INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 06 ISSUE: 06 | JUNE - 2022 IMPACT FACTOR: 7.185 ISSN: 2582-3930

A Review on Analysis and Prediction of Air Quality Using Machine Learning Techniques

Surabhi Sharma
Deptt. of Electronics & Communication Engineering
Shriram College of Engineering & Management (SRCEM)
Banmore, Gwalior(M.P)
ssurabhissharma@gmail.com

Prof. Ashish Duvey
Deptt. of Electronics & Communication Engineering
Shriram College of Engineering & Management (SRCEM)
Banmore, Gwalior(M.P)
adsrcem@gmail.com

Abstract*— Human health is greatly influenced by air quality. Air pollution causes a variety of health problems, particularly in youngsters. The capacity to anticipate air quality allows the government and other concerned groups to take the required precautions to protect the most vulnerable from being exposed to harmful air quality. To ensure optimum air quality, the air quality measurement system analyses different air contaminants in various areas. It is the most pressing problem in the current situation. The entry of harmful gases into the climate from industry, car emissions, and other sources pollutes the air. Conventional methods to this issue have had extremely little success due to a lack of access to significant longitudinal data for such techniques. Currently, air pollution has reached alarming levels, with several large cities exceeding the government's air quality standards. It has a significant influence on human health. With advances in machine learning techniques, it is now feasible to anticipate contaminants based on historical data. Machine learning methods are used to anticipate air pollution levels so that individuals may take preemptive steps to reduce air pollution. In this research, we provide a variety of ML techniques that can continue can take existing pollutants, however, with the aid of previous pollutants, we are running an algorithm based on machine learning to forecast the future data of pollutants. This approach can continue to take current pollutants.

Keywords*— Air pollution., Air Quality prediction, Artificial Intelligence, Machine Learning, supervised and unsupervised.

I. INTRODUCTION

In recent years, the huge reduction in air quality has been mostly attributable to a rapid expansion in industry and automobile density, both of which exacerbate air environment pollution. Accurate estimation of pollutant concentrations in the atmosphere is necessary with sufficient time-distance to manage the air standard up to a non-hazardous level and to design an air pollution management program. The majority of air-contaminated nations have implemented an active monitoring system to eliminate key air pollutants in severely polluted regions under their control.[1].

Analytical or statistical models may be used to arrive at an accurate estimate of air quality to determine the level of air pollution. Long-term forecasting as well as planning choices are frequently better served by analytical models[2]. However, such models fail to give good results for air pollutant series with fast dynamics[3]. Furthermore, analytical approaches are unable to provide a quantitative evaluation of environmental degradation in the absence of data on additional input elements such as temperature, wind, and traffic aspects for assessing the emission rate[4]. When further input factor data is lacking, stochastic modeling gives an alternate technique for dealing with time series of air contaminants[1].

In recent years, the application of Artificial Intelligence (AI), as well as Machine Learning, has aided in the growth of technology. ML refers to the process through which an AI system learns from data. Using machine learning approaches, several methodologies are used to forecast/classify the Air Quality Index (AQI). ML is a sub-field of AI. Its purpose is to let the computer learn on its own without having to be expressly loaded with the rules. To model as well as forecast the environment, a machine learning system may find and understand underlying patterns in observational data.

II. AIR QUALITY PREDICTION

Air is the material foundation on which humans live, as well as air quality, is a crucial predictor of whether or not a city is habitable. Air pollution is one kind of pollution that passes through a sample of the air, whether inside or outdoors. Pollutants infiltrate the atmosphere & making it more challenging for plants, animals, and even people to exist when the air gets polluted. The atmosphere, which is made up of many gases, is responsible for the survival of all living beings. Changes in these gases may generate an imbalance that is hazardous to survival.

Because of the fast growth of urbanization and industry in recent decades, air quality has emerged as a major concern in the context of environmental health. Because of the influence that air quality has on people's lives on a day-to-day basis, figuring out how to accurately forecast air quality has become

an important and pressing issue. The forecasting of air quality is a difficult subject because it relies on a number of intricate parameters, many of which are interdependent in other ways.

