Time Series Based LSTM Model to Predict Air Pollutant's Concentration for Prominent Cities in India

Vidushi Chaudhary, Anand Deshbhratar, Vijayanand Kumar, Dibyendu Paul
Samsung Research Institute India - Noida
{vidushi.c,a.deshbhrata,vijay.k2,d.paul}@samsung.com

ABSTRACT

Air pollution has been a major concern in India since last few years. Rapid industrialization has led to tremendous increase in air pollution. Growing air pollution has led to an increase in health related ailments like stroke, heart disease, lung cancer etc. To monitor air pollution, particularly in major pollution prone cities like Delhi, the government of India has installed pollutant's measuring sensors. The sensors regularly monitors the level of air pollutants like $PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, O_3. However, capturing current pollutants concentration is not helpful in decreasing/avoiding pollution exposure.

This paper proposes a deep learning based LSTM model to predict future air pollutant's concentration. We formulate the problem of predicting pollutant's concentration as a time series based problem where current pollutants level are dependent on previous pollutant, meteorological, traffic data, festivals and national holidays information. We create a sample dataset of historic pollutant levels, meteorological and traffic data, and identify discriminatory features to predict pollutants concentration for the next # of hours. Our empirical analysis reveals that certain meteorological features (temperature, pressure, humidity, wind speed, wind direction, UV Index, cloud cover, rain), traffic information, festivals information can be used to forecast pollutants. We conduct a series of experiments to validate the proposed solution approach and present evidences to demonstrate the effectiveness of proposed framework with average RMSE of pollutants is less than 15 for next 12 hour prediction, less than 8 for next 6 hour prediction and less than 5 for next hour prediction. We compare our model accuracy with real time data fetched from CPCB (Central pollution Control Board).

KEYWORDS

Pollutant's Prediction, Air Quality, AQI Prediction, LSTM, Regularization, Random Forest Regressor, Time Series Data

1 RESEARCH MOTIVATION AND AIM

Rapid increase of air pollution is a major concern considering its enormous impact on human life. Industrialization and urbanization have intensified environmental health risks and pollution, especially in developing countries like India. Study shows that air pollution poses a major health risks such as stroke, heart disease, lung cancer, and chronic and acute respiratory diseases. According to the World Health Organization (WHO) report [11], 14 out of the top 15 most polluted cities in the world are in India (in which Delhi is among the top list), an estimated 12.6 million people die from environmental health risks annually. According to the WHO, 92% of the world's population lives in areas where the air quality is below the WHO standards [25]. About 88% of premature deaths occur in the low and middle-income countries, where air pollution is escalating at an alarming rate. India is the third largest producer of greenhouse gases after China and the United States [22]. The severity of air pollution is so much that as per 2016 study conducted by the Indian Institute of Tropical Meteorology (IITM) and Atmospheric Chemistry Observations and Modeling Laboratory, National Center for Atmospheric Research, Boulder, Colorado, USA [8], life expectancy among Indians reduces by 3.4 years on an average while among the residents of Delhi it reduces by almost 6.3 years.

There are 6 prominent air pollutants present in the air, Particulate Matter ($PM_{2.5}$ and PM_{10}), Carbon Monoxide (CO), Ozone (O_3), Nitrogen dioxides (NO_2), Sulphur dioxide (SO_2). Table 1 shows the sources of air pollutants and their major effects on human health and environment [20]. To track the rising pollution trend in India, the government of India has installed pollutant's measuring sensors at various stations covering major pollution prone areas. Multiple steps has been taken by the government to control pollution such as metro facility, increase in public transport, and laws such as even-odd system for personal vehicles. Considering the current trend of pollution growth, these solutions are bound to fail in future. Therefore, air pollution forecasting and generating solutions to control it is today's need.

The focus of the work presented in this paper is forecasting air pollutants concentration and Air Quality Index (AQI) in prominent cities in India so that people are aware of pollution trends well in advance. Previous research shows that air pollution problem is prevalent and the problem of predicting pollutants concentration at sensor or non-sensor location has attracted researcher's attention [27] [10] [21]. For our research study and experiments, we choose 2 cities, Delhi and Agra. Delhi is among the most polluted city in India. Delhi is one of the most populated city also, forecasting air pollution in Delhi would have high impact on people's life. The Taj mahal is situated in Agra that makes it a tourist place where people from all over the world comes. According to Taj mahal's official website [15], The Taj attracts from 7 to 8 million visitors annually, with more than 0.8 million from overseas.

The research aim of the work presented in this paper is following:

Broad Objective : Understand air pollution problem in the environment and investigate effective solutions to improve the accuracy

UDM'18, Aug 2018, London, UK

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

Table 1: Air Pollutants - Emission Sources and Major Effects (Table has been adopted from Central Pollution Control Board website [20])

Pollutants	Emission Sources		Major Effects	
	Natural Sources	**Anthropogenic Sources**	**Health Effects**	**Environment Effects**
Sulphur Dioxide (SO_2)	Volcanic emissions	Burning of fossil fuels, metal melting etc.	Respiratory problems, heart and lung disorders, visual impairment	Acid rain
Nitrogen dioxide (NO_2)	Lightning, forest fires etc.	Burning of fossil fuels, biomass & high temperature combustion process	Pulmonary disorders, increased susceptibility to respiratory infections	Precursor of ozone formation in troposphere, aerosol formation.
Particulate matter (PM)	Windblown dust, pollen spores, photochemically produced particles	Vehicular emissions, industrial combustion processes, construction industries	Respiratory problems, liver fibrosis, lung/liver cancer, heart stroke, bone problems	Visibility reduction
Carbon monoxide (CO)	Animal metabolism, forest fires, volcanic activity	Burning of carbonaceous fuels, emission from IC engines	Anoxemia leading to various cardiovascular problems. infants, pregnant women and elderly people are at higher risk.	Effects the amount of greenhouse gases which are linked to climate change and global warming.
Ozone (O_3)	Present in stratosphere at 10-50 km height	Hydrocarbons and NO_x upon reacting with sunlight results in (O_3) formation.	Respiratory problems, asthma, bronchitis etc.	O_3 in upper troposphere causes green house effects, harmful effects on plants, death of plant tissues.

of forecasting pollution by mining past pollution trends, meteorological data, traffic information, which can be used to predict Air Quality.

