electronics

Article

A Radical Safety Measure for Identifying Environmental Changes Using Machine Learning Algorithms

Pravin R. Kshirsagar [1], Hariprasath Manoharan [2], Shitharth Selvarajan [3,*], Sara A. Althubiti [4], Fayadh Alenezi [5], Gautam Srivastava [6,7] and Jerry Chun-Wei Lin [8,*]

[1] Department of Artificial Intelligence, G.H Raisoni College of Engineering, Nagpur 414008, India; pravinrk88@yahoo.com
[2] Department of Electronics and Communication Engineering, Panimalar Institute of Technology, Poonamallee, Chennai 600123, India; hari13prasath@gmail.com
[3] Department of Computer Science and Engineering, Kebri Dehar University, Kebri Dehar P.O. Box 250, Ethiopia
[4] Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al-Majmaah 11952, Saudi Arabia; s.althubiti@mu.edu.sa
[5] Department of Electrical Engineering, College of Engineering, Jouf University, Sakaka 72388, Saudi Arabia; fshenezi@ju.edu.sa
[6] Department of Mathematics and Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada; srivastavag@brandonu.ca
[7] Research Center for Interneural Computing, China Medical University, Taichung 40402, Taiwan
[8] Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, 5063 Bergen, Norway
* Correspondence: shitharth.it@gmail.com (S.S.); jerrylin@ieee.org (J.C.-W.L.)

Citation: Kshirsagar, P.R.; Manoharan, H.; Selvarajan, S.; Althubiti, S.A.; Alenezi, F.; Srivastava, G.; Lin, J.C.-W. A Radical Safety Measure for Identifying Environmental Changes Using Machine Learning Algorithms. *Electronics* **2022**, *11*, 1950. https://doi.org/10.3390/electronics11131950

Academic Editor: George A. Papakostas

Received: 3 May 2022
Accepted: 19 June 2022
Published: 22 June 2022

Abstract: Due to air pollution, pollutants that harm humans and other species, as well as the environment and natural resources, can be detected in the atmosphere. In real-world applications, the following impurities that are caused due to smog, nicotine, bacteria, yeast, biogas, and carbon dioxide occur uninterruptedly and give rise to unavoidable pollutants. Weather, transportation, and the combustion of fossil fuels are all factors that contribute to air pollution. Uncontrolled fire in parts of grasslands and unmanaged construction projects are two factors that contribute to air pollution. The challenge of assessing contaminated air is critical. Machine learning algorithms are used to forecast the surroundings if any pollution level exceeds the corresponding limit. As a result, in the proposed method air pollution levels are predicted using a machine learning technique where a computer-aided procedure is employed in the process of developing technological aspects to estimate harmful element levels with 99.99% accuracy. Some of the models used to enhance forecasts are Mean Square Error (MSE), Coefficient of Determination Error (CDE), and R Square Error (RSE).

Keywords: air quality; predicting system; environment; air pollution; machine learning techniques

1. Overview of Contamination—An Introduction

The envelope of gases that surrounds humans every day is termed the atmosphere. Carbon emissions are one of the greatest sources of industrial pollution as they occur due to indiscretions in human activities and serious risks that are polluting the water. The composition of air pollutants in the surrounding atmosphere is affected by airspeed, wind patterns, and moisture levels. When there is a lot of humidity in the air, we sweat more because our perspiration cannot evaporate. Human activity, such as driving a combustion engine car is a major source of pollution due to increased transportation services [1]. Another major source of air pollution is mass production. The most prevalent pollutants are nitrogen oxide (NO), carbon monoxide (CO), particulate matter (PM), sulphur dioxide (SO_2), and others. Carbon monoxide is produced when a combustible, such as oil or gas, is not properly oxygenated. Nitrogen oxides create stomach pain; carbon dioxide causes headaches and vomiting; phenol causes breathing problems; nitrogen oxides cause headaches and nausea; microscopic matters, with a dimension of 2.5 mm or less have a greater impact on human health. Efforts must be taken to limit carbon emissions in the environment. The Air Quality Index (AQI) was used to assess the quality of the indoor environment. Predicting water quality using standard methods, such as mathematical and statistical methods is difficult due to the enormous amount of data required. Air pollution is a severe ecological calamity in both developed and developing economies. Nitrogen oxide is a pollutant that can harm humans, plants, or living organisms, as well as cause various problems with daily life or property [2]. The dispersion of carbon emissions is influenced by several variables. Predicting non-linear liveliness in carbon emissions, on the other hand, is a difficult problem that necessitates extensive knowledge of how air pollutants spread in the environment, which is costly [3].

Contaminants in urban environments may exceed what is considered safe, causing even more concerns. As a result, poor air quality has become a major worry for cities all over the world, prompting city planners to conduct studies as a primary priority. The public's awareness of the problem has prompted authorities to take action to reduce air pollution. One of the key tasks of urban planners is to educate the public about air quality assessments [4,5]. Municipal administrators may make public notifications concerning the frequency of average PM 2.5 and PM 10 particulates in response to air pollution [6]. People can use this information to avoid harmful areas and reduce pollution by taking public transportation. Municipal officials, on the other hand, may employ artificial intelligence to limit urban traffic and, indeed, polluting enterprises, as well as to improve public transit infrastructure to lower pollution levels. Computer vision technologies allow for reliable forecasting of future AQI levels, allowing for appropriate remedial action. Recurrent neural networks, transfer learning, and evolutionary computation are three different deep learning methods that all fall under the umbrella term of machine learning [4]. In the proposed study, a deep learning method was used. The approaches Support Vector Machine (SVM), Naive Bayes, and Random Forest are only a few of the many that fall under the umbrella phrase "machine learning techniques." We utilize Random Forest to anticipate air quality since it exceeds all of the other approaches in terms of accuracy.

1.1. Literature Survey

The researchers in [5] investigated water quality by using the Bias networking and forming a DAG using Kazakhstan's data. A subset of the database is used to develop a development or certification model. Consequently, the findings may differ depending on variables, such as geography and cultural setting. This technique has certain drawbacks where it is deciphered in [6], using an IoT-based vehicle emissions data collection method. The Internet of Things (IoT) based operation in vehicular systems is used for monitoring the amount of pollution that is produced by several vehicles where an automatic procedure of switching off the vehicles is enabled. Clean air prediction has been improved by using the Long Short-term Memory (LSTM) method, which reduces the amount of time it takes to train models. However, alternative methods, such as the Random Forest approach may be used to ensure efficiency. In [7–10] a specially engineered system is suggested; carbon dioxide and nitrous oxide are predicted using a nonlinear regression model. Toxic materials from a nearby industrial zone, such as Skikda have been considered, along with speed and altitude, air orientation, temperatures, and relative humidity. They utilized Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to judge the effectiveness; however, this technique only examined two components, NO and CO, and the other main contaminants, such as sulphur dioxide, PM 2.5, and PM 10 are just not examined. In [11] air quality using Nave Bayesian and J48 classification techniques has been analyzed. That one was 86.66% accuracy employing Naive Bayes, as well as 91.9% accuracy using the J48 random forest algorithm. The J48 method delivers more valid information than Nave Bayesian, and the inventor also justifies this.

In [4], improved identification and model accuracy were achieved by combining hybrid machine learning with Pareto-optimal solutions for a wide variety of information, such as standard performance and feature sets from a variety of growing domains [9–13]. The methodologies employed in numerous research projects were beneficial among the diverse assessment criteria in information technology, computational science, and cloud-based services. In [10] the K-means segmentation method is primarily used to examine Delhi's polluted air and determine the source of the substances that may pollute the atmosphere. Ashok Vihar, R.K. Puramand, and Punjabi Bagh are one of the most contaminated areas, according to the researchers. In [11] a technique for analyzing water quality using algorithms, such as Random Forest as well as multi-label classifiers has been developed. Multiclass classifiers were also shown to be greater than the corresponding forests by the researchers and in [12] a carbon emissions assessment approach for Bengaluru has been suggested. For the examination of air contaminants, the author used the ZeroR method. In addition, the writer shows how impurities are linked and interdependent. In [13–15] a new methodology for multimodal categorization of PM 10 levels has been presented to classify PM 10 concentrations where the research employs Back propagation classifiers and Random Forest classifiers. Randomized tree classification is also defended by the researcher. In [16,17] a classification algorithm is used where a way to forecast air pollution levels has been given. SVM, Logistic Regression, and Support Vector machines are some of the algorithms utilized by an author to solve a problem. Neural networks are more precise and reliable, according to the researchers.

