Prediction of Air Quality and Pollution using Statistical Methods and Machine Learning Techniques

Mr. V. Devasekhar[1], Dr. P. Natarajan[2]*
Research Scholar, SCOPE, VITU, Vellore, Tamil Nadu, India[1]
Associate Professor, SCOPE, VITU, Vellore, Tamil Nadu, India[2]

***Abstract*—Air pollution is a major environmental issue and machine learning techniques play an important role in analyzing and forecasting these data sets. Air quality is an outcome of the complex interaction of several factors involving the chemical reactions, meteorological parameters, and emissions from natural and anthropogenic sources. In this paper, we propose an efficient combined technique that takes the benefits of statistical techniques and machine learning techniques to predict/forecast the Air Quality and Pollution in particular regions. This work also indicates that prediction performance varies over different regions/cities in India. We used time series analysis, regression and Ada-boosting to anticipate PM 2.5 concentration levels in several locations throughout Hyderabad on an annual basis, depending on numerous atmospheric and surface parameters like wind speed, air temperature, pressure, and so on. Dataset for this investigation is taken from Kaggle and experimented with proposed method and comparison results of our experiments are then plotted.**

***Keywords*—Air quality; forecasting; machine learning; statistical techniques**

I. Introduction

Today, all cities have attracted significant attention in the context of urban development approaches [2,3]. The Internet and developments in broadband networking are seen as enablers of e-services and are becoming increasingly important for urban development and addressing the massive air pollution problem. Citizens and governments around the world have witnessed and expressed growing concern about the impact of air pollution [4] on human health, as well as advocated sustainable development to address air pollution challenges. The result of modern manufacturing is a mixture of liquid droplets, solid particles, and gas molecules that is dispersed throughout the atmosphere. The high concentration of particulate matter of size $PM2.5$ has a major negative impact on human health.

Recent studies have focused on rigorous statistical learning algorithms for evaluating air quality and predicting pollution levels. Neural networks have been employed by Raimondo et al. [5], Garcia et al. [6], and Park et al. [7] to build models for forecasting the occurrence of individual pollutants, such as particles less than 10 microns (PM10). To train their models, Raimondo et al. [5] employed a support vector machine (SVM) and an artificial neural network (ANN). Their best ANN model had a specificity of nearly 79 percent and a false-positive rate of only 0.82 percent, while their best SVM model had a specificity of 80 percent and a false-positive rate of only 0.13 percent. For AQI category prediction, Yu et al. [8] suggested RAQ, a random forest technique. After that, Yi et al. [9] used deep neural networks to predict AQI categories. For forecasting AQI levels, Veljanovska and Dimoski [10] used several settings to surpass k-nearest neighbour (k-NN), decision tree, and SVM. Their ANN model outperformed all other algorithms examined, with an accuracy of 92.3 percent.

The deep learning architecture is suitable for solving air pollution prediction problems and nonlinear problems, and to learn long-term dependencies from time-series data[23]. In [24]. authors presented deep learning solution to predict the hourly forecast of PM2.5 concentration in Beijing, China, based on CNN-LSTM, and other hybrid deep leaning techniques for air pollution analysis is presented in [25].

A. Air Quality Monitoring

The CPCB recommends a combination of physical, wet-chemical, and continuous online measuring procedures for each parameter. Analyzers for measuring PM10, PM2.5, SO2, CO, NO2, O3, NH3, and Benzene are provided in air quality monitoring systems [10,11,13]. Using filter-based air samplers, the metallic parameters Pb, Ni, and As are assessed offline.

The ambient air quality monitoring station (AQMS) having the following systems:

- PM10 & PM2.5:It Measures particle mass concentrations ranging from 0 to 5 mg/m3 with a lowest detection limit of 1 g/m3. It works on the principle of Beta Ray Attenuation. A PM10 intake and a PM2.5 inlet are included in the equipment.
- NOx and NH3: Based on the chemiluminescence technique, with a detection range of 0 to 2000 g/m3 and a minimum detection limit of 0.5 g/m3.
- SO2 Analyzer: Operates on the UV Fluorescence technique, with a detection range of 0 to 2000 g/m3 and a minimum detection limit of 0.5 g/m3.
- CO Analyzer: Uses the Non-Dispersive Infrared Spectrometry (NDIR) method to measure CO levels ranging from 0 to 100 mg/m3 with a detection limit of 0.03 g/m3.
- O3 Analyzer: Works on the UV Photometry principle, with a range of 0 to 2500 g/m3 and a minimum detection limit of 0.5 g/m3.

*Corresponding author.

- BTEX (Benzene, Toluene, Ethylbenzene, Xylene): GC/PID for automatic monitoring of BTEX in air with a minimum detection threshold of 10 ppt in ambient air.
- Multigas Calibrator: used to manually, remotely, or automatically calibrate gas analyzers for quality assurance. Up to 20 points of multi-calibration Ultrasonic Wind Sensor, Barometric Pressure, Temperature, Relative Humidity, Rainfall, Solar Radiation, and other features of an automatic weather station (AWS).
- Except for the AWS, all of these instruments are kept in a room or walk-through shelter with an appropriate sampling system for gaseous and particulate matter measurements.

B. Our Contribution

In this research,

1) We propose an effective combination method to forecast and anticipate air quality and pollution in specific areas by combining the advantages of statistical and machine learning methods.
2) Additionally, this research suggests that prediction accuracy differs between Indian cities and regions.
3) We predicted annual PM 2.5 concentration levels in different places throughout Hyderabad using time series analysis, regression, and Ada boosting, depending on a variety of meteorological and surface characteristics like wind speed, air temperature, pressure, and so on.
4) The investigation's data set was obtained via Kaggle, and the suggested strategy was tested.

C. Organization of the Paper

The remaining paper is structured as follows- Literature review in this domain is presented in Section II. Section III provides preliminaries related to machine learning and other statistical techniques. Section IV presents our proposed hybrid method. The experiment results with comparative results are mentioned in Section V. Conclusions are discussed in Section VI.

