2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) | 978-1-7281-8337-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICACCCN51052.2020.9362912

Air Quality Prediction using Machine Learning Algorithms –A Review

Tanisha Madan
PhD Scholar(CSE), Galgotias University &
Assistant Professor,
BPIT(GGSIPU)
tanishamadan@gmail.com

Shrddha Sagar
CSE
Galgotias University
Greater Noida, India
shrddha.sagar@galgotiasuniversity.edu.in

Deepali Virmani
CSE
BPIT(GGSIPU)
New Delhi, India
deepalivirmani@gmail.com

Abstract-Predicting air quality is necessary step to be taken by government as it is becoming the major concern among the health of human beings. Air quality Index measure the quality of air. Various air pollutants causing air pollution are Carbon dioxide, Nitrogen dioxide, carbon monoxide etc that are released from burning of natural gas, coal and wood, industries, vehicles etc. Air Pollution can cause severe disease like lungs cancer, brain disease and even lead to death. Machine learning algorithms helps in determining the air quality index. Various research is being done in this field but still results are still not accurate. Dataset are available from Kaggle, air quality monitoring sites and divided into two Training and Testing. Machine Learning algorithms employed for this are Linear Regression, Decision Tree, Random Forest, Artificial Neural Network, Support Vector Machine.

Keywords: Air quality Index, Decision tree, Support Vector Machine, Random Forest, Linear Regression

1. INTRODUCTION

Air pollution is dangerous for human health and should be decrease fast in urban and rural areas so it is necessary to predict the quality of air accurately. There are many types of pollution like water pollution, air pollution, soil pollution etc but most important among these is air pollution which should be controlled immediately as humans inhale oxygen through air.

There are various causes of air pollution. Outdoor air pollution caused by industries, factories, vehicles and Indoor air pollution is caused if air inside the house is contaminated by smokes, chemicals, smell. Two types of Pollutants that is causing air pollution are Primary Pollutants and Secondary Pollutants.

Primary Pollutants include: -

Carbon dioxide (CO_2): Carbon dioxide is playing an important role in causing air pollution. It is also named as Greenhouse gas. Global warming a major concern caused by increase in carbon dioxide in air.CO_2 is exhale by Human.CO_2 is also released by burning of fossil fuels.

Sulphur oxide (SO_X): Sulphur dioxide (SO_2) released by burning coal and petroleum. It is released by various industries. When react with Catalyst (NO_2), results in H_2SO_4 causing acid rains that forms the major cause of Air Pollution.

Nitrogen oxide (NO_X): Most commonly Nitrogen dioxide (NO_2) that is caused by thunderstorm, rise in temperature.

Carbon monoxide (CO): -Carbon monoxide is caused by burning of coal and wood. It is released by Vehicles.It is odorless, colorless, toxic gas. It forms a smog in air and thus a primary pollutant in air pollution.

Toxic metals –Example are Lead and Mercury

Chlorofluorocarbons (CFC): -Chlorofluorocarbons released by air conditioners, refrigerators which react with other gases and damage the Ozone Layer. Therefore, Ultraviolet Rays reach the earth surface and thus cause harms to human beings.

Garbage, Sewage and industrial Process also causes Air Pollution.

Particles originating from dust storms, forest, volcanoes in the form of solid or liquid causing air pollution.

Secondary Pollutants include: -

Ground Level Ozone: It is just above the earth surface and forms when Hydrocarbon react with Nitrogen Oxide in the sunlight presence.

Acid Rain: When Sulphur dioxide react with nitrgendioxide, oxygen and water in air thus

ISBN: 978-1-7281-8337-4/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:10:26 UTC from IEEE Xplore. Restrictions apply.

causing acid rain and fall on ground in dry or wet form.

The difference between Primary Pollutants and Secondary Pollutants is Primary Pollutants are those which are released into air directly from Source whereas Secondary Pollutants are those which are formed by reacting with either primary pollutants or with other atmospheric component. There are various pollutants causing air pollution but PM 2.5 being the major air pollutant as proposed by the author (J. Angelin Jebamalar & A. Sasi Kumar,2019) and comes out with the best results in predicting level of PM 2.5 in their research [13]. Logistic regression and autoregression help in determining the level of PM2.5 [4]. The day wise prediction of pollutant level [1] was removed by various authors further by predicting hourly wise data using different algorithm.

Benzene concentration can also account into air pollution and its concentration can be determined with CO [7].

