IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 August 2023, accepted 5 September 2023, date of publication 12 September 2023,
date of current version 18 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3314490

APPLIED RESEARCH

Transfer Learning-Driven Hourly PM2.5 Prediction Based on a Modified Hybrid Deep Learning

JUNZI YANG[1,2], AJUNE WANIS ISMAIL[1], YINGYING LI[1,3], LIMIN ZHANG[2], AND FAZLIATY EDORA FADZLI[1]

[1]Mixed and Virtual Reality Research Laboratory (Vicubelab), Faculty of computing, Universiti Teknologi Malaysia (UTM), Skudai, Johor 81310, Malaysia
[2]School of Mathematics and Computer Science, Hengshui University, Hengshui 053000, China
[3]School of Software, Shanxi Agricultural University, Jinzhong 030801, China

Corresponding author: Junzi Yang (yangjunzi@graduate.utm.my)

This work was supported in part by the Institute of Applied Mathematics, Hengshui University; in part by the 2022 Ministry of Education Industry-University Collaborative Education Project 220906627303255; and in part by the 2022 University Supply and Demand Matching Employment and Education Project of Ministry of Education under Grant 20220105953.

ABSTRACT Haze is a major problem in China's air pollution, which not only hinders economic development, but also causes harm to people's health. PM2.5 (fine particulate matter) is the primary cause of haze. Therefore, the timely prevention and control of haze benefit the precise forecast of PM2.5 concentration. Air quality has high-dimensional, non-linear and complex characteristics. In this paper, a modified hybrid deep learning model is proposed under the framework of transfer learning, which can solve the problem of air quality prediction in the case of sparse data. The research focuses on solving the problem of inadequate feature extraction in existing studies, and and predicts PM2.5 concentration at multiple sites. In the domain of adaptive extraction of air pollutant characteristics, long and short-term neural networks and multi-layer perceptron are used to realize the long-term dependence and nonlinear transformation of features, respectively. The learned features can be shared by PM2.5 prediction tasks at multiple sites. The channel and spatial attention mechanisms are added to extract the key information in the representation target. In the whole network, the residual neural unit is used to increase the depth of the network and improve prediction accuracy. This paper discusses the experimental results in Beijing dataset from 2013 to 2017 and Hengshui dataset from 2020 to 2022. Based on the findings, it shows that compared with the classical deep learning models, hybrid deep learning models and the most recent transfer learning approaches, the network can obtain higher accuracy and better robustness, especially for the prediction of sites with sparse data. The RMSE value of TL-Modified compared with TL-LSTM and TL-CNN-LSTM models decreased by 38 %, 16.5% and 25.6 % at different sites, respectively.

INDEX TERMS PM2.5, hybrid deep learning, domain adaption, feature extraction.

I. INTRODUCTION

The rapid development of the economy has brought about the improvement of people's material level, but also caused a series of air pollution. As a developing country, China is facing environmental pollution problems, especially in the special geographical location of Beijing-Hebei region, which has become the 'hardest hit' of air pollution [1], [2]. The main air quality pollutants in these areas are NO_2, SO_2, O_3, CO, PM10 and PM2.5, among which PM2.5 is the main pollutant [3]. The state and community have developed several strategies in recent years to reduce air pollution. For instance, the use of clean energy for winter heating in the north has replaced soot, and the overall quality of the air has progressively improved. However, the problem of particulate pollution remains serious, mainly caused by PM2.5. PM2.5 is

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.

the main reason for the formation of haze, resulting in low atmospheric visibility, which not only affects normal travel, but also endangers people's health. The number of deaths related to PM2.5 exposure in central and eastern China increases gradually from 2000 to 2010, so PM2.5 is an important factor causing deaths [4]. Therefore, if we can accurately predict the changing trend of PM2.5 concentration, it can reduce people's excessive exposure to pollution and has important guiding significance for air pollution control.

Many researchers employ statistical methods to identify variations in PM2.5 concentration. In the study of major formative factors, factor analysis is used to determine the sources of PM2.5 in the capital area of Seoul, and five sources are explained. Among them, coal combustion is the primary cause of PM2.5 production [5]. When examining the relationship between PM2.5 and other components, carbon monoxide (CO) is the pollutant highly correlated with PM2.5 in Hefei, Anhui Province, which is studied by Pearson correlation [6]. In addition, the maximum covariance analysisÂ (MCA) coupled model [7] is applied to analyze the temporal and spatial correlation of PM2.5, and it is found that PM2.5 and sulfur dioxide (SO_2) have a strong correlation in Northern China. The conventional statistical method autoregressive integrated moving average model (ARIMA) is frequently utilized in the prediction of PM2.5. When wind, temperature and other factors are added, the hybrid statistical model using ARIMA and multiple linear regression has better prediction accuracy than a single model [8]. Although the statistical model's accuracy in predicting PM2.5 over a short time frame is good, it is necessary to fully consider the diffusion mechanism of pollution sources, meteorological conditions and other factors, and the modeling are relatively complex.

There are also many researches on the trend prediction of PM2.5 concentration over time-based on machine learning, in which the support vector (SVM) appears more frequently. In [9], [10], and [11], the incremental support vector regression model based on time and space, SVM model based on nonlinear features, and the hybrid model of polynomial regression and support vector machine are proposed to predict PM2.5 concentration. The prediction outcomes are close to the actual value, but preprocessing and extracting characteristics before prediction requires a great deal of time, and the data processing methods are aimed at their own specific models, so the generalization ability of these models is insufficient. In addition, algorithms such as neural networks and random forests are also commonly used to predict PM2.5. In [12], the neural network is improved and the genetic algorithm is added to optimize the network, while in [13], the random forest and integrated neural networks are combined to predict PM2.5. These methods ultimately improve the prediction accuracy of the model, but the machine learning model can only capture the local spatial correlation between adjacent nodes, and there are still limitations for multi-scale and high-latitude air quality spatial data analysis. Researchers are gradually developing hybrid algorithms, such as ANFIS-GBO, SVR-SAMOA, and ELM-CRFOA, that integrate intelligent algorithms with machine learning algorithms to address the weaknesses of machine learning algorithms [14], [15], [16].

The deep learning model has more advantages in dealing with a huge amount of data and more complex nonlinear and non-stationary time series air quality data than the prior statistical model and machine learning model. It can automatically extract air quality features. In particular, the emergence of a recurrent neural network [17] enables the model to selectively memorize effective information from time series for prediction. Abbad et al. [18] suggest that the convolutional and long and short hybrid neural networks model (CNN-LSTM) is superior to the single model for PM2.5 prediction, but that the peak value prediction still has to be improved, which calls for the aid of the attention mechanism. Due to the hybrid deep learning model's increased complexity, there are more parameters, a longer training period, and a greater requirement for data. There are fewer data for the recently created air quality monitoring station, making it challenging to predict PM2.5. The transfer learning model can address the issue of limited data and expedite the parameter setting of the deep learning model. It has been widely used in the field of machine learning vision, but there are still few studies in the time series processing of air quality.

Aiming at the problems of data scarcity and poor performance of feature extraction in transfer learning models, we establish a modifeid hybrid deep learning model under the framework of transfer learning for PM2.5 prediction at multiple sites. The following sections make up the bulk of the paper. The first is the research motivation of model improvement. Then, this paper presents the related definitions and symbols under the transfer learning framework. We describe our research framework and detail out the modified feature extraction method. Next, we perform the comparison results of PM2.5 prediction experiments in Beijing and Hengshui data sets. In this paper, we also measure the validity of the model and further validated by the interpretability analysis of the attention mechanism. This paper presents the research findings, and possible future research directions.

II. RESEARCH MOTIVATION

Deep learning techniques, including recurrent neural networks (RNN) [19] and delay neural networks (TDNN) [20], are particularly popular in the prediction of air pollution concentration because they can extract complex and difficult abstract knowledge and low-dimensional elements from very vast high-dimensional data sets. In view of the strong long-term learning ability of LSTM (long-short-term memory), LSTME (consisting of two layers of LSTM and one fully linked layer) [21] is employed to forecast PM2.5 in Beijing and its MAPE (%) value is 14.94 % less than TDNN. The gate recurrent unit (GRU) model with a simpler structure can avoid the information redundancy caused by too many parameters, and achieves better prediction results than

LSTM in the PM2.5 prediction of three cities [22]. Besides, combined with the integrated learning method, remarkable prediction results are achieved. However, the prediction results of deep learning are easy to change greatly due to site changes or data changes [23]. If the amount of data rises, the GRU model's performance may not always be superior to the LSTM model. Through literature research, just a small number of academics apply a single deep learning model to predict time series, while the majority of researchers choose to utilize a hybrid deep learning model that incorporates the properties of time series data. In [24], combining the advantages of LSTM's time feature extraction and CNN's spatial feature extraction, the prediction of various types of data has achieved better prediction results than that of a single model. The attention mechanism is added to CNN-LSTM to capture the significance of various feature states over time, and the prediction effect is enhanced [25].

Although the hybrid deep learning network has a good prediction effect, it needs to adjust parameters and a large number of training data for network training. The absence of data on the recently built air quality monitoring sites also contribute to the low training effect. The transfer learning method uses the rich data sets of other non-target sites or the trained model for auxiliary training, which overcomes the problems mentioned above and can achieve better prediction results. From the perspective of different migration methods, Ma et al. [26] use a fine-tuning transfer learning method based on LSTM and stacked bidirectional long and short-term memory (Stacked BLSTM), and Fong et al. predicted air pollution in Macau using a pre-training method [27]. According to the different migration targets, many methods are proposed. Ma and Cheng et al., aiming at missing data, propose an iterative estimation based on transferring long short-term memory (TLSTM-IE) [28]. Lv et al. [29] adopt the regression algorithm to transfer air quality data from urban areas to non-urban areas. Dhole et al. collect the knowledge learned from the source domain data of multiple Beijing sites into a target site for prediction [30]. Despite some achievements that have been made in air quality transfer learning prediction, these methods use machine learning methods or simple deep learning modules in feature extraction, which cannot effectively extract feature representations of different dimensions of air quality.

