Analysis and Prediction of Air Quality

Reshma J

Assistant Professor, Department of Computer Science and Engineering, BNM Institute of Technology, Bengaluru, India

---***---

Abstract – In the modern world, air pollution is one of the growing environmental issue. It poses major threat to health and climate. Air quality in cities is deteriorating day by day. Pollutants in air also impacts both water and land. This paper presents a literature review of predicting air quality using Internet of Things and Machine learning techniques and forecast the air pollution levels so that preventive measures can be taken by the people in order to minimize the air pollution.

Key Words: Air pollution, criteria pollutants, climatology, air prediction, Machine Learning, Deep Learning, Artificial Neural Network

1. INTRODUCTION

Air pollution is the introduction of particulates, biological molecules or other harmful materials into the Earth's atmosphere. It causes disease, death to humans, damage to other living things or damage to the environment. Monitoring and preserving air quality is one of the most essential activities in many industrial and urban areas today. An air pollutant is a substance in the air that can create lot of adverse effects on humans and the ecosystem. The substance can be solid particles, liquid droplets or gases. The quality of air is affected due to many forms of pollution caused by transportation, electricity, fuel uses etc. The deposition of harmful gases including carbon dioxide, nitrous oxide, methane etc. is creating a serious threat for the quality of life in smart cities. According to a recent report European Environment Agency, by the year 2040, CO_2 and NO_x emissions from aviation are expected to increase by 21 and 16 per cent.

Pollutants are basically classified into primary or secondary pollutants. Primary pollutants are usually produced from a volcanic eruption. Other examples include carbon monoxide gas from motor vehicle exhaust or sulphur dioxide released from factories. Secondary pollutants are not emitted directly. Rather, they form in the air when primary pollutants react or interact. Ground level ozone is a prominent example of a secondary pollutant [1].

The six "criteria pollutants" are ground level ozone (O_3), fine particulate matter (PM2.5), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2) and lead. Among the above six pollutants, ground level O_3, PM2.5 and NO_2 (main component of NO_2) cause the most widespread health threat. Ground level O_3, a gaseous secondary air pollutant formed by complex chemical reactions between NO_x and volatile organic compounds (VOCs) in the atmosphere can have negative impacts on human health. Prolonged exposure of O_3 concentrations over certain level may cause permanent lung damage, asthma and other respiratory illnesses. Ground level O_3 can also have harmful effects on plants and ecosystems including reduction of crop yield, damage to plants and increase of vegetation vulnerability to disease.

Fine particulate matter (PM2.5) consisting of particles with diameter 2.5μm or smaller is an important pollutant among the above six pollutants. The microscopic particles in PM2.5 can penetrate deeply into the lungs and cause health problems including the decrease of lung function, development of chronic bronchitis and nonfatal heart attacks. Fine particles can be carried over long distances by wind and then deposited on ground or water through dry or wet deposition. The wet deposition is often acidic as fine particles containing sulphuric acid contribute to rain acidity or acid rain. The effects of acid rain include changing the nutrient balance in water and soil, damaging forests and farm crops and affecting the diversity of ecosystems [2].

Nitrogen oxides (NO_x) especially nitrogen dioxide (NO_2) are emitted from high temperature combustion. Nitrogen dioxide (NO_2) is one among the group of highly reactive gases known as "nitrogen oxides" (NO_x). NO_2 forms quickly from emissions of automobiles, power plants and off-road equipment. In addition to contributing to the formation of ground-level ozone and fine particle pollution, current scientific evidence links short-term NO_2 exposures ranging from 30 minutes to 24 hours will give adverse respiratory effects including airway inflammation in healthy people and increased respiratory symptoms in people with asthama.

Carbon monoxide (CO) is colourless, odourless, non-irritating but very poisonous gas. It is a product obtained by incomplete combustion of fuel such as natural gas, coal or wood. Vehicular exhaust is a major source of carbon monoxide. SO_2 is produced because of volcanoes and also by various industrial processes. Since coal and petroleum contains Sulphur compounds, their combustion generates sulphur dioxide. Further oxidation of SO_2, usually in the presence of a catalyst such as NO_2 forms H_2SO_4 and thus acid rain. The rise in the permissible concentrations of different pollutants is observed in past few years, adding up to the increasing pollution level. As a fact, the affected air quality level in the atmosphere impacts health of individual and may imbalance the economy. With increasing air pollution, it is necessary to implement efficient air quality monitoring models which collect information about the concentration of air pollutants and provide assessment of air pollution in each area. Hence, air quality evaluation, monitoring and prediction has become an important research area [3].