The entry of particles, biological molecules, or other toxic elements into the Earth's atmosphere causes sickness, death, or degradation to other living species including such food crops, as well as damages to the natural or man-made environment. An air pollutant is a chemical in the air that may harm individuals and the environment. Solid particles, liquid droplets, or gases are all examples of substances. Pollutants are divided into two types: primary and secondary. Primary pollutants are often created by a process, including such volcanic ash. Other instances are carbon monoxide gas emitted by automobiles and sulphur dioxide emitted by companies. Secondary pollutants are not directly released. Rather, they arise in the air as a result of main contaminants reacting or interacting. A common example of a simple pollutant is ground-level ozone. Ground level ozone (O3), fine particulate matter (PM2.5), carbon monoxide (CO), nitrogen dioxide (NO2), sulphur dioxide (SO2), as well as lead are the six "criteria pollutants," with ground level O3, PM2.5, as well as NO2 (the primary component of NOx) posing the most serious health risks.

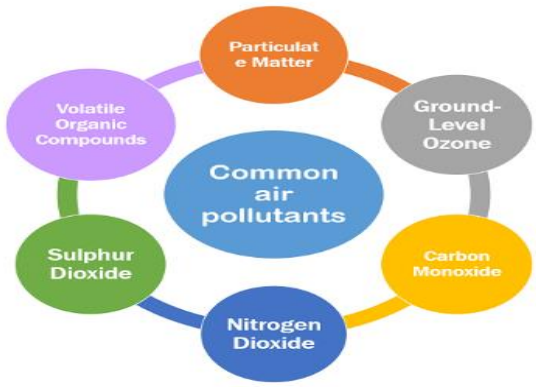

Figure 1: Different Types of Air Pollution

The state of the air in our surroundings is referred to as air quality. The degree to which the air is pure, clear, and devoid of contaminants such as smoke as well as dust, among some other gaseous impurities, is referred to as air quality. Table 1 outlines the major categories of contaminants and provides a brief explanation of each. Good air quality is essential for sustaining the delicate balancing act of life on Earth for people, plants, animals, or natural resources, which are jeopardised when pollution levels in the air reaches a critical level.

Table I: Types of pollutants

Name	Information
CO	The primary source of carbon monoxide is the burning of either natural gas, coal, or wood.
CO_2	At a stable level, carbon dioxide is both natural as well as needed, yet its levels have been steadily rising at an alarming rate.
SO_X	One of the things that should worry people about the ecological impact of using these fuels as power sources is the release of sulphur oxides, which may come from both volcanoes as well as industries.
NO_X	Combustion engines, such as those used in service and manufacturing, produce nitrogen oxides as a byproduct.
PM	The term "PM2.5" refers to particulate matter with a diameter of less than 2.5 microns, whereas "PM10" refers to particulate matter with a diameter of 10 microns. It is very dependant on the variables that are present in the area, such as the weather, the traffic, as well as the pollution. Long-distance transportation, the generation of road dust, and the burning of wood are the primary contributors in Norway.
O3	Ozone is a greenhouse gas that is produced when sunlight interacts with the atmosphere. Ozone may be formed in the atmosphere as a result of a reaction between hydrocarbons as well as nitrogen oxides either immediately at the source of the pollution or several kilometres downwind.
VOC	The release of volatile organic molecules, such as methane, into the atmosphere may contribute to an accelerated rate of global warming since methane is an exceptionally effective greenhouse gas.

2.1 What Causes Reduced Air Quality

Emissions from many sources continually degrade air quality. These are either natural or man-made sources. Natural causes include volcanic eruptions, windstorms, biological decomposition, and forest fires. Pollution from passing automobiles, industrial facilities, power plants, smelters, and burning wood or coal are examples of man-made sources. Pollutants from all these sources are emitted into the

atmosphere, posing serious health risks to people, animals, and the ecosystem. The amount of pollutants, the pace at that they are discharged into the atmosphere, as well as the length of time they are confined in a region all have an impact on air quality. If air contaminants are present in a location with excellent airflow, they will mix with the air and swiftly disperse. Because certain factors, such as weak breezes or impediments, prevent contaminants from being transported away from a location, they tend to linger in the air. As a result, the concentration of air pollution rises fast[5].

