IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

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], Carsten Maple[3], Saif ul Islam[4], Muhammad Nabeel Asghar[5]

[1] Department of Computer Science, COMSATS University Islamabad 44550, Pakistan (sabaameer13@gmail.com, mshah@comsats.edu.pk, abidkhan@comsats.edu.pk)
[2]Department of Electrical Engineering and Computer Science, Embry-Riddle Aeronautical University, 600 South Clyde Morris Blvd. Daytona Beach, FL 32114 USA (h.song@ieee.org)
[3] WMG, University of Warwick, Coventry, UK. CV4 7AL (cm@warwick.ac.uk)
[4]Department of Computer Science, Dr. A. Q. Khan Institute of Computer Science and Information Technology, Rawalpindi 47000, Pakistan.
[5] Department of Computer Science, Bahauddin Zakariya University, Multan 60800, Pakistan. (nabeel.asghar@bzu.edu.pk)

Corresponding author: Saif ul Islam (e-mail: saiflu2004@gmail.com).

"This work is supported by the Alan Turing Institute under EPSRC grant EP/N510129/1."

ABSTRACT Dealing with air pollution is one of the major environmental challenges in a smart city environment. Real-time monitoring of pollution data enables the metropolitans to analyze the current traffic situation of the city and take their decisions accordingly. Deployment of the Internet of things (IoT) 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 often required to have a better understanding of their processing time for multiple datasets. In this paper, we have performed the pollution prediction using four advanced regression techniques and have presented a comparative study to analyze the best model for accurately predicting the air quality with reference to data size and processing time. We have used Apache Spark for conducting experiments and performing pollution estimation using multiple available data sets. Mean Absolute error (MAE) and Root Mean Square Error (RMSE) have been used as evaluation criteria for the comparison of regression models. Furthermore, the processing time of each technique by standalone learning and by fitting the hyperparameter tuning on Apache Spark has also been calculated to find the best-fitted model in terms of processing time and least error rate.

INDEX TERMS IoT, Smart City, Air Quality Index (AQI), Data Mining, Apache Spark

I. INTRODUCTION

Air pollution is one of the main detriments to human health. According to World Health Organization, 7 million people are at health risk due to air pollution [1]. It is a leading risk factor for majority of health problems like asthma, skin infections, heart issues, throat and eye diseases, bronchitis, lungs cancer and respiratory system's diseases. Besides the health problems related to air pollution, it also poses a serious threat to our planet. Pollution emissions from the sources like vehicles and industry is the underlying cause of greenhouse effect, CO2 emissions are amongst the foremost contributors to the greenhouse phenomenon [2]. Climate change has been widely discussed at the global forums and has remained a burning issue for the world since last two decades as a result of increased smog and ozone damage.

Air pollution prediction problem has been solved in the past using statistical linear methods but these techniques are poor estimator for air pollution prediction due to complexity and variation in time-series data [3] [4]. Over the last 60 years, many machine learning techniques have been developed for handling the complex methods.

A. SMART CITY AND AIR POLLUTION:

A smart city is an urban municipality that utilizes information and communication technologies (ICT) to provide

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

better health, transport and energy related facilities to the citizens and enables the government to make efficient use of 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 population has grown rapidly due to industrialization and migration of the 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, needs of transportation and energy are also increased thus adding more industry and vehicles to the cities. This in turn increases the sources of pollution emissions which is a big concern for the authorities that intend to provide better lifestyle to its inhabitants by controlling pollution related diseases. Thus, coping with air pollution is one of the fundamental challenges in urban areas and a smart city.

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 while these particulate matters may also be formed by secondary sources like 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 longer in the air thus increasing the risk of its inhale by the human beings. Particulate matter 2.5 have much adverse effect on the human health as compared to the other pollution emissions. These particles can easily enter into the respiratory system through inhalation process where can badly effect the lungs and the breathing phenomenon. Moreover, it has the potential to cause cardiovascular diseases in people of almost every age group with children and people above 65 more 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 some lungs or heart disease need special precautionary measures in the polluted environment [9].

Effects of PM2.5 were analyzed over the last 25 years. It was estimated that approximately 4.2 million people have died due to long term exposure to PM 2.5 containing atmosphere while additional 0.25 million deaths have occurred because of exposure to ozone. 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, number of deaths due to air pollution have increased, especially in China and India [10]. House hold 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 its above mentioned adverse effects, PM 2.5 concentration is actively monitored by the municipalities around the globe and air quality index (AQI) is calculated on the basis of it. Air quality index is a function of pollutant concentration. It is a dimensionless number, different values of which exhibit different quantities of air pollution. If PM2.5 concentration is lower, it reflects lower value of AQI and healthy air while higher concentration refers to higher AQI and unhealthy to dangerous air depending upon the value. According to EPA, AQI is calculated from concentration of pollutant by the following method [11].

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

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}

Each country has its own standards of defining the quality of air based on Air Quality Index. Individual AQI is calculated for each of the separate pollutant concentration and 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. AQI is calculated and reported on hourly basis at most places to convey estimates of air pollution to general public. When AQI is high, people with heart and respiratory diseases may avoid outdoor activities or may use mask to protect them.

In the era of technology, where every object is connected has raised the concept of Internet of Things (IoT). New systems have been proposed which are based on IoT sensors [14]. Data gathered from sensors can play a vital role in helping the cities manage and measure air quality. With the help of sensors generating data, the smart city decisions have

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

been made much faster and easier but the processing of data brings its own challenges.

A big challenge in handling a smart city's information is to make it more efficient and reliable in addition to being rapidly analyzed. False alarm can lead to wrong decisions which can turn to be very dangerous. To ensure that the speed and processing of communication sources are made robust, prediction using artificial intelligence and use of right data to take appropriate decision is very important. Moreover, to analyze the pollution data, machine learning predictive algorithms are required to deliver the right information at the right place with low latency rate.

In this paper, we have performed pollution prediction with four different regression techniques mentioned in Section III. We have presented the comparative analysis of these techniques, based on evaluation criteria i.e. MAE and RMSE in order to find the best predictive model for pollution estimation. We have considered the multiple cities air pollution in order to find the best accurate model.

