Machine Learning Approach to Predict Air Quality– A Survey

Nimesh Mohanakrishnan, Tejaswini P R, Nashra Tanseer, Mohammed Saqlain, Mohammed Hussam Khatib

Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India

Email: nimeshmysore05@gmail.com, tejaswini.r.pattange@vvce.ac.in, nashdrell05@gmail.com, saqlain14.ms@gmail.com, hussamkhatib20@gmail.com

Abstract. Air pollution has become a threat to the lives of human, plant, and animal species. It's measured using the Air Quality Index. Various air pollutants such as Carbon Dioxide, Carbon Monoxide, Particulate Matter 2.5, and much more cause contamination of air. Therefore, it is beneficial to predict the air quality and make well-informed decisions to prevent any harm or effects. Multiple algorithms for machine learning have been devised and implemented to determine air quality index. This survey paper discusses several algorithms for machine learning that are used for predicting air quality. Machine learning algorithms used are Support Vector Machine, Linear Regression, Artificial Neural Networks, decision tree, and Random Forest.

1. Introduction

Air pollution is progressively increasing with the rapid growth of industries and human civilization. It affects the lives of living organisms on the planet. An estimation made by World Health Organization ascertains that air pollution kills almost seven million people every year. Polluted air is a primary concern for human beings because it elicits health factors such as lung cancer, stroke, heart disease, and many more. Moreover, it affects the environment on a large scale causing global warming, ozone layer depletion, and contaminating water and soil. Therefore, it is imperative to study and observe the air quality patterns caused by various pollutants.

Air pollutants in the earth's atmosphere result from anthropogenic processes caused by industries, factories, vehicles that run on fuel, and many more. They are classified based on origin, state of matter, and sources.

Depending on the Origin:

Primary pollutants: These pollutants are a result of natural disasters and human activities. The primary pollutants include:

Carbon Monoxide (CO): This pollutant is highly toxic and results from internal combustion engines, volcanoes, forest fires, and industries. It is also named a greenhouse gas. It produces carboxyhaemoglobin which reduces oxygen capacity in the blood.

Carbon Dioxide (CO2): Carbon Dioxide is heavier than air and results from volcanoes, fire, and many more. Humans too exhale CO2. It is also a greenhouse gas. Inhaling high concentrations of CO2 may cause dizziness and headache.

Chlorofluorocarbons (CFCs): Refrigerators, air conditioners, and aerosols use CFCs because of their physical structure and chemical nature. These are highly destructive to the ozone layer.

Nitrogen Oxide (NOx): There are present in various forms of their oxides such as NO2, NO3, etc. They are responsible for smog, acid rain, and the greenhouse effect.

Sulphur Dioxide (SO2): They are pungent-smelling colourless gas mainly produced from industrial processes and volcanic activities. They affect humans by causing respiratory issues and premature deaths.

Benzene: They are found in petrochemicals and used as additive fuel.

Asbestos: They occur naturally as a fibrous silicate mineral. Prolonged exposure to this pollutant can cause fatal illnesses.

Secondary Pollutants: The formation of these pollutants is because of chemical reactions between atmospheric elements and primary pollutants. Examples of secondary pollutants are sulphuric acid and carbonic acid.

Depending on the State of Matter:

Gaseous Pollutants: These pollutants exist in gaseous forms. Examples include Nitrogen Oxide, Sulphur Dioxide, Carbon Dioxide.

Particulate Air Pollutants: These are the suspended droplets or mixture of a few particles in the atmosphere.

