Forecasting Air Pollution Index in relation to Stubble Burning in Punjab using Machine Learning and Genetic Algorithm

[1] Dr. Rachhpal Singh, [2] Dr. Rupinder Singh, [3] Prabhjot Kaur

[1]AP , [2]Assistant Professor, [3]Assistant Professor
[1]Khalsa College, Amritsar,
[2]Khalsa College, Amritsar,
[3]Khalsa College, Amritsar,
[1] rachhpal_kca@yahoo.co.in, [2] rupi_singh76@yahoo.com, [3] prabhjotkaur@khalsacollege.edu.in

ABSTRACT

Tons of stubble is generated as residue after harvesting wheat and paddy agriculture crops. Farmers burn the stubble to dispose of it because there is not enough time for the next crop and take it as a quick and inexpensive solution. It harms our ecology and eco-system in numerous ways. For the survival of humanity, it is crucial right now in India to track and forecast the Air Quality Index (AQI). The most major and dangerous air contaminant in this area is particulate matter (PM). For making predictions, machine learning (ML) technology is more effective than earlier conventional methods. Numerous ML algorithms, such as Random Forest (RF), Support Vector Machine (SVM), classification methods, Regression analysis, etc., were widely used for maximum prediction. But here by using Random Forest with Genetic Algorithm (GA) a hybrid approach prediction takes place much better. In order to improve the output of the data-adaptive computation, GA was used. Presently, data on air pollution from the preceding five years have been analyzed and forecasted for a study on Punjab's key cities to estimate and forecast PM concentrations. It was examined how PM concentrations vary with the seasons and some air pollutants. Variable importance ranking (VIR) was used to assess the effectiveness of the presented model. Here, the main emphasis was on taking into account some of the data sets from important cities in Punjab for the prediction of ambient pollution and air quality by using machine learning with genetic algorithm. Various common metrics were used to compare the results of all the strategies.
Keywords: Particulate matter (PM), Air quality index (AQI), Correlation analysis, Machine learning (ML), Variable importance ranking (VIR), Random forests (RF), Support vector regression (SVR)

INTRODUCTION

Burning of agricultural residue inside fields during harvesting is known as stubble burning that is serious issues for a healthy ecosystem. Farmers do stubble burning, because this will cause a delay in sowing wheat or rice in their sessions as there is very little time available between harvesting and sowing. It means harvesting of one crop and sowing of next crop has three to four weeks' time window which is primary reason for this [1]. Although stubble burning is a worldwide concern, India is the world's greatest rice grower that originates this problem. Field fires deplete the soil's nutrients and pollute the ecosystem by adding more air pollution. Burning

paddy crops have a loss of nutritional values in millions of tones (59,000 t of nitrogen, 3.85 mt of organic carbon, 20,000 t of phosphorus and 34,000 t of potassium). By mixing smoke in the air with the emission of gases like nitrogen oxide, ammonia, and methane atmospheric pollution severely affects the air with stubble burning. These gases' release impairs lung function, exacerbates asthma and increases the risk of chronic bronchitis. As crop residue from burning caused Ozone pollution indirectly. The amount of organic gases in the air is a life support that crop burning disturbs. Many factors, such as the super-exploding population, modernization, deforestation, vehicle emissions, and industrialization cause pollution by releasing a variety of hazardous gases such as lead (Pb), sulphur dioxide (SO_2), ozone (O_3), carbon monoxide (CO) and nitrogen dioxide (NO_2) which raises the value of some particulate matter $PM_{2.5}$ and PM_{10} [2]. Airborne particles are suspended with minute liquid and solid liquid compositions. It contains a variety of substances, including SO_4, NO_3 and organic molecules [3]. The designation particulate matter (PM) for fine atmospheric environments with dimension values less than 2.5m indicates that $PM_{2.5}$ particles are the most and most harmful of all pollution particles [4]. Keep in mind that the $PM_{2.5}$ concentration is measured in g/m3 that intensely hazardous for humans and these particles quickly and profoundly erode along the alveolar wall of the lungs, irritate the lungs and impair lung function [5]. In addition to having a negative impact on asthma, lung inflammation and numerous cardiovascular disorders, $PM_{2.5}$ can also increase the risk of developing lung and skin cancer [6]. When tiny particles enter the lungs, they may target the respiratory system and lead to the development of the new coronavirus COVID-19 infection [7]. When the amount of pollution particles in the air is high enough, it can have a serious impact on people's health and quickly lead to life-threatening issues. According to research, these particulate materials may have an impact on human health [8]. In India and especially in North region i.e. Delhi, Haryana and Punjab, situation becomes worst in crop-burning seasons. There is air pollution 34% in Punjab, 32% in Haryana and 45% in Delhi as described by government meteorologists that will increased the Air Quality Index [9].

