IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DETECTION AND PREDICTING AIR POLLUTION LEVEL IN A SPECIFIC CITY USINIG MACHINE LEARNING MODELS

Aryan Agarwal, Pratik Dighole, Abhishek Sabnis, Dhananjay Thosar, Madhuri Mane

Computer Engineering
Pune Institute of Computer Technology, Pune, India

Abstract: In contexts of smart cities, dealing with air pollution is a significant environmental challenge. Real-time monitoring of pollution data enables local authorities to analyze the current situation of the city and make decisions accordingly. However, a comparison of various strategies is needed to better understand how long they take to process different datasets. Existing research has used a variety of machine learning algorithms for pollution prediction. There are various regression techniques for this purpose and a comparative study to determine the best model for accurately predicting air quality with reference to data size and processing time. In this project, we have selected the algorithms with fewer errors and higher accuracy, and the study combines those algorithms and creates a sample algorithm using the combination of these algorithms which involves the Random Forest Algorithm, Support Vector Machine Algorithm, Linear Regression, Decision Tree, etc. Using this, prediction of the level of pollution will be efficiently done.

Index Terms **- Regression techniques, Air quality prediction, Accuracy, Machine learning models.**

I. INTRODUCTION

[5] Most urban areas are seeing increased concentrations of ground-level air pollution due to global economic and social development, particularly in quickly developing nations like India and China. Everyone can be harmed by exposure to air pollution, but those with heart or lung conditions, the elderly, and children are most vulnerable. Even at concentration levels substantially below the conventional anatomical mean limit value, studies suggest that long-term exposure to fine particle air pollution or air pollution from traffic is linked to mortality from environmental causes. Building an early warning system that delivers accurate forecasts and also informs local residents to health alarms will therefore provide important information to protect humans from air pollution.

Air pollution poses a significant problem for communities since it causes millions of people to die prematurely each year across the globe. [13] The creation of real-time intelligent applications and services, such as the minimizing of exposure to bad air quality either on an individual or city scale, would be made possible by the widespread deployment of air quality sensor devices and data processing for the resulting data.

This document is a model and instructions for air pollution which tells that energy consumption and its consequences are inevitable in modern age human activities.

The anthropogenic sources of air pollution include emissions from industrial plants; auto-mobiles; planes; burning of straw, coal, and kerosene; aerosol cans, etc. Various dangerous pollutants like CO, CO_2, Particulate Matter (PM), NO_2, SO_2, O_3, NH_3, Pb, etc. are being released into our environment every day. Chemicals and particles constituting air pollution affect the health of humans, animals, and even plants. Various models have been exercised in the literature to predict AQI, like statistical, deterministic, physical, and Machine Learning (ML) models.

II. Related Works

Machine learning provides an important data base for monitoring air pollution in industrial parks and may manage exhaust emissions in industrial production on a human basis in response to the new circumstances and new needs of international industrial modernization transition.

A number of machine learning ensemble methods are investigated for the goal of predicting the fine-grained air quality level in the near future. [12] The majority voting, averaging, weighted averaging, and 16 various stacking strategies are all included in the ensemble approaches. Comprehensive comparative tests are carried out to examine the performances of different ensemble approaches. Traditional Autoregressive Integrated Moving Average (ARIMA), the well-known deep learning model Long Short-Term Memory (LSTM), and nine of the base-level models, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and several boosting models, are included as contrast models. All of the models are trained and validated using datasets collected from the inland city of Beijing and the coastal city of Hong Kong.

A. *Different Types of Input Data*

Accurate pollution in air can be achieved through a machine-learning-based technique that uses environmental monitoring data and meteorological measures to reliably predict the air quality index. The objective is to create an air quality estimate framework that employs an unique non-linear autoregressive neural network with exogenous input model that is enhanced for time series prediction. The framework is used to a case study using various monitoring locations in various cities, and comparisons with other conventional machine-learning-based predictive algorithms demonstrate the viability and reliable performance of the suggested method for various types of urban environments.

Using the complex original data, three sub modal time series were created by decomposing them into empirical mode decomposition (EMD) with equal change rates based on the volatility and periodicity of each intrinsic mode function (IMF); Second, using a back propagation algorithm to predict simpler sub-modal time series data is better than using a gated recurrent unit (GRU) network to predict more complex sub-modal time series data for the two groups. Finally, the prediction results from three sub-series were combined to get the prediction data from the initial data.