Furthermore, considering that the smart city's real-time data processing requires the process to be efficient, we have analyzed the processing time of these techniques by standalone learning and by fitting the hyper parameter tuning on Apache Spark. In this research, we have proposed the best fitted model in terms of processing time and least error rate.

The rest of paper is organized as follows. Section II consists of review of related studies. Section III describes the proposed architecture and estimation models. 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 conclusion & future work is presented in Section VII and VIII respectively.

II. LITERATURE REVIEW

In the last few decades, many machine learnings techniques have been proposed for solving air pollution prediction problems.

Authors in [15] have analyzed the urban pollution and mapped them according to geographical areas. They have analyzed the data of 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 as compared to other machine learning algorithms for better classifying the unknown classes of air quality. The paper presents good results in terms of Apache Spark processing time but the machine learning algorithm are not appropriate for real-time time series prediction. In [16], author addressed the prediction of air pollutants e.g. ozone, particle matter (PM2.5) and sulphur dioxide using optimization and regularization techniques to predict the next day air pollutant values. They have predicted the values of data set of two stations. One station predicts the values for O3 and SO2 and other for O3 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. Linear Regression Model fails forecasting or handling unforeseen events. Moreover, the data of only two stations is used, which is very limited.

Classification of air quality index based and its effect on health was studied in [17]. They implemented Decision tree and Naive Bayes J48 for classification and there results showed that decision tree algorithm perform with 91.9978%. There are many limitations to this research, e.g. dataset used was limited. Moreover, the decision tree cannot perform well on continuous variable and issue with overfitting. Another research for classifying of air quality index was proposed using K-mean algorithm, again the data set was limited. K-Mean technique unfits for predicting the future values [18].

Real-time Affordable Multi-Pollutant (RAMP) is a low-cost pollution monitoring system which is proposed in [19] for measuring pollutants. They have reduced the sensors cost and performed Random forest for prediction of future values. However, the data set consists of only 2 weeks which is very small. Random forest algorithm can have overfitting problem for small dataset.

Ilias Bougoudis, proposed a hybrid computational intelligence system for combined machine learning (HISYCOL) [20] for finding correlation of air pollutants with weather to find real cause pollutant and used. They gather data from wider Attica area. ANN and Random forest as an ensembles' learning is used. They claim the accuracy is increased but the feed forward neural network fails in accurately predicting the continuous values. Moreover, the training data is very limited in this research. Neural network with two phase's concept such that initially train meteorological parameters and then analyze it with air pollutants increases their accuracy [21]. They had considered only single stations and few hours' data, for small data set neural network faces over fitting issue.

Limitation of computational models for air quality is discussed by NASA Goddard Space Flight G'lIter [22]. They proposed machine learning techniques for forecasting the O3 in the different countries. They used sparse sampling, randomized matrix decompositions as a pre-processing to reduce the dimensionality of the data. They have used random forest regression technique for forecasting next 10 days. They only took one pollutant O3 for future prediction and the data subsample size is small. Air Pollution prediction using machine learning Dynamic Neural Network (DNN) approach was carried on data generated by their low cost sensors [23].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

They conducted experiments on two weeks data. Although learning has been carried out on limited amount of data. This solution is unable to forecast for smart city pollution challenges.

Prediction of Ozone Concentration in Smart City using Deep Learning is proposed in [24]. They have used Deep Learning using feed forward neural network on Aarhus city data set. They have performed comparison with SVM, NN machine learning algorithms and prove that deep learning neural networking perform well in accurately measuring the pollution value. They only took one pollutant and they solved the problem linearly. They did not mentioned how the real-time data will be maintained. In [25]. Ozone concentration is studied at Tunisia. They have used three monitoring stations for measuring ozone concentration and used Random forest and Support vector Regression for future prediction. They have found Random Forests to be more accurately estimator for predicting ozone. However, the data from three stations is of small amount and they have considered only one variable for future prediction.

Another study for forecasting air pollution in Canada was carried using machine learning Multilayer perceptron neural network (MLPNN) [26]. They addressed the issue of air quality prediction and model accuracy. However, they have used limited amount of data. Moreover, the computational caste for seasonally updating of the model is large.

Deep learning technique for decreasing the error rate of time-series analysis is proposed by [27]. They have made comparison of neural network with auto regression moving average (ARMA), and support vector regression (SVR) models. Although, accuracy has been increased but the processing time is not mentioned. In order, to process large data Big Data management architecture has been pro- posed to predict air quality in China [28]. But there is no implementation. In another study, big data architecture is proposed but there is no implementation [29].

Air simulation based on big data is proposed in [30]. They performed comparison of MapReduce Hadoop and Spark for simulating air quality. They have used the dataset of Texas 179 sensors and found that 20∼25 % performance benefits for the Spark solutions over MapReduce. They have mentioned that real time decision can be processed here but they did not mention the prediction accuracy. Another Apache Spark based AQI prediction system by Random forest is implemented using the Spark distributed on multiple clusters [31]. However, Random Forest can be used as classification problem. It does not perform for real-time analysis of time series data.

Recently, in China, air pollutants data of different cities has been analyzed using big data. They used ensemble Neural Network technique and analyzed data for 16 cities in China [31]. Although, accuracy of predictive model is improved but the processing time is not discussed. This technique can only be applicable to different regions comparison for offline-mode but not useful for real-time processing of within cities different point analysis.

A study proposed ETL (Extract-Transform-Load) framework on cloud platform for Air Quality analysis and prediction.[32] Authors worked on pre-processing of data which was collected from different source and used cloud platform for processing. They have achieved up to 81 % accuracy using RNN.

Another study [33] monitored data set of five cities in China to analyze the occurrence areas and percentage of various concentration ranges of PM2.5. Also, they have done assessment of air quality index of each city. Moreover, effects of winter-heating in the two cities i.e. Beijing and Shenyang. Authors have used statistical analysis to analyze the air quality of these cities. However, prediction and future data processing is missing in this study.

Another research used ground-based data of particulate matter in conjunction with a suite of remote sensing and meteorological data products [34]. Techniques are not discussed explicitly. Recent researchers have analyzed pollution in combination with metrological parameters. Using various machine learning techniques, they have used meteorological data to classify PM2.5 values. Moreover, they have also performed regression analysis to find the coherence [35].

