IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 8, 2021, accepted July 19, 2021, date of publication July 26, 2021, date of current version July 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3099111

Hybrid Time-Series Framework for Daily-Based $PM_{2.5}$ Forecasting

PEI-WEN CHIANG, (Member, IEEE), AND SHI-JINN HORNG
Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan

Corresponding author: Shi-Jinn Horng (horngsj@yahoo.com.tw)

This work was supported in part by the "Center for Cyber-Physical System Innovation" from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE), Taiwan; and in part by the Ministry of Science and Technology (MOST), Taiwan, under Grant 109-2221-E-011-115-, Grant 110-2221-E-011-125-, and Grant 110-2218-E-011-006-MBK.

ABSTRACT The impact of fine particulate matter on health has captured attention worldwide. Many studies have proven that fine particulate matter harms the respiratory system and the cardiovascular system. To prevent people from being harmed, many scientific research studies on $PM_{2.5}$ prediction have been conducted in recent years. Accurate $PM_{2.5}$ forecasting can not only alert people to stay away from concentrated areas but also provide the government with environmental policies in the future. In this paper, we propose a hybrid time-series prediction framework for daily-based $PM_{2.5}$ forecasting. The proposed framework consists of three components: the autoencoder, the dilated convolutional neural network, and the gated recurrent unit. The experimental dataset with 76 monitoring stations from the Taiwan Environmental Protection Administration is applied for comparison of the baseline and the proposed models. The proposed model is not only for the specified city-/county-wide region but also for the particular monitoring station/site to predict $PM_{2.5}$ concentration. By considering air quality data, meteorological data, and geographical data simultaneously, the proposed model can increase the accuracy of $PM_{2.5}$ prediction. In addition, the proposed $PM_{2.5}$ forecasting model can learn the location-centric spatial features and the daily-based temporal features simultaneously. The experimental results show that the prediction accuracy of the proposed model is superior to those of the baseline models.

INDEX TERMS Autoencoder, dilated convolutional neural network, gated recurrent unit, $PM_{2.5}$ forecasting.

I. INTRODUCTION

From many research studies of air quality in recent years, the results revealed that rapid climate change and serious environmental pollution have impacted human health worldwide. Additionally, many studies have shown that air quality prediction becomes more important for the living environment, particularly fine particulate matter ($PM_{2.5}$). From the viewpoint of the government, the warning of $PM_{2.5}$ concentration can be helpful to make environmental policies and to remind citizens to stay away from polluted areas. Hence, $PM_{2.5}$ forecasting and monitoring are not only national but also international topics for humans.

To live in a healthy environment, many studies have addressed air pollution intensity and air quality forecasting [1], [2]. Additionally, most of the research studies applied either theoretical methods or simulation models to highlight the situation under air pollution [3]. Machine learning has been applied to predict air quality. Dong *et al.* [4] proposed a method for $PM_{2.5}$ forecasting by using the hidden semi-Markov model (HSMM). Donnelly *et al.* [5] proposed real-time air quality forecasting, which is based on integrated parametric and nonparametric regression. Furthermore, weather and climate trends are considerably relevant to air quality; hence, applying traditional machine learning models is insufficient for air quality prediction.

In recent years, some studies have proposed a novel method that is used for prediction models. The equipment degradation processes possessed long-range dependence and multimode characteristics. The causes of the multimode characteristics include the external environment and operating conditions, as well as the equipment loads throughout its lifetime. Duan *et al.* [6] developed a multimodal fractional Lévy stable motion degradation model, which is used to predict the product technical life or the remaining useful life of equipment. Liu *et al.* [7] proposed a prediction model of

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

the remaining useful life based on the generalized Cauchy stochastic process. Reliable gearbox prediction is a complex problem. To overcome the gearbox reliability problem, a hybrid model based on the fractional Lévy stable motion, the gray model and the metabolism method was proposed [8]. In this model, the feature extraction method is used to reveal gearbox degradation and to solve gearbox insensitivity to weak faults. In 2021, a new long-range-dependent degradation model was proposed to predict the remaining useful life of rolling bearings [9]. The degradation model is based on the generalized Cauchy process and can describe the local irregularities and the global correlation characteristics of the time-series data. Linear mathematical models cannot adequately describe wind-speed characteristics. To overcome the limitations of the linearity assumption, a novel model based on the generalized Cauchy process was introduced [10]. In this model, the fractal dimension and Hurst parameter are combined for simulation and forecasting of the wind speed. Furthermore, the prediction model is applicable to describe the local irregularity and global correlation of wind speed.

Many studies have addressed the interrelation between air pollution factors such as $PM_{2.5}$ and meteorological data. Traditional machine learning lacks the adoption of data-driven approaches to process time-series air quality data [11], [12]. Nevertheless, deep learning models can apply to data-driven methods [13], in which the features of air quality data are competently extracted. In the domains of image classification, speech recognition, and natural language processing, deep learning models have made remarkable achievements [14]–[16]. Shallow learning models are utilized to predict air quality, and deep learning approaches are appropriately applied to predict time-series air quality data [17]–[20]. In [21], a multivariate time-series method was conducted for air quality prediction. In addition, deep learning models are most suitable for dealing with data-driven approaches and applicable to handle time-series data.

A hybrid time-series framework for daily-based $PM_{2.5}$ forecasting is proposed in this paper. The nationwide and city-/county-wide regions for air quality prediction are all covered. Both the temporal and spatial correlation dependencies of features are learned from air quality time-series data, such as wind speed, $PM_{2.5}$, and coordinates. We summarize the major contributions of this paper in the following paragraphs.

First, we develop a hybrid time-series deep learning model, which is composed of three components. Autoencoder (AE) and dilated convolutional neural network (CNN) learn spatial features through air quality and geographical data. Gated recurrent unit (GRU) extracts temporal features through air quality and meteorological data. Compared with the existing models, such as the ST-DNN model proposed by Soh *et al.* [22], the proposed model can decrease the average MAE and RMSE values by 16% and 18%, respectively. In addition, our model also shortens the average training time by 12%.

Second, according to the experimental results, the proposed model is not only accurate on a nationwide scale but also adequate for region-wide prediction of time-series data such as air quality. Furthermore, our model can be applied to predict the air quality of the target site. The proposed model is superior to the existing prediction models.

The rest of this paper is organized as follows. Section II describes the related research works. The methodology of the proposed $PM_{2.5}$ forecasting framework is presented in Section III. Section IV depicts the experimental setup and the prediction comparison, of which the prediction error and the training time are considered in our work. We show the conclusion and future work in Section V.

II. RELATED WORKS

For the literature on $PM_{2.5}$ forecasting, almost all existing works dealt with the prediction accuracy of air pollutants by using machine learning models and statistical methods [3], such as HMM [4], regression [5], artificial neural networks [23], and ARIMA [24]. Zhang *et al.* [1], [2] proposed a real-time air quality prediction approach, which focuses on the analysis of the major research trends and current status, as well as future directions. In Zhou *et al.* [25], the hidden temporal dependencies regarding $PM_{2.5}$ were addressed with Lasso-Granger by developing a probabilistic dynamic causal (PDC) model. The hybrid model is based on the regression neural network and empirical mode decomposition for the previous 24-hour $PM_{2.5}$ prediction proposed by Zhou *et al.* [23]. In Deleawe *et al.* [26], a machine learning model that conducts air quality measurements in the urban environment was used to predict the CO_2 levels.

In addition, many studies show that deep learning models have been used for air quality prediction. Air quality data possess time series and nonlinear characteristics, and thus, data-driven models are directed to address the topic of urban computing [27]. Moreover, many $PM_{2.5}$ forecasting studies are based on big data, which can obtain the predicted results by adopting many historical and multivariate data [28]. Zheng *et al.* [11] proposed a semisupervised learning model that combines CRF with ANN classifiers. Hsieh *et al.* [12] developed a semisupervised method to conduct fine-grained and real-time air quality data. An air quality prediction framework in real time that applies data-driven models was presented in Zheng *et al.* [29].

With regard to the capability of data-driven methods and nonlinear problems, deep learning models have generally been adopted to solve time-series and sequence data problems [6], [30], [31]. That is, air quality data possess characteristics such as time-series data. In [32], Li *et al.* [32] presented a novel air quality forecasting model by utilizing spatial-temporal deep learning (STDL), which considers the temporal and spatial correlations. To predict air pollution, Ong *et al.* [33] developed a deep recurrent neural network (DRNN) by adopting the autoencoder approach. In [18], Qi *et al.* [18] proposed a deep air learning (DAL) model to deal with feature analysis, interpolation and forecasting.

In [34], Zhang *et al.* [34] developed a deep residual neural network to extract the features of time-series data and analyzed the congestion of citywide crowds.

In [35], a hybrid deep learning framework was developed, which was combined with multiple deep neural network models. The hybrid framework has also been applied to the topics of video classification and face detection [36]. Additionally, hybrid deep learning frameworks have not been well fitted to handle air quality forecasting issue predictions [37]. Many researchers have shown that hybrid deep learning models have exceptional performances compared to classic deep learning models [36].