II. Related Work

Gopalakrishnan (2021) [26] used Google Street View data and machine learning to predict air quality in various locations throughout Oakland, California. The author created a web application that can predict pollution levels in any city and neighbourhood. Sanjeev (2021) [27] examined a set of data that would include pollutant concentrations as well as meteorological factors. Castelli et al. (2020) [28] used the Support Vector Regression (SVR) ML algorithm to forecast air quality in California in terms of pollutants and particulate levels. The researchers claimed to have created a novel method for modeling hourly atmospheric pollution. In [29] Doreswamy et al. (2020) investigated ML predictive models for PM concentration forecasting in the air. The authors examined six years of Taiwanese air quality monitoring data and applied existing models. They claimed that predicted and actual values were extremely close. Based on 11 years of data, Liang et al. (2020) [30] investigated the performance of six ML classifiers in predicting Taiwan's AQI. . Madan et al. (2020) [31] compared the performance of ML algorithms and twenty different literary works over pollutants studied. The authors discovered that many works used meteorological data such as humidity, wind speed, and temperature to more correctly estimate pollutant concentrations. They discovered that the Neural Network (NN) and boosting models outperformed the other leading machine learning (ML) algorithms. Monisri et al. (2020)[32] gathered air pollution data from a variety of sources in order to create a mixed model for predicting air quality. The proposed model, according to the authors, aims to assist people in small towns in analysing and forecasting air quality. Based on ML classifiers, Nahar et al. (2020) [33] created a model to predict AQI. Their proposed model accurately detected the most contaminated areas. Patil et al. (2020) presented some research papers on various machine learning techniques for AQI modeling and forecasting. Multi-agent systems[11,12] have been proposed as a helpful apparatus for huge scope frameworks like Important traffic and air quality control). The significant objective of such a framework is to help street administrators with traffic the board tasks while likewise further developing air quality on the course. Many examinations have proposed the use of MAS innovation in traffic signal and the executives frameworks, for example, (Namoun et al. 2013), which proposes an incorporated technique for demonstrating transportation framework and streamlining transportation in metropolitan regions to diminish fossil fuel byproducts.

In the hybrid classification PM2.5 fixation determining models, highlight determination is seldom applied in several techniques. Nonetheless, assuming that a PM2.5 fixation determining model's feedback incorporates an enormous number of elements (PM 2.5 etc.), it could be hard to prepare the model and increment the preparation time. This affects the PM2.5 fixation anticipating model's power [14]. At the same time, muddled info information might bring about overfitting of the model and a decrease in model precision [15]. The guideline parts investigation (PCA), stage space transformation (PSR), and angle helped relapse tree are at present famous component choice methodologies (GBRT).

Notwithstanding, on the grounds that these techniques assume a direct framework, they might be insufficient for air contamination focus arrangements, bringing about issues, for example, inability to accomplish worldwide ideal decrease. The fluffy hypothesis based unpleasant sets characteristic decrease (RSAR) strategy offers the benefits of unambiguous stop measures and no boundaries [16]. Through the reliance between particular ascribes, RSAR can decide the objective property's fundamental trait set. The RSAR calculation [17] is a well known review point. Information mining and examination as often as possible utilize grouping methods [33]. K-implies grouping (KC) [18], probabilistic c-means (PCM) [19], fix clustering [20], and other bunching approaches exist. The KC algorithm, when compared to others, gives an advantage based on easy procedure, quick computation speed, and great clustering results; as a result, it is currently the most extensively used clustering algorithm[21]. Combining the RSAR and KC algorithms allows the RSAR method to generate suitable, which is a promising exploration course.

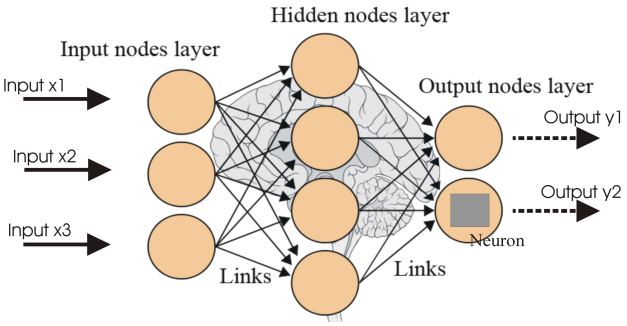


Fig. 1. Artificial neural network structure.

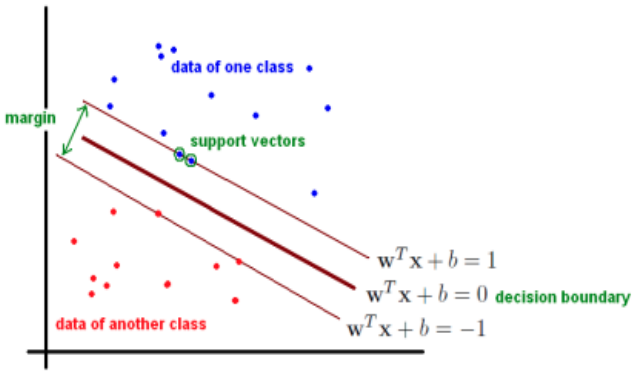


Fig. 2. SVM pictorial representation.

III. Machine Learning Techniques

Atmospheric particulate matter (PM) is one of the pollutant that may have a significant impact on human health. Data collected from various regions of country can be analysed using various models : a multiple linear regression model, a neural network models and other machine learning models. In this section, we present some significant statistical preliminaries, soft computing based methods and mathematical notations [2,4,7].

A. Artificial Neural Network

It is a collection of classifiers whose standard project for future and utility are similar to the algorithmic model of human cerebrum structure [1]. The precise organisation of neuronal organisation varies for order enquiry. To begin, the topological structure and number of organisation hubs present in the core layer are resolved throughout the preparation. For example, n-dimensional planes and hyperplanes have no eccentricity, unlike SVM. In any event, preparing informational collection measures takes time and results in less precise and proficient results. Fig. 1 provides ANN framework and its input parameters.

- For singleton data sequence, $x_0, x_1, x_2, x_3 \cdots x_{(n)}$ depicts several data items to the computational network. Each input is being multiplied with a corresponding weights. These weights are depicted as $w_0, w_1, w_2, w_3 \cdots w_{(n)}$. These weight represents the major strength of any specific node.
- Here b is considered as a bias value. A certain bias value permits to move the enactment increasing or decreasing work.
- In most straightforward cases, these items are added, taken care of a transfer function (activation function) to produce an outcome, and this outcome is sent as yield.

$$x_1.w_1 + x_2.w_2 + x_3.w_3 \cdots x_n.w_n = \sum xi.wi$$

- Presently, enactment (activation) function is applied, i.e. $\phi(\sum x_i.w_i)$

B. Support Vector Regression

Its well known that the categorization method comes in either supervised method or unsupervised method. Therefore, in the field of ML, support vector network architecture comes in the category of supervised machine learning standards. An SVM [2, 3] is the description of features / attribute states in the plane, alongside the non-direct hyperplanes for detachment task in order. A few boundaries like gaussian parts, standard deviation and fluctuation of information, bit capacities are some critical boundaries which influence the exhibition of SVM.

Fig. 2 provides the SVM framework and its representation.

1) Linearly-separable data aided Binary classification:

- **Goal:** we need to discover the hyperplane (for example choice limit) directly isolating to listed class. The limit of listed the condition: $w^T x + b = 0$.
- Any of the classes that falls this decision boundary must be label to 1.
  i.e., x_i such that $w^T x + b > 0$ will be having respective $y_i = 1$.
- Likewise, anything underneath the choice limit ought to have labeling - 1. i.e., x_i s.t. $w^T x + b < 0$ will possesses respective $y_i = -1$.