State and federal governments have made several decisions over the past few years to discourage stubble burning by outlawing it. The Indian Penal Code Section 188 on stubble burning was implemented in 1981 and made it an offence with severe punishment. Punjab produced over 180 lakh tones of paddy straw each year [10]. Recent years have seen an increase in the use of stubble by numerous industries for compressed feed stock in biogas plants, paper mills, power plants, card mills, worm farms, poultry litter, packing materials, thermal power plants co-firing, 2G ethanol plants feed stock, biomass power projects, industrial fuel boilers, WTE plants, etc. Table 1 is the burnt data on hectares by PPCB (Punjab Pollution Control Board) from 2018 to December 2022 (31.5% reduction than previous year).

Years	**2018**	**2019**	**2020**	**2021**	**2022**
Paddy stubble burn in lakh hectares	17.81	18.95	7.96	15.47	9.21

Table (1) - Burnt data (in lakh hectares) by PPCB

All air and soil contaminants were found to have significantly decreased in 2020, the pandemic year corona. In Punjab, there are currently 27,743,338 individuals breathing poisonous air because they are not adhering to the WHO's recommendations for air purification. With $PM_{2.5}$ levels predicted to be 303.4 g/m3, the Faridkot district in Punjab has the greatest air pollution and is therefore classified as seriously contaminated. Burning husk on fields depletes the soil's nutrients and lowers its fertility. Additionally, burning stubble generates a lot of heat, which seeps into the soil below and causes the loss of beneficial bacteria as well as moisture for the soil and air. A nationwide index called the AQI provides daily forecasts of air quality. It provides information on the ratio of clean to polluted air and the risks of breathing contaminated air [11] and complete diagram shown in figure 1 displaying the division of numerous color-coded categories representing various levels of health issues.

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0–50	Air quality is considered good, and air pollution poses little or no risk.
Moderate	51–100	Air quality may pose a moderate health risk, especially for those who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101–150	Members of sensitive groups, children and adults with respiratory and heart ailments, may experience health effects and should limit time spent outside. The general public is not likely to be affected.
Unhealthy	151–200	Everyone may experience health effects and should limit their outdoor activity; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201–300	Everyone may experience more serious health effects and should avoid outdoor activities, especially individuals with heart and breathing ailments, children, and older adults.

Figure - 1. AQI levels and impact on health

Table 2 is about the fire farm cases year wise in Punjab.

Years	**2018**	**2019**	**2020**	**2021**	**2022**
Satellite image data of farm fires by Government	51,766	52,991	36,765	21,921	17,542

Table (2) - Farm Fires cases in Punjab (By PPCB)

According to data published by air quality management commission in and around NCR area of India, Punjab has logged 12,112 (almost 80%) of the 15,461 cases that have been reported this season across the North Indian states and Madhya Pradesh. Haryana came in second with 1,813 cases, followed by Uttar Pradesh, Madhya Pradesh, 599, Rajasthan, 227, and Delhi, 5. With 1,397 instances, Ferozepur district in Punjab had the most stump burning incidences in 2021. Amritsar (1,195), Gurdaspur (1,090), and Moga are other districts (1,075). Table 3 shows various AQI level in various districts in Punjab is as (One day in the month of November 2022).

