IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received April 18, 2021, accepted April 30, 2021, date of publication May 4, 2021, date of current version May 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077574

PM2.5 Forecasting Model Using a Combination of Deep Learning and Statistical Feature Selection

ENDAH KRISTIANI[1,2], TING-YU KUO[3], CHAO-TUNG YANG[4,5,6], (Member, IEEE), KAI-CHIH PAI[7], CHIN-YIN HUANG[1], AND KIEU LAN PHUONG NGUYEN[8]

[1]Department of Industrial Engineering and Enterprise Information, Tunghai University, Taichung 407224, Taiwan
[2]Department of Informatics, Krida Wacana Christian University, Jakarta 11470, Indonesia
[3]Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 621301, Taiwan
[4]Department of Computer Science, Tunghai University, Taichung 407224, Taiwan
[5]Research Center for Smart Sustainable Circular Economy, Tunghai University, Taichung 407224, Taiwan
[6]Research Center for Nanotechnology, Tunghai University, Taichung 407224, Taiwan
[7]College of Engineering, Tunghai University, Taichung 407224, Taiwan
[8]Faculty of Environmental and Food Engineering, Nguyen Tat Thanh University, Ho Chi Minh City 70000, Vietnam

Corresponding author: Chao-Tung Yang (ctyang@thu.edu.tw)

This work was funded in part by the Ministry of Science and Technology (MOST), Taiwan, under Grant Number 109-2221-E-029-020 and 109-2625-M-029-001-. In addition, this work was also funded in part by The National Applied Research Laboratories (NARLabs), Taiwan, under Grant No. 03108F1106 and 03109F1106. We are grateful to the National Center for High-performance Computing for computer time and facilities.

ABSTRACT This paper proposed a PM 2.5 forecasting model using Long Short-Term Model (LSTM) sequence to sequence combined with the statistical method. Correlation Analysis, XGBoost, and Chemical processed are used as the methods to select the essential features. The air pollution data is extracted from Taiwan Environmental Protection Agency (EPA) for the Taichung City dataset in 2014–2018. The study points out that chemical processed model of particulate matter 10 micrometers or less in diameter (PM10), Sulfur Dioxide (SO2), and Nitrogen Dioxide (NO2) have the highest accuracy or lowest Root Mean Square Error (RMSE) and more short training and testing time among the other models. The chemical processed model of PM10, SO2, and NO2 (model B) has the highest accuracy (lowest RMSE), approximately 1 point lower RMSE values, and the shortest training and testing period among the other models. Furthermore, RMSE calculations based on the stations reveal that training with the entire station dataset has a 3 point higher RMSE value than training with each station dataset.

INDEX TERMS Air pollution monitoring, LSTM seq2seq, PM2.5, XGBoost, feature selection, correlation analysis, deep learning.

I. INTRODUCTION

Internet of Things (IoT) is an interconnection of various instruments, networks, techniques, and human resources for a common purpose. Various IoT-based apps are used in different industries and have been able to offer enormous advantages to users. The information produced from IoT devices is valuable only when the data is analyzed and presented in the graph, map, or table diagram [31]. Data analytics is used to examine large and small information sets with different data characteristics to draw significant findings and practical insights. These findings are generally in the form of trends, patterns, and statistics that support businesses in the proactive use of information for efficient decision-making.

Sequence prediction often includes predicting the following value for an input sequence in a real-value sequence or producing a class label. It is commonly referred to as a one-to-one or many-to-one sequence forecast problem. A more complicated form of sequence prediction problem takes a sequence as input and requires a sequence prediction as output [30]. These are referred to as sequence-to-sequence prediction (seq2seq) problems. The fact that the input and output sequences can vary in length poses a challenge. This problem is known as a sequence-to-many type prediction problem since there are many input time steps and many output time steps.

Feature selection is one of the key ideas in machine learning that greatly influence model efficiency. The feature characteristics used for training the data have a significant impact on the accuracy of the model [24], [26]. Irrelevant features or partially relevant ones can adversely affect the efficiency of the model. Feature selection should be one of model designing's most significant steps. Irrelevant features will decrease the accuracy of the model, also lead to false learning of the model. The feature selection functions have three advantages. First, it reduces over-fitting; having less

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

redundant information implies fewer noise-based choices. Second, it improves the accuracy by improving the precision of modeling implies less inaccurate information. Third, it reduces training time; fewer information points decrease the algorithms' complexity and more quickly train algorithms [3], [9], [23].

Predicting PM2.5 has been paid more attention to many scientists. A review study by Bai *et al.* [10] summarizes different models to forecast air pollutants (including PM2.5), such as statistical methods, artificial intelligence methods, and hybrid three-dimensional models, and other methods. Their statistical models demand less time to build models but depend on the data time series approach. Artificial Intelligence (AI) methods work well with nonlinear data; however, they are unstable and high dependent on data. Hybrid methods have good robustness, low risk, and strong adaptability [2], [6], [7], [24].

The PM2.5 forecasting model is proposed in this paper using the sequence-to-sequence Long Short-Term Model (LSTM) combined with the statistical method. Analysis of correlation, XGBoost, and processed chemistry is used to select the essential features. Air pollution data is collected in 2014-2018 from the Taiwan Environmental Protection Agency (EPA) for the Taichung City dataset. Accordingly, this paper is organized as follows. Section II describes the background review used to examine theoretical foundations, experimental theories, and prior works. Section III presented the methods applied in the paper. Section IV is the experimental results of our work. Finally, we summarized this research in Section V.

II. BACKGROUND REVIEW

PM2.5 in the atmosphere is derived from a primary and secondary source of precursors. Direct emissions are related to naturals, such as volcanoes, dust storms, forest fires, and anthropogenic, like fossil fuels' burning. Secondary emissions come from chemical reactions occurring in the atmosphere. Secondary particulates dominate most ambient monitoring stations of PM2.5 with the contribution of ammonium, sulfate, and nitrate, which are substances resulting from SOx and NOx emission correspondingly [16]. For instance, a study on PM2.5 composition in urban and rural areas in some cities in the United States, Canada, and Mexico reveals more than 77%, on average, generated from secondary sources [17]. Hence, there is a high correlation between PM2.5 and SOx, NOx in the ambient air. However, the relationship between ozone and PM2.5 is complicated and change in time and space and throughout the day. Zhao *et al.* [28], showed the trends of PM2.5 and O3 between 2015 and 2019 over 367 cities in China. As a result, the increase of O3 and decrease of PM2.5 concentrations simultaneously happen to range from 47.2% (spring) to 74.9% (summer) of the studied Chinese cities. In the time of the COVID-19 crisis in Baghdad, Iraq, for instance, NO2 and PM2.5 together decreased while O3 increased when comparing these factors before lockdown within partial and total lockdown [29]. Therefore, NOx, SOx, and O3 and temperature are critical factors affecting PM2.5 concentration in the atmosphere.

Neural networks like Long Short-Term Memory (LSTM) can handle model problems almost seamlessly with different input factors. It is a significant benefit in time series forecasting where traditional linear methods can be challenging to fit into multivariate or multiple problems of input forecasting. An LSTM auto-encoder is implemented to sequence time-series data using an Encoder-Decoder LSTM architecture. Once aligned, the sample encoder part can then be used to encode or compress sequence data for use as a feature vector input in a supervised learning model [5], [8]. Liu *et al.* [1] explained the slow pace of the seq2seq training to replace the first RNN encoder with a fully connected encoder to accelerate the training process. They also introduced position embedding to detect sequential relationships in the fully connected encoder between source sequences. The accumulation of mistakes generated by the recurrent prediction is another element. The n-step recurrent forecast has been suggested to solve this issue. Their experimental results verified that the AAQP with n-step recurrent forecasting had excellent performance since the accumulation of error was decreased, and when compared to the initial seq2seq attention model, the training time was substantially reduced. Viswanath *et al.* [4] have suggested that LSTM seq2seq models depend on deep learning to categorize monsoon days that are finally assembled to detect spells. Dry and wet days are classified with 95% and 87% of accuracies, respectively. It is observed that the prediction of break spells is more accurate than the active spells. The seq2seq model has also been shown to be more effective than the long-term memory model. They also perform better than typical monsoon spells at detecting classification models.