Depending on the Sources

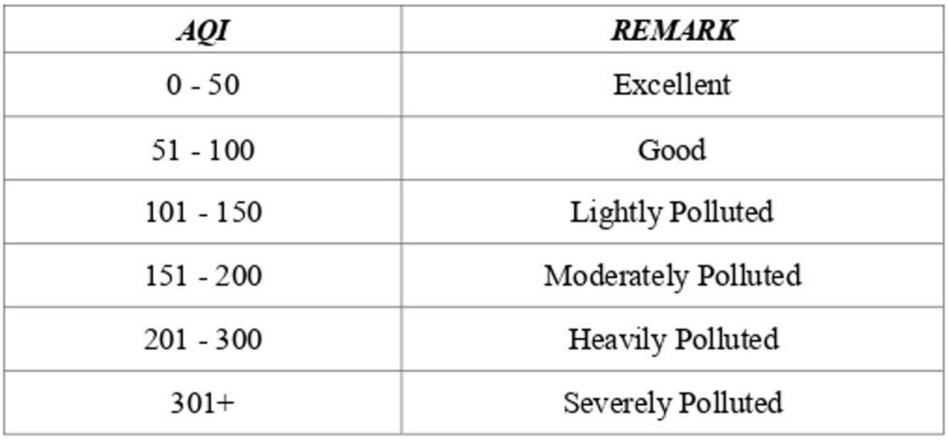
Natural Sources: Although there are multiple sources the major contributors include volcanic eruptions. These eruptions release sulphur gases which combine with water vapour to form sulfuric acid. Natural vegetation, under high heat can emit volatile organic compounds such as terpenes, which is a one of the precursor gases of ozone. Dust storms pick up fine grain particles that stay suspended in the subsurface airflow for a long time adding to the air pollution, forest or wildfires add smoke and ashes to the air, sulfur springs, organic and inorganic decays, natural geysers, vegetative decays, cosmic dust, marsh gases, pollen grains of flowers, photochemical reactions, soil debris, and so on are some more examples.

Man-made Sources: Power plants, factories, vehicles emit carbon dioxide, carbon monoxide, hydrocarbons, sulfur dioxide, nitrogen dioxides and particulate matter that consists of fine particles suspended in the air. Burning oil, coal, gasoline, and other fossil fuels is a major cause of man-made air pollution. A hefty number of the early nuclear tests were detonated in the atmosphere, which spread radioactive materials through the atmosphere

Air quality index (AQI) aids as a quotidian reporting measure of air quality. It works as a parameter to indicate how air pollution affects one's health over a defined period. The objective of AQI is to keep the individuals informed and aware about how local air quality can negative impact on their health. The Environmental Protection Agency (EPA) calculates AQI for five major air pollutants, this is calculated based on predefined national air quality guidelines aimed at protecting public health. The AQI number is directly proportional to contamination of air, thereby indicating an increased risk of human health. Many industrialised countries have adopted the application of AQI to make informed and responsible decisions over the past three decades. Air quality related information is easily acquired in real-time by the means of AQI.

To report air quality, various countries use different point systems. The United States, for example, employs a 500-point scale, with a score of 0 to 50 deemed satisfactory. A rating with values between 301 to 500 is considered dangerous. India uses this 500-point scale as well. Every day, sensors record the biggest contaminants' concentrations. EPA-developed standard equations are used to translate these raw combined values into a separate AQI value for individual pollutant (ground-level ozone, carbon monoxide, particle pollution and sulphur dioxide). AQI value for that day is determined by the highest of these AQI readings.

The figure below shows the remarks corresponding to the air quality index values for a 500-point scale measurement.



AQI	*REMARK*
0 - 50	Excellent
51 - 100	Good
101 - 150	Lightly Polluted
151 - 200	Moderately Polluted
201 - 300	Heavily Polluted
301+	Severely Polluted

Figure 1. AQI and Corresponding Remarks

Various air pollutants contribute to air pollution. However, numerous research and studies show that particulate matter (PM2.5) impels majorly and severely for the poor quality of air. Therefore, it is imperative to study and forecast air quality more accurately to safeguard from numerous harmful effects. The conventional and orthodox methods include massive statistical and mathematical calculations to measure air quality. Nonetheless, machine learning which is a segment of artificial intelligence (AI) proves to be better for predicting air quality. Since air quality prediction corresponds to time series prediction, choosing various machine learning algorithms is optimal.

Various research focuses on the estimation of the air quality index using numerous machine learning models. Many researchers have scrutinized machine learning algorithms such as decision trees, linear regression, random forest, support vector machines, and artificial neural networks. In section 2, we present a literature survey obtained by various researchers. In section 3, we examine the results of predicting air quality using different algorithms found by researchers. Lastly, in section 4, we conclude the survey paper.