City	**Mullanpur**	**Ludhiana**	**Faridkot**	**Khanna**	**Amritsar**	**Jalandhar**	**Mohali**	**Patiala**	**Ropar**

AQI	382	206	198	184	181	170	168	162	153

Table (3) - AQI level in various Punjab's districts

But the remote sensing Centre of Punjab Agricultural University showed 14,558 farm fire incidents till 31st October 2022. With the aid of artificial intelligence and related technologies, AQI forecasting is made possible. There are some prediction parameters in auto learning techniques that function similarly to those in pure statistics. The use of Artificial Neural Networks (ANN), etc., for AQI prediction is growing in popularity [12]. However, combining machine learning with genetic algorithms offers a hybrid strategy for precise prediction. The forecasting output from GA has been more optimized to acquire some correct initial values with some threshold values that apply to speed up training. The goal of suggested work here is to predict the air pollution particle concentration in Punjab so that preventative measures based on a hybrid approach may be done to protect human lives. A complete view of AQI (courtesy by **https://www.iqair.com/in-en/india/punjab)** is as shown in Figure 2:

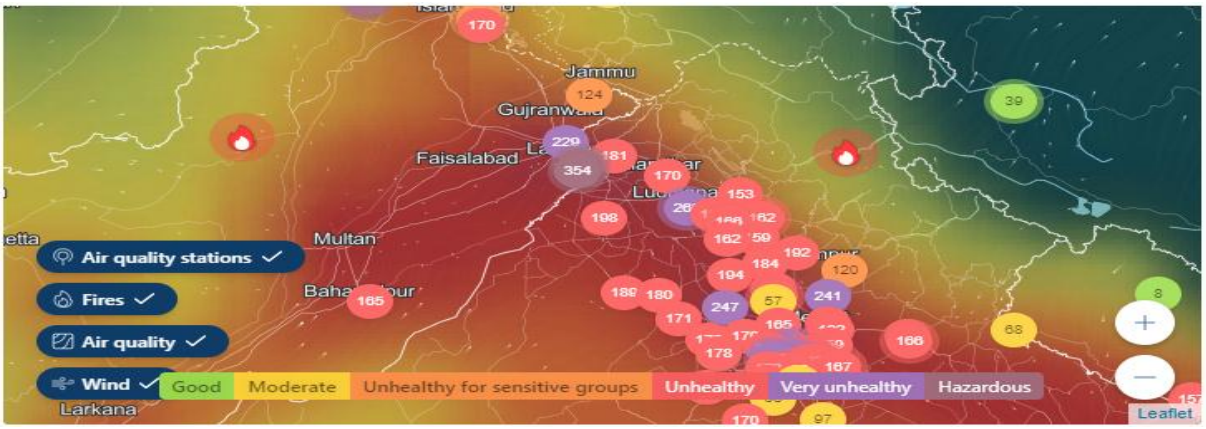

Figure – 2. A complete view of AQI (From iqair.com website)

LITERATURE REVIEW

Kumar et.al elaborated to generate the power from agriculture waste of paddy straw as biomass fuel that reduced the global warming and greenhouse effect with the help of PEDA (Punjab Energy Development Agency) [13]. Bellinger et.al surveyed the Europe, USA and China air pollution database by using machine learning and data mining techniques [14]. Rybarczyk and Zalakeviciute designed a hybrid model having traffic density correlation using air pollution database for getting maximum accuracy for controlling the pollution [15]. Sharma et.al analyzed Delhi city air quality data for observing the various pollution levels [16]. Sweileh et.al analyzed air pollution and health related literature by studying number of Scopus papers from year 1990 to year 2017 [17]. Dua et.al evaluated the air pollutants data for forecasting from Delhi area and presented air pollution real time online prediction system [18]. Kumar and B. P. Pande tarnished machine learning technique for finding Indian cities pollution SO_2 concentration prediction in Maharashtra state's environment and did conclusion of highly polluted Indian cities [19]. Mahalingam et.al intended a AQI prediction model using Support Vector Machine of different

