathrm{true}(2)}}\right|,\ldots,\left|\frac{\bar{\mathbf{Y}}_{\mathrm{pre}(n)}-\bar{\mathbf{Y}}_{\mathrm{true}(n)}}{\bar{\mathbf{Y}}_{\mathrm{true}(n)}}\right|\right)\times 100\% \tag{20}$$

where $\bar{\mathbf{Y}}_{\mathrm{true}(i)}$ and $\bar{\mathbf{Y}}_{\mathrm{pre}(i)}$ are the average reference and predicted NO_x values for the i^{th} combustion condition, respectively, n is the number of combustion conditions. In addition, the training time is considered as it is also an important criterion, particularly for the online model construction and practical uses.

It is worth mentioning that, in comparison to other commonly used scale-dependent evaluation metrics, the MAPE is one of the most accepted criteria for evaluating the model's performance in regression modelling. This is due to its advantages of scale-independency and intuitive interpretation in terms of relative error [33]. Meanwhile, the APE_{max} is very effective to evaluate the worst case of the proposed model, as it provides the maximum absolute percentage error among all the conditions.

B. Determination of hyper-parameters

The determination of optimum hyper-parameters such as training epoch, input image size, batch size, and the number of convolutional layers of the source DL model is crucial for

achieving satisfactory performance of the proposed in teams of the prediction accuracy, computational time and the usage of system resources. In this study, a grid search based trial-and-error method was deployed to determine the optimum hyper-parameters.

1) Training epoch

The training epoch needs to be defined to control the stop point of the training process. Fig. 7 shows the MAPEs of the training and validation losses when the source DL model is trained under different training epochs, with the input image size of 144×144, batch size of 64, and convolutional layers of 4. The validation loss is slightly higher than the training loss, which is expected as the validation dataset is not used to train the model. It can be seen that the training and validation losses decrease with the epochs and tend to be unchanged (around 0.37% and 0.59% MAPEs for the training and validation, respectively) after 7500. It was therefore decided to set a training epoch of 7500 throughout the construction process of the model.

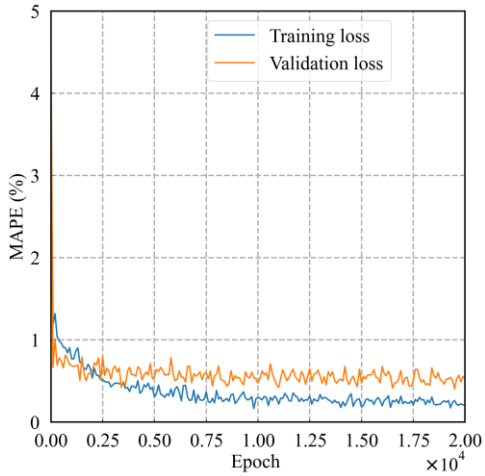

Fig. 7. Variation of MAPE for the training and validation losses against different epochs.

2) Input image size

The original flame image size is 719×719, which was scaled down to 288×288, 144×144, and 72×72, respectively. The scaled images were then used as the inputs of the source DL model with the batch size of 64, convolutional layers of 4, and training epoch 7500. Due to the influence of dropout, the model output slightly changes when the model was tested using the same Validation dataset, at different times. Hence, the model was tested 5 times for each Validation dataset, and their average output was used as the model output. Table II shows the variations of the MAPE and training time of the model for the different hyper-parameters. It is clear that the larger the input image size, the smaller the MAPE, and thus the higher the prediction accuracy, but the longer the training time. The results suggest that the input image size of 144×144 is a trade-off between the prediction accuracy (MAPE of 0.57%) and the training time.

3) Batch size

The batch size is the number of samples that are considered in each iteration during the model training process. The optimal batch size can lead to a better performance of the model and also reduce the training time. Thus, the batch size needs to be defined. As can be seen in Table II, the MAPE increases slightly with the batch size but remains almost unchanged when the batch size is 64 or greater. However, the larger the batch size is, the greater GPU capacity is required [34]. In practical uses, the model with less demanding computational resources is preferable. It is hence decided to use the batch size of 64 in the model construction.

