ISSN: 2348-2079

Volume-11 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

# Air Pollution Prediction Using Machine Learning

**M. Lakumanan[1], Gokul. R[2] , Lavanya. S[3] , ManojKumar. M[4]**

*[1]Assistant professor, Mca,Department of computer science -Data Science, K.S.Rangasamy college of arts and science (Autonomous),Tiruchengode-637215.*
*[2,3,4]Bachelor of Science in computer science-Data Science, Department of computer science -Data Science K.S.Rangasamy college of arts and science (Autonomous),Tiruchengode-637215.*

**Corresponding Author: M. Lakumanan**
**Published on: 12.05.2023**

## ABSTRACT

The fact that environment tracking is focused largely on the fundamental rights of people, lifestyles, and health makes it so important. As a result, this device tracks the quality of the air using excellent sensor nodes within that check for CO2, NOx, UV light, temperature, and humidity. The gad get is able to categorise automatically if a certain geographic area is going above the established gas emission restrictions thanks to the statistics assessment using device mastering algorithms. In order to choose the most contaminated sectors, the DB SCAN with LR,SVM, and NB set of rules delivered a noteworthy category overall performance. Monitoring air quality is a crucial concern in many commercial and physical areas ofthe world.
In areas with serious difficulties with air pollution, Air Quality Operational Centers (AQOCs) are established specifically for this purpose. The AQOCs are operational units responsible for managing tracking networks, analysing the gathered data, andeventually disseminating online assessments of air pollutants and their short- and long-term evolution. Up until recently, modelling of air pollution events has been focused mostly on dispersion models, which approximate the complexphysicochemical processes at play. Although the intricacy and complexity of these models have increased over time, their application in real-time atmospheric pollution tracking appears to no longer be acceptable in terms of performance, input data requirements, and compliance with the problem's time limitations.

**Keywords:** Air Pollution

## INTRODUCTION

Many nations across the world are affected by air pollution, which can have deadly consequences for human health. Over the past century, ouratmosphere has gotten worse as a result of an increased reliance on fossil fuels. A number of vehicle types have a significant impact on pollution.The primary contributors to air pollution, which can affect people's health in the short- and long-term, areRSPM, SO2, NO2, SPM, and otherpollutants. The objective of this study is to ascertain if analytics methods can be utilised to create a system thatcan approximatively predict future pollution levels with a high degree of accuracy. It is shown that strategies for rendered linear regression are insufficient for the time-dependentdata. With a high degree of confidence,we have predicted future levels of several pollutants using a time series forecasting method. The efficiency of our recommended method employing SARIMA is demonstrated by the experimental examination of the forecasting for the levels of air pollution in Bhubaneswar City.

### *MACHINE LEARNING*

Science's field of machine learning enables computers to learn without explicit programming. One of the most intriguing new technologies is machine learning. The computer's ability tolearn, as the name suggests, is what gives it a more human-like personality.Machine learning is now being actively used in more places than one may think. Machine learning programmes are capable of doing tasks without having them explicitly written. For certain tasks, computers use the data that is readily available to learn. For simple tasks given to computers, it is possible to create algorithms that tell the machine exactly what to do in order to solve the problem at hand; thecomputer doesn't need to learn anything. A human could find it challenging to manually create the necessary algorithms for more complicated tasks. In actual use, it could be more effective to assist the computer in developing its own algorithm as opposed to having human programmers specify every essentialstep.

## SARIMA

Science's field of machine learning enables computers to learn without explicit programming. One of the most intriguing new technologies is machine learning. The computer's ability to learn, as the name suggests, is what gives it a more human-like personality. Machine learning is now being actively used in more places than one may think. Machine learning programmes are capable of doing tasks without having them explicitly written. For certain tasks, computers use the data that is readily available to learn. For simple tasks given to computers, it is possible to create algorithms that tell the machine exactly what to do in orderto solve the problem at hand; the computer doesn't need to learn anything. A human could find it challenging to manually create the necessary algorithms for more complicated tasks. In actual use, it could be more effective to assist the computer in developing its own algorithm as opposed to having human programmers specify every essentialstep.

