A Survey of Machine learning Techniques in Air Pollutants and Hotspot Area Detection

Urvashi Parmar[1], Yoothika Patel[2], Mosam Patel[3]

Department of Computer Science

Parul Polytechnic Institute, Parul University, India

Email: *urvashi.parmar14384@paruluniversity.ac.in [1], yoothika.patel270041@paruluniversity.ac.in[2], mosam.patel270045@paruluniversity.ac.in[3]*

DOI:- https://doi.org/10.47531/SC.2022.12

Abstract

Air pollution worldwide is a growing threat to human health and the natural environment. The earth's atmosphere is full of air which contains gases such as Nitrogen, Oxygen, Carbon Monoxide and traces of some rare elements. Humans need an atmosphere of air free from contaminants, a condition crucial for human life and health. Any change in the natural composition of air may cause grave harm to life forms on earth. Air pollution is the presence of one or more contaminants in the atmosphere, such as gases in a quantity that can harm humans, animals and plants. This review focuses on the research done in data mining and machine learning for predicting and detecting the quality of air and air pollution hotspots.

Keywords: - *Air pollution, Hotspot area, Machine learning, Data mining*

INTRODUCTION

Air pollution is considered to be one of the most dangerous and common types of environmental pollution that have been reported in most industrial towns and metropolitans abroad and India such as New York, London, Tokyo, Mumbai, Delhi, Kolkata, Chennai, Jaipur, Hyderabad, Nagpur, Ahmedabad, Kanpur etc. Air pollution is generally a disequilibrium condition of air caused due to the introduction of foreign elements from natural and anthropogenic (man-made) sources that make it harmful or detrimental to biological communities.

Air pollution refers to any physical, chemical or biological change in the air. The air contamination by harmful gases, dust, and smoke drastically affects plants, animals, and humans. There is a certain percentage of gases present in the atmosphere. An increase or decrease in the composition of these gases is harmful to survival. This imbalance in the gaseous composition has increased the earth's temperature, known as global warming. There are two types of air pollutants:

Primary Pollutants

The pollutants that directly cause air pollution are known as primary pollutants. Sulphur- dioxide emitted from factories is a primary pollutant.

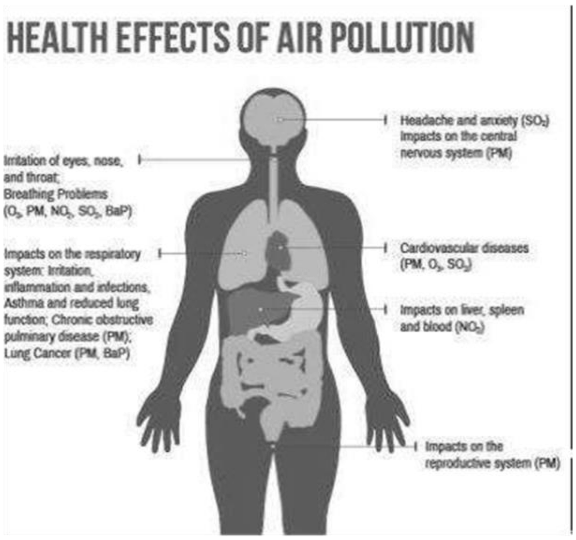
Secondary Pollutants

The pollutants formed by the intermingling and reaction of primary pollutants are known as secondary pollutants. Smog, formed by the intermingling of smoke and fog, is a secondary pollutant.

As by increment in the human population and the advancements in the industry and increased numbers of vehicles, the human race faces a never before threat: air pollution. Worldwide air pollution created havoc in nature and caused global warming, because of which climate change and extreme weather conditions are generated. In addition to this, the air quality in the residential area affects human life widely.

The main pollutants of air can cause a serious health hazards to humans and the environment.

The following diagram shows the effect of air pollution on the human body. The air quality index is used to work out the standard of the air within the particular city. Air quality is a measure of how clean or polluted the air is. Monitoring air quality is vital because polluted air is often insufficient for our health—and the health of the environment.



Air pollution can cause both short-term and long-term effects on health, and many people are concerned about pollution in the air they breathe. In the present scenario, Indian cities are failing to maintain the WHO guidelines for monitoring outdoor pollution for safe levels.

In India, air pollutants like sulfur dioxide have reached the edge point, the amount of PM10 and PM2.5 (Airborne particles smaller than 10 micrometres in diameter and a couple of .5 micrometres in diameter) also are increased in air and reduced the air quality, which leads to an adverse effect on living beings.

The following diagram shows the air quality index and its increment in the few days in the four major cities of India.

Days	Delhi	Kolkata	Mumbai	Hyderabad
28 Oct	368	160	104	106
29 Oct	400	219	112	92
30 Oct	419	205	89	93
31 Oct	410	244	76	57
1 Nov	484	231	65	52

In addition to this, Indian cities are also increasing their ranks in the global most polluted cities ranking. This raised a serious issue of detecting and predicting the air quality and pollutants hotspot area to maintain good air quality.