Environmental researchers have spent a lot of time and money in the past on this topic using traditional methods. Nevertheless, a wide range of variables impact air quality, including its location, time of day, and other unpredictability. Because of advances in big data technologies and the accessibility of atmospheric sensing networks including sensor data, researchers have recently begun to apply a big data analytics method to analyse, evaluate, and forecast air quality. When it comes to air pollution study, the most pressing issue is determining which modelling methods are appropriate for the task at hand. A variety of approaches, such as "Climatology," have been employed for air quality forecasting, predicated on the concept that the past is a good predictor of the future It is common for these methods to be used to forecast exceeding limits from specified thresholds rather than ambient amounts. Because of this, there is a lot of room for development in this area of prediction analysis. Most approaches are unable to accurately anticipate pollution levels because of the limitations of available data and the importance (priority) assigned to it. There are a variety of functions that may be used to fit pollution data as predictors using regression and machine learning approaches. Variable time periods, such as multi-leveled equations and visuals and tables, may be used to examine the distribution and concentration of various contaminants in the environment. These components must be created by the formatter using the following criteria.

The use of machine learning may open up new avenues for predicting pollution levels in the air. It's possible to use machine learning to teach computers to do things on their own, without having to provide them with explicit instructions. Currently, it is one of the most popular and rapidly growing areas of study. Simple pieces of data called training data are used by machine learning systems. This information is nothing more than previously acquired and archived data. In order to obtain the data, data mining methods are used. This data is used to train a machine learning technique, and the performance of the results improves as the amount of training data rises. Unsupervised learning as well as supervised learning are two types of machine learning applications[6].

III. MACHINE LEARNING TECHNIQUES

ML is a branch in AI. It is a large interdisciplinary discipline that incorporates concepts from programming language, mathematics, successfully developed, cognitive science, architecture, as well as a variety of other math and scientific areas. Currently, we employ machine learning in our everyday lives. Computer software that are ML-based may obtain information and then use it to learn for themselves. It suggests that in ML, past expertise is exploited to produce predictions. There are four sorts of ML methods: Supervised Learning, in that direct supervision is involved, restricts the boundaries of the techniques by having the programmer label the dataset. Unsupervised Learning does not need supervision, semi-supervised machine learning, that incorporates both forms of supervised learning & unsupervised learning, is performed in a mixed structure. The focus of this work is on supervised and unsupervised ML approaches, that are discussed in greater depth in following parts [7] [8] [9][10].

Figure 2: Machine Learning areas

3.1 Categories of Machine Learning

Machine Learning Algorithms divided into categories according to their purpose[11]:

1) Supervised learning

This really is the system that will be used to supply the input as well as the required output for any further data processing. Regression as well as classification are the two main types of learning tasks that are included in this. Support Vector Machines, also known as SVMs, Genetic algorithms, Decision Trees, also known as DTs, k-Nearest Neighbors, or k-NNs, or Artificial Neural Networks, or ANNs, are among the most prevalent types of methodologies.***Unsupervised learning***

This is done so that inferences may be drawn from datasets that just include input data and not tagged responses to those inputs. Association as well as clustering are the two types of learning tasks that are included in this. to discover the connections between the different items in a database. Apriori is the most well-known algorithm in use in association rules,

while clustering is the process of grouping together different kinds of datasets that are comparable. The k-means clustering technique as well as the association rule learning method are two of the most frequent types of techniques.

2) Semi-supervised Learning

This is a hybrid of data with labels and data without labels, which places it between supervised learning as well as unsupervised learning. The primary applications for this learning include the categorization of webpages, genetic sequencing, as well as voice recognition. Clustering & classification are the two primary learning activities that fall under the umbrella of semi-supervised learning.