These are the causes of air pollution. Air pollution is causing harmful effects on human beings and plants. It causes the less threatening diseases like irritation in throat, nose. Headache to most severe disease like Respiratory Problems, shortness of breath, Lungs Cancer, brain disease, kidney disease and even leads to death. There are masks which protect us from increasing air pollution and various acts are there to control air pollution. It is also necessary to create awareness among human being about air pollution.

It is necessary to predict the air quality accurately. Various traditional methods are there to measure it but results are not accurate and it involves a lot of mathematical calculations. Machine Learning a subset of Artificial Intelligence has an important role in predicting air quality. Various researches are being done on measuring Air quality Index by using Machine Learning algorithms. So, to control Air Pollution first necessary step is to measure accurately the Air Index Quality. Machine Learning algorithms plays an important role in measuring air quality index accurately. In this paper Various algorithm are compared on the basis of different condition in different areas and Neural Network comes out with best results [1].

In section 2 Literature survey is discussed and in section 3 result obtained by various researchers is discussed and in Section 4 conclusion is discussed

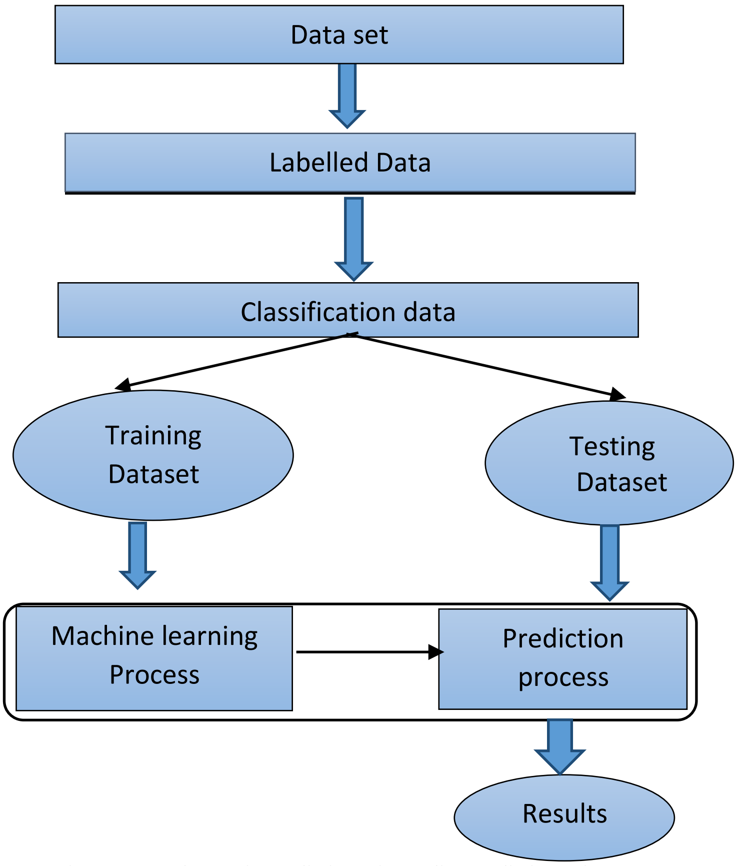

Fig. 1. General steps in predicting Air Quality Prediction using Machine Learning Algorithm

II. LITERATURE SURVEY

Lot of research had been done in this field. The research taken by various authors as follows: -

[1] Kostandina Veljanovska1 & Angel Dimoski (2018) in their research had compared unsupervised neural network algorithm in which output is not known with supervised algorithm K-Nearest Neighbour, Support vector machine, Decision Trees. Neural network performs better as compared to these algorithms but lacks in determining the hourly prediction of air pollutants level.

[2] Zhao et al. ,(2018) in their research had predicted the level of air pollutant with Recurrent neural network at any given time and remove the drawback of hourly prediction due to memorization power of algorithm but lacks in working without memory operation.

[3] Savita Vivek Mohurle, Dr. Richa Purohit and Manisha Patil(2018) had predicted Pm2 and Pm10 level using Fuzzy logic. Fuzzy logic helps removing the outliers that is unwanted gas present in atmosphere but in fuzzy logic clusters are created which can contain repeated data and can lead to inaccurate prediction.

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:10:26 UTC from IEEE Xplore. Restrictions apply.

[4] CR et al.(2018) in their research had used Autoregression to detect whether the air is polluted or not and linear regression can be used to determine the level of PM2.5 but the drawback is it does not able to determine the level of PM2.5 after when there is some change in atmospheric condition and it takes into account meteorological condition such as wind speed, temperature.