One of the study analyzed the personal health information, using techniques to ensure data confidentiality. It recorded the personal details as study identity numbers prior to uploading to the Cloud. Important information for urban planning was obtained using data mining techniques on the obtained environmental and behavioral data [36]. A study was carried out to analyze the PM2.5 pollution and its relationship with other meteorological factors like temperature, humidity etc. was studied in Chengdu, China for the purpose to improve local air quality. The results may help authorities to formulate future policies for control of emission in China [37]. In this paper, meteorological data and PM2.5 concentration data were obtained during the period January 1, 2013 to December 31, 2013. The spatial distribution of study area shows that the western part is more 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. [38]

Day wise air pollution predictions of 74 cities in China were studied with machine learning technique. Five different classification techniques were adopted with different features groups coming from WRF-Chem models forecast results.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

They worked on feature selection technique. [39] ANN has limitations of low convergence rate. A study proposed an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

TABLE I
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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

TABLE I
LIMITATIONS OF AIR POLLUTION FORECASTING TECHNIQUES (CONTD.)

	Problem Statement	Technique	Strength	Limitations
[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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

algorithm on Hong Kong data which showed better predictive ability, with increased R^2 and decreased RMSE. It was shown that Extreme Learning Machine performs well in terms of precision, generalization and robustness. No significant differences were noted between the prediction accuracies of each model. Extreme Learning Machine provided the best performance on indicators related to prediction such as R^2 and RMSE etc. The authors achieved 95 RMSE and training time of 0.07s [40].

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

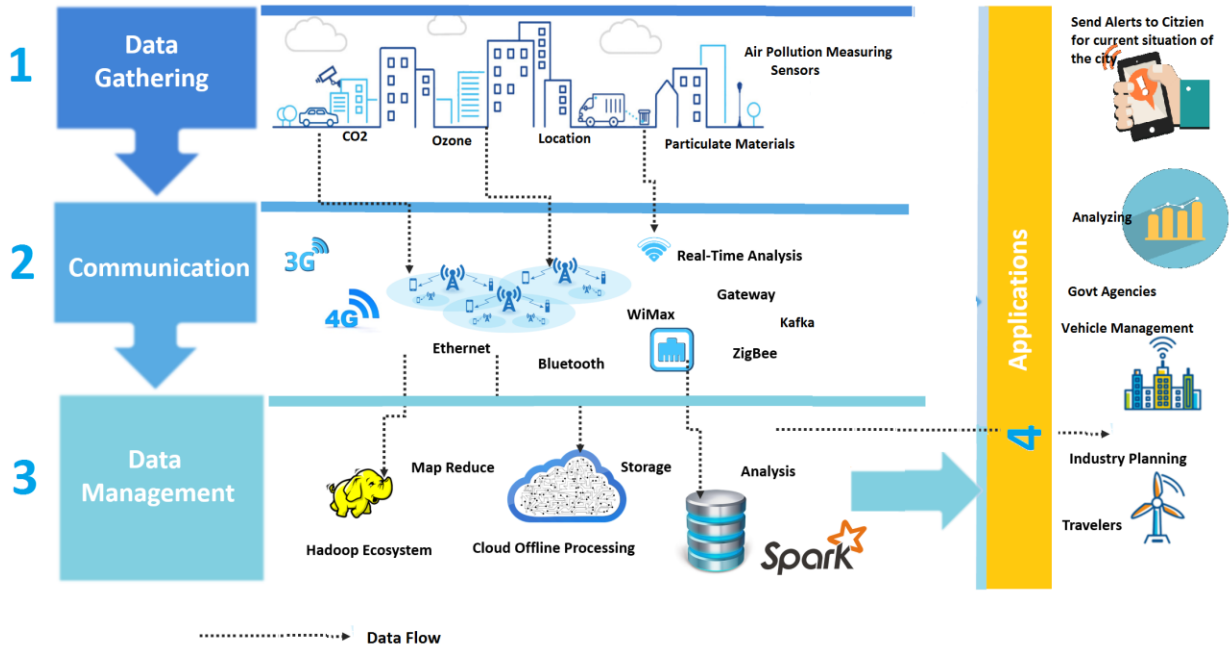

Figure 1 Smart City Air Pollution Monitoring Architecture.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

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) [41]. 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) [42]. 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 [44]. 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 [45]. 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 [47] [48]. 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.

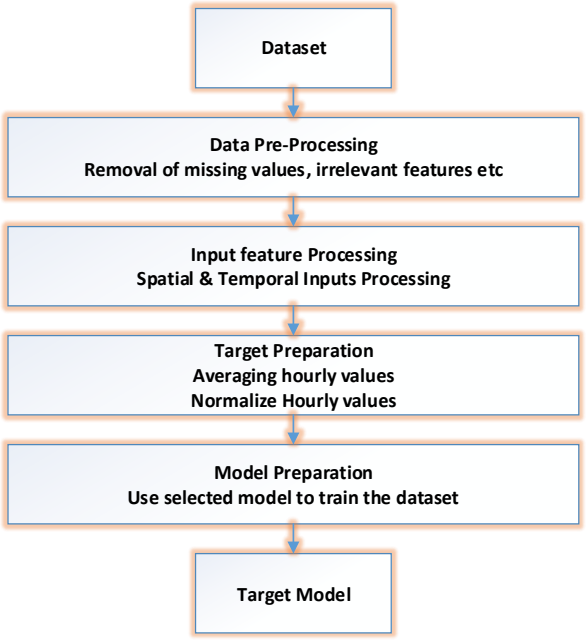

Figure 2 Implementation Model.

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.

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 [49]. It consists of different meteorological variables and PM2.5 recorded from different locations. Table II 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 III shows baseline characteristics of Data set showing mean and standard deviation. Among them, Guangzhou's PM2.5 was 53 ug/m3 ± 42 ug/m^3 which was recorded as lowest and PM 2.5 of Beijing and Shenyang were 85.6 ug/m^3

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

TABLE II
DATASET DESCRIPTION

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

TABLE III
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 IV
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

± 83.5 ug/m^3 and 78.7 ug/m^3 ± 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 IV 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 Shanghai city, data set of other cities may also be correlated in the same way.