III. Major Pollutants

The WHO has identified some cities that are on the verge of reaching a harmful threshold. Unfortunately, India is one of the nations with the greatest concentration of the world's most polluted cities.

A. *Carbon Monoxide (CO)*

CO is a odourless and colourless gas that, if inhaled in excessive quantities, can be dangerous. Any time something burns, CO is released. Trucks, cars and other machinery that consumes fossil fuels are the main sources of CO in the outdoor air. The quality of the air within your home can be impacted by a number of things, including gas stoves, enclosed space kerosene and gas space warmers, faulty furnaces and chimneys, and gas appliances.

B. *Carbon Dioxide (CO_2)*

You might not think of carbon dioxide, generally known as CO_2, as a typical air contaminant, as carbon dioxide doesn't lead to the typical haze that you might imagine, but it is when you consider of cities with high levels of air pollution, while having a significant impact on the greenhouse effect. Yet, studies suggest that, like more conventional types of air pollution, the consequences of global warming brought on by the release of carbon dioxide may have a direct influence on people's respiratory health.

C. *Particulate Matter (PM)*

Particulate matter or PM, often known as particle pollution, which is the word used to describe an airborne mixture of solid and liquid droplets. Dust, grime, soot, and smoke are examples of particles that are large enough or deep enough to be visible to the unaided eye.

1) *PM_{10}:* Airborne particles having a diameter of typically 10 micrometers or less.
2) *$PM_{2.5}$:* It refers to fine, inhalable particles with a typical diameter of 2.5 micrometers or less.

D. *Nitrogen Dioxide (NO_2)*

One of the strong oxidizing gases known as nitrogen oxides (NO_x) or oxides of nitrogen such as nitrogen dioxide (NO_2). Nitric acid and nitrous acid are two further nitrogen oxides. The signal for the bigger number of nitrogen oxides is NO_2, which is used.

The combustion of fossil fuels is the main source of NO_2 in the atmosphere. NO_2 is created as a result of emissions from automobiles, buses, trucks, power plants, and off-road vehicles.

E. *Sulphur Dioxide (SO_2)*

The national indoor air quality criteria for SO_2 set are intended to guard against exposure to all Sulphur oxides (SO_x). The most concerning component, SO_2, serves as a marker for the bigger number of atmospheric Sulphur oxides (SO_x). At far lower quantities than SO_2, the atmosphere contains other molecular SO_x (such SO_3).

People's exposure to all molecular SO_x should be reduced as a result of control actions that lower SO_2. A significant side benefit of this could be a decrease in the production of tiny sulphate particles and other particulate Sulphur pollution.

In most cases, emissions that result in high SO_2 concentrations also cause the creation of additional SO_x. The burning of fossil fuels in power plants and other industrial facilities is the main cause of SO_2 emissions.

F. *Ozone (O_3)*

Three oxygen atoms make up the gas known as ozone. Ozone exists both in the high atmosphere of the Earth and at ground level. Based on the location it's found; ozone can be either good or bad.

Good ozone, also known as stratospheric ozone, develops naturally in the earth's atmosphere where it creates a barrier that protects us from the sun's dangerous ultraviolet rays. A "hole in the ozone" is what is sometimes referred to as the result of man-made substances partially destroying this beneficial ozone.

Because of its negative effects on both individuals and the surroundings, ozone is a dangerous air pollutant at ground level and the primary component of "smog".

G. *Ammonia (NH_3)*

Ammonia is an odourless gas that becomes detectable at quantities greater than 50 ppm. The manufacture of fertilizer and the management of livestock manure account for the majority of the NH_3 emissions.

Inhaling large amounts of NH_3 can be dangerous, while smaller amounts can irritate the nose, throat and eyes. Additionally, the fine particulate matter is created in the atmosphere when it interacts with sulphates and nitrates ($PM_{2.5}$). It is well recognized that $PM_{2.5}$ is bad for the environment and people's health. Furthermore, NH_3 can help nitrify and eutrophize aquatic ecosystems.

H. *Lead (Pb)*

The sources of lead pollution differ from place to place. At the national level, processing of ores and metals as well as piston-engine aircraft using leaded aviation fuel are the main sources of lead in the air. Additional sources include power plants, utilities, and producers of lead-acid batteries. The areas closest to lead smelters typically have the greatest lead air concentrations.

