Air Pollution Prediction on Machine Learning

Amrutha V K, Harshita M Raj, Nisha M, Sindhu C, Sumathi M
Department of Information Science and Engineering, Don Bosco Institute of Technology,
Bengaluru, Karnataka, India

Abstract: The interpolation, prediction and feature analysis of fine air-quality are the three main important topics in the urban air computing, the solutions to these topics can provide useful information to support air pollution control. The economic and social impact of poor air quality in towns and cities is increasingly being recognized with the need for effective ways of creating awareness of real-time air quality levels and their impact on human health. With increasing air pollution we need to implement efficient quality monitoring models which collect information about the concentration of air pollutants and provide assessment of air pollution in each area. The system for monitoring and forecasting urban air pollution is presented in this paper. The models are responsible for receiving and storing the data, preprocessing and converting the data into useful information forecasting the pollutants levels. Two machine learning models are used to find accurate forecasting models. The Machine Learning models used are Artificial Neural Network (ANN) and Random Forest Algorithm.

Keywords: Air quality, forecasting, random forest.

I. Introduction

The interpolation, prediction and feature analysis of fine air quality are three important topics in the area of urban air computing. A good interpolation solves the problem that there are limited air quality monitor stations whose distribution is uneven in city, prediction of valuable information helps to protect humans from being damaged by air pollution and feature analysis reveals the variation of air quality. The solution to these topics can extract useful information to support air pollution control, and consequently generate great social and technical impacts [1]. There are insufficient air-quality monitor stations in a city due to the high cost of building and maintaining such a station, it is expensive to obtain labeled data set when dealing with air-quality. The labeled data of air-quality monitor station are incomplete, and there exists a lot of missing labels of the Air Pollution Prediction on Machine Learning data in some periods for some stations. Labeled data are difficult or expensive to obtain, but large amounts of unlabeled examples or data can be gathered cheaply and unlabeled data can help in providing useful information about the data [2].

It is widely believed that urban air pollution has a direct impact on human health especially on developing and industrial countries, where air quality measures are not available or minimally implemented [3]. Recent studies have shown evidences that exposure to atmospheric pollutants

has strong links to adverse diseases including asthma and lung inflammation [4]-[6]. In terms of economic impact, the association between air pollution and human health results in increase of healthcare services. Considering the significance of air quality on human lives, the World Health Organization (WHO) has developed guidelines for reducing the health effects of air pollution on public human health by setting limits of the concentrations of various air pollutants some of which are ground level ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2) [7]. The concentration of air pollutants are measured using air quality monitoring (AQM) stations that are highly reliable, precise, accurate and are able to measure a wide spectrum of pollutants [10].

The main aim of the AQM configuration is to move from mesco-scale to micro-scale coverage that improves spatiotemporal resolution of the collected air pollution data. Next Generation of Air Monitoring (NGAM) can help in reducing the cost of AQM networks, improves the public health by providing better air pollution data [15]. The network uses Artificial Neural Network (ANN) to study the effects of temperature, humidity, etc on pollutants concentrations. Three machine learning algorithms are investigated to get the proper concentration of pollutants that are: Artificial Neural Network (ANN) Support Vector Machine (SVM) and Random Forest algorithm. The performance of models are used to measure the actual value and predicted value of the taken data.

II. Related Work

Prediction of air quality levels is important for communicating pollution risks exposure level. However, it is a complex measure to calculate since the forma and dispersal patterns of pollutants are affected by environmental and meteorological factors. The early approach was human centered, where data collected from different monitoring stations were evaluated based on human experience. Hence it was unreliable. Currently computing approaches involve use of algorithms

Air Pollution Prediction on Machine Learning such as decision tree, neural network .For example based on 383 Chinese cities, Ye et al, found that although regional discrepancies exist, concentrations in locations where storms frequently occur are especially high. In accordance with the date of the Beijing-Tianjin-Hebei region in2014, Liu et al. concluded that spatial patterns of PM 2.5 concentrations in this area appeared to be high in the southeast and low in the northwest, and pollutants primarily arise from automobile emissions and coal burning. Feng et al. Researched the PM2.5 concentration data with a 6-hour-renewal interval in the Yangtze River Delta region and concluded that, instead of vehicle emissions, the burning of biomass mainly gives rise to high PM2.5 concentrations.