V. SYSTEM EVALUATION

A. TESTBED USED:

All the processing was carried on operating system Ubuntu 14.04 on i5 machine. Development was carried using Python Programming Language. Initially, pre-processing and time series evaluation was carried out using Pandas. Machine learning algorithms were implemented using scikit learn library which is an open source machine learning library for the Python programming language. Plotting of graph was done using plotly library. Evaluation was carried using sklearn metrics. All the code was written on Jupyter Notebook. All hyper-parameters are tuned using ten-fold cross validation method and the Grid SearchCV function. Grid SearchCV function has the capability to make an exhaustive search over specified parameter values defined by the user and detect automatically. In order to evaluate the performance on Spark, we have used spark learn library provided by DataBricks. So we used Grid SearchCV of spark learn library.

B. EVALUATION CRITERIA

Two mostly used metrics of measuring accuracy of continuous variables are MAE and RMSE.

1) MEAN ABSOLUTE ERROR (MAE):

Mean absolute error is the criteria which measures the average magnitude of the errors in a set of data values (predictions), without any consideration of direction [50]. In a test sample, MAE is the average of the absolute differences between actual and prediction observations. It is calculated as in Eq (2):

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

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

differences between actual vs predicted values and taking the square root of the final result [51]. It is calculated as:

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

In order to compare the datasets or modes having different scales, normalizing of the RMSE is done by the following method [51]:

$$Normalized\ RMSE = \frac{RMSE}{y_{max} - y_{min}} \quad (4)$$

Where,

y_{max} = Maximum value of data set
y_{min} = Minimum value of data set

VI. RESULTS AND DISCUSSION

The predicted vs actual results are shown by the line graphs which is a good visualization technique to assess the goodness-of fit of a regression model at a glance. Time is taken on x axis, while prediction values are shown on y axis.

A. BEIJING CITY

Beijing city is reported for the highest values of PM2.5 in China. We have applied the afore-mentioned regression techniques on Beijing city data set and predicted the maximum and minimum values of pollution in the city and compared it with the actual values. Data was trained from 1st Jan 2010 to 21 Dec 2015 pollution values and prediction was conducted for the next week on 22 Dec 2015 to 31 Dec 2015. Predicted analysis for Gradient Boosting Regression (GBR), Decision Tree Regression (DTR), Multi-layer Perceptron Regression (MLP) and Random Forest Regression (RFR) is conducted. As evident from the Figures 4(a) and 4(b), GBR has performed poor for identifying the peak values which can be seen from the difference of actual vs predicted values at locations of local maxima on December 25-26 & December 30, 2015. This inference is supported by the error rate calculation from the Figure 4 in which RMSE achieved is 0.08 after normalizing and MAE is 29.3 %, which is very high compared to other techniques. RMSE and MAE achieved by MLPR is also on higher side. Comparatively, DTR and RFTR has performed much better

for identifying the peak values. RMSE achieved for Random forest regression is 0.0725 after normalizing and MAE is 16 % which is much better as compared to Gradient Boosting regression and Decision tree regression.

B. SHANGHAI CITY

For Shanghai city, 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 to 31 Dec 2015. The results indicate that decision tree and gradient boosting regression were not able to accurately identify the maximum values. MAE & RMSE computed are displayed in Figure 6. MAE for DTR and GBR remained 22% and 17% respectively, while RMSE achieved was 0.09 and 0.07 respectively. These are higher as compared to the other two

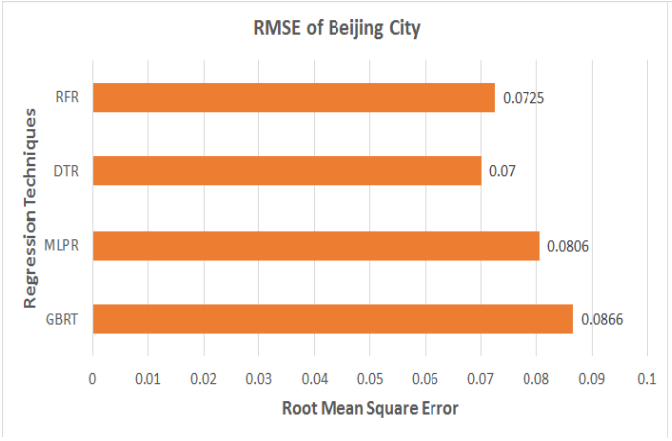

Figure 4(a) **RMSE for different regression techniques.**

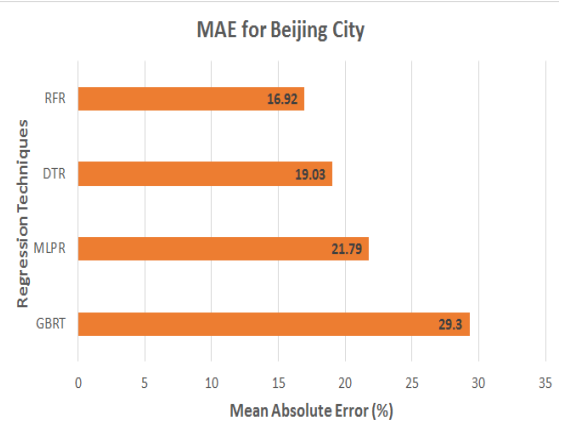

Figure 4(b) **MAE for different regression techniques.**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

