IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 30, 2019, accepted June 3, 2019, date of publication June 26, 2019, date of current version September 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925082

Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities

SABA AMEER[1], MUNAM ALI SHAH[1], ABID KHAN[1], HOUBING SONG[2], (Senior Member, IEEE), CARSTEN MAPLE[3], SAIF UL ISLAM[4], AND MUHAMMAD NABEEL ASGHAR[5]

[1]Department of Computer Science, COMSATS University Islamabad, Islamabad 44550, Pakistan
[2]Department of Electrical Engineering and Computer Science, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA
[3]WMG, University of Warwick, Coventry CV4 7AL, U.K.
[4]Department of Computer Science, Dr. A. Q. Khan Institute of Computer Sciences and Information Technology, Rawalpindi 47000, Pakistan
[5]Department of Computer Science, Bahauddin Zakariya University, Multan 60800, Pakistan

Corresponding author: Saif Ul Islam (saiflu2004@gmail.com)

This work was supported by the Alan Turing Institute under EPSRC Grant EP/N510129/1.

ABSTRACT Dealing with air pollution presents a major environmental challenge in smart city environments. Real-time monitoring of pollution data enables local authorities to analyze the current traffic situation of the city and make decisions accordingly. Deployment of the Internet of Things-based sensors has considerably changed the dynamics of predicting air quality. Existing research has used different machine learning tools for pollution prediction; however, comparative analysis of these techniques is required to have a better understanding of their processing time for multiple datasets. In this paper, we have performed pollution prediction using four advanced regression techniques and present a comparative study to determine the best model for accurately predicting air quality with reference to data size and processing time. We have conducted experiments using Apache Spark and performed pollution estimation using multiple datasets. The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) have been used as evaluation criteria for the comparison of these regression models. Furthermore, the processing time of each technique through standalone learning and through fitting the hyperparameter tuning on Apache Spark has also been calculated to find the best-fit model in terms of processing time and lowest error rate.

INDEX TERMS IoT, smart city, air quality index (AQI), data mining, Apache Spark.

I. INTRODUCTION

Air pollution is recognized as one of the main detriments to human health. According to the World Health Organization, 7 million people are at health risk due to air pollution [1]. It is a leading risk factor for a number of health problems such as asthma, skin infections, heart issues, throat and eye diseases, bronchitis, lung cancer and diseases of the respiratory system. Further to the health problems arising from air pollution, it also poses a serious threat to our planet. Pollution emissions from sources such as vehicles and industry is the underlying cause of the greenhouse effect, CO_2 emissions are amongst the foremost contributors to the phenomenon [2]. Climate change has been widely discussed at the global forums and has remained a burning issue for the world over the last two decades, as a result of increased smog and damage to the ozone.

The air pollution prediction problem has been addressed in the past using statistical linear methods but these techniques can provide poor estimations for air pollution due to the complexity and variation in time-series data [3], [4]. Over the last 60 years, a number of machine-learning techniques have been developed to help address the issues of complexity.

A. SMART CITY AND AIR POLLUTION

A smart city is an urban municipality that utilizes information and communication technologies (ICT) to provide better health, transport and energy related facilities to its citizens and enables the government to make efficient use

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

of its available resources, for the welfare of their people. Different types of data collection sensors are deployed at various points within the city which act as a source of information for management of city resources. Better traffic control, energy conservation, waste management, pollution control and improvement in public safety and security are among the fundamental objectives of developing a smart city.

In recent years, urban populations have grown rapidly due to industrialization and the migration of people from rural to urban areas. According to a UN report, approximately 54 to 66 percent of the world's population will move to urban areas by 2050 [5]. With the rise in population, reliance and demand of transportation and energy are also increased, thus adding further industry and vehicles to the cities. This, in turn, increases the sources of pollution emissions, which is becoming a major concern for local and national authorities as well as leaders on the global stage. Local and national governments wish to provide a better lifestyle for its inhabitants through controlling pollution-related diseases. Thus, coping with air pollution is one of the fundamental challenges in urban areas and key goal for smart cities.

B. AIR QUALITY INDEX AND PM2.5

PM2.5 (which means particles less than 2.5 microns in diameter) is a term which is used for the suspended solid and liquid particles in the air e.g. ash, dust and soot [6]. These particles may be emitted in combustion process from power generation or domestic heating or from the vehicles' emissions. Vehicles and industry are primary sources of PM 2.5 pollution, although such particulate matter may also be formed by secondary sources such as the interaction of various gases in the atmosphere. For example sulphur emissions from industry may react with oxygen and water droplets in the atmosphere to form sulphuric acid which is thus a secondary source of particulate matter [7].

These particles, being extremely small and light, have a tendency to stay in the air much longer than larger and heavier particles. This increases the risk of them being inhaled by human beings. Particulate matter of less than 2.5 microns is recognized to have a more adverse effect on human health than other pollution emissions. These particles can easily enter the respiratory system through the inhalation process, and there can badly affect the lungs and breathing. Moreover, it has the potential to cause cardiovascular diseases in people of almost every age group, with children and people above 65 particularly sensitive to its harmful effects [8]. It may cause plaque in arteries or may result in hardening of arteries thus leading to a heart attack. People who are already suffering from lung- or heart-related disease require special precautionary measures to be taken in polluted environments [9].

The effects of PM2.5 were analyzed over the last 25 years [10]. It was estimated that approximately 4.2 million people have died due to long term exposure to PM 2.5 in the atmosphere, while an additional 250,000 deaths have occurred due to ozone exposure. In global rankings of mortality risk factor, PM 2.5 was ranked as 5th and attributed for 7.6 % of total deaths all over the world. From 1990 to 2015, the number of deaths due to air pollution have increased, especially in China and India [10]. Household air pollution resulting from consumption of solid fuels in the underdeveloped and developing countries is also a major cause of mortality and possess a significant health challenge in conjunction with ambient air pollution.

Due to the above-mentioned adverse effects, PM 2.5 concentration is actively monitored by municipalities around the globe, and an air quality index (AQI) is calculated on the basis of it. The air quality index is a function of the concentration of pollutants, but the derivation of the value of AQI varies across nations. It is a dimensionless number, different values of which exhibit different quantities of air pollution. If the PM2.5 concentration is lower, this is reflected in a lower value of AQI while higher concentrations lead to a higher. According to the United States Environmental Protection Agency (EPA), there are six categories of AQI, from Good to Hazardous. The value for the AQI is calculated from concentration of pollutant by the following method [11].

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}} (C - I_{low}) + I_{low} \quad (1)$$

where,

I = Air Quality Index

C = Pollutant concentration

C_{low} = the concentration breakpoints that is $< C$

C_{high} = the concentration breakpoint that is $>= C$

I_{low} = the index breakpoint corresponding to C_{low}

I_{high} = the index breakpoint corresponding to C_{high}

Individual air quality indices are calculated for each separate pollutant concentration and the highest of all the values classify the location's AQI at that given point in time. Particulate matter, sulphur dioxide, ground-level ozone, nitrogen dioxide and carbon monoxide are important contributors for AQI calculations. The AQI is calculated and reported on hourly basis at most places to convey estimates of air pollution to general public. When the AQI is particularly high, people with heart and respiratory diseases may need to avoid outdoor activities or may need to a use mask to protect themselves. Observing people wearing masks is becoming commonplace in some of the largest cities in the world, particularly in China.

In an era of increasing connectivity, reducing costs and size of technology, the concept of the Internet of Things (IoT) is gaining significant traction. New systems that are based on IoT sensors are continually being proposed [14]. Data gathered from sensors can play a vital role in helping cities manage and measure air quality. With the help of sensors that generate data, decisions in smart cities can be made much faster and easier than before. However, it should be recognized that the processing of data brings its own challenges.

A significant challenge for smart cities concerns the handling of information; it is necessary to make sure analysis of data is both efficient and reliable. False interpretations of data can lead to erroneous decisions, which can turn to

be very dangerous. To ensure that the speed and processing of communication sources are made robust, taking care with predictions using artificial intelligence and the use of rigorous data is required.

