Environmental Science and Pollution Research (2020) 27:38155–38168
https://doi.org/10.1007/s11356-020-09855-1

RESEARCH ARTICLE

An ensemble learning based hybrid model and framework for air pollution forecasting

Yue-Shan Chang[1] · Satheesh Abimannan[2] · Hsin-Ta Chiao[3] · Chi-Yeh Lin[1] · Yo-Ping Huang[4]

Received: 13 April 2020 / Accepted: 22 June 2020 / Published online: 3 July 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
As advance of economy and industry, the impact of air pollution has gradually gained attention. In order to predict air quality, there were many studies that exploited various machine learning techniques to build predictive model for pollutant concentration or air quality prediction. However, enhancing the prediction performance always is the common problem of existing studies. Traditional templates based on machine learning and deep learning methods, such as GBTR (gradient boosted tree regression), SVR (support vector machine-based regression), and LSTM (long short-term memory), are most promising approaches to address these problems. Some previous researches showed that ensemble learning technology can improve predictive performance of other domains. In order to improve the accuracy of forecasting, in this paper, we propose a hybrid model and framework to improve the forecasting accuracy of air pollution. We not only exploit stacking-based ensemble learning scheme with Pearson correlation coefficient to calculate the correlation between different machine learning models to integrate various forecasting models together, but also construct a framework based on Spark+Hadoop machine learning and TensorFlow deep learning framework to physically integrate these models to demonstrate the next 1 to 8 h' air pollution forecasting. We also conduct experiments and compare the result with GBTR, SVR, LSTM, and LSTM2 (version 2) models to demonstrate the proposed hybrid model's predictive performance. The experimental results show that the hybrid model is superior to the existing models used for predicting air pollution.

Keywords Air pollution forecasting · Ensemble learning · LSTM · Pearson correlation coefficient · PM2.5 · SVR · GBTR

Introduction

Air pollution is one of key factors for global warming that lead to the causes of environment. The World Health Organization (WHO) (Who.int 2019) claimed that the polluted air inhaled by 95% of world's people. The contaminant particles are contained in the $PM_{2.5}$ and PM_{10} are the diameter of 2.5 μm and 10 μm (US EPA. 2019). It can easily pass through the human respiratory system through nose and throat that damage the physical system and stimulate many diseases (UN Environment 2019; Yang et al. 2018; Fan et al. 2016; Guo et al. 2020). As a precaution, the air pollution forecast, forms the basis for taking effective pollution control measures, and hence, accurate forecasting of air pollution has become an important issue (Delavar et al. 2019).

The pollution information primarily includes $PM_{2.5}$ and PM_{10} values. These values are useful for the authorities to take prevention action to reduce the pollution control. Therefore, prediction of the $PM_{2.5}$ and PM_{10} concentration value is the managerial solution to prevent and mitigate the malevolent ramifications. Hence, predicting the concentration values of $PM_{2.5}$ and PM_{10} continuously needs novel methods.

For government, local and tribal air quality planners to address these levels of pollution backgrounds are extremely difficult and need to be seen. Governments are therefore increasingly concerned about the prediction of $PM_{2.5}$ concentrations. In recent years, many studies (Zhou et al. 2014; Elangasinghe et al. 2014; Hu et al. 2014; Deng et al. 2019; Maharani and

Responsible editor: Marcus Schulz

✉ Yue-Shan Chang
ysc@mail.ntpu.edu.tw

[1] Department of Computer Science and Information Engineering, National Taipei University, New Taipei City, Taiwan

[2] Galgotias University, Greater Noida, Uttar Pradesh, India

[3] Tunghai University, Taichung City, Taiwan

[4] National Taipei University of Technology, Taipei City, Taiwan

Murfi 2019; Wang and Song 2018; Soh et al. 2018; Cho et al. 2019; Yi 2018; Mahajan et al. 2018; Rybarczyk and Zalakeviciute 2018; Franceschi et al. 2018; Corani 2005; Bai et al. 2018; Fielding, R. T. Chapter 5 2000; Zhang et al. 2019) have been conducted on the prediction of air pollution with the exception of air pollution gas concentration. Indeed, physical model, statistical model, machine learning technique, and deep learning are used to predict the air pollution. Each approach has its own pros and cons in different situations. In addition, as well known that ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem. Unlike ordinary machine learning approaches that try to learn one hypothesis from training data, ensemble methods try to build a set of hypotheses and combine them to predict what them want (Zhou 2019; Polikar 2006; Behera and Roy 2016; Mitchell 1997; Breiman 1996; Usmani et al. 2018; Zheng and Zhong 2011; Siwek et al. 2010; Zhang et al. 2015; Shang and He 2018; Verma et al. 2018; Rijal et al. 2018; Li et al. 2018; Liu et al. 2019; Franceschi et al. 2018; Ventura et al. 2019).

According to existing studies (Smola and Schölkopf 2004; Tsai et al. 2018; Chen et al. 2018), there are various machine learning methods, such as general machine learning and deep learning, have different precisions in different areas. Therefore, how to combine various methods to improve the accuracy of air quality prediction will be an important issue. In addition, different methods require different computing environments and resources, such as traditional machine learning, which can perform predictions on big data platforms, while deep learning methods often have better performance on the GPU. Therefore, how to integrate these different environments to predict air quality in real time is also an important issue.

In this work, we will propose a hybrid model and framework to improve the air pollution forecasting accuracy. We not only exploit stacking-based ensemble learning scheme with Pearson correlation coefficient to calculate the correlation between different machine learning models to integrate the forecasting results of various forecasting models (GBTR, SVR, LSTM, and LSTM2 (version 2)) together, but also construct a framework based on Spark+Hadoop machine learning and TensorFlow deep learning framework to physically integrate these models to demonstrate the next 1 to 8 h air pollution forecasting. In the work, the main difference between LSTM and LSTM2 is the scheme of filling the missing value. In the LSTM, the missing values are filled up with zero, while the LSTM2 exploit the Akima's interpolation (Akima 1970). Obviously, the forecast result of LSTM and LSTM2 will be not the same. The Pearson correlation coefficient is used to find the best model for different time and different monitoring station. We use the historical prediction results to find the linear regression equations of these models, as a weight for adjusting the prediction results, and then correct the prediction results. In this work, we retrieve the air pollution data from the Environmental Protection Administration Executive Yuan, R.O.C (Taiwan), from year 2012 to 2018 to train and test various forecasting model, and finally use the data of year 2019 to verify the result. The mean absolute error (MAE) and the root mean square error (RMSE) are exploited as performance metrics of the work. We also conduct many experiments and compare with GBTR, SVR, LSTM, and LSTM2 (version 2) models to demonstrate the predictive performance of the proposed model. The experimental results show that the hybrid model based on the hybrid framework is superior to single air pollution model.

The main contributions of this work are as follows:

1. Building a computing framework for the proposed hybrid model, which is referring to author's previous work (Tsai et al. 2018; Chen et al. 2018; Chang et al. 2018).
2. Proposing a hybrid model that exploits stacking ensemble learning model to integrate various machine learning models for improving the air pollution forecasting accuracy
3. Applying Pearson correlation coefficient to decide correlation with the four kinds (GBT, SVR, LSTM, LSTM2) of model and exploiting linear regression equation to find the best model
4. Substituting the forecast value with the result obtained from best model and evaluating the result using MAE and RMSE factors

The rest of this paper is arranged as follows. Section II presents background and related work. Section III describes the proposed hybrid framework method. Section IV shows the experimental results, and finally Section V provides the conclusions of the work.

Background and related work

This section discusses the background of ensemble learning in various approaches to predict the concentration of $PM_{2.5}$. The concepts GBT (gradient boosted tree regression), Support Vector Machine (SVM), and long short-term Memory (LSTM) are also outlined and compare with the hybrid model suggested in this paper.

Ensemble learning

Behera (Behera and Roy 2016) stated that the ensemble learning is a kind of supervised algorithm; it can train and predict the accurate value. The study also shows that there are different types of ensemble learning such as Bayes optimal classification, Bootstrap aggregation, Boosting, Bayesian averaging parameter, Bayesian model combination, Bucket of models, Stacking, and Remote sensing.

Mitchell (Mitchell 1997) used the optimal classifier from Bayes, which is a classification technique that assumes that the data is conditionally independent from the class to make the calculation more feasible. Breiman (1996) has conducted a study in order to obtain a better learning model for the ensemble. In Bagging predictors, each model has the same weight that trains each model to achieve high classification accuracy with a randomly drawn subset of the training set. By training each model data, boosting builds an ensemble model that can deliver better accuracy.