IV. Different Techniques Applied for Air Pollution Detection

It has been demonstrated that the ML-based AQI prediction models are more dependable and consistent. Data collecting was made simple and accurate by modern technologies and sensors. Only ML algorithms are capable of handling the rigorous analysis needed to make accurate and trustworthy predictions from such vast environmental data.

A. *Machine Learning*

A subtype of artificial intelligence called machine learning uses historical data to teach machines how to perform certain tasks they were not specifically built to perform. While machine learning methods are well suited for tasks like picture classification, image segmentation, and object recognition, traditional programs lack the ability to comprehend images.

Machine learning is being widely used in a wide range of activities in many different industries of business, industry, and science due to its better prediction capacity.

1) *Supervised Learning:*

This kind of machine learning (ML) uses supervision, where computers are trained on labelled datasets and allowed to make predictions based on the training data. According to the labelled dataset, some input and output parameters have already been mapped. Thus, the input and related output are used to train the machine. At later stages, a tool is created to forecast the result using the test dataset.

Think about an input dataset that contains photos of various birds. The computer is initially trained to recognize the images, including the shape, colour, and size of the bird's eyes. After training, an image of a bird is used as input, and the computer is supposed to recognize the object and forecast the result. To arrive at a final forecast, the trained machine looks for the many characteristics of the object in the input image, such as colour, eyes, shape, etc. In supervised machine learning, object identification works like this.

2) *Unsupervised Learning:*

Unsupervised learning is a learning method where no supervision is provided. In this case, the machine has been taught using an unlabeled information and is given the ability to predict the results independently. Unsupervised learning algorithms attempt to classify the input's patterns, similarities and differences into groups that correspond to the unsorted dataset.

Take, for instance, a collection of input as various photos of same type, like fruits. The machine learning algorithm used in this case is unfamiliar with the photos. When we feed the data into the machine learning (ML) model, the model's job is to categorize the objects in the input photos based on their characteristics, such as their colour, form, or differences. After categorization, the machine predicts the result while being put to the test against a test dataset.

3) *Reinforcement Learning:*

The process of reinforcement learning is feedback-based. Here, the AI element works by scanning its environment using the hit-and-trial method and gains knowledge from mistakes, and enhances performance. Every wrong one is punished and right move is rewarded, for the component. So, the reinforcement learning component's goal is to maximize rewards through doing well.

B.Random Forest Algorithm

The Random Forest Algorithm, a well-known machine learning algorithm, belongs to the supervised learning approach. It can be used for ML problems involving both regression and classification. It was discovered using the concept of ensemble learning, which is the process of combining different classifiers to tackle a challenging problem and improve the performance of the model. As its name suggests, Random Forest is a classifier that averages several decision trees applied to various subsets of the supplied information to improve the predicted accuracy of the dataset. The random forest uses predictions out of each tree and predicts the result according to the votes of the majority of projections rather than relying solely on one decision tree.

1) *Working:*

First step is to choose any number of random samples from a provided data collection or training set.

Secondly, for each training set of data, this algorithm will build a decision tree.

In the next step, Voting will be conducted using an average of the decision tree.

Lastly, choose the prediction result that received the most votes as the final classification result.

2) *Ensemble:*

Ensemble is a term used to describe a collection of various models. Ensemble employs two techniques:

Bagging: Bagging is the process of generating a different training subset via replacement from a sample training dataset. Majority number of votes decide the outcome.

Boosting is the process of turning weak classifier into strong ones by building successive models with the goal of achieving the maximum accuracy possible. Like XG BOOST and ADA BOOST.

3) *Bagging:*

In random forest, bagging is sometimes referred to as Bootstrap Aggregation. Starting with any initial random data, the process begins. After rearranging, it is divided into Bootstrap Sample samples. Bootstrapping is the name for this procedure. Also, each model is trained separately, producing distinct outcomes known as Aggregation. The final stage combines all the findings, and the output that is produced is predicated on majority voting. The Bagging phase of the process makes use of an Ensemble Classifier.

C.Support Vector Machine

SVM is an algorithm that classifies the data points. Once a separator between the categories is found, the data are changed to allow the hyperplane representation of the separator, even if they are not linearly separable by mapping the data to a high-dimensional feature space. The properties of fresh data can then be utilized to determine which group a new data should correspond to and to suggest mitigation strategies.