Recently the authors in [14] came up with a method for predicting pollutant concentrations. To obtain accurate predictions, the author had been using a hybrid strategy that blended the stochastic optimization procedure with something, such as a random forest classifier. A study [7] offered a synthetic-based approach where methane gas and oxides are predicted using a quadratic regression model. A variety of parameters, including velocity, air flow, heat, moisture, and dangerous constituents from construction plants including Skikda, were also studied. Their model was assessed using RMSE and MAE but only NO and CO should be included in this technique. Nitrogen oxide PM 2.5 and PM 10 will not be included in this method. In [18] Vehicle emissions forecasts were made in Spain using an SVM-based logistic regression that included the most important inductive reasoning in order to provide a good prediction of the main pollutants. The major findings of the existing methods [1–18] are that many different forms of pollution that are caused due to human activities are continuously monitored using several techniques, but as per the technological aspect, it is not possible to stop the spread of polluting contents. However, the presence of several chemicals in the atmosphere can be reduced by preventing the burning ratio of fossil fuels and other residues that introduces pollution to the surroundings. To prevent the abovementioned fuels, it is necessary to implement an automatic monitoring system that takes immediate action against the burning of fossil fuels.

1.2. Objectives of the Proposed Method

The existing models [1–18] are used for checking the amount of fossil fuels in the atmosphere where each method has its advantages and disadvantages, such as choosing the best optimization algorithm for reducing the amount of pollutants, selecting the correct automobile for reducing the amount of air quality, etc. However, the goal of this study is to ensure the efficiency of various approaches and pick the right one for vehicle emissions predictions. In addition, the estimation of carbon emissions by choosing the best predictive model and improving it, and finding the best refining process for carbon emissions and weather data for prediction and determination are the most considered factors for preventing the amount of pollutants. For the abovementioned objective, learning algorithms are incorporated to improve their performance by gaining knowledge from past experiences, improvising, and adapting to new circumstances. Machine learning methods may be used to construct accurate pollution forecasts.

2. Air Pollution Management System: System Model

Contaminants can be eliminated from atmospheric gases by using equipment, changing the commodities used in air quality operations, or changing operating practices to reduce environmental pollution where all the above-mentioned procedures are termed as monitoring approaches. They are still the cost components in the business since they have price costs connected with themselves. Different goods or procedures that deliver that very same benefit to society while emitting less pollution are almost always available. Products and services like this will have their distinct optimization model [1]. Monitoring or failing to manage the number of pollutants in the environment has consequences. A cash value may be assigned to all of the negative impacts of fossil fuels on the public, including harm to plants, materials, buildings, wildlife, the environment, and people's health. Destruction capabilities are the technical terms for these expenses. For as much as we know about the link between costs and damages, we can assess the purchase price of control techniques and tactics [7]. They are still business cost components because they have a pricing tag associated with them. Almost often, different commodities or techniques that provide the same benefit to society while releasing less pollution are accessible. This type of product or service will have its optimization model [1]. We should choose the most premium control options when we can accomplish these objectives using a variety of control options.

It is necessary to employ air quality to evaluate whether contaminants inside this air are consistent with desirable levels of economic security. It is hard to fathom that any authority would tolerate environmental contamination that is recognized as harmful to health by the government. In any case, it is a matter of personal opinion as to what constituted "harm" to one's health. The subject of what is reasonable in terms of harm to one's health is much more contentious [8]. When it comes to deciding on period control mechanisms, the same democratic factors apply as they did in the subject of episodes control scheme. Health-related harm may be tolerated regardless of relative costs, but general well-being cannot be without economic feasibility. Some countries may choose an emission threshold that permits some phytotoxicity, creatures, minerals, buildings, and the environment, as long even if they are confident that our inhabitants' safety will not be harmed. It is termed a vehicle emissions benchmark if the intensity is chosen by the authority. This is the standard that the government claims to want to keep [3]. In analytical terms, the periodic representation of different PM levels can be determined [4] using Equation (1) as follows,

$$l_i(2.5) = \sum_{i=1}^{n} (mat_2 + \vartheta_2) \times log(mat_1 + \vartheta_1) \times I_{in} \tag{1}$$

where,

$mat_2 + \vartheta_2$ denotes the summation of the second matrix representation and weight produced in the same matrix

$mat_1 + \vartheta_1$ represents the logarithmic values of the first matrix representation and weight produced in the same matrix

I_{in} indicates the normalized values of biological proportions that are present in the air

Equation (1) denotes the PM value if the level of pollution exceeds the level of PM 2.5. Whereas for other cases, the level of indications [3] is represented in Equation (2) as follows.

$$l_i(10) = \sum_{i=1}^{10} (mat_{10} + \vartheta_{10}) \times log(mat_1 + \vartheta_1) \times I_{10} \tag{2}$$

where,

$mat_{10} + \vartheta_{10}$ denotes the summation of the tenth matrix representation and weight produced in the same matrix

I_{10} indicates the maximized normalized values

At the output state, the normalized values must be converted to the original represented values; therefore, there is a need to define the maximum and minimum limits [4] which are represented using Equation (3) as follows,

$$O_i(2.5) = \sum_{i=1}^{n} \frac{l_i(2.5+1)+178}{a_i} \tag{3}$$

where, a_i represents the average value of different biological elements that are present in the air.

In Equation (3) four different elements are considered and the value of 178 indicates that normalized values are averaged for a period of 178 delay timings. Similarly, the original values of PM 10 can be formulated as given in Equation (4).

$$O_i(10) = \sum_{i=1}^{10} \frac{l_i(10+1)+402}{10} \tag{4}$$

The original values in Equation (4) denote a delay of 402 s with an average of 10 different biological elements that are present in the entire system. Since the maximum limits are measured from historical data it is necessary to denote a regularization parameter that controls high variations in the PM parameters. Therefore, the minimization of the regularization parameter [2] can be represented in Equation (5) as follows,

$$Gen_i = min \sum_{i=1}^{n} (P_i(\vartheta_i, C_i))^2 \times r_i \tag{5}$$

where,

P_i indicates the number of parameters

ϑ_i, C_i represents the number of weights and concentration levels in the environment

r_i denotes the number of regularization parameters

Consequently, the monitoring parameters depend on the number of nodes that are used in the connection pathway, where they minimize the cost of implementation [3] as represented in Equation (6).

$$cost_i = min \sum_{i=1}^{n} nodes_i \times roots_{in} + P_i \tag{6}$$

The amount of pollutants that are present in the air depends on the strength which is represented in the three-dimensional form [2] as follows,

$$PA_i(a,b,c) = \sum_{i=1}^{n} e^{\frac{-b^2}{\sigma^2}} \tag{7}$$

where, σ^2 denotes the standard deviation of three co-ordinate axis pollutants.

Figure 1 deliberates the model of an air pollution system that consists of several blocks that are used for selecting the criteria of occurrence in the atmosphere. Additionally, in the schematic representation, all different requirements, such as quality, materials, type of pollutant, aspects of speed, and mobility are interconnected with each other. In addition, if any one aspect is affected then the constraint will not be satisfied, thus, new material must be transformed with targeted pollutants.

Process for Estimating Air Quality

Furthermore, as seen in Figure 2, fast urbanization in neighbouring regions has made it more difficult to wash filthy air, resulting in an even greater concentration of pollutants inside the municipality. The municipality has warm summers moderated by the rainy season, with an average precipitation of 700 mm, the majority of which falls during the city's extended rainy season. Pollutant data from several known air quality measurement stations were taken into account while performing this investigation [9]. They are situated in even more polluting areas of the city. Among the other reasons for selecting these facilities was to highlight the complexities and variation in environmental predictions [10]. CO, NO_2, SO_2, O_3, PM 2.5, and PM 10 polluting amounts were gathered from either the Central Pollution Control Board (CPCB) site and an "Industrial emissions Environmental Tracking Systems" that was created to collect impurities' percentages [11]. A Wi-Fi module was used to transport cloud data, while an SD card was used to store files and documents immediately. Thing Speak's IoT network stored the data remotely, where this could be accessed by anybody.