In this paper, we have performed pollution prediction using the four different regression techniques mentioned in Section III. We present a comparative analysis of these techniques, based on recognized evaluation criteria. In this case we use MAE and RMSE in order to determine the best predictive model to estimate pollution. We have considered air pollution in a number of cities in order to identify the most accurate model.

Furthermore, given the requirement for real-time data processing of smart city data to be efficient, we have analyzed the processing time of these techniques through standalone learning and through fitting the hyperparameter tuning on Apache Spark. In this research, we have proposed the optimal model in terms of both processing time and least error rate.

The remainder of this paper is organized as follows. Section II consists of a review of related studies. Section III describes the proposed architecture and estimation models. The Data Analysis is presented in Section IV while results and discussions have been addressed in Section V. Section VI contains the system evaluation and the conclusions & future work are presented in Section VII and VIII, respectively.

II. LITERATURE REVIEW

In recent years there have been a number of machine learning methods proposed for solving air pollution prediction problems. In this section, we present and analyse some of the key work in the field.

Asgari *et al.* [15] have analyzed the urban pollution and mapped them according to the geographical areas considered. They analyzed data in Tehran from the period of 2009 to 2013, using Apache Spark. Moreover, they have compared the prediction accuracy of Logistic Regression and Naive Bayes algorithm. They have found the Naive Bayes to predict data more accurately than other machine learning algorithms for classifying unknown classes of air quality. The paper presents good results in terms of Apache Spark processing time, however, the algorithm is not appropriate for real-time time series prediction. In [16], the authors address the prediction of air pollutants such as ozone, particle matter (PM2.5) and sulphur dioxide. They use optimization and regularization techniques to predict level of air pollutant for the next day. They have predicted the values using the datasets from two stations. One station predicts the values for O_3 and SO_2 and the other holds values for O_3 and PM2.5. They have modelled the data based on similarity and have used liner regression for grouping. Root- mean-squared error (RMSE) was the evaluation criteria they employed. The limitations with this work arise from Linear Regression models being unable to forecast or handle unforeseen events. Moreover, the data of only two stations is used in this study, which is also limiting in its generality.

The classification of air quality index, and its effect on health, was studied in [17]. The authors implemented a Decision tree method and Naive Bayes J48 for classification. The results they obtained showed that decision tree algorithm performs with 91.9978% accuracy. However, there are many limitations with this research, including the issue that the dataset used was limited. Moreover, the decision tree methods are not to perform poorly over continuous variables and can have issues with overfitting. Another research for classifying of air quality index was proposed by the authors of [18]. In their work the authors employed K-means algorithm; again, in this research the dataset used was limited. Further issues arise when attempting to predict future values, a weakness in K-means methods.

Real-time Affordable Multi-Pollutant (RAMP) is a low-cost pollution monitoring system for measuring pollutants, first proposed in [19]. The authors devise a scheme that reduces the cost of sensors, and utilize a random forest method for predicting future values. However, the dataset consists of data collected over only 2 weeks, which makes it difficult to reliably assess performance. Furthermore, random forest algorithm can encounter problems with overfitting, especially when used with small datasets.

Bougoudis *et al.* have proposed a hybrid computational intelligence system for combined machine learning (HISY-COL) [20]. The method is used to identify correlation of air pollutants levels with weather patterns in an attempt to find the underlying cause of pollutants. They gather data from the wider Attica area to examine the issue. The methods they apply are ensemble methods using artificial neural networks (ANNs) and Random Forests. They claim the accuracy is increased using such an approach, however the feed-forward neural network fails in accurately predicting the continuous values. Moreover, the training data is also very limited in this research. Neural networks with two phases have been employed to train meteorological parameters and then used for analysis of air pollutants with some success, resulting in increased accuracy [21]. Unfortunately the authors have only considered single stations with a few hours' data. Again, the neural network is susceptible to faces overfitting when using small datasets.

Some of the limitations of computational models for air quality are discussed in [22]. The authors propose machine learning techniques for forecasting the O_3 in different countries. They used sparse sampling and randomized matrix decompositions as a pre-processing stage to reduce the dimensionality of the data. They then use a random forest regression technique to forecast for the next 10 days. However, the authors only consider one pollutant, O_3, and the data subsample size is small. Dynamic Neural Network (DNN) have also been used for Air Pollution prediction. In [23], the authors use such an approach on data generated from their low cost sensors. Experiments were conducted on two weeks of data.

Ghoneim has developed a method to predict ozone concentration in smart cities based on a Deep Learning approach,

using a feed-forward neural network [24]. The data used in the study was from the city of Aarhus in the Netherlands. The author performed comparisons between the new method, Support Vector Machines (SVM), and Neural Network machine learning algorithms. The results demonstrate that deep learning neural network schemes perform well, accurately measuring the pollution value. The author only considers one pollutant and solves the problem using a linear method. There is no mention of how the real-time data will be maintained. In [25], the concentration of ozone in Tunisia is studied. The authors have used three monitoring stations for measuring ozone concentration and used Random Forests and Support Vector Regression for future prediction. They have found Random Forests to be a more accurate estimator for predicting ozone. However, the data from three stations is limited and only one variable is considered for future prediction.

Another study for forecasting air pollution in Canada utilizes a multilayer perceptron neural network (MLPNN) [26]. The authors address the issue of air quality prediction and model accuracy. However, the amount of data used in the study is limited and the computational cost for seasonally updating of the model is large.

A deep learning technique for decreasing the error rate of time-series analysis is proposed by authors in [27]. They have made comparison between a neural network with auto regression moving average (ARMA), and support vector regression (SVR) models. Although, the accuracy has been increased, the processing time is not mentioned. Two recent studies propose management schemes to handle the large volumes of data by providing a Big Data management architecture [28], [29], the former specifically considering prediction of air quality in China. Unfortunately the investigators have not implemented the system in either case.

A method for air simulation based on big data is proposed in [30]; the authors perform a comparison of MapReduce Hadoop and Spark for simulating air quality. They have used the dataset of Texas 179 sensors and found that performance benefits of 20∼25 % for the Spark solutions over MapReduce. They have mentioned that real time decision-making can be performed, but did not mention the prediction accuracy. Another Apache Spark based AQI prediction system using Random Forest, implemented using the Spark distributed on multiple clusters is given in [31]. However, while Random Forests can be used for classification of data, the method is not used for real-time analysis of time series data.

Recently, in China, air pollutant data of different cities has been analyzed using an ensemble Neural Network technique for 16 cities in China [31]. Although, accuracy of predictive model is improved, the processing time is not discussed. Furthermore, this technique is only applicable to comparison of different regions in an offline-mode, and not useful for real-time processing of data within cities.

Chang *et al.* [32] propose a cloud-based ETL (Extract-Transform-Load) framework for Air Quality analysis and prediction. The authors worked on pre-processing of data collected from a variety of sources and have achieved up to 81% accuracy using RNN.

One study has monitored data in five cities in China and analyzed the occurrence areas and percentage of various concentration ranges of PM2.5 [10]. The authors derive an assessment of air quality index, using statistical approaches, for each city and determine effects of winter-heating in the two cities - Beijing and Shenyang. However, prediction and future data processing are omitted from this study.

Research recently published in the Journal of Thoracic Disease used ground-based data of particulate matter in conjunction with a suite of remote sensing and meteorological data products [13]. Given that this is published in a medical journal, rather than a computer sciences (and related) journal, the techniques are not discussed explicitly. This is in contrast to recent research analyzing pollution in combination with metrological parameters [33]. Using various machine learning techniques, they have used meteorological data to classify PM2.5 values and also performed regression analysis to find the coherence.

One study has analyzed personal health information, using techniques to ensure data confidentiality [34]. The investigators recorded the personal details as study identity numbers prior to uploading to a cloud-based system for analytics. Important information for urban planning was obtained using data mining techniques on the obtained environmental and behavioral data. A study was conducted to analyze PM2.5 pollution, and its relationship with other meteorological factors such as temperature and humidity [35]. The data was from Chengdu, China and the purpose of the study was to provide insight to improve local air quality. It is argued the results will help authorities to formulate future policies for control of emission in China. In [36], meteorological data and PM2.5 concentration data were obtained during the period January 1, 2013 to December 31, 2013. The spatial distribution of the study area shows that the western part is most seriously affected by PM2.5 pollution. The correlation between PM2.5 concentration data and meteorological data depicts that temperature is negatively correlated with PM2.5 concentration while precipitation is positively correlated with PM2.5.