Indian cities to find high accuracy [20]. Singh et.al classified and predicted model for atmospheric pollution using RPART and C5.0 supervised ML algorithms by collection data from ITO station, Delhi [21]. Castelli et.al designed forecasted model for finding air quality using different pollutants and particulate levels in California with machine learning Support Vector Regression *[22]*. Bamrah et.al computed AQI by considering various concentration pollutants levels using machine learning and many regression methods showing 81% accuracy [23]. Kumar et.al forecasted $PM_{2.5}$ concentration levels in Delhi regions by using regression with time series examination using various parameters to get the efficacy in the proposed system [24]. Harishkumar et.al inspected machine learning prediction ways with predicted and actual values by taking air pollution database of Taiwan [25]. Liang et.al deliberated Taiwan's AQI data for prediction by using various classifiers of machine learning [26]. Madan et.al compared number of pollutants data by using machine learning by considering various parameters and predicted air pollution accuracy [27]. Madhuri et.al found air pollutants concentration levels with machine learning RF approach [28]. Monisri et.al composed data related with air pollution from different resources by developing air quality predictions model [29]. Patil et.al retrieved machine learning AQI data for forecasting and modeling using Linear Regression, Logistic Regression and Artificial Neural Network [30]. Chhapariya et.al identified stubble burning database with machine learning and fuzzy approach in Punjab sites for identification and finding best classifier approach [31]. Sanjeev examined pollutants datasets and predicted air quality with Random Forest classifier [32]. Arif et.al summarized forest fire happing and did prediction for detection in burned areas by using ML approaches [33]. Barthwal et.al projected a model for forecasting PM concentrations at NCR locations and evaluated variable importance ranking (VIR) for finding root mean square error, absolute mean error and mean error [34]. Kaur et.al perceived air quality data of Indian cities and predicted AQI using data visualizations, correlation and statistical outliers with ML approach [35]. Pardasani and Raghav investigated impact of stubble burning in Punjab on NCR area with multiple regression analysis [36]. Keil et.al examined the various burn data practices and advised farmers to follow no-burn techniques like 'Happy Seeder'. Also analyzed the cost management and identified all the influencing factors by adopting Happy Seeder [37]. Sangwan and Deswal studied $PM_{2.5}$ modeling by comparing artificial intelligence methods like Artificial Neural Network, Random Forest and Support Vector Machine on various pollutants parameters by considering Rohtak area during stubble burning [38]. Pant et.al focused on AQI prediction using supervised ML techniques in Dehradun having pollutants like SO_2, PM_{10}, NO_2, $PM_{2.5}$ etc. and found the 98.63% of accuracy in prediction [39]. Aruna et.al simplified stubble aggregation task and disposal task [40].

METHODOLOGY

Today's agricultural areas are under threat from air pollution, which is having a severe impact on several Indian cities, particularly in the Punjab region. Increased AQI has a negative impact on health and hinders economic progress in India. Increased industrial energy production, vehicle traffic, road and soil dust, power plants, open waste burning, waste incineration, etc. are major pollutant emitters. Data was obtained from the Punjab Pollution Control Board of India [41] for the research study from 2017 to 2022. The Punjab Pollution Control Board's air pollution data [42] are the subject of the current study. This dataset includes 10 characteristics with 29,531 examples from 05 different Punjab cities and contains observations from January 2017 to

December 2022. Brief descriptive statistics of the pollutants/particles showing AQI from available data-set is shown in Table 4 below.

Statistics	**Pollutants**									
	$PM_{2.5}$	**PM_{10}**	**NO**	**NO_2**	**NH_3**	**CO**	**SO_2**	**O_3**	**Toluene**	**Benzene**
Count	23,989	18,222	24,999	24,209	18,999	27,202	26,656	24,999	21,987	22,986
Mean	66,980	119,019	17,409	28,100	22,200	2100	15000	34,101	8100	3200
Standard Deviation	63999	91000	22989	23980	24879	7000	18932	22910	19878	15000
Minimum	0.040	0.010	0.020	0.010	0.010	0.255	0.011	0.011	0.599	0.119
50%	48000	95000	9999	21010	15999	0.890	9100	31000	2989	3000
Maximum	99999	1000	400	465	35100	17500	19300	25500	450.0	455.00

Table (4) - Statistics of different AQI and Pollutants in Punjab Pollution Control Board Dataset

Some more statistics values showing a relationship between AQI and Xylene pollutant are in Table 5.