Usmani et al. (2018) predicted the performance of Karachi Stock Exchange (KSE), using comprised model of four kinds of machine learning technique such as single layer perceptron (SLP), multilayer perceptron (MLP), radial basis function (RBF), and support vector machines (SVM), respectively. The accuracy of the result was up to 95.7%. A study has been conducted by Zheng and Zhong (2011) to exploit the ensemble model incorporating with ARIMA and ANN for improving time series forecast. The result indicated that ensemble model has been the effective way to decreasing the forecast error.

The authors in (Siwek et al. 2010) merged multilayer perceptron (MLP), support vector machine for regression (SVR), Elman network (EN), and radial basis function network (RBF). They integrated predicted values of those models into one final forecast value with additional neural network to forecast the daily average values of PM_{10} and showed that neural predictor ensemble improved the accuracy of air pollution forecast. The result demonstrated that the hybrid model was effective for short-term forecast of air pollutant.

Zhang et al. (2019) launched a nonlinear autoregressive with exogenous input (NARX) network and Auto Regressive Moving Average (ARMA) composed a hybrid model to predict air pollutants in short terms. NARX network solved the problem of nonlinear and multidimensional, an ARMA improved the flexibility of the model. The result showed that hybrid model was effective for air pollutant forecast of short term. Further, Shang and He (2018) developed a combined random forests and ensemble neural network model to predict $PM_{2.5}$ concentrations in every hour. It had shown that the ensemble neural network had better performance than the random forest.

Verma et al. (2018) constructed an ensemble of three Bi-directional LSTMs (BiLSTM) to improve the prediction of $PM_{2.5}$. It showed that the ensemble model of BiLSTM performed better than single BiLSTM in most of case. Rijal et al. (2018) worked with an ensemble of deep neural network-based regression with outdoor images to estimate $PM_{2.5}$ concentrations, consisted of three CNN learners, which were VGG-16, Inception-v3, and ResNet50. The result showed that ensemble of three learners generated a better $PM_{2.5}$ concentration prediction in comparison to individual learner.

Li et al. (2018) built a wavelet neural network ensemble model, composed of predictive products of environmental weather models CUACE, BREMPS, and WRF-Chem. It showed that ensemble model effectively reduced deviation and had higher accuracy in comparison with four kinds of neural network models (BP, RBF, Elman, and T-S fuzzy). Liu et al. (2019) proposed a model consists of five algorithms: wavelet packet decomposition (WPD), gradient boost regression tree (GBRT), linear programming boosting (LPBoost), multilayer perceptron (MLP), and Dirichlet process mixture model (DPMM). Based on four pollutant data in Tangshan, the proposed model has satisfactory forecasting performance.

Franceschi et al. (2018) conducted a study to develop a model to forecast PM_{10} and $PM_{2.5}$ to help local authorities to prevent human exposure to high levels of pollution using combined model with data mining algorithms, artificial neural networks (ANN) and k-means clustering with multilayer perceptron for hourly forecasting of air pollution. The results of the study are useful to anticipate and take measures to control air pollution and protect human population.

Ventura et al. (2019) developed a model with an aim to anticipation of air pollution episodes in different areas (rural, industrial, and urban) using two models: Holt–Winters (HW) and artificial neural network (ANN), using $PM_{2.5}$ concentration time series. The result of both forecast models proved that it is accurate enough to be considered as a useful tool and to help to make decisions about air quality management.

Methodology

Data source

The air pollution data is retrieved from Environmental Protection Administration of Executive Yuan, R.O.C (Taiwan), from year 2012 to 2018 in this work. These data includes 17 attributes, such as CO, NO, NO_2, NO_x, O_3, *PH Rain*, PM_{10}, $PM_{2.5}$, *Rainfall*, *Rain Cond*, *RH*, *SO2*, *Wind HR*, *Wind Direction*, *Wind Speed*, *and Wind HR*. This work considers air pollution prediction concentration of $PM_{2.5}$ and PM_{10}. We have collected the data from EPA of Taiwan and have built a visualized platform based on Google Map, as shown in Fig. 1.

As we know that there are some missing values while retrieving the data from data source. In this work, the method used to fill in the missing value is the Akima. Techniques of gradient boosted tree regression (GBT), support vector machine (SVR), and long short-term memory (LSTM) machine learning algorithms are used to predict the concentration of $PM_{2.5}$ every hour.

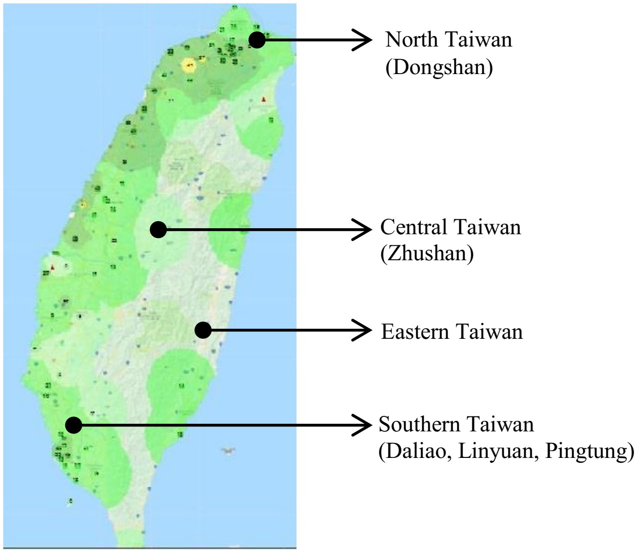


Fig. 1 Google Map of various parts of Taiwan (http://120.126.151.156/national/index.html?at = 25.0000,121.0000,10&m = pm25)

Model evaluation

Mean absolute error (MAE) and root mean square error (RMSE) are used to evaluate the forecasting performance of the proposed hybrid model. The MAE value reveals the average deviation between the actual data and forecasting data. The RMSE is sensitive to the relatively close to the ground and carrying a lot of weight error and reflects refined average departure from the norm of forecasting data. MAE and RMSE defined as in (1) and (2),

$$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}\left|f_n - R_n\right| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(f_n - R_n\right)^2} \tag{2}$$

where N is the number of data, f_n is the forecast value of the model, and R_n is observed value.

Gradient boosted tree

Gradient boosted tree regression (GBT) (Friedman 2002) is a kind of machine learning technique for regression and classification problem, which produces a prediction model.

In general, the technique of gradient boosting is used with decision trees as base learners. Gradient boosting combines weak "learners" into a single strong learner in an iterative fashion. Generic gradient boosting at the m-th step would fit a decision tree $h_m(x)$, to pseudo-residuals. Let L_m be the number of its leaves. The tree partitions the input space into L_m disjoint regions $R_{1m}, \ldots, R_{L_m m}$, and predicts a constant value in each region, he output of $h_m(x)$ for input x can be as formula (3):

$$h_m(x) = \sum_{i=1}^{L_m} b_{im} 1 R_{im}(X), \tag{3}$$

where b_{im} is the value predicted in the region R_{im}. This optimization algorithm builds the model in a stage-wise fashion. Further, $\mathrm{PM}_{2.5}$ concentration is exploited as input of GBT model to predict $\mathrm{PM}_{2.5}$ concentration in every 1 h. Then, autoregressive integral moving average (ARIMA) (Hyndman and Athanasopoulos 2018) forecast technique is used to predict a future of future value based on past data. The formula of ARIMA is expressed as (4):

$$\left(1-\sum_{i=1}^{p}\varnothing_i L^i\right)(1-L)^d X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t, \tag{4}$$

where p and φ_i are the order of autoregressive model and parameter of the autoregressive part. L^i, d, X_t, q, θ_i, and ε_t are log operator, degree of differencing, time series of data, order of the moving average, parameter of the moving average, and error terms.

Further, to predict $\mathrm{PM}_{2.5}$ concentration in next 2 to 8 h, ARIMA model is used. Then, the input data is generated by ARIMA and GBT model to predict $\mathrm{PM}_{2.5}$ concentration. If there are missing data from the Environmental Protection Administration, the observed value is used as a missing value.

Support vector machine (SVM)

Support vector machine (SVM) is a machine learning technique, which constructs a set of hyperplane in multidimensional space for regression analysis (Cortes 1995). It is used for minimize the error and individualize the hyperplane for maximizes the margin. In this training, data model is same as GBT model, as shown in (5).

$$\mathrm{f}(x) = w^T\varphi(x) + b \tag{5}$$

where x is the data, w^T is a maximum-margin hyperplane, $\varphi(x)$ is the radial basis function kernel (RBF kernel) (Chang et al. 2010), and b is the bias. The RBF kernel represented as shown in (6),

$$\begin{aligned}\mathrm{k}\left(x,x'\right) &= \exp\left(\frac{\|\mathrm{x}-\mathrm{x}'\|_2^2}{2\sigma^2}\right)K\left(x,x'\right)\\ &= \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)\end{aligned} \tag{6}$$

where x are the data, x' are the data after mapping, ||x-x ′‖^2$\|\mathrm{x}-\mathrm{x}'\|_2^2$ is the squared Euclidean distance between x and x', and σ is the parameter.