[5] Zhang et al. (2018) proposed Wavelet neural network considered as robust method to determine the level of Air pollutants but lack in determining the appropriate wavelet function and no of proper hidden layers required in their research which leads to inaccurate prediction of air pollutants.

[6] Timothy M. Amado and Jennifer C. Dela Cruz(2018) in their study considered neural network with integrated sensor giving accurate level of air pollutants but slow because it take all training set and cannot work with incomplete data set.

[7] Arwa et al. (2018) in their research had determined Benzene concentration by using Artificial Neural Network(ANN) and Support Vector Machine but does not predict the error accurately between actual and predicted value and its association with CO can be determined using method of Correlation.

[8]In this paper the author Kang et al.(2018), had compared the various models like ANN, Random Forest, Decision Trees, Least squares support vector machine model, Deep belief network and Deep belief network comes out to be superior as it takes into account hourly prediction of data but drawbacks are there are lots of issues in sensor quality of data due to faults in device .

[9] Mejía et al. (2018),had determined PM10 level best with Random Forest but does not accurately predict the level of dangerous pollutant but can work with incomplete data set.·

[10] Lidia et al.(2015) had determined Airvlc Model not only to predict the air quality level CO,NO,PM2.5 but also warns the public about dangerous air pollutant outside using his model through sensors

[11] Nandini K & G Fathima (2019) had used used Multinominal Logistic regression and decision trees to predict air quality pollutant level. In this Interpolation, prediction and feature analysis is done. The author in this uses K-means clustering to form the three clusters and classified as them as high , Medium, low level on air pollution level and dataset is splitted into training and testing. After k means clustering,MultinominalLogistic regression and decision trees are used to predict the upcoming values and in multinominal logistic regression the actual values are much closer to predicted values and considering multinominal logic regression giving better accurate result·

[12] Desislava Ivanova and Angel Elenkov (2019) had used Rasberry Pi platform with Multilayer perceptron algorithm of machine learning to predict the air pollutant accurately.Multilayer perceptron overcome problem of classification which is used for discrete values and regression which is used for continuous value. In this author uses discrete values and had used multilayer perceptron using backpropagation and therefore input did not pass to the activation function resulting in 0 or 1 indicating how big the difference between the predicted and actual value. The coefficient of determination(R^2) obtained is better but more can be improved by increment feeding.

[13] J. Angelin Jebamalar & A. Sasi Kumar(2019) in this paper uses hybrid light tree and light gradient boosting model .The proposed method PM2.5 level is captured using the sensor with raspberry PI and the cloud is used to store it and then hybrid model is used for predicting. Hybrid model is compared with Linear Regression ,Lasso Regression ,Support Vector Regression , Neural Network , Random Forest , Decision Tree , XGBoost and Hybid model comes out to be best to detect PM2.5 as Boosting is sequential ensemble as it correct the errors from the previous model. It takes less space and can handle large amount of data but the limitation is it takes more time and image can be used further to predict PM2.5 Level more accurately.

[14] Soubhik et al. (2018) had compared various algorithms like Linear regression, Neural network regression, Lasso regression, ElasticNet regression, Decision Forest, Extra trees, Boosted decision tree, XGBoost, KNN, and Ridge regression to predict air pollutant level.Better accuracy is obtained by extra trees because features are arranged in decreasing order of importance to predict the upcoming value·.

[15] Burhan BARAN (2019) proposed Air quality prediction using Extreme learning machine (ELM). ELM is single layer and feed forward artificial neural network. ELM uses faster learning speed .Activation used are sigmoidal, sine and hard-limit and the accuracy obtained by sigmodial function is better. More number of hidden neurons are used

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:10:26 UTC from IEEE Xplore. Restrictions apply.

and 10 cross validation is used to calulate the air pollutant level using ELM.

[16] In this the author Pasupuleti et al. (2020) compares the decision tree, linear regression, random forest. Major air pollutants are taken and meterological condition are taken using Arduino Platform. Random forest gives better accurate result due to overfitting that reduces errors But drawback is Random forest uses more memory and high cost.

[17] Haotian Jing & Yingchun Wang(2020) had predicted the air quality index using XG Boost.It uses the weak classifier and shortcoming of the previous weak classifier to form a strong classifier thus reducing the error between predicted and actual values .It uses the K- cross validation .The mean absolute error and coefficient of determination is determined to predict the difference between actual and predicted value.The drawback faced is that it takes the previous value and is affected by outlier unwanted pollutant in the air.