Luo *et al.* [20] developed and implemented a high-precision real-time PM2.5 forecasting system in Taiwan. Their paper suggests a predictive method called the Adaptive Iteration Forecast (AIF) that can forecast the value of PM2.5 for the next couple of hours based on historical data patterns. They have shown through various comparative studies that their model can generate significant results. A gradient-boosting-based machine learning method was proposed by Lin *et al.* [21] and Lee *et al.* [22]. The proposed mechanism is tested using Taiwan's EPA and Central Weather Bureau (CWB), which contains data from 77 stations of air monitoring and 580 weather stations that took hourly assessments for a year. According to their findings, the most notable increase in predictive efficiency was found in central Taiwan. They also compared the performance of the prediction model in Taiwan, Taipei, and London. Since Taipei and London have similar topography (basin), the findings show that these two cities have similar prediction results.

III. METHODS

A. CORRELATION ANALYSIS

PM2.5 might be classified as main or secondary precursors in the atmosphere. Primary PM2.5 is formed directly by anthropogenic and natural pollutants, while secondary PM2.5 is emitted due to chemical reactions in the atmosphere. The existence of primary PM2.5 and suitable gaseous precursors affects secondary aerosols. Secondary PM2.5 formation is significantly linked to SO2, NO2 precursor gases for PM2.5. This connection is demonstrated in the 2005-2015 Taiwan Air Quality Studies of Lee *et al.* [14]. Besides, in 31 Chinese cities between 2013 and 2014, Xie *et al.* [15] in 2015 asserted a medium to elevate the relationship of PM2.5 with SO2, NO2 accumulation in 286 surveillance locations. In the attempt to evaluate secondary of the forming of PM2.5 for the air spreading model, the US EPA also proposed using SO2, NO2/NOx.

B. XGBoost

XGBoost is a highly effective, versatile, and portable distributed gradient boosting library [13], [27]. Under the Gradient Boosting structure, it uses machine learning algorithms. XGBoost offers a parallel boost to the tree (also known as GBDT, GBM), which quickly and accurately solves many data science issues. The same code operates on a significant distributed setting (Hadoop, SGE, MPI) and can solve issues beyond billions of examples. The advantage of using decision-tree methods such as gradient boost is that they can automatically provide feature-scale estimates from a trained predictive model.

C. LSTM SEQUENCE TO SEQUENCE

The LSTM encoder-decoder [18] is a recurrent neural network that is intended to deal with sequence to sequence issues, often known as seq2seq. Prediction issues from sequence to sequence are challenging since the input and output sequence items can differ. Examples of seq2seq issues are text translation and learning to run programs. In particular, input sequences and output sequences have distinct lengths (for example, machine translation), and it requires the entire input sequence to begin predicting the target. They need a more sophisticated configuration, which is frequently referred to by individuals when they mention "sequence to sequence models" without any other context. LSTM sequence to sequence operates as following [3]:

1) The RNN layer (or stack of it) serves as an "encoder" to process the input sequence. Note that we discard the RNN encoder output only when the status is recovered. In the next stage, this state serves as the decoder's "context". The equation of the encoder is as follows:

$$h\_t = \int (W^{(hh)} h\_(t-1) + W^{(hx)} x\_t) \tag{1}$$

2) Another RNN layer (or stack) functions as a "decoder": it gives prior characters of the destination sequence, and it is trained to predict the next target sequence characters. In particular, the training process is trained to convert the target sequences into the same sequences but will be offset in a single step in the future. The encoder utilizes the encoder's vectors as its original state; how the decoder gets data is about what it should produce. Indeed, the decoder learns to generate targets [t + 1...], depending on the input sequence, given targets [...t]. The equation of the decoder is as follows:

$$h\_t = \int (W^{(hh)} h\_(t-1)) \tag{2}$$

D. ROOT MEAN SQUARED ERROR (RMSE)

The RMSE is the standard deviation of residuals (prediction errors). The residuals are an indicator of how far these data points are from the regression line. As well as the RMSE is an estimate of how far these residuals are. In other words, it shows us how close the real-values are to the best-fit axis. RMSE is commonly used to verify experimental results in climatology, prediction, and regression analysis. The following is a description of the equation:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)} \tag{3}$$

where n is the sample size, y_i is the actual expected output, and $\hat{y}_i$ is the model's prediction.

E. MACHINE LEARNING SYSTEM FLOW

A system workflow was designed to perform model training. There are total 30 training set were conducted in this experiment based on five categories of feature selections and six regions of all stations, Chungming, Dali, Fengyuan, Shalu, and Xitun station. In each training experiment, first, when the missing data is found, the preprocessing phase is necessary to prepare the data for the mode of training. The information Not Available (NA) was substituted by 0 in this stage. Then the dataset is divided into components of practice and testing. The training and testing were carried out after setting the parameters outlined in detail in the Table. Finally, RMSE calculations continued. Figure 1 shows the flow diagram of this project.

The model's training experiments are divided into five category models, as follows.

- Model A: Training using 17 parameters
- Model B: Training based on PM10, SO2, and NO2
- Model C: Training based on O3 and Ambient Temperature
- Model D: Training based on Correlation Analysis
- Model E: Training based on XGBoost feature selection

F. TRAINING NETWORK AND PARAMETERS

Table 1 shows the training and network parameters by sequencing 30 hours to forecast the next 2 hours using the Adam optimizer. The Adam optimization algorithm is a variant of stochastic gradient descent, which has recently gained

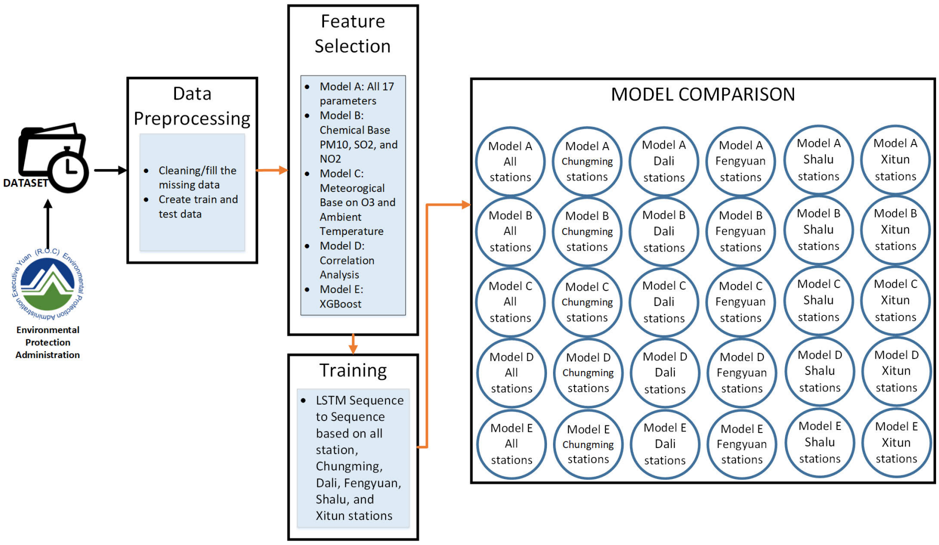

DATASET
Environmental Protection Administration
Executive Yuan (R.O.C) Environmental Protection Administration
Data Preprocessing
Cleaning/fill the missing data
Create train and test data
Feature Selection
Model A: All 17 parameters
Model B: Chemical Base PM10, SO2, and NO2
Model C: Meteorological Base on O3 and Ambient Temperature
Model D: Correlation Analysis
Model E: XGBoost
Training
LSTM Sequence to Sequence based on all station, Chungming, Dali, Fengyuan, Shalu, and Xitun stations
MODEL COMPARISON
Model A All stations
Model A Chungming stations
Model A Dali stations
Model A Fengyuan stations
Model A Shalu stations
Model A Xitun stations
Model B All stations
Model B Chungming stations
Model B Dali stations
Model B Fengyuan stations
Model B Shalu stations
Model B Xitun stations
Model C All stations
Model C Chungming stations
Model C Dali stations
Model C Fengyuan stations
Model C Shalu stations
Model C Xitun stations
Model D All stations
Model D Chungming stations
Model D Dali stations
Model D Fengyuan stations
Model D Shalu stations
Model D Xitun stations
Model E All stations
Model E Chungming stations
Model E Dali stations
Model E Fengyuan stations
Model E Shalu stations
Model E Xitun stations

FIGURE 1. System flow diagram.