Akima's interpolation method (Akima 1970) is used for fill-up the missing values, similar to GBT model. ARIMA is also used to generate input data of predicting $PM_{2.5}$ concentration in next 2 to 8 h. Then, the SVR model is used to predict $PM_{2.5}$ concentration in next 1 h.

Further, the input data generated by ARIMA and SVR model is employed to predict $PM_{2.5}$ concentration in next 2 to 8 h.

Long short-term memory

Long short-term memory (LSTM) (Hochreiter and Schmidhuber 1997) is an artificial recurrent neural network architecture used in the field of deep learning. It consists of a cell, input gate, output gate, and forget gate. The cell remembers the values of time intervals and the three gates are control of the flow of information, which enter and exit the cell. The forget gate decides which information will be discarded from the cell. The function of the forget gate is shown in (7),

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] = b_f\right) \tag{7}$$

where W_f is the weight matrices, h_{t-1} is the hidden layer vector at the previous unit $(t-1)$, b_f is the bias vector parameters, and x_t is the input vector which is added into *Sigmoid* function $S(t)$ to generate the value between 0 and 1. Zero means information is completely forgotten, while one represents that information are completely remembered. Hence, *Sigmoid* function $S(t)$ is shown in (8),

$$S(t) = \frac{1}{1+e^{-t}}, \tag{8}$$

Further, the input gate decides which new information needs to be remembered in cell state. The function of deciding how much new information cell state needs to be remembered is shown in (9),

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{9}$$

Then, the value of i_t is between 0 and 1. It multiply by i_t and election message $\widetilde{C}_t$ is represented in (10) to obtain the information which want to add into the cell state C_t as shown in (11).

$$\widetilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \tag{10}$$

$$C_t = f_t \times C_{t-1} + i_t, \tag{11}$$

In addition, we employ past 72 h' data retrieved from Environmental Protection Agency of Taiwan to predict $PM_{2.5}$ concentrations in next 1 to 8 h.

Pearson correlation coefficient

A measure of the linear correlation between two variables X and Y is the Pearson correlation coefficient (Pearson 1895). According to the inequality between Cauchy-Swarz (Steele 2004), Pearson's coefficient of correlation ranges from − 1 to + 1. If the value exceeds 0 represents a positive linear correlation, less than 0 is a negative linear correlation, and no linear correlation is equal to 0. We assume that X and Y's value is the model's real and predicted value. These values replace them with Pearson's coefficient formula of correlation to forecast one of the best models out of four in eight different time intervals. The formula of the Pearson correlation coefficient is displayed as (12),

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \tag{12}$$

where r is the Pearson correlation coefficient, n is the number of the forecast value, $\overline{X}$ is the average value of observed value, and $\overline{Y}$ is the average value of forecast value of the model.

In four models, we select the model's largest Pearson correlation coefficient value as the best model. The largest coefficient value of Pearson correlation represents the model's forecast value having a high relation to the value of reality. The reason we use Pearson's coefficient formula to find the best model is that we use the best model's forecasted value and the observed value to perform linear regression analysis.

Linear regression

Linear regression is a method of analyzing the relationship between explanatory variable and dependent variable (Freedman 2009). The explanatory variable and dependent variable relationship is expressed in two unknowns as a linear equation. One variable that is unknown is intercept, and another is the coefficient of regression. The linear regression formula is expressed as (13),

$$Y = \alpha X + \beta \tag{13}$$

where X is explanatory variable, Y is dependent variable, α is regression coefficient, and β is intercept.

We are building the model of linear regression (Seal 1967) in this research work to find the linear equation that showing the relationship between the best model's forecast value and the actual value. It is unknown the coefficient of regression and intercept that we need to acquire by model of linear regression. The model of linear regression is shown in (14).

$$y_n = \alpha x_n + \beta, n = 1, 2, \ldots, N \tag{14}$$

where N is the number of data, x_n is the forecast value of the best model, and y_n is the adjusted value.

Hybrid model development

The hybrid model is built by integrating the forecasting results of following models: GBT, SVR, LSTM, and LSTM2, and is used for forecasting the next 1 to 8 h of concentration of $PM_{2.5}$. Since the forecasting of each model is autonomous, the Pearson correlation coefficient is calculated and the linear regression design is constructed individually. Pearson's index of correlation calculation, the ratio of regression, and the interception of distinct time intervals are not the same. The model design of each monitoring stations has individual configuration. For instance, a site's model of $PM_{2.5}$ forecasting may comprise more than two types of model.

Therefore, by constructing a linear regression model, this hybrid model acquires a coefficient of correlation and intercept of each moment series. We calculate the correlation coefficient from the prediction results of every hour in all prediction models, and calculate the linear regression equation of the hybrid model with the highest value of the correlation coefficient from 1 to 8 h, respectively. Then use the calculated value of the obtained linear regression equation as the prediction result. The value resulting from the linear equation calculation is used as the Hybrid model's prediction value.

The observed value and the forecasted value of four types of model must exist at the same time so that we can do the Pearson correlation coefficient calculation. We use the predicted and observed values of the model with the highest correlation coefficient to calculate the coefficient and intercept of the linear regression equation. The flowchart of the hybrid model is shown in Fig. 2.

In Algorithm 1, GBT_n, SVR_n, $LSTM_n$, and $LSTM2_n$ are the forecasted values of GBT, SVR, LSTM, and LSTM2. Where n is the number of input data and $Reality_n$ is the observed value. Line 2 sets initial value of hr is one. Up to 8 h will be predicted by the loop. Then, calculate the coefficient of Pearson correlation between four models and the value of reality. The $GBT^{hr}_{Pearson}$, $SVR^{hr}_{Pearson}$, $LSTM^{hr}_{Pearson}$, and $LSTM2^{hr}_{Pearson}$ are Pearson correlation coefficient of GBT, SVR, LSTM, and LSTM2 respectively. Then, it selects the highest Pearson correlation coefficient as the best model. The $Best^{hr}_{Pearson}$ is the Pearson correlation coefficient of the best model. Then it builds the model of linear regression with the predicted value of the best models and the observed value. The coefficient of regression and interception is $\alpha[hr]$ and $\beta[hr]$, respectively. Finally, these values are replaced by the hybrid model and get the next 8 h forecast value.

Algorithm 1: The Hybrid model algorithm

```
      Input: $GBT_n$, $SVR_n$, $LSTM_n$, $LSTM2_n$, $Reality_n$
      Output: Hybrid
1     Begin
2     hr ←1
3     while (hr ≤ 8 ) do
4         $GBT^{hr}_{Pearson} \leftarrow Pearson(GBT^{hr}_n, Reality_n)$
5         $SVR^{hr}_{Pearson} \leftarrow Pearson(SVR^{hr}_n, Reality_n)$
6         $LSTM^{hr}_{Pearson} \leftarrow Pearson(LSTM^{hr}_n, Reality_n)$
7         $LSTM2^{hr}_{Pearson} \leftarrow Pearson(LSTM2^{hr}_n, Reality_n)$
8         $Best^{hr}_{Pearson} \leftarrow \mathbf{max}(GBT^{hr}_{Pearson}, SVR^{hr}_{Pearson}, LSTM^{hr}_{Pearson}, LSTM2^{hr}_{Pearson})$
9         $\alpha[hr], \beta[hr] \leftarrow$ Linear Regression $(Best^{hr}_n, Reality_n)$
10        hr ← hr + 1
11    while (hr ≤ 8 ) do
12        $Hybrid^{hr}_n \leftarrow \alpha[hr] \times Best^{hr}_n + \beta[hr]$
13        hr ← hr + 1
14    End
```