C. Learning

This section lists out various types of learning and tries to find out the answer of the question that, why deep learning outperforms over other traditional algorithms? The discussion is given below.

Types of learning:-
Various learning methods are mentioned below:-

1) Active learning: It picks a subset of an unstructured and basic event for reason for marking. The dynamic learner acquires bigger precision utilizing decreased measure of events.

2) Kernel-based learning: It is demonstrated to be a predominant approach to upgrade the computational potential proficiently. It is profitable with respect to that, both direct just as non-straight vector bit utilitarian strategies are available

to manage the non-linearity of information in N-dimensional element space.

3) Transfer learning: It is primarily advantageous as it can productively apply information, which has been adapted already so as to discover answer for new issues in quick, optimal and successful way.

4) Distributed learning: This sort of learning restrains the bunch arrangement, in which one processes thread is designated to each group in plan to perform multi-stringing in parallel and distributed fashion.

5) Deep learning: This learning considers more muddled, compartmented measurable examples of data sources and figures out how to be robust for new areas when contrasted with customary learning frameworks.

6) Supervised learning: It deals with learning a function from available training set data. A supervised learning algorithm uses the accessible training information and makes an induced capacity, which can be used further for the purpose of mapping the new ones.

7) Unsupervised learning: It manages unlabeled information without taking any predefined dataset for its preparation. This learning can be considered as an intense apparatus for look for patterns and trends and analysing available data. It is commonly employed for clustering similar input into logical groups.

8) Classification: is learning an specific function that maps (orders) an information thing into one of a few predefined categories [24]. Instances of grouping techniques utilized as a feature of information revelation applications remember the arranging of patterns for money related business sectors and the computerized recognizable proof of objects of revenue in enormous image oriented data sets [25]. The bank may exploit the characterization districts to consequently choose whether future credit candidates will be given an advance or not.

9) Regression: is learning a limit that maps a data thing to a real-valued esteemed forecast variable. Relapse applications are many, for instance, predicting the proportion of biomass present in a backwoods given distantly detected microwave estimations, assessing the likelihood that a patient will endure given the after effects of a bunch of analytic tests, anticipating shopper interest for another item as a component of promoting use, and foreseeing time arrangement where the information factors can be time-slacked adaptations of the prediction variable.

10)Clustering: is a typical graphic assignment where one looks to distinguish a limited arrangement of classifications or bunches to portray the information. The classes can be commonly exclusive and thorough or comprise of a more extravagant portrayal, for example, various leveled or covering classifications. Instances of bunching applications in an information disclosure setting incorporate finding homogeneous sub populaces for various customers in advertising data sets and distinguishing subcategories of spectra from infra-red sky estimations.

D. Information and Decision System

An instance of *information system* is represented as pair (U, A) where,

U : denotes - a kind of non-empty set.
A : denotes - a non-empty type finite set of features.
However, A *decision table* is a system possessing the form $S = (U, C \cup \{d\})$ where -
C : denotes set of conditional variables.
d : denotes the decision variable.

E. Uncertainty Approximation

The approximation extent of an uncertain variety of concept in a knowledge space is performed as follows - Let $S = (U, R)$ be a certain approximation space, X be a type of concept in that space, then, the *lower approximation(inf)* is-
$\underline{R}X = \{x \in U | [x] \subseteq X\}$
upper approximation(sup) is-
$\overline{R}X = \{x \in U | [x] \cap X \neq \emptyset\}$
where, $[x]$ can be considered as equivalence class, possessing an element e.

IV. Proposed Method

We propose an efficient method that uses statistical method and machine learning techniques for Air Quality Predication. In this section, we present our proposed method in phased manner as follows:

A. Data Preprocessing

Various preprocessing procedures, in overall, come before the learning phase. In India, O3, PM2.5, PM10, CO, SO2, and NO2 pollutants are monitored with respect to its concentration. Our method used the air quality data [22] between Nov 2019 to April 2020, that are collected from several monitoring stations across major metropolitan cities such as Delhi, Mumbai, Chennai, Bangalore and Hyderabad of India and reported via the Govt website [22]. At any given time, an inaccurate parameter will not influence the global data group.

B. Feature Selection

To get an efficient representation of data, the auto encoder procedure must capture very important features using several measures which are part of the method.

Given sequence of data points $x(1), x(2), x(3), ..., x(N)$, where $x^i \in R^D$, an auto encoder first encodes the info vector x to a more elevated level secret portrayal y in view of condition (1), and after ward it unravels the portrayal y back to a remaking z, determined as in condition (2):

$$y = f(W_1 x + b) \tag{1}$$

$$z = g(W_2 y + c) \tag{2}$$

where W_1 and W_2 are weights chosen for the procedure and b and c are the vectors of corresponding function. We utilized the strategic sigmoid capacities for $f(x)$ and $g(x)$ in this method.

Sigmoid function = $1/(1 + exp(-x))$

The boundaries of this neural organization are enhanced to limit the normal remaking error:

$$j(\theta) = \frac{1}{N}\sum_{i=1}^{N} M(x^{(i)}, z^{(i)}) \quad (3)$$

In this a loss function is M. We involved the customary squared blunder in our method.

C. Statistical Modeling

The underlying dividing step can be done at least numerous times to wipe out the non-stationarity of the mean capacity, and the accompanying quantifiable approach for showing is used as the info data exhibits proof of non-stationarity in the sensation of mean.

BEGIN PROCEDURE {
Consider, process PROC(α, β, γ) where; α, β, γ are +ve integers.
α: no. of lags.
β: difference degree
γ: order of the moving average model
for an input time series pollution data D_t, here index of integer is t and D_t are real numbers.
PROC(α', γ) is given by -

$$D_t - a_1 D_{t-1} - \cdots - a_{\alpha'} D_{t-\alpha'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_\gamma \varepsilon_{t-\gamma} \quad (4)$$

above equation can be equivalently represented as -

$$\left(1 - \sum_{i=1}^{\alpha'} a_i M^i\right) D_t = \left(1 + \sum_{i=1}^{\gamma} \theta_i M^i\right) \varepsilon_t \quad (5)$$

here, M is lag operator
a_i - autoregressive parameters
θ_i - average part parameters
ε_t are represented as error terms
$\hookrightarrow$ ε_t are considered to be iid (independent, identically distributed) variables and these are normal distribution samples with zero mean.