Specific Objective : To examine the application of LSTM network [9] [5] on time series data for forecasting air pollutants concentration based on several meteorological features (like temperature, wind, humidity, visibility etc.), traffic information and miscellaneous features like festivals, national holidays information. To conduct an empirical analysis on a real world dataset to measure the effectiveness of the proposed solution is the objective of this work.

2 RELATED WORK AND RESEARCH CONTRIBUTIONS

In this section, we discuss closely related work and present novel research contributions in context to existing work. We categorize the related work in 2 lines of research: Computer Science Solution and Environment Science Analysis based Solutions

2.1 Computer Science Solution

Reddy et al. [21] investigate the use of LSTM framework for forecasting pollution in future based on time series pollutant and meteorological data in Beijing area. The main aim of this paper is the application of LSTM sequence to scalar model to forecast pollution. The paper is predicting only PM2.5 concentration for next timestep. In our paper, we are predicting each pollutant concentration using stacked LSTM model. We are predicting each pollutants concentration for next 1 hour timestep, next 3^{rd} hour timestep, next 6^{th} hour timestep and next 12^{th} hour timestep. Reddy et al. have considered only meteorological data to predict next timestep pollutant concentration while we are taking traffic information, festivals and national holidays information along with meteorological and pollutants data to forecast pollutants. Reddy et al. predicted the pollutants concentration for Beijing area but our work is based on India (because of India's increasing pollution trends since last few years).

Zheng et al. [27] address the issue of air quality inference based on air quality reported by existing sensor stations. Meteorological data, traffic flow, human mobility, point of interests (POIs) are other features used to infer AQI at non-sensor locations. In our paper, we are predicting air pollutants for future time steps at sensor locations. Previous hours pollutant concentration, meteorological data, traffic data is used to forecast next # of hours pollutants. Zheng et al. paper implemented a co-training based semi supervised learning approach that consists of 2 separate classifiers, Artificial Neural Network (ANN) based spatial classifier and Conditional Random Field (CRF) based temporal classifier. While in this work, we have used stacked LSTM network which takes time series data as input and predict values for future timesteps based on past data. Zheng et al. evaluated their approach on 5 real data sources obtained in Beijing and Shanghai. But our research and experimental results are based on Indian cities, Delhi and Agra.

He et al. [7], Roy et al. [23] provide a method to predict PM_{10} concentration. Pérez et al. [19] proposed a method to predict $PM_{2.5}$ concentration for the next 24 hours. In our work, we are providing a method to predict each pollutant concentration and AQI upto next 12 hours.

He et al. [7] have used Changsha, China air quality data for their research study. They monitored daily average concentration of PM10 using TEOM 1400a during 2008 at the railway station monitoring

station. Roy et al. [23] used Mill tailings at Kolar Gold Fields data for their experiments. Monitoring was carried out at the National Institute of Rock Mechanics (NIRM), Kolar Gold Fields (KGF). Pérez et al. [19] performed their experiments for Malaysia. The data was provided by Malaysian Meteorological Department (MMD) and Department of Environment (DOE). Our research study is based on 2 most prominent Indian cities Delhi and Agra where pollutants data is downloaded from CPCB website. Though our model can be extended to any location by providing that location dataset, the experiment results are shown for Delhi and Agra location.
He et al. [7] develop a hybrid methodology to forecast PM10. The paper combines both AutoRegressive Integrated Moving Average (ARIMA) and ANN models to improve forecast accuracy. The paper used ARIMA to model the linear component and then ANN model is used to take care of the residuals from ARIMA model. They report that hybrid model can be a effective way to improve PM10 forecasting accuracy compared to single ARIMA model. Roy et al. [23] present an ANN based approach as predictive and data analysis tool for the evaluation of air pollutant. The paper proposes a multilayer feed forward network to predict PM_{10} concentration using meteorological data. Pérez et al. [19] proposed a three layer neural netwrok to predict $PM_{2.5}$ concentration. They used previous 24 hours $PM_{2.5}$ data for prediction. In this work, we have developed a time series based stacked LSTM model to forecast air pollutant concentration. Past pollutants concentration, meteorological data, traffic data, holidays information is provided to LSTM network in time series sequence to predict future pollutants concentration.

2.2 Environment Science Analysis based Solutions

Nancy Murray [17] et al. developed a method to combine estimates for PM2.5 using satellite retrieved aerosol optical depth (AOD) and simulations from the Community Multiscale Air Quality (CMAQ) modeling system. Their aim is to leverage the advantages offered by both methods in terms of resolution and coverage by using bayesian model averaging.
Adam Leelossy et al. [13] provides a brief review on the mathematical modeling of dispersion of air pollutants in the atmosphere. They discuss the advantages and drawbacks of several models like Gaussian, Lagrangian, Eulerian and CFD models. Their main focus is on several recent advances in the multidisciplinary research field like parallel computing using graphical processing units or adaptive mesh refinement.

Research Contributions: In context to closely related work, this paper makes the following novel contributions:
1. The work presented in this paper is the first step in the direction of predicting each air pollutant concentration ($PM_{2.5}$, PM_{10}, CO, O_3, NO_2, SO_2) and Air Quality Index. Previous research has predicted just PM pollutant concentration. This paper is the first research study of predicting air pollutant concentration for Indian stations (Delhi and Agra).
2. There has been work done in the direction of analyzing satellite data to prediction pollution concentration. Our study is using time series based pollutant, meteorological data, traffic information and miscellaneous features to predict each pollutant concentration in the environment.
3. We conduct an empirical analysis on real time dataset (acquired from Central pollution Control Board (CPCB) for air pollutant, Accuweather for meteorological data, "Here maps" [16] for traffic information and Google calendar for festivals and national holidays information) to train and test the effectiveness of the proposed feature set and prediction algorithm.

3 PROPOSED SOLUTION APPROACH

In this section, we first define our problem statement and then discuss our novel solution approach for predicting pollutant's concentration.