4) The number of convolutional layers

The number of convolutional layers is an important hyper-parameter in the construction of the source DL model. Generally, the more convolutional layers, the deeper features can be extracted, while it could lead to longer training time and overfitting problems. Therefore, it is crucial to determine the optimum convolutional layers for the model. As can be seen in Table II, the MAPE is almost unchanged for the different numbers of convolutional layers. However, the more convolutional layers resulted in increased training time considerably. It is therefore 4 convolutional layers were chosen in this study.

TABLE II

MAPES AND TRAINING TIMES AGAINST DIFFERENT HYPER-PARAMETERS

		MAPE (%)	Training time (h)
Input image size (Batch size: 64, Convolutional layer: 4)	72×72	0.72	1.20
	144×144	**0.57**	**4.93**
	288×288	0.31	19.50
Batch size (Input image size: 144×144, Convolutional layer: 4)	32	0.42	5.15
	64	**0.57**	**4.93**
Number of convolutional layers (Input image size: 144×144, Batch size: 64)	**4**	**0.57**	**4.93**
	8	0.56	11.58
	12	0.63	17.66

V. RESULTS AND DISCUSSION

A. Source DL model

Following the determination of the hyper-parameters, the source DL model was evaluated using the testing datasets from all 10 test conditions as given in Table I. Fig. 8 illustrates the predicted NO_x based on the source DL model against the reference NO_x (ground-truth) for the 'seen' and 'unseen' conditions. Note that, in the figure, dash lines 'y=x-4' and 'y=x+4' represent the ideal line y=x plus and minus the maximum standard deviation (STD=4 ppm, Test 8) of the reference NO_x. As can be seen, the predicted NO_x values for the 'seen' conditions are mostly within the STD of ±4 ppm. The MAPE of the predicted NO_x is 0.60% for all the 'seen' conditions, suggesting that the source DL model can predict NO_x accurately if the conditions have been 'seen'. The source DL model gives higher prediction accuracy for the testing dataset from 'unseen' condition 1 (Test 4 as given in Table I), as the reference NO_x value under Test 4 is close to that under the 'seen' condition Test 3. However, the MAPE increases dramatically up to 15.49%, 14.01% for the testing dataset from 'unseen' condition 2 (Test 5) and 'unseen' condition 3 (Test 8), respectively. The results suggest that the source DL model provides poor prediction accuracy for the 'unseen' conditions, and thus an updated model is required.

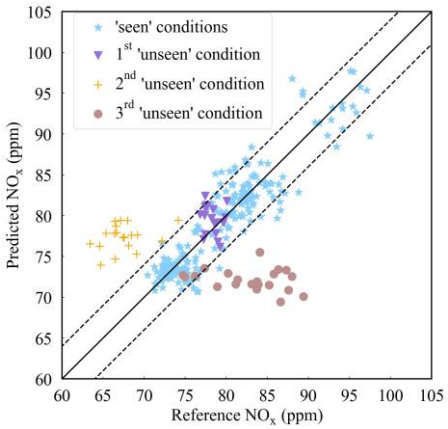

Fig. 8. Predicted and reference NO_x values based on the source DL model.

B. ODL model

As the ODL model is constructed based on the source DL model, the hyper-parameters are kept the same as that of the source model, i.e., the input image size of 144×144, the batch size of 64, and the convolutional layers of 4. However, the training epoch applied is 100, which is determined after a careful examination of the training process to ensure its convergence after the selected training epoch. It is worth noting that the epoch of 100 is much smaller than that used in the source model training (7500). This is due to the fact that, in the updating process, only the training dataset from one 'unseen' condition is used to maintain a high computational efficiency.

In this study, the 1st update is performed based on the source DL model and the training dataset from the 1st 'unseen' condition (Test 4); the 2nd update is based on the 1st ODL model and the training dataset from the 2nd 'unseen' condition (Test 5); the 3rd update is based on the 2nd ODL model and the training dataset from the 3rd 'unseen' condition (Test 8). The validation datasets from 'seen' and 'unseen' conditions were applied to examine the performance of the updated model. It is worth mentioning that, in this study, the model is updated continuously for three 'unseen' conditions due to the availability of 'unseen' conditions. However, the model can be updated continuously if more 'unseen' conditions are available.