Consider the polynomial $\left(1 - \sum_{i=1}^{\alpha'} a_i M^i\right)$ has a unit root of multiplicity β. The mentioned polynomial also can rewritten as -

$$\left(1 - \sum_{i=1}^{\alpha'} a_i L^i\right) = \left(1 - \sum_{i=1}^{\alpha'-\beta} \varphi_i M^i\right)(1 - M)^\beta \quad (6)$$

with polynomial factorization property,
$\alpha = \alpha' - \beta$ and given as -

$$\left(1 - \sum_{i=1}^{\alpha} \varphi_i M^i\right)(1 - M)^\beta D_t = \left(1 + \sum_{i=1}^{\gamma} \theta_i M^i\right) \varepsilon_t \quad (7)$$

this can be further generalized as -

$$\left(1 - \sum_{i=1}^{\alpha} \varphi_i M^i\right)(1 - M)^\beta D_t = \delta + \left(1 + \sum_{i=1}^{\gamma} \theta_i M^i\right) \varepsilon_t \quad (8)$$

where, drift $\dashrightarrow \frac{\delta}{1 - \sum \varphi_i}$

} *END PROCEDURE*

D. Model Evaluation

To build a model to predict concentrations, we used two different machine learning algorithms including a simple linear regression model and non linear regression. The NO2 dataset was split into test/train data and a cross-validation approach was applied to the training dataset. We use Linear Regression and A da-boosting for improving efficiency of the results.

When the quantity of PM 2.5 and PM 10 pollutant particles in the atmosphere is indeed very high, it has a negative impact on our health and can cause life-risk issues in a less period of time. Particulate matter has been shown to have an effects on peoples health, often at the genetic level, according to studies. So we are emphasising our work to forecast the concentration of PM 2.5 levels in the atmosphere. "Hyderabad Weather with Air Quality index and Covid" dataset is used. The dataset [22] is downloaded into .csv format. Brief about the datasets are as follows: It consists of a total of 5 months of data between October 2019 to April 2020 as described below:

- Date: dd/mm/yyyy
- Humidity
- Wind Speed
- Dew Point
- Temperature
- Pressure
- Festival (Rating out of 5)
- Lockdown (0 for No, 1 for Yes)
- Covid-19 Cases in Hyderabad
- Air quality (PM2.5)

Fig. 3 shows the first 10 rows along with the column values of the dataset when reading the csv file into the colab file.

```
df = pd.read_csv('/content/Hyderabad-AirQ -2019-20.csv')
df.head(10)
```

	Date	Humidity	Wind Speed	Dew Point	Temperature	Pressure	Festival	Lockdown	Covid-Case	PM2.5
0	01-10-2019	83.0	4.5	76.0	81.9	28.1	0	0	0	84
1	02-10-2019	81.6	4.6	77.4	83.8	28.1	0	0	0	83
2	03-10-2019	82.0	3.7	75.3	81.7	28.1	0	0	0	81
3	04-10-2019	85.4	2.7	73.9	78.6	28.1	0	0	0	94
4	05-10-2019	87.4	3.5	75.3	79.4	28.1	0	0	0	112
5	06-10-2019	86.3	2.5	73.3	77.9	28.1	4	0	0	97
6	07-10-2019	82.9	2.4	76.0	81.6	28.1	0	0	0	126
7	08-10-2019	86.2	3.1	76.3	80.8	28.0	3	0	0	125
8	09-10-2019	88.7	1.7	74.5	78.1	28.0	0	0	0	132
9	10-10-2019	85.6	3.4	72.9	77.3	28.1	0	0	0	138

Fig. 3. First 10 rows along with column of the dataset.

We pre-processed it, check if there is any null value inside it and removed it. After the data has been cleaned and pre-processed, it is submitted to later experiment, which includes time series analysis and determining the total impact

of each characteristic on the PM 2.5 value. Date is also an impacting factor which shows the PM 2.5 values increasing or decreasing with time. So we grouped them and find the values of each attribute. Fig. 4 shows the grouped values of each attribute. With the help of these values we are plotting the variation of PM 2.5 w.r.t date. Fig. 5 shows graph plot of PM 2.5 vs Date.

```
[ ] df= df.groupby("Date").mean()
    df.head(10)
```

Date	Humidity	Wind Speed	Dew Point	Temperature	Pressure	Festival	Lockdown	Covid-Case	PM2.5
01-03-2020	62.5	5.3	67.3	82.7	28.1	0	0	1	113
01-04-2020	62.0	3.9	72.0	92.6	28.1	0	1	96	120
01-10-2019	83.0	4.5	76.0	81.9	28.1	0	0	0	84
01-11-2019	82.7	4.8	75.3	81.3	28.1	1	0	0	57
01-12-2019	80.7	5.5	71.7	78.2	28.2	0	0	0	147
02-03-2020	58.9	4.6	67.3	84.0	28.1	0	0	1	114
02-04-2020	55.0	4.7	72.0	96.8	28.1	0	1	96	130
02-10-2019	81.6	4.6	77.4	83.8	28.1	0	0	0	83
02-11-2019	79.8	4.3	75.3	82.4	28.1	2	0	0	61
02-12-2019	86.3	5.8	71.1	75.3	28.2	0	0	0	144

Fig. 4. Grouped values of the dataset.

The format of date is $xx - zz - yy$ where xx is date, zz is month and yy is year. The value of PM 2.5 is increasing yearly. In December its value is higher as compared to other months. December is a winter month, so we say that in the winter months PM 2.5 value is high.

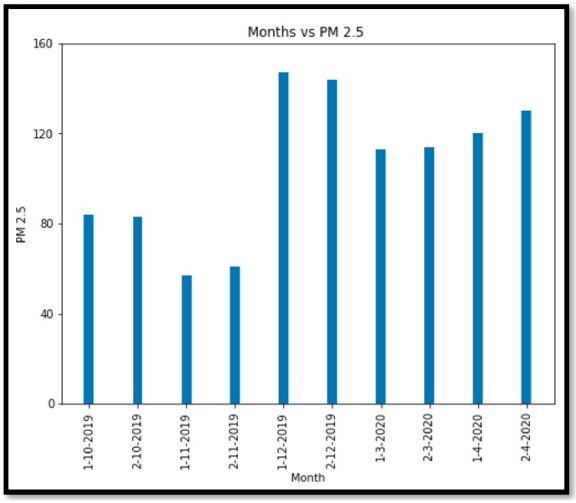


Fig. 5. PM 2.5 vs Month.

The effect of wind on PM 2.5 is also investigated. The PM 2.5 value was found to be lower when the wind speed was high, and vice versa. Fig. 6 shows a scatter plot of wind speed vs PM 2.5 value, which confirms that we have higher wind speed concentrations. It's also worth noting that when the wind speed is between $6-10m/s$, PM 2.5 levels are essentially non-existent.