Studies focusing on small scale cities have been published, but with significantly fewer in number. Liu et al. used the data of seventeen nationally controlled monitoring stations from 2014 to 2015 in Chongqing Province and concluded that PM2.5 concentrations in winter were highest, followed by autumn and spring, and with summer having the lowest concentrations. Kang et al. utilized the data from 4 cities in Heilong Province during 2014 and showed that PM2.5, the main source of pollution was negatively correlated with humidity, wind speed, and precipitation, while it was positively correlated with the air pollution level and that areas near water have a negative correlation with the air pollution level.

This proved the importance of preserving water areas for gaining a better air quality. The shortcomings of the existing PM2.5 studies based on ground monitoring station data are as follows. Firstly, the studies focused on large-scale regions, such as mainland China or metropolitan areas, with little research on small-scale areas, such as prefectures. With an increase in regional differentiation, analysis on small district cities can be more valuable. Secondly, the existing studies used mostly rough and historical data, causing the research results to lack accuracy and practicality. have been recorded since 2012. Studies based on a time span longer than 5 years are few, and most of them used data retrieved from remote-sensed imagery. International studies focusing on exposure to air pollution such as total exposure assessment methodology (TEAM) studies, the National Human Activity Pattern Survey (NHAPS) and the Population Exposure to Air Pollutants in Europe (PEOPLE) project relied on diary- based instruments (e.g., time- activity diaries (TADs), questionnaires, California Household Travel Survey, National Household Travel Survey, etc.) to categorize the environments where exposure occurred and sources of air pollutants, and to derive information on the temporal sequencing of human activities during the study period. However, such time- activity information does not account for the movement of the individual

and mostly lacks the exact “activity- space” where a specific activity is executed by the individual and consequently, the sequence of exposure events is not considered (Figure 1).

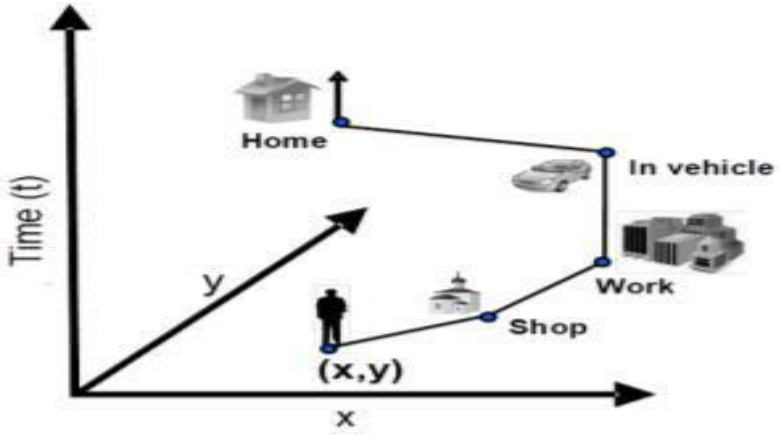

Figure : Trajectory of an individual in space (x, y) and time (t).

To overcome some of the uncertainties related to the human mobility during the exposure assessment period, the availability of GPS for human tracking presents an enormous opportunity for our understanding of how time-activity patterns can influence individual exposure and subsequent health effects. GPS is a freely accessible and promising technology which may answer crucial questions. Real time geographic positions, thus providing new insights in the field of personal exposure assessment to air pollution in urban areas. A literature review on various approaches currently available to quantify individual-level exposure to urban air pollution was conducted based on journal articles published in English from 2006 to June 2017 and indexed by ISI and SCOPUS. The search was performed considering the combination of three hundred and sixty four articles were identified from web of Science database.

III. Machine Learning Approach

ML involves computational methods that improve the performance of acquiring knowledge from experience [18]. Machines learn from complex data to be able to solve problems, answer questions and be more intelligent.