In the past, many environmental researchers have dedicated their research efforts on this subject using conventional approaches. However, the quality of air is affected by multidimensional factors including location, time and uncertain variables. Recently, many researchers began to use the big data analytics approach for studying, evaluating and predicting air quality due to the advancements in big data applications and the availability of environmental sensing networks and sensor data. One of the prime concerns for researchers is the use of adequate modelling tools that permit interpretation and validation of the data collected from multiple resources regarding air pollution. The various methods like 'Climatology' (based on the assumption that the past is a good predictor of the future) have been used for air quality forecasting. These approaches are usually used to predict exceeding limits from specific thresholds, not ambient concentrations. As a result, a lot of improvement is still required in this field for prediction analysis. With incomplete data parameters and their significance (priority), most of the methods fail to predict the pollution levels significantly. By using the regression and machine learning techniques, a collection of functions can be evaluated and utilized in order to fit the data of pollution as the selected predictors. The different pollutants present in the atmosphere can be observed, dispersed or concentrated during varied time period such as multi - levelled equations, graphics and tables which are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

Machine Learning provides one approach that can offer new opportunities for prediction of air pollution. Machine learning is a scientific way of getting computers to act without giving explicit instruction; it learns and improves based on experience of the previous data or instructions. It is the most emerging and commonly used field across the world. Machine learning algorithms works on a simple data called training data. This data is nothing but previously existing data collected and stored for long period. Data Mining techniques is used for the collection of the data. Machine learning system is trained based on this training data; the systems accuracy increases with the increase of training data. Machine learning tasks can be classified as supervised learning and unsupervised learning [4].

Supervised learning provides a function that maps from input to output based on training examples. The training examples consists of labelled training data. Supervised learning includes classification and regression. Random forest and Support Vector Machines are some of the most common algorithms of supervised learning.

Unsupervised learning is a kind of machine learning algorithm which make use of unlabeled data to draw some patterns or inferences. Unsupervised learning includes clustering and association. K-means algorithm is the most common algorithm used for clustering [4].

2. RELATED WORKS

Air pollution has been eventually increasing day by day and it is a very important problem that is to be considered. Prediction of air pollution becomes an important factor to control air pollution. It can affect individuals and their health and can also cause certain diseases like asthma, lung infections, breathing problems and so on. Asthma patients suffer very much by air pollution. Machine learning is a technique that provides unique and new methods for the prediction and analysis of air pollution. In this paper by Ziyue Guan [5], the air pollution data, particularly the particulate matter of less than 2.5 micro-meters (PM2.5) was collected from multiple web - based resources and after data cleaning, it is analysed with various machine learning models including linear regression, Artificial Neural Networks (ANN) and Long Short Term memory (LSTM), recurrent neural networks (RNN). Long Short Term Memory performed best and was able to predict high PM2.5 values with high accuracy. Whereas Artificial Neural Network and Linear Regression offered reasonable overall performance. The issue of the Long Short Term Memory is regarding the slow training process compared to other algorithms.

Smart cities (SCs) help in making the city services and monitoring of cities more aware, interactive, efficient and helps in perception of a situation or fact [6], Internet of things (IOT) is known as a system of interrelated computing devices, mechanical and digital machines and objects that are provided with unique identifiers and the ability to transfer the data over a network without requiring human-to-human or human-to-computer interaction. Since, there is an enormous change in the environment, climate and weather conditions, many cities are not able to elude from the problems that are influenced by air pollution problem. Air Quality Index (AQI) is used for extrapolating the local air quality that tell us how clean or unhealthy the air is. Therefore, forecasting parameters of air pollution such as ozone and nitrogen dioxide is important. While there are many gases that are harmful to human health and the environment, Nitrogen Dioxide (NO2) and ozone (O3) are of great concern. The main contribution of Ibrahim KOK [6] is bipartite: First they accustom a Deep learning (DL) model that can be applied to smart cities IOT data. Second, they design and implement a novel long short term memory (LSTM) based prediction model that makes it easier to solve future air quality problems in Smart cities. The development of Smart Cities has benefitted researchers to target their efforts on enhancing to manage resources of Smart Cities. The results indicate that Long Short Term Memory based prediction model takes its advantages of memorizing of long historical data and achieve higher prediction accuracy even with the simple network structure.