1) *Classification:*

Complex mathematical issues are frequently resolved by SVMs. In contrast, smooth SVMs are favoured for data classification tasks, where smoothing techniques are utilized to lessen data outliers and highlight the pattern.

Because smooth SVMs can handle bigger datasets than traditional SVMs can, they are used to solve optimization issues that cannot be solved by conventional SVMs. Even with non-linear data, smooth SVM types frequently investigate mathematical aspects such strong convexity for easier data classification.

2) *Linear SVM:*

For categorizing data that can be linearly separated, a linear SVM is utilized. This indicates that a dataset is considered to be linearly distinct or separable when it can be divided into categories or classes with the use of a single straight line. Moreover, the classification algorithm that uses such data is known as a linear SVM classifier.

Problems involving classification and regression analysis are frequently solved with a straightforward SVM.

3) *Kernal SVM:*

A kernel or non-linear SVM is used to classify data that cannot be divided into multiple groups with the aid of a straight line. A non-linear classifier is what is being discussed here. Instead of relying just on 2D space, classification can be carried out with a non-linear data type by inserting features into higher dimensions. Here, freshly introduced features are in line with a hyperplane that makes it simple to divide classes or groups.

Multiple variable optimization problems are often handled using kernel SVMs.

4) *Working:*

Assume that we have red and black labels with the x and y attributes. With these tags, we want to be able to categories data into the red or black categories using a classifier.

Plotting the labelled data on an x-y plane will look like this:

A standard SVM uses the hyperplane, which in this case is a two-dimensional line, to divide these data points into red and black tags. The decision boundary line, along which data points fall into the red or black category, is indicated by the hyperplane.

A line known as a "hyperplane" is one that tends to increase the distance between two tags or labels that are near together (red and black). The maximum distance between the hyperplane and the closest label makes it simpler to classify the data.

For data that can be separated linearly, the situation above is applicable. A straightforward straight line, however, is unable to separate the many data points for non-linear data.

D. Linear Regression

One of the basic and most widely used method or algorithm for machine learning is linear regression. It is a statistical tool for executing predictive analysis. For numeric/continuous/real variables like salary, sales, product price, and age, among others, linear regression makes predictions.

The linear regression algorithm, often known as linear regression, demonstrates a linear relationship between a dependent (y) and one or more independent (x) variables. Given that linear regression demonstrates a linear relationship, it may be used to determine how the dependent variable's value changes as a function of the independent variable's value.

A sloped straight line (beat fit line) illustrating the connection between the variables is provided by the linear regression model.

Formula for simple linear regression can be written in terms of dependent Y and independent variable X as in:

$$Y = \beta X + \varepsilon \tag{1}$$

1) *Scalability:*

Scaling is frequently required, and since linear regression requires little computational effort, it works well in these circumstances. On the use of big data, model scales well.

2) *Interpretability:*

The linear regression model is simpler than other deep learning models (neural networks). As a result, this approach outperforms black-box models that are unable to explain how an input variable changes an output variable.

3) *Ease of Implementation:*

As there are little engineering overheads required, both before and after the model is launched, the linear regression model is computationally easy to construct.

4) *Optimization:*

These algorithms can be employed in online contexts due to how easily they can be computed. The algorithm may be taught and properly trained with each new instance to produce predictions in real-time, as opposed to support vector machines or neural networks, which must be computationally complex, resource-heavy, and wait a long time to retrain on a new dataset. Such compute-intensive models are costly and unsuited for real-time applications due to all these issues.

5) *Algorithm:*

Firstly, in regression, a series of records with X and Y values are present, and these values are used to train a function that may be used to predict Y from an unknown X. Regression requires us to identify the value of Y, hence we need a function that forecasts Y provided that XY is continuous.

Thus, X is referred to as the predictor variable and Y is referred to as the criterion variable. Regression can be utilized with a wide variety of modules or functions. The basic type of curve is a linear function. In this case, X could be a unique trait or a group of features that collectively point to the problem.

The task of predicting a dependent variable's value (y) based on a specified independent variable (x) is carried out using linear regression. Hence, the term "linear regression" was coined. For example, x can represent a person's job history and y represents their wage. The regression line signifies the line that fits our model the best.