Air Pollution Prediction on Machine Learning

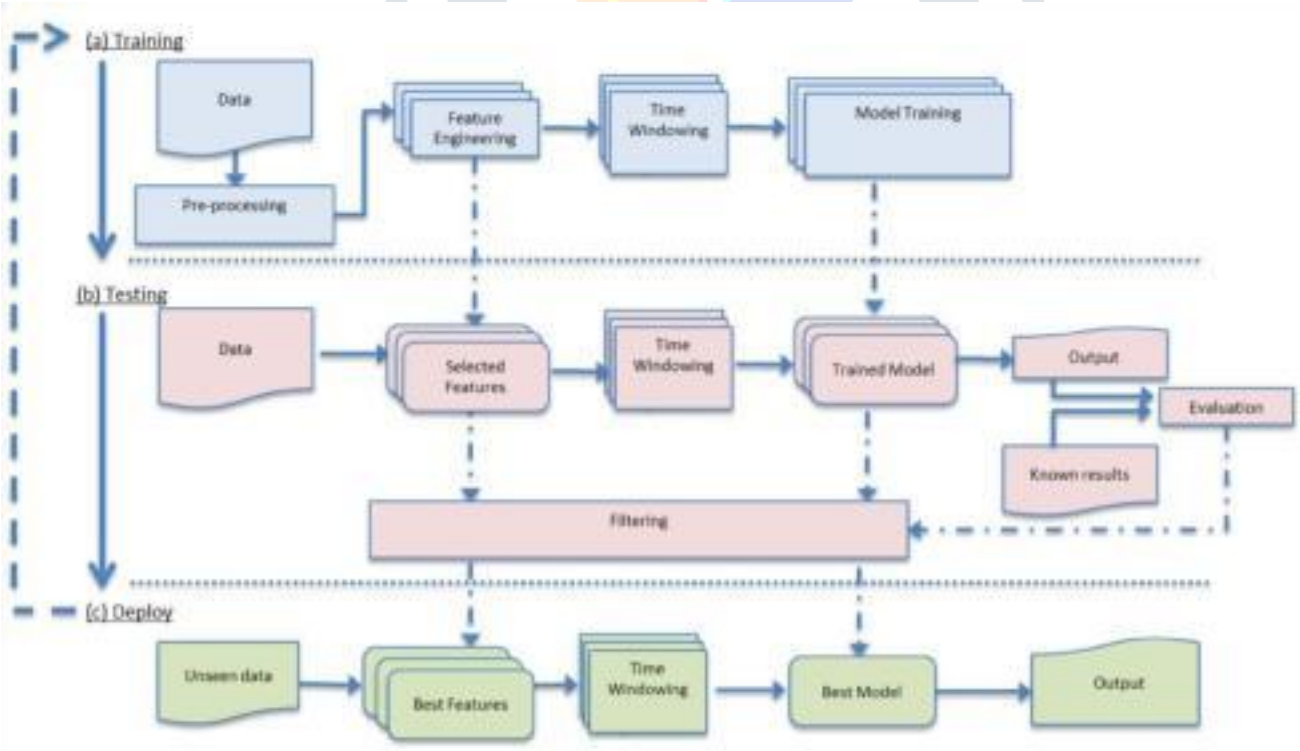

Figure 1: Process of constructing and predicting ML based models

In training, data with known target values are collected, a feature is selected, and then used to construct a model. There are many features that are selected and various ML algorithms are used to predict the values of data.

In testing, models from the training phrase are validated and evaluated. Several methods are used in model validation, such as different sliding windows, in which two windows are used for training and testing which has its own size. This validation method guarantees that instances used for testing are not known before to the model through training, hence reliable performance measures are calculated such as prediction trend accuracy (PTA) and root mean square error (RMSE).

PTA is a time series measurement of how close is the predicted data trend of the actual data. The actual value and predicted value are calculated as,

Actual value = label [i] – label [i-horizon]

Predicted value = Predicted [i] – label [i-horizon]

RMSE is common performance metric in model evaluation, and is calculated as [29],
Air Pollution Prediction on Machine Learning

RMSE =(i=1 to n)(y-y^)2/n

Where n is the number of instances, y is the actual value and y^ is the predicted value of a feature.