TABLE 1. The training and network parameters.

Parameter	Value
Number of stacked LSTM layers	2
learning rate	0,01
lambda reg	0,003
Gradient Clipping	2,5
Size of LSTM cell	24
epoch	100
batch size	16
keep rate	0,5

TABLE 2. Dataset parameters.

No.	Parameter	Unit	Name
1	PM2.5	μg/m3	Fine aerosol
2	CH4	ppm	Methane
3	AMB_TEMP	Celsius	Ambient temperature
4	NMHC	ppm	Non-methane hydrocarbons
5	NO	ppb	Nitrogen monoxide
6	NOx	ppb	Nitrogen oxide
7	CO	ppm	Carbon monoxide
8	NO2	ppb	Nitrogen dioxide
9	O3	ppb	Ozone
10	PM10	μg/m3	Aerosol
11	RAINFALL	mm	ppb
12	RH	%	ppm
13	SO2	ppb	Sulfur dioxide
14	THC	ppm	Tetrahydrocannabinol
15	WIND_DIREC	degree	Wind direction hourly
16	WS_HR	m/s	Wind speed hourly
17	WIND_SPEED	m/s	Wind speed
18	WD_HR	degree	Wind direction hourly

popularity in deep learning applications for computer vision and natural language processing.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the experimental results are divided into four parts: the dataset, the selected parameters, the accuracy, the training and testing time, and the PM2.5 visualization.

A. DATASET

The dataset was acquired for 2014-2018 from the Taiwan EPA [11]. It comprises five stations in Taichung City, Taiwan, including Fengyuan, Chungming, Xitun, Shalu, and Dali. The characteristics of air monitoring consist of 18 parameters as outlined in Table 2.

B. SELECTED PARAMETERS

This study compares five models of PM2.5 prediction in terms of accuracy, training time, and testing time. The five models consist of (1) model A: using LSTM seq2seq for 17 parameters; (2) model B: using LSTM seq2seq for PM10, NO2, SO2; (3) model C: using LSTM seq2seq for PM10, O3, AMB TEMP; (4) model D: using LSTM seq2seq for top 5 parameters selected by correlation analysis; and (5) model E: using LSTM seq2seq for top 5 parameters selected by feature selection. For model D and model E, five parameters that have the most crucial role in predicting PM2.5 are presented. Model D is based on correlation analysis shown in a heat-map matrix in Fig. 2.

Model E is based on XGBoost Feature selection as described in Fig. 3

The five feature selections of Model D and Model E are listed in Table 3.

C. PREDICTION RESULTS

A plot diagram is used to visualize the comparison of real value and prediction. In this diagram, we can see how close the prediction against the real value. In the following graphs, we can see that the predictions are more close to the real

TABLE 3. Top 5 parameter selected by model D and model E.

Model	ALL Stations	Xitun	Fengyuan	Dali	Chungming	Shalu
Model D	1. PM10	1. PM10	1. PM10	1. PM10	1. PM10	1. PM10
	2. NO2	2. NO2	2. NO2	2. CO	2. CO	2. CO
	3. CO	3. CO	3. SO2	3. SO2	3. NO2	3. RAINFALL
	4. NOx	4. SO2	4. NOx	4. NMHC	4. NMHC	4. NO2
	5. SO2	5. NMHC	5. O3	5. NOx	5. NOx	5. NOx
Model E	1. PM10	1. PM10	1. PM10	1. PM10	1.PM10	1. PM10
	2. WIND_SPEED	2. O3	2. CO	2. O3	2. O3	2. WIND_SPEED
	3. WS_HR	3. CO	3. O3	3. SO2	3. CO	3. WS_HR
	4. O3	4. SO2	4. SO2	4. CO	4. SO2	4 NOx
	5. NOx	5.AMB_TEMP	5. RH	5.AMB_TEMP	5. WD_HR	5. RH

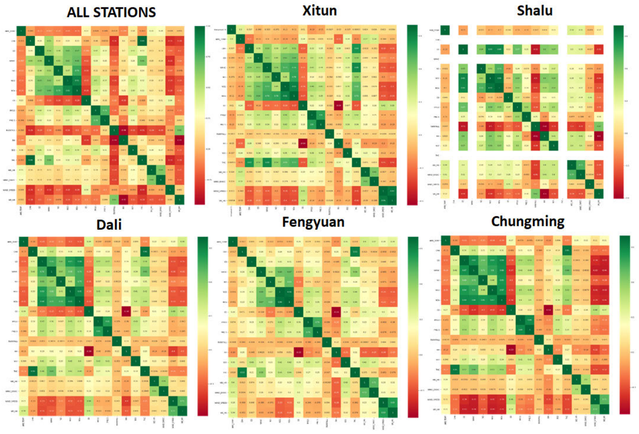


FIGURE 2. The correlation analysis matrix.

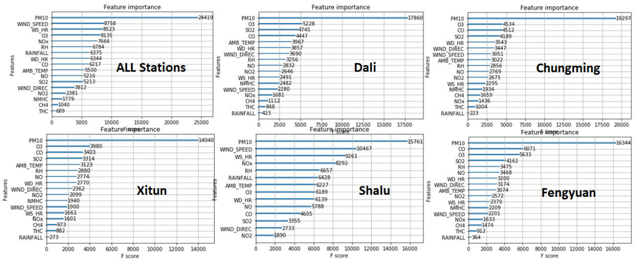


FIGURE 3. The XGBoost of 10 important features.

values. However, for Model E (XGBoost), the Chungming station has a very low accuracy among the other stations. Also, Model E is not good in Xitun, Dali, Shalu, and Fengyuan stations. Figure 4, 6, 8, 10, 12, and 14 describe the comparison of the real and the predicted values.

1) Figure 4 shows the plot diagram of PM2.5 real values and predictions based on all stations.
   Figure 5 presents the list of PM2.5 real values and predictions based on all stations.
2) Figure 6 shows the plot diagram of PM2.5 real values and predictions based on Chungming stations.
   Figure 7 presents the list of PM2.5 real values and predictions based on Chungming stations.
3) Figure 8 shows the plot diagram of PM2.5 real values and predictions based on Dali stations.
   Figure 9 presents the list of PM2.5 real values and predictions based on Dali stations.
4) Figure 10 shows the plot diagram of PM2.5 real values and predictions based on Fengyuan stations.
   Figure 11 presents the list of PM2.5 real values and predictions based on Fengyuan stations.
5) Figure 12 shows the plot diagram of PM2.5 real values and predictions based on Shalu stations.
   Figure 13 presents the list of PM2.5 real values and predictions based on Shalu stations.
6) Figure 14 shows the plot diagram of PM2.5 real values and predictions based on Xitun stations.
   Figure 15 presents the list of PM2.5 real values and predictions based on Xitun stations.

D. MODEL ACCURACY

RMSE is applied to evaluate the accuracy of these models. As shown in Fig. 16, model B and model D has the lowest value of RMSE. However, the value in model B seems to be homogeneous in different stations. In terms of training all stations, the RMSE values are high in all models. It means that all trained stations have not good performance compared to each station. Model A and B have a similar pattern in RMSE values, but model B is more good than model A. While in model C, the Shalu station has a weak accuracy than other models. In model D, the Dali station is the most inferior performance among other models. In model E, some stations like Xitun and Chungming have a lousy performance. In detail, the average RMSE value in training all stations for model A, B, C, D, and E have 12.177 RMSE values. Xitun stations have 9.941, Fengyuan stations reach 9.260, Dali stations are at 10.188, Chungming stations have 9.920, and Shalu stations have 10,412 RMSE rate values. From the point of each model, models A, B, C, and E have an average of 10.499, 9.065, 10.771, 10.280, and 10.969, respectively. It can be seen that model B has the lowest RMSE value that reflected the excellent model.