Many works have shown that convolutional neural networks (CNNs) have excellent performance in video processing and image recognition [16]. Indeed, it is also applied to time-series data prediction [31]. Among deep learning models, CNNs are superior to time-series data and multivariate data. Recurrent neural networks (RNNs) are applicable for learning the time dependencies and extracting the temporal features of time-series data. Furthermore, to mitigate the vanishing gradient problems of RNNs, Hochreiter *et al.* [38] developed a variant of RNN, LSTM, which refers to the internal states of the memory cells for mitigating the vanishing gradient problems. RNN can predict time-series data; moreover, LSTM is excellent in time-dependent feature extraction.

Du *et al.* [37] proposed a hybrid deep learning model to predict $PM_{2.5}$ in Beijing, and the hybrid model was composed of 1D-CNN and Bi-LSTM. In his work, two datasets were applied to $PM_{2.5}$ forecasting: the Beijing $PM_{2.5}$ Dataset from the US Embassy in Beijing, and the Urban Air Quality Dataset from the Urban Air Project of Microsoft Research. The experimental results show that 1D-CNN can effectively extract the local trends and spatial features from the air quality time-series data. Soh *et al.* [22] developed an adaptive deep learning model to forecast $PM_{2.5}$ in Taiwan and Beijing. In his work, the Taiwan dataset was collected from the Taiwan Environmental Protection Administration, and the Beijing dataset was provided by the Urban Air Quality Dataset from the Urban Air Project of Microsoft Research. The experimental results also showed that CNN can extract the surrounding targets and spatial features from the air quality data.

Du *et al.* [37] proposed a deep air quality forecasting framework (DAQFF), which is composed of various deep learning models. The main idea of the DAQFF is not only to deal with time-series forecasting issues but also to address spatiotemporal data features. In addition, the DAQFF correlates the multivariate air quality data. The DAQFF can extract the temporal features as well as the spatial features from air quality data. One-dimensional CNN and bidirectional LSTM are two components of DAQFF; the former deals with the spatial data features, and the latter deals with the temporal data features [39], [40].

In [41], Yu and Koltun proposed a convolutional network approach to conduct dense forecasting, and the proposed model applies dilation factors to aggregate multiscale contexts without degrading the resolution. The proposed model introduces dilation factors for the sake of expanding the receptive fields. In addition, dilated convolution can increase the accuracy and efficiency of dense forecasting.

Soh *et al.* [22] developed an adaptive air quality prediction model that includes multiple deep learning models. The spatial-temporal deep neural network (ST-DNN) considers terrain and meteorological data concurrently, which means that ST-DNN can extract spatial and temporal features from air quality data. ST-DNN combines three deep learning models: the first two models are artificial neural network (ANN) and long short-term memory (LSTM), which extracts the temporal features from the air quality data; the last one is the convolutional neural network (CNN), which extracts the spatial features from the air quality data. In summary, ST-DNN can extract spatial correlations and the temporal dependencies of neighboring locations. Zhang *et al.* [42] combined CNN and LSTM to achieve higher forecasting accuracy of air pollution.

III. METHODOLOGY

In this section, we first present the framework of this paper and then state three main components in the framework. That is, the autoencoder, one-dimensional CNN and GRU are described in order.

In this work, our motivation is to develop a deep learning framework for time-series $PM_{2.5}$ forecasting. To consider the comparability and fairness of the model performances, we compare the classic and existing deep learning models and choose a traditional machine learning model for comparison. In this paper, $PM_{2.5}$ forecasting considers both the location correlation of multiple monitoring stations and the time dependency of a single monitoring station. CNNs can effectively extract the local trends and spatial features of different districts. The GRU possesses a memory mechanism, so it can effectively extract the short- and long-term temporal features of a particular district.

Fig. 1 depicts the components and functionalities of our framework. On the left-hand side, the air quality and geographical data are first input into the autoencoder layers and then into the dilated convolution layers. On the right-hand side, the air quality and meteorological data are input into the GRU layers. Afterward, the air quality input data are output from the abovementioned layers and merged with the concatenate layer for data fusion. One flattened layer is then introduced to feed the following dense layer with the input data. Finally, the predicted $PM_{2.5}$ values are generated.

By considering the interrelated factors of variant data sources, we consider the correlation of geographical areas, meteorological conditions, and air quality time-series data simultaneously. AE and dilated CNN can effectively extract $PM_{2.5}$ concentrations from the specified district based on historical air quality, and GRU can effectively extract $PM_{2.5}$ concentrations from the seasonal climate based on historical air quality. To predict the $PM_{2.5}$ concentration of particular monitoring stations under various climate circumstances,

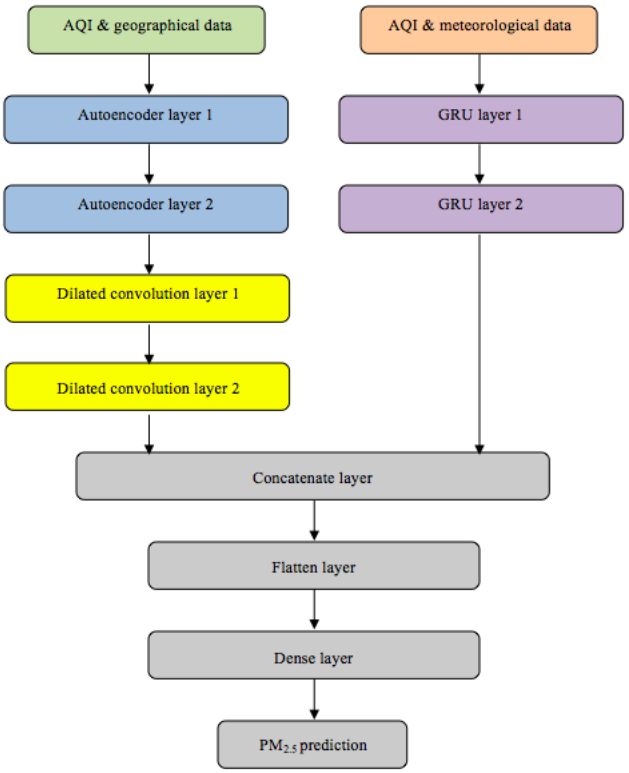

FIGURE 1. Framework diagram.

we combine the learning results of AE and dilated CNN with GRU for all the time periods.

The proposed model is named the hybrid time-series framework (HTSFW), which combines unsupervised and supervised models for daily-based $PM_{2.5}$ forecasting. The HTSFW model applies AE and dilated CNN to extract the local trends and spatial features. Additionally, HTSFW utilizes GRU to extract the long dependencies and temporal features.

For conducting the spatiotemporal features of the air quality data, we extract the spatial features from the $PM_{2.5}$ values correlated with the monitoring station locations, such as the coordinates; in addition, we concurrently retrieve the temporal features from the $PM_{2.5}$ records interrelated to the weather and climate factors, such as the wind speed.

The air quality-related dataset is composed of geographical data and meteorological data, such as $PM_{2.5}$, longitude, latitude, SO_2, CO, O_3, NO_2, PM_{10}, wind speed, temperature and humidity. To increase prediction accuracy, the HTSFW model consolidates the training results of the geographical data and the meteorological data. In the HTSFW model, the missing values of the air quality dataset were padded with zeros. In other words, the same experimental dataset was applied to all the baseline and HTSFW models. The data contents are recorded by day; hence, $PM_{2.5}$ forecasting generates daily time frames.

As is known, the innovation of the HTSFW model can not only consider a single monitoring site for time-dependent meteorological factors but also examine wide regions for location-interrelated geographical factors. In addition, the HTSFW model combines the unsupervised with the supervised models, which has excellent performance in $PM_{2.5}$ forecasting. The details of the HTSFW model are discussed in the following subsections.

A. AUTOENCODER

The first model used in the proposed framework for the air quality and the geographical data and autoencoder is an unsupervised deep learning model [10]. The simplest autoencoder network has one hidden layer. The input layer first encodes the high-dimensional data; then, the hidden layer stores the low-dimensional codes as the intended data features. Additionally, the functionality of the output layer is to use the low-dimensional codes to reconstruct the high-dimensional input vectors.

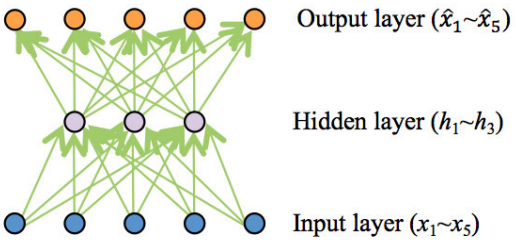

FIGURE 2. Autoencoder operation.

Fig. 2 depicts the autoencoder operation, which possesses only one hidden layer. The dimensionality of the input layer is five, and the input data are encoded and then stored in the hidden layer. Therefore, fewer neurons of the hidden layer lead to the outcome of data compression or dimensionality reduction. With the same dimensionality as the input layer, the purpose of the output layer is to decode the hidden representation from the previous layer and reconstruct it to the original input data.

Furthermore, multiple hidden layers can be constructed to form the stacked autoencoder (SAE) network. In the encoding phase, the dimensionality of the next layer has fewer neurons than the previous layer, which means that each neuron ignores useless data and keeps meaningful data. During the training process, backpropagation can be utilized for fine-tuning the connection weights. In the decoding phase, in contrast to the encoding phase, the dimensionality of the next layer possesses more neurons than the previous layer; each neuron learns the data features and reconstructs the original input.