The following table shows the global ranking of the Indian cities in the global pollution index ranking.

With the advancement of machine learning algorithms and some datelining techniques, now we can predict the level of pollutants in the air on any given day by using historical data.

Date lining and machine learning techniques

The following section will overview the same techniques used in various research papers to predict air pollutants and air quality.

In the review process, we have avoided the IoT devices' techniques devices' techniques, as we are only focused on pure machine learning techniques.

DISCUSSION

In 2019 U. Mahalingam, K.Elangovan, H. Dobhal, C. Valliappa, S. Shrestha and G. Kadam proposed a machine learning method based on ANN to predict the air quality of Delhi[1]. In the said methodology, they have used Particulate Matter PM10, PM2.5, Nitrogen dioxide (NO2), Sulphur dioxide (SO2), Carbon monoxide (CO), Ozone(O3), Ammonia (NH3), Lead (Pb) as the main pollutants, and based on the historical data of the same and predicted the air quality index. The results showed some improvements in the prediction accuracy and suggested that the same methodology can also be used for other cities.

In 2018 similar work but for different cities has been published by Nicolás Mejía Martínez, Laura Melissa Montes, Ivan Mura, Juan Felipe Franco.[2] They used three other machine learning techniques, i.e. logistic regression tree and random forest, to predict the PM10 value in the air of Bogota`. The parameters used in their research are Mean PM10, Mean Wind Speed, Max Wind Speed, Mean Wind Direction, Mean Temperature, Max Temperature, and Max Precipitation.

According to these techniques, the random forest has the most accuracy but lower sensitivity, but logistic regression has the highest sensitivity. The results showed that more twerking in the regression algorithm could increase the accuracy with higher sensitivity.

In 2019 Khushi Maheshwari and Sampada Lamba proposed a similar kind of work [3]. They proposed the use of multilayer perceptrons and supervised regression to predict air quality. They focused their research on the PM2.5 values of Shanghai. Considered parameters were Temp, Pressure, Wind speed, Snow_hours and Rain hours. The study showed that the Multi-layered perceptron gives better accuracy of around 95.4 with a valid positive rate. In addition to this, some more parameters and variants of decision making can be used to improve the accuracy.

In 2019 K. Nandini and Fatima suggested work based on Clustering, Classifiers, Decision Tree, Multinomial Logistic Regression to predict and analyse urban air quality [4]. In their research, they predicted the values of pollutants such as Nitrogen dioxide (NO2), Sulphur dioxide (SO2), Carbon monoxide (CO), and Ozone (O3). In this research, they have used a k-mean clustering algorithm and used Multinomial Logistic.

Regression and Decision Tree algorithms to analyse the results based on available data: They further compared their results based on the error rate and accuracy and determined that the multinomial logistic regression model gives the higher accuracy.

In 2018 Chavi Srivastava, Shyamli Singh and Amit Prakash Singh suggested using AI and regression models to predict the air pollutants in the Delhi city [5]. Their research used different classification and regression techniques like Linear Regression, SDG Regression, Random Forest Regression, Decision Tree Regression, Support Vector Regression, Artificial Neural Networks, Gradient and Adaptive Boosting Regression. They predicted the values of the pollutants like PM2.5, PM10, CO, NO2, SO2 and O3.

They used wind speed (WS), vertical wind speed (VWS), wind direction (WD), temperature (Temp) and relative humidity (RH) as their parameters. They determined that Support vector regression and ANN gives the best result. In further they suggested that the data of longer span can be used to improve the accuracy further and can introduce more parameters like precipitation, minimum and maximum temperature, solar radiation, vapour pressure.

In 2019 Dan Wie suggested Logistic regression, live Bayes classification and support vector machines to predict the air pollution level in any specific city[6]. The research uses the data from China Meteorological Data Sharing Service System, Beijing Transportation Research Center and US Embassy in Beijing. Parameters like temperature, wind speed, relative humidity, traffic index and air quality of the previous day are used to predict the value of PM2.5. The results show that the SVM has the best performance index in all. He further suggested the larger dataset covering long term data can be used to increase the results further.

In 2017 Jan Kleine Deters, 1 Rasa Zalakeviciute, 2Mario Gonzalez, 2 and Yves Rybarczyk used a spatial visualisation of the distribution of fine particulate matter trends according to the wind (speed and direction) and precipitation parameters in two locations in Quito. They include a description of the preparation of the data for classification. Then, various machine learning models are exploited to classify different levels of PM 2.5, namely, Boosted Trees and Linear Support Vector Machines. Finally, a Neural Network regression and a time series analysis are applied to provide insight into the parametric boundaries in which the classification models perform adequately. In 2017 Xi Mao, Taoshen, Xiaofeng pursued a different kind of university study to explore the effectiveness of satellite data in predicting hourly PM 2.5 (Respirable particulate matter with aerodynamic diameter below 2.5 μm) concentrations for a chosen number of forwarding time steps over eastern China.