3.1 Problem Statement

(1) Let P_i^t denote the concentration of pollutant i at t^{th} time. Let $\{M_1^t \ldots M_n^t\}$ be the meteorological features, T^t be the traffic information at the same t^{th} time. Given a tuple (P_i^t, $\{M_1^t \ldots M_n^t\}$, T^t), our goal is to find discriminatory feature set f $\subseteq$ $\{M_1^t \ldots M_n^t \quad T^t\}$ s.t. P_i^t and f has co-relation with each other.
(2) Let P_i^t denote the concentration of pollutant i at t^{th} time. Let f ($f_1 \ldots f_n$) be the feature set determined from the co-relation between air pollutant's and meteorological, traffic features. Given a tuple (p,f),our goal is to predict P_i^{t+1}, P_i^{t+2},..., P_i^{t+m}, where m is the number of hours for which prediction is required.

3.2 Proposed Approach

Fig 1 represents the research methodology adopted in our study and the proposed solution approach. Our proposed methodology is a four step process as described below.

3.2.1 ***Data Collection***. The first step consists of acquiring training and testing dataset. We are predicting pollutant concentration in real time. The air pollutant concentration is dependent on many factors like meteorological data and traffic at that time, working day vs. holiday information. We have downloaded pollutant, meteorological and traffic data from various sources starting from 12^{th} Feb 2018, 8am till 6^{th} May 2018, 7am (more details about the dataset are contained in section 4). The dataset is used to train the model. Next hour's forecast meteorological and traffic data is also captured to predict the pollutant concentration in future.

3.2.2 ***Empirical Analysis and Feature Extraction***. Previous studies [21] [27] claims that there is a direct co-relation between meteorological factors, traffic on pollutant's concentration. In this section, we present our empirical analysis to prove the co-relation between them.

(1) **Meteorological Features:** We hypothesize that the concentration of air pollutants is highly influenced by meteorological factors. Our hypothesis is based on the observation that there is a direct co-relation between meteorological factors and pollutant concentration. Fig 3 shows the co-relation between CO and temperature, dew point and wind speed. As the temperature decreases, CO concentration increases and as the wind speed increases, CO concentration also increases.

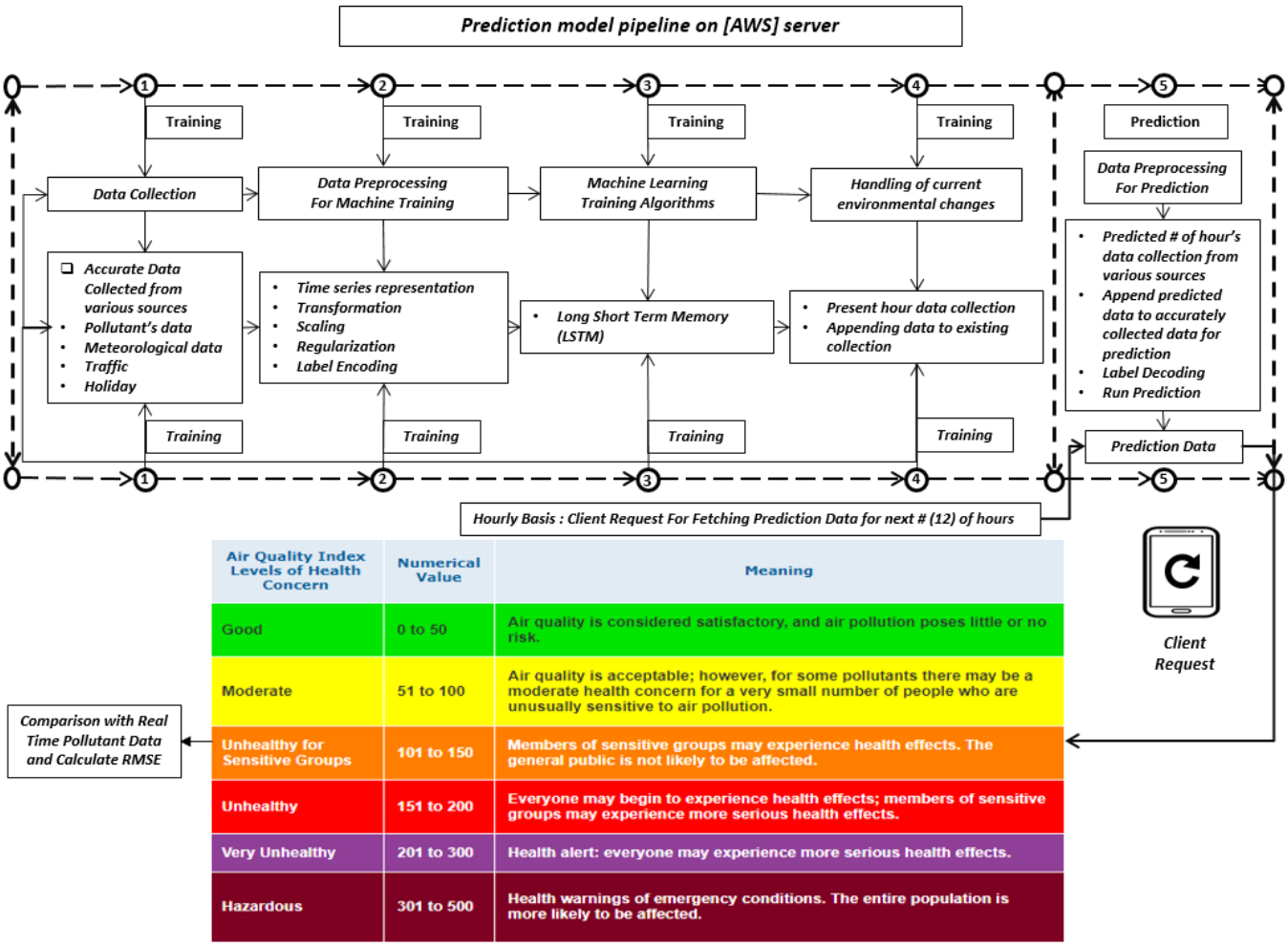

Figure 1: Framework of Pollution Prediction

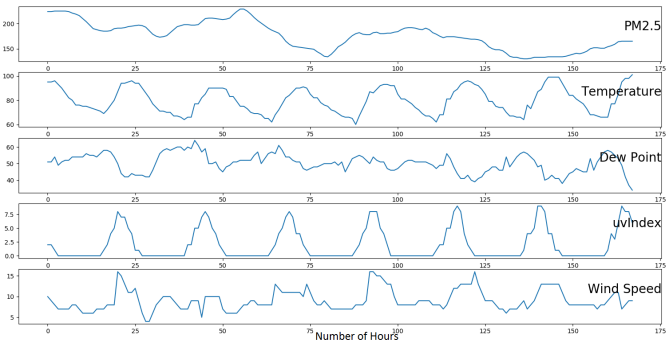

Figure 2: Co-Relation between PM2.5 pollutant Gas and Temperature,Dew Point and Wind Speed