1) *Impact of loss balance weight on the model's performance*

As described in Section II, the loss balance weight is introduced to the objective function of the ODL model to 'balance' the performance of the ODL model between the 'seen' and 'unseen' conditions. A larger η_i tunes the i^{th} ODL model to the $(i\text{-}1)^{th}$ ODL model, leading to a better performance of the i^{th} ODL model for the 'seen' conditions. On the other hand, a smaller η_i tunes the $(i\text{-}1)^{th}$ ODL model to the i^{th} ODL model, leading to a better performance of the i^{th} ODL model for the 'unseen' conditions.

To determine the appropriate value of η_i in each update, a grid search technique is applied, where η_i is set between 0 to 1 with an interval of 0.1. The training and validation datasets are then applied to the model for the different values of η_i. Fig. 9 illustrates the MAPEs of the ODL model for different values of η_i when the model was updated by the training and validation datasets from three 'unseen' conditions incrementally. Note that the number of the 'seen' conditions varies depending upon the number of updates as the 'unseen' condition which is used to update the ODL model becomes one of the 'seen' conditions in the next updated ODL model. Thus, the 'seen' conditions shown in Fig. 9 (a) are Tests 1-4, 6, 7, 9-10, in Fig. 9 (b) are Tests 1-7, 9-10, and in Fig. 9 (c) are Tests 1-10.

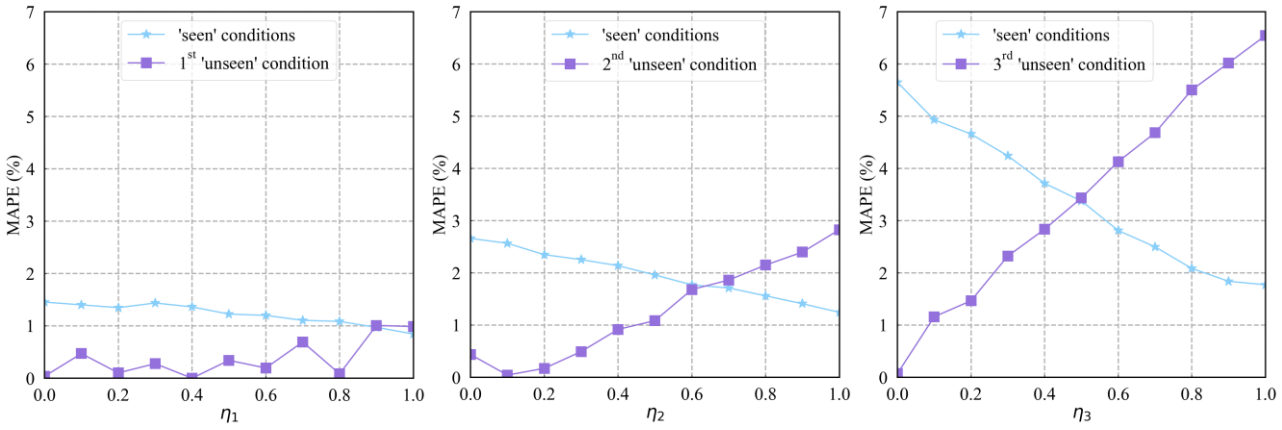

Fig. 9. Variation of MAPE with η_i for the i^{th} updates (i=1, 2, 3). (a) the 1st update, (b) the 2nd update, (c) the 3rd update.