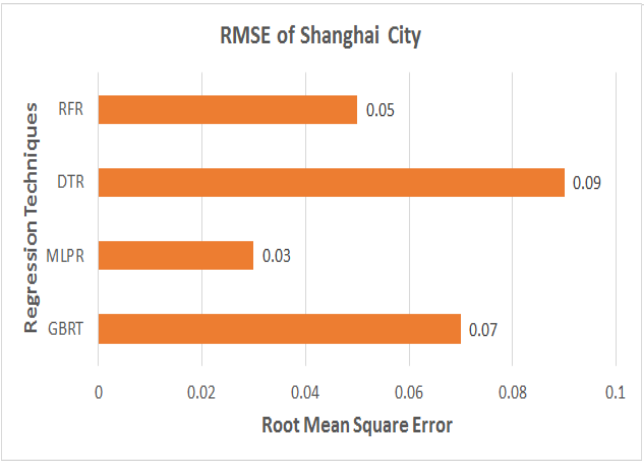

Figure 6(a) **RMSE for different regression techniques.**

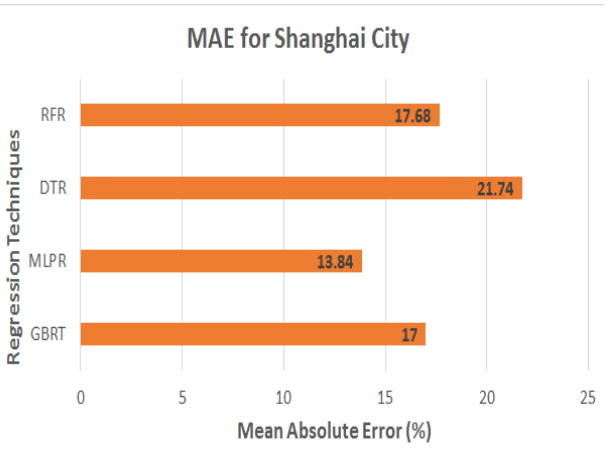

Figure 6(b) **MAE for different regression technique.**

techniques. MLP performed much better not only in identifying the peak values but also achieved lowest RMSE of 0.03 and lowest MAE of 13.84%. Random forest regression was also not much behind with RMSE of 0.05 and MAE of 17%.

C. SHENYANG CITY

Regression analysis was performed on Shenyang city's data set and prediction was performed. The MAE and RMSE of different techniques for this city were calculated and shown in Figure 8. From the graph, it is clear that MLPR and RFR have

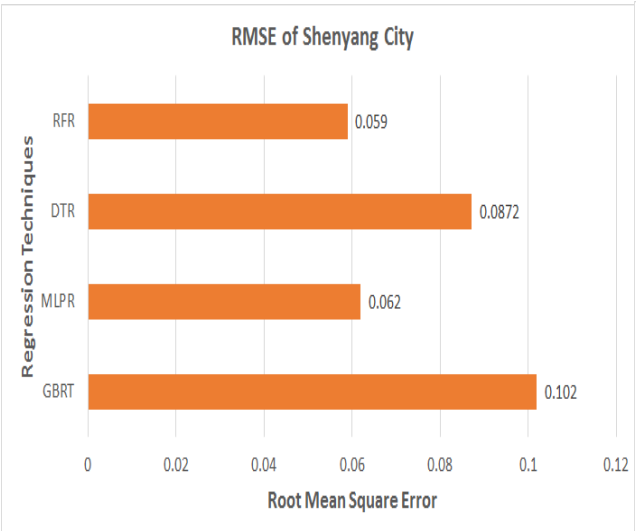

Figure 8(a) **RMSE for different regression techniques.**

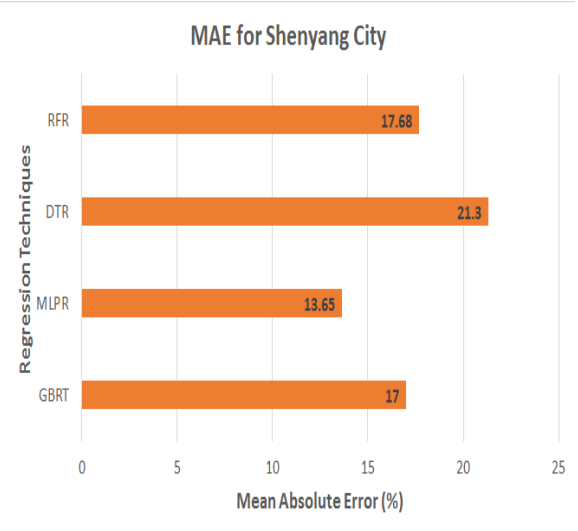

Figure 8(b) **MAE for different regression techniques.**

remained best in accurately predicting the pollution. MLPR has the mean absolute error of 13.65 % while RMSE achieved was 0.062. Although, MAE and RMSE of GBR and DTR are also not much higher but they have performed poor compared to other two techniques. Gradient boosting regression was not able to correctly identify the peak values. Decision tree gave the mean absolute error of 21.3% and RMSE of 0.087 which is on higher side.

D. GUANGZHOU CITY

Regression analysis was performed on Guangzhou city's data set and prediction was performed. The data set is smooth with only one extraordinary peak at the start on December 22, 2015. This data was also trained from 1st Jan 2010 to 21Dec 2015 pollution values and prediction was conducted for the next week on 22 Dec 2015 to 31 Dec 2015. From the Figure 10, it is evident that GBR could not perform well to predict the pollution values in this data. Neither it could identify the peak values nor accurately predict pollution as a whole. MAE for gradient boosting regression remained 27.8% while RMSE achieved was 0.17 which is very high. MLP and random forest regression again remained the top performers. MLP achieved MAE and RMSE of 12.2% and 0.045 respectively while random forest was also close with MAE of 13.1% and RMSE of 0.05. They were able to identify the peaks with the least error. Decision tree also performed better than GBR with MAE of 14.36% and RMSE of 0.06.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

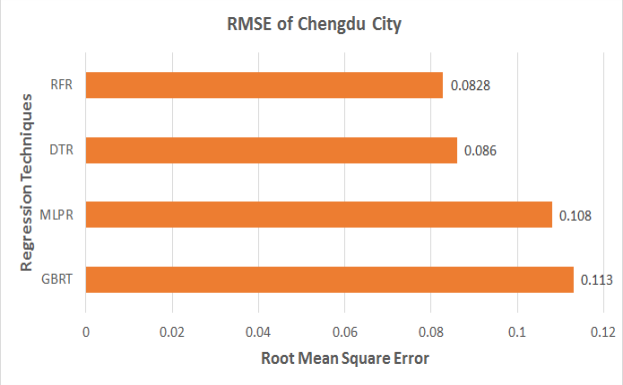

Figure 12(a) RMSE for different regression techniques.