Normalized RMSE (NRMSE) is used to compare the performance of models predicting different target variables and is calculated as [30],

RMSE = NRMSE (ymax ymin)

Where ymax and ymin are the maximum and minimum values of collected data respectively.

In the deployment phase, the best model and features will be used to process the data and produce predicted results. The model performance is kept on check to validate its prediction results. In changing environment, the process of training, testing and deployment are periodically repeated to maintain high accuracy of reults.

The most successful algorithms used are ANN, SVM.

ANN is a network nodes connected via different layers, input layer, hidden layer, and output layer. In a feed forward neural network, input data are fed to the nodes in the input layer and then it is propagated through the network passing by the hidden layer nodes and then to the output layer [20]. The value of each hidden node is calculated as the activation function of total weighted input of that node [21]. The activation function used is the sigmoid function. ANN uses a back

propagation algorithm to train the network. In this phase, the output of the network is compared to the actual output.

SVM is a supervised learning method that can be used for solving classification and regression problems. SVM regression aims to find an approximation to a non-linear function that maps the input data in high dimensional space.

Random forest is an advanced method, used to model the regression relationships between concentrations of the pollutants NO2, NOx and PM2.5 and other variables describing meteorological conditions, temporal conditions and traffic flow. The most important explanatory variable in the models of concentrations of nitrogen oxides was traffic flow, while in the case of PM2.5 the most important were meteorological conditions, in particular temperature, humidity. Temporal variables were found no significant effects on the concentrations of the pollutants.

Air Pollution Prediction on Machine Learning

IV. Methodology

Our methodology consists of following steps:

1. **Data Preprocessing**: Here, data are cleaned by removing anomalies and outliers. Data are also prepared to be in proper format for the ML algorithms to use. In this step, the erroneous and missing data are estimated and replaced with new data points using interpolation process.
2. **Feature analysis:** This step is concern with selecting the features to be included in the prediction processing along with each target gas such as temperature, humidity etc. The included feature includes two modeling methods such as,
   a. Univariate Modeling: In this step, only the concentrated value of one gas/ feature is used.
   b. Multivariate Modeling: Here different features are used.
3. **Model Training and Testing:** The collected, preprocessed data are used to train prediction models based on ML algorithms including ANN, SVM, Random forest algorithm. In testing, the produced data are validated and evaluated.
4. **Building Models:** After validating and evaluating the data best feature is selected and model is designed.

V. Result and Discussions

For the proposed AQI, a maximum operator system is selected:

AQI = Max (I1, I2, I3,……, In)

There are two reasons for adopting a maximum operator:

1. Free from eclipsing and ambiguity (Ott 1978).
2. Health effects of combination of pollutants (synergistic effects) are not known and thus a health-based index cannot be combined or weighted.

Air Pollution Prediction on Machine Learning

AQI Category, Pollutants and Health Breakpoints

AQI Range	**PM10 (24hr)**	**PM2.5 (24hr)**	**NO2 (24hr)**	**O3 (24hr)**	**CO (24hr)**	**SO2 (24hr)**	**NH3 (24hr)**	**Pb (24hr)**
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.5-1.0
Moderately Polluted (101-200)	101-250	61-90	81-18/0	101-168	2.1-10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10-17	381-800	801-1200	2.1-3.0
Very Poor (301-400)	3.51-430	121-250	281-400	209-748	17-34	801-1600	1200-1800	3.1-3.5
Severe (401-500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

AQI Range and their associated Health Impacts on Humans

AQI Range	Associated Health Impacts
Good (0-50)	Minimal Impact
Satisfactory (51-100)	May cause minor breathing discomfort to sensitive people.
Moderately Polluted (101-200)	May cause breathing discomfort to people with lung disease such as asthma, discomfort to people with heart disease, children and older adults.
Poor (201-300)	May cause breathing discomfort to people on prolonged exposure, and discomfort to people with heart disease.
Very Poor (301-400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart disease.