Daily air pollution predictions of 74 cities in China were studied using a machine learning technique in [37]. Five different classification techniques were adopted with different features groups coming from WRF-Chem models to forecast results. They worked on feature selection technique and the results showed that ANN has the limitation of a low convergence rate. In [38], the authors present a study that proposed an algorithm that showed better predictive ability, with increased R2 and decreased RMSE, when conducted on Hong Kong data. It was shown that Extreme Learning Machine (ELM) performs well in terms of precision, generalization and robustness. No significant differences were found between the prediction accuracies of each model. ELM provided the best performance on indicators related to prediction such as R2 and RMSE etc. The authors achieved 95 RMSE and training time of 0.07s [38].

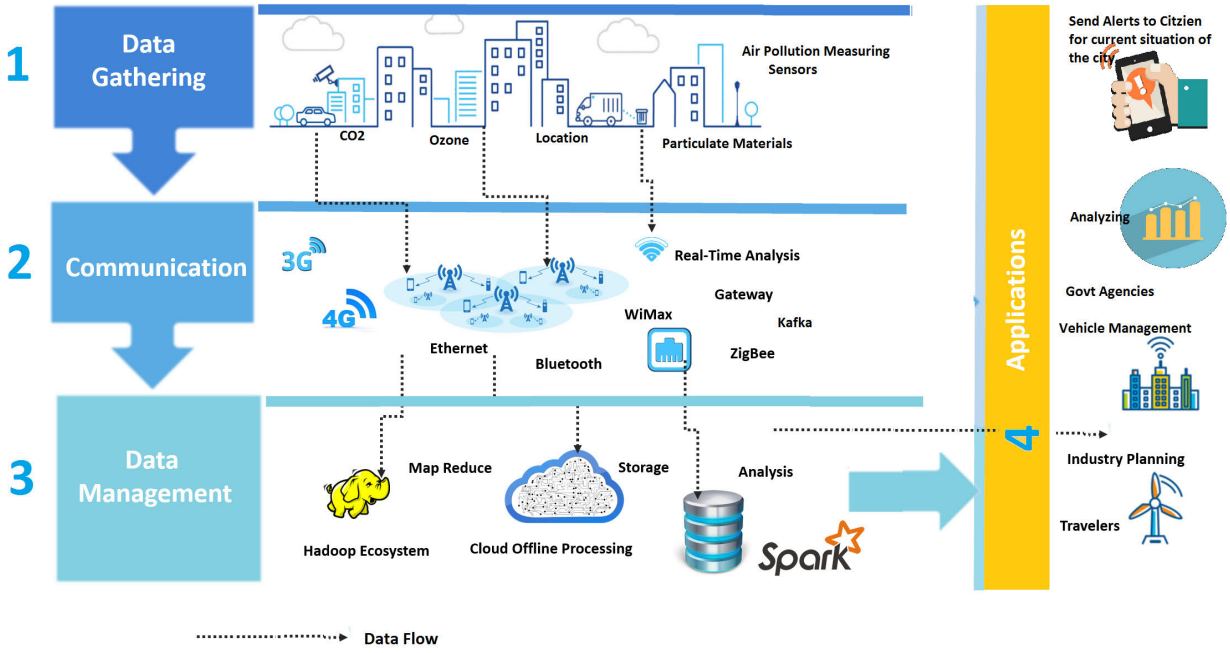


FIGURE 1. Smart city air pollution monitoring architecture.

III. PROPOSED ARCHITECTURE

A. PROPOSED ARCHITECTURE

In this paper, a 4-layer architecture for predicting air pollution has been proposed as shown in Fig 1. These layers are:

- Layer 1 - Data gathering
- Layer 2 - Communication
- Layer 3 - Data Management
- Layer 4 – Application

Different layers in the architecture have different functionalities as described below:

1) DATA GATHERING

This layer gathers data from different heterogeneous devices connected in smart city. Different air pollutants for example ozone, nitrogen dioxide, sulphur dioxide and particulate materials etc. are calculated by sensors deployed at different places in the city. Since lots of data is gathered from different sources, so collection and aggregation takes place here. Data may vary in formats, thus all the pre-processing and initial filtration takes place here. Pre-processing is carried out and the unnecessary information is detected and removed at this layer.

2) COMMUNICATION

This layer is responsible for transferring all the data from data collection layer to further layers. This layer consists of different technologies like 3G, 4G, LTE, Wi-Fi, ZigBee and other communications technologies. All the data transfer from IoT devices to data processing layer takes place here. This layer can also be used for gateways that are efficient enough to process real- time processing. Fog Computing can be used to increase the latency rate. Initial data processing and real time decision can be processed here.

3) DATA MANAGEMENT / STORAGE LAYER

This is the main layer which is responsible for storing and analyzing data. Since real-time processing is required in analysis, so different third party tools can be combined here. For example Spark, VoltDb, Storm etc. can be used for real time processing. This layer is also capable for handling and storing large amount of data in HDFS system. Different other systems can be used for historical data query and analysis. Both In-memory and offline data analysis takes place at this layer. It can also be used for learning through different machine learning algorithms. Predictions and pattern finding also takes place in this layer.

4) APPLICATION

This layer is interface of all the meaning full information. This last layer is connected with the real-time devices; hence events generated are transferred to them. Reports and data in form of charts and dashboards are displayed using this layer. End users of this layer are government agencies who are responsible to monitor pollution. This data is then utilized to take important decisions. This layer can also announce all the pollution related information. It is the interface where people interact to monitor the pollution statistics and make decisions.

IV. METHODOLOGY AND ESTIMATION MODELS

A. REGRESSION TECHNIQUES

1) DECISION TREE REGRESSION

The process of non-parametric supervised learning for the purpose of regression and categorization/classification is termed as Decision Trees (DTs) [39]. The primary objective

of the DTs is to yield a predictive model for the values of the outcome variable with the help of simple decision rules that have been derived from the essential features of the data.

Classification and regression trees (CART) do not calculate the sets of decision rules, however, they are used for the quantitative outcome variable(s) [40]. By using the threshold and characteristics that spawn the greatest amount of information at each node, binary trees are developed by the CART.

2) RANDOM FOREST REGRESSION

The random forest ensures that every tree in the ensemble is generated from a sample with replacement (bootstrapping) from the training set [42]. Moreover, while a tree is being generated, the selected split is the best split in a random subset of features instead of being the best split among all alternatives. As a corollary to this randomness, the bias of the forest may increase a little bit, however, owing to the averaging, its variance is usually reduced, which may compensate the rise in the bias, leading to a superior model on the whole.

3) GRADIENT BOOSTING REGRESSION

The generalization of boosting to an arbitrary differentiable loss function is termed as the GRBT [43]. It constitutes an effective and precise solution that could be utilized for the classification as well as the regression problems. Numerous fields have found the pertinent applications of GRBT including the ecology and the web search ranking.

4) ANN MULTI-LAYER PERCEPTRON REGRESSION

Through training on a dataset, the supervised learning algorithm that learns a function f ():$R_m \rightarrow R_o$ is termed as Multi-layer Perceptron (MLP), where o is the number of output dimensions and m is the number of input dimensions [44], [45]. Given a target y and set of features X=x_1, x_2, x_m, it may come to learn a non-linear function approximator for regression as well as classification. The presence of one or more hidden layers between input and output layers differentiates multi-layer perceptron from the logistic regression.

B. IMPLEMENTATION MODEL

We have also designed an implementation model as shown in Fig 2 which follows the architecture. Data is generated by sensors through different devices located in the city for example, pollution calculating devices. Initially, the data is filtered and processed at the layer 1 to remove all the metadata.

This streaming data is provided to the system on layer 2 where real-time processing events take place. Moreover, this data is stored to Hadoop application of different machine learning algorithm, which help them in making real time decisions.