In addition to dealing with the geographical relationship set of the target stations for $PM_{2.5}$ forecasting, we develop an approach to select the target region or site, which is simultaneously applicable for extracting the local trends and long dependencies. Algorithm 1 depicts the proposed method.

Here, d_i and s_j indicate the particular district and station, respectively, and i and j denote the district and station

Algorithm 1 Select the Geographical Relationship Set for the Target Stations, Where Target Districts Are Given by d_{target}

```
Let district D = {d_i | i = 1, ..., m}, station S = {s_j | j = 1, ..., n}, time T = {t_k | k = 1, ..., o}
if ∀d_i ⊂ D then
  sort d_i by ascending order
  for i = 1 to m do
    if d_i = d_target then
      sort s_j contained in d_i by ascending order
      sort t_k recorded in s_j by ascending order
    end if
  end for
end if
```

identifiers, respectively. Moreover, t_k denotes the timestamp of the data record, and k represents the timestamp index.

Algorithm 1 first sorts the district identifiers in ascending order; the HTSFW model defines the target district identifier afterward. Then, the proposed algorithm searches the matching district identifier and sets the matching district to be the target district. After that, the monitoring stations of the target district are sorted by the station identifiers in ascending order. Finally, each target monitoring station is sorted by the timestamps in ascending order.

As shown in Algorithm 1, the target district identifier can refer to a single district or multiple districts. In other words, the regional coverage of $PM_{2.5}$ forecasting can be dynamically conducted in accordance with the proposed algorithm.

In this paper, an experimental dataset with 76 monitoring stations (Jan. 2014 to Jun. 2019) in Taiwan is downloaded, which was provided by the Taiwan Environmental Protection Administration (TWEPA). In the downloaded dataset, the missing values of the data items were recorded to be empty. To deal with the missing values in the dataset, we padded the mentioned data items with zeros.

B. ONE-DIMENSIONAL DILATED CONVOLUTIONAL NEURAL NETWORK

SVM was developed in 1992 and is a supervised machine learning model. It has been widely used for data classification and regression. SVM is also a forecasting method based on a statistical learning framework. One of the SVM models, LSSVM, is a least-squares version of SVM, which is to minimize the sum of the squared errors of the objective function. SVR, another version of SVM, was proposed in 1996 and is mainly used for data regression. Du *et al.* [37] proposed the DAQFF model, which can accurately predict $PM_{2.5}$ concentrations. To compare the performances of the proposed and baseline models, SVR, ARIMA, LSTM, GRU, CNN and RNN were tested in his work. For the next one-hour prediction of the Beijing $PM_{2.5}$ Dataset, RMSE values were 42.61, 41.86, 30.60 and 12.21 for SVR-POLY, SVR-RBF, SVR-LINEAR and CNN, respectively, while MAE values were 31.82, 34.93, 20.47 and 9.09 for SVR-POLY, SVR-RBF, SVR-LINEAR and CNN, respectively. For the next one-hour prediction of the Urban Air Quality Dataset, RMSE values were 56.35, 50.51, 29.23 and 20.95 for SVR-POLY, SVR-RBF, SVR-LINEAR and CNN, respectively, while MAE values were 47.20, 42.26, 18.82 and 16.36 for SVR-POLY, SVR-RBF, SVR-LINEAR and CNN, respectively. Based on the previously mentioned work, we, therefore, choose CNN to build our $PM_{2.5}$ forecasting framework. In our work, the data interval is 24 hours, and the forecasting horizons are one, two, three, six, 12, 24, 36, 48, 60 and 72 day(s).

The next model used in the proposed framework for the air quality and geographical data is a one-dimensional dilated CNN. The main consideration is stated as follows. CNNs are well suited for spatial feature extraction, while one-dimensional CNNs apply to time-series data. The dilated convolution network can extract air quality features from the input time-series data. In addition, using dilation factors can expand the receptive fields and further enhance the training efficiency. Both are discussed in the following subsections.

1) 1D-CNN FOR TIME-SERIES DATA FEATURE EXTRACTION

In the image processing field, convolutional neural networks are commonly adopted [16]. Nevertheless, CNN is also applied to time-series data prediction. The classical CNN commonly consists of convolutional, activation and pooling layers. Furthermore, the two-dimensional CNN is popularly utilized for image classification [35]. In this work, the HTSFW model utilizes the one-dimensional CNN to predict the $PM_{2.5}$ concentration. In general, the activation function of the one-dimensional convolutional layer is depicted as follows:

$$s_i^t(x^t) = ReLU(s_{i-1}^t(x^t) * w_i^t + b_i^t). \tag{1}$$

As shown in Equation (1) is the activation function of the convolutional layer, where $*$ indicates a convolution operator, w_i^t and b_i^t represent the weights and biases, respectively, and x^t represents the t-th time step of the input data.

The HTSFW model uses two connected one-dimensional convolutional layers to extract the spatial features from the geographical data, and the two connected layers constitute the hierarchy to represent the local trend features. In other words, the two connected one-dimensional CNNs can learn the local trend features of a single monitoring station and can also extract the hidden spatial correlation features from multiple stations.

In Fig. 3, the time-series features of the input air quality data are first filtered through the one-dimensional convolution kernel, and the activation function processes the input features, weights and biases. After that, the extraction process of the activation function generates the output targets.

Since the air quality input dataset contains time-series data items, the one-dimensional CNN is adopted to compress the length of multivariate input data and learn the air quality data features. Due to the local perception and weight sharing of the one-dimensional convolution network, the number of parameters decreases, and the learning efficiency improves.

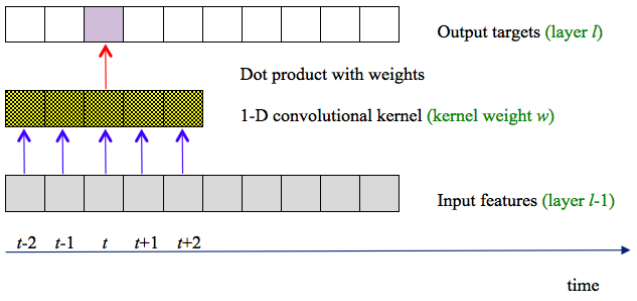

FIGURE 3. One-dimensional convolution process.

2) DILATION FOR LEARNING PHASE EFFICIENCY

In our model, we applied 1D-CNN to extract the local trend and spatial features from the time-series data. Since air quality data are related to the time sequence, we apply the dilation factors in the hidden layers to expand the receptive fields. When the air quality data are inputted to 1D-CNN, by using the dilation convolution, our model can speed up the learning process both in the single time step and the multiple time steps.

Yu and Koltun [41] developed a convolutional network module that applies dilation factors to aggregate multiscale contextual information. The dilated convolution module can exponentially expand the receptive fields without losing coverage. The dilated convolution is defined as follows:

$$(F *_l k)(p) = \sum_{s=\{\alpha,\alpha+l\}} F(s)\, k(s). \tag{2}$$

Equation (2) states the one-dimensional dilated convolution, in which dilation rate l convolves input F with kernel k, where $*_l$ denotes a dilated convolution operator and $p = \alpha + l; \alpha = \{l-1, 3l-1, 5l-1, 7l-1, \ldots, Ml-1-l\}; M$ is the input bucket size.

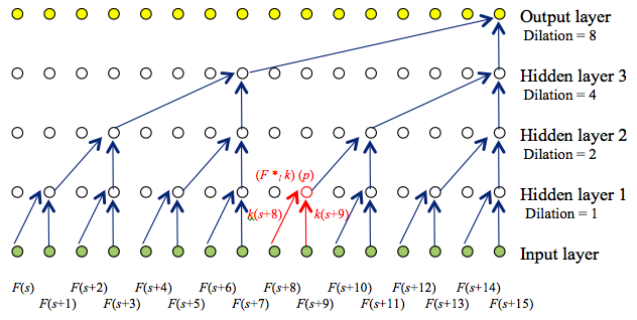

FIGURE 4. Dilated convolution operation.

Fig. 4 represents the dilated convolution operation. First, the one-dimensional dataset is input into the dilated convolution layers. Next, the dilation factors are implemented in the hidden layers; to put it differently, dilation factors are set to one, two, four and eight. Afterward, as previously mentioned, the receptive fields of the next layers can expand exponentially. As a result, the one-dimensional dilated CNN can decrease the number of parameters and reduce the training time. For example, Fig. 4 shows three hidden layers with 16 input data. Based on the depiction of the elapsing time sequence in Fig. 3, the indices start from zero and pass through the left to the right.

In [43], Zhen *et al*. presented a dilated CNN approach for sequence prediction. The approach utilizes dilation factors to extend the receptive fields and introduces residual connections to form a deeper network. The experimental video analysis results revealed performance enhancements with fewer parameters and shorter running times. Hence, inspired by the convolutional neural network with dilation factors, this work proposes a one-dimensional dilated CNN to expand the receptive fields and increase the training efficiency for air quality forecasting.