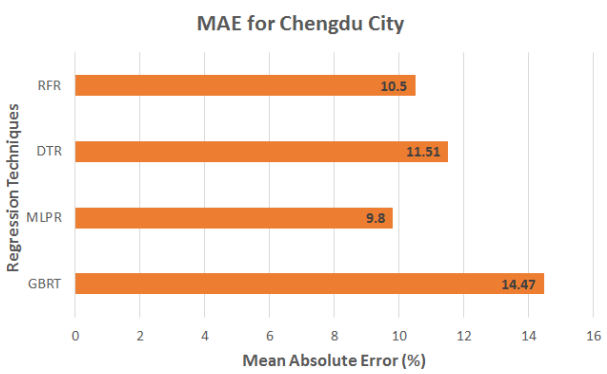

Figure 12(b) MAE for different regression technique.

E. CHENGDU CITY

Predictive analysis for Chengdu city applying different regression techniques is plotted in figure 12. From the graph, it is clear that gradient boosting regression remained the poor performer of all the four techniques. It could not identify the data peaks. MAE for gradient boosting regression was 14 % and RMSE achieved was 0.113 showed in 5.10. Random forest regression performed the best in accurately predicting the results. It achieved RMSE of 0.08 and MAE of 10.5 %. MLP was identified as the second best technique with MAE of 9.8 % and RMSE of 0.108.

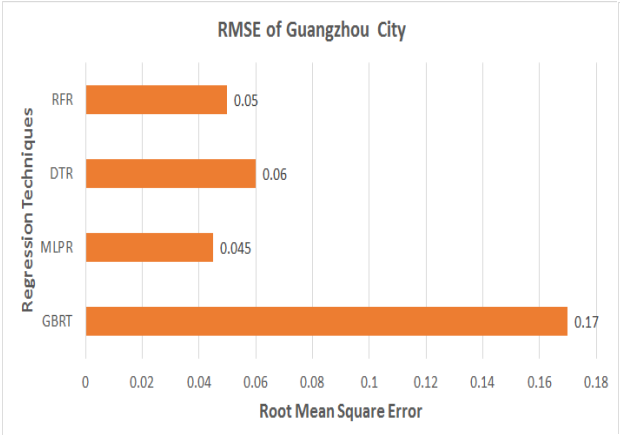

Figure 10(a) RMSE for different regression techniques.

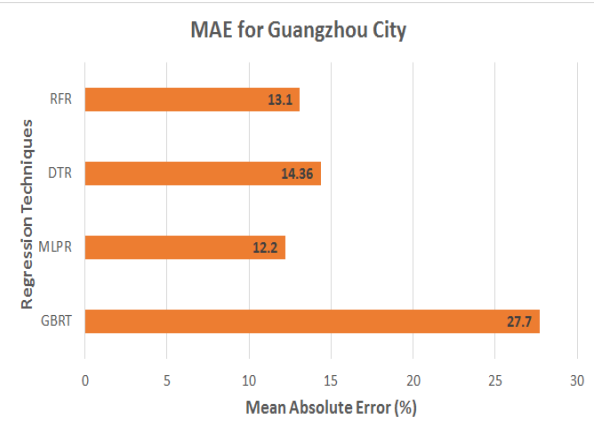

Figure 10(b) MAE for different regression techniques.

F. DISCUSSION

In the previous section, results of all the four regression techniques on different data sets have been presented. A relation of PM2.5 with other 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 depicts that lower the temperature of the city, the higher will be the amount of PM 2.5 concentration in the city. This can be explained from the fact 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 the longer time as compare to light air, hence the concentration of PM 2.5 is reportedly 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.

Hence, the concentration of particulate matter will be reported lower at the particular location where wind is blowing at higher speeds. In winters, the PM2.5 level in China also increases due to burning of fossil fuels.

Out of the four techniques used, decision tree has the advantage of simplicity to understand and implement. Although, accuracy of decision tree is not up to the mark with other techniques but due to simplicity of tree regression, the processing time of this technique is less as compared to other models. Error rate in mean absolute error is between 8 to 21 % while RMSE is between 0.06 to 0.24.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

Random Forest Regression is the 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 as compared to other models. MAE for different dataset ranges from 6 to 18 % while RMSE ranges from 0.05 to 0.18. Random Forest Regression performed well after hyper-parameter tuning.

However, the Random Forest Error prediction is almost comparable to the Multi Linear Perception Regression. For the dataset which consists of large amount of historic data, Random Forest Regression performs the best. It is found that Random Forest Regression have better estimation performance among all four regression 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 the table V. It is evident that Gradient Boosting Regression has the highest error rate as compared to other three regression techniques in most of the data sets. Random Forest Regression achieves the lowest mean absolute error and **RMSE**.

Table VI depicts the time in seconds taken by different regression models used while learning and running it on test data. Learning For performance tuning and evaluation, we have initially run the algorithms without setting the parameters. The lowest processing time was consumed by Decision Tree Regression and Random Forest Regression as compared to Gradient Boosting Regression and Multi Linear Perception. Then we did the hyper parameter tuning on single node of Spark 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 air pollution prediction issue. The techniques included 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 which can be perform well for pollution prediction of data sets of varying size and location having different characteristics. Its processing time was found much lesser as compared to the gradient boosting and multi-layer perceptron algorithms. Whereas, its error rate was found least among all the four techniques. Although processing time of decision tree was recorded as lowest; however, its error rate remained high in most of the techniques and also it was not able to properly identify the data peaks in almost all data sets. In comparison, random forest regression took very less time as compared to other techniques which is just higher than decision tree and it also performed well in identifying the peak values and accurately predicting the results with less error rate. Therefore, we can deduce the conclusion that random forest regression was the best technique among the considered four algorithms.

TABLE V
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 VI
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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2925082, IEEE Access

Gradient boosting regression has performed worst 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 performances of techniques on 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." [Online]. Available: https://www.who.int/phe/eNews_63.pdf

[2] Moore, Frances. "Climate change and air pollution: exploring the synergies and potential for mitigation in industrializing countries." Sustainability 1.1 (2009): 43-54.

[3] Hsieh, Hsun-Ping, Shou-De Lin, and Yu Zheng. "Inferring air quality for station location recommendation based on urban big data." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 437-446. ACM, 2015.

[4] Johnson, M., Isakov, V., Touma, J.S., Mukerjee, S. and Özkaynak, H.,. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. Atmospheric Environment, 44(30), (2010) pp.3660-3668.