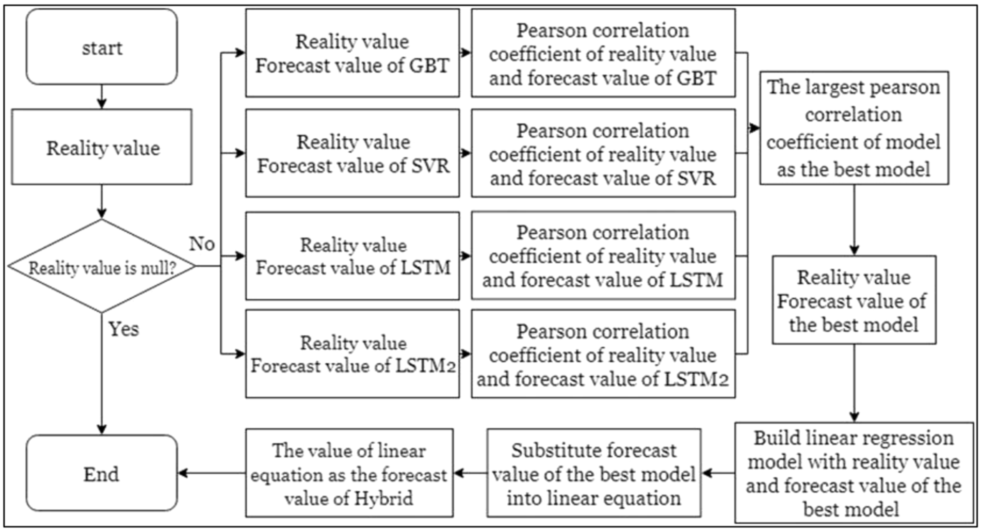

start
Reality value
Reality value is null?
No
Yes
Reality value Forecast value of GBT
Reality value Forecast value of SVR
Reality value Forecast value of LSTM
Reality value Forecast value of LSTM2
Pearson correlation coefficient of reality value and forecast value of GBT
Pearson correlation coefficient of reality value and forecast value of SVR
Pearson correlation coefficient of reality value and forecast value of LSTM
Pearson correlation coefficient of reality value and forecast value of LSTM2
The largest pearson correlation coefficient of model as the best model
Reality value Forecast value of the best model
Build linear regression model with reality value and forecast value of the best model
Substitute forecast value of the best model into linear equation
The value of linear equation as the forecast value of Hybrid
End

Fig. 2 Flow graph of hybrid system

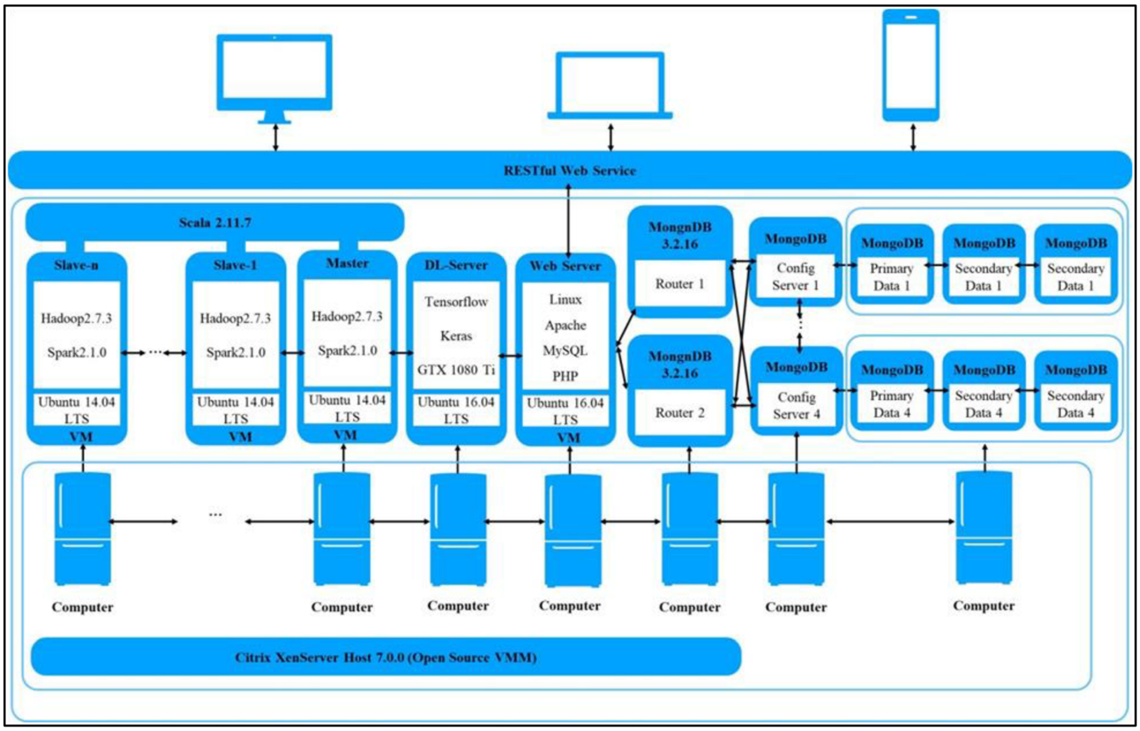


Fig. 3 Hybrid framework of air pollution forecasting

Hybrid model framework

The hybrid model forecasting framework is modified and revised from (Chen et al. 2018) and is shown in Fig. 3. It is comprised of MongoDB, Hadoop, Spark, Tensorflow, Keras, and RESTful it embraces. MongoDB is a type of NoSQL database that is also a cross-platform and open-source document-oriented database with fault tolerance capability. Hadoop (https:/hadoop.apache.org) is a distributed, clustered system processing framework that processes and stores data. In this context, the programming of Hadoop Distributed File System (HDFS) is used to split the files and distribute them in a cluster into different nodes to make data processing faster. Spark (http://spark.apache.org/) is used to analyze big data. It not only provides Hadoop's accelerated analytics service, but also a library for machine learning (MLlib). In the work, both the GBT and the SVR models are implemented on the Spark and Hadoop platform. In addition, Tensorflow is used to build and to run deep learning models, such as LSTM and LSTM2. Representational State Transfer (RESTful) (Fielding, R. T. Chapter 5 2000) is web service architecture to observe the outcome of each model. Web server allows the forecast value of models to be shown on the web through RESTful technique in this research work. Furthermore, GTX 1080Ti's four Graphics Processing Unit (GPU) cluster is used to accelerate deep learning computation.

Case study

Evaluation

In this work, the data includes 67 air quality monitoring stations and is retrieved from the EPA of Taiwan. The training data is from year 2012 to 2017. One of four models' forecasted values in 2018 are used as the values that fit reality value to build linear regression model. Pearson correlation coefficient is calculated to find the best model, which has highest relation with observed value. Table 1 shows the best forecasting model and coefficient value for Pearson correlation.

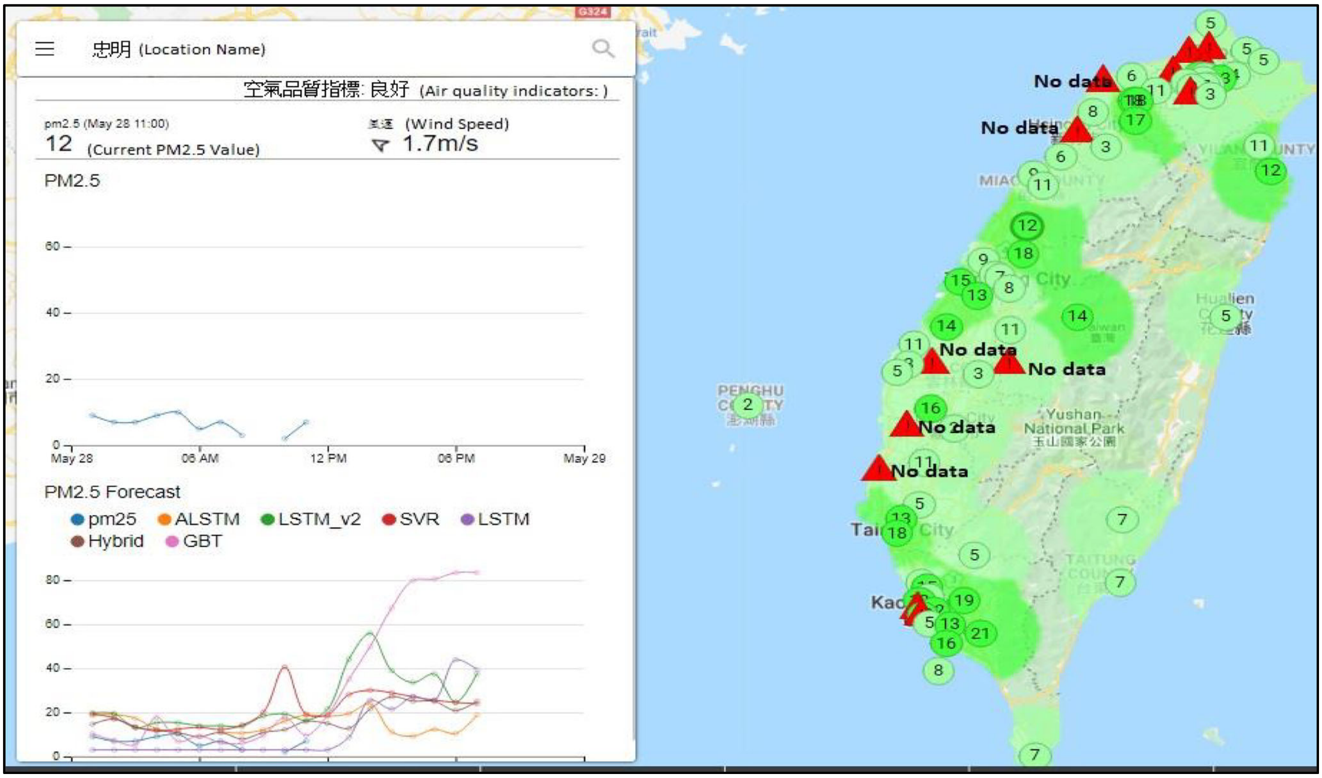
For instance, the best forecast model in Daliao district is GBT for the first 2 h. The coefficient of GBT's Pearson correlation is 0.92 and 0.83. The remaining hours are SVR's best model for predicting $PM_{2.5}$ concentration. Pearson's correlation coefficient is 0.78, 0.75, 0.73, 0.72, 0.72, and 0.71 respectively. Therefore, the hybrid model of Daliao consists of a linear regression model of GBT and SVR. The hybrid model may contain more than two types of linear regression model. The *r* value of Linyuan, Dongshan, Nanzi and Linyuan, Xiaogan, Dongshan, and Nanzi in LSTM and LSTM2 are 0.41–0.71 and 0.4–0.73. The *r* value therefore shows that there is no significant difference between LSTM and LSTM2. Hence, the research work is not included LSTM2 method for PM10 concentration forecasting. The *r* value of GBT and SVR are 0.63–0.92 and 0.63–0.94. Therefore, there is no big difference between GBT and SVR. The best model