In hopes of efficiently extracting the complex components of air quality data from various locations, this research proposes the attention mechanism module to develop a modified hybrid deep learning network structure depending on domain adaptation. The innovations in model structure are mostly comprised of the following elements:

1) The nonlinear complicated air pollutant features can be extracted using the cyclic neural network with strong memory and multi-layer perceptron with nonlinear feature conversion ability.
2) The residual module is adopted to deepen the depth of the network, and the attention mechanism is added, which ameliorates the unstable prediction effect of the deep learning model.
3) Multi-task learning is instrumental in narrowing the distribution differences in air quality characteristics at different stations, reducing the scale of overall model parameters, and improving the prediction effect of the model.
4) We construct the hybrid deep learning prediction model is driven by transfer learning to adapt to the multi-scale spatio-temporal characteristics. The model's PM2.5 forecast precision is superior to that of the baseline methods, which can provide police decisions for the prevention and control of PM2.5 pollution in a large range of areas.

III. RELEVANT DEFINITIONS AND SYMBOLS

In this section, the relevant definitions of transfer learning are introduced. Besides, the symbols used in this study are listed to provide explanations for the subsequent research on multi-task domain adaptive transfer learning strategies.

1) DEFINITION OF TRANSFER LEARNING

Given the source domain $\mathcal{D}_s$ and learning task $\mathcal{T}_s$, target domain $\mathcal{D}_t$ and learning task $\mathcal{T}_t$, the goal of transfer learning is to reduce the generalization error of target domain prediction model $f_t(x)$ under the condition of $\mathcal{D}_s \neq \mathcal{D}_t$ or $\mathcal{T}_s = \mathcal{T}_t$ by acquiring the knowledge of the source domain and learning task [31].

2) DEFINITION OF DOMAIN ADAPTION

The feature space $\mathcal{X}$ and label space $\mathcal{Y}$ are the same, but the probability distribution is different, that is $P_s(x, y) \neq P_t(x, y)$. This scenario is called domain adaptation.

3) DEFINITION OF MULTI-TASK LEARNING

Zhuang et al. extend the above definition of transfer learning and apply $\left\{\left(\mathcal{D}_s^i, \mathcal{T}_t^i\right) \mid i = 1, 2, \cdots, m^s\right\}$ and $\left\{\left(\mathcal{D}_s^i, \mathcal{T}_t^i\right) \mid i = 1, 2, \cdots, m^t\right\}$ to denote $(\mathcal{D}_t, \mathcal{T}_t)$ and $(\mathcal{D}_s, \mathcal{T}_s)$ respectively, where $m^s, m^t \in \mathbb{N}^+$, so the new definition covers the case of multi-task transfer learning. m^t indicates the number of tasks for transfer learning [32]. When $m^t = 1$, this scenario is single task learning. If $m^t \geq 2$, it is multi-task learning. In this paper, air quality data from three sites are predicted at the same time, so $m^t = 3$. The symbol $\mathcal{X}$ denotes feature space of domain $\mathcal{D}$. $\mathcal{X}_s^1, \mathcal{X}_s^2, \mathcal{X}_s^3$ and $\mathcal{X}_t^1, \mathcal{X}_t^2, \mathcal{X}_t^3$ depict the feature spaces of the source domain and the target domain, respectively. $X = \left\{x \mid x_i^{lj} \in \mathcal{X}, i = 1, 2, \cdots, n; l = 1, 2, 3; j = 1, 2, \cdots, m\right\}$ denotes n sample sets of l tasks that has j features, where x is the feature of data, so the three task features of the source domain and the target domain are $(x_{si}^{lj}(T), x_{ti}^{lj}(T))$ as the inputs of proposed method at time T. The task $\mathcal{T}^l$ label space $\mathcal{Y}_{si}^l, \mathcal{Y}_{ti}^l$ and the prediction model $f_t(x_{si}^{lj}, x_{ti}^{lj})$, thatÂ is, $\mathcal{T}^l = \left\{(\mathcal{Y}_{si}^l, \mathcal{Y}_{ti}^l), f_t(x_{si}^{lj}, x_{ti}^{lj})\right\}$. From the statistical theory of view, the forecasting function $f(x_{si}^{lj}, x_{ti}^{lj})$ is expressed as $f(x_{si}^{lj}, x_{ti}^{lj}) = P(Y \mid X)$, where

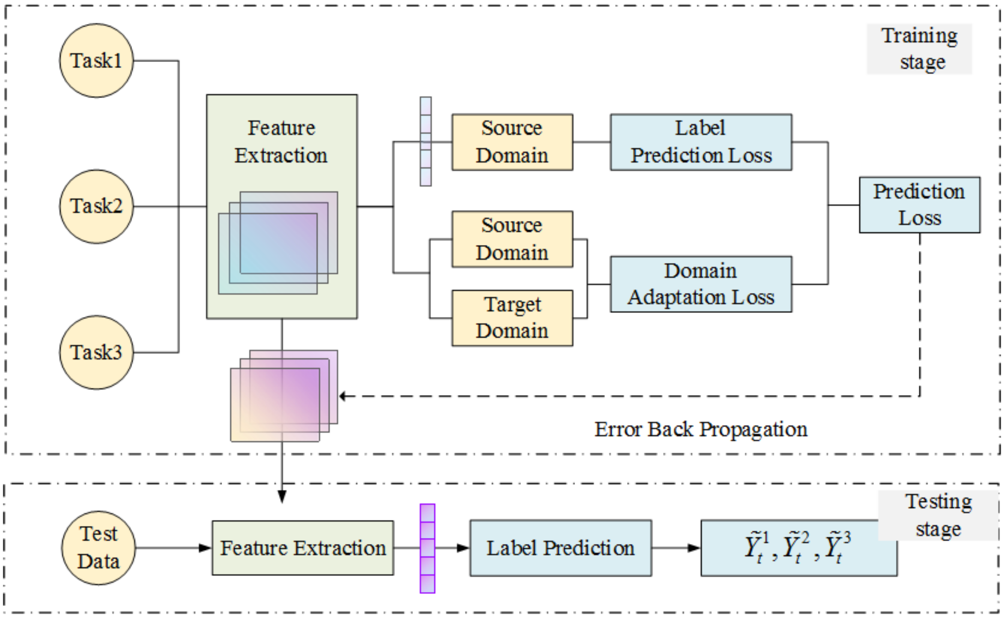


FIGURE 1. **Multi-task domain adaptive PM2.5 prediction framework.**

$Y = \{y_1, y_2, \cdots, y_n\} \in \mathcal{Y}$. At time $T + 1$, the predicted result of PM2.5 is $\widetilde{Y}_t^1, \widetilde{Y}_t^2, \widetilde{Y}_t^3$ as outputs.

IV. RESEARCH FRAMEWORK OF PM2.5 PREDICTION BASED ON TRANSFER LEARNING

High dimension, nonlinearity, multi-channel, and spatiotemporal correlation are properties of the air quality time series data. Therefore, the research on air quality prediction is much more complex, especially when the establishment of a new air quality monitoring site does not have enough data, and the prediction effect of deep learning algorithms in this area is very poor. In addition, the characteristics of air quality data at different sites are similar, and there are also differences between data due to geographical location, climate and different sources of pollution in local areas. Because of the above problems, this paper introduces the migration learning strategy of multi-task domain adaptation to predict the air quality of the three stations synchronously (as shown in Fig. 1). The model can transfer existing knowledge to new sites and effectively address the issue of data scarcity. Moreover, when Multitasking prediction is carried out, the distribution difference of air quality characteristics of different sites is reduced, and the scale of overall model parameters is reduced, making PM2.5 prediction more efficient.

Data from different sites in the city to be predicted are used as the target domain $\mathcal{D}_t$, and monitoring data from multiple sites of urban air quality to be migrated are used as the source domain $\mathcal{D}_s$. A domain adaptation transfer model is established for l tasks to realize the shared parameters of multiple tasks. The training phase and the testing phase comprises the entirety of the model. There are three primary components to the training phase: shared feature extraction, feature domain adaptation, and label prediction. The l tasks are input into the network structure together to extract features. In this study, $l = 3$, the input n sample data at time T is $(x_{si}^{lj}(T), x_{ti}^{lj}(T))$, and the output is $\widetilde{Y}_t^1, \widetilde{Y}_t^2, \widetilde{Y}_t^3$ at time $T + 1$. In the modified feature extraction part, an attention-based long-short-term memory(LSTM) [33] network is proposed to extract both long-term as well as short-term characteristics, and more effective features are extracted from time and space by attention module. The multi-layer perceptron (MLP) [34] is introduced to transform features from data dimension, and the residual block [35] is used to prevent gradient disappearance. Feature domain adaptation maps the features of multiple tasks in two domains to the same subspace by using the maximum mean difference (MMD) [36] method based on statistical criteria to learn the common feature representation and reduce the distribution difference between them. In addition, the regression prediction loss of multiple site labels in the source domain is measured by mean square error (MSE) loss [37], [38]. Two losses are combined as the total forecast loss.