Deep Learning models are discerned as powerful models for showing excellent performance on difficult learning tasks. Urban air prediction becomes a substantial alternative to restrain its detrimental consequences. In recent times, the accelerated growth, urbanisation and improved lifestyle have intensified the air pollution in urban areas considerably. In the paper by Chavi Srivastava [7] they make use of meteorological features for predicting the air quality. Because the meteorological data is easily available for urban areas. In this work, various machine learning models, namely Linear Regression (LR), Stochastic Gradient Descent (SGD) Regression, Random Forest Regression (RFR), Decision Are Regression (DTR), Support Vector Regression (SGR), Multi-layer Perceptron (MLP) or neural Networks, Gradient Boosting Regression (GBR) and Adaptive Boosting Regression (ABR) are exploited to predict the levels of pollutants. The Multi-layer Perceptron gave least errors in estimation and provided maximum accuracy with fair-low range of errors compared to others models. This technique would be helpful for the requisite authorities in taking adequate measures and providing information to the general public as safety actions. The dataset was used for shorter duration which limits the capability of the model. The meteorological factors used were less that affected the accuracy of the system.

Since several decades, industrial pollution and especially atmospheric pollution receive an important interest and concern due to the fact that industries contain more and more pollutants that are very hazardous. In this paper by Nadjet_Djebbri and Mounira Rouainia [8], they aim to predict industrial pollution by non-linear auto regressive model (NARX) based Artificial Neural Network (ANN) by studying the influence of inclusion of meteorological variables (WD,WS,T and HR) to predict concentrations of two pollutants COx and NOx. The database used to train the neural network corresponds to historical time series of meteorological variables wind speed, wind direction, temperature and relative humidity and concentration of pollutants in the petrochemical plants. The inclusion of meteorological variables with pollutant concentration leads to a better prediction performance and great accuracy in numerical approximation. It is an efficient way to estimate pollutants concentration by improving performance of model. The disadvantage is that the unclear trend and wide fluctuations of air pollutants affects the performance and accuracy of the model.

In this paper by Ping-Wei Soh, Jia-wei Chang, Jen-Wei Chang [9] they use Dynamic Time Warping (DTW), convolutional neural network (CNN), LSTM, spatio - temporal analysis and big data for air quality forecast. The model aims to forecast air quality for up to two days using a combination of multiple neural networks including an artificial neural network, convolutional neural network and long short term memory to extract spatial-temporal relations. The proposed predictive model considers various meteorology data from the previous few hours as well as information related to the elevation space to extract terrain impact on air quality. The model includes trends from multiple locations extracted from correlations between adjacent locations. The advantage of proposed model is that including an long short term memory module enhanced the first hour prediction with convolutional neural network module inclusion being more useful for longer time frame prediction. Issue is with the inclusion of all locations causing increased model noise and hence poorer prediction performance.

In [10], a model is proposed based on the Multiple Linear Regression (MLR) and Support Vector Regression (SVR) to predict air pollutants. These models prove to be the successful approach to monitor the quality of air in many major cities using environmental data. Support Vector Regression and Linear Modules like multiple linear regression consisting of gradient descent, stochastic gradient descent and mini-batch gradient descent were implemented. The air quality index is dependent on pollutant concentration of NO2, CO, O3, PM2.5, PM10 and SO2. Support Vector Regression outperforms the other regression models with the highest accuracy. It makes the separation of data samples much easier that increase the accuracy of regression and reduce the generalization error. The issue with the model is that Support Vector Regression is not suitable for large datasets and it has more noise.

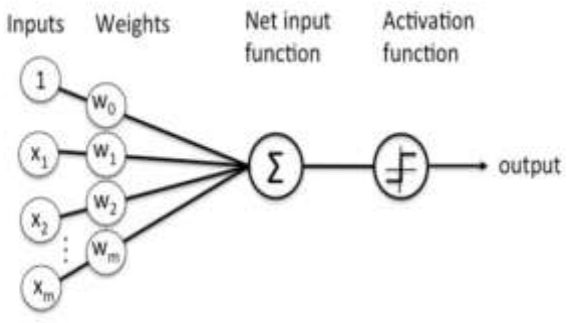
In this paper by Adven Masih [11], Random Forest Algorithm was used to predict the concentration of Nitrogen Dioxide (NO2) in air. The dataset of emission inventory and meteorological parameters was analysed as input predictors. The model performance was compared with two other classification algorithms known as M5P and Support Vector Machine. The performance of Random Forest to predict the concentration of Nitrogen Dioxide (NO2) is better than the other two approaches.

Temesegan Walelign Ayele proposes an IOT based air prediction model using a machine learning algorithm called Recurrent Neural Network. Here, the data is collected from DHT11 sensor and generates real time digital temperature and humidity. The implementation is carried out using python using tensorflow backend. The results demonstrate that it has a quick convergency with reduced training cycles and yields good accuracy [12].

From the above survey, it can be concluded that in comparison to other methods, Artificial Neural Network is found to be an appropriate method of forecasting air pollutant. Artificial Neural Network models are found to be accurate in predicting the air pollutant concentrations. The model can be interfaced with the web application for the user that could benefit from the work and take precautions in order to minimize the air pollution.