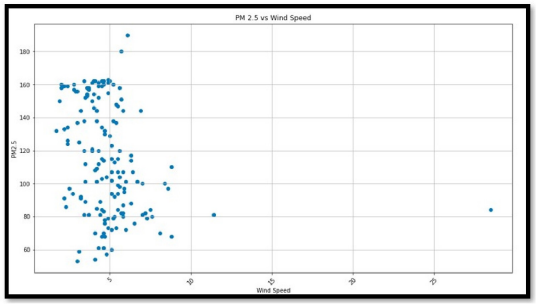


Fig. 6. Effects of PM 2.5 due to wind speed.

The impacts of humidity on PM 2.5 is investigated. The value of PM 2.5 value was found to be lower when the humidity value ranges between 30 $gm^3 to 60 gm^3$. Its value increases to 120 μm when humidity value ranges between 60 $gm^3 to 80 gm^3$. Fig. 7 shows a scatter plot of humidity vs PM 2.5 value.

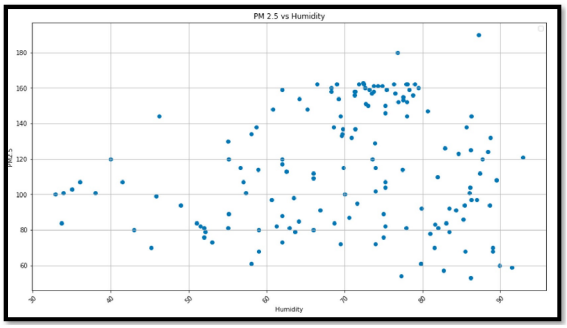


Fig. 7. Effect of humidity on PM 2.5 values.

The effect of dew drops on PM 2.5 was found to be lower when dew drops value ranges between $60°F - 65°F$and PM 2.5 value increases when dew points value ranges between $65° to 75°$. Fig. 8 shows a scatter plot between PM 2.5 vs Dew Points

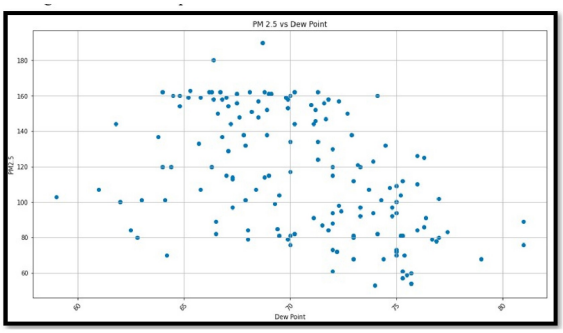


Fig. 8. Effect of dew points on PM 2.5.

Similarly the impacts of temperature on PM 2.5 is investigated. Its value was found lower when the value of temperature

increases whereas the value of PM 2.5 increases when the temperature is lower. It shows that at winter season PM 2.5 value increase into the atmosphere. Fig. 9 shows a scatter plot between PM 2.5 vs Temperature.

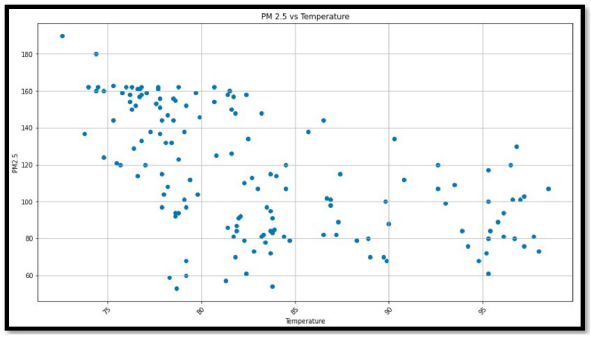


Fig. 9. PM 2.5 vs temperature.

Also the effects of pressure is investigated on PM 2.5. Fig. 10 shows a scatter plot between PM 2.5 and Pressure.

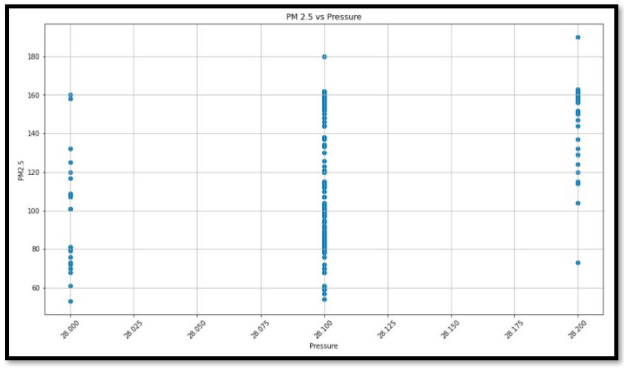


Fig. 10. PM 2.5 vs pressure.

V. Experiment Results and Comparative Analysis

We present experiment results of our proposed methods and efficiency analysis among several methods.

A. Metrics

Firstly, we brief the metrics used for measuring the results are discussed below:

1) Mean Absolute Error (MAE):: It refers to the size of the difference between the predicted and true value of an observation. The average of absolute errors for the entire group is used to calculate the size of mistakes for a group of forecasts and observations.

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}|\bar{\Upsilon}_i - Y_i|$$

In this N is the total number of data items, y_i is i-th measurement, and $\bar{\Upsilon}_i$is its respective prediction.

2) Root Mean Square Error (RMSE):: One of the most often used approaches for assessing the validity of estimates is root mean square error, also known as root mean square deviation. It uses Euclidean distance to demonstrate how far predictions differ from observed true values.

$$RMSE = \sqrt{\frac{\sum_i^n |Y_i - \bar{\Upsilon}_i|^2}{n}}$$

In this N is the total number of data items, y_i is i-th measurement, and $\bar{\Upsilon}_i$is its respective prediction.

3) R2_score:: The coefficient of determination, often known as the R2 score, is used to assess the performance of a linear regression model.

$$R2_{score} = 1 - \frac{\Sigma_{i=1}(Y_i - \bar{\Upsilon}_i)^2}{\Sigma_{i=1}(Y_i - \mu)^2} \quad (9)$$

In this N is the total number of data items, y_i is i-th measurement, and $\bar{\Upsilon}_i$is its respective prediction and μ is the mean of actual values.

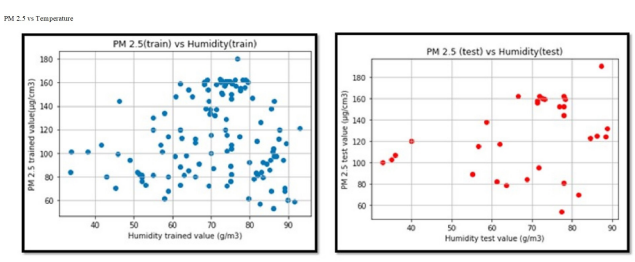
B. Methods: Experiment Results

We present experiment results of our proposed methods by adopting linear regression, Ada-boosting and XG-boosting on standard data sets available on Kaggle [22].