With global climate warming, the air quality time series data changes every day, and noise points are easy to appear in the prediction task. When the error back propagation [39] is used to optimize the parameters, the adaptive moment estimation (Adam) algorithm [40] is adopted to efficiently search the parameter space, which can dynamically adjust the learning rate and is ideal for handling the air quality prediction problem with high noise. At the same time, multiple tasks can interact with each other when training together. Sometimes, adding noise to an additional output of the back propagation network can improve the generalization ability of the model. When the model enters the test phase after training, feature extraction and label prediction

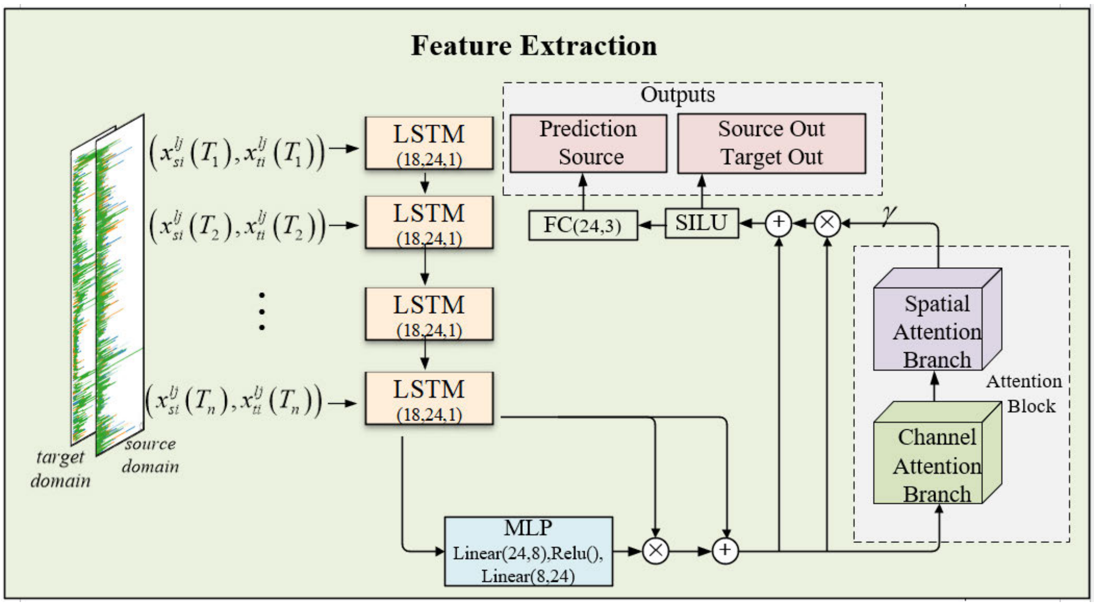


FIGURE 2. The network structure of the modified hybrid deep learning feature extraction method.

complete the regression prediction task of multiple air quality monitoring sites. The following is a detailed introduction to each part of the model.

V. MODIFIED SHARED FEATURE EXTRACTION METHOD

When dealing with data-scarce air quality prediction tasks, the main challenge of transfer learning is that domain adaptation requires deep neural networks to learn the data feature representation of different domains, especially the simultaneous prediction of multiple sites, which puts forward higher requirements for the generalization ability of the network. This part proposes a modified hybrid deep learning feature extraction method, see Fig.2, to extract potential features to complete the target domain's and the source domain's distribution adaption. Firstly, LSTM is used to learn the correlation of time in the sequence, and then MLP utilizes a multi-layer network and nonlinear transformation of features to complete the automatic cross combination of features to form higher-order features, but there is still a problem of insufficient spatial information extraction ability. Next, the spatiotemporal attention mechanism [41] is applied to capture the spacio-temporal characteristics of the time series. The residual modules are incorporated into the entire network architecture to keep the gradient from vanishing. Finally, the extraction of complex, high dimensional and nonlinear air quality features are realized.

A. NEURAL NETWORKS FOR SHORT AND LONG TERM MEMORY

The air quality prediction process is a time-dependent dynamic process, and its state is not only related to the current pollutant concentration value, but also related to the historical time value. In this part, a single LSTM [33] network is used to learn the long-term time dependence in historical sequence data. The n-dimensional data of source domain and target domain are respectively input into the network at time T after processing. The values of input gate $i'(T)$, forgetting gate $f(T)$ output gate $o(T)$ of LSTM is calculated by the fully connected layer with Sigmoid activation function δ, and the number of neurons in the hidden layer $h_{T-1}$is k. $\left[h_{T-1}, x_i^{lj}(T)\right]$ is the concatenation and fusion of the hidden layer information of the previous sequence and the input information of the current sequence. We can get the formulas (1)-(3):

$$f(T) = \delta\left(W_{hf}\left[h_{T-1}, x_i^{lj}(T)\right] + b_f\right), \tag{1}$$

$$i'(T) = \delta\left(W_{hi'}\left[h_{T-1}, x_i^{lj}(T)\right] + b_{i'}\right), \tag{2}$$

$$o(T) = \delta\left(W_{ho}\left[h_{T-1}, x_i^{lj}(T)\right] + b_0\right), \tag{3}$$

where $W_{hf}, W_{hi}, W_{ho} \in \mathbb{R}^{(k+m)\times k}$, $b_f, b_i, b_o \in \mathbb{R}^k$.

The forgetting gate selects the historical information of the previous moment and determines the balance of the retention of the internal cell state C_{T-1} of the long-term information. The input gate, which establishes the amount of candidate cell state $\widetilde{C}_T$ information, is used to calculate C_T. Thus, the memory cell state $c(T)$ is achieved at time T, which is formula (4).

$$c(T) = f(T) \odot C_{T-1} + i'(T) \odot \widetilde{C}_T, \tag{4}$$

where $\widetilde{C}_T = tanh(W_{hc}[h_{t-1}, x_i^{lj}(T) + b_c])$, $W_{hc} \in \mathbb{R}^{(k+m)\times k}$, $b_c \in \mathbb{R}^k$. $\odot$ is Hadamard product [42]. The output gate takes the current moment input information and the updated memory cell state together as the current

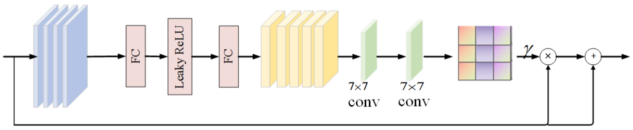


FIGURE 3. The network structure of attention mechanism.

moment output (5)

$$h_T = o(t) \odot tanh(c(T)). \tag{5}$$

B. NEURAL NETWORKS FOR MULTI-LAYER PERCEPTRON

MLP [34] uses the multi-layer network to complete an automatic cross combination of features, to learn the impact of different air pollutant characteristics on the prediction of target PM2.5. The ReLU activation function is $f(\cdot)$ is used to realize the nonlinear mapping of data features and improve the fitting ability of the network. The purpose is to mix, transform, and filter the features of each task to obtain a feature representation that multiple tasks can share. The calculation formula of network layer nodes is the long-term dependent air quality information $h(T)$ obtained from (5), and the features are further extracted by inputting the multi-layer perceptron to the final output of $\alpha(T)$, whose calculation formula is (6):

$$\alpha(T) = W_1 f(W_0 h(T) + b_0) + b_1, \tag{6}$$

where $W_0 \in \mathbb{R}^{(k/r)\times k}, b_0 \in \mathbb{R}^{k/r}, W_1 \in \mathbb{R}^{k\times k/r}, b_1 \in \mathbb{R}^k, r = 3$.

The residual module is used to deepen the depth of the network, while preserving the important features extracted by the MLP layer to get the final output $M(T)$ in(7) at time T.

$$M(T) = h(T) + h(T) \odot \alpha(T). \tag{7}$$

C. NEURAL NETWORKS FOR ATTENTION MECHANISM

Air quality prediction transfer learning is a dynamic time series prediction related to time and space. Combining channel and spatial attention mechanism, it helps the network focus on important feature channels, and further aggregates the spatial dependence of each position in the feature map to form a spatial attention map. It can effectively suppress the noise of different tasks, highlight the target region, and then more precisely extract the target information. The network structure is shown in Fig. 3.

The feature dimension k output from the MLP layer is transformed into $W \times C$ by spatial dimension transformation, and the feature map is obtained as the input feature vector $X_c \in \mathbb{R}^{H\times W\times C}$ of the channel attention mechanism module. In order to decrease the amount of parameters and enhance the migration ability of the module, the module uses two fully connected layers to obtain the weight parameters W_{c0} and W_{c1} of their respective layers, and the Leaky ReLU activation function as the ReLU variant is used. The output of Leaky ReLU [43] can prevent the emergence of more silent neurons when the input value is negative. During the training of the channel attention module, the feature weights of each channel domain and the correlation between channels can be learned. Equation (8) illustrates how different weights can be used to the convolutions channel features to extract important data from the representation target.

$$Z_c(T) = W_{c1} LeakyRelu(W_{c0} X_c(T) + b_{c0}) + b_{c1}, \tag{8}$$

where $W_{c0} \in \mathbb{R}^{1\times C}, W_{c1} \in \mathbb{R}^{C\times 1}, b_{c0} \in \mathbb{R}^1, b_{c1} \in \mathbb{R}^C$, $Z_c(T)$ is a channel attention parameter matrix at time T.

The feature graph output of the channel attention mechanism module is transformed into the dimension as the input feature graph $X_s \in \mathbb{R}^{C\times H\times W}$ of the spatial attention mechanism module. Firstly, the multi-channel features are compressed into a single channel by 7 × 7 two-dimensional convolution $f_s^{7\times 7}$ to eliminate the influence of information distribution between channels of different tasks on the spatial attention mechanism [44]. Then, the spatial weight information is normalized by Leaky ReLU [43] activation function. Next, the features are mapped to multiple channels by 7 × 7 two-dimensional convolution $f_s^{7\times 7}$. Finally, the batch normalization layer is used to restore the feature distribution to be learned by the original network, which reduces the dependence of the subsequent network on the previous network and improves the feedback ability of the network. The operation process is shown in (9).

$$Z_s(T) = BN(f_s^{7\times 7} LeakyReLU(f_s^{7\times 7})), \tag{9}$$

where $Z_s(T)$ is a spatial attention parameter matrix at time T.