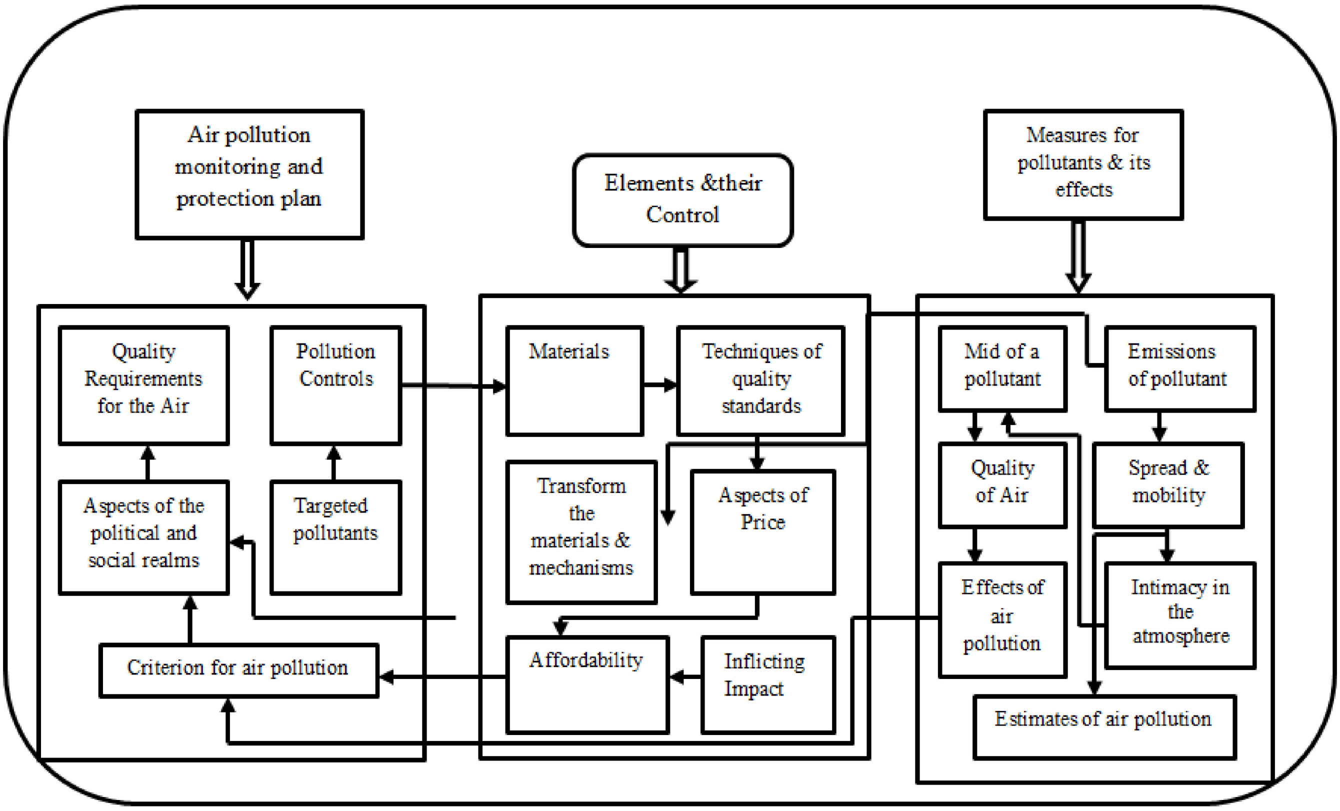

Air pollution monitoring and protection plan
Elements &their Control
Measures for pollutants & its effects
Quality Requirements for the Air
Pollution Controls
Aspects of the political and social realms
Targeted pollutants
Criterion for air pollution
Materials
Techniques of quality standards
Transform the materials & mechanisms
Aspects of Price
Affordability
Inflicting Impact
Mid of a pollutant
Emissions of pollutant
Quality of Air
Spread & mobility
Effects of air pollution
Intimacy in the atmosphere
Estimates of air pollution

Figure 1. Model of the air pollution management system.

Temperature, velocity vector, high humidity, air velocity, and other such influential parameters were indeed gleaned from the aforementioned sources [12]. Dependent variables were removed from the research design before they could be used for analysis. Options, such as pollutants, are approximated by utilizing an imputer program to estimate the null values; the normal distribution estimate is applied in this case. All characteristics are converted for ease of calculation just before the input is homogenized [13]. As a result, the degree-based performance parameters for wind conditions have been transformed into a wind speed index. To ensure that all qualities have greater validity, it is easy to boost the input's properties to a certain range.

An essential quality cannot be overshadowed by a less essential one that has a wider range of values [19–31]. Predictive output data may be better predicted by narrowing down the original collection of attributes to those that are most useful. Image enhancement is used when there is additional information [14]. To extract features from a collection, the best input variables must be picked from the image database. For subsequent investigation, the compressed information is referred to. There are a total of five inputs that may be analyzed, thus all of the variables are used in the calculations.

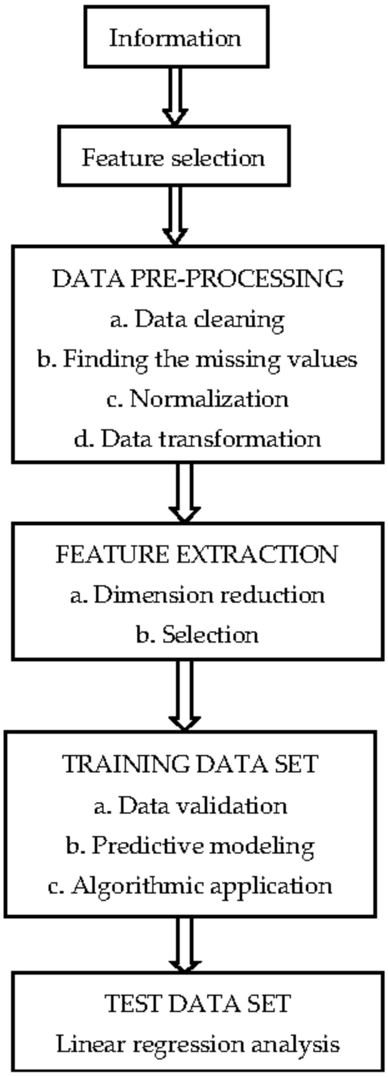


Figure 2. Process for estimating air quality.

3. Optimization Using Machine Learning Algorithm

Leaders are supported by some of these structures because they provide criteria for evaluating possibilities or for justifying their decisions. Something between action with several perpetrators must be simple, fast, and efficient. In Figure 3, the uppermost layer may be seen. New metrics for assisting customers in ecological fundamental administration are now available thanks to the growing advancement in Artificial Intelligence techniques, notably those involved in Information Architecture [15]. According to several natural paradigms, algorithms and quantitative measurements are inadequate. Such structures need the use of a variety of disparate factors, to accurately predict their behaviour. These vulnerabilities may be reduced over time by using various problem-solving methods (such as Circumstance Arguments and Commitment Gratification). This situation is often described as having an unorganized environment in Machine Learning [16]. There is a lack of knowledge among experts about the linkages between the concepts or attributes of the region. The program's linkages between these marvels are poorly understood. When faced with a choice between a plethora of possible solutions, an ever-evolving picture of the environment and wildlife emerges. Because of this, the ML (Machine Learning) approach is capable of being learned without any difficulty, although the full-time capacity may be poor [17].