Air Pollution Prediction on Machine Learning

Severe (401-500)	May cause respiratory impact even on healthy people, and serious health impact on people with lung/heart disease. The health impacts may be experienced even during light physical activity.

Actual and predicted values for some parameters

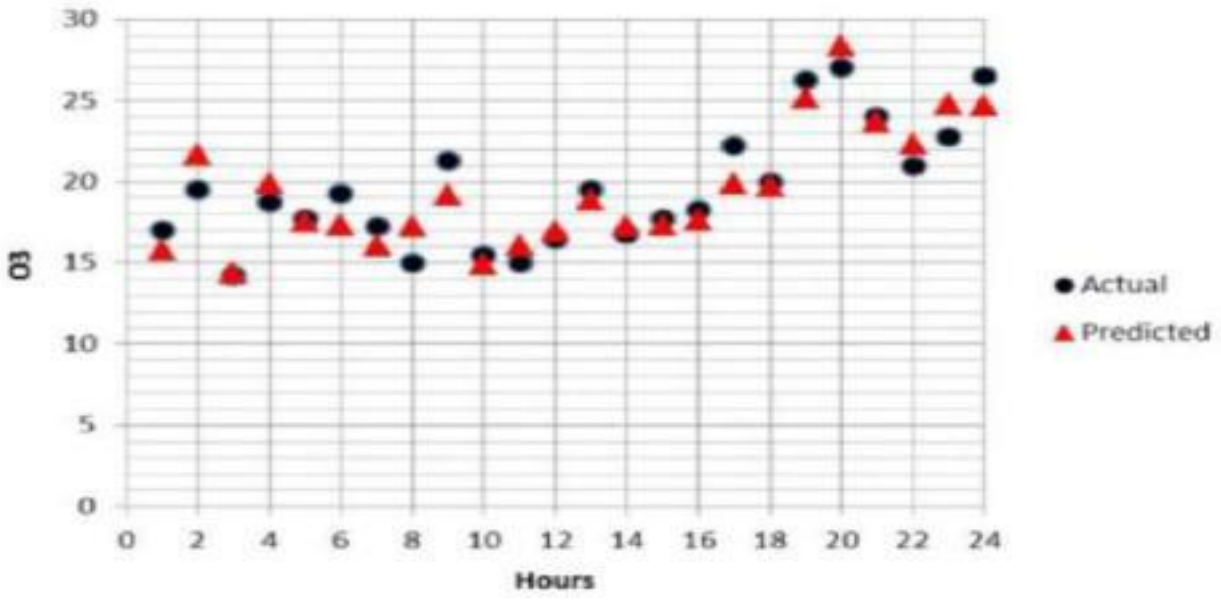

Figure 1.2 Actual and Predicted values of O3

Air Pollution Prediction on Machine Learning

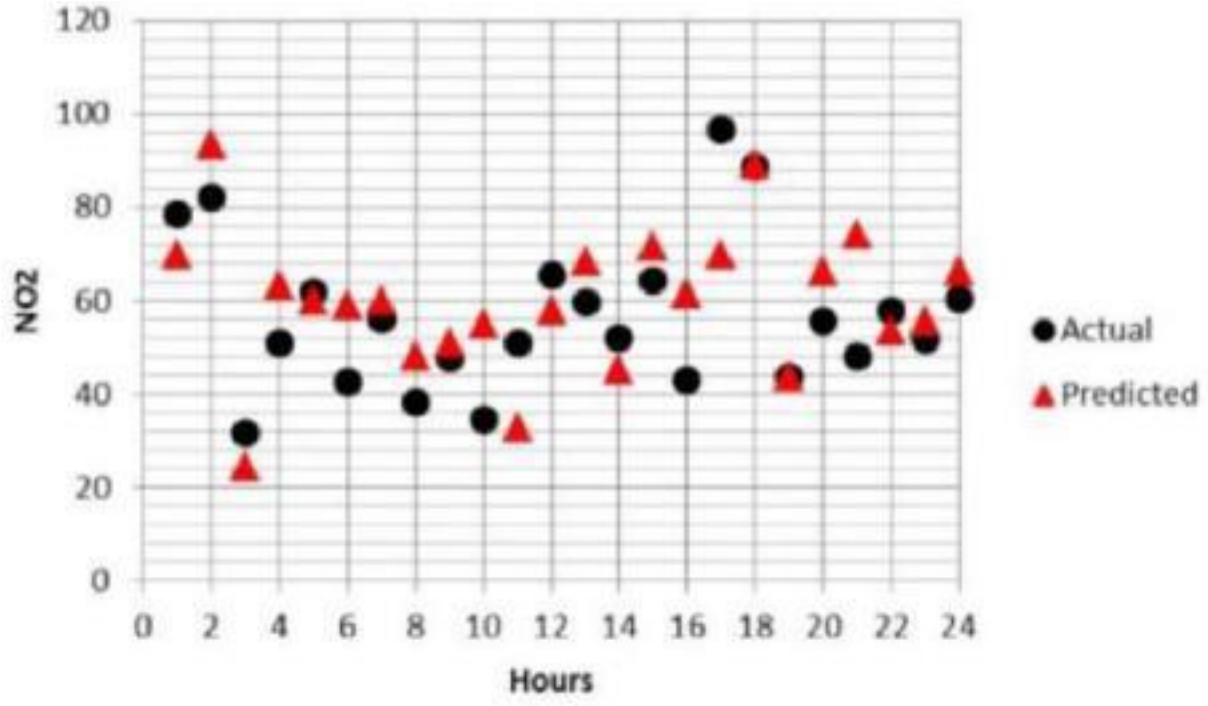

Figure 1.3 Actual and predicted values of NO2

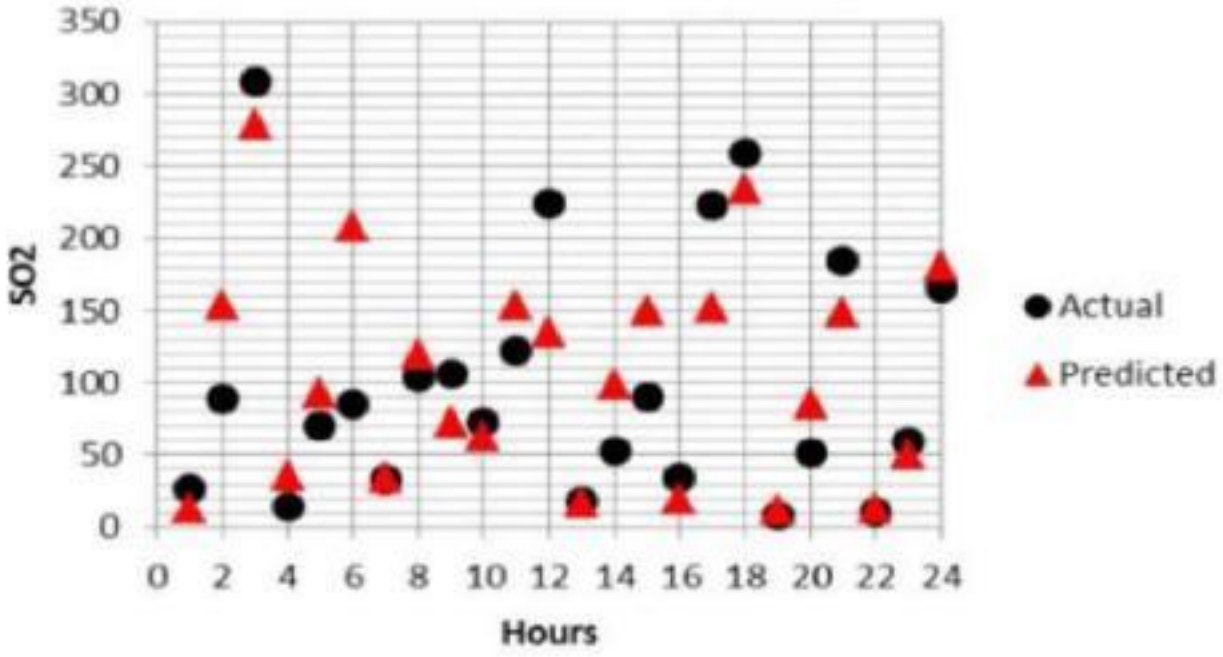

Figure 1.4 Actual and Predicted values of SO2

Air Pollution Prediction on Machine Learning

VI Conclusion

In this paper we are using efficient machine learning methods for air pollution prediction. Air quality is an important problem that directly effects on human health. Air quality data are collected from monitoring models. These data collected are analyzed and used in forecasting concentration values of pollutants using machine to machine platform. The platform uses ML algorithms to build the models by learning from collecting the data. These models predict 1, 8, 12, and 24 hours ahead of concentrating values. ANN algorithm is sufficient for small dataset and it leads to a complex network that overfit the data, while having SVM and Random Forest algorithm better than ANN as it works good for larger dataset and gives