As we train the provided model: input training data (one input variable, or parameter), x Labels to data, y (Supervised learning) The model works the best line to estimate the value or amount of y for a given x during training. By identifying the optimum values for the line's slope and minimizing its error, the model produces the best regression fit line. Here, error is the line's y-intercept, and the slope is the coefficient of x. The best fit line is obtained once the slope and minimum error values are determined to be optimal.

Although it has significant drawbacks, linear regression is a potent tool for comprehending and forecasting a variable's behavior. Its assumption of a linear relationship between the dependent and independent variables, which may not always be the case, is one of its limitations. Moreover, outliers—data points that considerably deviate from the norm—are taken into account using linear regression. These outliers may affect the fitted line disproportionately, which could result in incorrect predictions.

The model seeks to predict y values such that the error difference between predicted value and true value is minimal by attaining the best-fit regression line. To find the optimal value that minimizes the error between the predicted y value (pred) and true y value, it is crucial to change the error and slope values. It is sometimes referred to as the line's cost function.

The cost function (J) of linear regression is the Root Mean Squared Error (RMSE) between the true y value (y) and the predicted y value (pred). Descent in Gradients: The model employs gradient descent to update error and slope variables in order to lower Cost function (minimizing RMSE value) and obtain the best-fit line. Starting with random slope and error values, the goal is to iteratively update the parameters until the minimal cost is reached.

E.Decision Tree

A supervised learning method called a decision tree can be used to solve regression and classification problems, but it is typically preferable for doing so. It is a forest classifier, where nodes in the network stand to represent a dataset's features, branches for the decision-making process, and each leaf node for the classification result.

The Leaf Node and dec Node are the two nodes of a decision tree. here, Leaf nodes are the results of decisions and do not have any more branches, and Decision nodes are utilized to create decisions and have numerous branches.

It is a graphical or visual illustration for obtaining all feasible answers to a choice or problem based on predetermined conditions.

The Classification and Regression Tree algorithm or the CART algorithm, is used to construct a tree.

A decision tree only poses a question and divides the tree into subtrees according to the response (Yes/No).

1) CART Algorithm:

Both classification and regression issues can be resolved with the CART algorithm. Also, it divides the datasets using the Gini index measure, as opposed to the ID3 and C4.5 algorithms, which employ information gain or entropy and gain ratio.

The goal of the greedy strategy used in the CART splitting procedure is to minimize the cost function. The purity of the leaf nodes is calculated for classification tasks using the Gini index as a cost function. To select the most accurate forecast, the algorithm uses the sum squared error as the cost function for regression.

2) Features to Split:

A top-down greedy method is used by decision trees to determine the appropriate feature split. In greedy approaches, all points in the same decision region are split, and further splits are done methodically. The branch (sub-tree) that results has a higher metric value than the preceding tree.

3) Entropy:

Entropy quantifies the degree of uncertainty in the processed information and determines its randomness. The more entropy there is, the harder it is to draw inferences from a situation. The general goal is to reduce entropy and create decision zones that are more homogeneous and contain data points that are members of the same class.

The formula, gives the entropy (E) as in:

$$Entropy(E) = -\sum p_i \log_2\left(p_i\right) \tag{2}$$

where, p is the probability of a class or an element in the data.

4) *Gini Index:*

The metric calculates the odds that a randomly chosen data point will be incorrectly labelled by a specific node. The Gini index serves as the cost function for assessing feature splits in a dataset.

The formula yields the Gini Index as in:

$$Gini = 1 - \sum \left(p_i \right)^2 \qquad (3)$$

5) *Information Gain:*

The Gini index or reduction in entropy as a result of a feature split is measured using the IG metric. When tree-based algorithms use Entropy or Gini index as criteria, informative splits are obtained. In other words, a split like that only reduces the needs by a certain percentage.

The formula provides information gain (IG) as in:

$$IG = E_{before\ splitting} - E_{after\ splitting} \qquad (4)$$

6) *Residual Sum of Squares:*

A top-down greedy method is used by decision trees to determine the appropriate feature split. In greedy approaches, all points in the same decision region are split, and further splits are done methodically. The branch (sub-tree) that results has a higher metric value than the preceding tree.

7) *Algorithm:*

The first step is to decide on the decision at the top of your decision tree as the target or goal (root): Decision trees can be applied in a variety of real-world situations, so, it is essential to determine the main goal of a decision tree, which implies determining what you are trying to identify.

The second step is to make a list of all potential options or moves:

You can make a list of all the options and actions accessible in the following step. It's frequently recommended to limit these in a decision tree scenario.