In this paper, two one-dimensional convolution layers are concatenated, and the dilation rate is applied to two for both layers. The number of output filters is 64, and the length of the 1D convolution window is one in the first dilated convolution layer. In the second dilated convolution layer, we set the number of output filters to 128, and the length of the 1D convolution window is one. The padding of the two dilated convolution layers is parameterized to the same dimension for both the input and output.

C. GATED RECURRENT UNIT

The exploding and vanishing gradient problems of traditional RNNs are inevitable. To mitigate gradient problems, a long short-term memory (LSTM) network was developed in 1997 [38]. LSTM refers to the internal states within the memory cells for mitigating the mentioned gradient problems. In 2014, another RNN variant, the gated recurrent unit (GRU), was proposed by Cho *et al.* [14]. LSTM is composed of three main building blocks: the input gate, forget gate, and output gate. In other words, GRU has a simpler network than LSTM. The GRU consists of two main components: the reset gate and update gate. For this reason, the GRU training process results in better efficiency.

The reset gate is responsible for the short-term memory, while the update gate is in charge of the long-term memory. In this work, GRU is used for certain reasons: one is the functionality of hidden state handling, which implies that GRU is well suited to time-series data prediction; the other is a simpler network, which results in a faster learning process. Therefore, GRU is feasible for extracting the long-term temporal dependency features.

Fig. 5 shows the main components of the GRU building block. The main components are combined to handle the hidden states and retain the temporal features over a period of time. The main components of the single GRU block are represented as follows:

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right), \tag{3}$$
$$z_t = \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right), \tag{4}$$
$$\hat{h}_t = \phi_h\left(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h\right), \tag{5}$$
$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \hat{h}_t. \tag{6}$$

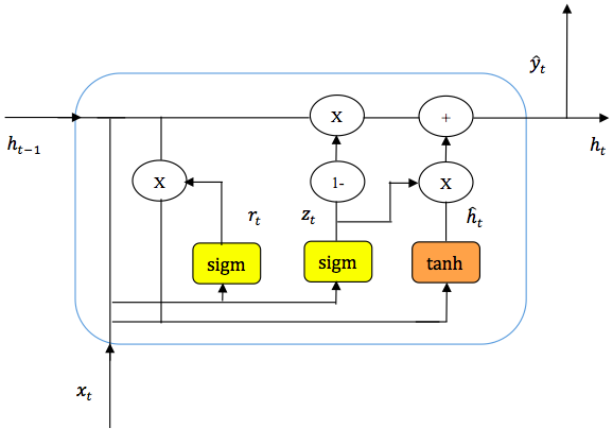

FIGURE 5. The structure of the GRU unit.

As denoted in the above formulas, r_t represents the reset gate that determines the amount of previous information to be ignored, z_t represents the update gate that determines the amount of previous information to be passed, $\hat{h}_t$ denotes the candidate hidden state, h_{t-1} represents the previous hidden state, and h_t represents the current hidden state. In addition, x_t indicates the input data, W and U are the weights, and b is the bias.

In this work, GRU is applied to retrieve the long-term and time-series features of the air quality data. The HTSFW model extracts the local trend features using AE and a one-dimensional dilated CNN; additionally, the long-term spatial-temporal correlation features hidden in multivariate time-series data are extracted by using GRU.

The local trend features and the long-term correlation features are concatenated afterward, followed by the flattened layer, which transforms the features into a vector. In addition, the vector dimensionality is reduced by the fully connected layer. Eventually, the output of the training process results in $PM_{2.5}$ forecasting.

IV. EXPERIMENTS

In this section, we first describe the experimental dataset of this work and interpret the experimental parameters and the settings afterward. Moreover, we conduct performance comparisons of the baseline and proposed models, the ST-DNN [22] and the proposed models.

A. DATASET

In this paper, the experimental dataset is sourced from the Taiwan Environmental Protection Administration (TWEPA). The dataset contains air quality-related data that cover all of the cities and counties in Taiwan, and each data record is a daily average. The data features include air quality, geographical and meteorological data such as $PM_{2.5}$, longitude, latitude, SO_2, CO, O_3, NO_2, PM_{10}, wind speed, temperature and humidity.

The time of the dataset is from Jan. 2014 to Jun. 2019, and the data interval is 24 hours. There are 76 monitoring stations built around all cities and counties in Taiwan. To deal with the missing values in the dataset, those recorded as empty were padded with zeros.

Many works show that air quality is highly related to meteorological circumstances. For instance, moderate air pollution is due to high wind speed, good air quality can be due to high atmospheric pressure, and high humidity deteriorates the $PM_{2.5}$ concentration [11], [29].

To prepare the experimental dataset, we sort the data items by geographical location after downloading the raw data from the TWEPA. That is, the neighboring cities/counties are sorted in order. For data preprocessing, we extract the spatial correlation features between the air quality and the geographical regions.

In the experimental dataset, each monitoring station possesses air quality, geographical and meteorological data features. In addition, each data item contains a recorded timestamp. The 76 monitoring stations are distributed in Taiwan; therefore, this work can predict the $PM_{2.5}$ concentration in Taiwan and the city-/county-wide regions, as well as the particular sites.

B. EXPERIMENTAL SETUP

This section describes the system setting of the experimental environment and the parameters of the HTSFW model. The experimental environment is configured on a Windows 10 system, and the resources are equipped with an Intel(R) Xeon(R) CPU E5-2620 2.10 GHz and 32 GB memory. Our experiments are completed in Python 3.7.3 with the deep learning libraries Keras 2.2.4 on top of TensorFlow 1.9.0 to execute the baseline and the proposed models.

The HTSFW model is compared with five deep learning models and one machine learning model. The baseline models embrace convolutional neural networks (CNNs), recurrent neural networks (RNNs), and two variants of RNNs, which include long short-term memory (LSTM) and gated recurrent units (GRUs), and support vector regression (SVR). To further compare with the ST-DNN model [22], it is also considered a baseline model.

In our work, the default parameters in Keras are applied to weight initialization. During the training phase, for overfitting prevention, a dropout rate of 0.3 is configured. In addition, the lookup size, batch and epoch are 1, 32 and 100, respectively. The activation functions of CNN and RNN (including LSTM and GRU) are *ReLU* and *tanh*, respectively. Adam is used as the optimizer. The learning rate is set to 0.001. The values of beta one and beta two are 0.9 and 0.999, respectively. The epsilon is 1e-7.

For the baseline models, the number of hidden layers is set to one by default, and each hidden layer possesses 128 neurons. In the HTSFW model, we first utilize two AE layers that concatenate with two one-dimensional convolution layers with a dilation factor of two for air quality and geographical feature extraction. In addition, the number of output filters and the length of the convolution window of each layer are applied to (64, 1) and (128, 1), respectively.

TABLE 1. Prediction error comparison between the baseline and proposed models.

Model	Prediction Size	MAE	Improvement (%)	RMSE	Improvement (%)
CNN/Ours	1	4.220/3.101	26.52	6.508/5.039	22.57
	2	5.937/5.073	14.55	8.122/7.424	8.59
	3	7.450/5.924	20.48	9.691/8.833	8.85
	6	7.842/6.709	11.78	10.508/10.009	2.47
	12	8.344/7.243	13.20	10.840/10.267	5.29
	24	9.191/7.513	18.26	11.913/10.346	13.15
	36	9.943/8.551	14.00	12.987/11.247	13.40
	48	10.514/8.612	18.09	13.265/11.526	13.11
	60	10.602/9.147	13.72	13.639/12.106	11.24
	72	11.954/10.742	5.21	14.960/13.666	5.47
RNN/Ours	1	4.186/3.101	25.92	5.437/5.039	7.32
	2	6.253/5.073	18.87	7.869/7.424	5.66
	3	7.137/5.924	17.00	9.074/8.833	2.66
	6	8.114/6.709	14.74	10.172/10.009	1.60
	12	8.496/7.243	14.75	10.439/10.267	1.65
	24	8.726/7.513	13.90	10.769/10.346	3.93
	36	9.942/8.551	13.99	12.136/11.247	7.33
	48	10.304/8.612	16.42	12.501/11.526	7.80
	60	11.409/9.147	19.83	13.692/12.106	11.58
	72	11.806/10.742	4.02	14.192/13.666	2.97
LSTM/Ours	1	5.179/3.101	40.12	8.704/5.039	42.11
	2	6.841/5.073	25.84	9.706/7.424	23.51
	3	7.540/5.924	21.43	10.070/8.833	12.28
	6	8.010/6.709	13.63	10.512/10.009	2.51
	12	8.252/7.243	12.23	10.546/10.267	2.65
	24	8.471/7.513	11.31	10.751/10.346	3.77
	36	8.907/8.551	4.00	11.770/11.247	4.44
	48	10.175/8.612	15.36	12.663/11.526	8.98
	60	10.550/9.147	13.30	13.217/12.106	8.41
	72	12.492/10.742	9.29	15.585/13.666	9.27

Note: prediction sizes are the following time steps measured by day(s); improvement denotes the reduction ratio of the prediction error; the measuring unit of MAE and RMSE is $\mu g/m^3$.