## *RELATED WORK*

### *Air quality evaluation and pollutant concentration prediction in a air quality monitoring and early warning system.*

Wang Jian Due to the industrialization and urbanisation of many nations, air pollution is getting worse, which complicates the work of policymakers and contributes to climate change and poor public health. Develop a better scientific early warning and monitoringsystem for air quality in order to quantify air pollution levels objectively and anticipate pollutant concentrations. It is uncommon, nevertheless, for an air quality system to include both air quality assessment and air pollutant concentration forecasting. In this article, we suggest a novel air quality monitoring and earlywarning system including modules for assessment and forecasting. The principal pollutants are recognised andthe level of air pollution is more thoroughly analysed utilising fuzzycomprehensive evaluation in the air quality assessment module. To increase the precision of six major air Elman neural network, pollutantconcentrations, a modified cuckoo search and differential evolution approach, an innovative hybridization model with complementary ensemble empirical mode decomposition, and allof these are discussed. Utilizingpollution data from two Chinese cities, the method's efficacy is confirmed. The fuzzy comprehensive evaluation's findings indicate that PM10 and PM2.5are the main air pollutants in Xi'an and Jinan, respectively, and that Xi'an has superior air quality than Jinan. The forecasting outcomes show that the suggested hybrid model, with its improved prediction accuracy and stability, is unquestionably superior to all benchmark models. The air pollution problem in China is getting worse as a result of the country's burgeoning economy, increasing industry, and increased automobile ownership. More fossil fuels and natural resources must be consumed in order for a country with 1.4 billion people to prosper economically,changing the chemical make-up of the atmosphere. According to the Environmental Performance Index (EPI), which evaluates 180 nations for their environmental performance this year, China has the second-worst air quality in the world. The 10 nationswith the worst air quality in 2016 are listed in Fig. 1 together with the associated EPI values. According to Wang et al., air pollution has a substantial effect on the biological environment and has the potential to destroy flora and historic sites (2016). Numerous epidemiological studies have consistently demonstrated a relationship between air pollution and cardiovascular and respiratory disorders, as well as the potential for air pollution to induce lung cancer andother associated illnesses. There is a tremendous need for an effective, precise, yet user-friendly air quality monitoring and early warning system to maintain people's health and enhance their quality of life. In terms of long-term development and scientific decision-making, environmental monitoring and early warning are crucial components of environmental protection efforts. The Chinese peoplewish to clear the air and get rid of the fog. More than 2700 monitoring stations, more than 268,000 pieces of monitoring equipment, and more than 60,000 monitoring staff have all been established to achieve this (MEP Ministry of Environmental Protection, 2015). Even though China has made great strides toward clearing haze and enhancing air quality, persistent issueswith reducing air pollution still need tobe addressed.

### *A Case Study Applied to Tehran, Iran's Capital: A Novel Approach for Improving Air Pollution Prediction Based on Machine Learning Approaches*

Abdullah Mohammed Delavar Urbanization and the expansion of industry are primarily blamed for global environmental deterioration. In particular, Tehran, the capital of Iran, where its government and people havelong struggled with air pollution harm such as the health concerns of its citizens, has been acknowledged as one of the major issues in urban areas throughout the world. According to the research's study area, PM10 and PM2.5 particles are responsible for a sizable portion of Tehran's air pollution.The current study was carried out to create prediction models for estimating Tehran's air pollution levelsbased on PM10 and PM2.5 pollution concentrations. Information on the geography, the weather, the day of the week, the month of the year, and the degree of pollution in the two nearest neighbours were among the input factors. Calculations of the air toxicity were made using machine learning techniques. As a machine learning method for the prediction of air pollution, these techniques include regression support vector machines, spatially weighted regression, artificial neural networks, and auto-regressive nonlinear neural networks with an external input. The error rate was then decreased and improved by 57%, 47%, 47%, and 94%, respectively, by adding a prediction model to the previously mentioned methodologies. The recommended autoregressive nonlinear neural network with externalinput has a one-day prediction error of 1.79 g/m3, making it the mostaccurate method for predicting airpollution. Using the genetic algorithm, it was discovered that information on the day of the week, month of the year,terrain, wind direction, maximum temperature, and pollutant rate of the two nearest neighbours were the most reliable indicators for forecasting airpollution.