Statistics	**AQI Value**	**Xylene pollutant**
Count	24,989	11,222
Mean	166,98	370,01
Standard Deviation	140.99	6100
Minimum	13000	133.0
50%	118.00	0.987
Maximum	2100.00	0.976

Table (5) - Statistics values of AQI with Xylene

Table 6 below illustrates an exact connection of AQI with each value from the provided dataset:

Pollutants	**Correlation values**
PM_{10}	0.81221

$PM_{2.5}$	0.65123
CO	0.67999
NO_2	0.52346
SO_2	0.51999
NH_3	0.25212
O_3	0.20000
Xylene	0.16557
Benzene	0.04123
Toulene	0.27897

Table (6) - Correlation or association between pollutants and AQI

It should be noted that many ML algorithms perform better if the data has a normal distribution, which aids in skewed distribution. So it's crucial for Skewness identification from present characteristics available from data-sets. These were mapped and some transformations done from this Skewness that are important for converting the distribution from skewed to normal. Figure 3 show that the characteristics of CO, Benzene, Xylene and Toluene have significantly skewed values.

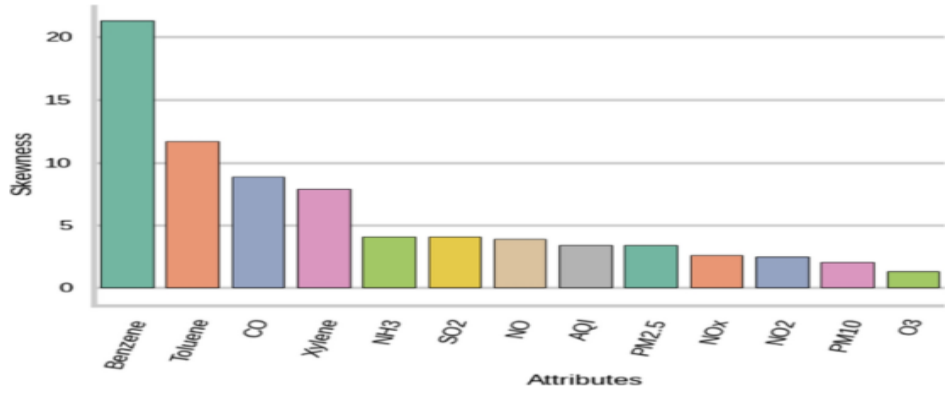

Figure – 3. Skewness having dataset features

A. Problem definition:

After accurate measurement, fine particles level found in air is a crucial component. Its level is also affected by other elements including solar illumination, wind speed and wind direction among others. Specific particles known as $PM_{2.5}$ are important in the prediction of pollution. Numerous approaches were used to find $PM_{2.5}$ concentration levels, but it a very difficult job to find such levels accurately due to their variable dependence and time-dependent behavior on a number of additional factors in Punjab and its surrounding areas, such as stubble burning and vehicle CO_2 emissions. Therefore, the primary task at a higher level performed time-based regression that has helpful continuous $PM_{2.5}$ prediction. It is dependent on prior $PM_{2.5}$ values or datasets as well as numerous meteorological characteristics that have time-series format records. The suggested strategy specifically addresses the issue of having accurate $PM_{2.5}$ levels in Punjab and the neighboring areas. Historical data is employed as the starting point for machine learning processes, which are then placed through a genetic approach to produce predictions for the

future. Pre-processing was done after the initial data collecting step, further applying analysis process and then discussed the machine learning (Feature selection) and genetic algorithm. These all are discussed as:-

i. Acquisition of Data (Data Collection)

The Punjab Pollution Control Board (PPCB), Patiala, official website was used to collect data for this model. PPCB operates independently of the Central Pollution Control Board of India but reports to it [43]. The information was in.csv format, with values separated by commas to illustrate numerous features and to remove unnecessary information from the entire database file.

ii. Pre-processing of Data (Data Cleaning)

By pre-processing data, using any of the most popular machine learning techniques produces superior results. Therefore, several approaches are implemented in this suggested approach for the observation of outlier data movement, which is extremely beneficial for viewing data relevance and serves as a decision-maker for $PM_{2.5}$ values concentration [44]. Extreme values can occasionally occur as a result of inaccurate readings, inaccurate data gathering, or incorrect detection that was missed before final data processing. Additionally, sample data with missing values was eliminated to ensure accurate prediction. Remember that these characteristics were extremely rare and insignificant. The most important prerequisite depends on the quality of the data for efficient machine learning using genetic hybrid models that is the production of an effective visualization. Preprocessing measures that improve the processing speed with general capabilities for ML and its related hybrid approaches can eliminate or lessen any type of noise or error existing in the acquired data. The two most frequent mistakes for monitoring applications following data extraction are missing data and outliers. Filling out non-required data, modifying outlier data, eliminating outlier data, and other operations are frequently performed on the existing data for data preparation procedures.