As can be seen, a larger η_i leads to higher validation's MAPEs for the 'unseen' condition and lower validation's MAPEs for the 'seen' conditions. It is found that the appropriate value of η_i would be in a range between 0.6 and 0.8. Fig. 10 shows the variation of the APE_{max} with η_i for three updates using the validation datasets from the 'seen' and 'unseen' conditions. The APE_{max} decreases with η_i in the 1st and 2nd updates. This is because, as η_i increases, the contribution of the regression loss is reduced, and the distillation loss dominates the total loss of the model (refer to 8). It also means that the performance of the i^{th} ODL model is close to that of the $(i\text{-}1)^{th}$ ODL model (or the source model). When η_i reaches 1, the contribution of the regression loss is 'zero', and only the distillation loss exists. In other words, the model remains unchanged. It is also noted that the mean ground truth (NO_x) of the 2nd 'unseen' condition (Test 5) is significantly different from other 'seen' conditions (maybe due to the reduced primary to total flow). This explains that the APE_{max} of the 2nd update is

higher than that of the 1st update. The APE_{max} of the 3rd update is consistently higher than that of the 1st and 2nd updates because the 3rd 'unseen' condition (Test 8) is very different from the 'seen' conditions (refer to Table I). However, it does not show a monotonically decreasing trend. This is believed to attribute a greater standard deviation in comparison with the 'seen' and the 1st and 2nd 'unseen' conditions. Therefore, a loss balance weight of 0.7 is considered to be a trade-off among the 'unseen' conditions examined.

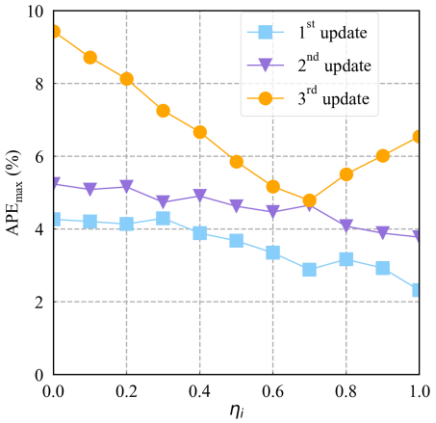

Fig. 10. Variation of APE_{max} with η_i for the i^{th} updates (i=1, 2, 3).

2) Performance of ODL model

Figs. 11 illustrates the NO_x emissions predicted using the ODL model for three updates, respectively (η_1=η_2=η_3=0.7). It can be seen that the predicted NO_x emissions for 'seen' and 'unseen' conditions are around the reference line 'y=x' for the three updates. Comparing to the results in Fig. 8, it is evident that the prediction accuracy of the ODL model was significantly improved through the incremental updating.

Table III summarises the prediction errors of the source DL and ODL models for the 'seen' and 'unseen' testing datasets. As can been seen, although the prediction errors of the models have increased due to the updating, they remain small, e.g., the MAPE is 2.56% with a standard deviation of 1.45%, and the APE_{max} is no greater than 5%, after three updates. The results have revealed that the proposed technical approach is effective to allow the ODL model to be learnt incrementally based on the datasets only from 'unseen' conditions. In addition, the updated model enables to predict the NO_x emissions under 'seen' and 'unseen' conditions with acceptable errors. Furthermore, the ODL model can be updated by an 'unseen' condition dataset in a very short period (0.01 hours). Note that the training time for each update varies slightly, but is rounded to 0.01 hours. In comparison, it will take more than 4.93 hours when more and more stream data from 'unseen' conditions are combined with the data from 'seen' conditions to train a DL model. This has shown great potentials of the proposed ODL approach for the online NO_x prediction in practical combustion systems with variable operation conditions.

TABLE III

PREDICTION ERRORS OF THE SOURCE DL AND ODL MODELS FOR THE 'SEEN' AND 'UNSEEN' TESTING DATASETS

	Source DL	1st updated ODL	2nd updated ODL	3rd updated ODL
Test	1-3, 6, 7, 9, 10	1-4, 6, 7, 9, 10	1-7, 9, 10	1-10
MAPE (%)	0.60	1.05	1.46	2.56
APE_{max} (%)	1.47	2.42	3.84	4.87
σ (%)	0.59	0.77	1.25	1.45
Training time (hour)	4.93	0.01	0.01	0.01