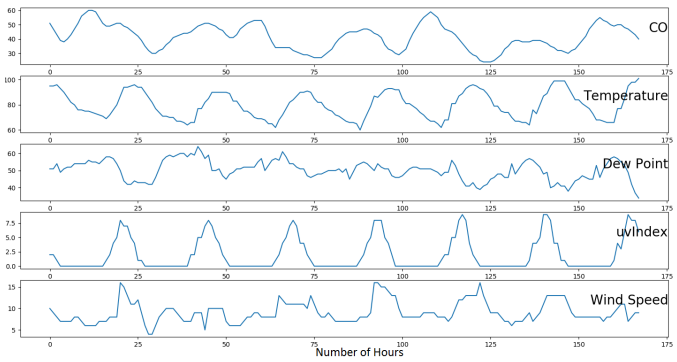

Figure 3: Co-Relation between CO pollutant Gas and Temperature,Dew Point and Wind Speed

Figure 2 and 4 shows the effect of meteorological features on PM2.5 and Ozone gas respectively. For representation purpose, we have shown co-relation graphs on 1 week data.

(2) **Traffic Data :** Previous research [26] [27] shows that traffic is one of the major source of air pollution. Increase in traffic rapidly increases the pollution. Our empirical analysis

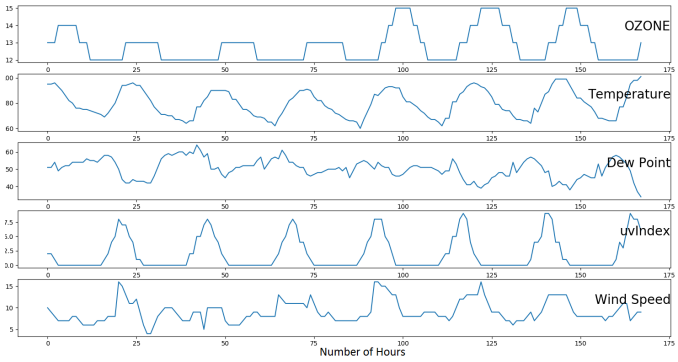

Figure 4: Co-Relation between Ozone pollutant Gas and Temperature,Dew Point and Wind Speed

shows that there is a co-relation between traffic and CO, NO2 concentration. Fig 5 shows the co-relation between traffic condition and CO, NO_2. For representation purpose, we have shown co-relation graphs on 120 hours data. As traffic increases, CO concentration also increases. However, CO is not only dependent on traffic condition but also depends on meteorological factors so behavior is different at some time stamps. As shown in figure 5, pattern of NO_2 change is also dependent on traffic conditions.

(3) **Festivals and National Holidays :** In India, many festivals are celebrated using crackers, colors etc. which eventually increases pollution levels. We observe that on festivals like Diwali and Holi, pollution level reaches new heights. We also observed that PM 2.5 and PM10 concentration is very less on weekend compared to weekdays. The reason could be less vehicles on road that eventually leads to less PM levels. However, we also observed that at visiting places like monuments and historical places, pollutant's concentration (especially PM concentration) is quite high compared to working days. So providing festivals and national holidays information as input to our model helped us in taking care of such patterns.

3.2.3 *Pre-Processing*.

(1) **Pre-Processing of Data in Training Phase**
The pollutant gases, meteorological and traffic data have been merged chronologically. The first pre-processing step is the arrangement of data in time series format where data of each station is maintained hourly in chronological order. This is the most important pre-processing step because input to LSTM model is time series representation of data. Next, all categorical features like wind direction, festivals & national holidays information have been encoded using encoder of sklearn library. MinMaxScaler of sklearn library [18] is used to convert whole dataset into float datatype. Due to scaling, each value vary between 0 and 1. Scaling is performed because LSTM model works best on values in 0-1 range. The last step of preprocessing is the regularization. Since each pollutant is dependent on different meteorological features, regression is performed on each pollutant individually to get

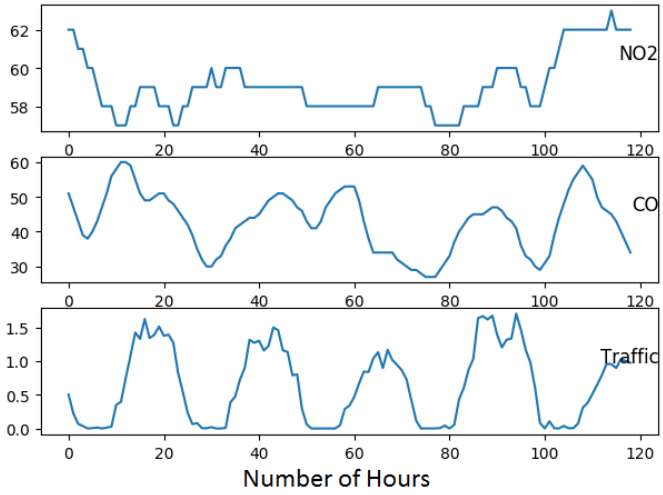

Figure 5: Co-Relation between Traffic Condition and Air Pollutants

Table 2: Selected Meteorological Features Using Regularization

Pollutant	Meteorological feature
NO_2	Temperature, Dew Point, Wind Speed, Cloud Cover
SO_2	Temperature, Dew Point, Cloud Cover
PM2.5	Wind Direction, Wind Speed, Temperature, Dew Point
PM10	Wind Speed, Wind Direction, DewPoint, Ceiling Value, UV index, Cloud Cover
Ozone	Temperature, Visibility, Dew Point, Wind Speed, Wind Direction
CO	Temperature, Dew Point, Wind Speed, Wind Direction

meteorological features on which the pollutant varies. We use Random Forest Regressor [14] with n_estimator set to 7 to perform regularization. After regularization, we get indexes of features important for prediction of that particular pollutant gas. Table 2 shows the important meteorological features for pollutant concentration prediction.