iii. Analysis of Data

After data collection, error correction, and pre-processing (data cleaning), time-series analysis used for further analyses the data. Additionally, each characteristic has an overall impact on the readings for $PM_{2.5}$, PM_{10} etc. during analysis [45]. An analysis of different PM concentrations in various Seasons from the given data is discussed as:

a. PM concentrations Seasonality

Due to climatic and geographic considerations, Punjab region PM levels based on seasonal patterns. As according to various levels of risk, full year pollutant concentrations were investigated in three seasons. Winters (20 October to 28 January), spring and summer (29 January to 21 June) and monsoons are among them (22 June to 19 October). The forecast models are created using daily average $PM_{2.5}$ and PM_{10} concentrations from Ludhiana and Patiala for the period of January 2021 and January 2022. Table 7 displays the PM time-series having minimum, maximum, range, median, mode, standard deviation and mean.

Parameters	Maximum Value	Minimum Value	Mean	Median	Mode	Standard Deviation	Range
$PM_{2.5}$	722.22	4.9	122.33	92.11	51.22	102.12	710.99
PM_{10}	991.12	12.99	244.43	211.44	111.11	149.88	923.32

Table (7) - PM concentration's basic statistics from 1st January 2021 and 31st January 2022 winter session)

According to Table 8, the highest daily average PM_{10} value for the spring season is observed on May 1, 2022, while the peak daily average $PM_{2.5}$ value is observed on February 1, 2021. The mean daily $PM_{2.5}$ and PM_{10} concentrations of 97.96 g/m3 and 276.05 g/m3 respectively are much higher than levels recommended by WHO. Median PM concentrations values for the two seasons are 89.98 g/m3 and 250.44 g/m3 respectively.

Parameters	Maximum Value	Minimum Value	Mean	Median	Mode	Standard Deviation	Range
$PM_{2.5}$	322.22	14.9	97.96	89.98	41.21	47.12	234.98
PM_{10}	891.11	88.91	276.05	250.44	98.12	119.81	897.12

Table (8) - PM concentration's basic statistics from 31st January 2021 and 1st May 2022 spring session)

The fundamental data on PM variations during monsoons is presented in Table 9. The period from the first day of July 2022 to the fourteenth of August 2022 has the lowest average PM values of the three seasons.

Parameters	Maximum Value	Minimum Value	Mean	Median	Mode	Standard Deviation	Range
$PM_{2.5}$	190.23	4.9	49.31	39.91	40.02	37.32	184.18
PM_{10}	471.12	15.99	146.19	112.41	95.22	98.11	449.92

Table (9) - PM concentration's basic statistics from 1st July, 2022 and 14th August, 2022 (Monsoon session)

b. Relationship of Environmental parameters and PM values

A region's PM levels are known to be influenced by meteorological factors including relative humidity (RH), rainfall (RF), temperature (TEMP), solar radiation (SR), wind direction (WD), wind speed (WS), atmospheric pollutants and atmospheric pressure (AP) like sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO) and ozone (O_3). The prediction models in this work use meteorological as well as some pollutant inputs as predictor variables to more accurately anticipate future PM concentrations. Table 10 lists the fundamental statistics of the 11 parameters that this study's forecast model used as inputs.