One of the most significantenvironmental problems has an influence on both developed and developing nations. "Air pollution" is defined as the presence of one or more pollutants in the air, inside or outdoors, for varied lengths of time and in varying quantities, which may harm people,

plants, or animals or have unexpected effects on daily life or property. The process through which air pollution is spread is intricate and dependent on several variables. Inactuality, predicting air pollution, whichhas a non-linear dynamic, is a highly challenging endeavor
that necessitates a thorough comprehension of how air pollutants disperse in the atmosphere at a large expense. The high air pollution inmegacities, which periodically exceeds the legal limit, raises concerns. It isacknowledged that air pollution is acrucial issue in urban management since it has become a concern in manyplaces throughout the world. As a result of the public's awareness of thisissue, policymakers were compelled toenact legislation to cut air pollution.Giving the people the knowledge they need to understand the levels of air quality is one of the goals of urban managers. By incorporating the densityof daily PM2.5 and PM10 pollutants, municipal managers may use the pollution data to inform the public about the problem of air pollution. It ispossible for individuals to lessen pollution by taking publictransportation and avoiding polluted places. In an effort to reduce pollution, concerned local authorities can use the data to restrict urban traffic, control polluting businesses, and expand the reach of public transportation. The right air pollution prediction tools must be used in order to accomplish this aim. The Tehran Air Quality Control Company (AQCC) maintains 21 stations, while the Iranian Environmental Protection Agency is in charge of 16 stations. According to themost recent statistics available, PM10and PM2.5 account for the majority of the air pollution concentration in Tehran. PM2.5 is the most prevalent pollutant, followed by CO, O3, NO2, SO2, PM10, and PM2.5. According to an AQCC research conducted in 2017 and the technical report created on the Tehran Air Pollution Prediction System, over 5% of PM2.5 pollutants originate from nearby inhabited areas in the west, particularly the cities of Karaj,Shahryar, and Rey. This fraction has been noted to be larger in the summer due to higher wind speeds that transport the dust blown from the westand trapped in the Greater Tehran basin. According to the AQCC's expert opinion, when compared to other pollutants that have been found to be under 5%, the amount of PM2.5 pollution shown here is the highest.Furthermore, the PM2.5 observedabove during the winter is not made up of wind- or naturally-occurring dust from far-off deserts.

### *Air Pollution Monitoring and Forecasting System Based on IOT*

Chen Xiaojun According to empiricalresearch, standard air automated monitoring systems provide excellent precision, but are inappropriate for large-scale installation due to their bulky design, expensive price, andsingle datum class. This paper proposes a strategy based on the usage of the Internet of Things (IOT) for environmental protection for real- time monitoring and forecasting of air pollution. The hardware cost of thissystem might be cut in half by utilisingIOT. The technique allows for the construction of a network of monitoring sensors in an extensive monitoring region. In addition to performing the duties of a standardautomated air monitoring system, it also demonstrates the capacity to anticipate the evolution of air pollutionover a certain time period by using neural network technology to analyse the information gathered by a front-end perception system. To reducelosses during actual use, targetedemergency disposal strategies may beimplemented.