C. DATA SET DESCRIPTION

We have used data set of different cities of China to evaluate and compare the prediction performance of above mentioned regression techniques for data of different regions and size.

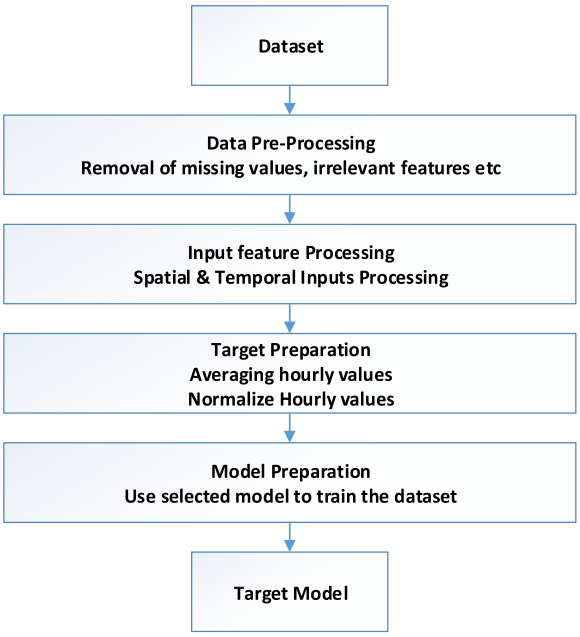


FIGURE 2. Implementation model.

The dataset consists of five cities of China which include Guangzhou, Chengdu, Beijing, Shanghai and Shenyang. The data set period is from 1 Jan 2010 to 31st Dec 2015 [46]. It consists of different meteorological variables and PM2.5 recorded from different locations. Table 2 contains the details of dataset.

Data set contains 15 parameters mentioned as:

No (row number), year, month, day, hour, season, PM2.5 concentration (ug/m^3), Dew Point (Celsius Degree), Temperature (Celsius Degree), Humidity (%), Pressure (hPa), cbwd: Combined wind direction, Iws: Cumulated wind speed (m/s), precipitation: hourly precipitation (mm),Iprec: Cumulated precipitation (mm).

D. BASIC STATISTICAL ANALYSIS

Table 3 shows baseline characteristics of Data set showing mean and standard deviation. Among them, Guangzhou's PM2.5 was 53 ug/m^3 $\pm$ 42 ug/m^3 which was recorded as lowest and PM 2.5 of Beijing and Shenyang were 85.6 ug/m^3 $\pm$ 83.5 ug/m^3 and 78.7 ug/m^3 $\pm$ 75.9 ug/m^3 which were recorded as the highest respectively.

In order to evaluate the relation of PM2.5 with other metrological variables, we have computed correlation matrix of all the five cities. Table 4 shows the co-relation matrix results of PM2.5 with other meteorological variables in Shanghai city only.

During our experiments we have found that PM2.5 has a negative correlation with temperature and also a negative correlation with wind speed which depicts that lowering the temperature increase the amount of PM2.5. In winters, the PM2.5 level increases due to burning of fossil fuels in China. We have showed the co-relation matrix of only

TABLE 1. Limitations of air pollution forecasting techniques.

	Problem Statement	Technique	Strength	Limitations
[8]	Predictive air quality for next 24 hr. in Tehran with efficient way.	Apache Hadoop + Naïve Bayes and Logistic regression	They find Logistic Regression to best estimator	• Logistic Regression can perform well for predicting classes. However, it fails to explain find continuous out comes.
[9]	Analyzing air quality using machine learning	Regularization and Optimization	Minimizes the error rate using Closed Regularization	• Amount of data is small. • Accuracy is discussed but processing time is not mentioned
[10]	Machine Learning techniques for classifying air quality	Decision tree and Naïve Bayes algorithm	91% Accuracy for decision tree.	• Short data amount • Decision tree are not good classifier for time series.
[11]	IoT Sensors AQI Prediction	K-Means	Increase the accuracy as compared to PFCM	• Data size is limited • K-mean poor classifier for time-series
[12]	Low cost AQI Measuring sensors deployment and use machine learning analysis	Random Forest	They decreased the cost	• Data Handling is not discussed • Processing time not discussed
[13]	HISYCOL a hybrid computational intelligence system for combined machine learning	Unsupervised clustering Ensemble ANN	Proposed method increases the computational accuracy.	• Computational cost and processing time is not discussed • Data is in small amount
[14]	Air pollution forecast for short period of time	Co-relation and Neural Network	Improve the accuracy of air pollutant prediction	• Sampling station is considered only on, thus the dataset is very small consisting of few hours.
[15]	Machine Learning and Air Quality Modeling	Randomized Matrix Decompositions & Random Forest Regression	Reduce the computation power consumption compared to GEOS-Chem model and forecast the values for O3	• Sub-sample size for which the training is conducted using Random forest is very small. • The only O3 prediction is mentioned in this paper.
[16]	Low cost sensors and efficiently predicting pollution	Dynamic Neural Network (DNN)	Proposed method decreased the sensors cost , efficiency and prediction accuracy	• Data set is only of two weeks. Training set is short
[17]	Predicting Ozone using deep learning	Deep Neural Network	Increase in accuracy	• Only ozone factor • Lack big data handling
[19]	Machine Learning techniques for AQI in Canada	Multilayer neural networks with nonlinear regression	Proposed algorithm have reduces the error rate.	• Data used is very short amount for multilayer neural network
[20]	Analysis of pollution in China	spatiotemporal deep learning (STDL)	Improve accuracy with comparison to ARMA and Regression	• Data size is small • Linear methods for classification are used
[22]	Monitoring of health is discussed for Big Data	No implementation	Architecture is proposed for big data	• Implementation is not discussed
[23]	Air simulation programming models MapReduce and Spark comparison	K-means Big data Spark	In comparison of MapReduce and Spark, found Spark is fast in processing	• No results related to pollution prediction are presented. • How this spark will help in air pollution prediction is not discussed.
[24]	Air Quality Index Level Prediction Using Random Forest	Random Forest Apache Spark	They have performed the experiments for training the dataset on clusters. And found the system to be more accurate and time efficient.	• Random forest performs well only on classification problems. Thus, this system can only classify the system.
[25]	Analysis of China 16 air pollution data and predicting the air pollution value.	They used PMI based separate IVS scheme for predictors (pollutants) selection. And Ensemble Neural Network for prediction.	Comparison of different regions of China and predicted the value for one day ahead	• The comparison is between regions pollutants variables not within the different points within the city. • Real-time analysis within the city is not discussed

TABLE 2. Dataset description.

City	No. of instances	Parameters	Duration
Guangzhou, Beijing, Chengdu, Shanghai, Shenyang	52584	15	Jan 1st, 2010 to Dec 31st, 2015

Shanghai city, data set of other cities may also be correlated in the same way.

V. SYSTEM EVALUATION

A. TESTBED USED

We have conducted our experimentation on an i5 machine running the Ubuntu 14.04 operating system. The algorithms have been implemented using the Python Programming Language and pre-processing and time series evaluation was conducted using Panda. Machine learning algorithms employed the scikit learn library - an open source machine learning library. To plot graphs the plotly library has been used. Evaluation of performance has been conducted using sklearn metrics. All hyper-parameters are tuned using ten-fold cross validation method and the Grid SearchCV function. This function is capable of making an exhaustive search over

TABLE 3. Baseline characteristics of dataset.

City	Pollutant	Mean ug/m^3	Std ug/m^3	Min ug/m^3	25% ug/m^3	50% ug/m^3	50% ug/m^3	75% ug/m^3
Guangzhou	PM2.5	53.5	42.6	1	26	41	41	67
Chengdu	PM2.5	84.3	59.0	1.0	44.0	68.0	68.0	107.0
Beijing	PM2.5	85.6	83.5	3.0	25	62	62	117
Shanghai	PM2.5	75.5	68.9	1	31	56	56	96
Shenyang	PM2.5	78.7	75.9	2	32	58	58	101

TABLE 4. Co-relation matrix of shanghai city.