E. TRAINING TIME AND TESTING TIME

The training and testing time are displayed in Fig. 17 and Fig. 18. The training and testing time of model B with PM10,

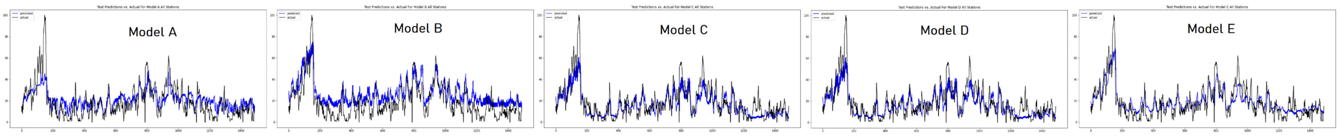
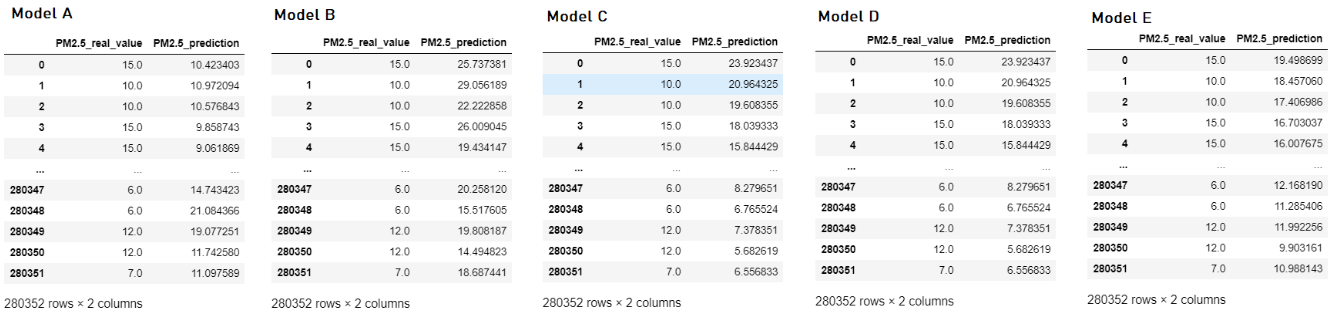


FIGURE 4. The plot diagram of PM2.5 real values and predictions based on all stations.



Model A

	PM2.5_real_value	PM2.5_prediction
0	15.0	10.423403
1	10.0	10.972094
2	10.0	10.576843
3	15.0	9.858743
4	15.0	9.061869
...	...	...
280347	6.0	14.743423
280348	6.0	21.084366
280349	12.0	19.077251
280350	12.0	11.742580
280351	7.0	11.097589

280352 rows × 2 columns

Model B

	PM2.5_real_value	PM2.5_prediction
0	15.0	25.737381
1	10.0	29.056189
2	10.0	22.222858
3	15.0	26.009045
4	15.0	19.434147
...	...	...
280347	6.0	20.258120
280348	6.0	15.517605
280349	12.0	19.808187
280350	12.0	14.494823
280351	7.0	18.687441

280352 rows × 2 columns

Model C

	PM2.5_real_value	PM2.5_prediction
0	15.0	23.923437
1	10.0	20.964325
2	10.0	19.608355
3	15.0	18.039333
4	15.0	15.844429
...	...	...
280347	6.0	8.279651
280348	6.0	6.765524
280349	12.0	7.378351
280350	12.0	5.682619
280351	7.0	6.556833

280352 rows × 2 columns

Model D

	PM2.5_real_value	PM2.5_prediction
0	15.0	23.923437
1	10.0	20.964325
2	10.0	19.608355
3	15.0	18.039333
4	15.0	15.844429
...	...	...
280347	6.0	8.279651
280348	6.0	6.765524
280349	12.0	7.378351
280350	12.0	5.682619
280351	7.0	6.556833

280352 rows × 2 columns

Model E

	PM2.5_real_value	PM2.5_prediction
0	15.0	19.498699
1	10.0	18.457060
2	10.0	17.406986
3	15.0	16.703037
4	15.0	16.007675
...	...	...
280347	6.0	12.168190
280348	6.0	11.285406
280349	12.0	11.992256
280350	12.0	9.903161
280351	7.0	10.988143

280352 rows × 2 columns

FIGURE 5. The plot diagram of PM2.5 real values and predictions based on all stations.

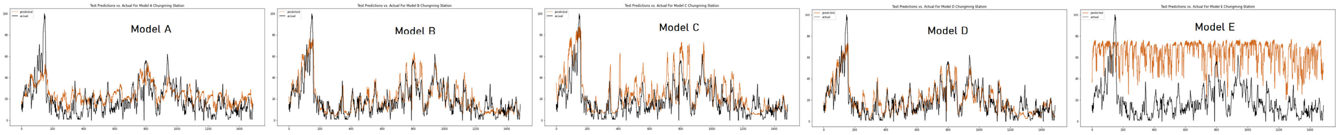
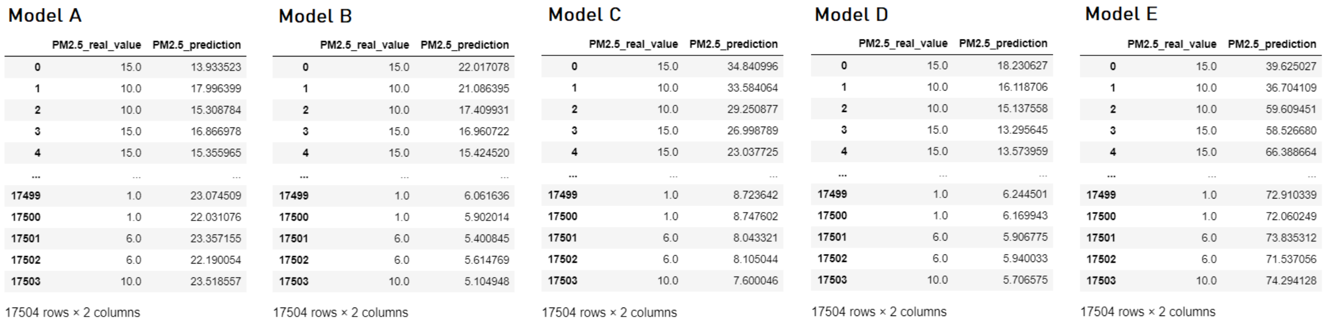


FIGURE 6. The plot diagram of PM2.5 real values and predictions based on Chungming stations.



Model A

	PM2.5_real_value	PM2.5_prediction
0	15.0	13.933523
1	10.0	17.996399
2	10.0	15.308784
3	15.0	16.866978
4	15.0	15.355965
...	...	...
17499	1.0	23.074509
17500	1.0	22.031076
17501	6.0	23.357155
17502	6.0	22.190054
17503	10.0	23.518557

17504 rows × 2 columns

Model B

	PM2.5_real_value	PM2.5_prediction
0	15.0	22.017078
1	10.0	21.086395
2	10.0	17.409931
3	15.0	16.960722
4	15.0	15.424520
...	...	...
17499	1.0	6.061636
17500	1.0	5.902014
17501	6.0	5.400845
17502	6.0	5.614769
17503	10.0	5.104948

17504 rows × 2 columns

Model C

	PM2.5_real_value	PM2.5_prediction
0	15.0	34.840996
1	10.0	33.584064
2	10.0	29.250877
3	15.0	26.998789
4	15.0	23.037725
...	...	...
17499	1.0	8.723642
17500	1.0	8.747602
17501	6.0	8.043321
17502	6.0	8.105044
17503	10.0	7.600046

17504 rows × 2 columns

Model D

	PM2.5_real_value	PM2.5_prediction
0	15.0	18.230627
1	10.0	16.118706
2	10.0	15.137558
3	15.0	13.295645
4	15.0	13.573959
...	...	...
17499	1.0	6.244501
17500	1.0	6.169943
17501	6.0	5.906775
17502	6.0	5.940033
17503	10.0	5.706575

17504 rows × 2 columns

Model E

	PM2.5_real_value	PM2.5_prediction
0	15.0	39.625027
1	10.0	36.704109
2	10.0	59.609451
3	15.0	58.526680
4	15.0	66.388664
...	...	...
17499	1.0	72.910339
17500	1.0	72.060249
17501	6.0	73.835312
17502	6.0	71.537056
17503	10.0	74.294128

17504 rows × 2 columns

FIGURE 7. The plot diagram of PM2.5 real values and predictions based on Chungming stations.