In order to produce features with varying weights, the dimension of the spatial weight information is modified and then multiplied by the relevant components of the input characteristics. We get $Z'_s(T) \odot M(T)$. Based on the idea of a residual network, a certain proportion γ is applied to the features and then added to the input features, and the final prediction results of source domain $\widetilde{Y}_s^l$ and target domain $\widetilde{Y}_t^l$ are obtained through SILU activation function [45] as following:

$$\widetilde{Y}_s^l(T), \widetilde{Y}_t^l(T) = SILU(\gamma(Z'_s(T) \odot M(T)) + M(T)). \tag{10}$$

Through the complete connection layer, the source domain regression's predicted value is produced. Equation (11) depicts the operation procedure:

$$\widetilde{Y}_s'^l(T) = FC(\widetilde{Y}_s^l(T)). \tag{11}$$

VI. FEATURE DOMAIN ADAPTATION AND BACK PROPAGATION

In terms of the adaptive distribution between feature domains, the statistical criteria-based methods [46], [47], [48] are commonly used in deep transfer learning algorithms, which are more interpretable. Based on the features extracted above, the MMD [36] method is adopted in this migration framework to measure the difference between different task distributions in the two domains $\mathcal{D}_s, \mathcal{D}_t$. It is assumed that the set of continuous functions on the air quality sample space is $\mathcal{F}$, and $f \in \mathcal{F}$ is the function that maximizes the difference between the mean values of samples with different distributions. Let Y_s, Y_t be the data sets containing m_1 and m_2 elements produced from arbitrary sampling inside the source domain and target domain sample spaces. The distributions are p_s and p_t, respectively. The MMD distance's empirical formula is as follows:

$$\begin{aligned} MMD[\mathcal{F}, p_s, p_t] &= \sup_{\mathrm{f}\in\mathcal{F}} \left(E_{p_s}(f(x_{si}^{lj})) - E_{p_t}(f(x_{ti}^{lj}))\right) \\ &= \sup_{f\in\mathcal{F}} \left(\frac{1}{m_1}\sum_{i=1}^{m_1} f(x_{si}^{lj}) - \frac{1}{m_2}\sum_{i=1}^{m_2} f(x_{ti}^{lj})\right). \end{aligned} \tag{12}$$

Mapping data to regenerative Hilbert Spaces (RKHS) [46] has $f(x_i^{lj}) = \langle f(x_i^{lj}), \varphi(x_i^{lj})\rangle_{\mathcal{H}}$, and $\|f\|_{\mathcal{H}} < 1$. Then, square both sides of equation (12) to get:

$$\begin{aligned} MMD^2[\mathcal{F}, p_s, p_t] &= \left\| \frac{1}{m_1}\sum_{i=1}^{m_1} \varphi(x_{si}^{lj}) - \frac{1}{m_2}\sum_{i=1}^{m_2} \varphi(x_{ti}^{lj}) \right\|_{\mathcal{H}}^2 \\ &= \left\| \frac{1}{m_1^2}\sum_{i=1}^{m_1}\sum_{i'=1}^{m_1} \varphi(x_{si}^{lj})\varphi(x'^{lj}_{si}) \right. \\ &\quad - \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{i'=1}^{m_2} \varphi(x_{si}^{lj})\varphi(x_{ti}^{lj}) \\ &\quad \left. + \frac{1}{m_2^2}\sum_{i=1}^{m_2}\sum_{i'=1}^{m_2} \varphi(x_{ti}^{lj})\varphi(x'^{lj}_{ti}) \right\|_{\mathcal{H}}. \end{aligned} \tag{13}$$

In higher dimensional space, the inner product is replaced by the Gaussian kernel function $k(x_i^{lj}, x'^{lj}_i)$ to get the final formula:

$$\begin{aligned} MMD^2[\mathcal{H}, p_s, p_t] &= \left\| \frac{1}{m_1^2}\sum_{i=1}^{m_1}\sum_{i'=1}^{m_1} k(x_{si}^{lj}, x'^{lj}_{si}) \right. \\ &\quad - \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{i'=1}^{m_2} k(x_{si}^{lj}, x_{ti}^{lj}) \\ &\quad \left. + \frac{1}{m_2^2}\sum_{i=1}^{m_2}\sum_{i'=1}^{m_2} k(x_{ti}^{lj}, x'^{lj}_{ti}) \right\|_{\mathcal{H}} \\ &= \left\| \frac{1}{m_1^2}\sum_{i=1}^{m_1}\sum_{i'=1}^{m_1} e^{-\frac{\|x_{si}^{lj}-x'^{lj}_{si}\|}{2\sigma^2}} \right. \\ &\quad - \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{i'=1}^{m_2} e^{-\frac{\|x_{si}^{lj}-x_{ti}^{lj}\|}{2\sigma^2}} \\ &\quad \left. + \frac{1}{m_2^2}\sum_{i=1}^{m_2}\sum_{i'=1}^{m_2} e^{-\frac{\|x_{ti}^{lj}-x'^{lj}_{ti}\|}{2\sigma^2}} \right\|_{\mathcal{H}}. \end{aligned} \tag{14}$$

When the MMD square distance is calculated by the adaptation layer, it is combined with the regression prediction loss function as the optimization goal L_{total} of the multi-task domain adaptation network to minimize the distribution of the source domain and the target domain. The mean square error (MSE) loss used to determine the mean square error between the source domain's real value and forecast value is the regression loss function in the model. The regression loss function is defined as (15). The difference between the source domain's actual value y_{si}^l and forecasting value $\hat{y}_{si}^l$ is calculated using the mean square error (MSE) loss.

$$L_{s\mathcal{Y}} = \frac{1}{nl}\sum_{l=1}^{3}\sum_{i=1}^{n}(y_{si}^l - \hat{y}_{si}^l)^2. \tag{15}$$

The feature domain adaptation network continuously reduces the distribution difference between the source domain and the target domain data during the training process in order to achieve domain adaptation. This is done by using the total loss function L_{total} as the optimization objective to update the network parameters as demonstrated in (16).

$$L_{total} = L_{s\mathcal{Y}} + \beta MMD^2[\mathcal{H}, p_s, p_t], \tag{16}$$

where β controls the distribution distance how much to participate in the training of the network.

In the back propagation, the Adam optimization [40] technique is used to optimize the network model by selecting distinct learning rates for each separate parameter. This increases prediction accuracy and speeds up the model's convergence speed in the early stage. The Adam algorithm uses the first moment to estimate the adjustment direction of $m(T)$, control weight parameter value or threshold. The second-order moment estimation $\vartheta(T)$ is used to adjust the size of the learning rate, so that the learning rate is adaptive to the gradient change. The first-order moment estimation and the second-order moment estimation are modified by the deviation to $m'(T)$ and $\vartheta'(T)$. The weight parameter value or threshold θ are updated as follows: $\theta(T) = \theta(T-1) - \frac{\lambda m'(T)}{\sqrt{\vartheta'(T)}+\varepsilon}$, where λ is the learning rate; ε is a small constant added to maintain numerical stability.

Algorithm 1 introduces the training process of the improved transfer learning model for multi-site PM2.5 prediction.

VII. AIR QUALITY PREDICTION EXPERIMENTS

This section provides a detailed introduction to the data set, data preprocessing, and model evaluation criteria. The multi-task domain adaptive transfer learning model's parameters are established. The detailed analysis of the experimental findings is followed by a comparison of the suggested model framework with other comparison models. The effectiveness of PM2.5 predictions at various places is assessed, and the predictions outcomes are examined.

A. AIR QUALITY DATASETS INTRODUCTION

The air quality pollution in Beijing is serious [49] and has improved in recent years, but it is still the focus of environmental protection. Hengshui City is located in the southeast of Hebei Province. Although its geographic location is crucial, it is a city with poor economic strength in Hebei Province. Hengshi City also faces the problem of air pollution control. Therefore, selecting Beijing and Hengshui air quality as the research object is conducive to promoting

Algorithm 1 Training Process for PM2.5 Prediction

```
Input: $x_{si}^{lj}$ multiple sites' training data of source domain;
$x_{ti}^{lj}$ multiple sites' training data of target domain;
$Epoch_m$ maximum number of iterations;
$bachsize_m$ the number of data passed to the program for training at a time;
learning rate $\lambda$;
$\gamma$ feature weight;
$\beta$ proportion of distribution distance in training of participating network.
Initialization: initialize transfer learning network parameters: $\theta \rightarrow W, b$.
for $i$ to $Epoch_m$ do
    set total loss $L_{total} = 0$
    for $(x_{si}^{lj}, x_{ti}^{lj})$ in source domain training data set $X_s^{train}$ do
        clear past gradient
        From formulas (1) to (11), the air quality characteristics $\widetilde{Y}_s^l, \widetilde{Y}_t^l$ of source domain and target domain in the adaptation layer and the predicted value $\widetilde{Y}_s^{\prime l}$ of source domain label are obtained.
        The objective loss function of (16) $L_{s\mathcal{Y}} + \beta MMD^2[\mathcal{H}, p_s, p_t]$ is calculated by (14) and (15).
        The gradient of network parameters is calculated according to the total target loss.
        The weight parameters and threshold values are updated and optimized by the Adam optimizer: $\theta(T) \leftarrow \theta(T-1) - \frac{\lambda m'(T)}{\sqrt{\vartheta'(T)}+\varepsilon}$.
    end for
end for
```

the coordinated development of Beijingand Hebei. In this paper, transfer learning is first carried out on Beijing, and then its generalization ability is studied on Hengshui dataset. For this purpose, the experiment in this paper is based on two different real data sets, and the hourly air quality index time series data is selected as the research object. The Beijing dataset [50], [51] consists of 35064 bits of data, representing hourly pollution gas concentration data from 12 sites in Beijing from March 1, 2013, to February 28, 2017. Hengshui dataset comes from the real-time air quality publishing platform of China Environmental Protection Administration. It records the hourly air quality data of three monitoring stations of Hengshui City Shichengguanju, Shijiancezhan and Dianjibeizhan from January 1, 2020 to April 30, 2022, with a total of 20,424 pieces of data. The two data sets mainly include the concentration values of oxides such as NO_2, SO_2, O_3 and CO and suspended particulate matter such as PM2.5 and PM10. Among them, the main pollutant of air quality is PM2.5, which is also the main object considered by other researchers [52], [53], [54] when predicting air quality. Therefore, we selected the main six pollutant indicators to predict PM2.5 and record the experimental results.

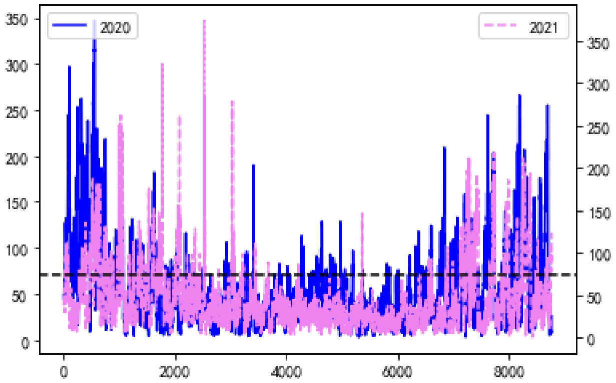


FIGURE 4. The distribution trend of PM2.5 in Hengshui in 2020 and 2021.