(2) **Pre-Processing of Data in Prediction Phase**
Initially, the forecasted meteorological and traffic data is captured, categorical features are encoded using sklearn library. After that, only those features, which are selected using regularization, are selected. Current pollutant, meteorological and traffic data are appended and scaling is performed using same MinMaxScaler. Finally, whole input is fed into the LSTM network to predict next hour pollutant gases. The results are re-scaled to get the actual value. To get further forecast, the above task is performed recursively by appending the previous hour's pollutant, meteorological and traffic results to the next hour (predicted) meteorological and traffic data.

3.2.4 ***Long Short Term Memory (LSTM) Model***. In this paper, we propose a time series based LSTM model [9] [5] to predict pollutants concentration and calculate Air Quality Index (AQI) for future timesteps. LSTM is a specific type of RNN model that has memory and takes not only current data as input, but also considers what they got previously. So the input of LSTM model at time t is the model output at time *t-1* along with new input at time *t*. Since we are using this model to predict future pollutants concentration, which is highly dependent on previous pollutants, meteorological and traffic data, so a time series based model is a perfect match to achieve the goal.
There are 2 phases of LSTM model, training phase and prediction phase.

(1) **Training Phase :** Fig 6 represents our LSTM training model used to predict the pollutant's concentration. Our LSTM model has 3 types of layers, input layer, hidden layer, dense layer. Input layer is used to provide input to the LSTM model. Input to LSTM model is a vector containing current hour pollutants, meteorological, traffic, festivals data. This feature vector is denoted by X in the diagram, where X_t denotes feature vector at time t. Our framework has 1 hidden layers. Output of LSTM model at time t is an initial parameters vector which also an input for the model for time t+1. The hidden units are internally connected where output h_0 of LSTM at time t is the input of next hidden unit. Hidden layer is used to adjust the weights assigned to initial parameters based on the gradient descent difference. As shown in fig. 6, output of LSTM model at time t is also the input for the model for time t+1. This is because of the LSTM behaviour that next hour output is dependent on previous hour's output. Last layer is the dense layer. The output of all units in the hidden layer (h_i) are connected to a dense layer whose output (P_i) has 6 units which represents each pollutant. These predicted values are then compared with actual pollutant's value at that time.
(2) **Prediction Phase:** Fig 7 represents prediction module. In prediction phase, we are using trained LSTM model to predict pollutant's concentration for next # of hours. Feature vector at time t is the input to trained LSTM model that predict pollutant's concentration for number of hours passed as argument in the function. Next hour predicted values are appended with corresponding hour's meteorological data to predic further hour's pollutants. The whole function is recursively called n times, where n is the number of hours for which prediction is required. Since we are predicting future n number of hours pollutant's (where n can vary from 1 to 12), instead of training separate LSTM model for different values of n, model is trained to predict the next hour value. That next hour predicted value is extrapolated for further hour's prediction.

3.2.5 ***Running Environment***. We have used AWS (Amazon Web Service)[24], a cloud computing platform by Amazon, to run our training and prediction model for the following reasons.

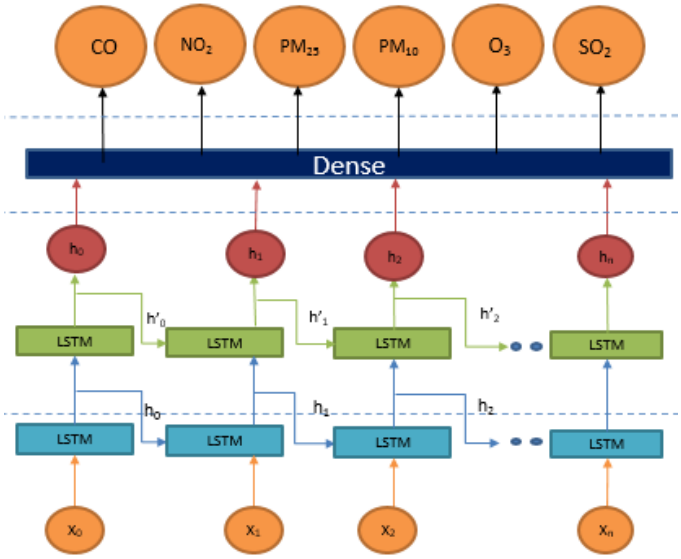

Figure 6: LSTM Training Model with Input, Hidden and Dense Layer

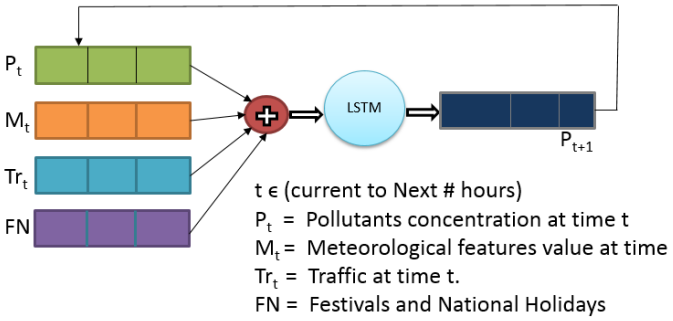

Figure 7: LSTM Prediction Model

(1) Elastic Cloud Computing(EC2) instances can be upgraded very easily, so in future we can make more complex but accurate models without worrying about the training time.
(2) It is much more reliable and maintainable than physical system so AWS is best suited for continuous data downloading and integration.

The basic architecture is defined in fig 8. A python server is running continuously on EC2 instance. As soon as the client makes a "Post" call to the server, the server first parses the incoming JSON object and gets the location and no of hours to predict the pollution level. With these parameters the server calls another script, which is actually responsible for the prediction. As soon as the function call returns to the server script, the server combines the resultant data as a JSON response and sends back to the client. A python script for data collection is running continuously in the background. It fetches the meteorological, pollutant, traffic data from predefined sources and stores the values in csv format on the server itself.