Table 1 Pearson correlation coefficient between hybrid models

		Daliao	Linyuan	Pingtung	Zhushan	Xiaogang	Dongshan	Nanzi
1 h	Model	GBT	SVR	GBT	SVR	GBT	LSTM2	GBT
	Pearson	0.92	0.89	0.91	0.94	0.91	0.68	0.93
2 h	Model	GBT	LSTM2	GBT	GBT	GBT	LSTM	GBT
	Pearson	0.83	0.73	0.85	0.86	0.84	0.5	0.84
3 h	Model	SVR	LSTM	GBT	GBT	GBT	LSTM	GBT
	Pearson	0.78	0.71	0.78	0.8	0.78	0.53	0.77
4 h	Model	SVR	LSTM	GBT	SVR	GBT	LSTM	GBT
	Pearson	0.75	0.64	0.74	0.75	0.74	0.52	.71
5 h	Model	SVR	LSTM2	GBT	SVR	GBT	LSTM	LSTM
	Pearson	0.73	0.63	0.7	0.71	0.7	0.41	0.7
6 h	Model	SVR	LSTM2	SVR	SVR	GBT	LSTM2	GBT
	Pearson	0.72	0.62	0.67	0.67	0.66	0.45	0.63
7 h	Model	SVR	LSTM2	SVR	SVR	GBT	LSTM2	LSTM2
	Pearson	0.72	0.49	0.65	0.65	0.63	0.4	0.6
8 h	Model	SVR	LSTM2	SVR	SVR	LSTM2	LSTM	LSTM2
	Pearson	0.71	0.56	0.64	0.63	0.62	0.41	0.6

for Daliao, Pingtung, and Zhushan is the combination of GBT and SVR. The other best model for Linyuan is the combination of SVR and LSTM. LSTM is well performed for the district of Dongshan. The combination of GBT and LSTM is good for Nanzi. Another experimental observation, SVR is Daliao's best model, and from 4 to 8 h Zhushan.

In Fig. 4, Google Map of Taiwan shows on the right side. It indicates that the current location $PM_{2.5}$ value and red color indicate that the data is not available at the particular point of time. The graphical representation of the current $PM_{2.5}$ value and history of the past and prediction of the future value is shown on the left side of Fig. 4. The reader can refer to the


忠明 (Location Name)

空氣品質指標: 良好 (Air quality indicators:)

pm2.5 (May 28 11:00)
12 (Current PM2.5 Value)

風速 (Wind Speed)
1.7m/s

Fig. 4 Online daily basis $PM_{2.5}/PM_{10}$ forecasting screen shot

website for further understanding of the concept. (http://120.126.151.156/national/index.html?at = 25.0000,121.0000,10&m = pm25).

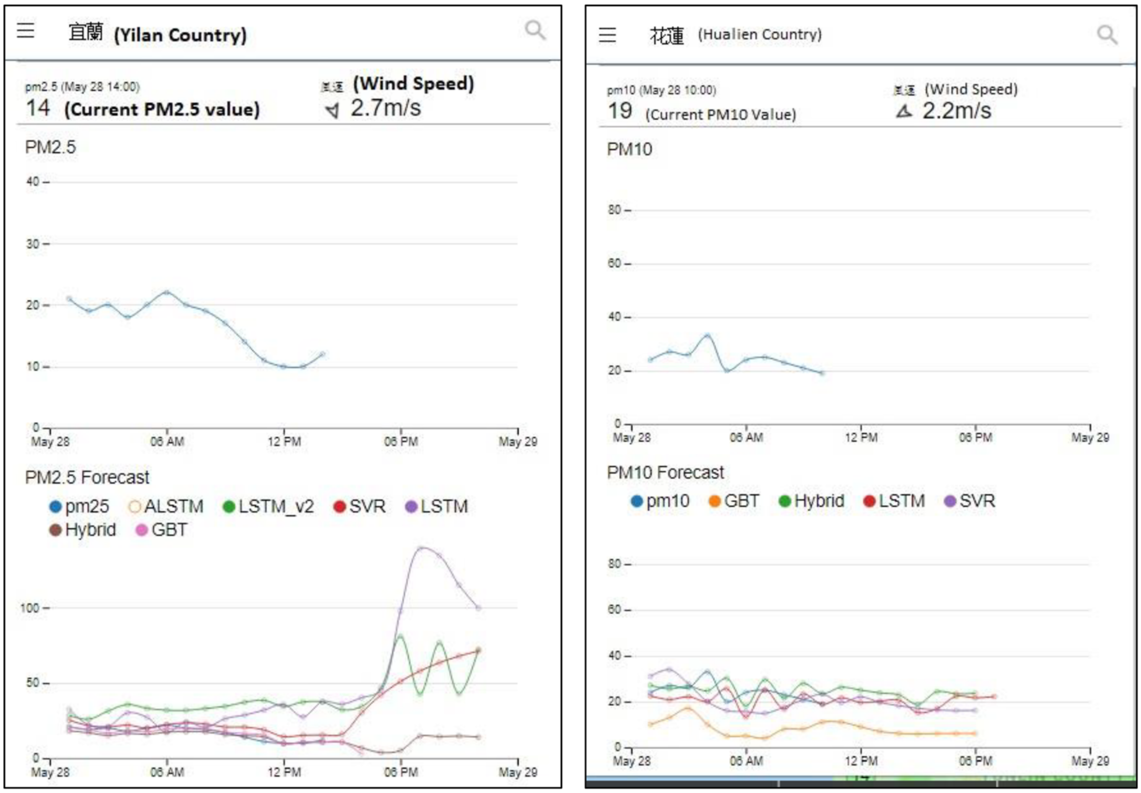
The current $PM_{2.5}$ reading and forecasted value for the next 8 h are shown in Fig. 5a. The current reading of $PM_{2.5}$ at the specific location and its wind speed are shown in the top of the diagram. At the end of the diagram shows how different algorithms are performed in the graphical form. The GBT, SVR, LSTM, LSTM2, and the hybrid model are expected to provide the $PM_{2.5}$ forecast for the next 24 h. Figure 5 a shows one more algorithm ALSTM (Chang et al. 2020). We are not considering the performance of ALSTM in this paper. In Fig. 5 a, the hybrid model predicts more accurately than other algorithms. In some ways, the GBT algorithm also performs very close to the hybrid model. The SVR, LSTM, and LSTM2 algorithms perform well up to 8 h. As shown in the figure, the performance of the SVR, LSTM, and LSTM2 is not good after 8 h.