3) Reinforcement Learning

This is an example of a technique known as machine learning. It is associated with how software agents autonomously decide the best behaviour within a certain environment in order to optimise its effectiveness. This is done in order to make the software as efficient as possible. The reinforcement signal communicates the reward feedback to the agent, which enables the agent to learn how to improve its behaviour. Classification and control are the two components that make up this learning problem. Applications of this technology include computer-controlled board games, robotic hands, including automobiles that drive themselves. The algorithms Q-learning, Temporal differences, and Deep Adversarial Networks are the ones that are used most frequently.

3.2 Machine Learning algorithms

In order to do the data analysis, machine learning methods were used. The classifier algorithms are the ones responsible for the classification. The data from the reviews are utilised for creating predictions on the air quality. By using a wide variety of algorithms as well as machine learning ideas, these methods provide an illustration of how the study is performed on forecasting air quality. [12][13][14][15].

1) Naive Bayes (NB)

NB is dependent on the Bayes theorem using speculation amongst predictions. The Naive Bayes approach makes it possible to rapidly build frameworks that can forecast the future while also offering a new way to explore & comprehend the data. When using Naive Bayes to develop a predictive model, this method may be used for predictive analysis. Utilizing naive Bayes as a predictor is not a bad idea.

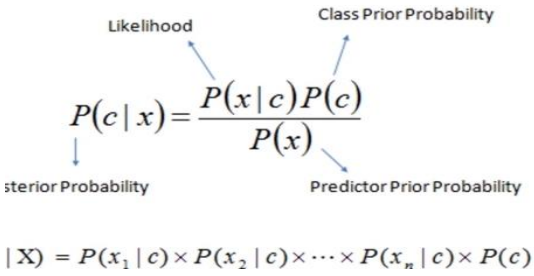

Figure 3: NBC algorithm

2) Decision tree

The decision tree approach is perhaps the most often used in DM. One of the most effective classifiers is the decision tree, which can be implemented quickly and easily. A decision tree is a prediction model for data mining that makes use of a decision tree. A decision tree and a categorization algorithm are both utilized in this study to estimate illness based on patient data.

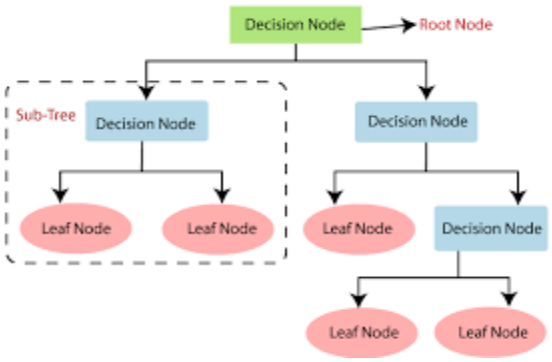

Figure 4: Structure of Decision Tree

3) Artificial Neural Network

This information processing unit, inspired by human brains, is known as an Artificial Neural Network (ANN). Usually, neural networks are arranged in layers, with every layer consisting of several linked nodes, each with an activation function. It is the input layer that presents patterns to the network & interacts with one or more hidden layers, where the processing is carried out using a system of connection weights. To get the detecting outcome, the hidden layers are connected to an output layer.

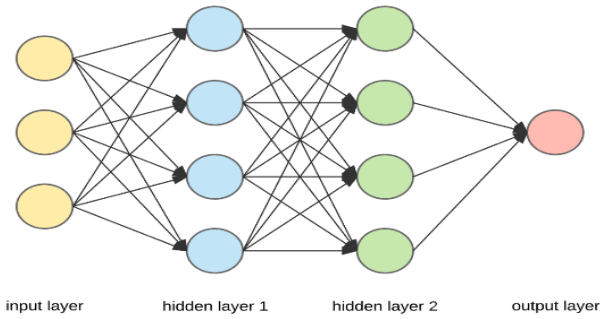

Figure 5: Structure of Artificial Neural Network

4) Random forest

RF is a classifier of tree-defined collection, in which individually assigned random vectors are identically distributed and input x is the unit for each tree. Most of the time, random forest yields good results. Increasing its efficiency is challenging, and it can also deal with many sorts of data like numerical, binary, & nominal).