high accuracy for the concentrated values. This paper studies three important topics in the urban air computing: the interpolation, prediction and feature analysis of air quality. The solution to these topics can provide crucial information to support air pollution control, and consequently provides great social and technical impacts.

References

1. Y. Zheng, F. Liu and H. P. Hsieh, "U-air: When urban air quality inference meets big data," in Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining, ser. KDD '13, 2013, pp. 1436-1444.
2. H. P. Hsieh, S. D. Liu and Y. Zheng, "Inferring air quality for station location recommendation based on urban big data, "in Proceeding of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '15, 2015, pp. 437-446.
3. World Health Organization, "Monitoring ambient air quality for health impact assessment," WHO Regional Officer Eur., Copenhagen, Denmark, Tech. Rep. 85, 1999.
4. U. Gehring et al., "traffic-related air pollution and the development of asthma and allergies during the first 8 years of life," Amer J. Respiratory Cricital Care Med., vol. 181, no.6, pp. 596-603, 2010.Air Pollution Prediction on Machine Learning
5. L. E. Plummer, S. Smiley-Jewe H and K. E. Pinkerton, "Impact of air pollution on l;ung inflammation and the role of toll-like receptors," Int. J. Interferon, Cytokine Meditor Res., vol. 4, pp. 43-57, May 2012.