Figure 5 b shows the current concentration value of PM_{10} and the forecast for the next 8 h. Compared with other algorithms, the proposed hybrid model performs well from the first hour onwards. Its predictive result is very close to the observed value. The difference between the observed value and the forecasted value is very small. LSTM also performs very close to the hybrid model. But the GBT's performance is not good compared with other algorithms. But the same GBT performs well in predicting the value of $PM_{2.5}$, as shown in Fig. 5a. This shows that the performance of GBT, SVR, and LSTM changes constantly each frequent hour. However, the hybrid model constantly performs the result. The forecasted result of hybrid model is close to observed value.

The forecasting accuracy is measured using MAE and RMSE. Figure 6 a, c, e, and g and Fig. 7 a, c, e, and g show the comparison of MAE and RMSE between single model, such as GBT, SVR, LSTM, LSTM2, and proposed hybrid model. Hybrid model of forecasting concentration of $PM_{2.5}$ and PM_{10} in 1 to 8 h can be observed to have lower MAE compared with single model. Figure 6 b, d, f, and h and Fig. 7 b, d, f, and h are the average RMSE comparison between single model and hybrid model. The experimental result reveals that the hybrid model performance is good in both MAE and RMSE for 1 to 8 h $PM_{2.5}$ and PM_{10} prediction. Hence, the proposed hybrid model is therefore an effective way to improve air pollution.

Comparison of existing work with hybrid model

Table 2 shows the comparison of the existing works and proposed hybrid model with respect to various factors. The existing hybrid models (Mahajan et al. 2018; Liu et al. 2019; Jiang et al. 2018) use various combinations of algorithms. These models use only limited variables, maximum of 7



(a) (b)

Fig. 5 Hybrid model compared with GBR, SVR, LSTM, and LSTM2 for $PM_{2.5}$ and PM_{10}

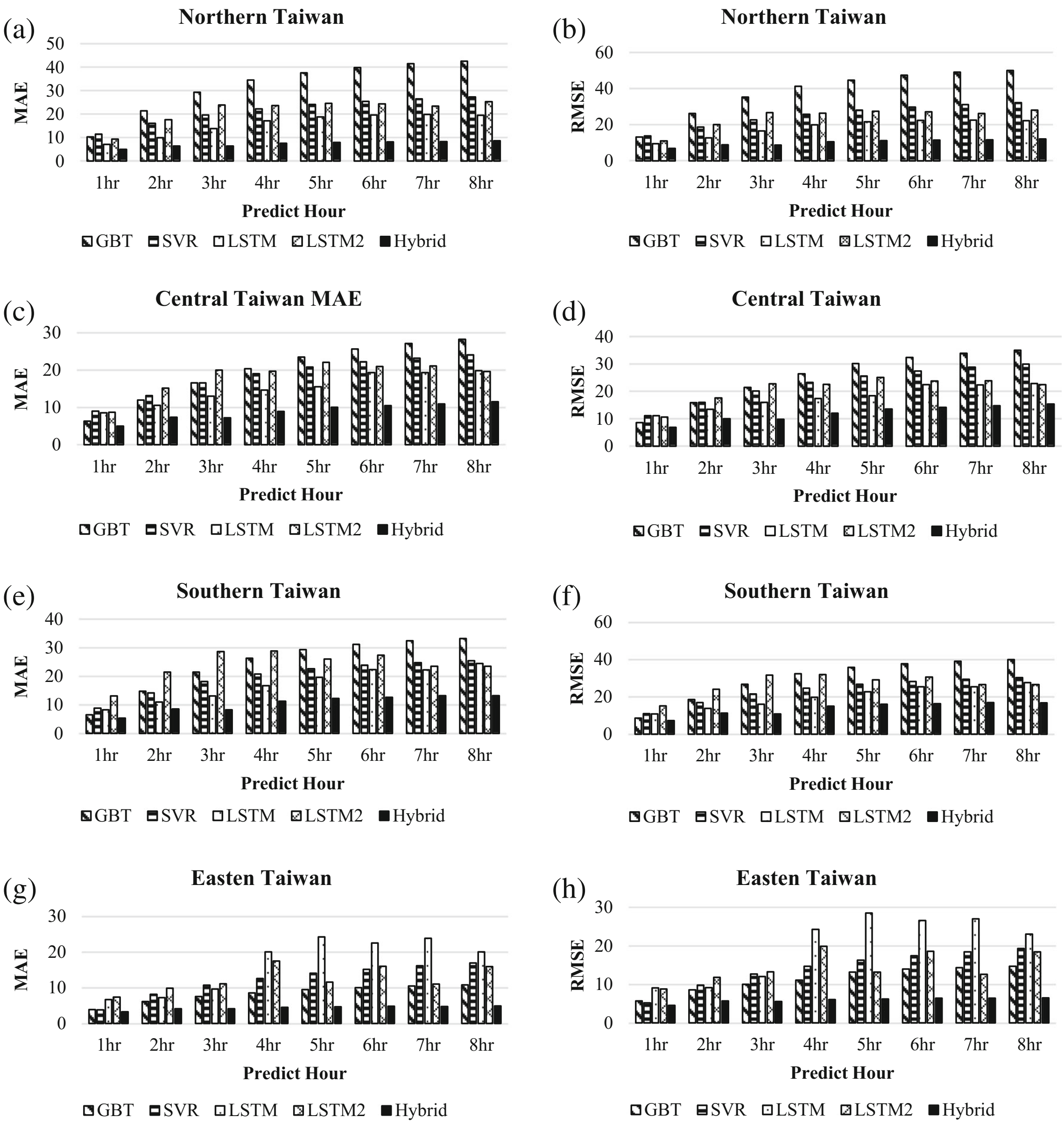


Fig. 6 **a** Northern Taiwan—MAE ($PM_{2.5}$). **b** Northern Taiwan—RMSE ($PM_{2.5}$). **c** Central Taiwan—MAE ($PM_{2.5}$). **d** Central Taiwan—RMSE ($PM_{2.5}$). **e** Southern Taiwan—MAE ($PM_{2.5}$). **f** Southern Taiwan—RMSE ($PM_{2.5}$). **g** Eastern Taiwan—MAE ($PM_{2.5}$). **h** Eastern Taiwan—RMSE ($PM_{2.5}$)

futures, as the input for forecasting the $PM_{2.5}$, and the data collected from the locations are limited up to 7 sites. The training data set used by the existing hybrid models is a period of maximum 3 years and minimum 2 months. In the proposed hybrid model, we ensemble the four well-known algorithms: LSTM, SVR, GBT, and LSTM2, and up to 17 variables as the input for forecasting the $PM_{2.5}$. The data set used in this work are collected from 67 monitoring station that include data for 7 years. In addition, compared with the existing model, we have built a practical platform, which consisting of MongoDB, Hadoop, Spark, Tensorflow, Keras, and RESTful web service server, for computing the data to obtain the more accurate forecasting value. Hence, the proposed hybrid model is considered to be better than the existing one.

Conclusion and future work

In this work, we have proposed a hybrid model to improve the prediction accuracy of air pollution, particularly for $PM_{2.5}$ and PM_{10}. In this work, we build a computing framework for the proposed hybrid model and propose a hybrid model that exploits stacking ensemble learning model to integrate various machine learning models for improving the air pollution