[5] Malalgoda, Chamindi, Dilanthi Amaratunga, and Richard Haigh. "Local governments and disaster risk reduction: a conceptual framework." Massey University/The University of Auckland, 2016.

[6] Kioumourtzoglou, Marianthi-Anna, Joel D. Schwartz, Marc G. Weisskopf, Steven J. Melly, Yun Wang, Francesca Dominici, and Antonella Zanobetti. "Long-term PM2. 5 exposure and neurological hospital admissions in the northeastern United States." Environmental health perspectives 124, no. 1 (2015): 23-29.

[7] World Health Organization, and UNAIDS Air quality guidelines: global update 2005. World Health Organization, 2006.

[8] Kim, Ki-Hyun, Ehsanul Kabir, and Shamin Kabir. "A review on the human health impact of airborne particulate matter." Environment international 74 (2015): 136-143.

[9] Hvidtfeldt, Ulla Arthur, Matthias Ketzel, Mette S_rensen, Ole Hertel, Jibran Khan, J_rgen Brandt, and Ole Raaschou-Nielsen. "Evaluation of the Danish AirGIS air pollution modeling system against measured concentrations of PM2.5, PM10, and black carbon." Environmental Epidemiology 2, no. 2 (2018): e014.

[10] A. J. Cohen et al., \Articles 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. 6736, no. 17, pp. 1{12,2017.

[11] En:wikipedia:org:(2018):Airqualityindex:[online]Availableat : https :==en:wikipedia:org=wiki=Airqualityindex[Accessed2Dec:2018].

[12] Yi, Wei, Kin Lo, Terrence Mak, Kwong Leung, Yee Leung, and Mei Meng. "A survey of wireless sensor network based air pollution monitoring systems." Sensors 15, no. 12 (2015): 31392-31427.

[13] Y. Xing, Y. Xu, M. Shi, and Y. Lian, \The impact of PM2 . 5 on the human respiratory system," vol. 8, no. I, pp. 69{74, 2016.

[14] M. M. Rathore, A. Paul, A. Ahmad, and S. Rho, \US CR," Comput. Networks, no. 2016, 2015.

[15] Asgari, Marjan, Mahdi Farnaghi, and Zeinab Ghaemi. "Predictive mapping of urban air pollution using Apache Spark on a Hadoop cluster." In Proceedings of the 2017 International Conference on Cloud and Big Data Computing, pp. 89-93. ACM, 2017.

[16] D. Zhu, C. Cai, T. Yang, and X. Zhou, \A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization," no. December,pp. 1{14, 2017.

[17] R. W. Gore, \An Approach for Classi_cation of Health Risks Based on Air Quality Levels," pp. 58{61, 2017.

[18] K. G. Ri, R. Manimegalai, G. D. M. Si, R. Si, U. Ki, and R. B. Ni, \Air Pollution Analysis Using Enhanced K-Means Clustering Algorithm for Real Time Sensor Data," no. August 2006, pp. 1945{1949, 2016.

[19] N. Zimmerman et al., \Closing the gap on lower cost air quality monitoring:machine learning calibration models to improve low-cost sensor performance," no. 2, pp. 1~36, 2017.

[20] I. Bougoudis, K. Demertzis, and L. Iliadis, \EANN 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," Proc. - 5th Int. Conf. Adv. Cloud Big Data, CBD 2017, pp. 385{390, 2017.

[22] C. A. Keller, M. J. Evans, J. N. Kutz, and S. Pawson, \Machine Learning and Air Quality Modeling," pp. 4488{4494, 2017.

[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," Sensors Actuators, B Chem.,vol. 231, pp. 701{713, 2016.

[24] O. A. Ghoneim, \Forecasting of Ozone Concentration in Smart City using Deep Learning," pp. 1320{1326, 2017.

[25] A. Ben Ishak, M. Ben Daoud, and A. Trabelsi, \Ozone Concentration Forecasting Using Statistical Learning Approaches," 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," pp. 195{211, 2017.

[27] X. Li and L. Peng, \Deep learning architecture for air quality predictions," Environ. Sci. Pollut. Res., pp. 22408{22417, 2016.
T. Huang, L. Lan, X. Fang, P. An, J. Min, and F. Wang, \Accepted

[28] Manuscript Promises and Challenges of Big Data Computing in Health Sciences Reference: To appear in: Revised date: Graphical abstract," Big Data Res., 2015.

[29] H. Ayyalasomayajula, E. Gabriel, P. Lindner, and D. Price, \Air Quality Simulations Using Big Data Programming Models," 2016 IEEE Second Int. Conf. Big Data Comput. Serv. Appl., pp. 182{184, 2016.

[30] C. Zhang and D. Yuan, \Fast _ne-grained air quality index level prediction using random forest algorithm on cluster computing of spark," Proc. – 2015 IEEE 12th Int. Conf. Ubiquitous Intell. Comput. 2015 IEEE 12th Int. Conf. Adv. Trust. Comput. 2015 IEEE 15th Int. Conf. Scalable Comput. Commun. 20, pp. 929{934, 2016.

[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," vol. 27, no. 2, pp. 565{578, 2018.

[32] Chang, Yue Shan, Kuan-Ming Lin, Yi-Ting Tsai, Yu-Ren Zeng, and Cheng-Xiang Hung. "Big data platform for air quality analysis and prediction." In Wireless and Optical Communication Conference (WOCC), 2018 27th, pp. 1-3.IEEE, 2018.

[33] A. J. Cohen et al., \Articles 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. 6736, no. 17, pp. 1{12, 2017.

[34] Y. Xing, Y. Xu, M. Shi, and Y. Lian, \The impact of PM2 . 5 on the human respiratory system," vol. 8, no. I, pp. 69{74, 2016.

[35] J. K. Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, \Modeling PM 2 .5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters," vol. 2017, 2017.

[36] Y. 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," no. 2, pp. 309{312, 2015.

[37] Y. Li, Q. Chen, H. Zhao, L. Wang, and R. Tao, \Variations in PM10, PM2.5 and PM1.0 in an Urban Area of the Sichuan Basin and Their Relation to Meteorological Factors," pp. 150{163, 2015.

[38] J. Wang and S. Ogawa, \E_ects of Meteorological Conditions on PM 2.5 Concentrations in," pp. 9089{9101, 2015.