We also adopt each of the two GRU layers with 128 hidden neurons for air quality and meteorological feature extraction. The mean square error (MSE) is used as the loss function in the training process, and the activation function used in the output layer for target prediction is *tanh*. Furthermore, the input time-series data are normalized to [0, 1] by using the min-max function, while the missing values in the dataset of the data items padded the data items with zeros. The dataset is divided into two parts: the first four years of data, and the last 18 months of data. The former part is used for training, and the latter part is used for testing.

For training process evaluation, two metrics, MAE and RMSE, are applied to measure the learning performances. The two error indices are denoted as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}, \tag{8}$$

where n stands for the number of testing data, y_i indicates the actual $PM_{2.5}$ value, and $\hat{y}_i$ represents the predicted $PM_{2.5}$ value.

C. PERFORMANCE COMPARISON OF BASELINE AND PROPOSED MODELS

In this section, the HTSFW model is compared with the baseline models. It is noted that the coverage area of $PM_{2.5}$ prediction is all the cities and counties in Taiwan, and the air quality data are from the 76 monitoring stations. In addition, both the single and multiple time steps of $PM_{2.5}$ forecasting are shown in Tables 1 and 2.

Since all the data items are 24-hour averages, one step indicates one day later, and multistep represents the specified days later. The prediction size in Tables 1 and 2 specifies a single time step and multiple time steps. That is, the predicted size of one means the $PM_{2.5}$ prediction one day later. In this experiment, CNN, RNN, LSTM, GRU, SVR, and ST-DNN

TABLE 2. Prediction error comparison between the baseline (ST-DNN [22]) and proposed models.

Model	Prediction Size	MAE	Improvement (%)	RMSE	Improvement (%)
GRU/Ours	1	4.855/3.101	36.13	7.903/5.039	36.24
	2	6.659/5.073	23.82	9.008/7.424	17.58
	3	7.299/5.924	18.84	9.542/8.833	7.43
	6	7.643/6.709	9.49	10.203/10.009	2.16
	12	8.179/7.243	11.44	10.426/10.267	1.53
	24	8.580/7.513	12.44	10.861/10.346	4.74
	36	8.709/8.551	1.81	11.490/11.247	2.11
	48	9.793/8.612	12.06	12.175/11.526	5.33
	60	10.474/9.147	12.67	13.066/12.106	7.35
	72	11.781/10.742	3.82	14.514/13.666	2.57
SVR/Ours	1	9.191/3.101	66.26	11.572/5.039	56.46
	2	9.364/5.073	45.82	11.704/7.424	36.57
	3	9.627/5.924	38.46	11.752/8.833	24.84
	6	9.982/6.709	32.79	12.705/10.009	21.22
	12	10.558/7.243	31.81	13.079/10.267	21.55
	24	10.622/7.513	29.27	13.087/10.346	20.94
	36	11.919/8.551	28.26	14.448/11.247	22.16
	48	11.956/8.612	27.97	14.495/11.526	20.48
	60	12.037/9.147	24.01	14.715/12.106	17.73
	72	12.321/10.742	12.82	14.828/13.666	7.84
ST-DNN/Ours	1	6.616/3.101	53.13	9.370/5.039	46.22
	2	7.462/5.073	32.02	11.002/7.424	32.52
	3	7.766/5.924	23.72	11.261/8.833	21.56
	6	8.222/6.709	15.86	11.865/10.009	13.63
	12	8.427/7.243	14.05	12.184/10.267	15.73
	24	8.878/7.513	15.38	12.780/10.346	19.05
	36	8.934/8.551	4.29	12.844/11.247	12.43
	48	9.527/8.612	9.60	13.133/11.526	12.24
	60	10.201/9.147	10.33	14.283/12.106	15.24
	72	10.874/10.742	1.21	14.497/13.666	5.73

Note: prediction sizes are the following time steps measured by day(s); improvement denotes the reduction ratio of the prediction error; the measuring unit of MAE and RMSE is $\mu g/m^3$.

of baseline models and the proposed HTSFW model are compared by indices MAE and RMSE, respectively.

As shown in Table 1, the performance of $PM_{2.5}$ forecasting 24 hours later shows that the HTSFW model can decrease the ratio of MAE by 26.52%, 25.92% and 40.12% by using CNN, RNN and LSTM, respectively. The HTSFW model can reduce the percentage of RMSE by 22.57%, 7.32% and 42.11% by applying the respective CNN, RNN and LSTM models.

In Table 2, for the $PM_{2.5}$ prediction one day later, the performance indicates that the HTSFW model can decrease the proportion of MAE by 36.13%, 66.26% and 53.13% by utilizing GRU, SVR and ST-DNN, respectively. In addition, the HTSFW model can reduce the scale of RMSE by 36.24%, 56.46% and 46.22% with GRU, SVR and ST-DNN, respectively.

In addition, we highlight the preprocessing of the missing data items in the experimental dataset. The lack or absence of data item values is inevitable, such as the malfunction of the monitoring station over a five-year duration. The handling of the missing data values is important because it is related to the experiment proceeding. In this work, the missing values of the experimental dataset are padded with zeros. The missing values with the zero-padding approach are applied to all the baseline models and the proposed model.

As shown in Tables 1 and 2, the prediction performances of the proposed model are compared with four classic deep learning models, one traditional machine learning model and one existing deep learning model. The comparison includes single- and multistep forecasting performances, and the prediction sizes are parameterized to ten time steps. In addition, MAE and RMSE are used for performance comparison. In our model, the columns of improvement show a reduction in the ratio of the prediction errors. Based on the improvement values, the proposed model is superior to the six baseline models in the ten time steps by using MAE and RMSE.

The recording period of the experimental dataset is from 01/01/2014 to 06/30/2019, and the data interval is 24 hours. There are 148,399 data samples in total. The time of the training data is from 01/01/2014 to 12/31/2017, and the number of data samples is 107,357 (72%); the time of the testing data is from 01/01/2018 to 06/30/2019, and the number of data

samples is 41,042 (28%). To consider all the seasonal factors affecting the $PM_{2.5}$ concentration, we use a forecasting period between 01/01/2018 and 12/31/2018. In terms of the performance comparison of all the forecasting models, the HTSFW model can also extract the time-series data features from the seasonal interrelation.

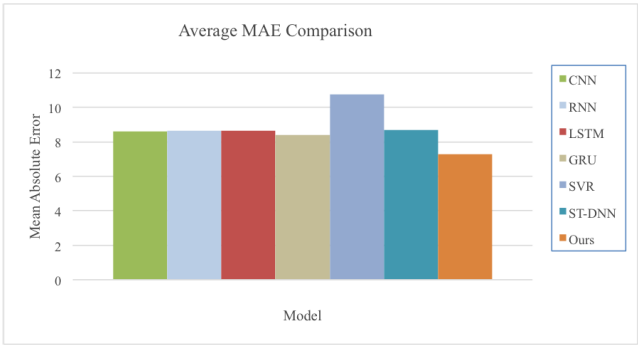

FIGURE 6. Average MAE comparison. Note: average MAE in ten different time steps for each model; the measuring unit of MAE is $\mu g/m^3$.

Fig. 6 depicts the average values of the MAE metric in ten different time steps for each baseline and our model. For $PM_{2.5}$ prediction analysis, we further average the MAE values for each model based on MAE in Tables 1 and 2. According to the bar charts corresponding to MAE averages plotted in Fig. 6, the prediction errors of the HTSFW model are clearly lower than those of the baseline models. SVR has the highest MAE average of 10.76, while HTSFW has the lowest MAE average of 7.26, which means that the proposed model can reduce the prediction errors for both short-term and long-term periods.

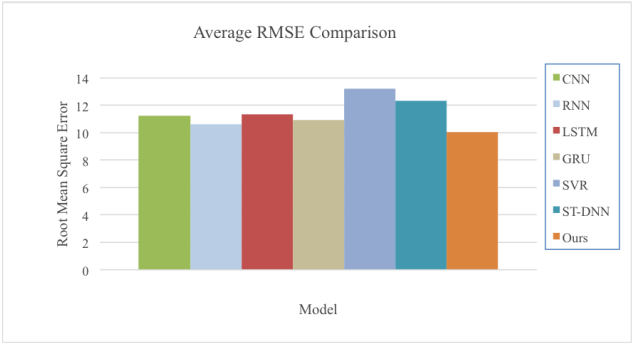

FIGURE 7. Average RMSE comparison. Note: average RMSE in ten different time steps for each model; the measuring unit of RMSE is $\mu g/m^3$.

Fig. 7 denotes the average values of the RMSE index in ten different time steps for the entire baseline and the proposed model. As shown in the bar charts, HTSFW reveals the lowest prediction error. The highest RMSE average of 13.24 is SVR, and the lowest MAE average of 10.05 is HTSFW. In summary, we compare the prediction errors of HTSFW with the existing baseline models by MAE and RMSE metrics, and the resulting figures indicate that the HTSFW model can remarkably decrease the $PM_{2.5}$ prediction errors.

TABLE 3. Training time comparison between the baseline (ST-DNN [22]) and proposed models.