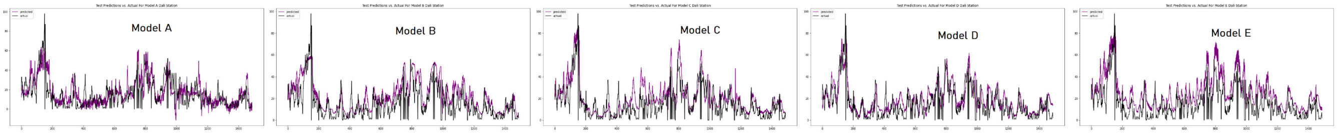
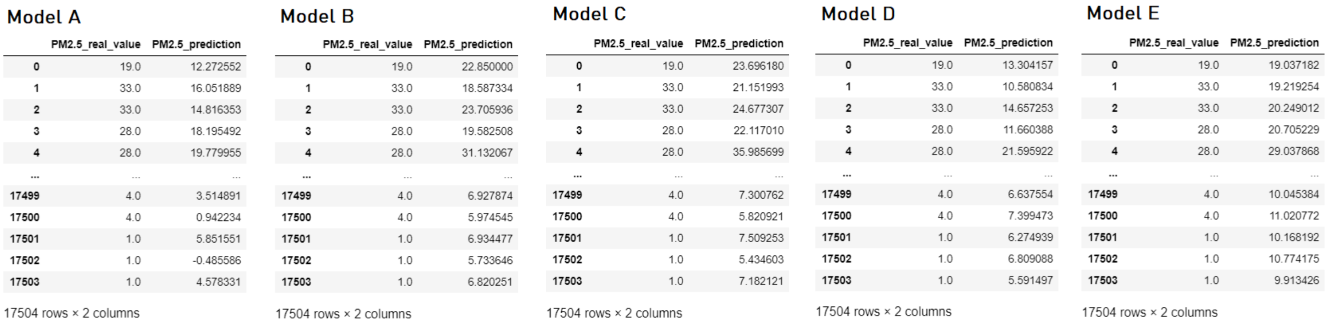


FIGURE 8. The plot diagram of PM2.5 real values and predictions based on Dali stations.

NO2, SO2 execute with the shortest value, about 15 seconds and 20 seconds, respectively. However, there is not much different time consumption between these models.

F. DISCUSSION

The proposed model implemented feature selection methods and analyzed the model between all stations compared to each station in machine learning implementation. The experiments show that a separation model centered on each station is superior to a model of unity in all stations. Furthermore, the use of the feature selection method has improved the model's accuracy and speed. The model performance was the most stable when secondary particulates of PM10, SO2, and NO2 (model B) were applied. As compared to other



Model A

	PM2.5_real_value	PM2.5_prediction
0	19.0	12.272552
1	33.0	16.051889
2	33.0	14.816353
3	28.0	18.195492
4	28.0	19.779955
...	...	...
17499	4.0	3.514891
17500	4.0	0.942234
17501	1.0	5.851551
17502	1.0	-0.485586
17503	1.0	4.578331

17504 rows × 2 columns

Model B

	PM2.5_real_value	PM2.5_prediction
0	19.0	22.850000
1	33.0	18.587334
2	33.0	23.705936
3	28.0	19.582508
4	28.0	31.132067
...	...	...
17499	4.0	6.927874
17500	4.0	5.974545
17501	1.0	6.934477
17502	1.0	5.733646
17503	1.0	6.820251

17504 rows × 2 columns

Model C

	PM2.5_real_value	PM2.5_prediction
0	19.0	23.696180
1	33.0	21.151993
2	33.0	24.677307
3	28.0	22.117010
4	28.0	35.985699
...	...	...
17499	4.0	7.300762
17500	4.0	5.820921
17501	1.0	7.509253
17502	1.0	5.434603
17503	1.0	7.182121

17504 rows × 2 columns

Model D

	PM2.5_real_value	PM2.5_prediction
0	19.0	13.304157
1	33.0	10.580834
2	33.0	14.657253
3	28.0	11.660388
4	28.0	21.595922
...	...	...
17499	4.0	6.637554
17500	4.0	7.399473
17501	1.0	6.274939
17502	1.0	6.809088
17503	1.0	5.591497

17504 rows × 2 columns

Model E

	PM2.5_real_value	PM2.5_prediction
0	19.0	19.037182
1	33.0	19.219254
2	33.0	20.249012
3	28.0	20.705229
4	28.0	29.037868
...	...	...
17499	4.0	10.045384
17500	4.0	11.020772
17501	1.0	10.168192
17502	1.0	10.774175
17503	1.0	9.913426

17504 rows × 2 columns

FIGURE 9. **The plot diagram of PM2.5 real values and predictions based on Dali stations.**

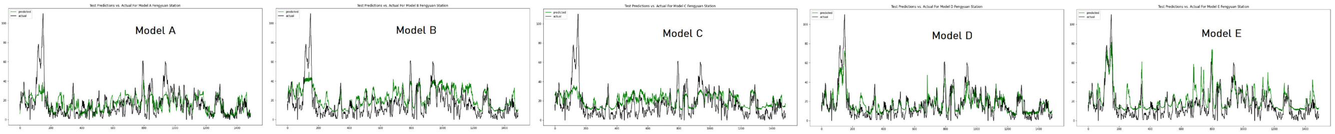
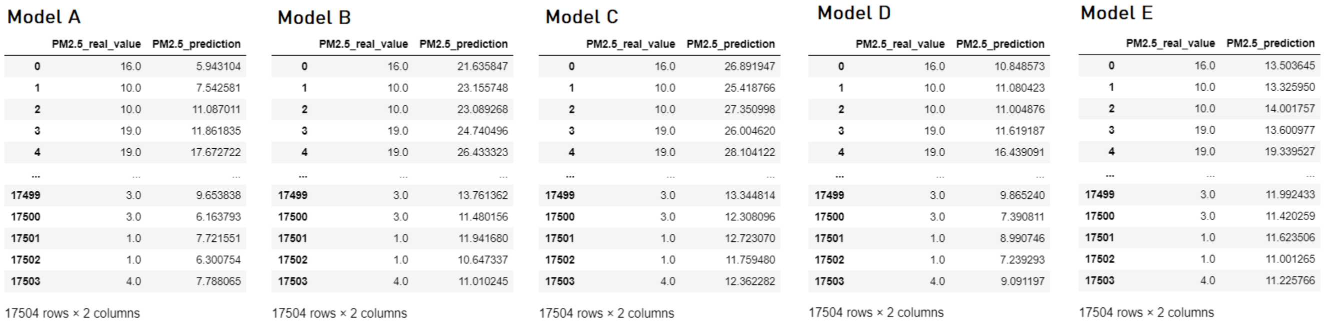


FIGURE 10. **The plot diagram of PM2.5 real values and predictions based on Fengyuan stations.**



Model A

	PM2.5_real_value	PM2.5_prediction
0	16.0	5.943104
1	10.0	7.542581
2	10.0	11.087011
3	19.0	11.861835
4	19.0	17.672722
...	...	...
17499	3.0	9.653838
17500	3.0	6.163793
17501	1.0	7.721551
17502	1.0	6.300754
17503	4.0	7.788065

17504 rows × 2 columns

Model B

	PM2.5_real_value	PM2.5_prediction
0	16.0	21.635847
1	10.0	23.155748
2	10.0	23.089268
3	19.0	24.740496
4	19.0	26.433323
...	...	...
17499	3.0	13.761362
17500	3.0	11.480156
17501	1.0	11.941680
17502	1.0	10.647337
17503	4.0	11.010245

17504 rows × 2 columns

Model C

	PM2.5_real_value	PM2.5_prediction
0	16.0	26.891947
1	10.0	25.418766
2	10.0	27.350998
3	19.0	26.004620
4	19.0	28.104122
...	...	...
17499	3.0	13.344814
17500	3.0	12.308096
17501	1.0	12.723070
17502	1.0	11.759480
17503	4.0	12.362282

17504 rows × 2 columns

Model D

	PM2.5_real_value	PM2.5_prediction
0	16.0	10.848573
1	10.0	11.080423
2	10.0	11.004876
3	19.0	11.619187
4	19.0	16.439091
...	...	...
17499	3.0	9.865240
17500	3.0	7.390811
17501	1.0	8.990746
17502	1.0	7.239293
17503	4.0	9.091197

17504 rows × 2 columns

Model E

	PM2.5_real_value	PM2.5_prediction
0	16.0	13.503645
1	10.0	13.325950
2	10.0	14.001757
3	19.0	13.600977
4	19.0	19.339527
...	...	...
17499	3.0	11.992433
17500	3.0	11.420259
17501	1.0	11.623506
17502	1.0	11.001265
17503	4.0	11.225766