[39] X. Xi et al., \A Comprehensive Evaluation of Air Pollution Prediction Improvement by a Machine Learning Method," pp. 176{181, 2015.

[40] H. Kong, \Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of," pp. 1{19, 2017.

[41] "Airqualityindex";En:wikipedia:org; 2018:[Online]:Available : https : ==en:wikipedia:org=wiki=Airqualityindex:

[42] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classi_cation and Regression Trees. Wadsworth, Belmont, CA, 1984

[43] "DecisionTreesjscikit learn0:20:1documentation"; Scikit learn:org; 2018:[Online]:Available: https: scikit-learn.org/stable/modules/tree.html

[44] L. Breiman, \Random Forests", Machine Learning, 45(1), 5-32, 2001.

[45] Ridgeway, \Generalized Boosted Models: A guide to the gbm package", 2007.

[46] Prettenhofer, Peter, and Gilles Louppe. "Gradient boosted regression trees in scikit-learn." (2014).

[47] Backpropagation" Andrew Ng, Jiquan Ngiam, Chuan Yu Foo, Yifan Mai, Caroline Suen - Website, 2011

[48] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." Cognitive modeling 5.3 (1988): 1.

[49] Liang, X., S. Li, S. Zhang, H. Huang, and S. X. Chen (2016), PM2.5 data reliability, consistency, and air quality assessment in five Chinese cities, J. Geophys.Res. Atmos., 121.

[50] Chai, Tianfeng & Draxler, R.R.. (2014). Root mean square error (RMSE) or mean absolute error (MAE)Arguments against avoiding RMSE in the literature. Geoscienti_c Model Development. 7. 1247-1250. 10.5194/gmd-7-1247-2014.

[51] In Root mean squaredeviation; "Wikipedia; 28 Aug 2018:[Online]:Available :https://en.wikipedia.org/wiki/Root-mean-square_deviation:

Saba Ameer is a citizen of Rawalpindi, Pakistan. She is a student of master's in computer science at the Department of Computer Science, COMSATS University Islamabad, and Islamabad, Pakistan. Her area of research includes Big Data, IoT, machine learning, digital image processing, artificial intelligence and algorithms.

Munam Ali Shah received B.Sc. and M.Sc. degrees, both in Computer Science from University of Peshawar, Pakistan, in 2001 and 2003 respectively. He completed his M.S. degree in Security Technologies and Applications from University of Surrey, UK, in 2010, and has passed his Ph.D. from University of Bedfordshire, UK in 2013. Since July 2004, he has been a Lecturer, Department of Computer Science, COMSATS Institute of Information Technology, Islamabad, Pakistan. His research interests include MAC protocol design, QoS and security issues in wireless communication systems. Dr. Shah received the Best Paper Award of the International Conference on Automation and Computing in 2012. Dr. Shah is the author of more than 50 research articles published in international conferences and journals.

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, in August 2012. In August 2017, he joined the Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL, where he is currently an Assistant Professor and the Director of the Security and Optimization for Networked Globe Laboratory (SONG Lab, www.SONGLab.us). He served on the faculty of West Virginia University from August 2012 to August 2017. In 2007 he was an Engineering Research Associate with the Texas A&M Transportation Institute. He serves as an Associate Technical Editor for IEEE Communications Magazine. He is the editor of four books, including Smart Cities: Foundations, Principles and Applications, Hoboken, NJ: Wiley, 2017, Security and Privacy in Cyber-Physical Systems: Foundations, Principles and Applications, Chichester, UK: Wiley-IEEE Press, 2017, Cyber-Physical Systems: Foundations, Principles and Applications, Boston, MA: Academic Press, 2016, and Industrial Internet of Things: Cyber-manufacturing Systems, Cham, Switzerland: Springer, 2016. He is the author of more than 100 articles. His research interests include cyber-physical systems, cybersecurity and privacy, Internet of things, edge computing, big data analytics, unmanned aircraft systems, connected vehicle, smart and connected health, and wireless communications and networking. (PDF) A Trust and Priority based Code Updated Approach to Guarantee Security for Vehicles Network. Available from: https://www.researchgate.net/publication/328009957_A_Trust_and_Priority_based_Code_Updated_Approach_to_Guarantee_Security_for_Vehicles_Network [accessed Apr 29, 2019].

Castren Maple is Professor of Cyber Systems Engineering at WMG's Cyber Security Centre (CSC), University of Warwick. He is the director of research in Cyber Security working with organisations in key sectors such as manufacturing, healthcare, financial services and the broader public sector to address the challenges presented by today's global cyber environment. He is a member of several professional societies including the Council of Professors and Heads of Computing (CPHC) whose remit is to computing departments. He is an elected member to the Committee of this body. He is an Education Advisor for TIGA the trade association representing the UK's games industry. He is also a Fellow of the British Computer Society, the Chartered Institute for IT and is a Chartered IT professional. He also holds two Professorships in China, including a position at one of the top two control engineering departments in China. His interests include Information Security and Trust and Authentication in Distributed Systems.

Saif ul Islam received his Ph.D. in Computer Science at the University Toulouse III Paul Sabatier, France in 2015. He is Assistant Professor at the Department of Computer Science, Dr. A. Q. Khan Institute of Computer Science and Information Technology, Rawalpindi, Pakistan. Previously, he served as Assistant Professor for three years at the COMSATS University, Islamabad, Pakistan. He has been part of the European Union-funded research projects during his Ph.D. He was a focal person of a research team at COMSATS working in O2 project in collaboration with CERN Switzerland. His research interests include resource and energy management in large-scale distributed systems

(Edge/Fog, Cloud, Content Distribution Network (CDN)) and the Internet of Things (IoT).

Dr. Muhammad Nabeel Asgahr is an Assistant Professor in the Department of Computer Science, Bahauddin Zakariya University Pakistan. He received his PhD from University of Bedfordshire, UK where his study mainly emphasis on modelling for machine vision, specifically digital imagery and its wide spread application in all vistas of life. He has been investigating machine learning approaches for analysing 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. 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 & visualization, graphics modelling & simulation, CR MAC protocol design, Internet of Things and security issues in wireless communication systems.