Model	Prediction Size	Time	Improvement (%)
ST-DNN/ Ours	1	19.61/ 18.00	8.23
	2	20.34/ 17.72	12.88
	3	19.90/ 17.84	10.39
	6	20.22/ 17.80	11.96
	12	21.80/ 18.38	15.67
	24	20.00/ 18.16	9.22
	36	20.87/ 18.16	12.97
	48	21.27/ 18.06	15.08
	60	20.48/ 18.54	9.46
	72	21.19/ 18.42	13.08

Note: prediction sizes are the following time steps measured by day(s); time denotes the training time of the experiments measured in seconds; improvement denotes the reduction ratio of the training time.

According to ten different time step performance comparisons in Tables 1 and 2, the HTSFW model is superior to all the baseline models overall. Based on the experimental results, the contribution of this paper is stated as follows. First, the HTSFW model can accurately predict $PM_{2.5}$ concentrations in various regions concurrently, for example, urban and rural areas. In other words, HTSFW can predict $PM_{2.5}$ concentrations for monitoring stations located nationwide. Second, the HTSFW model can also accurately predict $PM_{2.5}$ concentrations in not only a short period of time but also a long period of time. $PM_{2.5}$ forecasting in those periods is superior to the existing deep learning models. Therefore, HTSFW is applicable for predicting $PM_{2.5}$ concentrations in both single time step and multiple time steps.

D. PERFORMANCE COMPARISON BETWEEN ST-DNN AND PROPOSED MODELS

In this section, the HTSFW model is further compared to the ST-DNN model [22]. First, we compare the training time for all 76 monitoring stations in ten different time steps. Next, we select four cities/counties in Taiwan and compare the prediction errors of the monitoring stations located at each of the four cities/counties. In addition to comparing the prediction errors, we further select four particular sites that are located in the four cities/counties.

As shown in Table 3, we compare the training time of $PM_{2.5}$ forecasting of ST-DNN with that of HTSFW. In other words, the training time of all the monitoring stations in Taiwan is evaluated. In accordance with the training time comparison, HTSFW can reduce the time by at most 15.67% and at least 8.23%. The average training time in ten different

TABLE 4. Prediction error comparison between ST-DNN [22] and proposed models.

Region	Model	Prediction Size	MAE	Improvement (%)	RMSE	Improvement (%)	Peak Error
TPE	ST-DNN/Ours	1	4.390/3.349	23.71	6.216/4.519	27.30	55.14/35.20
	ST-DNN/Ours	2	5.099/4.331	15.06	6.939/5.648	18.60	48.71/40.83
	ST-DNN/Ours	3	5.746/5.547	3.46	7.741/6.893	10.95	52.58/42.51
	ST-DNN/Ours	6	6.129/5.887	3.95	8.137/7.465	8.26	52.61/47.10
	ST-DNN/Ours	12	6.529/6.173	5.45	8.688/7.872	9.39	54.15/50.03
TCH	ST-DNN/Ours	1	5.720/4.360	23.78	8.114/5.770	28.89	63.59/37.64
	ST-DNN/Ours	2	7.762/5.790	25.41	9.909/8.012	19.14	54.00/43.36
	ST-DNN/Ours	3	8.293/7.005	15.53	10.757/9.482	11.85	51.56/46.18
	ST-DNN/Ours	6	8.977/7.212	19.66	11.403/9.669	15.21	65.87/52.01
	ST-DNN/Ours	12	9.280/7.540	18.75	11.667/10.549	9.58	66.51/49.27
TAN	ST-DNN/Ours	1	6.987/4.539	35.04	9.781/5.808	40.62	81.82/42.83
	ST-DNN/Ours	2	7.892/5.825	26.19	11.144/8.383	24.78	68.65/67.99
	ST-DNN/Ours	3	8.977/7.072	21.22	12.512/10.249	18.09	72.79/77.14
	ST-DNN/Ours	6	9.026/7.289	19.24	12.643/10.260	18.85	80.05/82.49
	ST-DNN/Ours	12	9.505/7.621	19.82	13.223/10.685	19.19	81.04/82.25
TAT	ST-DNN/Ours	1	3.141/1.103	64.88	3.675/1.575	57.14	15.82/8.56
	ST-DNN/Ours	2	3.194/1.842	42.33	4.057/2.287	43.63	16.39/10.36
	ST-DNN/Ours	3	4.007/2.494	37.76	4.967/3.427	31.00	18.53/15.84
	ST-DNN/Ours	6	4.094/2.664	34.93	5.230/3.656	30.10	19.48/16.80
	ST-DNN/Ours	12	4.516/2.815	37.67	5.652/3.881	31.33	20.48/17.99

Note: prediction sizes are the following one, two, three, six, and twelve day(s); improvement denotes the reduction ratio of the prediction error; the measuring unit of MAE, RMSE, and peak error is $\mu g/m^3$; peak errors indicate the differences between the actual peak value and the two predicted peak values.

time steps indicates that ST-DNN takes 20.57 seconds and HTSFW takes 18.11 seconds. As a result, HTSFW can decrease the ratio of the average training time by 11.96%.

Fig. 8 shows the geographical map of all the cities/counties in Taiwan. The upper part of the map in green is the northern region, the middle part in blue is the central region, the lower area in yellow is the southern region, and the right side of Taiwan in red is the eastern region. To predict the $PM_{2.5}$ concentration of the specified regions, we apply the developed Algorithm 1 to select the target districts. This work selects one city/county from the four regions mentioned previously indicated by the red arrow signs on the map. The four target districts are Taipei City (TPE), Taichung City (TCH), Tainan City (TAN) and Taitung County (TAT). Our Algorithm 1 can define arbitrary target district identifiers, although we select the four cities/counties in consideration of their representativeness of the four regions in Taiwan. In addition, the four selected target districts possess different weather conditions and climate circumstances.

Table 4 represents the experimental comparison of prediction errors between ST-DNN and HTSFW. We train the two models in five different time steps and compare the forecasting errors of the four target districts. In the testing data of the four cities/counties, the highest maximum $PM_{2.5}$ concentration value of 74 is in TAN, while the lowest maximum value of 23 is in TAT. For the $PM_{2.5}$ forecasting one day later, the HTSFW model can decrease the ratio of MAE by 35.04% and 64.88% in TAN and TAT, respectively; additionally, HTSFW reduces the scale of RMSE by 40.62% and 57.14%.

With respect to the predicted size of 12, the HTSFW model decreases the percentage of MAE by 5.45%, 18.75%, 19.82% and 37.67% in TPE, TCH, TAN and TAT, respectively, while HTSFW can reduce the percentage of RMSE by 9.39%, 9.58%, 19.19% and 31.33%, respectively. According to the prediction comparison in the five different time steps, the HTSFW model is conspicuously better than the ST-DNN model. From the $PM_{2.5}$ forecasting of the nationwide coverage and the regional districts, HTSFW performs well in the time series with air quality data prediction.

In Table 4, the experimental results of the HTSFW model show that the forecasting errors of the four regions in five prediction sizes decrease. Since ST-DNN is an adaptive deep learning model, it is composed of traditional machine learning and classic deep learning models. Our hybrid model is further compared to the ST-DNN model by the peak error values. The peak error indicates that the peak predicted $PM_{2.5}$ values of the two models are different from the observed $PM_{2.5}$ values. Based on the experimental results, HTSFW possesses lower peak errors than ST-DNN in the TPE, TCH and TAT regions of five predicted sizes. For the TAN region, HTSFW has lower peak errors than ST-DNN of predict sizes

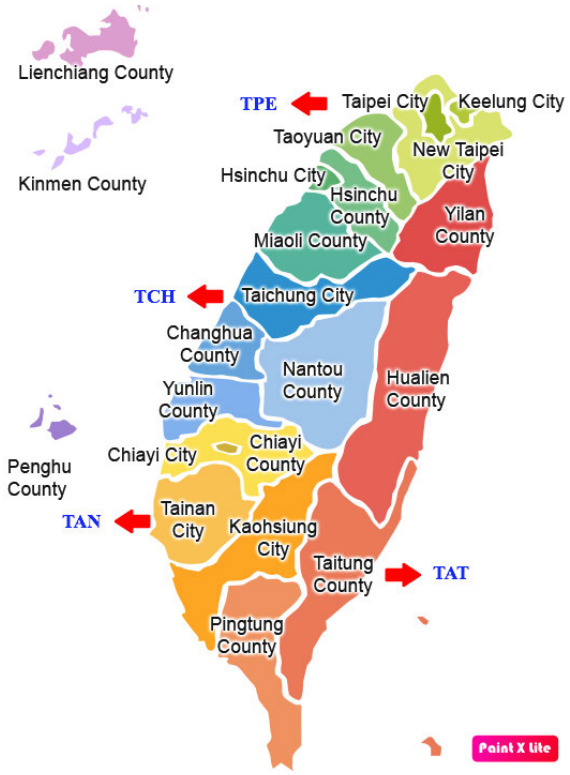

FIGURE 8. Regional city/county for $PM_{2.5}$ forecasting in Taiwan. Note: https://eego.epa.gov.tw/english/tour1/index1.asp? Parser = 99,10,27,,,,,1.