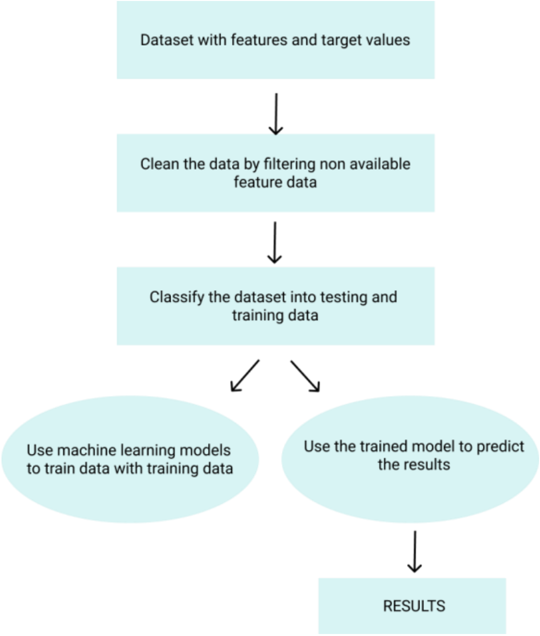

Dataset with features and target values
Clean the data by filtering non available feature data
Classify the dataset into testing and training data
Use machine learning models to train data with training data
Use the trained model to predict the results
RESULTS

Figure 2. General Steps Involved in AQI Prediction

2. Literature Survey

A lot of research is conducted to predict the air quality. Numerous authors have researched the prediction of air quality with differing pollutants and dataset as follows:

[1] Authors Wang Zhengua and Tian Zhihui have used an improved BP neural network by integrating the genetic algorithm to predict the AQI value. Their model has three layers. The input layer is 6 dimensions, and the output layer is 1 dimension. The hidden layer dimension is calculated by an empirical formula: ($\sqrt{n+m}$ + a) where n corresponds to input node number, and m is the output node number. a is a constant number between 1 to 10. However, the accuracy rate of prediction using the improved model was 80.44% only. The accuracy levels will have to be improved further.

[2] In this paper, the authors examine and evaluate various air contaminants such as CO, SO2, PM2.5, PM10, NO2, and O3. They use three machine learning algorithms: linear regression, decision tree, and random forest regression, and conclude by stating that the Random Forest algorithm gives a better prediction of air quality. Multitude of decision trees are constructed during the training time. It works as an meta estimator wherein it can either classify or predict results. The total number of functions at each node is split depending on the hyper parameter percentage. The drawback of this paper is that the prediction accuracy of various pollutants using random forest lies between 70%-86% only. And other major pollutants such as benzene, toluene, xylene are not considered.

[3] Kostandina and Angel have examined the accuracy differences between Neural Network, K-Nearest Neighbors, Decision Trees, and Support Vector Machine. They have considered the unsupervised neural network where the output value is unknown. The neural network constructed by them contains 6 input attributes, 1 hidden layer with 10 neurons, and the output layer outputs 3 classes - high, medium, and low. When the total dataset is partitioned into 70% training, 10% validation, and 20% testing, they attain the highest accuracy of 92.3%. The neural network performs better than other algorithms for daily predictions and not hourly basis.

[4] Arwa Shawabkeh et al. (2018), in their study, estimated the concentration levels of benzene in correlation with CO using Support Vector Machine (SVM) and Artificial Neural Networks (ANN). They found ANN to result in fewer errors. They used 5 hidden layers in their proposed methodology and Levenberg-Marquardt algorithm - an algorithm designed to work with loss functions specifically which take the form of a sum of squared errors. However, with increase in data samples the mean square value and mean absolute value of errors in SVM decreased.

[5] Yuelai Su has utilized the Light Gradient Boosting Machine and eXtreme Gradient Boosting Machine to predict the air quality (through PM2.5 measurements) and used 50,000 data samples. They concluded by stating that Light GBM works more optimally compared to eXtreme. While segmenting data points, the Light GBM does not use pre-sorting algorithm - an algorithm that pre-sorts all features by values and uses cost to find optimal segmentation points of each feature. Instead, they use method of sorting buckets such as histogram algorithm that splits eigenvalues according to the intervals. This mechanism ensures that only small amount of precision is lost, and massive computing memory is saved. However, with increasing data samples Light GBM would reduce the running time of machine learning. And hence, short-term prediction becomes almost impossible.

[6] Lidia Contreras Ochando et al. (2015) developed an application - Airvlc, that employs a regression model to predict real-time levels of CO, NO, PM2.5. They use mean squared error as performance measure. The application also provides information to people about air pollutant concentrations through sensors.