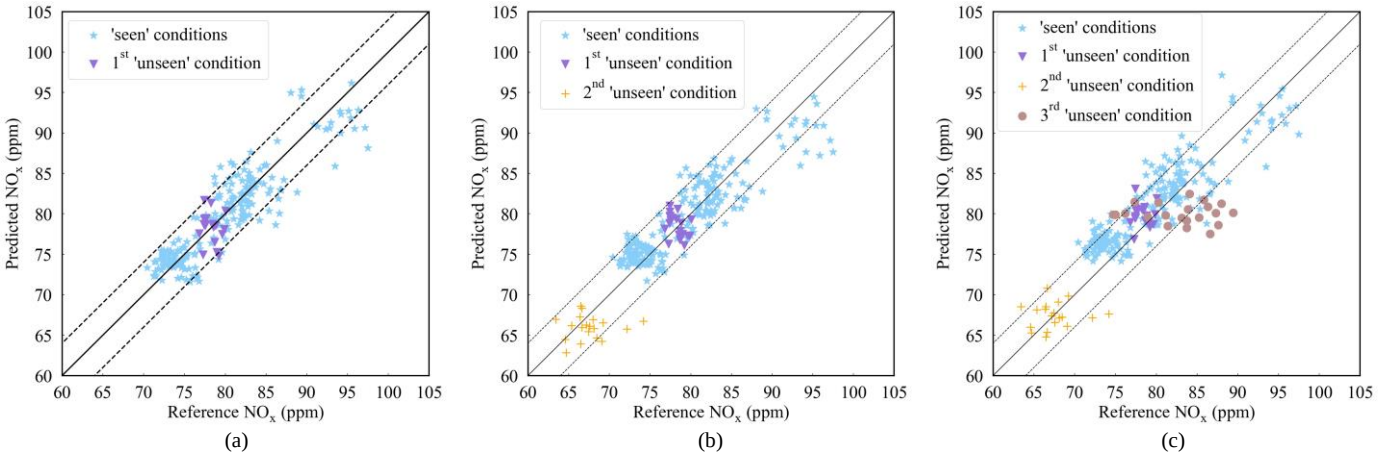

Fig. 11. Predicted and reference NO_x values based on the 1st, 2nd and 3rd updated ODL models. (a) the 1st updated ODL model, (b) the 2nd updated ODL model, (c) the 3rd updated ODL model.

It is also worth mentioning that, at the training stage, the input (flame temperature maps) and ground truth (NO_x values) at each batch are randomly sampled (by the model being trained) from a 'seen' or 'unseen' training dataset. There is no particular input order among the different 'seen' and 'unseen' training datasets during the training process. At the testing stage, on the other hand, the model was tested by the testing datasets from 'seen' and 'unseen' conditions. The testing data were fed into the model one by one with no particular order. Thus, the input order of different 'seen' and 'unseen' testing datasets to the model has no influence on the model's performance.

It is thought that the problem of increased error with the number of updates could be minimised by introducing a small portion of the dataset from 'seen' conditions to the updating process, which would be a trade-off between the accuracy and

efficiency of the ODL model. It has been recognised, however, that further work needs to be done to examine the performance of the ODL model for a greater number of updates.

VI. Conclusion

A flame imaging based ODL model has been developed for predicting NO_x emissions from an oxy-biomass combustion process. Flame temperature maps obtained under a range of operation conditions on an Air/Oxy-fuel Combustion Test Facility were used as the input datasets. The ODL model has been constructed based on CNN (the source model) by updating the model incrementally only using 'unseen' datasets to reach improved accuracy and efficiency. This has also been achieved by introducing a loss balance weight into the objective function. The criteria for choosing the loss balance weight have carefully been set to ensure that the ODL model would not lose information from 'seen' and 'unseen' conditions. The test results have revealed that the proposed ODL model is capable of predicting NO_x emissions for 'seen' and 'unseen' conditions with a mean absolute percentage error less than 3% even after three updates. In comparison with the conventional CNN, the ODL model has an incremental learning ability which has given a great potential of the proposed ODL model for the online NO_x prediction in practical combustion systems under variable operation conditions with improved accuracy and efficiency. Further work will focus on examining the performance of the ODL model for more updates so that the accuracy, reliability, and practicability of the model can be further assessed. Finally, the proposed ODL approach has potentials not only for the NO_x prediction of combustion processes, but also other engineering applications where online condition monitoring and prediction are required.