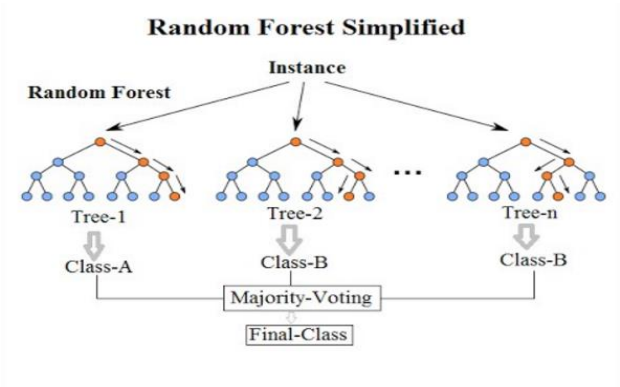

Figure 6: Structure of random forest

5) Logistic Regression (LR)

It's made up of logistic functions & looks like a sigmoid curve with an output of 0 and 1. In the form of an S curve, LR demonstrates growth & increases between 0 and 1 in the range.

6) XGboost (XGB)

(XGB) depending on the gradient boosting method, it has been built in an improved version to increase efficiency and speed. There are three primary parameters to the methodologies: boosters, learners, & general. In regression & tree, boosters' parameters are in charge of making the booster work. whereas learning variables are accountable for optimization & general variables are in charge of how well a method works as a complete.

7) Support Vector Machines (SVMs)

SVMs have supervised learning techniques for classifying, regression, & detecting outliers, among other things. Employing SVM has a lot of advantages, including the following: i) It works well in high-dimensional spaces, ii) Because it only uses a subset of decision function's training points (known as support vectors), it uses less memory, iii) It is flexible, since alternative kernel functions may be given for decision function.

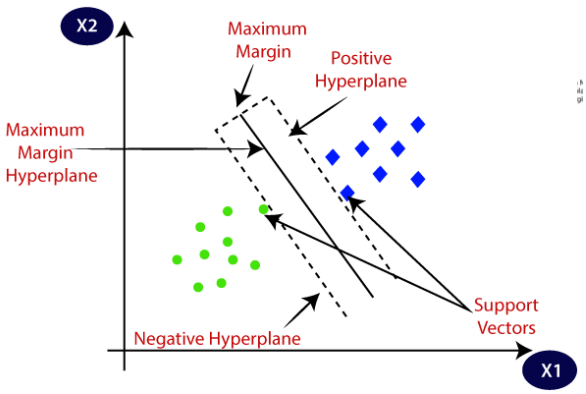

Figure 7: Structure of SVM

IV. LITERATURE REVIEW

In this part of the paper, we will review the several publications that are linked to the ML approach for predicting air quality. We take all of the papers from the most recent years. Several research on predicting air quality using machine learning techniques have been carried out. These investigations have been undertaken.

This paper [16], presents a model for predicting the grade of air quality relying on the K neighbour technique. Initially, the appropriate meteorological website's historic air quality monitoring info is crawled & stored to a local CSV file; the information is then read as well as the statistical method is being used to graphically present the six variables that impact the air quality level assessment; last, the K closest neighbour method is picked and the discrepancy is corrected. By parameters training and subsequent testing, a 95.10 percent accuracy rate was achieved. It is now possible to forecast the air quality level based on a random collection of data, which is in accordance with the predicted findings.

In this paper[17], Extraction as well as prediction features of Beijing's air quality monitoring data were extracted and predicted using two approaches, LightGBM's & XGB. The predictive performance as well as operation time of both approaches were examined. LightGBM's accuracy &

production performance were found to be much superior to those of XGB, it was concluded at the end of the day.

In this paper[18], system for the prediction of air quality that is geared toward revenue and is built on cloud microservices is offered for smart cities. An air quality prediction system founded on RF is built for each city and used to forecast the air quality attributes, including such SO2, NO2, RSPM, as well as the like, collected data. Revenue-oriented cloud microservices developed in golang are started for each connection to the conceptual methodology for air quality predictions, while credits for city stations are made depending on the predictions for each query. In the studies, the methodology for forecasting SO2, NO2, as well as RSPM achieved above 70% to 90% prediction accuracy percentage.