The probability of environmental pollution accidents, particularly air pollution accidents, has grown due to the economy's rapid expansion and theincreased frequency of building and production operations in chemical industrial parks. Air pollution will beheavily concentrated quickly after it occurs due to geographic and climatic factors, inflicting significant injury or perhaps catastrophic devastation to both humans and the ecosystem. Therefore, it is crucial to instal a real-time air pollution monitoring system.
According to a laboratory examination, the traditional automated monitoringsystem for air is made up of expensive, sizable, and sophisticated machinery. Installation on a big scale is notfeasible because of the high cost and size. Due to the fact that this equipment can only be deployed in key monitoring sites for a small number ofkey businesses, it is impossible to acquire system data that would allow for the prediction of the entire pollution picture. The method described in this research solves the drawbacks of conventional monitoring systems and detection approaches while lowering test costs. It combines environmental monitoring with IOT technologies. By substituting sensor networks for the typical empirical analysis's monitoring equipment, IOC technology enables the flexible and economical placement of sensors throughout the region for omnidirectional monitoring and datasupport.

### *New Multilinear Regression Equations for Lateral Spread Displacement Prediction.*

Empirical formulations for predicting lateral spread displacement were first developed by Youd, Leslie Bartlett, and Youd in 1992 and 1995; they are now commonly used in engineering practise. A sizable case history record was utilised to create the equations using multilinear regression (MLR). The initial analysis is updated and corrected in this report. Corrections and changes include. These blunders are rectified here. Bartlett and Youd incorrectly overstated the reported displacements for the lateral spreads caused by the Nihonkai-Chubu, Japan earthquake in 1983. Due to boundary shear's restriction on unrestricted lateral movement, some places were eliminated. Three more earthquakes' worth of data were submitted. To avoid needlessly overestimating displacements as R goes smaller, the functional form of the mean-grain-sizecomponent has been changed to depend on the earthquake's magnitude. New MLR equations were produced by regressing the updated data. It isadvised that the new equations be used in engineering practise.
For the purpose of forecasting lateral spread displacement in liquefiable locations, Bartlett and you developed an empirical equation in the early 1990s. The equation has since been used often in engineering practise. Theequation was created using multilinearregression and a substantial case history record that these researchers had acquired.

### *Categorization, fuzzy sets, and multilayer perceptron's*

Mitra S This study provides a multilayer perceptron-based fuzzy neural network model that use backpropagation to classify patterns in a fuzzy manner. The output vector is described in terms of fuzzy classmembership values, whereas the inputvector consists of membership values to linguistic characteristics. Therefore, fuzzy uncertain patterns may be accurately characterised by suitably weighting the back

propagated errors based on the membership values atthe different outputs. The learning rateis steadily reduced during the course of training in discrete phases until the network converges on a minimumerror solution. The effectiveness of thealgorithm is shown in relation to a speech recognition issue. Theoutcomes are contrasted with those ofstandard MLPs, Bayes classifiers, and other related models.

## *EXISTING SYSTEM*

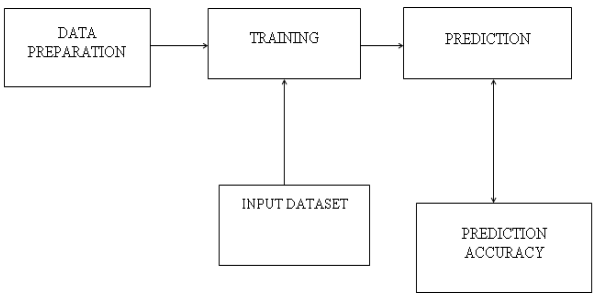
Making decisions on the system's software and hardware architecture, as well as which functionalities should be implemented in software running on programmable components and which should be implemented in more specialized hardware, is a crucial element of the design challenge. Embedded systems are frequently used in situations where reliability andsafety take precedence over performance. Today, manual design and past experience with goods that are similar to them have a significant impact on the on-the-fly creation ofembedded systems.

## *PROPOSED SYSTEM*

Effectively, however in our project, we used a convex hull to generate the statistics shape, allowing us to use a set of statistics that included date, time, temperature, CO2, NO2, O3, and PM10 in addition to SO2. With this method, we were able to create a structural database out of a sizable dataset by combining DBSCAN with LR, SVM, and NB to obtain statistics on the structural arrangement of air molecules. Now that DBSCAN cansupply a specified value, we are able toform clusters. (DBSCAN) is a collection of suggested statistical clustering rules. It is a density-based, fully clustering, non-parametric set of rules: given a fixed number of factors in a particular area, it groups together factors that are tightly packed together (factors with many near neighbours), designating as outlier's factors those that locate solely in low-density areas.