	PM2.5	Dew	Pressure	Temp	Wind Speed
PM2.5	1	-0.269954	0.192088	-0.255804	-0.196961
Dew	-0.269954	1	- 0.847333	0.87451	-0.040213
Pressure	0.192088	-0.847333	1	-0.835328	0.0341899
Temperature	-0.255804	0.87451	-0.835328	1	-0.0578183
Wind Speed	-0.196961	-0.040213	0.0341899	-0.0578183	1

TABLE 5. Comparison of regression techniques.

	GBRT		MLP		DTR		RFTR	
City	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Shanghai	17	0.07	13.84	0.03	21.74	0.09	17.68	0.05
Guangzhou	27.7	0.17	12.2	0.045	14.36	0.06	13.1	0.05
Chengdu	14.47	0.113	9.8	0.108	11.51	0.086	10.5	0.0828
Shenyang	17	0.102	13.65	0.062	21.3	0.0872	17.68	0.59
Beijing	29.30	0.0866	21.79	0.0806	19.03	0.07	16.92	0.0725

TABLE 6. Processing time in secs.

	GBRT		MLP		DTR		RFTR	
City	With spark	Without Spark	With spark	Without Spark	With spark	Without Spark	With spark	Without Spark
Shanghai	13	9.52	7.9	9.23	0.16	0.14	0.83	0.87
Guangzhou	10.6	7.7	9.6	9.5	0.12	0.12	0.95	0.8
Chengdu	11	22.35	8.	17	0.35	0.35	2.2	2.8
Shenyang	11	9.3	5.9	6.6	0.13	0.22	0.75	1.62
Beijing	14	10	9.1	9.2	0.14	0.12	0.70	0.8

specified parameter values, defined by the user. In order to evaluate the performance on Spark, we have used spark learn library provided by DataBricks.

B. EVALUATION CRITERIA

We evaluate performance using the standard measures, MAE and RMSE.

1) MEAN ABSOLUTE ERROR (MAE)

Mean absolute error is used to measure the average magnitude of the errors in a set of data values (predictions), without any consideration of direction [48]. In a test sample, MAE is the average of the absolute differences between actual and prediction observations. It is calculated as in Eq (2):

$$MAE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \acute{y}_j) \tag{2}$$

where,

n = Number of observations

y_j = Actual value

$\acute{y}_j$ = Predicted value

2) ROOT MEAN SQUARED ERROR (RMSE)

The RMSE is also used to calculate the average magnitude of the error. It is obtained by taking the average of squared differences between actual vs predicted values and taking the

(a) RMSE for different regression techniques.

(b) MAE for different regression techniques.

FIGURE 3. (a) RMSE for different regression techniques. (b) MAE for different regression techniques.

(a) RMSE for different regression techniques.

(b) MAE for different regression technique.

FIGURE 4. (a) RMSE for different regression techniques. (b) MAE for different regression techniques.

square root of the final result [48]. It is calculated as:

$$RMSE = \sqrt{\frac{1}{n}(\sum_{j=1}^{n}(y_j - \acute{y}_j)^2)} \tag{3}$$

In order to compare the datasets we normalize the RMSE as follows [49]:

$$Normalized\ RMSE = \frac{RMSE}{y_{max} - y_{min}} \tag{4}$$

where,

y_{max} = Maximum value of data set

y_{min} = Minimum value of data set

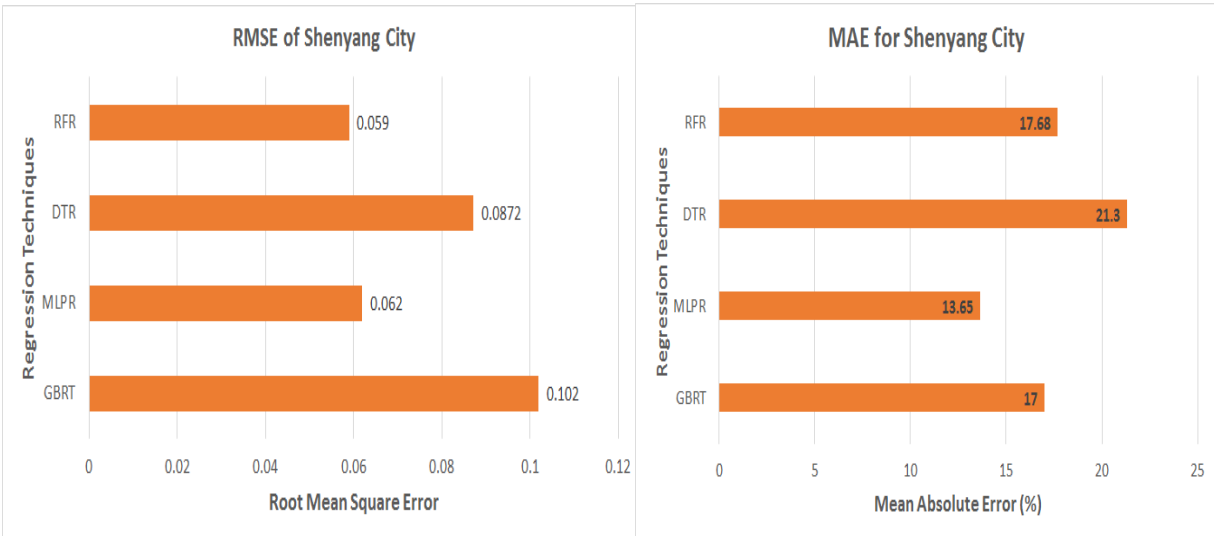
VI. RESULTS AND DISCUSSION

A. BEIJING CITY

Beijing city has the highest reported values of PM2.5 in China. We have applied the aforementioned regression techniques on Beijing city data set and predicted the maximum and minimum values of pollution in the city and compared it to actual values. To do this we used the data from 1st Jan 2010 to 21 Dec 2015 for training, and predicted for the next week on 22 Dec 2015 to 31 Dec 2015. Prediction results for Gradient Boosting Regression (GBR), Decision Tree Regression (DTR), Multi-layer Perceptron Regression (MLP) and Random Forest Regression (RFR) are presented. The poor perfomrance is supported by considering the error rate calculation shown in Figure 4, showing the RMSE is 0.08 after normalizing, and a MAE of 29.3%, much higher than other techniques. The RMSE and MAE obtained by using MLPR is also high. Comparatively, DTR and RFTR have performed much better for identifying the peak values on this dataset. The RMSE achieved using Random Forest regression is 0.0725 after normalizing and gives an MAE of 16%, much better than both Gradient Boosting regression and Decision Tree regression.

B. SHANGHAI CITY

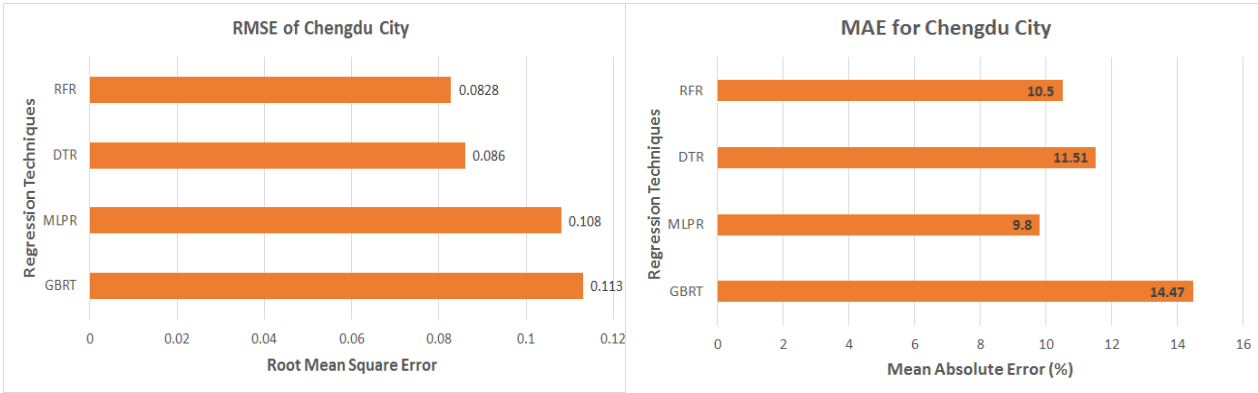
For the city of Shanghai, the data was trained on pollution values obtained from 1st Jan 2010 to 21 Dec 2015 and predictions were conducted for the next week on 22 Dec 2015



(a) RMSE for different regression techniques.