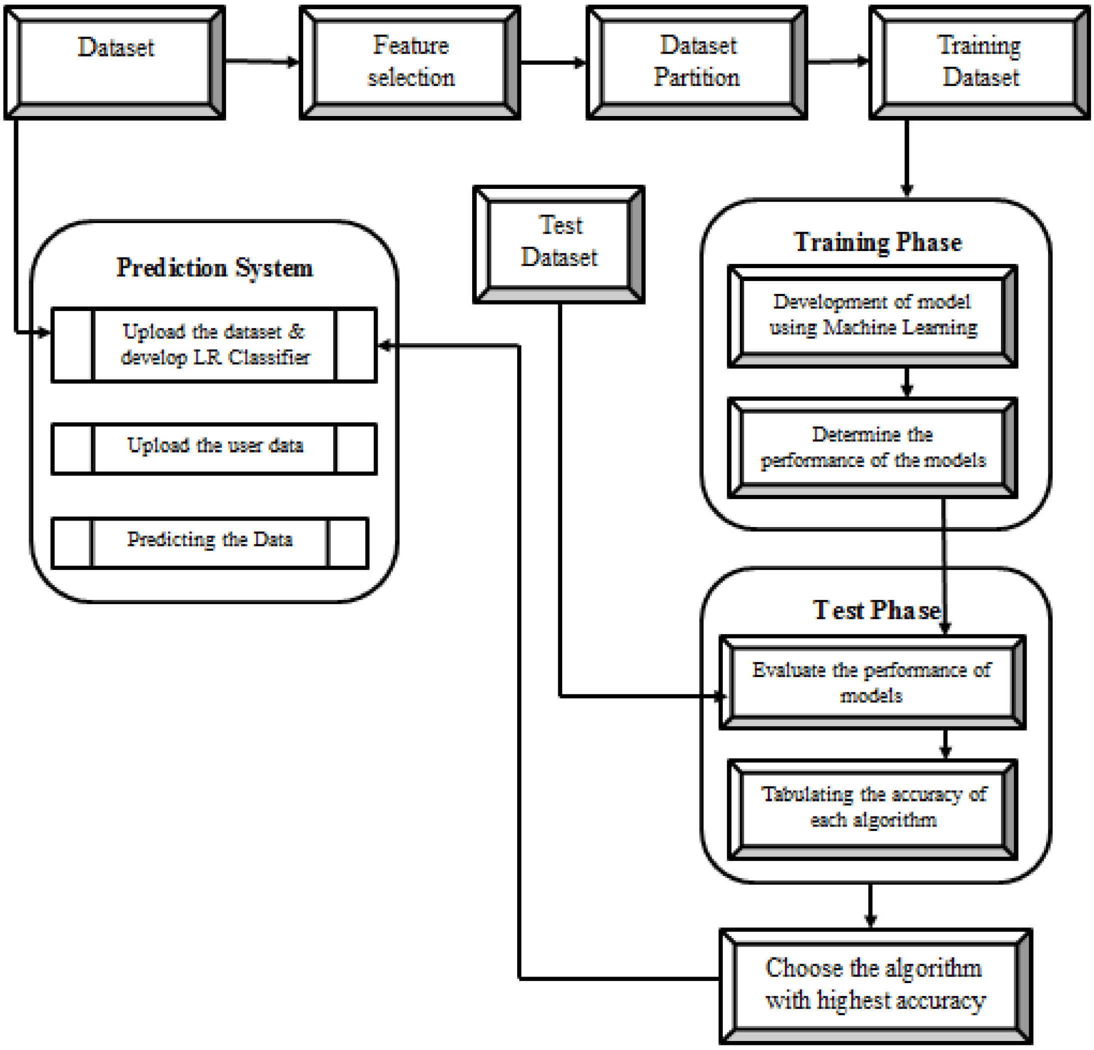

Dataset
Feature selection
Dataset Partition
Training Dataset
Test Dataset
Training Phase
Development of model using Machine Learning
Determine the performance of the models
Prediction System
Upload the dataset & develop LR Classifier
Upload the user data
Predicting the Data
Test Phase
Evaluate the performance of models
Tabulating the accuracy of each algorithm
Choose the algorithm with highest accuracy

Figure 3. Forecasting and monitoring of air pollution based on machine learning techniques.

The machine weights must be adjusted if the quantity of the effort components in the training examples varies considerably. Barbells generated by the training method will have a wide range of magnitudes. The issue can be solved with input data cleaning. Formalized knowledge, for contrast, has the means worth deleted, at around that point split by the error margin, resulting in components with a Gaussian distribution and unit statistical significance in this study The daytime cycle does not need to be removed from the data since separate authors have suggested accounting for the different hours of this week [17]. To eliminate irrational reduced sensations and ensure fair pacing of alterations in estimates, we used criteria, such as controlling all nearby areas with the least preoccupation, the most intense attention, and the fastest tempo possible. For the most part, this study is the first to look at the use of continuous learning to improve the accuracy of filling out a form, and it aims to do so by selecting the best method for predicting air pollution [18]. Many studies have also not examined the differentiating evidence of valid factors in exhaust prediction relying on a conceptual framework, which is the focus of this investigation. Choose and produce the best factual portrayal for the anticipation of air pollution; modify air pollution and weather using the best diagnostic and therapeutic options to almost predict overlooked material while also funneling its uproar. The most important factor in determining air pollution expectations is shown in Figure 4.

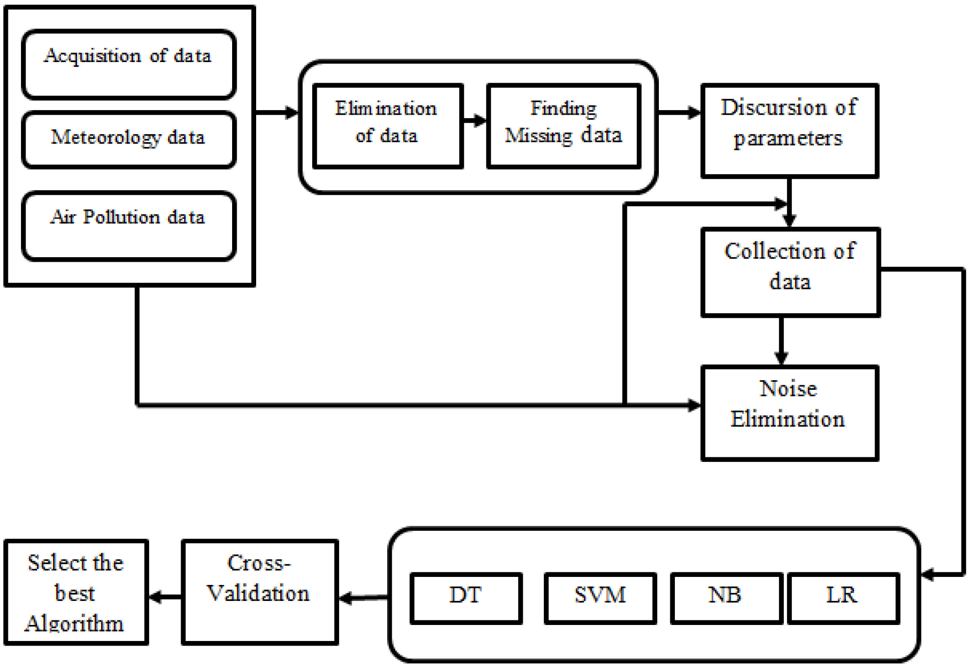


Figure 4. Air pollution model and quality predicting model.

4. Methodology

Sensors measure air contaminants that were then analyzed according to a standardized methodology and saved as a collection. Several preprocessing functions have been applied to this data collection, including standardization, classification techniques, and finite difference. Training and test datasets are created again when the database is ready [19]. The trained model is then subjected to the further Classification Techniques. Analyzing the findings requires comparing them to the validation set. The suggested model's design is shown in Figure 5. Four data mining techniques, such as LR, SVM, DT, and NB, are taken into account when predicting air quality utilizing the Unsupervised Data Mining technique.

4.1. Decision Trees

It is well known that the classification tree controller [24] belongs to the computer vision class. Figure 6 shows labeled data evaluations and leaves indicate categorizing decisions in the trees shown in this section (classes) [20]. Normally, a pessimistic computational intelligence strategy is used to generate a small predictive model again from the testing data by constant segmentation based on predicted algorithms. This method correctly categorizes the empirical values. Pruning is a current system functionality that removes the problem of well over. The C4.5 technique [23] was implemented in Weka [24]. Careful thought trees are distinguished by their simple representational structure, which might be described as a set of regulations.

4.2. Support Vector Machine

In mathematics, an SVM is referred to as a heavier model encoder. Classification and discrimination in SVM are accomplished by the use of different circuits or graphics. The separation of students based on academic standing [25]. How much SVM is doing is exactly what it says it does. It uses lines to designate categories, similar to kernel multiple regression. Using SVMs, you may classify data using kernels. An additional feature of the SVM is that it has a primary aim of achieving the best prediction performance. As they do not fit within a Bayesian framework, we will briefly examine SVMs here [26]. If Kotler, Keller, and places of great importance used SVM and obtained positive results, mentioning those places may help us. The two classes are often denoted by the labels +1 and −1 in the SVM investigation. wTx + b >= 0 class +1, 0 class −1 for a dimensional feature space described by parameters seems to be a straight difference. Historically, SVM has been used since the pattern may not be that huge while it is being developed.

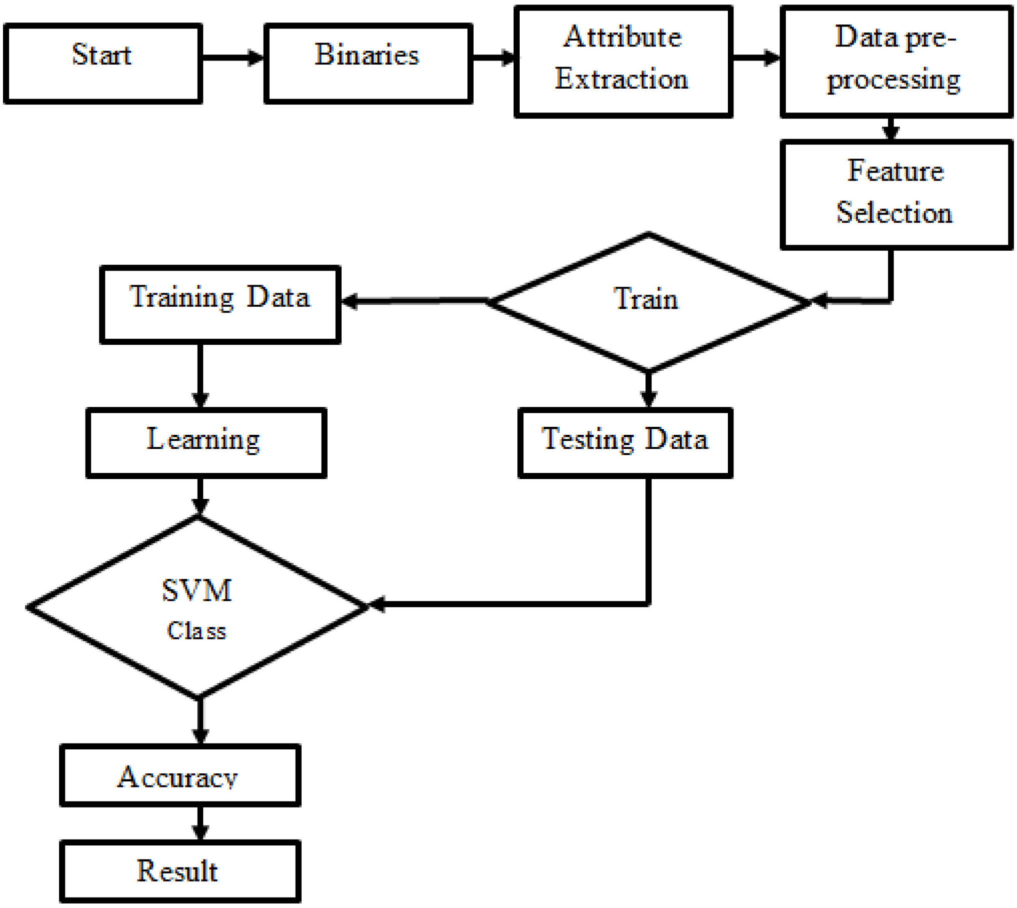


Figure 5. Flow diagram for malware detection and classification using artificial intelligence techniques.