In this paper[19], AQI is utilized as a decision element for PM2.5, PM10, SO2, NO2, CO as well as O3 8h. In order to build a prediction model, multiple regression techniques are chosen and their accuracy and generalizability are assessed. AQI as well as air quality level may be accurately predicted using RFR or GBR, according to the findings. This document serves as a resource for developing an air quality modelling framework.

This paper [20], predicts the AQI, a measure of pollution levels, using a variety of ML methods. The AQI measures the degree of pollution in the air. In addition to particle matter, NO2, SO2, as well as carbon monoxide make up the bulk of the air pollution in the US. Air quality can be predicted using probability or statistics; however, these approaches are difficult to predict. To address the limitations of prior methods, machine-learning methods provide a superior strategy for estimating levels of air pollution. RF Regression, SVR, while LR are all examples of ML methods. The RMSE approach is used to assess the accuracy of various models.

To be able to accurately anticipate the city's air quality[21], model for predicting the air quality index is developed using the upgraded BP neural network in this article. Because air quality is affected by so many variables and cannot be predicted linearly, the model relies on the nonlinear fitting approximation properties of the BP neural network. Genetic method is used to calculate the issue of slow convergence as well as the tendency to slip into a local optimal solutions of the BP neural network. The city of Xuchang serves as a case study for this research. According to findings, the air quality indicator has an average relative inaccuracy of 22%. 80% of the time, the results were correct. Air quality measurements have an accuracy of 82.5 percent. BP neural network predictions enhanced is more accurate than the original BP neural network. The design is adaptable and feasible in certain ways.

Table II: Comparison Table of Related Work[22]

Publication	Title	Method	Limitation
IEEE, 2016	Predicting Trendsin Air pollution inDelhi usingData Mining.	LR, MLP, Time series analysis	Using LR, only linear connections between variables are considered. Thesis might be erroneous at times.
IEEE, 2016	Air Pollution Monitoring System with Forecasting Models.	SVM, ANN	Designed to fill in the blanks and turning categorical input into numeric value are necessary for neural networks. The structure of the NN has to be defined.
AMCS,2016	Data mining methods for prediction of air pollution	SVM Regression RF_fusion	For huge datasets, the SVM method does not work as well as expected. The more noise in the data the less effective SVM is.
Springer, 2018	Pollution prediction using extreme learning machine: a case study on delhi.	ELM(Extreme Machine Learning)	However, ELM can only store one layer of abstraction, therefore it cannot be described as "deep."
Elsevier, 2018	Forecasting airpollution load inDelhi usingdata analysis tools.	Time series regression	Regression on time series data is used in this example.

V. CONCLUSION

Air Quality (AQ) is a measure of how polluted our surroundings as well as homes are. The prediction of AQ values based on previous data may assist us in Analysing and mitigating pollution levels. AQ values may be categorized into predefined categories, and machine learning methods can be used to increase the classification performance of the derived Air Quality value. The primary goal of this work is to inform future researchers about the significance of different Machine Learning algorithms used for forecasting air quality. Based on the results of the study, it can be stated that ML is the best approach for predicting air pollution. The accuracy of ML in forecasting air pollution has been shown. As an output, air pollutant concentrations such as SO2, NO2, O3, CO, and particle matter have been predicted. As input parameters for predictions, meteorological elements such as temperature, humidity levels, absolute humidity, and wind direction may be employed. Additionally, the model may be interfaced with online apps so that users can profit from the work as well as take care to reduce air quality.

VI. References

[1] I. Nadeem, A. M. Ilyas, and P. S. Sheik Uduman, "Analyzing and forecasting ambient air quality of chennai city in india," *Geogr. Environ. Sustain.*, 2020, doi: 10.24057/2071-9388-2019-97.