(b) MAE for different regression techniques.

FIGURE 5. (a) RMSE for different regression techniques. (b) MAE for different regression techniques.



(a) RMSE for different regression techniques .

(b) MAE for different regression technique.

FIGURE 6. (a) RMsE for different regression techniques. (b) MAE for different regression techniques.

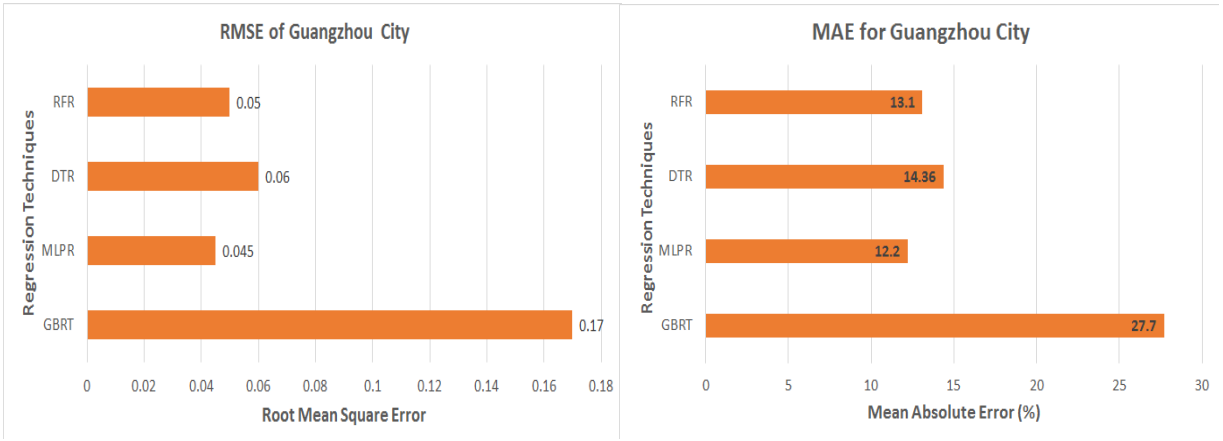
to 31 Dec 2015. The results indicate that both Decision Tree and Gradient Boosting regression were unable to accurately predict the maximum values. The MAE and RMSE for all methods are presented in Figure 6. The MAE obtained for DTR and GBR was 22% and 17% respectively, while the RMSE was 0.09 and 0.07 respectively. These values are higher than the other two techniques. MLP performed much better not only in identifying the peak values but also achieved the lowest RMSE (0.03) and the lowest MAE (13.84%). Random Forest regression performed almost as well as MLP, resulting in an RMSE of 0.05 and an MAE of 17%.

C. SHENYANG CITY

Regression analysis has been performed on Shenyang city's data set and predictions obtained. The MAE and RMSE for each of the techniques is presented in Figure 8. From the graphs, it is clear that MLPR and RFR achieved the best results for predicting the pollution levels in terms of RMSE. MLPR resulted in a RMSE of 0.062, only slightly poorer than the 0.059 obtained using Random Forests. The RMSE results for GBR and DTR are similar (0.102 and 0.0872) and much worse than the other two techniques. In terms of MAE, MLPR performs far better than all other methods, with a result of 13.65%, far superior to the results of Decision Tree regression which achieved 21.3%.

D. GUANGZHOU CITY

Regression analysis was also performed on data obtained from the city of Guangzhou and predictions obtained. Again, the data for pollution values from 1st Jan 2010 to 21Dec 2015 was used to train the model and predictions calculated for the next week on 22 Dec 2015 to 31 Dec 2015. The data set is smooth with only one extraordinary peak at the start on December 22, 2015. Observing Figure 10, it is evident that GBR does not perform well in predicting the pollution values in this data. Neither it could identify the peak values



(a) RMSE for different regression techniques . (b) MAE for different regression techniques.

FIGURE 7. (a) RMSE for different regression techniques. (b) MAE for different regression techniques.

nor accurately predict pollution as a whole. The MAE for Gradient Boosting regression was 27.8% while the RMSE was 0.17, both much higher than all other methods, and indeed poor performance when compared to all methods on other cities. MLP and Random Forest regression were again the best performing methods. MLP achieved an MAE and RMSE of 12.2% and 0.045 respectively, while Random Forest regression performed only slightly worse with an MAE of 13.1% and an RMSE of 0.05. The Decision Tree method performed much better than GBR, ond only slightly worse than Random Forests, with an MAE of 14.36% and an RMSE of 0.06.

E. CHENGDU CITY

Predictions for the city of Chengdu, obtained by applying the various regression techniques, are presented in figure 12. The MAE for Gradient Boosting regression was 14% and the RMSE achieved was 0.113. Random Forest regression performed the best in accurately predicting the results. It achieved an RMSE of 0.08 and an MAE of 10.5 %. MLP can be seen to be the second best technique with an MAE of 9.8 % and an RMSE of 0.108.

F. DISCUSSION

In the previous section, results of all the four regression techniques on different data sets have been presented. The relationship between PM2.5 levels and meteorological parameters has also been calculated and presented in preceding section. It was found that PM2.5 has a negative correlation with temperature and also a negative correlation with wind speed which implies that lower the temperature of the city, the higher the amount of PM 2.5 concentration in the city will be. This can be explained by considering the fact that the lower temperature causes the density of air to increase, thus increasing the potential of more suspended particles in the air. Dense air stays in the atmosphere for a longer time than light air, and so the concentration of PM 2.5 is recorded as higher at low temperatures. Similarly, when the wind speed is high, PM 2.5 concentration is lower in the city. This is due to the fact that higher wind speed causes the particles to be washed away from the atmosphere of a particular location where sensors are located. In winter, we observe the PM2.5 level in China increases, and this is likely to be due to the burning of fossil fuels.

Of the four techniques tested, Decision Tree regression has the advantage of being simple to understand and implement. Furthermore, the processing time of Decision Trees compares favorably to other techniques. However, it must be recognized the performance does not compare well with other models – the mean absolute error is between 8% to 21%, while RMSE is between 0.06 to 0.24.

Random Forest regression is an ensemble method of multiple trees. It reduces the overfitting of single trees by combining several trees. This model was able to identify the peak values. Moreover the processing time was also less than other models. The MAE for the different data sets ranged from 6% to 18 % while the RMSE ranged from 0.05 to 0.18. Random Forest regression also performed well after hyper-parameter tuning.

For the dataset, with a large amount of historic data, Random Forest regression performs the best among all four regression algorithms. We have validated our approach with field trials and have shown the performance comparison against different algorithms.

We have validated our approach with field trials and have shown the estimation performances between different algorithms.

The error rate computed for different data sets is presented in table V. It is evident that the Gradient Boosting regression method has the highest error rate compared to other three regression techniques, for most of the data sets. The Random Forest regression technique achieves the lowest mean absolute error and RMSE.

Table VI presents the time, in seconds, taken by the different regression models for learning and prediction on test data. For evaluation, we have initially run the algorithms without setting the parameters. The lowest processing time was taken by Decision Tree regression and Random Forest regression

when compared to Gradient Boosting regression and Multi Linear Perception. We then performed hyper parameter tuning on a Spark single node and calculated the results. It was found that Random Forest regression has performed best overall in terms of error time and processing rate.