17504 rows × 2 columns

FIGURE 11. **The plot diagram of PM2.5 real values and predictions based on Fengyuan stations.**

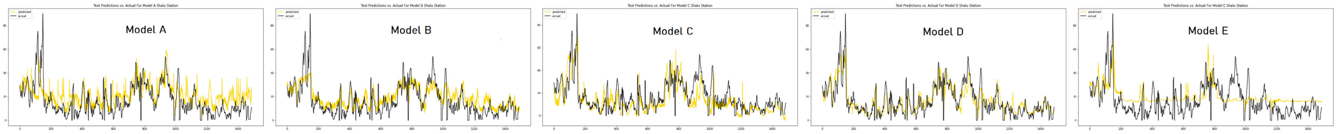
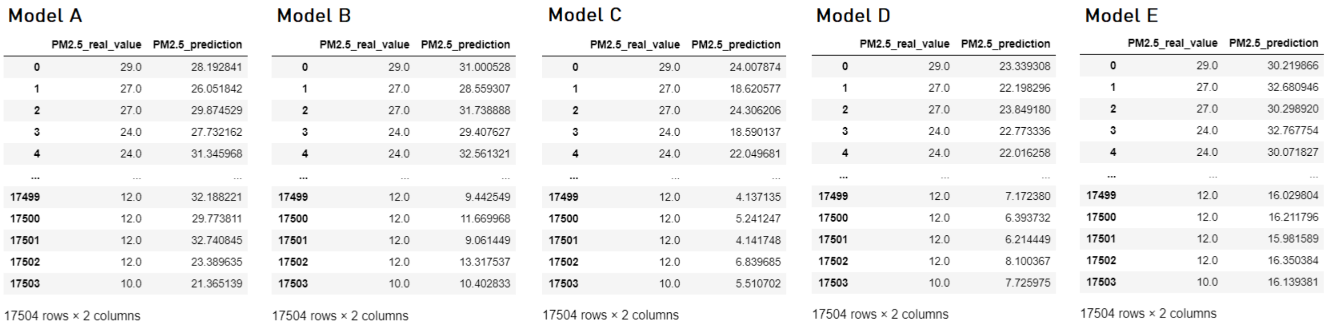


FIGURE 12. **The plot diagram of PM2.5 real values and predictions based on Shalu stations.**



Model A

	PM2.5_real_value	PM2.5_prediction
0	29.0	28.192841
1	27.0	26.051842
2	27.0	29.874529
3	24.0	27.732162
4	24.0	31.345968
...	...	...
17499	12.0	32.188221
17500	12.0	29.773811
17501	12.0	32.740845
17502	12.0	23.389635
17503	10.0	21.365139

17504 rows × 2 columns

Model B

	PM2.5_real_value	PM2.5_prediction
0	29.0	31.000528
1	27.0	28.559307
2	27.0	31.738888
3	24.0	29.407627
4	24.0	32.561321
...	...	...
17499	12.0	9.442549
17500	12.0	11.669968
17501	12.0	9.061449
17502	12.0	13.317537
17503	10.0	10.402833

17504 rows × 2 columns

Model C

	PM2.5_real_value	PM2.5_prediction
0	29.0	24.007874
1	27.0	18.620577
2	27.0	24.306206
3	24.0	18.590137
4	24.0	22.049681
...	...	...
17499	12.0	4.137135
17500	12.0	5.241247
17501	12.0	4.141748
17502	12.0	6.839685
17503	10.0	5.510702

17504 rows × 2 columns

Model D

	PM2.5_real_value	PM2.5_prediction
0	29.0	23.339308
1	27.0	22.198296
2	27.0	23.849180
3	24.0	22.773336
4	24.0	22.016258
...	...	...
17499	12.0	7.172380
17500	12.0	6.393732
17501	12.0	6.214449
17502	12.0	8.100367
17503	10.0	7.725975

17504 rows × 2 columns

Model E

	PM2.5_real_value	PM2.5_prediction
0	29.0	30.219866
1	27.0	32.680946
2	27.0	30.298920
3	24.0	32.767754
4	24.0	30.071827
...	...	...
17499	12.0	16.029804
17500	12.0	16.211796
17501	12.0	15.981589
17502	12.0	16.350384
17503	10.0	16.139381

17504 rows × 2 columns

FIGURE 13. **The plot diagram of PM2.5 real values and predictions based on Shalu stations.**

models, model B has a significant improvement in all stations and each station. Therefore, when the model is implemented in the application, the precision is more excellent, and the load time can be reduced. According to the results, among the other models, model B has the highest accuracy (lowest RMSE), approximately 1 point lower RMSE values, and the

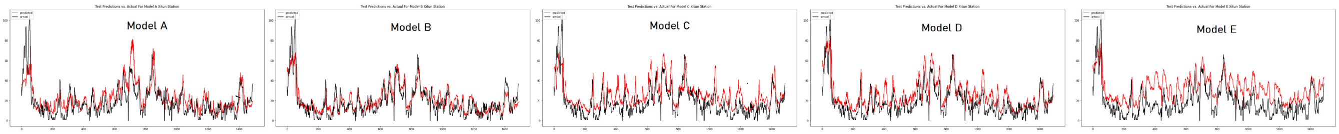
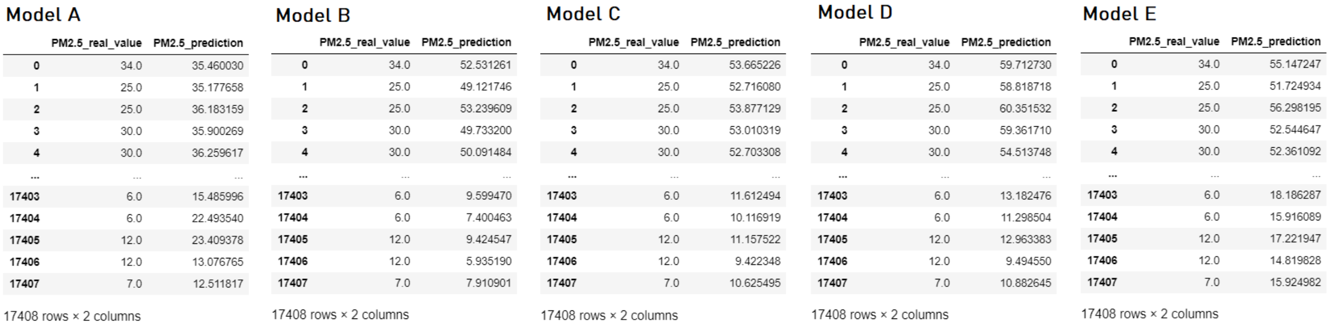


FIGURE 14. The plot diagram of PM2.5 real values and predictions based on Xitun stations.



Model A

	PM2.5_real_value	PM2.5_prediction
0	34.0	35.460030
1	25.0	35.177658
2	25.0	36.183159
3	30.0	35.900269
4	30.0	36.259617
...	...	...
17403	6.0	15.485996
17404	6.0	22.493540
17405	12.0	23.409378
17406	12.0	13.076765
17407	7.0	12.511817

17408 rows × 2 columns

Model B

	PM2.5_real_value	PM2.5_prediction
0	34.0	52.531261
1	25.0	49.121746
2	25.0	53.239609
3	30.0	49.733200
4	30.0	50.091484
...	...	...
17403	6.0	9.599470
17404	6.0	7.400463
17405	12.0	9.424547
17406	12.0	5.935190
17407	7.0	7.910901

17408 rows × 2 columns

Model C

	PM2.5_real_value	PM2.5_prediction
0	34.0	53.665226
1	25.0	52.716080
2	25.0	53.877129
3	30.0	53.010319
4	30.0	52.703308
...	...	...
17403	6.0	11.612494
17404	6.0	10.116919
17405	12.0	11.157522
17406	12.0	9.422348
17407	7.0	10.625495

17408 rows × 2 columns

Model D

	PM2.5_real_value	PM2.5_prediction
0	34.0	59.712730
1	25.0	58.818718
2	25.0	60.351532
3	30.0	59.361710
4	30.0	54.513748
...	...	...
17403	6.0	13.182476
17404	6.0	11.298504
17405	12.0	12.963383
17406	12.0	9.494550
17407	7.0	10.882645