4.3. Naïve Bayes

Predicted mostly using the naïve Bayes method, which states that the approximate solution of something, such as classification, is proportional to its maximal prevalence and even with the modified possibilities of the features supplied in this classification environment, the Naïve Bayesian framework has been developed. If there are no concerns about the independence of the parts, the predictive performance should always be computed using a Bayes technique [27]. Naive Bayes makes this approach more efficient by assuming that property is statistically independent having respect for some characteristics and needing mostly a nonlinear lot of factors to be estimated. Any current class prior probability can be easily derived from either classifier and these percentages should be used to determine the classifier's subsequent correctness considering a set of characteristics. According to research, Naive Bayes can accurately classify data in a broad range of fields [25].

4.4. Linear Regression Model

Regressions and statistical modeling are two of the most used methodologies. A linear methodology and formulation among two factors are related to them. Multivariate linear regression, or MLR for short, is a statistical technique that uses a large number of factors to predict the outcome of the same need [28]. Multiple linear regression models are used to explain the linear connections between independent (cause) and able-to-respond (response) elements (MLRs). Therefore, the formula for the linear regression model [11] can be formulated as given in Equation (8).

$$c_i = \alpha_0 + \alpha_1 z_{i1} + \alpha_2 z_{i2} + \ldots + \alpha_y z_{ip} + \tau \tag{8}$$

The engine is hardly any or non-existent: the motor is the replication of the postponement of its operation. The relationship between the variables is not considered in regression models since it is presumed to be absent [29]. Dependent Remnant mistakes arise whenever there is an excessive amount of interdependence between the predictor variables.

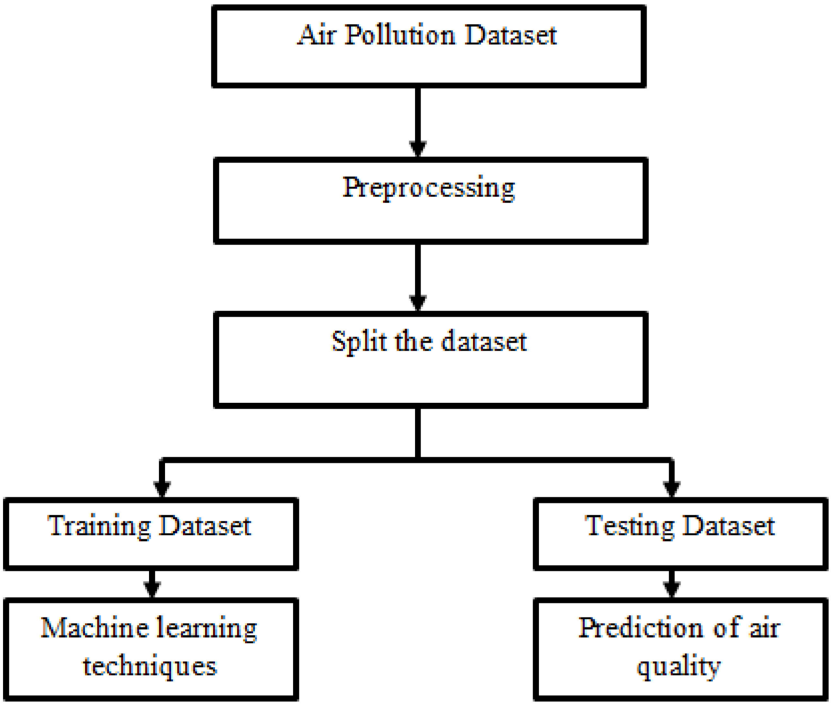


Figure 6. Flow diagram of air quality using ML techniques.

It may be expressed as Yp = Xi(a + b), where Yp is the expected variable, Xi denotes the parameter, a denotes the slopes, and b denotes the interception. The mistake E is as follows:

$$E = \sum_{k=1}^{m} (AP - PO)^2 \tag{9}$$

In this case, AP is the proportional gain, PO is the expected activity, and the rectangle of the discrepancy between performance and anticipated output is called the variance square [4]. The workflow of LR, which is employed in the calculation of AQI, is shown in Figure 6. When we conduct our studies, we use certain parameters to determine whether or not the multivariate regression that was employed was successful, and whether or not there are any probable connections between the coefficient of determination (R^2 that exists). Here, n describes the number of observations examined, and the amounts estimated and measured are denoted by R_t and R'_t. The mathematical expression of the presume for the diagonal of predicted and real value disparities are calculated [14] and use the root mean square error formula.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} |R_t - R'_t|^2} \tag{10}$$

It keeps track of the differences between two successive time series [10], regardless of whether or not their recommendations are taken into consideration [30–38]. Specifically, it demonstrates that the standard and ultimate differences between a prediction and actual gathered data are the same rolling average [11] and the relative disparities in the equations for the MAE computation are represented in Equation (11).

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|R_t - R_t'\right| \tag{11}$$

It is one of the simplest and easiest metrics used in the regression [3]. It is defined as the sum of the squares of the difference between the actual value and the predicted value or it is the average squared errors of the prediction made. It is given by the Equation (12).

$$\text{MSE} = \frac{1}{n} + \sum_{t=1}^{n}\left(R_t - R_t'\right) \tag{12}$$

According to the significance level, the percentage of all fluctuations of the dependence variance is explained by the additional factor via the predictive connection and is expressed as 1. Generally speaking, the closest this same value of R^2 gets to one, the greater the independent variable's ability to justify the regression model growth. Take a look at the following calculating methodology [4],

$$R^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{13}$$

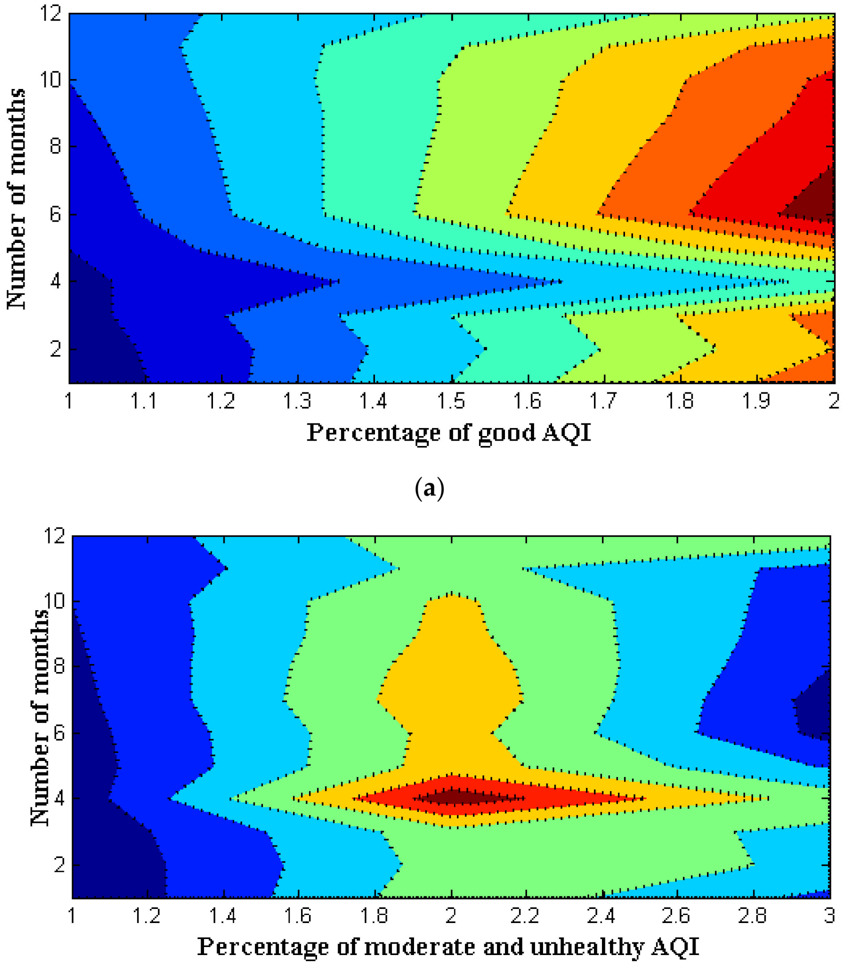
5. Results and Discussion

Testing the supplied equations with data from the information used only to anticipate the Air Pollution Levels during the next few hours is required, as illustrated in Figure 7. As may be seen in Tables 1 and 2, the expected values have been provided. When looking at the statistics, the intermediate AQI number has to be the most common in any single period (Figure 7a). Because harmful pollution happens more often from December to April, it is reasonable to assume that the greatest emissions occur during the cold months. Year-based groupings (Figure 7b) indicate an overall decrease in air quality from 2015 to 2021, with a little increase in pollutant levels in 2016. The intermediate class accounted for 50% of all instances, although the desirable and harmful classes answer approximately 30% and 20% of all episodes, respectively.