The distribution trend of PM2.5 in Hengshui in 2020 and 2021 is shown in Fig. 4. With a standard 24-hour mean concentration of pm2.5 below 75 micrograms per cubic meter ($\mu g/m^3$) [50], the dotted line is plotted as the baseline level. Compared to baseline levels, PM2.5 concentration is higher in winter and spring and lower in summer. Areas with moderate and high concentrations of PM2.5 increase significantly during the winter period. This phenomenon is associated with seasonal activities (e.g. crop burning, heating) resulting in higher emissions of pollutants. In addition, due to the influence of meteorological factors, pollutants accumulate in the lower atmosphere, which is difficult to disperse and dilute, thus aggravating the accumulation of PM2.5 [50].

B. DATA PREPROCESSING

1) DATA CLEANING

In the process of data collection, there are abnormal phenomena such as missing duplicate data, which requires data preprocessing of Beijing dataset and Hengshui dataset. The time-based information necessary for the experiment as well as the properties of six air contaminants are retained after merging the data files kept at several locations. Next, the data is cleaned. the repeated site names are replaced, and the repeated sample data is deleted. Two columns of data, the date and time, are first transformed to string format for merging, and then to timestamp format. Since the first column of the Beijing dataset is missing, the observed values behind the missing values are filled in using the backward-filling method. The Hengshui dataset employs the forward-filling approach, in which the observation value is filled and the time is finished before the missing value. The value exceeds the set interval as a result of equipment malfunction, human error, and other circumstances. It is required to process these data values because they are invalid and will significantly affect the test results. The Hengshui dataset's ozone data value surpasses the range. The treatment method for this situation is that firstly, the invalid values are replaced with data from neighboring monitoring stations due to the strong correlation between the air quality characteristics of the stations in Hengshui City [55]. After that, the mean is used to replace the unprocessed invalid data.

2) STANDARDIZATION AND SEGMENTATION OF DATA

There are dimensional differences between air quality indicators. In addition, the data sets of Beijing and Hengshui have different change intervals as a result of the different times and locations at which they were collected. This difference will affect the calculation of loss values in deep learning training. At the same time, the gradient transmitted to the input layer will become large during back propagation if the input value is large, which is not conducive to finding the optimal value. Hence, normalization is crucial to enhancing the transfer learning model's accuracy and rate of convergence. As the data is converted to floating-point data in the experiment, the minimum-maximum normalization method [56] is used to map the values of raw sample data between 0 and 1. The original air quality data is partitioned, with 30% being used as a test set to assess the model and 70% being used as a training set for the multi-task domain adaptive migration model.

3) SLINDING WINDOW PROCESSING

The time correlation between the time series of air quality data is high. In order to fit the data to the time series model based on LSTM, time-based sliding windows [57] are used to process the data. Suppose the sample sequence length is n. In the subsequence segmentation of time series, a sliding window with a length of l is used to segment the time series with equal length, and then a step length r is moved backward to continuously slide $(n-l)/r$ times to form $(n-l)/r+1$ subsequence fragments of equal length. The hourly data of air quality in the past every 12 hours are integrated and encapsulated as window data as an input of the model. The hourly data of PM2.5 concentration at the subsequent time is set as the model's output once the step size r is set to 1. After 1 hour, the concentration of PM2.5 is predicted using the air quality information from the previous 12 hours.

C. EXPERIMENT SETTING

1) EXPERIMENTAL ENVIRONMENT PARAMETER SETTINGS

All experiments in this study are carried out on a PC server with a CPU processor based on the python language under the pytorch 1.10 framework. The training set sequence is altered in the experiment to improve the model's capacity for generalization. In Algorithm 1, the parameters involved in the input are set to $\gamma = \beta = 0.1, Epoch_m = 100$. When extracting data in batches, set the batch size to 1000, and the final portion of the data that does not fit into the allotted number of small batches is discarded in order to avoid training and prevent overfitting. We set the learning rate λ to 0.0001. The value of bachsize is 1000, and discard method is adopted in the training process.

2) EVALUATION INDEX

In this study, three evaluation indices [50] that are often used in regression tasks-root mean square error (RMSE), mean absolute error (MAE), and mean square error (MSE)-are utilized to compare the performance of several methods in the PM2.5 concentration time series prediction job. The discrepancy between the real value Y_t^l and the predicted value $\widetilde{Y}_t^l$ of the three monitoring sites is gauged using these three indicators. The value of each index should be as low as possible, indicating that the prediction error of PM2.5 concentration in the future is small and the model accuracy is high. The mathematical process is given in (17)-(19).

$$MSE = \frac{1}{n}\sum(Y_t^l - \widetilde{Y}_t^l)^2; \tag{17}$$

$$RMSE = \sqrt{\frac{1}{n}\sum(Y_t^l - \widetilde{Y}_t^l)^2}; \tag{18}$$

$$MAE = \frac{1}{n}\sum\left|Y_t^l - \widetilde{Y}_t^l\right|. \tag{19}$$

D. BASELINE METHODS

Some classical deep learning time series prediction models are combined with the current advanced deep learning model in the field of predicting air quality, and sufficient experiments are conducted based on the aforementioned two actual data sets to illustrate the effectiveness of the modified model. The fundamental models listed above are:

- *LSTM [33]:* It has a gated structure that can effectively overcome the problem in RNN that gradients may disappear quickly in the process of back propagation, so as to obtain long-term dependence and process long time series data.
- *GRU [58]:* The three gate functions of the LSTM network are optimized by the GRU network, a variation of the LSTM network. The structure of GRU is simpler and the training parameters are reduced, thus shortening the training time of the model.
- *CNN-LSTM [59]:* With the CNN-LSTM hybrid neural network design, the advantages of both CNN and LSTM are merged. The high-dimensional mapping space's temporal feature vector is first constructed using CNN, and then LSTM is used to learn lengthy time series and train features on historical data.
- *AdaRNN [51]:* This algorithm defines the time covariate shift problem, and uses the time distribution characterization and time distribution matching algorithm to construct the model. The one step air quality prediction displays the state-of-art experimental results.

E. PM2.5 PREDICTION RESULT ANALYSIS

In this experiment, the pollutants such as NO_2, SO_2, O_3, CO, PM2.5, and PM10 are chosen as the input. For all models, the hidden layer size is set to 24, and the input size is 18. In the CNN-LSTM model, the kernel size of 1D convolution is set to 3, and paddling is set to 1. The 2D convolution's kernel size in the modified hybrid deep learning model is set to 7, and padding is set to 3. Firstly, the multi-task transfer learning between stations is carried out on the concentration of PM2.5 pollutants in Beijing. Then, the superiority of the proposed model is further studied. The

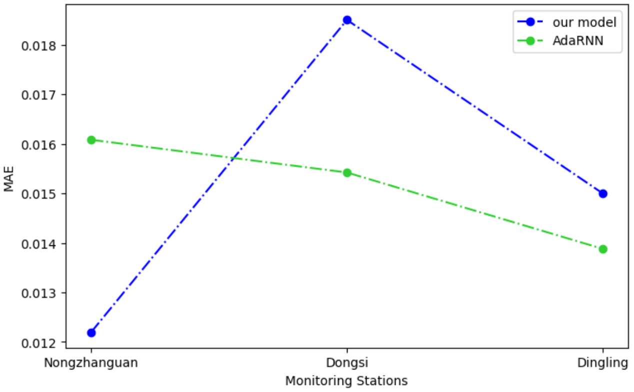


FIGURE 5. MAE of AdaRNN [51] and our model.

Beijing monitoring station is used as the source domain to study the spatio-temporal transfer learning of air quality in Hengshui City. The following is a detailed introduction.

1) RESULT ANALYSIS OF BEIJING PM2.5 PREDICTION

The Beijing data set is used in this experiment. By applying Changping, Aotizhongxin and Guanyuan as source domain data, the PM2.5 of Nongzhanguan, Dongsi, Dingling in the target domain is predicted. Our modified model is based on domain-adapted migration learning between sites, and predicts the PM2.5 concentration of multiple sites at the same time. The AdaRNN algorithm is based on domain generalization to study the migration of historical data of a site and predict the PM2.5 concentration of a site. The feature extraction is improved in our model due to the complicated characteristics of air quality data, including high-dimensionality and non-linearity. Not only the LSTM and MLP are used to extract the features, but also channel attention and spatial attention are added to realize the time and spatial feature extraction of time series. Although the MAE value of our improved model at Nongzhuang site is 0.01608, which is higher than 0.0122 based on AdaRNN algorithm, the MAE values at Dongsi and Dingling sites are 0.01542 and 0.01388, which are lower than 0.0185 and 0.015 based on AdaRNN algorithm [51], as shown in Fig.5. Without GPU acceleration, our air quality prediction method is more applicable.

2) RESULT ANALYSIS OF HENGSHUI PM2.5 PREDICTION

In order to test the effectiveness and applicability of the modifeid model, the small data set of Hengshui City from January to April 2022 is chosen to forecast PM2.5, and six pollution indicators are also picked as input. Aotizhongxin, Dingling and Changping, are randomly selected from the environmental monitoring sites of the Beijing dataset as the source domain sites to predict the three sites Shichengguanju, Shijiancezhan and Dianjibeizhan in the Hengshui dataset as the target domain. In Table 1, the experimental findings are presented.