Acknowledgement

The authors acknowledge that the UKCCSRC PACT Facilities, funded by the Department for Business, Energy & Industrial Strategy, and the Engineering and Physical Sciences Research Council, UK, were used for the experimental work reported in this publication. Dr Janos Szuhanszki at the University of Sheffield is thanked for his contribution to the data collection of this study by conducting the tests described. Dr Xue Wang is also acknowledged for her technical advice to this study.

References

[1] R. Saidur, E. A. Abdelaziz, A. Demirbas, M. S. Hossain, and S. Mekhilef, "A review on biomass as a fuel for boilers," *Renew. Sustain. Energy Rev.*, vol. 15, no. 5, pp. 2262–2289, Jun. 2011.

[2] L. Wang, N. Karimi, T. Sutardi, and M. C. Paul, "Combustion characteristics and pollutant emissions in transient oxy-combustion of a single biomass particle: a numerical study," *Energy and Fuels*, vol. 33, no. 2, pp. 1556–1569, Jan. 2019.

[3] A. Zhou, H. Xu, M. Xu, W. Yu, and Z. Li, "Numerical investigation of biomass co-combustion with methane for NO_x reduction," *Energy*, vol. 194, p. 116868, Mar. 2020.

[4] I. A. Shah, X. Gou, Q. Zhang, J. Wu, E. Wang, and Y. Liu, "Experimental study on NO_x emission characteristics of oxy-biomass combustion," *J. Clean. Prod.*, vol. 199, pp. 400–410, Oct. 2018.

[5] H. Liu, J. Chaney, J. Li, and C. Sun, "Control of NO_x emissions of a domestic/small-scale biomass pellet boiler by air staging," *Fuel*, vol. 103, pp. 792–798, Jan. 2013.

[6] J. Chang, X. Wang, Z. Zhou, H. Chen, and Y. Niu, "CFD modeling of hydrodynamics, combustion and NO_x emission in a tangentially fired pulverized-coal boiler at low load operating conditions," *Adv. Powder Technol.*, vol. 32, no. 2, pp. 290–303, Feb. 2021.

[7] J. F. Tuttle, L. D. Blackburn, and K. M. Powell, "On-line classification of coal combustion quality using nonlinear SVM for improved neural network NO_x emission rate prediction," *Comput. Chem. Eng.*, vol. 141, p. 106990, Oct. 2020.

[8] G. Wang, O. I. Awad, S. Liu, S. Shuai, and Z. Wang, "NO_x emissions prediction based on mutual information and back propagation neural network using correlation quantitative analysis," *Energy*, vol. 198, p. 117286, May 2020.

[9] T. Yang, K. Ma, Y. Lv, and Y. Bai, "Real-time dynamic prediction model of NO_x emission of coal-fired boilers under variable load conditions," *Fuel*, vol. 274, p. 117811, Aug. 2020.

[10] W. Fan, F. Si, S. Ren, C. Yu, Y. Cui, and P. Wang, "Integration of continuous restricted Boltzmann machine and SVR in NO_x emissions prediction of a tangential firing boiler," *Chemom. Intell. Lab. Syst.*, vol. 195, p. 103870, Dec. 2019.

[11] P. Tan, J. Xia, C. Zhang, Q. Fang, and G. Chen, "Modeling and reduction of NO_X emissions for a 700 MW coal-fired boiler with the advanced machine learning method," *Energy*, vol. 94, pp. 672–679, Jan. 2016.

[12] F. Wang, S. Ma, H. Wang, Y. Li, and J. Zhang, "Prediction of NO_x emission for coal-fired boilers based on deep belief network," *Control Eng. Pract.*, vol. 80, pp. 26–35, Nov. 2018.

[13] N. Li, G. Lu, X. Li, and Y. Yan, "Prediction of pollutant emissions of biomass flames through digital imaging, contourlet transform, and support vector regression modeling," *IEEE Trans. Instrum. Meas.*, vol. 64, no. 9, pp. 2409–2416, Sept. 2015.

[14] N. Li, G. Lu, X. Li, and Y. Yan, "Prediction of NO_x emissions from a biomass fired combustion process based on flame radical imaging and deep learning techniques," *Combust. Sci. Technol.*, vol. 188, no. 2, pp. 233–246, 2016.