Table 1. Month-based of AQI.

Number of Months	Good (%)	Moderate (%)	Unhealthy (%)
1	39	37	17
2	35	34	29
3	37	36	28
4	21	66	35
5	35	45	19
6	48	44	7
7	46	48	6
8	44	46	10
9	42	43	13
10	41	42	14
11	38	33	17
12	29	37	37

RMSE, average error percentage, mean exponential error, and R^2 are just the productivity statistics that were utilized throughout this work to assess the performance of various algorithms. R^2 stands for mean square root. It is indeed a common strategy for determining the accuracy of a model's prognosis when dealing with empirical information. The following are the photographer's effectiveness values, as shown in Table 3.



(**a**)

(**b**)

Figure 7. Month-based AQI: (**a**) Good state (**b**) Moderate and unhealthy state. The blue color map in Figure 7 represents the good state whereas the yellow and red indicates the moderate and unhealthy states of AQI values.

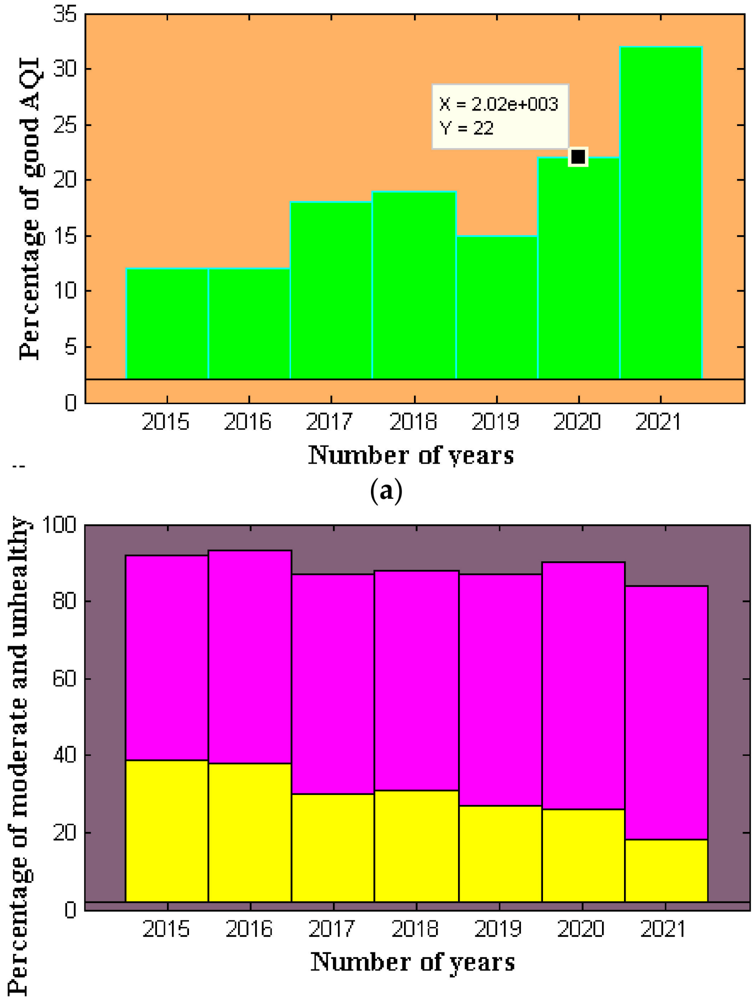
Table 2. Year-based of AQI.

Year	Good (%)	Moderate (%)	Unhealthy (%)
2015	12	53	39
2016	12	55	38
2017	18	57	30
2018	19	57	31
2019	15	60	27
2020	22	64	26
2021	32	66	18

Table 3. Performance metrics values for Machine learning models.

Model	RMSE	MAE	MSE	R^2
SVM	19.892	17.982	16.923	0.990
DT	09.563	11.971	8.912	0.992
NB	08.157	10.307	7.135	0.953
LR	03.116	05.125	4.123	0.923

To forecast the air quality index (AQI), this article employs a variety of ML Algorithms, containing algorithms, such as Regression Analysis, the SVM, the DT blueprint, and thus the NB framework. By reviewing the outcomes of all versions' quality metrics, it is possible to determine that the LR hybrid learning has the minimum output values, as seen in Figure 8. As a result, this methodology has been adopted to anticipate the Air Quality Status for the region during the next 5 min.



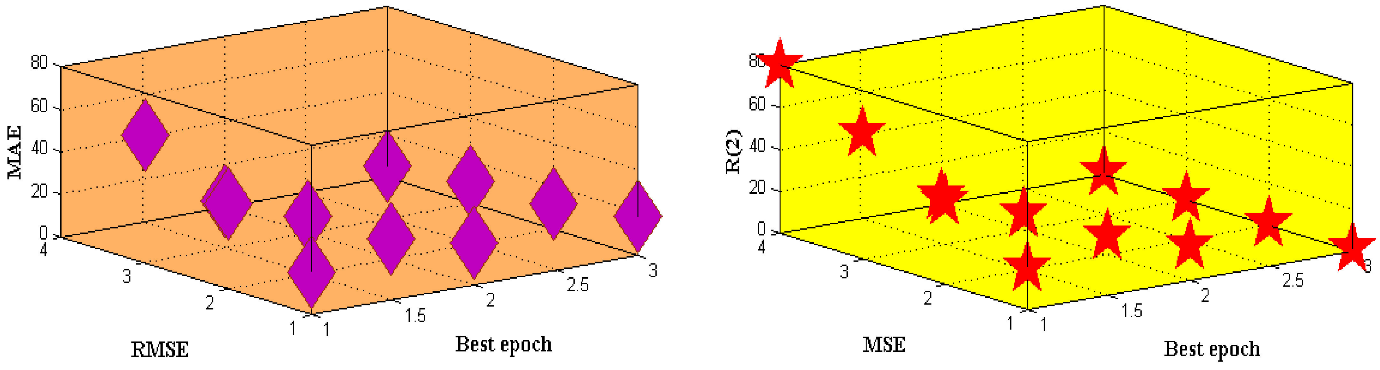
(**a**)

(**b**)

Figure 8. Year-based AQI (**a**) Good state. The green color indications proves that only good AQI values are achieved throughout the years. (**b**) Moderate and unhealthy state. The yellow and pink color indications proves that only moderate and unhealthy AQI values are achieved throughout the years.

To more naturally assess the predicted effectiveness of the Coefficient Of determination, and the SVM Classification framework, DT and NB become full, and concentrations of different air contaminants acquired via predicting the future were picked for assessment. As illustrated in Figure 9, the competence of each simulation is indeed assessed by other assessment criteria: MSE, RMSE, MAE, and R^2, which are all derived from the mean square error.

According to Tables 4–6 which are simulated in Figures 10–12 when predicting the concentrations of each component, the LR figure's mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and R^2 are also the smallest whenever they are compared to some other three techniques. That is, the LR system can produce the least overall error in predicting, while also exhibiting the best reference value.



(**a**) (**b**)

Figure 9. Performance metrics values for Machine learning models (**a**) The diamond marks indicate the RMSE and MAE values for iteration numbers in the entire marked area. (**b**) The star marks indicate appropriate MSE and R^2 values for proper iteration values.

Table 4. Error comparison for PM 10.

Pollutant		**PM 10**		
Parameter	**RMSE**	**MSE**	**MAE**	R^2
SVM	0.3232	0.35345	0.3390	0.6564
DT	0.30765	0.3186	0.4254	0.6381
NB	0.41 42	0.4014	0.40 15	0.6152
LR	0.2145	0.2345	0.2041	0.5412

Table 5. Error comparison for PM 2.5.