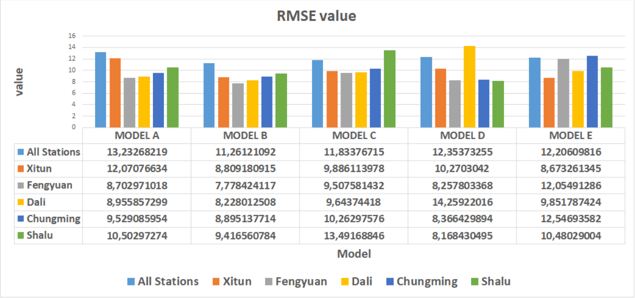
17408 rows × 2 columns

Model E

	PM2.5_real_value	PM2.5_prediction
0	34.0	55.147247
1	25.0	51.724934
2	25.0	56.298195
3	30.0	52.544647
4	30.0	52.361092
...	...	...
17403	6.0	18.186287
17404	6.0	15.916089
17405	12.0	17.221947
17406	12.0	14.819828
17407	7.0	15.924982

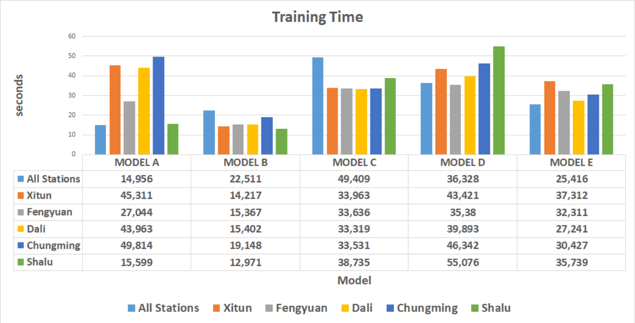
17408 rows × 2 columns

FIGURE 15. The plot diagram of PM2.5 real values and predictions based on Xitun stations.



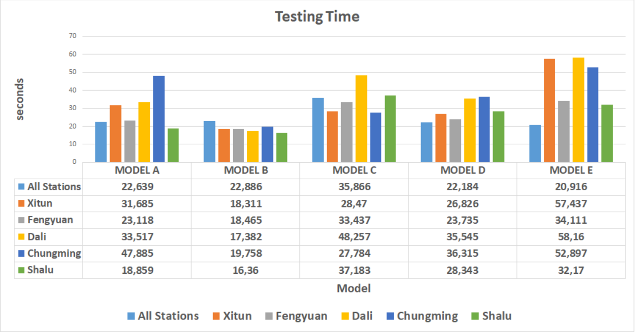
	MODEL A	MODEL B	MODEL C	MODEL D	MODEL E
All Stations	13,23268219	11,26121092	11,83376715	12,35373255	12,20609816
Xitun	12,07076634	8,809180915	9,886113978	10,2703042	8,673261345
Fengyuan	8,702971018	7,778424117	9,507581432	8,257803368	12,05491286
Dali	8,955857299	8,228012508	9,64374418	14,25922016	9,851787424
Chungming	9,529085954	8,895137714	10,26297576	8,366429894	12,54693582
Shalu	10,50297274	9,416560784	13,49168846	8,168430495	10,48029004

FIGURE 16. Comparison of RMSE of five models.



	MODEL A	MODEL B	MODEL C	MODEL D	MODEL E
All Stations	14,956	22,511	49,409	36,328	25,416
Xitun	45,311	14,217	33,963	43,421	37,312
Fengyuan	27,044	15,367	33,636	35,38	32,311
Dali	43,963	15,402	33,319	39,893	27,241
Chungming	49,814	19,148	33,531	46,342	30,427
Shalu	15,599	12,971	38,735	55,076	35,739

FIGURE 17. Comparison of training time of five models.



	MODEL A	MODEL B	MODEL C	MODEL D	MODEL E
All Stations	22,639	22,886	35,866	22,184	20,916
Xitun	31,685	18,311	28,47	26,826	57,437
Fengyuan	23,118	18,465	33,437	23,735	34,111
Dali	33,517	17,382	48,257	35,545	58,16
Chungming	47,885	19,758	27,784	36,315	52,897
Shalu	18,859	16,36	37,183	28,343	32,17

FIGURE 18. Comparison of testing time of five models.

shortest training testing period. The RMSE calculations based on the stations indicate that training with the entire stations dataset has less accuracy, with an RMSE value of about 3 points higher than training with the individual station's dataset.

V. CONCLUSION

This paper demonstrated the PM2.5 forecasting model using Long Short-Term Model (LSTM) seq2seq combined with the statistical method. The air pollution data were extracted from Taiwan EPA for the Taichung City dataset in 2014 - 2018. Correlation Analysis, XGBoost, and Chemical processed are used to select the important feature. Five models of PM2.5 forecasting involve LSTM seq2seq for 17 parameters (model A), LSTM seq2seq for PM10, NO2, SO2 (model B), LSTM seq2seq for PM10, O3, AMB TEMP (model C), LSTM seq2seq for top 5 parameters selected by correlation analysis (model D), and LSTM seq2seq for top 5 parameters selected by XGBoost (model E). The study points out that the chemical processed model of PM10, SO2, and NO2 (model B) has the highest accuracy (lowest RMSE), approximately 1 point lower of RMSE values, and the shortest training testing time among the other models. The RMSE calculations based on the stations show that training using all stations dataset has less accuracy at around 3 points higher RMSE value than training based on each station dataset. In the future, the comparison of the deep learning framework and the network could be examined more with other methods, such as multi-step LSTM. Also, the ensemble learning method could be applied in the model application.

REFERENCES

[1] B. Liu, S. Yan, J. Li, G. Qu, Y. Li, J. Lang, and R. Gu, "A sequence-to-sequence air quality predictor based on the n-step recurrent prediction," *IEEE Access*, vol. 7, pp. 43331–43345, 2019.

[2] Y. F. Xing, Y. H. Xu, M. H. Shi, and Y. X. Lian, "The impact of $PM_{2.5}$ on the human respiratory system," *J. Thoracic Disease*, vol. 8, no. 1, p. E69, 2016.

[3] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: A tutorial," *Statist. Sci.*, vol. 14, no. 4, pp. 382–401, 1999.

[4] S. Viswanath, M. Saha, P. Mitra, and R. S. Nanjundiah, "Deep learning based LSTM and SeqToSeq models to detect monsoon spells of India," in *Proc. Int. Conf. Comput. Sci.* Cham, Switzerland: Springer, Jun. 2019, pp. 204–218.

[5] V. Reddy, P. Yedavalli, S. Mohanty, and U. Nakhat, "Deep air: Forecasting air pollution in Beijing, China," Environ. Sci., Tech. Rep., 2018. [Online]. Available: https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/deep-air-forecasting_final.pdf

[6] C.-T. Yang, S.-T. Chen, W. Den, Y.-T. Wang, and E. Kristiani, "Implementation of an intelligent indoor environmental monitoring and management system in cloud," *Future Gener. Comput. Syst.*, vol. 96, pp. 731–749, Jul. 2019.

[7] C.-T. Yang, C.-J. Chen, Y.-T. Tsan, P.-Y. Liu, Y.-W. Chan, and W.-C. Chan, "An implementation of real-time air quality and influenza-like illness data storage and processing platform," *Comput. Hum. Behav.*, vol. 100, pp. 266–274, Nov. 2019.

[8] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for time series classification," *Neural Netw.*, vol. 116, pp. 237–245, Aug. 2019.

[9] H. Tran, J. Kim, D. Kim, M. Choi, and M. Choi, "Impact of air pollution on cause-specific mortality in Korea: Results from Bayesian model averaging and principle component regression approaches," *Sci. Total Environ.*, vol. 636, pp. 1020–1031, Sep. 2018.

[10] L. Bai, J. Wang, X. Ma, and H. Lu, "Air pollution forecasts: An overview," *Int. J. Environ. Res. Public Health*, vol. 15, no. 4, p. 780, Apr. 2018.