6. International Agency for Research on Cancer (IARC), "Outdoor air pollution a leading environmental cause of cancer deaths," World Health Org., Geneva, Switzerland, tech, Rep. 221, 2013.

7. World health Organization, “WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide, sulfur dioxide,” WHO.,Geneva, Switzerland, Tech. Rep. WHO/SDE/PHE/OEH/06.02, 2005.
8. M. F. Othmana and K. Shazalii, “Wireless sensor network applications: A study in environment monitoring system,” Proc. Eng., vol. 41., pp. 1204-1210, Aug. 2012.
9. M. I. Mead et al., “The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks,” Atoms. Environment, vol. 70, pp. 186-203, May 2013.
10. J. Y. Kim, C. H. Chu and S. M. Shin, “ISSAQ. An integrated sensing system for real-time indoor air quality monitoring,” IEEE Sensors J., vol. 14, no. 12, pp. 42230-4244, Dec. 2014.
11. M. A. Hearst, S. T. Dumais, E. Osman, J. Platt and B. Scholkopf, “Support Vector Machines,” IEEE Intell. Syst. Applications, vol. 13, no. 4, pp. 18-28, Jul/Aug. 2008.
12. C. M. Bishop, Neural Networks for pattern recognition. Oxford, U. K. Khaled Bashir Shaban received the Ph. D. degree Clarendon Press, 1995 from the Department of electrical and Computer.
13. X. Zhou, W. Huang, N. Zhang, W. Hu, S. Du, G. Song and K. Xie, “Probabilistic dynamic casual model for temporal data,” in Neural Networks (IJCNN), 2015 International Joint Conference on, July 2015, pp. 1-8.