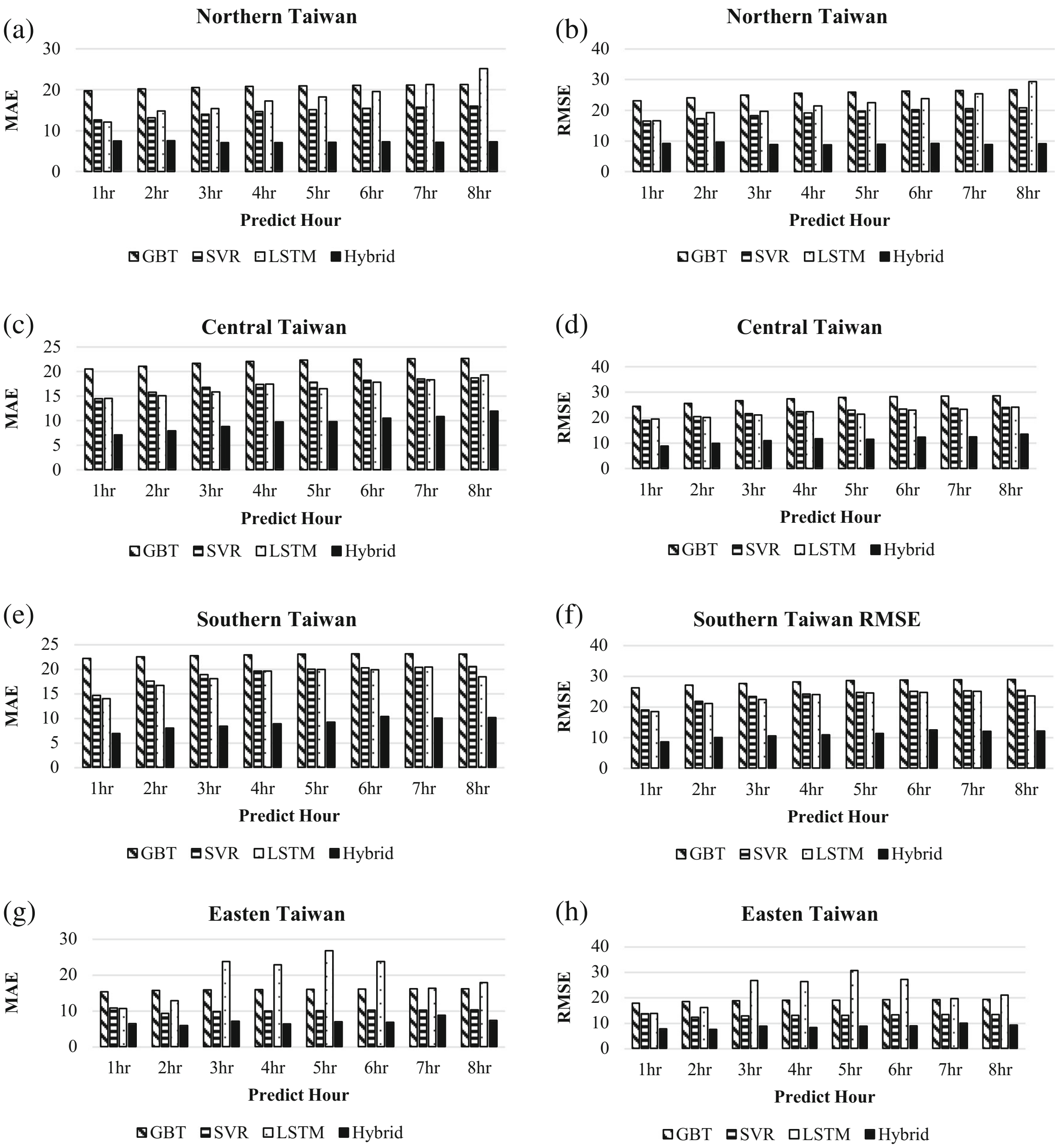


Fig. 7 **a** Northern Taiwan—MAE (PM_{10}). **b** Northern Taiwan—RMSE (PM_{10}). **c** Central Taiwan—MAE (PM_{10}). **d** Central Taiwan—RMSE (PM_{10}). **e** Southern Taiwan—MAE (PM_{10}). **f** Southern Taiwan—RMSE (PM_{10}). **g** Eastern Taiwan—MAE (PM_{10}). **h** Eastern Taiwan—RMSE (PM_{10})

forecasting accuracy. In the hybrid model, Pearson correlation coefficient is applied to decide correlation with the four kinds (GBT, SVR, LSTM, LSTM2) of model and exploiting linear regression equation to find the best model. The forecasted value will be substituted by the forecasted value of proposed best model. The proposed hybrid model is run on the cloud-based big data platform, which comprises Spark+Hadoop machine learning environment and TensorFlow-based deep learning framework to physically integrate these models to demonstrate the next 1 to 8 h air pollution forecasting. The evaluation results reveal that the proposed hybrid model is superior to single traditional machine learning techniques in terms of MAE and RMSE. We think that the proposed framework can be easily applied to other country.

The concentration of air pollutants remains difficult to predict; however, because of the multiplicity of sources and the complexity of physical and chemical processes which influence air pollutant formation and transportation. We will develop new methods for predicting concentrations of $PM_{2.5}$ and PM_{10} in the future. For example, we know that meteorological data is an important parameter that affects air quality prediction. We can design an adaptive weighting method based on parameters of various attributes, especially meteorological data at the location of the station (such as wind direction, wind

Ta5ble 2 Comparison of LSTM and hybrid model of air pollution forecasting

Model	LSTM	SVR	GBT	LSTM2	No. of input variables	Data collected (number of monitoring sites/ locations)	Data collected (number of years)	Implementation
3-D CTM simulation (Kim et al. 2019)	√	√	–	–	6	7	3	Simulation
Hybrid neural network model (Mahajan et al. 2018)	-	-	-	-	1	4	0.2	Grid-based clustering approach
Hybrid model (Jiang et al. 2018)	-	-	-	-	5	2	0.8	MATLAB
Hybrid model (Liu et al. 2019)	-	-	√	-	7	1	1	MATLAB
Hybrid model (proposed in this work)	√	√	√	√	14	67	7	MongoDB, Hadoop, Spark, Tensorflow, Keras, and RESTful web services

speed, rainfall). Give different weights according to different values, and then calculate which weight can get the smallest MAE, in order to get the best model, and finally add the prediction results obtained by various weight models to our ensemble learning model to get better result.

Funding information This work was partially supported by Ministry of Science and Technology of Taiwan, Republic of China under Grant No. MOST 106-3114-M-305-001-A, MOST 108-2119-M-305-001-A, MOST 109-2119-M-305-001-A, and MOST108-2321-B-027-001-; and by National Taipei University under Grant No. 106-NTPU_A-H&E-143-001, 107-NTPU_A-H&E-143-001, and 108-NTPU_A-H&E-143-001.

References

Akima H (1970) A new method of interpolation and smooth curve fitting based on local procedures. J ACM 17(4):589–602

Bai L, Wang J, Ma X, Lu H (2018) Air pollution forecasts: an overview. Int J Environ Res Public Health 15(4):780. https://doi.org/10.3390/ijerph15040780

Behera RN, Roy MD (2016) Ensemble based hybrid machine learning approach for sentiment classification-a review. Int J Comput Appl 146(6):31–36. https://doi.org/10.5120/ijca2016910813

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140. https://doi.org/10.1023/A:1018054314350

Chang YW Hsieh CJ Chang KW Ringgaard M, Lin C, Chih-Jen J (2010) Training and testing low-degree polynomial data mappings via linear SVM. Journal of Machine Learning Research, 11, 1471–1490, 2010. [online] Available at: http://www.jmlr.org/papers/volume11/chang10a/chang10a.pdf [Accessed 26 May 2019]

Chang, Y.-S., Lin, K.-M., Tsai, Y.-T., Zeng, Y.-Z. and Hung, C (2018) Big data platform for air quality analysis and prediction. In: 2018 27th Wireless and Optical Communication Conference (WOCC). IEEE Xplore,1–3. https://doi.org/10.1109/WOCC.2018.8372743

Chang Y-S, Chiao H-T, Abimannan S, Huang Y-P, Tsai Y-T, Lin K-M (2020) An LSTM-based aggregated model for air pollution forecasting. Atmos Pollut Res 11(8):1451–1463. https://doi.org/10.1016/j.apr.2020.05.015

Chen L, Huang H, Wu C, Tsai Y and Chang Y-S (2018) LoRa-based air quality monitor on unmanned aerial vehicle for smart city. In: 2018 International Conference on System Science and Engineering (ICSSE). IEEE Xplore, pp 1–5. https://doi.org/10.1109/ICSSE.2018.8519967

Cho K, Lee B, Kwon M, Kim S (2019) Air quality prediction using a deep neural network model. J Korean Soc Atmos Environ 35(2):214–225. https://doi.org/10.5572/KOSAE.2019.35.2.214

Corani G (2005) Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. Ecol Model 185(2–4):513–529. https://doi.org/10.1016/j.ecolmodel.2005.01.008

Cortes, C. Vapnik, V (1995) Support-vector networks. Mach Learn, 20(3), 273–297. https://doi.org/10.1023/A:1022627411411

Delavar MR, Gholami A, Shiran GR, Rashidi Y, Nakhaeizadeh GR, Fedra K, Afshar SH (2019) Novel method for improving air pollution prediction based on machine learning approaches: a case study applied to the capital city of Tehran. Int J Geo-Inf 8(2):89–109. https://doi.org/10.3390/ijgi8020099

Deng F, Ma L, Gao X, Chen J (2019) The MR-CA models for analysis of pollution sources and prediction of PM2.5. IEEE Trans Syst Man Cybernet Syst 49(4):814–820. https://doi.org/10.1109/TSMC.2017.2721100

Elangasinghe M, Singhal N, Dirks K, Salmond J, Samarasinghe S (2014) Complex time series analysis of PM10 and PM2.5 for a coastal site using artificial neural network modelling and k-means clustering. Atmos Environ 94:106–116. https://doi.org/10.1016/j.atmosenv.2014.04.051