3. METHODOLOGY

Artificial Neural Network is a connected network made up of simple processing units called the neurons and large number of weighted links. A perceptron is a single layer neural network and multi-layer perceptron is called Neural Networks. The acquired knowledge is stored in neurons and the signal and information are passed over the weighted links of the large network [13]. An artificial neurons take multiple inputs and gives out one output as seen in Fig 2.1



2.1 Artificial Neural network structure

The model operates in two modes namely learning mode and testing mode. The neurons are modelled such that it consists of inputs that are multiplied by weights and then calculated by a mathematical function. This calculation subsequently determines the activation of the neurons. Another function is then used to calculate the output of the artificial neuron. Higher the weight of the artificial neuron, stronger is the input with which it is multiplied. The process of adjusting the weights is called learning or training the model [13]. By feeding the network with a data set of desired inputs and set of desired outputs, the network can be made to repeatedly adjust the weights of each interior link to model more accurately and determine the correct output for a given input. By exploiting the behaviour of input and output of network, it can be trained with known data to predict the outcomes for new data. The network thus trained can be used to predict the outcome of new independent input data. The data sets that are used in network can now be divided into two distinct sets called training and testing sets. The largest amongst the sets is the one for training and is used by neural network to study patterns present in the data. The testing set is then used to by the network constructed to evaluate the generalization capability of the trained network [13].

4. CONCLUSION

From the above survey, it can be concluded that in comparison to other methods Artificial Neural Networks (ANN) is found to be the appropriate method for forecasting the air pollutants. Artificial Neural networks is found to be accurate in predicting the air pollution. Air pollutants concentration like SO2, NO2, O3, CO and particulate matter have been used for prediction as output. The meteorological factors like temperature, relative humidity, absolute humidity, wind direction can be used for prediction as input variables. Furthermore, the model can be interfaced with the web applications for the user that could benefit from the work and take precautions in order to minimise the air pollution [14].

REFERENCES

[1] www.wikipedia.com

[2] Gaganjot Kaur, Jerry Zeyu, Shengqiang Lu, “Air Quality Prediction: Big data and Machine Learning Approach”, Index of Community Socio-Educational Advantage, pp. 150-158, May 2017.

[3] Rajeev Tiwari, Shuchi Upadhyay, Parv Singhal, “Air Pollution Level Prediction System”, International Journal of Innovative Technology and Exploring Engineering, vol.8, pp. 201-207, April 2019.

[4]https://expertsystem.com/machine-learning-definition/

[5] Ziyue Guan, Richard O. Sinnott “Prediction of Air Pollution through Machine Learning Approaches on the Cloud”, Institute of Electrical and Electronics Engineers International Conference, Zurich Switzerland, pp. 51-60, December 2018.

[6] Ibrahim KOK, Mehmet Ulvi, Suat Ozdemir, “A Deep Learning Model for Air Quality in Smart Cities”, Institute of Electrical and Electronics Engineers International Conference, Boston USA, pp. 1983-1990, December 2017.

[7] Chavi Srivastava, Amit Singh, Shymali Singh, “Estimation of Air Pollution in Delhi using Machine Learning Techniques”, Institute of Electrical and Electronics Engineers International Conference, Noida India, pp. 304-309, September 2018

[8] Nadjet Djebbri, Mounria Rouainia, “Artificial Neural Networks Based Air Pollution Monitoring in Industrial Sites”, Institute of Electrical and Electronics Engineers International Conference, Antalya Turkey, August 2017.

[9] Ping Wei Soh, Jia Wei Chang, Jen Wei Huang, “Adpative Deep Learning-Based Air Quality Prediction Model using the most Relevant Spatial-Temporal Relations”, Institute of Electrical and Electronics Engineers, vol. 6, pp. 38186-38199, June 2018.

[10] Sankar Ganesh, Sri Harsha Modali, Soumith Reddy, “Forecasting Air Quality Index using Regression Models”, Institute of Electrical and Electronics Engineers International Conference, Tirunelveli India, pp. 248-254, May 2017.

[11] Adven Masih, “Application of Random Forest Algorithm to Predict the Atmospheric Concentration of NO2”, Institute of Electrical and Electronics Engineers International Conference, Yekaterinburg Russia, April 2019.

[12] Temesegan Walelign Ayele, Rutvik Mehta, “Air Pollution Monitoring and Prediction using IOT’, Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT) Coimbatore, India, pp.1741-45, April 2018.

[13] https://towardsdatascience.com

[14] C. Ma, Real time Mobile Pollution Detection Monitoring for Improved Route Planning, Master Dissertation, University of Melbourne, 2016. 38.