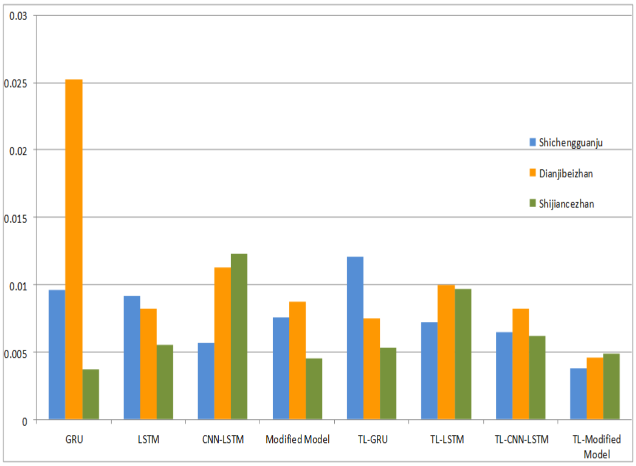


FIGURE 6. MSE of PM2.5 prediction by different deep learning methods.

From Table 1, it can be found that the changing trend of the MAE index of the eight models at each site is consistent with the MSE and RMSE indicators. The MSE values of the PM2.5 predictions of the three stations are presented in a compound bar chart to help make the result more understandable (see Fig.6). Without domain adaptation, it can be seen from Fig.6 that the modified model's MSE at two sites is smaller than those of GRU, LSTM, and CNN-LSTM, which proves that the modified model can extract air quality feature information more effectively. The GRU model performs the poorest in Dianjibeizhan and the best in Shijizhan in terms of learning outcomes. It has poor adaptability in multi-task learning and cannot learn the shared characteristics of each task well.

Under the common framework of transfer domain adaptation, TL-GRU, TL-CNN-LSTM and TL-modified model learning performance have been improved, especially TL-CNN-LSTM and TL-modified model, except that TL-LSTM learning effect has decreased slightly. In Dianjibeizhan and Shijiancezhan, the RMSE of TL-CNN-LSTM is decreased by 14.9% and 29.2%, respectively, in comparison to CNN-LSTM. The MSE of TL-modified model is reduced by 29.2% in Shichengguanju and 27.3% in Dianjibeizhan compared with the modified model. On the whole, the prediction effect of TL-modified model at three sites is not much different. Among them, Shichengguanju and Dianjibeizhan have significant better prediction effects than other models. Its RMSE values at two sites are 18.3% lower than CNN-LSTM and 21.8% lower than TL-GRU. These are enough to prove that the TL-modified model put forth in this paper can effectively decrease the distribution error between the source domain and the target domain data, and learn the shared features of multi-task well, which is more suitable for the simultaneous prediction of PM2.5 at multiple sites.

The results of PM2.5 prediction adopting all the data of Hengshui dataset from January 2020 to April 2022 are shown in Table 2. With the increase of data, except

TABLE 1. Comparison of the prediction results of several deep learning techniques for PM2.5.

Model	Shichengguanju			Dianjibeizhan			Shijiancezhan		
	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
GRU	0.0096	0.0982	0.0755	0.0252	0.1589	0.1452	0.0037	0.0608	0.0503
LSTM	0.0092	0.0960	0.0822	0.0082	0.0904	0.0768	0.0055	0.0744	0.0566
CNN-LSTM	0.0057	0.0753	0.0562	0.0113	0.1062	0.0906	0.0123	0.1109	0.0956
Modified Model	0.0076	0.0869	0.0722	0.0087	0.0931	0.0786	0.0045	0.0673	0.0543
TL-GRU	0.0121	0.1102	0.0957	0.0075	0.0864	0.0665	0.0053	0.0727	0.0611
TL-LSTM	0.0072	0.0850	0.0684	0.0100	0.1000	0.0845	0.0097	0.0983	0.0795
TL-CNN-LSTM	0.0065	0.0807	0.0688	0.0082	0.0904	0.0738	0.0062	0.0785	0.0634
TL-Modified Model	0.0038	0.0615	0.0443	0.0046	0.0676	0.0524	0.0049	0.0697	0.0534

TABLE 2. Comparison of PM2.5 prediction results after adding data.

Model	Shichengguanju			Dianjibeizhan			Shijiancezhan		
	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE
TL-GRU	0.0099	0.0994	0.0866	0.0135	0.1163	0.0985	0.0078	0.0883	0.0704
TL-LSTM	0.0072	0.0851	0.0699	0.0153	0.1236	0.1000	0.0046	0.0678	0.0493
TL-CNN-LSTM	0.0082	0.0907	0.0822	0.0068	0.0824	0.0711	0.0038	0.0617	0.0495
TL-Modified Model	**0.0028**	**0.0528**	**0.0428**	**0.0047**	**0.0688**	**0.0575**	**0.0021**	**0.0459**	**0.0341**

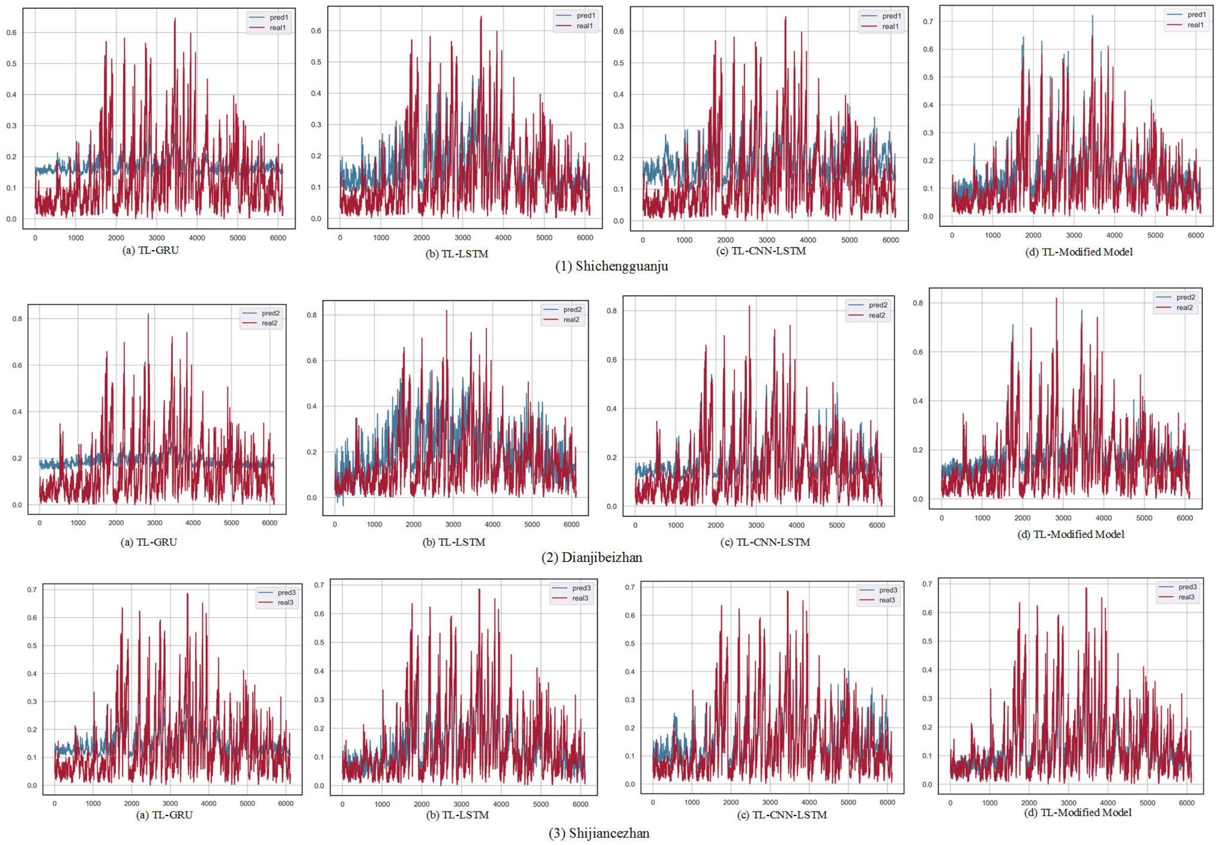
TL-GRU, the prediction ability of the other three transfer learning models has been improved. The TL-Modified Model is the best of the three evaluation indexes, and TL-GRU has the worst prediction effect. Compared with TL-LSTM, the RMSE value of the TL-Modified Model in Shichengguanju decreased by 38%, and that in Dianjibeizhan and Shijiancezhan decreased by 16.5% and 25.6% compared with TL-CNN-LSTM. At three sites, the forecast results of TL-GRU, TL-GRU and TL-CNN-LSTM are quite different, while the difference of the TL-Modified Model is small. It can be seen that TL-Modified Model has strong generalization ability, can be applied to different data sets, and has strong robustness.

For three air pollution monitoring sites, the comparison curve analysis of four models (TL-GRU, TL-LSTM, TL-CNN-LSTM, and TL-Modified Model) between the predicted value and the actual value is shown in Fig.7. As can be seen from the figure, the improved model is better than the benchmark method in predicting the position of trough and peak, and the two lines are almost in agreement. Comparative investigation demonstrates that the efficiency of the benchmark hybrid deep learning model TL-CNN-LSTM is superior to the performance of the TL-GRU and TL-LSTM single deep learning models. The model TL-GRU is unable to effectively predict the development trend of PM2.5 pollution at the three sites, especially when pollution is severe. TL-LSTM only has a good prediction effect in Shijiancezhan, so its generalization ability is poor. Although TL-CNN-LSTM has superior prediction accuracy, the prediction performance of TL-Modified Model is further improved compared with TL-CNN-LSTM model, which can effectively predict the peak and trough trend of PM2.5 concentration, and the overall prediction performance of TL-Modified Model is the best in three sites.

In summary, the TL-Modified Model can maintain the best prediction effect under various data sets (including historical data sets, latest data, small data, big data, etc.), real data can be closely matched with the projected value of PM2.5. This demonstrates that the enhanced multi-task deep learning PM2.5 prediction model put forth in this paper is capable of efficiently learning a variety of features in the air quality time series data, including time trend of the sequence, long-term dependence features and spatial features, and shared features between multi-tasks. The proposed PM2.5 prediction model based on an improved transfer learning framework provides a valuable model reference for the analysis and prediction of PM2.5 pollution in the application of air pollution prevention and control.