Fan J, Li S, Fan C, Bai Z, Yang K (2016) The impact of PM2.5 on asthma emergency department visits: a systematic review and meta-analysis. Environ Sci Pollut Res 23:843–885. https://doi.org/10.1007/s11356-015-5321-x

Fielding, R. T. Chapter 5 (2000) Representational State Transfer (REST). Architectural styles and the design of network-based software architectures (Ph.D.). University of California, Irvine, 2000. [online] Available at: https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf

Franceschi F, Cobo M, Figueredo M (2018) Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá, Colombia, using artificial neural networks, principal component analysis, and k-means clustering. Atmos Pollut Res 9(5):912–922. https://doi.org/10.1016/j.apr.2018.02.006

Freedman DA (2009) Statistical models: theory and practice revised. Cambridge University. ISBN: 978-0-521-74385-3

Friedman JH (2002) Stochastic Gradient Boosting. Comput Stat Data Analysis 38(4):367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Guo C, Xu Y, Tian Z (2020) Inversion of PM2.5 atmospheric refractivity profile based on AlexNet model from the perspective of electromagnetic wave propagation. Environ Sci Pollut Res. https://doi.org/10.1007/s11356-020-07703-w

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu X, Waller L, Lyapustin A, Wang Y, Al-Hamdan M, Crosson W, Estes M, Estes S, Quattrochi D, Puttaswamy S, Liu Y (2014) Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. Remote Sens Environ 140:220–232. https://doi.org/10.1016/j.rse.2013.08.032

Hyndman RJ, & Athanasopoulos G (2018) Forecasting: principles and practice, 2nd, OTexts: Melbourne. OTexts.com/fpp2. [accessed on 12^{th} may 2018]

Jiang P, Li C, Li R, Yang H (2018) An innovative hybrid air pollution early-warning system based on pollutants forecasting and Extenics evaluation. Knowl-Based Syst 164:174–192. https://doi.org/10.1016/j.knosys.2018.10.036

Kim HS, Park I, Song CH, Lee K, Yun JW, Kim HK, Jeon M, Lee J (2019) Development of daily PM10 and PM2.5 prediction system using a deep long short-term memory neural network model. Atmos Chem Phys Discuss 19:12935–12951. https://doi.org/10.5194/acp-19-12935-2019

Li, T, Li, X, Wang, L, Ren, Y, Zhang, T, Yu, M (2018) Multi-model ensemble forecast method of PM2.5 concentration based on wavelet neural networks. In: *2018* 1st international cognitive cities conference (IC3*)*, Okinawa, Japan ,81–86, 7–9. https://doi.org/10.1109/IC3.2018.00026

Liu H, Duan Z, Chen C (2019) A hybrid framework for forecasting PM2.5 concentrations using multi-step deterministic and probabilistic strategy. Air Qual Atmos Health 12(7):785–795. https://doi.org/10.1007/s11869-019-00695-8

Mahajan S, Liu H-M, Tsai T-C, Chen L-J (2018) Improving the accuracy and efficiency of PM2.5 forecast service using cluster-based hybrid neural network model. IEEE Access 6:19193–19204. https://doi.org/10.1109/ACCESS.2018.2820164

Maharani D, Murfi H (2019) Deep neural network for structured data - a case study of mortality rate prediction caused by air quality. J Phys Conf Ser 1192:012010. https://doi.org/10.1088/1742-6596/1192/1/012010

Mitchell T (1997) Machine learning. Singapore: McGraw-Hill, 1997. ISBN-13: 978–0070428072

Pearson K (1895) Notes on regression and inheritance in the case of two parents. Proc R Soc Lond 58(347- 352):240–242. https://doi.org/10.1098/rspl.1895.0041

Polikar R (2006) Ensemble based systems in decision making. IEEE Circ Syst Mag 6(3):21–45. https://doi.org/10.1109/MCAS.2006.1688199

Rijal N, Gutta RT, Cao T, Lin J, Bo Q, Zhang J (2018) Ensemble of deep neural networks for estimating particulate matter from images. In: 2018 IEEE 3rd international conference on image, Vision and Computing (ICIVC*)*, 733-738, 27–29. https://doi.org/10.1109/ICIVC.2018.8492790

Rybarczyk Y, Zalakeviciute R (2018) Machine learning approaches for outdoor air quality modelling: a systematic review. Appl Sci 8(12): 2570. https://doi.org/10.3390/app8122570

Seal HL (1967) Studies in the history of probability and statistics. XV: the historical development of the Gauss linear model. Biometrika 54(1–2):1–24. https://doi.org/10.2307/2333849

Shang Z, He J (2018) Predicting hourly PM2.5 concentrations based on random forest and ensemble neural network. In: 2018 Chinese Automation Congress (CAC). pp 234–2345. https://doi.org/10.1109/CAC.2018.8623175

Siwek K Osowski S. Sowinski M (2010) Neural predictor ensemble for accurate forecasting of PM10 pollution. In: The 2010 International joint conference on neural networks (IJCNN), 1-7. https://doi.org/10.1109/IJCNN.2010.5596900

Smola A, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Soh P, Chang J, Huang J (2018) Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access 6:38186–38199. https://doi.org/10.1109/ACCESS.2018.2849820

Steele JM (2004) The Cauchy–Schwarz master class: an introduction to the art of mathematical inequalities, The Mathematical Association of America. ISBN-13 978–0–521-83775-0

Tsai Y, Zeng Y and Chang Y (2018) Air pollution forecasting using RNN with LSTM. In: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 1074–1079. https://doi.org/10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00178

UN Environment (2019). Air pollution: Africa's invisible, Silent Killer [online] Available at: https://www.unenvironment.org/fr/node/20803 [Accessed 26 May 2019]

US EPA (2019). Particulate matter (PM) pollution | US EPA. [online] available at: https ://www.epa.gov/pm-pollution [Accessed 26 May 2019]

Usmani M Ebrahim M Adil SH Raza K (2018) Predicting market performance with hybrid model. In: 2018 3rd international conference on emerging trends in engineering, sciences and technology (ICEEST), 1-4. https://doi.org/10.1109/ICEEST.2018.8643327

Ventura L, de Oliveira Pinto F, Soares L, Luna A, Gioda A (2019) Forecast of daily PM2.5 concentrations applying artificial neural networks and Holt–Winters models. Air Qual Atmos Health 12(3): 317–325. https://doi.org/10.1007/s11869-018-00660-x

Verma I Ahuja R Meisheri H, Dey L (2018) Air pollutant severity prediction using Bi-directional LSTM Network. In: 2018 IEEE/WIC/ACM international conference on web intelligence (WI*),* 651-654. https://doi.org/10.1109/WI.2018.00-19

Wang J, Song GA (2018) Deep spatial-temporal ensemble model for air quality prediction. Neurocomputing 314:198–206. https://doi.org/10.1016/j.neucom.2018.06.049

Who.int (2019) How air pollution is destroying our health. [online] Available at: https://www.who.int/air-pollution/news-and-events/how-air-pollution-is-destroying-our-health [Accessed 26 May 2019]

Yang B, Guo J, Xiao C (2018) Effect of PM2.5 environmental pollution on rat lung. Environ Sci Pollut Res 25:36136–36146. https://doi.org/10.1007/s11356-018-3492-y

Yi X (2018) Deep distributed fusion network for air quality prediction. In: 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. [online] London, United Kingdom: ACM New York, 965–973. https://doi.org/10.1145/3219819.3219822

Zhang X, Rui X Xia X Bai X Yin W Dong T (2015) A hybrid model for short-term air pollutant concentration forecasting. In:2015 IEEE International Conference on Service Operations and Logistics, And Informatics (SOLI), 171–175. https://doi.org/10.1109/SOLI.2015.7367614

Zhang Y, Wang Y, Gao M, Ma Q, Zhao J, Zhang R, Wang Q, Huang L (2019) A predictive data feature exploration-based air quality prediction approach. IEEE Access 7:30732–30743. https://doi.org/10.1109/ACCESS.2019.2897754

Zheng F, Zhong S (2011) Time series forecasting using an ensemble model incorporating ARIMA and ANN based on combined objectives. In: 2011 2nd international conference on artificial intelligence, management science and electronic commerce (AIMSEC), 2671-2674. https://doi.org/10.1109/AIMSEC.2011.6011011

Zhou Z-H. Ensemble learning. In: Li, SZ (eds) Encyclopedia of biometrics, Springer, Berlin. [online] Available at: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication /springerEBR09.pdf [Accessed 26 May 2019]

Zhou Q, Jiang H, Wang J, Zhou J (2014) A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Sci Total Environ 496:264–274. https://doi.org/10.1016/j.scitotenv.2014.07.051

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.