[15] Y. Zhai, X. Ding, X. Jin, and L. Zhao, "Adaptive LSSVM based iterative prediction method for NO_x concentration prediction in coal-fired power plant considering system delay," *Appl. Soft Comput. J.*, vol. 89, p. 106070, Apr. 2020.

[16] Y. Lv, T. Yang, and J. Liu, "An adaptive least squares support vector machine model with a novel update for NO_x emission prediction," *Chemom. Intell. Lab. Syst.*, vol. 145, pp. 103–113, Jul. 2015.

[17] G. Lu, Y. Yan, S. Cornwell, M. Whitehouse, and G. Riley, "Impact of co-firing coal and biomass on flame characteristics and stability," *Fuel*, vol. 87, no. 7, pp. 1133–1140, Jun. 2008.

[18] D. Sun, G. Lu, H. Zhou, and Y. Yan, "Condition monitoring of combustion processes through flame imaging and kernel principal component analysis," *Combust. Sci. Technol.*, vol. 185, no. 9, pp. 1400–1413, 2013.

[19] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.

[20] P. Dhar, R. V. Singh, K. C. Peng, Z. Wu, and R. Chellappa, "Learning without memorizing," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 5133–5141.

[21] N. M. Aszemi and P. D. D. Dominic, "Hyperparameter optimization in convolutional neural network using genetic algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 269–278, 2019.

[22] W. E. L. Ilboudo, T. Kobayashi, and K. Sugimoto, "Robust stochastic gradient descent with student-t distribution based first-order momentum," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–14, 2020.

[23] D. Sun, G. Lu, H. Zhou, and Y. Yan, "Flame stability monitoring and characterization through digital imaging and spectral analysis," *Meas. Sci. Technol.*, vol. 22, no. 11, 2011.

[24] H. Wang, S. Member, J. Xu, R. Yan, S. Member, and R. X. Gao, "A new intelligent bearing fault diagnosis method using SDP representation and SE-CNN," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2377-2389, May 2020.

[25] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *32nd Int. Conf. Mach. Learn. ICML 2015*, 2015, pp. 448–456.

[27] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013.

[28] N. Akhtar and U. Ragavendran, "Interpretation of intelligence in CNN-pooling processes: a methodological survey," *Neural Comput. Appl.*, vol.

32, no. 3, pp. 879–898, 2020.
[29] W. J. Lee, H. Wu, A. Huang, and J. W. Sutherland, “Learning via acceleration spectrograms of a DC motor system with application to condition monitoring,” *Int. J. Adv. Manuf. Technol.*, vol. 106, no. 3–4, pp. 803–816, 2020.
[30] C. Garbin, X. Zhu, and O. Marques, “Dropout vs. batch normalization: an empirical study of their impact to deep learning,” *Multimed. Tools Appl.*, vol. 79, no. 19–20, pp. 12777–12815, 2020.
[31] A. W. Qurashi, V. Holmes, and A. P. Johnson, “Document Processing: Methods for semantic text similarity analysis,” *Int. Conf. Innov. Intell. Syst. Appl. Proc.*, Novi Sad, Serbia, 2020, pp. 1-6.
[32] O. F. Moguel, J. Szuhánszki, A. G. Clements, D. B. Ingham, L. Ma, and M. Pourkashanian, “Oscillating coal and biomass flames: a spectral and digital imaging approach for air and oxyfuel conditions,” *Fuel Process. Technol.*, vol. 173, pp. 243–252, May 2018.
[33] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, “Mean Absolute Percentage Error for regression models, ” Neurocomputing, vol. 192, pp. 38–48, 2016.
[34] L. J. Ma, G. Montague, J.Y. Ye, Z.W. Yao, A. Gholami, K. Keutzer, M. W. Mahoney, “Inefficiency of K-FAC for large batch size training,” AAAI 2020 - 34th AAAI Conf. Artif. Intell., New York, USA, 2020, pp. 5053–5060.