F. MODEL INTERPRETABILITY OF ATTENTION MECHANISM

Fig.8 shows the transformation trend of air quality characteristics caused by channel and spatial attention mechanisms. The horizontal axis represents the characteristics of air



(a) TL-GRU (b) TL-LSTM (c) TL-CNN-LSTM (d) TL-Modified Model

(1) Shichengguanju

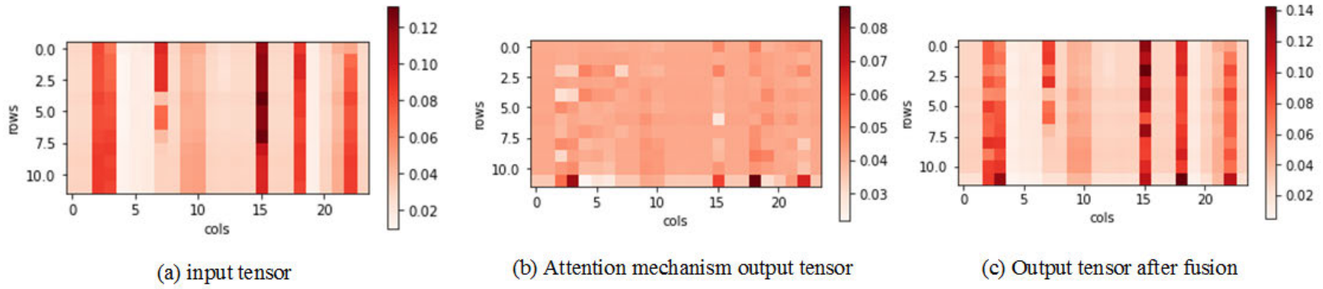
(a) TL-GRU (b) TL-LSTM (c) TL-CNN-LSTM (d) TL-Modified Model

(2) Dianjibeizhan

(a) TL-GRU (b) TL-LSTM (c) TL-CNN-LSTM (d) TL-Modified Model

(3) Shijiancezhan

FIGURE 7. Graphical comparison of PM2.5 predicted values (dark blue) and actual observed values (dark red) of four models (TL-GRU, TL-LSTM, TL-CNN-LSTM, and TL-Modified Model) in Shichengguanju,Dianjibeizhan, and Shijiancezhan.



(a) input tensor

(b) Attention mechanism output tensor

(c) Output tensor after fusion

FIGURE 8. Variation trend of feature output tensor with attention mechanism in model training.

quality after transformation, and the vertical axis represents 12 time series data of a window. Fig.8(a) is the feature extracted in the previous stage. Fig.8(b) represents the feature tensor extracted on the basis of Fig.8(a), and Fig.8(c) is the result of feature fusion with Fig.8(b) and Fig.8(a). From the graphs, it is clear that the attention mechanism can extract the importance of different channels and spaces, learn the feature representation between different tasks, and improve the robustness of network extraction features.

VIII. CONCLUSION

The domain-adapted modified hybrid deep learning model described in this paper can forecast PM2.5 concentrations simultaneously at a number of sites. As the time series data has multi-scale spatial characteristics and time-period dynamic offset characteristics, the LSTM is used to learn the time-period dynamic offset characteristics of air quality data in the feature extraction time column prediction model, and the MLP is used to transform the characteristics. In addition, to capture various channels and spatial variables, the attention mechanism is included, and the residual network is employed to increase the model's depth. The performance of the suggested model is contrasted with different deep learning techniques using actual data from Beijing and Hengshui. The model's excellent measurement accuracy and ability to better extract features under various types of data sets are demonstrated by the results of PM2.5 prediction, which make it acceptable for predicting air quality in regions with sparse data. Given that various variables, including weather and temperature, have an impact on air quality [60], numerous variables will be incorporated in the ensuing research to enhance the experiment's ability to predict outcomes. Reducing levels of PM2.5 and other air pollutants requires a range of individual and collective efforts to promote clean energy and encourage the use of public transport. Improving air pollution prediction model can provide accurate predictions and insights into pollution patterns for these actions, helping to inform policy making and advise people on how to travel.

REFERENCES

[1] Y. Miao, X.-M. Hu, S. Liu, T. Qian, M. Xue, Y. Zheng, and S. Wang, ''Seasonal variation of local atmospheric circulations and boundary layer structure in the Beijing–Tianjin–Hebei region and implications for air quality,'' *J. Adv. Model. Earth Syst.*, vol. 7, no. 4, pp. 1602–1626, Dec. 2015.

[2] L. Wu, N. Li, and Y. Yang, ''Prediction of air quality indicators for the Beijing–Tianjin–Hebei region,'' *J. Cleaner Prod.*, vol. 196, pp. 682–687, Sep. 2018.

[3] L. Wang, F. Zhang, E. Pilot, J. Yu, C. Nie, J. Holdaway, L. Yang, Y. Li, W. Wang, S. Vardoulakis, and T. Krafft, ''Taking action on air pollution control in the Beijing–Tianjin–Hebei (BTH) region: Progress, challenges and opportunities,'' *Int. J. Environ. Res. Public Health*, vol. 15, no. 2, p. 306, Feb. 2018.

[4] J. Wu, J. Zhu, W. Li, D. Xu, and J. Liu, ''Estimation of the $PM_{2.5}$ health effects in China during 2000–2011,'' *Environ. Sci. Pollut. Res.*, vol. 24, no. 11, pp. 10695–10707, Apr. 2017.

[5] S. R. Won, I.-K. Shim, J. Kim, H. A. Ji, Y. Lee, J. Lee, and Y. S. Ghim, ''$PM_{2.5}$ and trace elements in underground shopping districts in the Seoul metropolitan area, Korea,'' *Int. J. Environ. Res. Public Health*, vol. 18, no. 1, p. 297, Jan. 2021.

[6] K. Mi, R. Zhuang, Z. Zhang, J. Gao, and Q. Pei, ''Spatiotemporal characteristics of $PM_{2.5}$ and its associated gas pollutants, a case in China,'' *Sustain. Cities Soc.*, vol. 45, pp. 287–295, Feb. 2019.

[7] K. Li and K. Bai, ''Spatiotemporal associations between $PM_{2.5}$ and SO_2 as well as NO_2 in China from 2015 to 2018,'' *Int. J. Environ. Res. Public Health*, vol. 16, no. 13, p. 2352, Jul. 2019.

[8] X. Yan and X. Enhua, ''ARIMA and multiple regression additive models for $PM_{2.5}$ based on linear interpolation,'' in *Proc. Int. Conf. Big Data Artif. Intell. Softw. Eng. (ICBASE)*, Oct./Nov. 2020, pp. 266–269.

[9] L. Song, S. Pang, I. Longley, G. Olivares, and A. Sarrafzadeh, ''Spatio-temporal $PM_{2.5}$ prediction by spatial data aided incremental support vector regression,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 623–630.

[10] Y. Dong, H. Wang, L. Zhang, and K. Zhang, ''An improved model for $PM_{2.5}$ inference based on support vector machine,'' in *Proc. 17th IEEE/ACIS Int. Conf. Softw. Eng., Artif. Intell., Netw. Parallel/Distrib. Comput. (SNPD)*, May 2016, pp. 27–31.

[11] S. Yang, ''Seasonal prediction of $PM_{2.5}$ based on support vector machine model and multiple regression model,'' in *Proc. Int. Conf. Algorithms, High Perform. Comput., Artif. Intell. (AHPCAI)*, Dec. 2021, pp. 304–316.

[12] X. Wang and B. Wang, ''Research on prediction of environmental aerosol and $PM_{2.5}$ based on artificial neural network,'' *Neural Comput. Appl.*, vol. 31, no. 12, pp. 8217–8227, Dec. 2019.

[13] Z. Shang and J. He, ''Predicting hourly $PM_{2.5}$ concentrations based on random forest and ensemble neural network,'' in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 2341–2345.

[14] R. M. Adnan, R. R. Mostafa, A. Elbeltagi, Z. M. Yaseen, S. Shahid, and O. Kisi, ''Development of new machine learning model for streamflow prediction: Case studies in Pakistan,'' *Stochastic Environ. Res. Risk Assessment*, vol. 36, no. 4, pp. 999–1033, Apr. 2022.

[15] R. M. Adnan, O. Kisi, R. R. Mostafa, A. N. Ahmed, and A. El-Shafie, ''The potential of a novel support vector machine trained with modified mayfly optimization algorithm for streamflow prediction,'' *Hydrol. Sci. J.*, vol. 67, no. 2, pp. 161–174, Jan. 2022.

[16] R. M. A. Ikram, H.-L. Dai, M. M. Chargari, M. Al-Bahrani, and M. Mamlooki, ''Prediction of the FRP reinforced concrete beam shear capacity by using ELM-CRFOA,'' *Measurement*, vol. 205, Dec. 2022, Art. no. 112230.

[17] W. Zaremba, I. Sutskever, and O. Vinyals, ''Recurrent neural network regularization,'' 2014, *arXiv:1409.2329*.

[18] L. Abbad, D. Brahmia, and M. N. Cherfia, ''Study and comparison of machine learning models for air $PM_{2.5}$ concentration prediction,'' in *Proc. 5th Int. Symp. Informat. Appl. (ISIA)*, Nov. 2022, pp. 1–6.

[19] F. Biancofiore, M. Busilacchio, M. Verdecchia, B. Tomassetti, E. Aruffo, S. Bianco, S. Di Tommaso, C. Colangeli, G. Rosatelli, and P. Di Carlo, ''Recursive neural network model for analysis and forecast of PM_{10} and $PM_{2.5}$,'' *Atmos. Pollut. Res.*, vol. 8, no. 4, pp. 652–659, Jul. 2017.

[20] B. T. Ong, K. Sugiura, and K. Zettsu, ''Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting $PM_{2.5}$,'' *Neural Comput. Appl.*, vol. 27, no. 6, pp. 1553–1566, Aug. 2016.

[21] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi, ''Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation,'' *Environ. Pollut.*, vol. 231, pp. 997–1004, Dec. 2017.