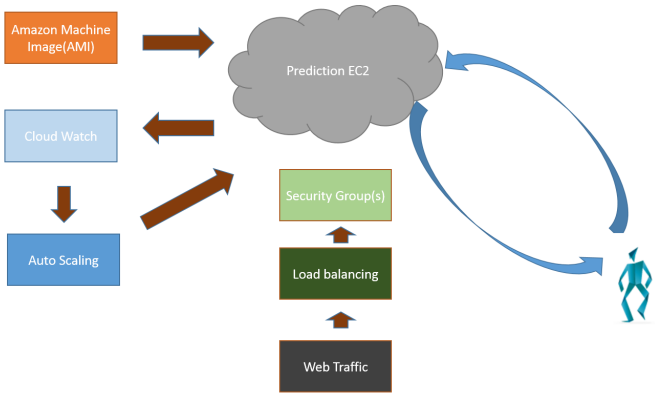

Figure 8: AWS Architecture

4 EXPERIMENTS

In this section, we describe the databases, evaluation metrics and the experimental parameters used in the evaluation.

4.1 Datasets

For training and testing purpose, we use the following real time datasets. Table **??** shows the details of the dataset with downloading source. Whole dataset is divided into 4 categories

1. *Pollutant's data:* We collect real time data of 6 air pollutants, $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3 for Delhi and Agra. The data is downloaded from Central pollution Control Board (CPCB) website every hour. The data is available in xml format that is parsed by a script and combined with other data in csv format.

2. *Meteorological Data :* For each sensor location, we collect corresponding meteorological data from Accuweather website. Coordinates of each sensor location are provided in Accuweather api to get exact location meteorological data. The data consists of temperature, humidity, pressure, wind speed, wind direction, uv index, visibility, cloud cover, rain value and dew point. The data is collected every hour. The Accuweather provides data in json format, which is parsed and merged with existing data.

3. *Traffic Information :* Corresponding traffic information of each sensor location is captured using "Here" maps APIs. Here maps updates traffic information with a traffic value and a confidence score. Traffic information with confidence score less than 0.75 is discarded, the final traffic value is calculated by multiplying traffic value to confidence score as described in equation 1.

$$TrafficCondition = TF * CN \tag{1}$$

where TF = Traffic value and CN = Confidence Score

Since we have coordinates of sensor locations only and traffic information is provided w.r.t roads so from sensor location all roads linked to sensor location are considered to evaluate traffic at that location. Traffic condition of each linked road is averaged out to get the final traffic value at the sensor location.

4. *Others like Holidays Information:* Google calendar is used to create a database of festivals and public holidays. Our empirical analysis shows that there is huge difference in pollution level on holiday compared to a working day.

Table 3: Details of the datasets

Category	Source
Pollutant's Data	https://data.gov.in/sites/default/files/datafile/data_aqi_cpcb.xml
Meteorological Data	https://developer.accuweather.com/apis
Traffic Information	http://traffic.cit.api.here.com
Festivals and National Holidays	Google, Personalized dataset created

The dataset is collected for ITO, Delhi and Sanjay Palace, Agra stations for time span of approx. 3 months starting from 12^{th} Feb 2018 till 6^{th} May 2018.

4.2 Evaluation Metrics

We report *Root Mean Square Error (RMSE)* as the evaluation metric. RMSE is the difference between air pollutant values predicted by a model or an estimator and the values actually observed. Since the model is trained on past data, we report the RMSE for future pollutant prediction value. Equation 2 represents the equation used to calculate RMSE, where n is the test sample size.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(Predicted_i - Actual_i\right)^2} \tag{2}$$

Ground truth: We separate training data from testing data by time. We trained our model on 3 months dataset starting from 12^{th} Feb 2018, 8am till 6^{th} May 2018, 7am. Next 1 hour, next 3^{rd} hour, next 6^{th} hour and next 12^{th} hour air pollutants concentration are predicted to check the prediction accuracy. The actual pollutant concentration measured at that time is used as the ground truth to measure the RMSE.

4.3 Experimental Parameters

Our framework is designed with multiple python packages and libraries like tensorflow [1], keras [4] and sklearn library [18]. Encoder and minmaxscaler of sklearn library is used to perform label encoding and scaling on our dataset respectively. The architecture of LSTM is used to predict next # of hour pollutant concentration. There are multiple parameters (number of epochs, hidden layers, hidden units, learning rate etc.) on which LSTM model work. Tuning all these parameters results in different training time or RMSE. We perform many experiments to find optimal value of parameters to get least RMSE and training/prediction time. There are 2 gradient descent optimization algorithms, Adam [12] and SGD [2] (Stochastic Gradient Descent). The advantage of Adam algorithm over SGD is that global minima is achieved in less number of epochs [3] [12]. In SGD algorithm number of epochs used to reach global minima are way more than Adam algorithm as shown in fig. 9 and 10. Classical stochastic gradient descent maintains a single learning rate for all weight updates and the learning rate does not change during training. Adam method computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients [3]. Fig 9 shows the effect of input parameters on number of epochs to achieve global minima using SGD gradient descent optimization algorithm. As the graphs shows, global minima