The forecasting model for hourly PM2.5 concentration three days in advance using satellite-derived AOD is proposed. The model is built by a MLP type of back-propagation neural network. The model has the capability of predicting the space-time distribution of pollutant concentration. The model shows potential to be adapted for other regions. In 2017 the case of Hong Kong prediction of air pollutants concentration based on an extreme learning Machine in Hong Kong. The major sources of air pollutants are mobile, stationary, and transboundary sources. This research proposes predicting the concentration of r pollutants using trained extreme learning machines based on the data obtained from eight air quality parameters in two monitoring stations, including Sham Shui Po and Tap Munin Hong Kong, for six years. The experimental results show that our proposed algorithm performs better on the Hong Kong data both quantitatively and qualitatively. Particularly, our algorithm shows better predictive ability, with R2 increased and root mean square error values decreased, respectively.

In 2017, similar kind of work for PM 2.5 Analysis and prediction by Ping Jiang Qingli Dong Peizhi Li, the analysis and prediction of air pollutants are of great significance in environmental research today since airborne pollution is a substantial threat, especially in urban agglomerations of China. To develop more effective warning systems and management advice, the authorities and city dwellers need more accurate air pollution forecasts. They work on a novel and practical framework, termed HML- AFNN was successfully developed to analyse and forecast the concentration of particular matter (PM 2.5) for a selected number of forward time steps. In a simulation of the trajectory of air pollutants, the high-dimension association rules (HDAR) approach considered the tempo-spatial relations and the meteorological and geographical factors of the ambient regions as parameters.

In addition, the learning vector quantisation (LVQ) network was adopted to select the appropriate inputs to improve the efficiency of the training process. Moreover, an adaptive fuzzy neural network (AFNN), a combination of neural and fuzzy logic, was utilised to analyse and predict the PM2.5concentration.

CONCLUSION

As observed in all the above papers in recent years, we can see that various datelining and machine learning methods can be used to predict the pollution level in the air. The accuracy of the prediction model can vary depending on the size and span of the dataset, and the parameters are chosen. This paper only focused on machine learning techniques. Still, we can further use the IoT devices such as pollutant sensors to get accurate data on the air quality of the current location, and we can use that data to predict the air quality. Furthermore, we can also use satellite Images and image processing and computer vision techniques to identify the pollution Hotspots and determine the future hotspots' future values by using the same techniques. In addition to these techniques, more variants of the said algorithms, different approaches and a more detailed dataset with more parameters can also help to increase the accuracy.

REFERENCES

1. U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha and G. Kadam, "A Machine Learning Model for Air Quality Prediction for Smart Cities," 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), Chennai, India, 2019, pp. 452-457,DOI: 10.1109/WiSPNET45539.2019.9032 734.
2. N.M.Martínez, L.M.Montes, I.Mura and J. F. Franco, "Machine Learning Techniques for PM10 Levels Forecast in Bogotá," 2018 ICAI Workshops (ICAIW), Bogota, 2018, pp. 1-7, doi: 10.1109/ICAIW.2018.8554995.
3. K. Nandini and G. Fathima, "Urban Air Quality Analysis and Prediction Using Machine Learning," 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India,

2019, pp. 98-102, DOI: 109/ICATIECE45860.2019.9063845

4. M. Shahbaz, S. A. Taqvi, A. C. Loy, A. Inayat, F. Uddin, A. Bokhari, S. V. R. Naqvi, Artificial neural network approach for the steam gasification of palm oil waste using bottom ash and CaO. Renewable Energy vol. 132, 243– 254, 201 Mar1
5. B. C. Liu, et al., "Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing- Tianjin- Shijiazhuang", PLOS, 2017.
6. C. Srivastava, S. Singh and A. P. Singh, "Estimation of Air Pollution in Delhi Using Machine Learning Techniques, "2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, Uttar Pradesh, India, 2018, pp.304-309, DOI: 10.1109/GUCON.2018.8675022.
7. Wei, D.: Predicting Air Pollution Level in a Specific City. Stanford Publication(2019)
8. R. Pasupuleti, Uhasri, P. Kalyan, Srikanth and H. K. Reddy, "Air Quality Prediction Of Data Log By Machine Learning," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 1395-1399, DOI: 10.1109/ICACCS48705.2020.9074431
9. Aditya C R#1 , Chandana R Deshmukh*2 , Nayana D K*3, Praveen Gandhi Vidyavastu *4 #Associate Professor, Department of Computer Science and Engineering,
   *B.E. Student, Department of Computer Science and Engineering, Vidya Vikas Institute of Engineering and Technology, Mysuru, Karnataka, India570028.