One of the most popular common place clustering methods and one thatis also frequently cited in medical literature is DBSCAN. After the systemis trained, the input data are analysed using a variety of machine learning methods to give accurate findings. Weutilise the Indian dataset, whichconsists of diverse data about air quality.

## *BLOCK DIAGRAM*




## IMPLEMENTATIONDATA

The records had come from reliablesources. The other carries meteorological data, whereas the first provides air quality records. Air qualityrecords From the Air Now website, we got the historical air fine data for New Delhi. The functions for our usage included in this dataset are year, month, day, hour, and AQI cost for each three hours starting at three am on 1-1-2015 to 24-4-2017, among othercolumns. 6700 values make up this. There are several columns, such as conc, conc. devices, etc., that may not be very useful to us. We only extend the meteorological records bycombining the hour, month, and dayvalues in the AQI column for ourpainting needs.

## METEOROLOGICAL DATA

The Indian data collection known as the Delhi Weather Dataset was acquired from Kaggle. It contains hourly weather data for Delhi from 1997 to 2017. From January 1st, 2015, through April 24th, 2017, we reduced itto every three hours of data. Date, time,cords, dewpt, fog, hail, heatindexm, hum, precpm, pressure, rain, snow, tamp, thunder, tornado, odd, wind gust,wind-chill, and wisdom are the columns in the climatic facts. Conditions such as fog, partly fog, mist,haze, light drizzle, rain, etc. are described by Cond's. Dew point is provided by dewpt, and fog and hail arealso indicated by the presence orabsence of fog. Precept denotes the precipitation; hum represents the humidity. pressure is used to describe pressure. The direction of the wind is represented by odd, and its speed by wisdom.

## DATA VISUALIZATION

First, we remove a few introductory rows from the dataset as their AQI values were no longer available. Then, we make an effort to visualise the dataand see how certain capabilities affectthe AQI tiers over time. We want toidentify the desired and unnecessary skills for our task when we plot the characteristic versus AQI graph. graphs showing the capabilities thatseem to affect AQI values.

## DATA CLEANING

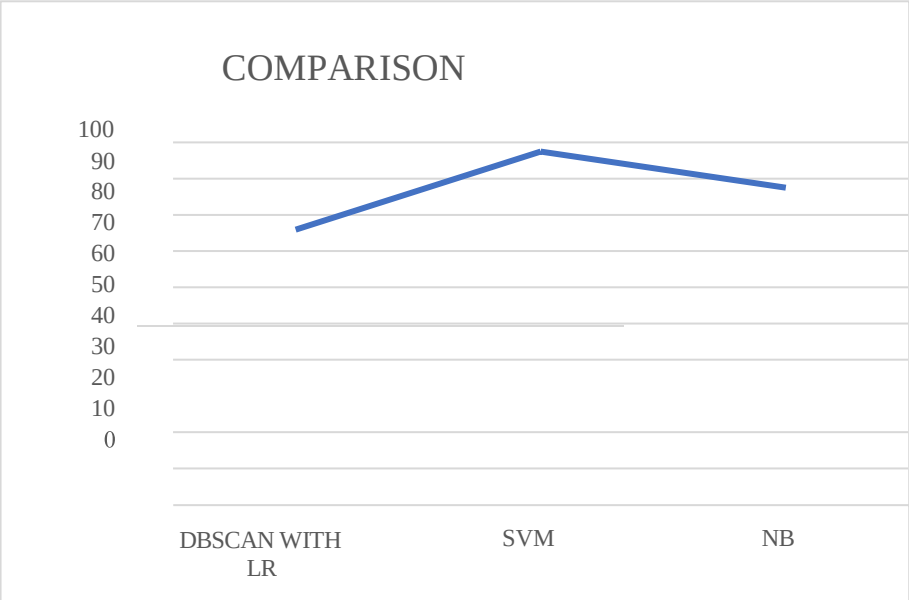
We can see from the plots that ourdesired capabilities include date and time, conditions, dewptm, hum, pressure, temperature, wdird, wspdm, month, day, and hour, and AQI. It isclear that the AQI is lower with higher wind speeds. A good predictor isprobably wind speed. Similar to how AQI scores are a little better in the winter, wind direction also influences. All other capacities typically have subpar results or no longer significantly affect the AQI rate. The ones columns are therefore removed. So that we may predict the AQI rate forthe current hour based on the past fivesupplied climates and AQI records, wenext establish previous fee capabilitieswith the help of transferring periods. Next, we deal with missing values or NaNs by removing those rows or, if it makes sense, by filling them withvalues from the imply column. These actions prepare our records for incorporation into the models.