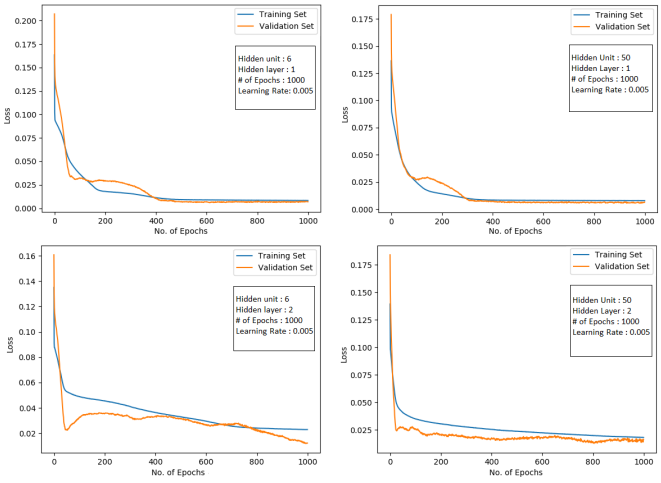

Figure 9: Global minima vs. number of epochs using SGD algorithm

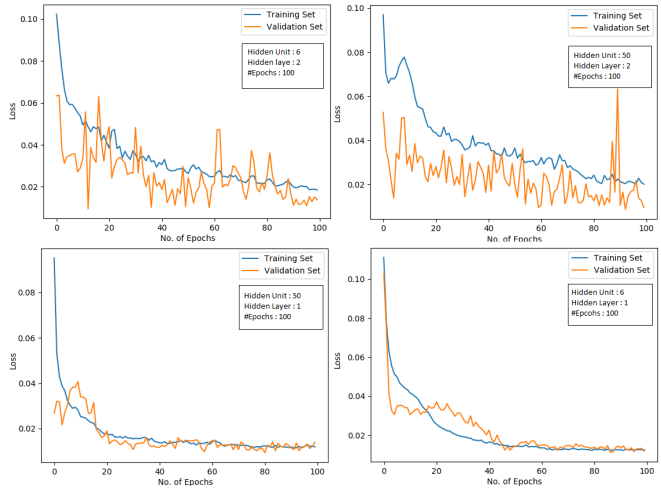

Figure 10: Global minima vs. number of epochs using Adam algorithm

is achieved with least number of epochs when number of hidden units are 50, hidden layer is 1 and learning rate is 0.005. With these values, global minima is achieved with 250 epochs. Fig 10 shows the effect of input parameters on number of epochs using Adam gradient descent optimization algorithm. As shown in figure 10, global minima is achieved in less than 50 epochs with 50 hidden units and 1 hidden layer.

We used Adam algorithm with 1 hidden layer, 50 hidden units and 100 epochs in our framework to obtain global minima. Adam algorithm computes individual adaptive learning rates for different parameters so we don't have to define learning rate in Adam algorithm.

We also performed some experiments to check the effect of input parameters on RMSE. Table 4 shows the experimental results. As shown in the table, least RMSE is achieved with 50 hidden

Table 4: Effect of Input Parameters on RMSE: HU-Hidden Unit, HL-Hidden Layer for ITO, Delhi

Input Parameters				RMSE		
HU	HL	Epochs	Optimizer	RMSE (1hr)	RMSE (6hr)	RMSE (12hr)
6	1	100	Adam	3.162	8.743	14.856
50	1	100	Adam	3.157	7.452	10.644
6	2	100	Adam	19.365	18.206	18.707
50	2	100	Adam	23.993	21.342	21.435

Table 5: Comparison Results of Our Model with Real Time Pollutant Data for Sanjay Palace, Agra

Time		AQI	AQI Class
6^{th} May, 8a.m.	Actual	211	Very UnHealthy
	Predicted	210	Very UnHealthy
6^{th} May, 9a.m.	Actual	210	Very UnHealthy
	Predicted	204	Very UnHealthy
6^{th} May, 10a.m.	Actual	211	Very UnHealthy
	Predicted	201	Very UnHealthy
6^{th} May, 2p.m	Actual	212	Very UnHealthy
	Predicted	198	UnHealthy
6^{th} May, 8p.m	Actual	216	Very UnHealthy
	Predicted	215	Very UnHealthy

Table 6: Comparison Results of Our Model with Real Time Pollutant Data for ITO, Delhi

Time		AQI	AQI Class
6^{th} May, 8a.m.	Actual	265	Very UnHealthy
	Predicted	263	Very UnHealthy
6^{th} May, 9a.m.	Actual	272	Very UnHealthy
	Predicted	266	Very UnHealthy
6^{th} May, 10a.m.	Actual	280	Very UnHealthy
	Predicted	273	Very UnHealthy
6^{th} May, 2p.m	Actual	292	Very UnHealthy
	Predicted	281	Very UnHealthy
6^{th} May, 8p.m	Actual	195	UnHealthy
	Predicted	189	UnHealthy

units, 1 hidden layer, 100 epochs using Adam optimizer. The above experiments are performed on approx. 3 months training dataset (12^{th} Feb 2018 to 6^{th} May 2018).

5 RESULTS

The experimental results of our solution approach are shown in table 5, 6 and 7. We compare our results with real time pollutant data. Central Pollution Control Board (CPCB) [6] website of Indian government is considered as a source of real time data which provides real time pollutant's concentration at each sensor location. We present comparison results for 2 sensor location, Sanjay palace, Agra and ITO, Delhi. Since Delhi is one of the most polluted city

Table 7: RMSE Of Each Pollutant at Sanjay Palace (S.P), Agra and ITO, Delhi [Training Time : 12th Feb 2018, 8am to 6th may 2018, 7am]

Station	Forecast Hour	NO2	SO2	$PM_{2.5}$	PM_{10}	O_3	CO
S.P, Agra	Next 1hr.	0.6	1	2.513	NA	0.43	1.7
	Next 6hr.	0.86	1.8	6.82	NA	0.86	11.3
	Next 12hr.	2.4	3.79	11.4	NA	1.67	15.4
ITO, Delhi	Next 1hr.	2.0	NA	3.183	4.29	NA	NA
	Next 6hr.	4.06	NA	12.37	5.8	NA	NA
	Next 12hr.	10.8	NA	17.58	7.19	NA	NA

in India and Agra is a tourist place, we selected these two stations for our experiments. We trained our model on 3 month historical data (12^{th} Feb 2018,8am till 6^{th} May 2018, 7am) and prediction is made for next 1 hour (6^{th} May 2018, 8 am), next 2^{nd} hour (6^{th} May 2018, 9 am), next 3^{rd} hour (6^{th} May 2018, 10 am), 6^{th} hour (6^{th} May 2018, 2 pm) and 12^{th} hour (6^{th} May 2018, 8 pm). As shown in table, at 2pm, there is difference in AQI class but the difference in AQI value is very less. Table 6 shows the comparison results for ITO Delhi location. Our experimental results show that most of the times, our predicted AQI class is same as actual AQI at that time.
We also report the root mean square error (RMSE) of each pollutant concentration prediction for next 1 hour, next 6^{th} hour and next 12^{th} hour in table 7. Since we are getting forecast meteorological data for next 12 hours only so the prediction is limited to 12 hours. Our model can be extended for further hours if the corresponding data is available. As shown in table 7, prediction accuracy for next 1 hour is better compared to next 6^{th} and 12^{th} hour. Our results shows that compared to other gases, pm2.5 is varying the most at both stations which is validated by our empirical analysis also. At some stations, few pollutant concentration is not available hence the prediction could not be possible for those pollutants and corresponding values are marked as NA.