[7] Soubhik Mahanta et al. (2019) predicted air quality compared the efficiency using Linear Regression, Lasso regression, Neural Network Regression, Decision Forest, ElasticNet Regression,

Extra Trees, XGBoost, Boosted decision tree, KNN, and Ridge regression. In their research, they found that the performance of the Extra Tree Regression model was better and resulted in an accuracy of 85%. The reason is that the arrangement of features was in decreasing order.

[8] The authors, in the quest to predict air quality have used the CERL hybrid ensemble model in-order to exploit the working of recurrent neural networks as well as the working of forwarding neural networks. The pollutants used for measuring and predicting air quality are CO, PM2.5, NO2, SO2 and O3. They have also used AQI values while predicting the air quality. CERL improves prediction performance using recurrent neural networks. Furthermore, the predictions work better for the hour level. Hybrid CERL model is formed by combining forward neural networks with recurrent neural network. Here, the prediction results are grouped together. These groupings occur as the features of training and test sets to build the hybrid model. They evaluation methodology metrics used are mean absolute deviation (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE), and correlation coefficients(R). However, the accuracy for long-term prediction is low. The authors have also suggested future work to be explored using a convolutional neural network for air quality prediction.

$$MAE = \left(\frac{1}{n}\right)\sum_{i=0}^{n} |xi - x'|$$

$$RMSE = \sqrt{\left(\frac{1}{n}\right)\sum_{i=0}^{n} x^2 - x'^2 - 2x^2x'^2}$$

Where xi and x' represent the actual and predicted values and n represents the number of test samples.

[9] The authors considered five different machine learning algorithms: Random Forest, k-nearest neighbors (KNN), Support Vector Machine (SVM), Naive Bayesian, and Neural Network. Their results show that neural networks with integrated sensors give the highest accuracy of detecting air pollutants. They have compared the working of the neural network by varying the number of hidden layers weight decays. When the number of hidden layers is 5 and weight decay is 1 x 10^{-4}, the neural network yields higher accuracy. For prototyping purposes, they have used DTH 11 Arduino sensors. However, the response time increases with increase of data the dataset, and it is incapable to work with deficient and partial data set.

[10] Author Burhan Baran in his research, used Extreme Learning Machine (ELM) to predict the air quality. They also made use of three different activation functions for estimation: sine, sigmoid, and hard limit. The dataset included temperature, wind speed, humidity, pressure, PM10, and SO2. The hard-limit function had the highest test accuracy of 74.17 and a test period of 0,0004156 seconds for 50 neuron counts in the hidden layer.

[11] Limei Ma et al. (2020) state that it is very effective to predict the air quality using dependent variables than independent variables. They also state that multivariate regression is significantly more practical than univariate regression. They have used PM2.5, PM10, SO2, NO2, CO, and 03 values to predict the values. Multivariate regression mainly considers correlation between a dependent and independent variable. The equation consists of x values that correspond to the independent variables, and a random error. The only drawback is that they have used one year's data. Insufficient data may lead to incorrect prediction results.

$$\gamma = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta pxp + \in$$

Where x represents independent variables, xp represents cut off, and ∈ representers random error.

[12] Chuanting Zhang & Dongfeng Yuan in their research have used spark technology to parallelly distribute real-time meteorological data values for prediction of the air quality. They have used the random forest algorithm for predicting the air quality. In their methodology, they have used spark's in-memory computation model to overcome massive computation problems caused by trees. Spark has master node, worker node, and cluster manager. The cluster manager allocates resources and communicates between masters and worker nodes. The master node performs data partition and worker nodes are used for execution purposes. While implementing random forest algorithm, they eliminate features that do not have meteorological information and the missing values in the data set are filled with attribute's average value. Their methods allow faster prediction of the results. However, the accuracy rate is only 79.25%.

[13] Liying et al. (2019) have utilized Amazon S3, MongoDB, and Apache Spark as means for distributed computing model. They have used random forest and reported an accuracy of 81%. They have also showed that a standalone system is not very sufficient in processing real-time air quality data for better prediction results.

[14] Krittakom Srijiranon & Narissara Eiamkanitchat proposed a neuro-fuzzy model that includes 14 input features. These features are further divided into meteorological and air pollution data. They have used ensemble neural network with neighborhood component analysis which gives highest accuracy of 79.79%.