[2] K. Juda-Rezler and P. N. Cheremisinoff, "Air pollution modeling," 1989, doi: 10.1201/ebk1439809624-c3.

[3] G. J. Cats and A. A. M. Holtslag, "Prediction of air pollution frequency distribution - Part I. The lognormal model," *Atmos. Environ.*, 1980, doi: 10.1016/0004-6981(80)90285-1.

[4] J. R. Zimmerman and R. S. Thompson, "User's guide for HIWAY, a highway air pollution model," *EPA PUBL.*, 1975.

[5] A. J. Lepper, "Air Quality Prediction with Machine Learning," no. June, 2019.

[6] J. Reshma, "Analysis and Prediction of Air Quality," *Int. Res. J. Eng. Technol.*, pp. 266–270, 2020, [Online]. Available: www.irjet.net.

[7] M. Limbitote, "A Survey on Prediction Techniques of Heart Disease using Machine Learning," *Int. J. Eng. Res. Technol.*, vol. 9, no. 06, 2020.

[8] Y. Nikhate, "SURVEY ON HEART DISEASE PREDICTION USING MACHINE LEARNING," *Int. J. Creat. Res. Thoughts*, vol. 8, no. 8, 2020.

[9] H. Hormozi, E. Hormozi, and H. R. Nohooji, "The Classification of the Applicable Machine Learning Methods in Robot Manipulators," *Int. J. Mach. Learn. Comput.*, 2012, doi: 10.7763/ijmlc.2012.v2.189.

[10] H. Animesh, K. M. Subrata, G. Amit, M. Arkomita, and A. Mukherje, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review," *Adv. Comput. Sci. Technol.*, 2017.

[11] S. Kusuma and D. U. J, "Machine Learning and Deep Learning Methods in Heart Disease (HD) Research," vol. 119, no. 18, pp. 1483–1496, 2018.

[12] T. Mehmood and H. B. M. Rais, "Machine learning algorithms in context of intrusion detection," 2016, doi: 10.1109/ICCOINS.2016.7783243.

[13] T. Madan, S. Sagar, and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms-A Review," *Proc. - IEEE 2020 2nd Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2020*, pp. 140–145, 2020, doi: 10.1109/ICACCCN51052.2020.9362912.

[14] V. R. Pasupuleti, Uhasri, P. Kalyan, Srikanth, and H. K. Reddy, "Air Quality Prediction of Data Log by Machine Learning," *2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020*, pp. 1395–1399, 2020, doi: 10.1109/ICACCS48705.2020.9074431.

[15] R. Murugan and N. Palanichamy, "Smart city air quality prediction using machine learning," *Proc. - 5th Int. Conf. Intell. Comput. Control Syst. ICICCS 2021*, no. Iciccs, pp. 1048–1054, 2021, doi: 10.1109/ICICCS51141.2021.9432074.

[16] Y. Gong and P. Zhang, "Research and Realization of Air Quality Grade Prediction Based on KNN," in *2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM)*, 2021, pp. 299–304, doi: 10.1109/AIAM54119.2021.00068.

[17] Y. Su, "Prediction of air quality based on Gradient Boosting Machine Method," 2020, doi: 10.1109/ICBDIE50010.2020.00099.

[18] S. Benedict, "Revenue oriented air quality prediction microservices for smart cities," 2017, doi: 10.1109/ICACCI.2017.8125879.

[19] C. Li, Y. Li, and Y. Bao, "Research on Air Quality Prediction Based on Machine Learning," 2021, doi: 10.1109/ICHCI54629.2021.00022.

[20] K. Mahesh Babu and J. Rene Beulah, "Air quality prediction based on supervised machine learning methods," *Int. J. Innov. Technol. Explor. Eng.*, 2019, doi: 10.35940/ijitee.I1132.0789S419.

[21] W. Zhenghua and T. Zhihui, "Prediction of air quality index based on improved neural network," 2018, doi: 10.1109/ICCSEC.2017.8446883.

[22] gayathri, Shankar, and Duraisamy, "Air Pollution Prediction using Data Mining Technique," pp. 4292–4297, 2020.