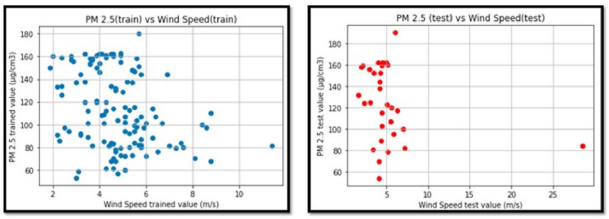
1) Linear regression: We get the values of R2 is very less or nearly equal to 0.5. Absolute mean error is 27.33 when temperature vs PM 2.5 Linear Regression calculated. Root mean square error is 30.31 when temperature vs PM 2.5 Linear Regression is calculated. Table I shows the comparative result of mean absolute error, mean square error and r2 score. Fig. 11(a), (b) show the comparative analysis graph. As absolute mean error and root mean square error are the errors so when their value is small, then the model is good. And $r2\_score$ calculates the accuracy of the model so when their value is high, model is good.

TABLE I. Comparative Result of Evaluation Metrics.

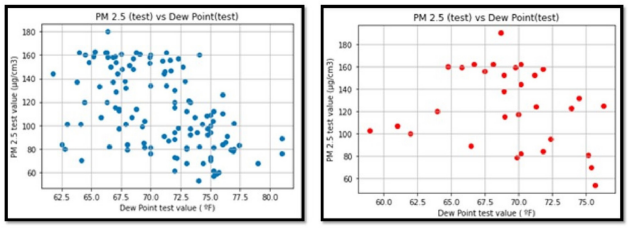
Factors/Error	MAE	RMSE	$R2\_sccore$
Humidity	29.51	35.99	-0.07
Wind Speed	35.39	41.93	-0.54
Dew Point	31.58	34.58	-0.06
Temperature	**27.33**	**30.31**	**0.2**
Pressure	29.28	33.86	0.0



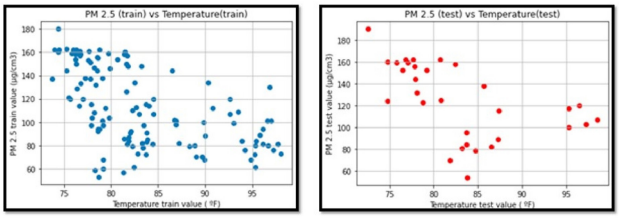
(a) Training (left) and testing (right) results for PM2.5 vs. Humidity



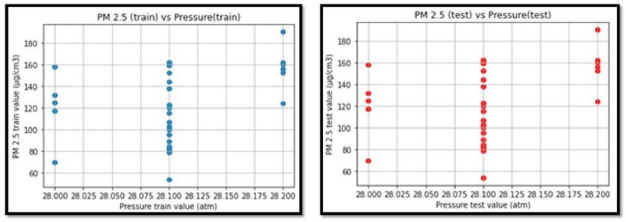
(b) Training (left) and testing (right) results for PM2.5 vs. Wind speed



(c) Training (left) and testing (right) results for PM2.5 vs. Dew-point



(d) Training (left) and testing (right) results for PM2.5 vs. Temperature



(e) Training (left) and testing (right) results for PM2.5 vs. Dew Pressure

Fig. 11. Training and testing results.

Fig. 11(a) represents training and testing results between PM2.5 and Humidity. Fig. 11(b) represents training and testing results between PM2.5 and Wind speed. Fig. 11(c) represents training and testing results between PM2.5 and Dew-point. Fig. 11(d) represents training and testing results between PM2.5 and Temperature. Fig. 11(e) represents training and testing results between PM2.5 and Pressure.

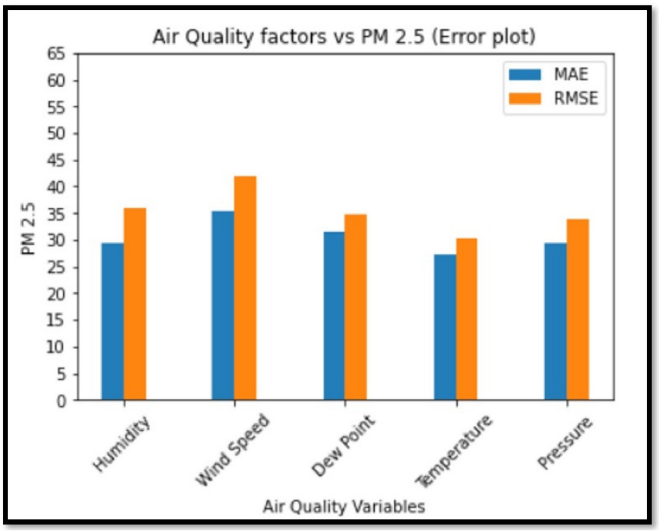


Fig. 12. Comparative analysis of mean square errors and root mean square error.

Fig. 12 shows the comparative analysis of mean square errors and root mean square error.

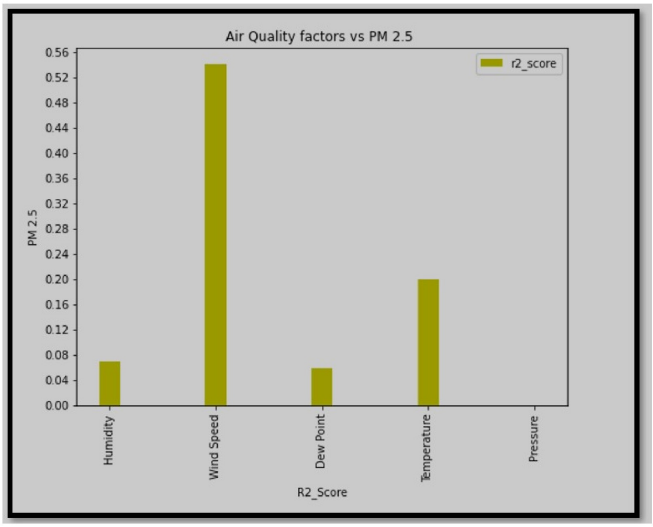


Fig. 13. Comparative analysis of $r2\_score$ on atmospheric factors.

2) Ada-Boosting: We get the values of R2 is very less or nearly equal to 0.38. Absolute mean error is 22.39 when temperature vs PM 2.5 Ada-Boosting Regression calculated. Root mean square error is 26.42 when temperature vs PM 2.5 Ada-Boosting Regression is calculated.

Table II shows the comparative result of mean absolute error, mean square error and r_2 score. Fig. 13 and 14 shows the comparative analysis graph.

TABLE II. COMPARATIVE RESULT OF EVALUATION METRICS

Factors/Error	MAE	RMSE	$R2\_sccore$
Ada- Humidity	27.68	35.55	-0.10
Ada-Wind Speed	29.33	36.41	-0.16
Ada-Dew Point	27.35	30.76	-0.17
Ada-Temperature	**22.39**	**26.42**	**0.38**
Ada-Pressure	28.99	33.44	-0.02

Fig. 14. Comparative analysis of mean square errors and root mean square error using Ada-Boosting.