[22] Y. Li, Z. Liu, and H. Liu, ''A novel ensemble reinforcement learning gated unit model for daily $PM_{2.5}$ forecasting,'' *Air Qual., Atmos. Health*, vol. 14, no. 3, pp. 443–453, Mar. 2021.

[23] R. K. Choudhary and S. K. Singh, ''A deep learning approach to estimate air pollutants concentration levels in Delhi's aerosphere,'' in *Proc. IEEE Global Conf. Comput., Power Commun. Technol. (GlobConPT)*, Sep. 2022, pp. 1–8.

[24] H. Dai, G. Huang, J. Wang, H. Zeng, and F. Zhou, ''Prediction of air pollutant concentration based on one-dimensional multi-scale CNN-LSTM considering spatial–temporal characteristics: A case study of Xi'an, China,'' *Atmosphere*, vol. 12, no. 12, p. 1626, Dec. 2021.

[25] S. Li, G. Xie, J. Ren, L. Guo, Y. Yang, and X. Xu, ''Urban $PM_{2.5}$ concentration prediction via attention-based CNN–LSTM,'' *Appl. Sci.*, vol. 10, no. 6, p. 1953, Mar. 2020.

[26] J. Ma, Z. Li, J. C. P. Cheng, Y. Ding, C. Lin, and Z. Xu, ''Air quality prediction at new stations using spatially transferred bi-directional long short-term memory network,'' *Sci. Total Environ.*, vol. 705, Feb. 2020, Art. no. 135771.

[27] I. H. Fong, T. Li, S. Fong, R. K. Wong, and A. J. Tallón-Ballesteros, ''Predicting concentration levels of air pollutants by transfer learning and recurrent neural network,'' *Knowl.-Based Syst.*, vol. 192, Mar. 2020, Art. no. 105622.

[28] J. Ma, J. C. P. Cheng, Y. Ding, C. Lin, F. Jiang, M. Wang, and C. Zhai, ''Transfer learning for long-interval consecutive missing values imputation without external features in air pollution time series,'' *Adv. Eng. Informat.*, vol. 44, Apr. 2020, Art. no. 101092.

[29] M. Lv, Y. Li, L. Chen, and T. Chen, ''Air quality estimation by exploiting terrain features and multi-view transfer semi-supervised regression,'' *Inf. Sci.*, vol. 483, pp. 82–95, May 2019.

[30] A. Dhole, I. Ambekar, G. Gunjan, and S. Sonawani, ''An ensemble approach to multi-source transfer learning for air quality prediction,'' in *Proc. Int. Conf. Comput., Commun., Intell. Syst. (ICCCIS)*, Feb. 2021, pp. 70–77.

[31] S. Pan and Q. Yang, ''A survey on transfer learning,'' *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Nov. 2010.

[32] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, ''A comprehensive survey on transfer learning,'' *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[33] R. C. Staudemeyer and E. R. Morris, ''Understanding LSTM—A tutorial into long short-term memory recurrent neural networks,'' 2019, *arXiv:1909.09586*.

[34] N. Tathawadekar, N. A. K. Doan, C. F. Silva, and N. Thuerey, ''Modeling of the nonlinear flame response of a Bunsen-type flame via multi-layer perceptron,'' *Proc. Combustion Inst.*, vol. 38, no. 4, pp. 6261–6269, 2021.

[35] M. Shafiq and Z. Gu, ''Deep residual learning for image recognition: A survey,'' *Appl. Sci.*, vol. 12, no. 18, p. 8972, Sep. 2022.

[36] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, ''Domain adaptation via transfer component analysis,'' *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[37] C. M. Hurvich, R. Deo, and J. Brodsky, ''The mean squared error of Geweke and Porter–Hudak's estimator of the memory parameter of a long-memory time series,'' *J. Time Ser. Anal.*, vol. 19, no. 1, pp. 19–46, Jan. 1998.

[38] L. S. de Oliveira, S. B. Gruetzmacher, and J. P. Teixeira, ''COVID-19 time series prediction,'' *Proc. Comput. Sci.*, vol. 181, pp. 973–980, Jan. 2021.

[39] J. Li, J.-H. Cheng, J.-Y. Shi, and F. Huang, "Brief introduction of back propagation (BP) neural network algorithm and its improvement," in *Advances in Computer Science and Information Engineering* (Advances in Intelligent and Soft Computing). Berlin, Germany: Springer, 2012, pp. 553–558.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[41] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck Attention Module," 2018, *arXiv:1807.06514*.

[42] R. Ye and Q. Dai, "Implementing transfer learning across different datasets for time series forecasting," *Pattern Recognit.*, vol. 109, Jan. 2021, Art. no. 107617.

[43] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, Jun. 2013, vol. 30, no. 1, pp. 1–6.

[44] *Assessment of LSTM, Conv2D and ConvLSTM2D Prediction Models for Long-Term Wind Speed and Direction Regression Analysis*. Accessed: 2021. [Online]. Available: https://www.researchsquare.com/article/rs-1011778/latest

[45] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, pp. 3–11, Nov. 2018.

[46] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[47] S. Zhu, B. Chen, and Z. Chen, "Exponentially consistent kernel two-sample tests," 2018, *arXiv:1802.08407*.

[48] A. Schrab, I. Kim, B. Guedj, and A. Gretton, "Efficient aggregated kernel tests using incomplete U-statistics," in *Proc. 36th Adv. Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. Red Hook, NY, USA: Curran Associates, 2022, pp. 18793–18807. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/774164b966cc277c82a960934445140d-Paper-Conference.pdf

[49] D. Z. Antanasijević, V. V. Pocajt, D. S. Povrenović, M. D. Ristić, and A. A. Perić-Grujić, "PM_{10} emission forecasting using artificial neural networks and genetic algorithm input variable optimization," *Sci. Total Environ.*, vol. 443, pp. 511–519, Jan. 2013.

[50] S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen, "Cautionary tales on air-quality improvement in Beijing," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 473, no. 2205, Sep. 2017, Art. no. 20170457.

[51] Y. Du, J. Wang, W. Feng, S. Pan, T. Qin, R. Xu, and C. Wang, "AdaRNN: Adaptive learning and forecasting of time series," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 402–411.

[52] W. Qiao, W. Tian, Y. Tian, Q. Yang, Y. Wang, and J. Zhang, "The forecasting of $PM_{2.5}$ using a hybrid model based on wavelet transform and an improved deep learning algorithm," *IEEE Access*, vol. 7, pp. 142814–142825, 2019.

[53] W. Ban and L. Shen, "$PM_{2.5}$ prediction based on the CEEMDAN algorithm and a machine learning hybrid model," *Sustainability*, vol. 14, no. 23, p. 16128, Dec. 2022.

[54] H.-C. Chen, K. T. Putra, and J. Chun-WeiLin, "A novel prediction approach for exploring $PM_{2.5}$ spatiotemporal propagation based on convolutional recursive neural networks," 2021, *arXiv:2101.06213*.

[55] *ST-MVL: Filling Missing Values in Geo-Sensory Time Series Data*. Accessed: 2016. [Online]. Available: https://www.microsoft.com/en-us/research/publication/st-mvl-filling-missing-values-in-geo-sensory-time-series-data/

[56] J. Ni, Y. Chen, Y. Gu, X. Fang, and P. Shi, "An improved hybrid transfer learning-based deep learning model for $PM_{2.5}$ concentration prediction," *Appl. Sci.*, vol. 12, no. 7, p. 3597, Apr. 2022.

[57] J. Ma, J. C. Cheng, C. Lin, Y. Tan, and J. Zhang, "Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques," *Atmos. Environ.*, vol. 214, Oct. 2019, Art. no. 116885.

[58] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[59] R. Yan, J. Liao, J. Yang, W. Sun, M. Nong, and F. Li, "Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114513.

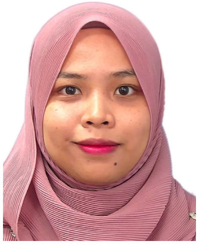
[60] Y. Liang, "AirFormer: Predicting nationwide air quality in China with transformers," in *Proc. 37th AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 12, pp. 14329–14337.

JUNZI YANG was born in 1986. She received the bachelor's and master's degrees in applied mathematics. She is currently pursuing the Ph.D. degree in computer science with Universiti Teknologi Malaysia (UTM), Johore, Malaysia. She is also with the School of Mathematics and Computer Science, Hengshui University, China, teaching statistics and data mining. Her research interests include time series data processing and statistical modeling.

AJUNE WANIS ISMAIL received the B.S. degree in computer graphics and computer vision and the M.S. and Ph.D. degrees from Universiti Teknologi Malaysia (UTM), Johore, Malaysia, in 2016. Her current research interest includes augmented reality. Her research was vision-based tracking in augmented reality. In 2013, she joined the Human Interface Technology Laboratory New Zealand (HITLabNZ), University of Canterbury, as a Researcher, where she has completed her research in the Ph.D. degree in three years. She is currently the Head of the Mixed and Virtual Environment Research Laboratory (Mivielab), UTM, where she is also a Senior Lecturer. Her research interests include augmented reality and mixed reality environments.

YINGYING LI was born in 1994. She is currently pursuing the Ph.D. degree in computer science with Universiti Teknologi Malaysia (UTM), Johore, Malaysia. In 2019, she was with the School of Software, Shanxi Agricultural University, China, mainly engaged in educational administration and laboratory construction. Her research interest includes virtual reality (VR).

LIMIN ZHANG received the Ph.D. degree in control science and engineering from Yanshan University, Hebei, China, in 2016. He is currently with Hengshui University as an Associate Professor. He has been involved in more than four projects supported by the National Natural Science Foundation of China, the National Natural Science Foundation of Hebei, and other important foundations. His research interests include system identication, data mining, and machine learning.



FAZLIATY EDORA FADZLI received the B.S. degree (Hons.) in computer science (graphics and multimedia) from Universiti Teknologi Malaysia, in 2019, and the Ph.D. degree in computer science from Universiti Teknologi Malaysia, by enhancing a life-size holographic telepresence framework with real-time 3D reconstruction for dynamic scene. She was offered a fast-track offer letter to a full-time study after receiving the B.S. degree.