[18]MaryamAljanabi, MohammadShkoukani and MohammadHijjawi (2020) depict the ozone layer depending upon the Temperature, humidity, windspeed,wind direction. Various machine learning algorithm used are MLP, XG Boost, SVR, DTR. In this data set is taken according to area selection then the preprocessing is done the fluctuation if any is removed in the dataset using holt winter, moving average, savitsky, Golay and Savitsky and it was observed that Golay gives better result in preprocessing. Then Feature selection is performed using forward feature wrapper selection as there are some unwanted features which is not to be taken into account to predict accurately the level of pollutant in air .Then the machine learning algorithm described above is performed and the coefficient of determination, root means square error, Mean absolute error are compared and MLP Comes out to be superior model. It predicts the Ozone layer using MLP on day to day basis.

[19] ShuWang et al. (2020) uses gas recurrent Neural Network to predict the air quality level and compared with MLP and SVR. Gas recurrent neural network performs better due to less invariant in sensor drift but Gas Recurrent Neural Network is more prone to variablity and humidity in atmosphere.

[20] Zhao et al.(2020) uses CERT Model which is formed by combining forward and recurrent neural but predicts the pollutant level day wise not hourly wise and cover north east china in its area and does not account into meterological data .

III. COMPARATIVE ANALYSIS OF RESULT OBTAINED BY VARIOUS RESEARCHERS WITH DIFFERENT ALGORITHMS

Table 1. Result obtained by various researchers with different algorithms

	Technique	**Prediction Performance**	**Pollutants**	**Areas**
[1]	Neural Network	Accuracy: 92.3%	SO_2, NO_2, O_3,CO, PM2.5, PM10	Republic of Macedonia
[2]	RNN	Accuracy: 80.27 %	CO,NO_2,O_3 ,SO_2, $PM_{2.5}$, PM10	Atlanta-Sandy Springs-Roswell,
[4]	Logistic regression, Autoregression	Mean Accuracy:- 0.99859 StandardDeviation Accuracy 0.000612	Temperature, Wind speed, Dewpoint, Pressure,PM2.5	Populating and Developing countries
[6]	Neural Network	Accuracy:99.56	Air temperature , Relative humidity, MQ2, MQ135,MQ5	General
[7]	Artificial Neural Network	MRE: -0.16	Benzene Concentration	General
[8]	Deep Belief Network	Error :1.6%	Pm 2.5	China

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:10:26 UTC from IEEE Xplore. Restrictions apply.

[9]	Random Forest	Accuracy: Between 70% and 90%	PM10, Wind speed, Wind direction, Temperature	Bogota
[10]	Random Forest	MSE:- 0.53(CO concentration) 29.517(NO concentration) 14.851(NO2 Concentration)	CO, NO,NO_2,PM2.5	Avd. Francia Station
[11]	Regression Model	Error rate:0.428	SO ,NOX	General
[12]	Multilayer Perceptron Regressor	Coefficiento f determination:0.65	NO,NO_2,O_3 and PM10	General
[13]	Hybridtree andlight gradient boosting model	Accuracy :Greater than 99%	PM 2.5	General
[14]	Extra trees	Accuracy:85.3%	Pressure, wind direction, wind temperature, humidity	Polluted cities
[15]	Extreme Learning Machine	Accuracy:75.5%	Temperature, humidity, pressure, wind speed, PM 10, SO2	General
[16]	Random Forest	Accuracy:79%	CO, SO_2, and O_3. Temperature, Wind Speed, Humidity and Wind Direction	General
[17]	XG Boost	R Squared:0.9951	So_2,NO_2,O_3,CO	General
[18]	MLP	R squared:98.653%	O_3	Jordanian Ministry of Environment
[19]	Gated Recurrent Neural Network (GRU)	25,17: RMSE	CO	General
[20]	CERT	AQI:0.9792	PM2.5, PM10, CO, SO2, NO2, and O3	China

Different results are obtained on the basis of different pollutants. Methods are chosen on the basis of different pollutants and different areas whether urban or rural and accuracy and error is predicted and seen how much predicted value is closer to exact value

R Squared can be calculated as: -

$$R^2 = \left[\frac{1}{N}\frac{\sum_{i=1}^{N}\left[(P_i-\bar{P})(A_i-\bar{A})\right]}{\sigma_P\sigma_A}\right]^2$$

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:10:26 UTC from IEEE Xplore. Restrictions apply.