Take a home plot as an illustration. When you want to purchase something, "money" is the first thing that springs to mind. We start by including a new decision node in the tree diagram as a result. This choice can be made in the form of a question.

If I want to buy a parcel of land in a residential neighborhood, for instance, do I have enough money in the bank to do so?

The decision formulation in this problem structure enables you to take into account all the relevant factors that could affect your choice.

The third step is to determine the standards for each decision:

You can concentrate on defining the evaluation method for each choice once you've determined all the options or factors that will be taken into account. Finalizing the decision points that establish the decision path you should think about to accomplish the goal is what this phrase refers to. Additionally, because decision trees are designed to help you reach an unambiguous decision, it is crucial to make sure that the decision variables are mutually exclusive.

You can move on to step 4 and draw the decision tree after completing step 3. In case 3, the nodes stand in for the decision variables or criteria while branches stand in for the decision actions. The root decision node of the decision tree structure serves as the objective node and serves as the node on which the final choice is made.

The final step is to review your decision-making structure:

After completing, the decision tree diagram can be examined and updated as necessary. This phase allows you to go back and look at the decision tree to see if the decision variables or criteria have been adjusted or updated. If the answer is affirmative, you can edit the tree diagram to show the updated modifications.

Also, you can alter or update the decision criteria and start over if a decision tree produces an inaccurate result. Sharing such tree diagrams with interested teammates and stakeholders can help brainstorming sessions become more efficient and effective while also bringing them closer to the decision tree's main goal. This approach makes sure that everyone on your team is familiar with the concepts behind the decision tree's design.

8) *Practices:*

When developing a decision tree, the following procedures should be taken into account:

Ensure simplicity: Avoid adding more text to the decision tree. Mark the critical decisions using precise terminology.

Predict outcomes using data: A decision tree is useful when taking into account actual facts while figuring out the potential outcomes. As a result, employing data and a straightforward flowchart-based action plan will let you make quick decisions.

Use professionally developed decision tree templates as a priority: Professionally created templates appeal to stakeholders more than amateur ones do. Thus, it is advised as a best practice while developing a decision tree.

F.Figures and Tables

The table below gives a brief about the algorithms used for different datasets and the accuracy that is shown with respect to its algorithm.

Table 1: Results of Different Approaches used for Air pollution Detection

Ref	Datasets Used	Algorithms/Approach	Performance/Accuracy
[1]	Shanghai $PM_{2.5}$ Air Pollution Historical Data.	Random Forest, Linear Regression Algorithm.	RF - 98% LR - 96%
[2]	Air Quality in India (2015-2020).	Linear Regression, K-nearest neighbor Algorithm.	LR - 97% KNN - 93%
[3]	$PM_{2.5}$ Data of five Chinese Cities, Custom modifications.	Random Forest, Stochastic Gradient Boosting, and Model Averaged Neural Network.	RF - 97% SGB - 95% MANN - 93%
[4]	1980-2021 Daily Air Quality Index from EPA.	Support Vector machines, Decision Tree Algorithm.	SVM - 98% DT - 96%