Li Qin received the B.Eng. degree in Thermal Energy and Power Engineering, and the M.Sc. degree in Power Engineering from North China Electric Power University, Beijing, China, in 2015 and 2018, respectively. She worked as a research assistant in the Institute of Mechanics, Chinese Academy of Sciences, Beijing, China, in 2018. She is currently pursuing the Ph.D. degree in Electronic Engineering in the School of Engineering, University of Kent, Canterbury, U.K. Her research interests include incremental learning, online learning, inverse problems, instrumentation and measurement, condition monitoring, emission prediction and Net-zero.

Gang Lu (Senior Member, IEEE) received the B.Eng. degree in Mechanical Engineering from Central South University, Changsha, China, in 1982. He worked as an engineer and senior engineer with the Institute of Metallurgical Equipment Design and Research, Beijing, China, from 1982 to 1996. He joined Teesside Technology Centre, British Steel, Middlesbrough, U.K., as a visiting research engineer for a year in 1994. Dr Lu returned to academia in 1996 and obtained the Ph.D. degree in Electronic Engineering from the University of Greenwich, Chatham, U.K., in 2000. He is currently a Reader in Electronic Instrumentation with the School of Engineering, University of Kent, Canterbury, U.K. Dr Lu’s research interests include imaging sensors, combustion instrumentation, condition monitoring, and machine learning for engineering applications.

Md Moinul Hossain (Senior Member, IEEE) received the B.Sc. degree in Computer Science and Engineering from International Islamic University, Bangladesh, in 2005, the M.Sc. degree in Wireless Communications and Systems Engineering from the University of Greenwich, Chatham, U.K., in 2008, and the Ph.D. degree in Electronic Engineering from the University of Kent, Canterbury, U.K., in 2014. He is currently a Lecturer in Electronic Engineering with the School of Engineering, University of Kent. His research interests include combustion diagnostics, sensors, instrumentation and measurement, condition process monitoring, machine learning and light field sensing. He serves as an Editorial Board member of IEEE Transactions on Instrumentation and Measurement and IEEE Access Journals.

Andy Morris received the B.Sc. degree in Mechanical Engineering from Nottingham Trent University, Nottingham, U.K., in 1985. He is a Fellow of the Institution of Mechanical Engineers and his career extends over 35 years with experience on plant life assessment, life extension, failure investigations, risk mitigation for operating plant, supporting operational generating sites on in-service problems and undertaking research in the Power Industry on Nuclear, Conventional and Renewables assets. Andy has worked in both technical and managerial positions in the Power Industry and has spent significant periods of his career to date at Rolls-Royce Marine Power, E.ON and currently with EDF Energy as Chief Mechanical Engineer in the Coal and Gas Operations business. He is a Visiting Professor at Imperial College London, U.K., and an Honorary Professor at the University of Nottingham, U.K., and has been actively researching plant life assessment and practical condition monitoring methods for high temperature materials developing field deployable optical strain measurement systems for high temperature materials, steam and gas turbine materials research and wind turbine blade composites in relation to strain mapping, inspection methods and failure modes. These research projects have been conducted with several U.K. universities and external research organisations.

Yong Yan (Fellow, IEEE) received the B.Eng. and M.Sc. degrees in Instrumentation and Control Engineering from Tsinghua University, Beijing, China, in 1985 and 1988, respectively, and the Ph.D. degree in Flow Measurement and Instrumentation from the University of Teesside, Middlesbrough, U.K., in 1992. He was an Assistant Lecturer with the Department of Automation at Tsinghua University from September 1988 to September 1989. He joined the University of Teesside as a Research Assistant in October 1989. After a short period of Post-Doctoral Research, he was a Lecturer with the University of Teesside from 1993 to 1996, and then as a Senior Lecturer, a Reader, and a Professor with the University of Greenwich, Chatham, U.K., from 1996 to 2004. He is currently a Professor of Electronic Instrumentation and the Director of Innovation at the School of Engineering, University of Kent, Canterbury, U.K. Prof. Yan was elected as a Fellow of the Royal Academy of Engineering in 2020. He was awarded the gold medal in 2020 by the IEEE Transactions on Instrumentation and Measurement as the most published author of all time from the U.K. His research interests include multiphase flow measurement, combustion instrumentation, and condition monitoring.