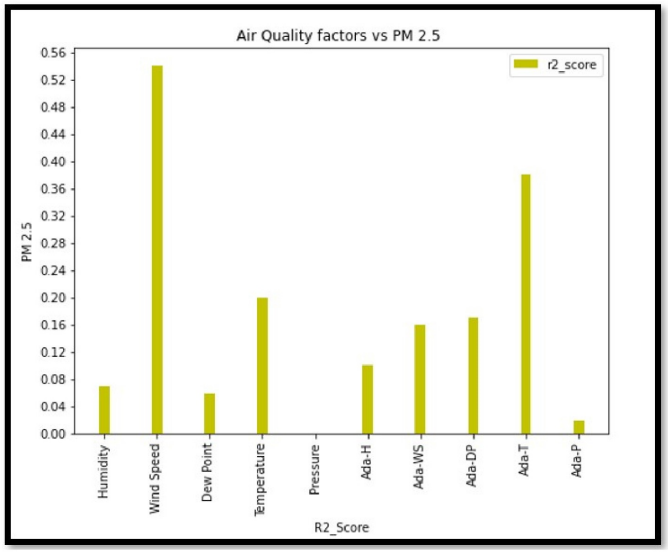


Fig. 15. Comparative analysis of $r2\_score$ using Ada-Boosting.

3) XG-Boosting regression: We get the values of $r2_$score is very less or nearly equal to 0.23 when PM 2.5 vs Humidity is calculated. Mean Absolute error is 28.09 when humidity vs PM 2.5 XG-Boosting Regression calculated. Root mean square error is 35.07 when dew points vs PM 2.5 XG-Boosting Regression is calculated. Table III shows the comparative result of mean absolute error, mean square error and r2 score. Fig. 15 and 16 show the comparative analysis graph.

TABLE III. COMPARATIVE RESULT OF EVALUATION METRICS

Factors/Error	MAE	RMSE	$R2\_sccore$
XGB- Humidity	**28.09**	37.62	**-0.23**
XGB-Wind Speed	32.31	42.10	-0.5
XGB-Dew Point	28.67	**35.07**	-0.07
XGB-Temperature	29.65	38.92	-0.32
XGB-Pressure	28.30	32.87	-0.05

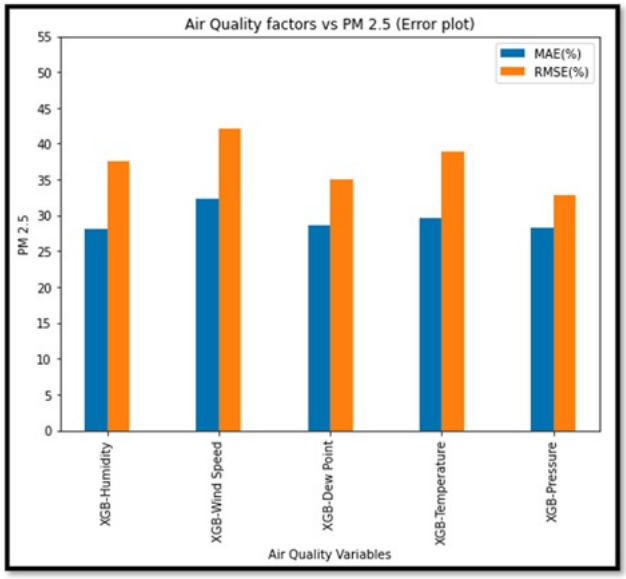


Fig. 16. Comparative analysis of mean square errors and root mean square error using XG-Boosting.

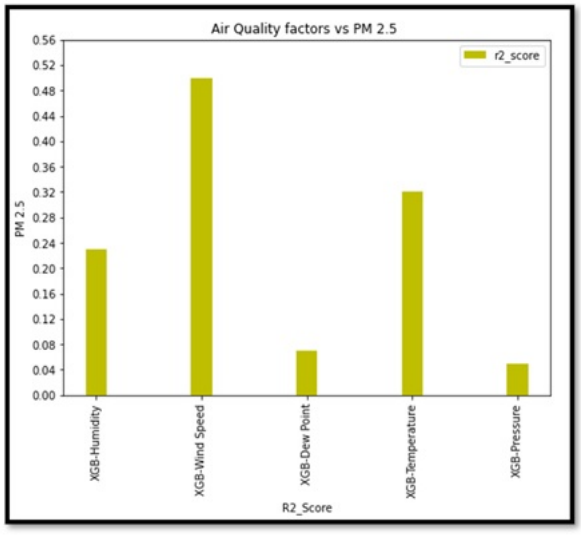


Fig. 17. Comparative analysis of $r2\_score$ using XG-Boosting.

4) Efficiency analysis of models: Final comparative result of all the models are given in a Table IV and the comparative graph is shown in Fig. 17.

TABLE IV. COMPARATIVE RESULT OF EVALUATION METRICS

Factors/Error	MAE	RMSE	$R2\_sccore$
LR-Humidity	29.51	35.99	-0.07
LR-Wind Speed	35.39	41.93	-0.54
Dew Point	31.58	34.58	-0.06
L-Temperature	**27.33**	**30.31**	**0.2**
Pressure	29.28	33.86	0.0
Ada- Humidity	27.68	35.55	-0.10
Ada-Wind Speed	29.33	36.41	-0.16
Ada-Dew Point	27.35	30.76	-0.17
Ada-Temperature	**22.39**	**26.42**	**0.38**
Ada-Pressure	28.99	33.44	-0.02
XGB- Humidity	28.09	37.62	-0.23
XGB-Wind Speed	32.31	42.10	-0.5
XGB-Dew Point	28.67	35.07	-0.07
XGB-Temperature	**29.65**	**38.92**	**-0.32**
XGB-Pressure	28.30	32.87	-0.05

VI. CONCLUSION

Worldwide, air pollution is responsible for around 1.3 million deaths annually according to the World Health Organization (WHO) [11]. The depletion of air quality is just one of harmful effects due to pollutants released into the air. In this paper, we propose an efficient combined technique that takes the benefits of multi-agent systems, statistical techniques and machine learning techniques for forecasting Air Quality utilizing supervised machine learning procedures. We used Linear Regression and Ada-boosting for improving efficiency of the results. As part of our future work, we will improve our results using various deep learning techniques. Further, we will explore whether adopting lag meteorological variables and tuning the hyper parameters to improve the accuracy of the model.

REFERENCES

[1] Yi X, Zhang J, Wang Z, Li T, Zheng Y. Deep distributed fusion network for air quality prediction. the 24th ACM SIGKDD International Conference: ACM; 2018. https://doi.org/10.1145/3219819.3219822.