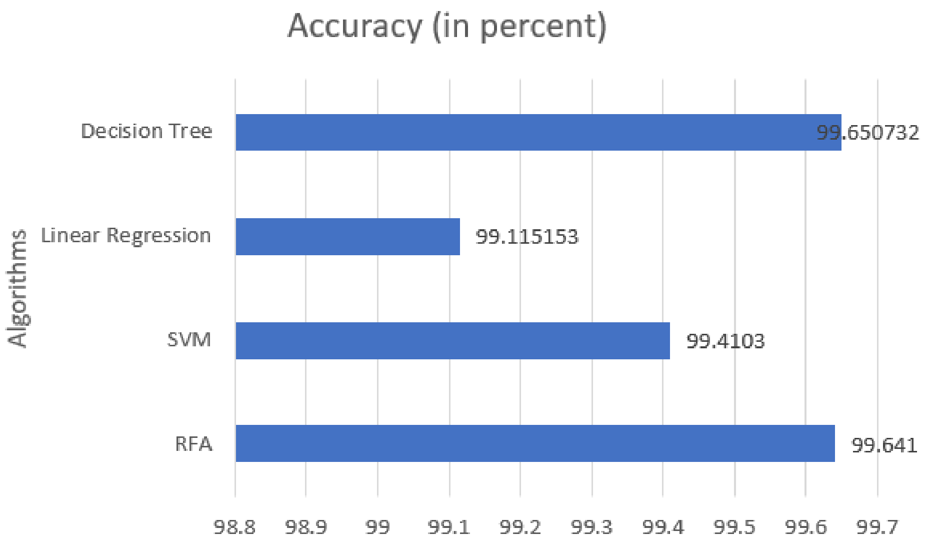

Figure 1: Bar Graph of Accuracy of Different Algorithms

V. CONCLUSION

In conclusion, predicting air quality is a challenging task due to the ever-changing environment, unpredictable nature, and diverse range of pollutants in different places and times. Given the severe consequences of air pollution on human health, wildlife, plants, historic sites, climate, and the environment, it is crucial to regularly monitor and analyze air quality, especially in developing countries. However, research on predicting AQI in India has been relatively limited. In this study, data from 23 Indian cities over a six-year period were utilized, and the dataset was cleaned, pre-processed, and normalized to ensure data quality. Furthermore, a feature selection approach was employed to identify the pollutants that have the greatest impact on AQI, and logarithmic transformations were suggested for future research. These findings highlight the importance of continued research in this field to better understand the factors that affect air quality and develop effective strategies to mitigate the impact of air pollution on public health and the environment.

VI. ACKNOWLEDGMENT

It gives us great pleasure in presenting the preliminary project report on ‘Detection and Predicting Air Pollution Level in a Specific City Using Machine Learning Models’.

We would like to take this opportunity to thank our guide Prof. Madhuri Mane for giving us all the help and guidance we needed. We are really grateful to her for her kind support. Her valuable suggestions were very helpful.

We are also grateful to Dr. G.V. Kale, Head of Department of Computer Engineering, Pune Institute of Computer Technology for her indispensable support and suggestions.

VII. REFERENCES

[1] Sk. Atik Tajwar Sihan 17301109, “Analysing Area Wise Air Pollution Level Using Machine Learning for Better Future”, Brac University, September 2021.

[2] K. Kumar, B. P. Pande, “Air Pollution prediction with Machine Learning: A Case Study of Indian Cities”, Int J Environ Sci Technol (Tehran), 2022.

[3] Stenka Vulova, Fred Meier, Daniel Fenner, and Birgit Kleinschmit “Summer Night in Berlin, Germany: Modelling Air Temperature Specially with Remote Sensing, Crowdsourced Weather Data, and Machine Learning”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, August 2020.

[4] Rasheed Olakunie Kelani, Zainal Ahmad, “Prediction of API using SVM”, Journal of Environmental Chemical Engineering, June 2019.

[5] Suparna De, “Data-Driven Air Quality Characterization for Urban Environments”, University of Surrey, December 2018.

[6] Xiaokai Wang, “Fusion Prediction Model of Atmospheric Pollutant Based on Self-Organized Feature”, Shanxi University, January 2021.

[7] Q. Zhou, Y. Cheng, J. Liu, Y. Chen, J. Ma, "Air Quality Prediction Based on Machine Learning Algorithm in Beijing", IEEE International Conference on Applied System Innovation (ICASI), 2020.

[8] D. Nair, A. Cherukara, R. Ramakrishnan, "Predicting Air Pollution with Machine Learning", International Conference on Communication and Signal Processing (ICCSP), 2020.

[9] Z. Han, Y. Liu, X. Yin, and Y. Zhou, "Predicting Hourly Air Pollution in Beijing Using Machine Learning Techniques", 14th International Conference on Computational Intelligence and Security (CIS), 2020.

[10] Y. Yao, S. Wang, W. Li, "Real-time Air Quality Prediction Using Machine Learning Techniques", International Conference on Big Data and Artificial Intelligence (BDAI), 2020.

[11] J. Chen, Y. Zhang, Y. Liu, "Air Quality Prediction in Smart Cities Using Machine Learning Techniques", International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021.

[12] Hong Zheng, Yunhui Cheng, Haibin Li, "Investigation of Model Ensemble for Fine-Grained Air Quality Prediction", East China University of Science and Technology, July 2020.

[13] Xiang Su, Petri Pellikka, Pan Hui, "Intelligent and Scalable Air Quality Monitoring With 5G Edge", University of Helsinki, April 2021.