[15] The author Kang et al. (2018) compared various models such as ANN, Random Forest, Decision Trees, Deep Belief Network (DBF), Least Squares Support Vector Machine Model, and found that DBF is superior because it considers hourly data prediction. The DBF inherently considers spatial and temporal correlations. Also, a stacked auto-recorder (SAE) model which is trained in greedy layer-wise manner extracts inherent air quality features. However, due to the device defects there were issues in recording high quality data.

[16] Jayant Kumar Singh & Amit Kumar Goel have used linear regression to predict the air quality. Their model shows an accuracy of 96%. However, the data attributes are collected from a specific zone of Delhi only.

3. Prediction Results of Algorithms

The table 1 below shows the results of numerous algorithms obtained by various researchers.

Table 1. Algorithms and Corresponding Results

Reference No.	Algorithm	Prediction Result
[1]	BP Neural Network	Accuracy: 80.44%
[2]	Random Forest	Accuracy: 70-86%
[3]	Neural Network	Accuracy: 92.3%
[4]	Artificial Neural Network	MRE: -0.16
[5]	Light GBM	MSE: 3762.021
[6]	Random Forest	MSE: 0.153(CO Concentration), 29.517(NO Concentration), 3.214(PM2.5 Concentration)

[7]	Extra trees	Accuracy: 85.3%
[8]	CERT	AQI accuracy: 0.9792
[9]	Neural Network	Accuracy: 99.86%
[10]	Extreme Learning Machine	Accuracy: 74.17%
[11]	SPSS Algorithm	Standard Deviation: 0.997 Error rate: 10%
[12]	Random Forest Algorithm	Accuracy: 79.25%
[13]	Random Forest Algorithm	Accuracy: 81%
[14]	Ensemble Neural Network with Neuro-Fuzzy Logic	Accuracy: 79.79%
[15]	Linear Regression	Accuracy: 96%
[16]	Deep Belief Neural Network	Error rate: 1.6%(PM2.5)

4. References

[1] Wang Zhenghua, Tian Zhihui, Prediction of air quality index based on improved neural network, 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC)

[2] Venkat Rao Pasupuleti, Uhasri, Pavan Kalyan, Srikanth, Hari Kiran Reddy, Air Quality Prediction Of Data Log By Machine Learning, 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS)

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

[4] Arwa Shawabkeh, Feda Al-Beqain, Ali Rodan, Maher Salem, Benzene Air Pollution Monitoring Model using ANN and SVM, 2018, The Fifth HCT INFORMATION TECHNOLOGY TRENDS (ITT 2018), Dubai, UAE, Nov., 28 - 29, 2018

[5] Yuelai Su, Prediction of air quality based on Gradient Boosting Machine Method, 2020 International Conference on Big Data and Informatization Education (ICBDIE)

[6] 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

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

[8] 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

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

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

[11] Limei Ma, Yijun Gao, & Chen Zhao, Research on Machine Learning Prediction of Air Quality Index Based on SPSS, International Conference on Computer Network, Electronic and Automation (ICCNEA), Shijiazhuang, China, 2020, IEEE

[12] Chuanting Zhang & Dongfeng Yuan, Fast Fine-Grained Air Quality Index Level Prediction

Using Random Forest Algorithm on Cluster Computing of Spark, UIC-ATC-ScalCom-CBDCom-IoP, 2015, IEEE

[13] Liying Li∗, Zhi Li∗, Lara G. Reichmann & Diane Myung-kyung Woodbridge, A Scalable and Reliable Model for Real-time Air Quality Prediction, SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing &Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, 2019, IEEE

[14] Krittakom Srijiranon & Narissara Eiamkanitchat, Neuro-fuzzy Model with Neighborhood Component Analysis for Air Quality Prediction, 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST), 2021, IEEE

[15] 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

[16] Jayant Kumar Singh & Amit Kumar Goel, Prediction of Air Pollution by using Machine Learning Algorithm, 7th International Conference on Advanced Computing & Communication Systems (ICACCS), 2021, IEEE

[17] NAAQS Table. (2015). [Online]. Available: https://www.epa.gov/criteria-air-pollutants/naaqs-table