[2] Lin Y, Mago N, Gao Y, Li Y, Chiang YY, Shahabi C, et al. Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning. In: The 26th ACM SIGSPATIAL International Conference; 2018. p. 359–68. https://doi.org/10.1145/3274895.3274907.

[3] Liao, Q., Zhu, M., Wu, L. et al. Deep Learning for Air Quality Forecasts: a Review. Curr Pollution Rep 6, 399–409 (2020). https://doi.org/10.1007/s40726-020-00159-z.

[4] Athira V, Geetha P, Vinayakumar R, Soman KP. DeepAirNet: applying recurrent networks for air quality prediction. Procedia Comput Sci. 2018;132:1394–403. https://doi.org/10.1016/j.procs.2018.05.068.

[5] Raimondo, G.; Montuori, A.; Moniaci, W.; Pasero, E.; Almkvist, E. A Machine Learning Tool to Forecast PM10 Level. In Proceedings of the Fifth Conference on Artificial Intelligence Applications to Environmental Science, San Antonio, TX, USA, 14–18 January 2007; pp. 1–9.

[6] Garcia, J.M.; Teodoro, F.; Cerdeira, R.; Coelho, R.M.; Kumar, P.; Carvalho, M.G. Developing a Methodology to Predict PM10 Concentrations in Urban Areas Using Generalized Linear Models. Environ. Technol. 2016, 37, 2316–2325.

[7] Park, S.; Kim, M.; Kim, M.; Namgung, H.-G.; Kim, K.-T.; Cho, K.H.; H, K.; Kwon, S.-B. Predicting PM10 Concentration in Seoul Metropolitan Subway Stations Using Artificial Neural Network (ANN). J. Hazard. Mater. 2018, 341, 75–82.

[8] Yu, R.; Yang, Y.; Yang, L.; Han, G.; Move, O.A. RAQ A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. Sensors 2016.

[9] 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 and Data Mining, London, UK, 19–23 August 2018; pp. 965–973.

[10] Veljanovska, K.; Dimoski, A. Air Quality Index Prediction Using Simple Machine Learning Algorithms. Int. J. Emerg. Trends Technol. Comput. Sci. 2018, 7, 25–30.

[11] Ivo T, Miguel A, Rosaldo R, Eugénio O (2011) Using TraSMAPI for developing multi-agent intelligent traffic management solutions. In: Demazeau Y, Pěchouček M, Corchado J, Pérez J (eds) PAAMS 2011: 9th international conference on practical applications of agents and multiagent systems.

[12] Salamanca, Spain, April 2011. Advances in intelligent and soft computing (advances on practical applications of agents and multiagent systems), vol 88, Springer, Heidelberg.

[13] Jin X, Jie L (2012) A study of multi-agent based model for urban intelligent transport systems. Int J Adv Comput Technol 4:126–134. doi:10.4156/ijact.vol4.issue6.15.

[14] L. Da, J. Wang, W. Hui Short-term wind speed forecasting based on spectral clustering and optimised echo state networks Renew Energ, 78 (2015), pp. 599-608.

[15] L. Xu, Y. Yu, J. Yu, J. Chen, Z. Niu, L. Yin, et al. Spatial distribution and sources identification of elements in PM2.5 among the coastal city group in the Western Taiwan Strait region, China Sci Total Environ, 442 (1) (2013), pp. 77-85.

[16] C. Li, Z. Zhu Research and application of a novel hybrid air quality early-warning system: a case study in China Sci Total Environ, 626 (2018), pp. 1421-1438.

[17] C. Wang, M. Shao, Q. He, Y. Qian, Y. Qi Feature subset selection based on fuzzy neighborhood rough sets Knowl-Based Syst, 111 (2016), pp. 173-179.

[18] S. Wang, Q. Li, H. Yuan, D. Li, J. Geng, C. Zhao, et al. δ-Open set clustering a new topological clustering method WIREs Data Mining Knowl Discov, 8 (6) (2018).

[19] S. Yu, S. Chu, C. Wang, Y. Chan, T. Chang Two improved k-means algorithms Appl Soft Comput, 68 (2018), pp. 747-755.

[20] Q. Zhang, L.T. Yang, Z. Chen, P. Li High-order possibilistic c-means algorithms based on tensor decompositions for big data in IoT Inf Fusion, 39 (2018), pp. 72-80

[21] Majumdar J, Udandakar S, Bai BM. Implementation of cure clustering algorithm for video summarization and healthcare applications in big data. In: Emerging Research in Computing, Information, Communication and Applications. Singapore: Springer; 2019. p. 553–64.

[22] https://www.kaggle.com/rohanrao/air-quality-data-in-india

[23] Ghufran Isam Drewil, Riyadh Jabbar Al-Bahadili, Air pollution prediction using LSTM deep learning and metaheuristics algorithms, Measurement: Sensors Volume 24, December 2022, 100546.

[24] Abdellatif Bekkar, Badr Hssina, Samira Douzi & Khadija Douzi, Air-pollution prediction in smart city, deep learning approach, Journal of Big Data volume 8, Article number: 161 (2021)

[25] Qiuju Xie, Ji-Qin Ni, Enlin Li, Jun Bao, Ping Zheng, Sequential air pollution emission estimation using a hybrid deep learning model and health-related ventilation control in a pig building, Journal of Cleaner Production, Volume 371, 15 October 2022, 133714

[26] Gopalakrishnan V (2021) Hyperlocal air quality prediction using machine learning. Towards data science. https://towardsdatascience.com/hyperlocal-air-quality-prediction-using-machine-learning-ed3a661b9a71.

[27] Sanjeev D (2021) Implementation of machine learning algorithms for analysis and prediction of air quality. Int. J. Eng. Res. Technol. 10(3):533–538

[28] Castelli M, Clemente FM, Popovi˘c A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in California. Complexity 2020(8049504):1–23. https://doi.org/10.1155/2020/8049504

[29] Doreswamy HKS, Yogesh KM, Gad I (2020) Forecasting Air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Comput Sci 171:2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

[30] Liang Y, Maimury Y, Chen AH, Josue RCJ (2020) Machine learning-based prediction of air quality. Appl Sci 10(9151):1–17.

[31] Madan T, Sagar S, Virmani D (2020) Air quality prediction using machine learning algorithms–a review. In: 2nd international conference on advances in computing, communication control and networking (ICACCCN) pp 140–145.

[32] Monisri PR, Vikas RK, Rohit NK, Varma MC, Chaithanya BN (2020) Prediction and analysis of air quality using machine learning. Int J Adv Sci Technol 29(5):6934–6943.

[33] Nahar K, Ottom MA, Alshibli F, Shquier MA (2020) Air quality index using machine learning—a jordan case study. COMPUSOFT, Int J Adv Comput Technol 9(9):3831–3840.