6 CONCLUSION AND FUTURE WORK

In this work, we predict air pollutant's concentration for prominent Indian stations (Delhi and Agra) based on past air pollutants value, meteorological data, traffic and miscellaneous factors like festivals and national holidays observed at sensor location. We identify features based on the real time dataset and propose a stacked LSTM network based on time series data to forecast pollutants. We evaluated our approach using the real time data obtained from Central pollution Control board, India. The results are shown for 2 sensor locations with RMSE less than 15 for next 12 hour prediction, less than 8 for next 6 hour prediction and less than 5 for next hour prediction. The current work predict air pollutants concentration for sensor location only. We plan to extend this work to come up with a prediction algorithm for predicting pollutants concentration at non-sensor locations.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A System for Large-scale Machine Learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation (OSDI'16)*. USENIX Association, Berkeley, CA, USA, 265–283. http://dl.acm.org/citation.cfm?id=3026877.3026899

[2] Antoine Bordes, Léon Bottou, and Patrick Gallinari. 2009. SGD-QN: Careful Quasi-Newton Stochastic Gradient Descent. *J. Mach. Learn. Res.* 10 (Dec. 2009), 1737–1754. http://dl.acm.org/citation.cfm?id=1577069.1755842

[3] Jason Brownlee. [n. d.]. Gentle Introduction to the Adam Optimization Algorithm for Deep Learning. ([n. d.]). https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/

[4] François Chollet et al. 2015. Keras. https://github.com/fchollet/keras. (2015).

[5] Felix A. Gers, Jürgen Schmidhuber, and Fred A. Cummins. 2000. Learning to Forget: Continual Prediction with LSTM. *Neural computation* 12 10 (2000), 2451–71.

[6] Indian government. [n. d.]. Central Pollution Control Board. ([n. d.]). https://data.gov.in/catalog/historical-daily-ambient-air-quality-data

[7] G. He and Qihong Deng. 2012. A Hybrid ARIMA and Neural Network Model to Forecast Particulate Matter Concentration in Changsha, China.

[8] hindustantimes.com. 2016. Air pollution shortens your life by 3.4 years, Delhiites worst hit. *Hindustan Times* (2016). https://www.hindustantimes.com/mumbai/air-pollution-shortens-your-life-by-3-4-years/story-L9VOawHyX4PCMfCuAjv4ML.html

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[10] Hsun-Ping Hsieh, Shou-De Lin, and Yu Zheng. 2015. Inferring Air Quality for Station Location Recommendation Based on Urban Big Data. In *KDD*.

[11] IndiaToday.in. 2018. 14 of world's most polluted 15 cities in India, Kanpur tops WHO list. *IndiaToday* (2018). https://www.indiatoday.in/education-today/gk-current-affairs/story/14-worlds-most-polluted-15-cities-india-kanpur-tops-who-list-1224730-2018-05\-02

[12] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980 http://arxiv.org/abs/1412.6980

[13] Adam Leelossy, Ferenc MolnÃąr, Ferenc IzsÃąk, ÃĄgnes Havasi, IstvÃąn Lagzi, and RÃ§bert MÃĺszÃąros. 2014. Dispersion modeling of air pollutants in the atmosphere: a review. 6 (09 2014), 257–278.

[14] Andy Liaw and Matthew Wiener. 2002. Classification and Regression by randomForest. *R News* 2, 3 (2002), 18–22. https://CRAN.R-project.org/doc/Rnews/

[15] The Taj Mahal. [n. d.]. The Taj Mahal Official Website. ([n. d.]). https://tajmahal.gov.in/taj_visitors.html

[16] Here Maps. [n. d.]. Here Maps: Location for Developers). ([n. d.]). https://developer.here.com/

[17] Heather Holmes Nancy Murray, Howard H. Chang and Yang Liu. 2018. Combining Satellite Imagery and Numerical Model Simulation to Estimate Ambient Air Pollution: An Ensemble Averaging Approach.

[18] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12 (Nov. 2011), 2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195

[19] Patricio Perez and Jorge Reyes. 2001. Prediction of Particlulate Air Pollution using Neural Techniques. *Neural Computing & Applications* 10, 2 (01 May 2001), 165–171. https://doi.org/10.1007/s005210170008

[20] Central pollution Control Board. [n. d.]. ENVIS Centre on Control of Pollution Water, Air and Noise. ([n. d.]). http://cpcbenvis.nic.in/envis_newsletter/Air%20pollution%20in%20Delhi.pdf

[21] Vikram Simha A Reddy, Pavan S. Yedavalli, Shrestha Mohanty, and Udit Nakhat. 2017. Deep Air: Forecasting Air Pollution in Beijing, China.

[22] reuters.com. [n. d.]. India says is now third highest carbon emitter. ([n. d.]). https://www.reuters.com/article/us-india-climate/india-says-is-now-third-highest-carbon-emitter-idUSTRE6932PE20101004

[23] Surendra Roy. 2012. Prediction of Particulate Matter Concentrations Using Artificial Neural Network. 2 (03 2012), 30–36.

[24] George Sammons. 2016. *Introduction to AWS (Amazon Web Services) Beginner's Guide Book: Learning the Basics of AWS in an Easy and Fast Way.* CreateSpace Independent Publishing Platform, USA.

[25] weforum.org. 2016. 92 % of us are breathing unsafe air. This map shows just how bad the problem is. (2016). https://www.weforum.org/agenda/2016/09/\92-of-the-world-s-population-lives-in-areas-with-unsafe-air-pollution-levels\-this-interactive-map-shows-just-how-bad-the-problem-is/

[26] Kai Zhang and S Batterman. 2013. Air pollution and health risks due to vehicle traffic. *The Science of the total environment* 450-451 (2013), 307–16.

[27] Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. 2013. U-Air: when urban air quality inference meets big data. In *KDD*.