VII. CONCLUSION

In this work, we have analyzed and compared four existing schemes for solving the air pollution prediction issue. The techniques were Decision Tree regression, Random Forest regression, Multi-Layer Perceptron regression and Gradient Boosting regression. We have compared the techniques with respect to error rate and processing time. The simulation results show that Random Forest regression was the best technique, performing well for pollution prediction for data sets of varying size and location and having different characteristics. Its processing time was found mush lower than the gradient boosting and multi-layer perceptron algorithms. Furthermore, its error rate was found to be the least among all four techniques. Although the processing time of Decision Trees was found to be the lowest, its error rate remained higher than most techniques and it was not able to properly identify the data peaks in almost all data sets. In comparison, Random Forest regression took less time than the other two techniques, and just higher than Decision Trees; it also performed well in identifying the peak values and accurately predicted the data with a low error rate. Therefore, we can deduce the conclusion that Random Forest regression was the best technique among the four algorithms considered. Gradient boosting regression has performed the worst of all algorithms, as it has achieved highest processing time in almost all data sets and has given a very high error rate in most cases.

VIII. FUTURE WORK

In the future, we aim to investigate the performance of these techniques on the multi-core environment of Spark. Furthermore, we also intend to investigate the other factors effecting the air pollution.

REFERENCES

[1] *7 Million Premature Deaths Annually Linked to Air Pollution*. Accessed: Apr. 27, 2019. [Online]. Available: https://www.who.int/phe/eNews_63.pdf
[2] F. C. Moore, "Climate change and air pollution: Exploring the synergies and potential for mitigation in industrializing countries," *Sustainability*, vol. 1, no. 1, pp. 43–54, 2009.
[3] H.-P. Hsieh, S.-D. Lin, and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 437–446.
[4] M. Johnson, V. Isakov, J. S. Touma, S. Mukerjee, and H. Özkaynak, "Evaluation of land-use regression models used to predict air quality concentrations in an urban area," *Atmos. Environ.*, vol. 44, no. 30, pp. 3660–3668, 2010.
[5] C. Malalgoda, D. Amaratunga, and R. Haigh, "Local governments and disaster risk reduction: A conceptual framework," in *Proc. 6th Int. Building Resilience Conf., Building Resilience Address Unexpected*. Palmerston North, New Zealand: Massey Univ., 2016, pp. 699–709.
[6] M.-A. Kioumourtzoglou, J. D. Schwartz, M. G. Weisskopf, S. J. Melly, Y. Wang, F. Dominici, and A. Zanobetti, "Long-term $PM_{2.5}$ exposure and neurological hospital admissions in the northeastern United States," *Environ. Health Perspect.*, vol. 124, no. 1, pp. 23–29, 2015.
[7] *World Health Organization, and UNAIDS Air Quality Guidelines: Global Update 2005*, World Health Org., Geneva, Switzerland, 2006.
[8] K.-H. Kim, E. Kabir, and S. Kabir, "A review on the human health impact of airborne particulate matter," *Environ. Int.*, vol. 74, pp. 136–143, Jan. 2015.
[9] U. A. Hvidtfeldt, M. Ketzel, M. Sørensen, O. Hertel, J. Khan, J. Brandt, and O. Raaschou-Nielsen, "Evaluation of the Danish AirGIS air pollution modeling system against measured concentrations of $PM_{2.5}$, PM_{10}, and black carbon," *Environ. Epidemiol.*, vol. 2, no. 2, p. e014, 2018.
[10] A. J. Cohen *et al.*, "Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the global burden of diseases study 2015," *Lancet*, vol. 389, pp. 1907–1918, May 2017.
[11] En.wikipedia.org. (2018). *Airqualityindex*. Accessed: Dec. 2, 2018. [Online]. Available: https:==en:wikipedia:org=wiki=Airqualityindex
[12] W. Yi, K. Lo, T. Mak, K. Leung, Y. Leung, and M. Meng, "A survey of wireless sensor network based air pollution monitoring systems," *Sensors*, vol. 15, no. 12, pp. 31392–31427, 2015.
[13] 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, pp. 69–74, Jan. 2016.
[14] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the Internet of Things using big data analytics," *Comput. Netw.*, vol. 101, pp. 63–80, 2016.
[15] M. Asgari, M. Farnaghi, and Z. Ghaemi, "Predictive mapping of urban air pollution using apache spark on a hadoop cluster," in *Proc. Int. Conf. Cloud Big Data Comput.*, 2017, pp. 89–93.
[16] D. Zhu, C. Cai, T. Yang, and X. Zhou, "A machine learning approach for air quality prediction: Model regularization and optimization," *Big Data Cogn. Comput.*, vol. 2, no. 1, p. 5, 2018.
[17] R. W. Gore and D. S. Deshpande, "An approach for classification of health risks based on air quality levels," in *Proc. Int. Conf. Intell. Syst. Inf. Manage. (ICISIM)*, Oct. 2017, pp. 58–61.
[18] G. R. Kingsy, R. Manimegalai, D. M. S. Geetha, S. Rajathi, K. Usha, and B. N. Raabiathul, "Air pollution analysis using enhanced K-means clustering algorithm for real time sensor data," in *Proc. IEEE Region Conf. (TENCON)*, Nov. 2016, pp. 1945–1949.
[19] N. Zimmerman, A. A. Presto, S. P. N. Kumar, J. Gu, A. Hauryliuk, E. S. Robinson, A. L. Robinson, and R. Subramanian, "Closing the gap on lower cost air quality monitoring: Machine learning calibration models to improve low-cost sensor performance," *Atmos. Meas. Tech. Discuss*, vol. 2017, pp. 1–36, 2017. doi: 10.5194/amt-2017-260.
[20] I. Bougoudis, K. Demertzis, and L. Iliadis, "HISYCOL a hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in Athens," *Neural Comput. Appl.*, vol. 27, no. 5, pp. 1191–1206, 2016.
[21] C. Yan, S. Xu, Y. Huang, Y. Huang, and Z. Zhang, "Two-phase neural network model for pollution concentrations forecasting," in *Proc. 5th Int. Conf. Adv. Cloud Big Data (CBD)*, 2017, pp. 385–390.
[22] C. A. Keller, M. J. Evans, J. N. Kutz, and S. Pawson, "Machine learning and air quality modeling," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 4570–4576.
[23] E. Esposito, S. De Vito, M. Salvato, V. Bright, R. L. Jones, and O. Popoola, "Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems," *Sens. Actuators B, Chem.*, vol. 231, pp. 701–713, Aug. 2016.
[24] O. A. Ghoneim, H. Doreswamy, and B. R. Manjunatha, "Forecasting of ozone concentration in smart city using deep learning," in *Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI)*, 2017, pp. 1320–1326.
[25] A. B. Ishak, M. B. Daoud, and A. Trabelsi, "Ozone concentration forecasting using statistical learning approaches," *J. Mater. Environ. Sci.*, vol. 8, no. 12, pp. 4532–4543, 2017.
[26] H. Peng, A. R. Lima, A. Teakles, J. Jin, A. J. Cannon, and W. W. Hsieh, "Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods," *Air Qual., Atmos. Health*, vol. 10, no. 2, pp. 195–211, 2017.
[27] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environ. Sci. Pollut. Res.*, vol. 23, no. 22, pp. 22408–22417, 2016.
[28] T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, "Promises and challenges of big data computing in health sciences," *Big Data Res.*, vol. 2, no. 1, pp. 2–11, 2015.