## RESULT ANALYS

We use conventional plotting tools andSclera packages to execute it on an x64 system running Python 2.7, with 8GB of RAM, and an Intel i7 processor.The majority of the work was implemented using the panda's data frame. Each model was trained using the train set and assessed using the test set. The data was divided into a train set and test set in the ratio 7:3, or around 4400 train values and 2000 test values. The regression graphs for eachof the models we used to predict the AQI are shown below. Additionally, the feature significance analysis has madeuse of the decision forest and Extratrees. The results' accuracy scores andRMSE values are also included in the table below. We can see from the regression graphs that the models produce quite accurate results. The blue line represents the correspondingpredicted values in ascending order,and the orange line represents the testset points (set in ascending order), according to DBSCAN with Linear Regression and SVM (support vector Machine) and (Navie Bayes) NB. Thesemethods each have a different level of accuracy for predicting air quality, with SVM (support vector Machine) offering the highest accuracy values. The features are listed by the Extra Trees model in decreasing order of relevanceas follows: month, temperature, conditions, hour, pressure, and humidity readings from the past.

**Comparison analysis table**

| ALGORITHM | COMPARISON |
|---|---|
| DBSCAN WITH LR | 75.5 |
| SVM | 96.75 |
| NB | 86.92 |




**Comparison Graph**

## CONCLUSION

In this study, we tested the ability of the sklearn library's pre-existing regression models to forecast the values of the air quality index using historical meteorological data. Additionally, we made an effort to identify the traits that some of these models' predictions might benefit the most from. The accuracy of the various machine learning algorithmsfor the Indian dataset is supplied by the DBSCAN, which offers accuracy of 75%, NB, which offers accuracy of 86%,and the support vector machine, whichoffers the greatest accuracy of roughly96%. Additionally, additional data may be used, and real-time prediction techniques can be implemented in settings like Azure ML.

## REFERENCES

1. Yang Z., Wang J. Environ Res. Air quality evaluation and air pollutant concentration prediction: a unique air quality monitoring and early warning system Yang. 2017;158:105-17. doi: 10.1016/j.envres.2017.06.002, PMID 28623745, Z., and Wang, J.
2. Nakhaeizadeh GR, Fedra K, HatefiAfshar S. ISPRS International Journal of Geoformation. 2019. An application to Tehran's capital city of an unique technology for improving air pollution prediction based on machine learning methodologies. Along with M. R. Delavar, A. Gholami, and G. R. Shiran, the authors are Y. Rashidi;8(2):99.
3. Jun CX, Peng LX, Peng X. IoT-based air pollution monitoring and forecasting system, 2015 international conference on computer and computational sciences. In: ICCCS. IEEE Publications; 2015. p. 257-60.

4. Youd TL, Hansen CM, SF. Bartlett claim that changes have been made to multilinear regression equations for forecasting lateral spread displacement in Journal of Geotechnical and geoenvironmental Engineering. Vol. 128(12); 2002. p. 1007-17.
5. In 1992, a study titled ”multilayer Perceptron, Fuzzy Sets, Classification” was published by S.K. Pal and S. Mitra.