Pollutant		**PM 2.5**		
Parameter	**RMSE**	**MSE**	**MAE**	R^2
SVM	0.4239	0.35345	0.3390	0.6564
DT	0.4187	0.3186	0.4254	0.6381
NB	0.4253	0.4014	0.4015	0.6152
LR	0.3256	0.2345	0.2041	0.4212

Table 6. Error comparison for $O_3/NO_2/CO/SO_2$.

Pollutant		$\mathbf{O_3/NO_2/CO/SO_2}$		
Parameter	**RMSE**	**MSE**	**MAE**	R^2
SVM	0.5348	0.4584	0.4489	0.6854
DT	0.5279	0.4696	0.5345	0.6785
NB	0.5364	0.5125	0.5268	0.6584
LR	0.43668	0.1315	0.1045	0.3212

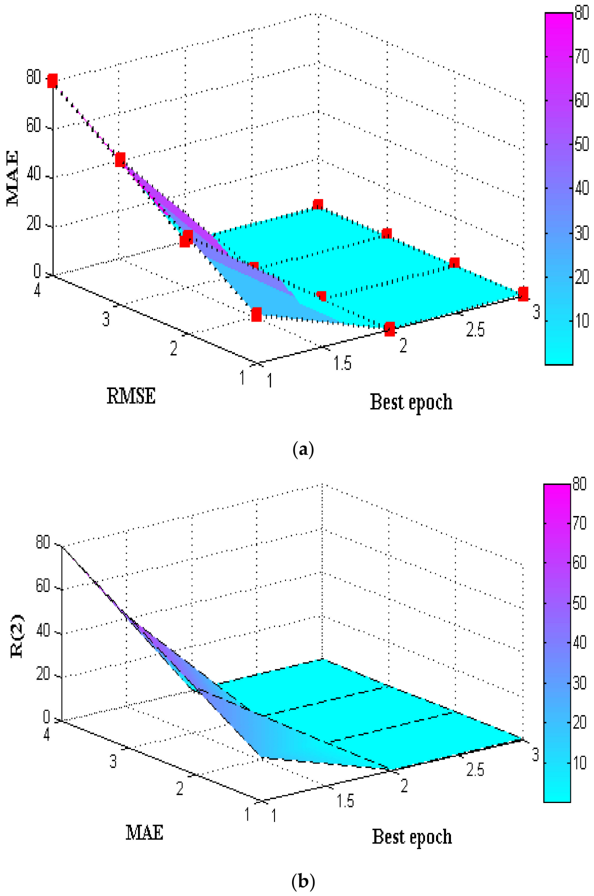


Figure 10. Comparison of errors (**a**) RMSE vs. MSE (**b**) MAE vs. R^2.

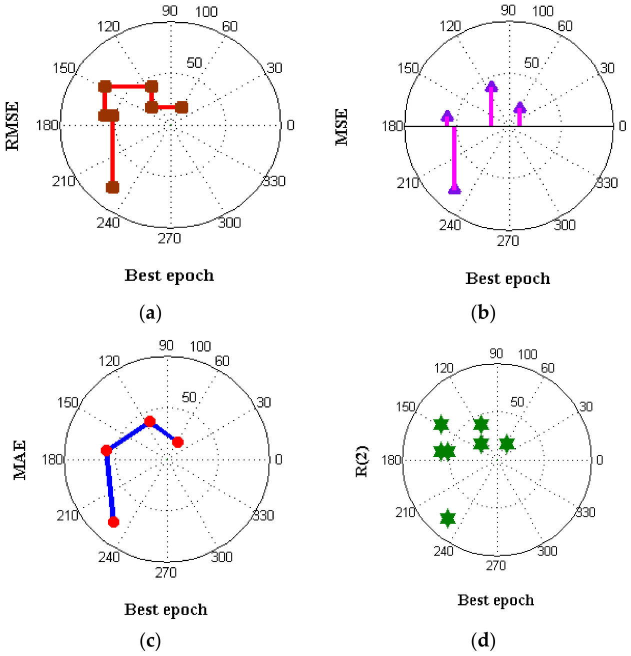
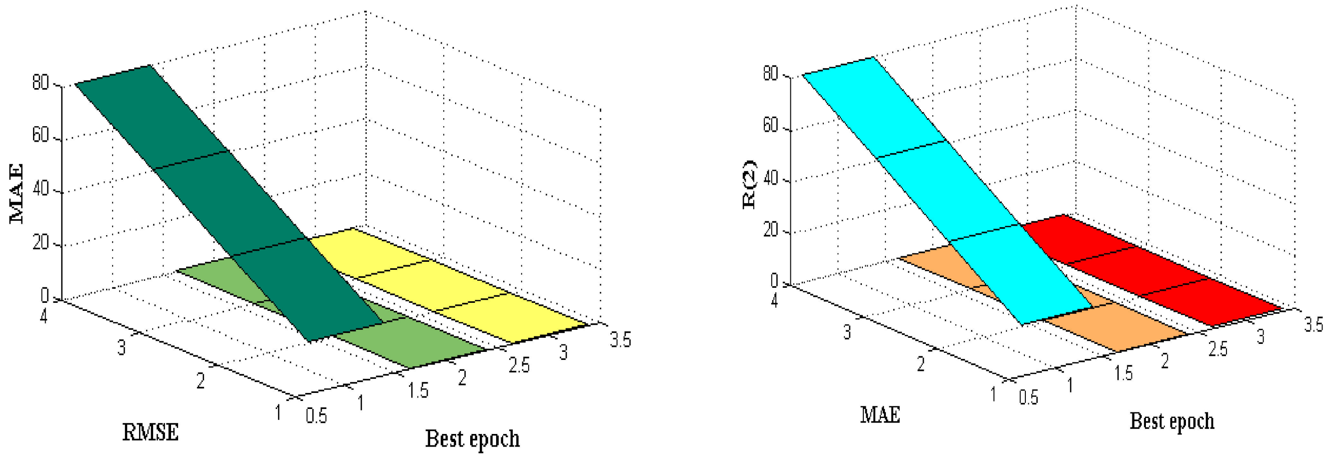


Figure 11. Comparison of errors using PM 2.5 (**a**) RMSE (**b**) MSE (**c**) MAE. The red dots that connect the blue line indicates that with respect to best epoch values the error values are reduced. (**d**) R^2. The star marks indicate the R^2 values achieved for corresponding iteration.



(**a**) (**b**)

Figure 12. Comparison of error using $O_3/NO_2/CO/SO_2$ (**a**) RMSE vs. MSE. The green and yellow color plots indicate that for every best epoch values the RMSE and MAE values are plotted for existing and proposed method. (**b**) MAE vs. R^2. The blue, brown and red color plots indicate that for every best epoch values the MAE and R^2 values are plotted for existing and proposed method.

6. Conclusions

In the air quality described in the proposed method, different types of air pollution circumstances are observed and an indication is provided to humans for choosing suitable conditions in the absence of high solid particles. With the use of machine learning (ML) modeling techniques, the vehicle emissions quality assessment and forecasting concept developed in this suggested study effort provides a legitimate and reasonable solution to the multidisciplinary nature of air quality (AQ levels). The percentage of air contaminants in airflow is influenced by climatic characteristics, such as the velocity and airflow, the moisture content, and the surrounding air. Additionally, the comparison analysis has been made in the proposed method using an Ambient Air Quality (AQI) tool to find the real-time values. In addition, the AQI toolbox incorporated more data with multiple detection techniques, thus parallel data effectiveness on the percentage of air pollution in the atmosphere is measured for LR, SVM, DT, and RF. The comparative results prove that LR is much more effective compared to other methods as the detection of pollution in the atmosphere is highly accurate for about 62%.

Author Contributions: Data curation: S.A.A. and F.A.; Writing original draft: H.M., P.R.K. and S.S.; Supervision: H.M., P.R.K. and S.S.; Project administration: S.S. and P.R.K.; Conceptualization: H.M. and P.R.K.; Methodology: S.S. and H.M.; Validation: S.A.A. and F.A.; Visualization: S.A.A. and F.A.; Resources: S.S. and H.M.; Review and Editing: G.S. and J.C.-W.L.; Funding acquisition: G.S. and J.C.-W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is partially supported by the Western Norway University of Applied Sciences, Bergen, Norway.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Djebbri, N.; Rouainia, M. Artificial neural networks based air pollution monitoring in industrial sites. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–5.
2. Irfan, S.A.; Irshad, K.; Algahtani, A.; Azeem, B.; Tirth, V.; Algarni, S.; Islam, S.; Abdelmohimen, M.A.H. Machine learning-based modeling of thermoelectric materials and air-cooling system developed for a humid environment. *Mater. Express* **2021**, *11*, 153–165.