S#	Environmental Parameters	Maximum value	Minimum value	Mean Value	Standard Deviation
1.	**RF**	955	0	65.88	155.98
2.	**RH**	98.09	22.23	62.78	15.16
3.	**TEMP**	42.89	12.12	27.33	6.20
4.	**WS**	8.1	0.1	1.1	0.80
5.	**SR**	333.3	9.9	88	44
6.	**WD**	276.9	66.9	171	44.45
7.	**CO**	6.9	0.33	2.00	0.77
8.	**AP**	749.99	722.11	711.22	7.22
9.	**O_3**	450	1.1	55	65
10.	**SO_2**	111.11	6.6	33.11	12.34
11.	**NO_2**	121.1	11.11	55.55	22.23

Table (10) - Statistic values for predicted variables (From 02nd January 2021 to 28th February 2022)

For better AQI prediction, air pollutants like $PM_{2.5}$, PM_{10}, CO, NO_2, O_3, SO_2, etc. were analyzed by applying a hybrid approach of machine learning method with genetic technique. Figure 4 show the complete methodological process using ML with GA.

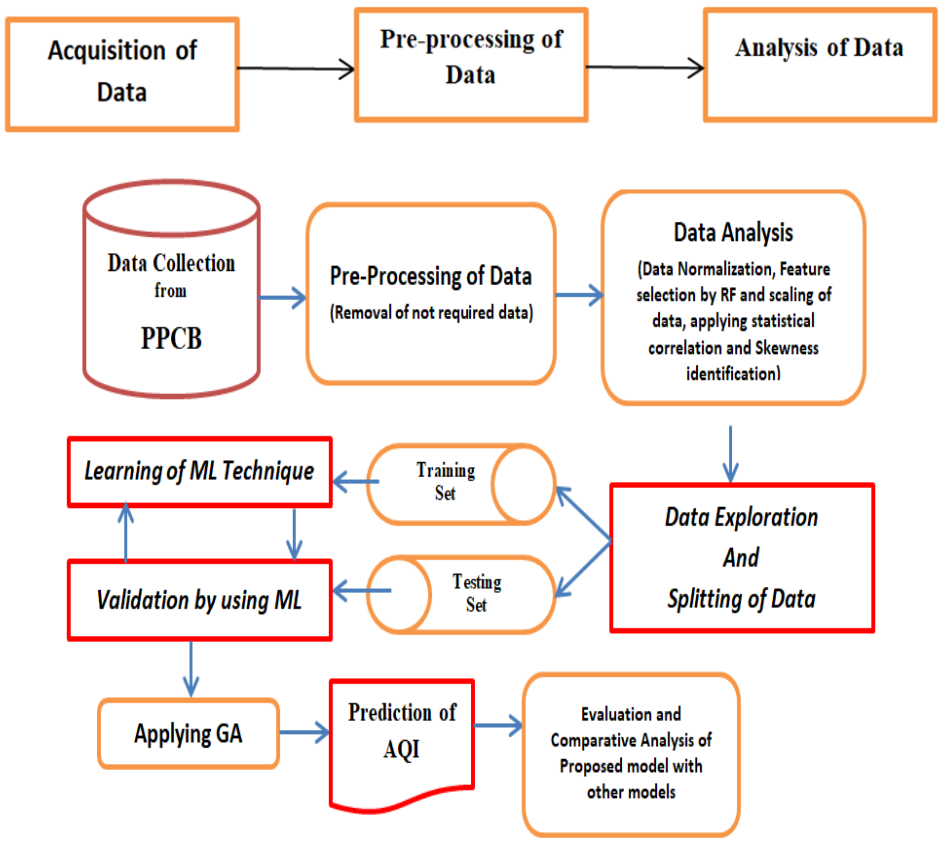

Figure – 4. A complete methodology of ML with GA

iv. Feature selection

For the objective of air quality forecasting and alerting the public, the PPCB dataset was examined for AQI-specific parameters. AQI classified in six standards as: first is acceptable (0 to 50), second is tolerable (51 to 100), third is moderate (101to 200), fourth is poor (201 to 300), fifth is extremely poor (301 to 400) and last sixth is severe having range 401 to 500 that was published by National Ambient board. Studies in this field revealed that by dropping some input variables and decreasing computational costs predictions accuracy can be improved. In the current study, a feature selection method employing correlation was used to compare each pair of initial and final variables in order to determine best values for pollutants as input variables. Be aware that many machine learning approaches are extremely sensitive to these outliers. Correlation study between AQI features and other contaminants feature value sets was performed for the selection of significant features.