[29] H. Ayyalasomayajula, E. Gabriel, P. Lindner, and D. Price, "Air quality simulations using big data programming models," in *Proc. IEEE 2nd Int. Conf. Big Data Comput. Service Appl.*, Mar./Apr. 2016, pp. 182–184.
[30] C. Zhang and D. Yuan, "Fast fine-grained air quality index level prediction using random forest algorithm on cluster computing of spark," in *Proc. IEEE 12th Int. Conf. Ubiquitous Intell. Comput., IEEE 12th Int. Conf. Adv. Trusted Comput., IEEE 15th Int. Conf. Scalable Comput. Commun.*, vol. 20, Aug. 2015, pp. 929–934.
[31] S. Chen, G. Kan, J. Li, K. Liang, and Y. Hong, "Investigating China's urban air quality using big data, information theory, and machine learning," *Polish J. Environ. Stud.*, vol. 27, no. 2, pp. 565–578, 2018.
[32] Y. S. Chang, K.-M. Lin, Y.-T. Tsai, Y.-R. Zeng, and C.-X. Hung, "Big data platform for air quality analysis and prediction," in *Proc. 27th Wireless Opt. Commun. Conf. (WOCC)*, 2018, pp. 1–3.
[33] J. K. Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, "Modeling $PM_{2.5}$ urban pollution using machine learning and selected meteorological parameters," *J. Elect. Comput. Eng.*, vol. 2017, Jun. 2017, Art. no. 5106045.
[34] K.-F. Ho, H. W. Hirai, Y.-H. Kuo, H. M. Meng, and K. K. F. Tsoi, "Indoor air monitoring platform and personal health reporting system: Big data analytics for public health research," in *Proc. IEEE Int. Congr. Big Data*, vol. 2, Jun./Jul. 2015, pp. 309–312.
[35] Y. Li, Q. Chen, H. Zhao, L. Wang, and R. Tao, "Variations in PM_{10}, $PM_{2.5}$ and $PM_{1.0}$ in an urban area of the Sichuan Basin and their relation to meteorological factors," *Atmosphere*, vol. 6, no. 1, pp. 150–163, 2015.
[36] J. Wang and S. Ogawa, "Effects of meteorological conditions on $PM_{2.5}$ concentrations in Nagasaki, Japan," *Int. J. Environ. Res. Public Health*, vol. 12, no. 8, pp. 9089–9101, 2015.
[37] X. Xi, Z. Wei, R. Xiaoguang, W. Yijie, B. Xinxin, Y. Wenjun, and D. Jin, "A comprehensive evaluation of air pollution prediction improvement by a machine learning method," in *Proc. IEEE Int. Conf. Service Oper. Logistics, Inform. (SOLI)*, Nov. 2015, pp. 176–181.
[38] J. Zhang and W. Ding, "Prediction of air pollutants concentration based on an extreme learning machine: The case of Hong Kong," *Int. J. Environ. Res. Public Health*, vol. 14, no. 2, p. 114, 2017.
[39] En.Wikipedia.org. (2018). *Airqualityindex*. [Online]. Available: https: ==en:wikipedia:org=wiki=Airqualityindex
[40] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, vol. 37, no. 15. Belmont, CA, USA: Wadsworth Int. Group, 1984, pp. 237–251.
[41] (2018). *DecisionTreesjscikit learn0:20:1documentation*. [Online]. Available: https: scikit-learn.org/stable/modules/tree.html
[42] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[43] G. Ridgeway. (2007). *Generalized Boosted Models: A Guide to the gbm Package. R Package Vignette*. [Online]. Available: http://CRAN.R-project.org/package=gbm
[44] P. Patiño, "Gradient boosted regression trees in scikit-learn," PyData, London, U.K., 2014.
[45] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cogn. Model.* vol. 5, no. 3, p. 1, 1988.
[46] X. Liang, S. Li, S. Zhang, H. Huang, and S. X. Chen, "$PM_{2.5}$ data reliability, consistency, and air quality assessment in five Chinese cities," *J. Geophys. Res., Atmos.*, vol. 121, no. 17, pp. 10–220, 2016.
[47] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, pp. 1247–1250, Jun. 2014. doi: 10.5194/gmd-7-1247-2014.
[48] (Aug. 28, 2018). *In Root Mean Squaredeviation*. [Online]. Available: https://en.wikipedia.org/wiki/Root-mean-square_deviation

SABA AMEER is currently pursuing the master's degree in computer science with the Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan. Her research interests include big data, the IoT, machine learning, digital image processing, and artificial intelligence and algorithms.

MUNAM ALI SHAH received the B.Sc. and M.Sc. degrees in computer science from the University of Peshawar, Pakistan, in 2001 and 2003, respectively, the M.S. degree in security technologies and applications from the University of Surrey, U.K., in 2010, and the Ph.D. degree from the University of Bedfordshire, U.K., in 2013. Since 2004, he has been a Lecturer with the Department of Computer Science, COMSATS Institute of Information Technology, Islamabad, Pakistan. He is the author of more than 50 research articles published in international conferences and journals. His research interests include MAC protocol design, QoS, and security issues in wireless communication systems. He received the Best Paper Award at the International Conference on Automation and Computing, in 2012.

ABID KHAN received the Ph.D. degree from the Harbin Institute of Technology. He is currently an Assistant Professor of computer science with COMSATS University Islamabad, Islamabad. His research interests include security and privacy of cloud computing (outsourced storage and computation), security protocols, digital watermarking, secure provenance, and information systems.

HOUBING SONG (M'12–SM'14) received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 2012. In 2007, he was an Engineering Research Associate with the Texas A&M Transportation Institute. In 2017, he joined the Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA, where he is currently an Assistant Professor and the Director of the Security and Optimization for Networked Globe Laboratory (SONG Lab). He served on the Faculty of West Virginia University, from 2012 to 2017. He is the Editor of four books: *Cyber-Physical Systems: Foundations, Principles and Applications* (Boston, MA: Academic Press, 2016), *Industrial Internet of Things: Cybermanufacturing Systems* (Cham, Switzerland: Springer, 2016), *Smart Cities: Foundations, Principles, and Applications* (Hoboken, NJ: Wiley, 2017), and *Security and Privacy in Cyber-Physical Systems: Foundations, Principles, and Applications* (Chichester, U.K.: Wiley-IEEE Press, 2017). He is the author of more than 100 articles. His research interests include cyber-physical systems, cybersecurity and privacy, the Internet of Things, edge computing, big data analytics, unmanned aircraft systems, connected vehicle, smart and connected health, and wireless communications and networking. He also serves as an Associate Technical Editor for the *IEEE Communications Magazine*.

CARSTEN MAPLE is currently the Director of research in cyber security and a Professor of cyber systems engineering with the Cyber Security Centre, University of Warwick, where he leads the GCHQ-EPSRC Academic Centre of Excellence in Cyber Security Research. He is the Privacy and Trust Stream Lead and leads the project constellation in transport and mobility with PETRAS, the U.K. Research Hub for Cyber Security of the Internet of Things. He is a principal or co-investigator of a number of projects in cyber security. He has published over 200 peer-reviewed articles and has provided evidence and advice to governments and organisations across the world including being a high-level Scientific Advisor for cyber security at the European Commission. He is currently or has recently been supported by a range of sponsors including the EPSRC, EU, DSTL, South Korean Research Agency, Innovate U.K., and private companies. He is a member of various boards and expert groups. He is a Fellow of the Alan Turing Institute. He is the Immediate Past Chair of the Council of Professors and Heads of Computing in U.K.

SAIF UL ISLAM received the Ph.D. degree in computer science from the University Toulouse III Paul Sabatier, France, in 2015. He was an Assistant Professor with COMSATS University Islamabad, Islamabad, Pakistan, for three years. He has been a part of the European Union funded research projects during his Ph.D. degree. He was a focal person of a research team at COMSATS working in O2 project in collaboration with CERN, Switzerland. He is currently an Assistant Professor with the Department of Computer Science, Dr. A. Q. Khan Institute of Computer Sciences and Information Technology, Rawalpindi, Pakistan. His research interests include resource and energy management in large-scale distributed systems [edge/fog, cloud, and content distribution network (CDN)] and the Internet of Things (IoT).

MUHAMMAD NABEEL ASGHAR received the Ph.D. degree from the University of Bedfordshire, U.K., with a focus on modeling for machine vision, specifically digital imagery, and its wide spread application in all vistas of life. He is currently an Assistant Professor with the Department of Computer Science, Bahauddin Zakariya University, Pakistan. He has been investigating machine learning approaches for analyzing video content ranging from broadcast news, sports, surveillance, personal videos, entertainment movies, and similar domains, which is increasing exponentially in quantity and it is becoming a challenge to retrieve content of interest from the corpora, and also on their applications, such as information extraction and retrieval. His recent work is concerned with multimedia, incorporating text, audio, and visual processing into one dynamic novel frame work. His research interests include information retrieval, computer graphics, computer vision, image processing and visualization, graphics modeling and simulation, CR MAC protocol design, the Internet of Things, and security issues in wireless communication systems.