[11] *Taiwan Air Quality Network*. Accessed: Jan. 4, 2019. [Online]. Available: https://taqm.epa.gov.tw/taqm/en/ YearlyDataDownload.aspx

[12] *Civil IoT Taiwan*. Accessed: Mar. 14, 2019. [Online]. Available: https://ci.taiwan.gov.tw/index_ne.aspx

[13] *XGBoost Documentation*. Accessed: Apr. 20, 2019. [Online]. Available: https://xgboost.readthedocs.io/en/latest/

[14] C.-S. Lee, K.-H. Chang, and H. Kim, "Long-term (2005–2015) trend analysis of $PM_{2.5}$ precursor gas NO_2 and SO_2 concentrations in Taiwan," *Environ. Sci. Pollut. Res.*, vol. 25, no. 22, pp. 22136–22152, Aug. 2018.

[15] Y. Xie, B. Zhao, L. Zhang, and R. Luo, "Spatiotemporal variations of $PM_{2.5}$ and PM_{10} concentrations between 31 Chinese cities and their relationships with SO_2, NO_2, CO and O_3," *Particuology*, vol. 20, pp. 141–149, Jun. 2015.

[16] W. M. Hodan and W. R. Barnard, "Evaluating the contribution of $PM_{2.5}$ precursor gases and re-entrained road emissions to mobile source $PM_{2.5}$ particulate matter emissions," MACTEC Federal Programs, Research Triangle Park, NC, USA, Tech. Rep., 2004. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.593.2783&rep=rep1&type=pdf

[17] P. H. McMurry, M. F. Shepherd, and J. S. Vickery, *Particulate Matter Science for Policy Makers: A NARSTO Assessment*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[18] F. Chollet. *A Ten-Minute Introduction to Sequence-to-Sequence Learning in Keras*. Accessed: Aug. 30, 2019. [Online]. Available: https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html

[19] C.-T. Yang, S.-T. Chen, C.-H. Chang, W. Den, and C.-C. Wu, "Implementation of an environmental quality and harmful gases monitoring system in cloud," *J. Med. Biol. Eng.*, vol. 39, no. 4, pp. 456–469, Aug. 2019.

[20] C.-H. Luo, H. Yang, L.-P. Huang, S. Mahajan, and L.-J. Chen, "A fast $PM_{2.5}$ forecast approach based on time-series data analysis, regression and regularization," in *Proc. Conf. Technol. Appl. Artif. Intell. (TAAI)*, Nov. 2018, pp. 78–81.

[21] L. Lin, C.-Y. Chen, H.-Y. Yang, Z. Xu, and S.-H. Fang, "Dynamic system approach for improved $PM_{2.5}$ prediction in Taiwan," *IEEE Access*, vol. 8, pp. 210910–210921, 2020.

[22] M. Lee, L. Lin, C.-Y. Chen, Y. Tsao, T.-H. Yao, M.-H. Fei, and S.-H. Fang, "Forecasting air quality in Taiwan by using machine learning," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, Dec. 2020.

[23] U. Pak, J. Ma, U. Ryu, K. Ryom, U. Juhyok, K. Pak, and C. Pak, "Deep learning-based $PM_{2.5}$ prediction considering the spatiotemporal correlations: A case study of Beijing, China," *Sci. Total Environ.*, vol. 699, Jan. 2020, Art. no. 133561.

[24] 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.

[25] S. W. Choi and B. H. S. Kim, "Applying PCA to deep learning forecasting models for predicting $PM_{2.5}$," *Sustainability*, vol. 13, no. 7, p. 3726, Mar. 2021.

[26] J. Wang, H. Li, H. Yang, and Y. Wang, "Intelligent multivariable air-quality forecasting system based on feature selection and modified evolving interval type-2 quantum fuzzy neural network," *Environ. Pollut.*, vol. 274, Apr. 2021, Art. no. 116429.

[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

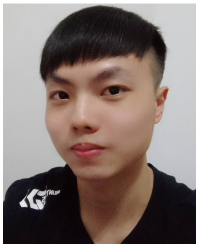
[28] S. Zhao, D. Yin, Y. Yu, S. Kang, D. Qin, and L. Dong, "$PM_{2.5}$ and O_3 pollution during 2015–2019 over 367 Chinese cities: Spatiotemporal variations, meteorological and topographical impacts," *Environ. Pollut.*, vol. 264, Sep. 2020, Art. no. 114694.

[29] B. M. Hashim, S. K. Al-Naseri, A. Al-Maliki, and N. Al-Ansari, "Impact of COVID-19 lockdown on NO_2, O_3, $PM_{2.5}$ and PM_{10} concentrations and assessing air quality changes in Baghdad, Iraq," *Sci. Total Environ.*, vol. 754, Feb. 2021, Art. no. 141978.

[30] E. Kristiani, Y.-A. Chen, C.-T. Yang, C.-Y. Huang, Y.-T. Tsan, and W.-C. Chan, "Using deep ensemble for influenza-like illness consultation rate prediction," *Future Gener. Comput. Syst.*, vol. 117, pp. 369–386, Apr. 2021.

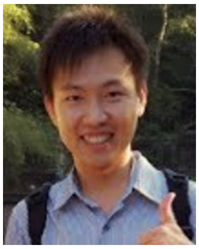
[31] E. Kristiani, C.-T. Yang, C.-Y. Huang, P.-C. Ko, and H. Fathoni, "On construction of sensors, edge, and cloud (iSEC) framework for smart system integration and applications," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 309–319, Jan. 2021.

ENDAH KRISTIANI received the M.S. degree in electrical engineering (information technology) from Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2007. She is currently pursuing the Ph.D. degree with the Department of Industrial Engineering and Enterprise Information, Tunghai University, Taichung, Taiwan. In August 2007, she joined the Department of Informatics Engineering, Faculty of Engineering and Computer Science, Krida Wacana Christian University (UKRIDA), Jakarta. She joins the High-Performance Computing Laboratory, Tunghai University. Her research interests include machine learning and edge AI.



TING-YU KUO received the B.S. degree from the Department of Computer Science, Tunghai University, Taichung, Taiwan, in July 2020. He is currently a Graduate Student with National Chung Cheng University, Taiwan. His work focuses specifically on machine learning and smart applications.

CHAO-TUNG YANG (Member, IEEE) received the Ph.D. degree in computer science from National Chiao Tung University, in July 1996. He is currently a Distinguished Professor of computer science with Tunghai University, Taiwan. In August 2001, he joined the Faculty of the Department of Computer Science, Tunghai University. He is serving in a number of journal editorial boards, including *Future Generation Computer Systems*, *International Journal of Communication Systems*, *KSII Transactions on Internet and Information Systems*, and *Journal of Cloud Computing*. He has published more than 300 articles in journals, book chapters, and conference proceedings. His present research interests include cloud computing, big data, parallel computing, and deep learning. He is a member of the IEEE Computer Society and ACM.



KAI-CHIH PAI received the Ph.D. degree from the Graduate Institute of Educational Information and Measurement, National Taichung University of Education, Taiwan. He is currently an Assistant Professor and working with Tunghai University. His research interests include intelligent tutoring systems, natural language processing, conversation-based assessment, and automated text analysis.

CHIN-YIN HUANG received the Ph.D. degree from Purdue University, USA. He is currently a Professor of industrial engineering and enterprise information with Tunghai University, Taiwan. His publications appear in the *International Journal of Production Research*, *International Journal of Production Economics*, *Computers in Industry*, *Computers and Industrial Engineering*, *Robotics and Computer-Integrated Manufacturing*, *Epilepsy Research*, *Production Engineering*, and *Engineering Computations*. He has coauthored chapters for *Handbook of Industrial Engineering*, *Handbook of Industrial Robotics*, and *Handbook of Automation*. He has coauthored two books in *Industrial Engineering and Management* published in Taiwan. His research interests include healthcare management, clinical data analysis, distributed manufacturing systems, manufacturing process optimization, and industry 4.0.

KIEU LAN PHUONG NGUYEN was born in Vietnam, in 1987. She received the B.S. degree in environmental technology and the M.S. degree in environmental science and management from the University of Liege, Belgium, in 2009 and 2014, respectively. She is currently pursuing the Ph.D. degree in environmental science with Tunghai University, Taiwan. From 2012 to 2016, she was a Lecturer with the Department of Environment, Nguyen Tat Thanh University. Her studies are related to environmental system analysis and management towards sustainable development through statistical techniques and programming in air quality control, water resources management, and solid waste management.