3. Verma, I.; Ahuja, R.; Meisheri, H.; Dey, L. Air Pollutant Severity Prediction Using Bi-Directional LSTM Network. In Proceedings of the 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, Chile, 3–6 December 2018; pp. 651–654.
4. Manoharan, H.; Selvarajan, S.; Yafoz, A.; Alterazi, H.A.; Uddin, M.; Chen, C.-L.; Wu, C.-M. Deep Conviction Systems for Biomedical Applications Using Intuiting Procedures With Cross Point Approach. *Front. Public Health* **2022**, *10*, 909628. [CrossRef] [PubMed]
5. Yang, R.; Yan, F.; Zhao, N. Urban air quality based on Bayesian network. In Proceedings of the 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN), Guangzhou, China, 6–8 May 2017; pp. 1003–1006.
6. Ayele, T.W.; Mehta, R. Air pollution monitoring and prediction using IoT. In Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 20–21 April 2018; pp. 1741–1745.
7. Shah, J.; Mishra, B. Analytical equations based prediction approach for PM2.5 using artificial neural network. *SN Appl. Sci.* **2020**, *2*, 1516. [CrossRef]
8. Gore, R.W.; Deshpande, D.S. An approach for classification of health risks based on air quality levels. In Proceedings of the 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), Aurangabad, India, 5–6 October 2017; pp. 58–61.
9. Selvarajan, S.; Manoharan, H.; Hasanin, T.; Alsini, R.; Uddin, M.; Shorfuzzaman, M.; Alsufyani, A. Biomedical Signals for Healthcare Using Hadoop Infrastructure with Artificial Intelligence and Fuzzy Logic Interpretation. *Appl. Sci.* **2022**, *12*, 5097. [CrossRef]
10. Paulose, B.; Sabitha, S.; Punhani, R.; Sahani, I. Identification of Regions and Probable Health Risks Due to Air Pollution Using K-Mean Clustering Techniques. In Proceedings of the 2018 4th International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, India, 9–10 February 2018; pp. 1–6.
11. Gore, R. Air Data Analysis for Predicting Health Risks. *IJCSN Int. J. Comput. Sci. Netw.* **2018**, *7*, 36–39.
12. Shitharth, S.; Meshram, P.; Kshirsagar, P.R.; Manoharan, H.; Tirth, V.; Sundramurthy, V.P. Impact of Big Data Analysis on Nanosensors for Applied Sciences using Neural Networks. *J. Nanomater.* **2021**, *2021*, 4927607. [CrossRef]
13. Kshirsagar, P.; Balakrishnan, N.; Yadav, A.D. Modelling of optimised neural network for classification and prediction of benchmark datasets. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2020**, *8*, 426–435. [CrossRef]
14. Raturi, R.; Prasad, J.R. Recognition of Future Air Quality Index Using Artificial Neural Network. *Int. Res. J. Eng. Technol. IRJET* **2018**, *5*, 3404–3407.
15. Rubal; Kumar, D. Evolving Differential evolution method with random forest for prediction of Air Pollution. *Procedia Comput. Sci.* **2018**, *132*, 824–833. [CrossRef]
16. Kaya, K.; Gunduz Oguducu, S. A Binary Classification Model for PM 10 Levels. In Proceedings of the 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, 20–23 September 2018; pp. 361–366.
17. Kshirsagar, P.; Akojwar, S. Optimization of BPNN parameters using PSO for EEG signals. In Proceedings of the International Conference on Communication and Signal Processing 2016 (ICCASP 2016), Lonere, India, 26–27 December 2016; Volume 137, pp. 385–394.
18. Liang, Y.C.; Maimury, Y.; Chen, A.H.L.; Juarez, J.R.C. Machine learning-based prediction of air quality. *Appl. Sci.* **2020**, *10*, 9151. [CrossRef]
19. Suárez Sánchez, A.; García Nieto, P.J.; Riesgo Fernández, P.; del Coz Díaz, J.J.; Iglesias-Rodríguez, F.J. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). *Math. Comput. Model.* **2011**, *54*, 1453–1466. [CrossRef]
20. Kang, G.K.; Gao, J.Z.; Chiao, S.; Lu, S.; Xie, G. Air Quality Prediction: Big Data and Machine Learning Approaches. *Int. J. Environ. Sci. Dev.* **2018**, *9*, 8–16. [CrossRef]
21. Sundaramurthy, S.; Saravanabhavan, C.; Kshirsagar, P. Prediction and Classification of Rheumatoid Arthritis using Ensemble Machine Learning Approaches. In Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 8–9 November 2020; pp. 17–21.
22. Yi, X.; Zhang, J.; Wang, Z.; Li, T.; Zheng, Y. Deep distributed fusion network for air quality prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 965–973.
23. Sun, W.; Sun, J. Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. *J. Environ. Manag.* **2017**, *188*, 144–152. [CrossRef] [PubMed]
24. Veljanovska, K.; Dimoski, A. Air Quality Index Prediction using Machine Learning Algorithms. *Int. J. Recent Technol. Eng.* **2019**, *8*, 7489–7492.
25. Teng, Y.; Huang, X.; Ye, S.; Li, Y. Prediction of particulate matter concentration in Chengdu based on improved differential evolution algorithm and BP neural network model. In Proceedings of the 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 20–22 April 2018; pp. 100–106.
26. Ge, S.; Wang, S.; Xu, Q.; Ho, T. Study on regional air quality impact from a chemical plant emergency shutdown. *Chemosphere* **2018**, *201*, 655–666. [CrossRef] [PubMed]
27. Kleine Deters, J.; Zalakeviciute, R.; Gonzalez, M.; Rybarczyk, Y. Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. *J. Electr. Comput. Eng.* **2017**, *2017*, 5106045. [CrossRef]

28. Akojwar, S.G.; Kshirsagar, P.R. A Novel Probabilistic-PSO Based Learning Algorithm for Optimization of Neural Networks for Benchmark Problems. *Wseas Trans. Electron.* **2016**, *7*, 79–84.
29. Tripathi, C.B.; Baredar, P.; Tripathi, L. Air pollution in Delhi: Biomass energy and suitable environmental policies are sustainable pathways for health safety. *Curr. Sci.* **2019**, *117*, 1153–1160. [CrossRef]
30. Liu, T.; Wu, T.; Wang, M.; Fu, M.; Kang, J.; Zhang, H. Recurrent Neural Networks based on LSTM for Predicting Geomagnetic Field. In Proceedings of the 2018 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES), Bali, Indonesia, 20–21 September 2018; Volume 5, pp. 56–60.
31. Chang, Y.S.; Lin, K.M.; Tsai, Y.T.; Zeng, Y.R.; Hung, C.X. Big data platform for air quality analysis and prediction. In Proceedings of the 2018 27th Wireless and Optical Communication Conference (WOCC), Hualien, Taiwan, 30 April 2018–1 May 2018; pp. 1–3.
32. Flores-Cortez, O.O.; Adalberto Cortez, R.; Rosa, V.I. A Low-cost IoT System for Environmental Pollution Monitoring in Developing Countries. In Proceedings of the 2019 MIXDES—26th International Conference "Mixed Design of Integrated Circuits and Systems", Rzeszow, Poland, 27–29 June 2019; pp. 386–389.
33. Montanaro, T.; Sergi, I.; Basile, M.; Mainetti, L.; Patrono, L. An IoT-Aware Solution to Support Governments in Air Pollution Monitoring Based on the Combination of Real-Time Data and Citizen Feedback. *Sensors* **2022**, *22*, 1000. [CrossRef]
34. Yang, R.; Hao, X.; Zhao, L.; Yin, L.; Liu, L.; Li, X.; Liu, Q. Design and implementation of a highly accurate spatiotemporal monitoring and early warning platform for air pollutants based on IPv6. *Sci. Rep.* **2022**, *12*, 4615. [CrossRef]
35. Kortoçi, P.; Motlagh, N.H.; Zaidan, M.A.; Fung, P.L.; Varjonen, S.; Rebeiro-Hargrave, A.; Niemi, J.V.; Nurmi, P.; Hussein, T.; Petäjä, T.; et al. Air pollution exposure monitoring using portable low-cost air quality sensors. *Smart Health* **2022**, *23*, 100241. [CrossRef]
36. Dhanalakshmi, M.; Radha, V. A Survey paper on Vehicles Emitting Air Quality and Prevention of Air Pollution by using IoT Along with Machine Learning Approaches. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 5950–5962.
37. Kumar Sai, K.B.; Mukherjee, S.; Parveen Sultana, H. Low Cost IoT Based Air Quality Monitoring Setup Using Arduino and MQ Series Sensors with Dataset Analysis. *Procedia Comput. Sci.* **2019**, *165*, 322–327. [CrossRef]
38. Ilieș, D.C.; Marcu, F.; Caciora, T.; Indrie, L.; Ilieș, A.; Albu, A.; Costea, M.; Burtă, L.; Baias, Ș.; Ilieș, M.; et al. Investigations of museum indoor microclimate and air quality. Case study from Romania. *Atmosphere* **2021**, *12*, 286. [CrossRef]