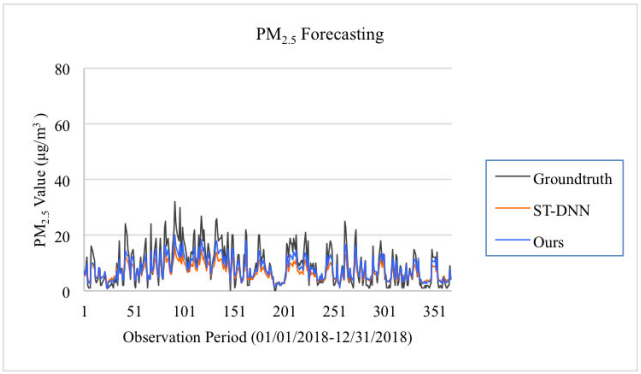

FIGURE 9. $PM_{2.5}$ forecasting comparison between two models (ST-DNN [22]) at station Yangming in Taipei City.

one and two. For a single time step in particular, HTSFW has almost half the peak error values of those of ST-DNN. For prediction sizes three, six and twelve, HTSFW has slightly higher peak error values than ST-DNN.

Fig. 9 depicts the particular monitoring station Yangming in Taipei City; the station Yangming is located in the mountain area. The maximum $PM_{2.5}$ concentration of the testing data of station Yangming is 32 in spring. The black trend represents the ground truth, the red line indicates the $PM_{2.5}$ prediction of ST-DNN, and the blue line represents the $PM_{2.5}$ prediction of HTSFW. From the plotted depiction, the one-year prediction of HTSFW is closer to the ground truth than that of ST-DNN.

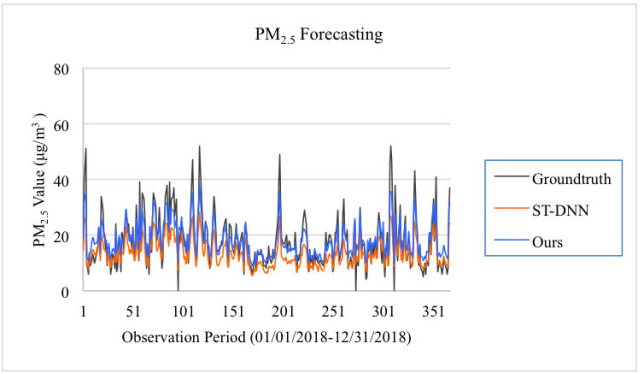

FIGURE 10. $PM_{2.5}$ forecasting comparison between two models (ST-DNN [22]) at the Fengyuan station in Taichung City.

Fig. 10 shows the $PM_{2.5}$ forecasting of station Fengyuan, which is located in the urban area of Taichung city. The maximum value of the $PM_{2.5}$ concentration of the testing data at the Fengyuan station is 52 in the spring. According to the prediction results, the ST-DNN model [22] is worse than the HTSFW model, especially the peak points of the trend, as shown in the forecasting figure.

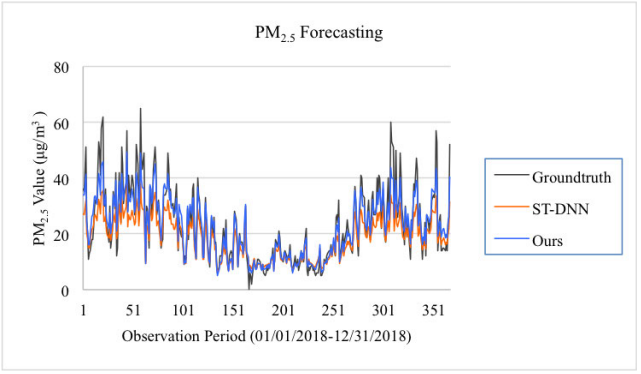

FIGURE 11. $PM_{2.5}$ forecasting comparison between two models (ST-DNN [22]) at station Tainan in Tainan City.

In Fig. 11, the figure indicates the $PM_{2.5}$ prediction of the Tainan station located in an urban area in Tainan City. The maximum value of the $PM_{2.5}$ concentration of the testing data at the Tainan station is 65 in the spring. From the plotted representation, the forecasting trend of the HTSFW model is quite close to the ground truth. In contrast, the $PM_{2.5}$ prediction of ST-DNN is close to the ground truth when the values are below 20. However, the prediction of the ST-DNN model [22] is evidently different from the ground truth when the values are above 20.

In Fig. 12, the figure represents the $PM_{2.5}$ forecasting of the Guanshan station, which is located in a rural area in Taitung County. In eastern Taiwan, there is a sparse population, and most of the people in the east earn a living by farming, particularly Taitung. The maximum $PM_{2.5}$ concentration of the testing data at station Guanshan is 21 in autumn. Due to lower pollution from industry and vehicular traffic, the annual

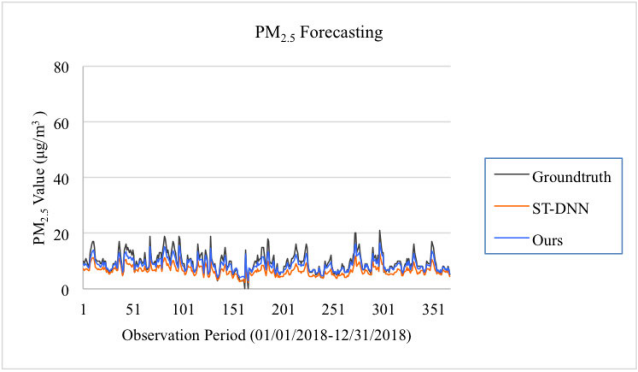

FIGURE 12. $PM_{2.5}$ forecasting comparison between two models (ST-DNN [22]) at the Guanshan station in Taitung County.

$PM_{2.5}$ concentrations in Taitung are mostly below 20. Despite the fact that both the forecasting trends of ST-DNN and HTSFW have no obvious difference from that of the ground truth, the figure shows that the prediction results of HTSFW are better than those of ST-DNN.

In this work, to evaluate the proposed model and the developed Algorithm 1, we first compare the baseline models with the proposed model by using the air quality data from the 76 monitoring stations in Taiwan. Furthermore, we compare the training time and the prediction errors of the existing ST-DNN model [22] with those of the proposed HTSFW model. For the comparison of the ST-DNN and HTSFW models, we compare their prediction errors of the four specified target districts and the sites belonging to each district. Moreover, we compare the prediction errors of selected particular sites belonging to each of the four target districts. According to the experimental results, the HTSFW model outperforms other models in different coverage regions and prediction periods.

Based on the experimental results, the innovation of the HTSFW model is expressed as follows. First, for air quality and geographical data training, the HTSFW model adopts the dilated CNN for feature extraction. Dilation factors are used to expand the receptive fields that lead to increased training efficiency. In addition, HTSFW applies GRU to time-series feature extraction for air quality and meteorological data learning. As shown in Table 3, the learning time decreases significantly because the GRU network is simpler than the LSTM network. Second, we develop an innovative algorithm to select the target district for $PM_{2.5}$ forecasting in Algorithm 1. By using the proposed Algorithm 1, the proposed HTSFW model can accurately predict the $PM_{2.5}$ concentration of a particular site as well as the specified regions. In Table 4 and Figs. 9-12, HTSFW can accurately predict the $PM_{2.5}$ concentration of the four cities/counties in northern, central, southern and eastern Taiwan. Furthermore, for the selected monitoring station in each of the four cities/counties, the $PM_{2.5}$ prediction of the HTSFW model is obviously better than that of the ST-DNN model [22].

V. CONCLUSION

In this paper, we propose a hybrid time-series deep learning model for daily-based $PM_{2.5}$ forecasting. The accuracy of the $PM_{2.5}$ prediction is enhanced by considering the air quality data and the meteorological data simultaneously. In addition, the $PM_{2.5}$ concentration is also related to geographical location and time frame. For the performance comparison, we first compare the MAE and RMSE of $PM_{2.5}$ prediction between the baseline and the proposed models. For ten different time steps, our model is superior to the baseline models. Moreover, to predict the air quality of the four city-/county-regions, the proposed model is further compared to the ST-DNN model [22]. The selected regions are located in northern, central, southern, and eastern Taiwan. Furthermore, we select one monitoring station in each region for the accuracy comparison. From the experimental results, the proposed model is accurate for all the regions and suitable for local regions and specified sites.

In our work, the proposed HTSFW model can accurately predict the $PM_{2.5}$ concentration of the specified monitoring station or regional district in Taiwan. In addition, in this COVID era and post-COVID period, both the administration and citizens have to avoid symptomatic infection and spread of the epidemic. For instance, timely announcements of the setup of screening stations and the confirmed case locations are both critical factors. The government needs to effectively manage and predict the COVID-19 vaccine distribution; hence, the forecasting models need to consider the spatial correlation, temporal dependency, air quality, and traffic transportation simultaneously. In other words, the forecasting models have to effectively extract the data features from variant sources and accurately predict the COVID-19 vaccine distribution in time.

In this work, we develop a hybrid deep learning model to increase the accuracy of $PM_{2.5}$ forecasting. Such hybrid models are based on the combination of multiple machine learning and/or deep learning models. The adaptive ST-DNN model is proposed to predict the $PM_{2.5}$ concentration of a single monitoring station, which is combined with machine learning and deep learning models. Our HTSFW model is composed of multiple deep learning models that enhance the prediction accuracy of the existing models. Since 2014, GANs have been widely applied to time-series data such as natural language processing [15], [17]. Air quality prediction is highly related to time-series data, and therefore, we will consider utilizing GAN for our future work.