RMSE can be calculated as: -

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - A_i)^2}$$

MAE can be calculated as: -

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - A_i|\,.$$

IV. CONCLUSION

Lots of research had been done and different algorithms comes out to be superior with different conditions like pollutant chosen and area selection. Various algorithms had taken meteorological data like temperature, wind speed, humidity in predicting accurately the upcoming pollutant level. Neural Network and boosting model comes out to be superior than other algorithms.

REFERENCES

[1]. Kostandina Veljanovska1 & Angel Dimoski2, Air Quality Index Prediction Using Simple Machine Learning Algorithms,2018, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS).

[2]. Xiaosong Zhao , Rui Zhang, Jheng-Long Wu, Pei-Chann Chang and Yuan Ze University, A Deep Recurrent Neural Network for Air Quality Classification, 2018, Journal of Information Hiding and Multimedia Signal Processing

[3]. Savita Vivek Mohurle, Dr. Richa Purohit and Manisha Patil, A study of fuzzy clustering concept for measuring air pollution index,2018, International Journal of Advanced Science and Research

[4]. Aditya C R , Chandana R Deshmukh , Nayana D K and Praveen Gandhi Vidyavastu , Detection and Prediction of Air Pollution using Machine Learning Models,2018, International Journal of Engineering Trends and Technology (IJETT)

[5]. Shan Zhang , Xiaoli Li & Yang Li , Jianxiang Mei, Prediction of Urban PM2.5 Concentration Based on Wavelet Neural Network,2018,IEEE.

[6]. Timothy M. Amado & Jennifer C. Dela Cruz, Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization, 2018, IEEE

[7]. Arwa Shawabkeh , Feda Al-Beqain , Ali Rodan, Maher Salem, Benzene Air Pollution Monitoring Model using ANN and SVM,2018,IEEE

[8]. Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie, Air Quality Prediction: Big Data and Machine Learning Approaches,2018, International Journal of Environmental Science and Development

[9]. Nicolás Mejía Martínez, Laura Melissa Montes, Ivan Mura and Juan Felipe Franco, Machine Learning Techniques for PM10 Levels Forecast in Bogotá,2018,IEEE

[10]. Lidia Contreras Ochando, Cristina I. Font Julian, Francisco Contreras Ochando, Cesar Ferri,Airvlc: An application for real-time forecasting urban air pollution,2015, Proceedings of the 2 nd International Workshop on Mining Urban Data, Lille, France.

[11]. Nandini K & G Fathima, Urban Air Quality Analysis and Prediction Using Machine Learning,2019,IEEE

[12]. Desislava Ivanova and Angel Elenkov, Intelligent System for Air Quality Monitoring Assessment using the Raspberry Pi Platform,2019, IEEE

[13]. J. Angelin Jebamalar & A. Sasi Kumar, PM2.5 Prediction using Machine Learning Hybrid Model for Smart Health,2019, International Journal of Engineering and Advanced Technology (IJEAT)

[14]. Soubhik Mahanta, T. Ramakrishnudu, Rajat Raj Jha and Niraj Tailor, Urban Air Quality Prediction Using Regression Analysis,2019, IEEE

[15]. Burhan BARAN, Prediction of Air Quality Index by Extreme Learning Machines,2019, IEEE

[16]. Venkat Rao Pasupuleti , Uhasri , Pavan Kalyan , Srikanth and Hari Kiran Reddy, Air Quality Prediction Of Data Log By Machine Learning,2020,IEEE

[17]. Haotian Jing & Yingchun Wang, Research on Urban Air Quality Prediction Based on Ensemble Learning of XGBoost, 2020, E3S Web of Conferences

[18]. Maryam Aljanabi, Mohammad Shkoukani and Mohammad Hijjawi, Ground-level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman, Jordan,2020, International Journal of Automation and Computing

[19]. Shu Wang, Yuhuang Hu, Javier Burgues´ Santiago Marco and Shih-Chii Liu, Prediction of Gas Concentration Using Gated Recurrent Neural Networks, 2020, IEEE

[20]. Zhili Zhao , Jian Qin, Zhaoshuang He, Huan Li, Yi Yang and Ruisheng Zhang, Combining forward with recurrent neural networks for hourly air quality prediction in Northwest of China, Environmental Science and Pollution Research,2020

Authorized licensed use limited to: Central Michigan University. Downloaded on May 14,2021 at 04:10:26 UTC from IEEE Xplore. Restrictions apply.