GAN was designed in 2014 and is one of the deep learning models [15], [17]. In this model, two neural networks compete with each other based on zero-sum game theory. The generative network generates new data through unsupervised learning, and the discriminative network evaluates the new data through a competitive process. It has been widely used for image, video, and natural language processing. Inspired by the sequential data of GAN applications, we will seriously consider including GANs in future work.

For future work, we will further study the heights of each monitoring station for $PM_{2.5}$ prediction. In addition, we will also develop a new $PM_{2.5}$ forecasting model to deal with seasonal climate change to enhance the air quality prediction accuracy.

REFERENCES

[1] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov, "Real-time air quality forecasting, Part I: History, techniques, and current status," *Atmos. Environ.*, vol. 60, pp. 632–655, Dec. 2012, doi: 10.1016/j.atmosenv.2012.06.031.

[2] Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, and A. Baklanov, "Real-time air quality forecasting, Part II: State of the science, current research needs, and future prospects," *Atmos. Environ.*, vol. 60, pp. 656–676, Dec. 2012, doi: 10.1016/j.atmosenv.2012.02.041.

[3] S. Vardoulakis, B. E. A. Fisher, K. Pericleous, and N. Gonzalez-Flesca, "Modelling air quality in street canyons: A review," *Atmos. Environ.*, vol. 37, no. 2, pp. 155–182, Jan. 2003, doi: 10.1016/S1352-2310(02)00857-9.

[4] M. Dong, D. Yang, Y. Kuang, D. He, S. Erdal, and D. Kenski, "$PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining," *Expert Syst. Appl.*, vol. 36, no. 5, pp. 9046–9055, Jul. 2009, doi: 10.1016/j.eswa.2008.12.017.

[5] A. Donnelly, B. Misstear, and B. Broderick, "Real time air quality forecasting using integrated parametric and non-parametric regression techniques," *Atmos. Environ.*, vol. 103, pp. 53–65, Feb. 2015, doi: 10.1016/j.atmosenv.2014.12.011.

[6] S. Duan, W. Song, E. Zio, C. Cattani, and M. Li, "Product technical life prediction based on multi-modes and fractional Lévy stable motion," *Mech. Syst. Signal Process.*, vol. 161, Dec. 2021, Art. no. 107974, doi: 10.1016/j.ymssp.2021.107974.

[7] H. Liu, W. Song, Y. Niu, and E. Zio, "A generalized Cauchy method for remaining useful life prediction of wind turbine gearboxes," *Mech. Syst. Signal Process.*, vol. 153, May 2021, Art. no. 107471, doi: 10.1016/j.ymssp.2020.107471.

[8] H. Liu, W. Song, and E. Zio, "Metabolism and difference iterative forecasting model based on long-range dependent and grey for gearbox reliability," *ISA Trans.*, May 2021, doi: 10.1016/j.isatra.2021.05.002.

[9] H. Liu, W. Song, Y. Zhang, and A. Kudreyko, "Generalized Cauchy degradation model with long-range dependence and maximum Lyapunov exponent for remaining useful life," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, Mar. 2021, doi: 10.1109/TIM.2021.3063749.

[10] H. Liu, W. Song, and E. Zio, "Generalized Cauchy difference iterative forecasting model for wind speed based on fractal time series," *Nonlinear Dyn.*, vol. 103, no. 1, pp. 759–773, Jan. 2021, doi: 10.1007/s11071-020-06150-z.

[11] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: When urban air quality inference meets big data," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, Aug. 2013, pp. 1436–1444.

[12] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, Aug. 2015, pp. 437–446.

[13] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[14] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078

[15] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, New York, NY, USA, 2012, pp. 1097–1105.

[17] L. Gonog and Y. Zhou, "A review: Generative adversarial networks," in *Proc. 14th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Xi'an, China, Jun. 2019, pp. 505–510.

[18] Z. Qi, T. Wang, G. Song, W. Hu, X. Li, and Z. Zhang, "Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 12, pp. 2285–2297, Dec. 2018, doi: 10.1109/TKDE.2018.2823740.

[19] B. Eravci and H. Ferhatosmanoglu, "Diverse relevance feedback for time series with autoencoder based summarizations," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 12, pp. 2298–2311, Dec. 2018, doi: 10.1109/TKDE.2018.2820119.

[20] J. Wang and G. Song, "A deep spatial-temporal ensemble model for air quality prediction," *Neurocomputing*, vol. 314, pp. 198–206, Nov. 2018, doi: 10.1016/j.neucom.2018.06.049.

[21] D. E. H. Zhuang, G. C. L. Li, and A. K. C. Wong, "Discovery of temporal associations in multivariate time series," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 2969–2982, Dec. 2014, doi: 10.1109/TKDE.2014.2310219.

[22] P.-W. Soh, J.-W. Chang, and J.-W. Huang, "Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations," *IEEE Access*, vol. 6, pp. 38186–38199, Jun. 2018, doi: 10.1109/ACCESS.2018.2849820.

[23] Q. Zhou, H. Jiang, J. Wang, and J. Zhou, "A hybrid model for $PM_{2.5}$ forecasting based on ensemble empirical mode decomposition and a general regression neural network," *Sci. Total Environ.*, vol. 496, pp. 264–274, Oct. 2014, doi: 10.1016/j.scitotenv.2014.07.051.

[24] L. A. Díaz-Robles, J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson, and J. A. Moncada-Herrera, "A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile," *Atmos. Environ.*, vol. 42, no. 35, pp. 8331–8340, Nov. 2008, doi: 10.1016/j.atmosenv.2008.07.020.

[25] X. Zhou, W. Huang, N. Zhang, W. Hu, S. Du, G. Song, and K. Xie, "Probabilistic dynamic causal model for temporal data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Killarney, Ireland, Jul. 2015, pp. 1–8.

[26] S. Deleawe, J. Kusznir, B. Lamb, and D. J. Cook, "Predicting air quality in smart environments," *J. Ambient Intell. Smart Environ.*, vol. 2, no. 2, pp. 145–152, Jan. 2010, doi: 10.3233/ais-2010-0061.

[27] Z. Yu, C. Licia, W. Ouri, and Y. Hai, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Sep. 2014, doi: 10.1145/2629592.

[28] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep distributed fusion network for air quality prediction," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, Jul. 2018, pp. 965–973.

[29] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, Aug. 2015, pp. 2267–2276.

[30] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognit. Lett.*, vol. 42, pp. 11–24, Jun. 2014, doi: 10.1016/j.patrec.2014.01.008.

[31] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, New York, NY, USA, 2015, pp. 3995–4001.

[32] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environ. Sci. Pollut. Res.*, vol. 23, no. 22, pp. 22408–22417, Nov. 2016, doi: 10.1007/s11356-016-7812-9.

[33] 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, doi: 10.1007/s00521-015-1955-3.

[34] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, New York, NY, USA, 2017, pp. 1655–1661.

[35] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proc. 23rd ACM Int. Conf. Multimedia (ACM MM)*, New York, NY, USA, 2015, pp. 461–470.

[36] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 1489–1496, Dec. 2013, doi: 10.1109/TPAMI.2015.2505293.

[37] S. Du, T. Li, Y. Yang, and S.-J. Horng, "Deep air quality forecasting using hybrid deep learning framework," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2412–2424, Jun. 2021, doi: 10.1109/TKDE.2019.2954510.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[39] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Artificial Neural Networks: Formal Models and Their Applications*. Berlin, Germany: Springer, 2005, pp. 799–804.
[40] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5, pp. 602–610, Jul. 2005, doi: 10.1016/j.neunet.2005.06.042.
[41] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, 2016, pp. 1–13.
[42] Q. Zhang, J. C. K. Lam, V. O. K. Li, and Y. Han, "Deep-AIR: A hybrid CNN-LSTM framework for fine-grained air pollution forecast," 2020, *arXiv:2001.11957*. [Online]. Available: http://arxiv.org/abs/2001.11957
[43] R. Chakraborty, X. Zhen, N. Vogt, B. Bendlin, and V. Singh, "Dilated convolutional neural networks for sequential manifold-valued data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct./Nov. 2019, pp. 10620–10630, doi: 10.1109/iccv.2019.01072.

PEI-WEN CHIANG (Member, IEEE) received the B.S. degree in information management from the National Yunlin University of Science and Technology, in 1997, and the M.S. degree in management information systems from the National Chengchi University, in 1999. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology. His research interests include machine learning, deep learning, cloud computing, information security, the IoT, and 5G networks.

SHI-JINN HORNG received the B.S. degree in electronics engineering from the National Taiwan Institute of Technology, in 1980, the M.S. degree in information engineering from the National Central University, in 1984, and the Ph.D. degree in computer science from the National Tsing Hua University, in 1989. He is currently a Chair Professor with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology. He has published more than 200 research articles. His research interests include deep learning, biometric recognition, image processing, and information security. He received many awards. In particular, the Distinguished Research Award from the National Science Council, Taiwan, in 2004.