Relationship of Environmental parameters forecasting by Random Forests

By comparing ANN or SVR with Random Forests (RFs), it is found that RFs are better suited for prediction because they are completely non-parametric and do not require knowledge of various input parameters distribution. Further to represent non-linear correlations between available classes and features that account for missing values, only Random Forests can accept both category and numerical inputs. Also RFs are favored due to explicit values and understandable simple regression structure. The Random Forest technique entails creating a collaborative regression tree values from training data and allowing them to select for forecasting. Random vectors control how each tree in the ensemble grows. For the nth regression tree, a random vector "n" is created that has the same distribution as the prior vectors "1," "...," and "n_1," but is unrelated to them. The training set is utilized to grow a tree together with n. This generates a vast number of trees to select most well-liked class. Regression/classification is decision tree examples used for handling continuous data values. Note that any individual decision tree look like tree structure having some attribute test that run on every node of tree. Every terminal node has feature label and each branch conveys the test result. The root node contains all input data at the beginning. The data set is then split into child nodes using a number of splitting variables. Entropy, expected entropy and information gain are calculated by the decision tree method to assess splitter from input variable values and also to decide for further split of inputted nodes. The following steps are detailed instructions for creating an RF model:

Algorithm

Step 1: Consider the forest contain t trees.

Step 2: Using seasonal data set values as,

Y_j (j=1,...,t) that was the boot-strap samples and obtained by row-sampling. Every sample having p prediction values randomly obtained from p predictors by column-sampling.

Step 3: The fresh training dataset Y_j is used to create a decision tree C_i (Y). By dividing each node into two smaller nodes and maximizing information gain, best split-point values and variables among the p predictors are selected.

Step 4: Last prediction is done by combining the predictions using K values of various decision trees as in formula -

$$\mathcal{F}_{RF}(Y) = \frac{1}{K}\sum_{i=1}^{K} C_i(Y)$$

v. Genetic Algorithm (GA)

GA [46] is an improved method that simulates hybridization, reproduction and mutation using a mechanism based on natural selection and population inheritance. The use of bio-inspired operators for efficient solutions to search and optimization issues comes from inheritance and natural selection processes. Individuals in GA have a potential remedy that was "chromosome" encoded. The population of the further solution domain includes every potential individual. A fitness function with a workable solution is computed from this population. Individual selection is produced in the following generation utilizing the specified fitness function after each individual's fitness value has been evaluated using the predetermined fitness function. Selection aims to keep the strong and eliminate the weak. By utilizing two additional operations crossover and mutation these chosen individuals' values further developed a new generation set of values. Individuals from the new generation inherit the good value sets from the old generation and they do better overall than the old generation, which was progressively progressing towards the best possible outcome. GA process is as shown in figure 5.

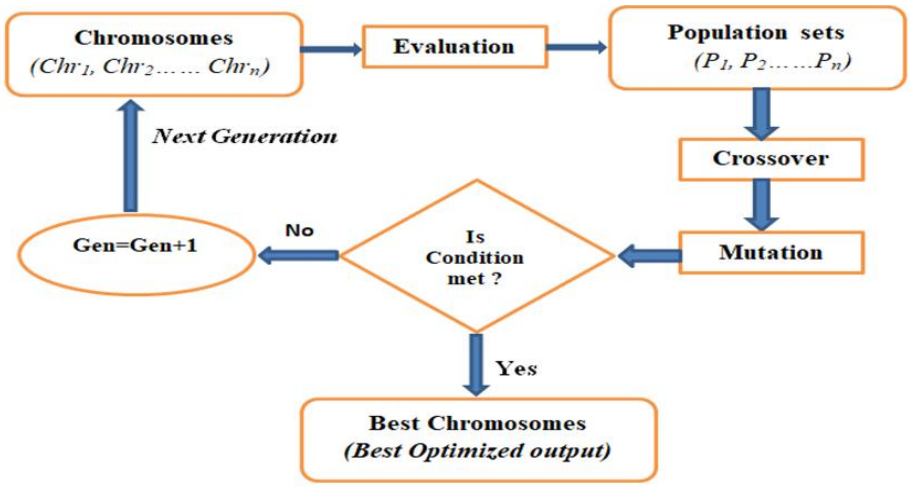

Figure – 5. A complete GA process

B. VIR (Variable Importance Ranking)

The significance of explanatory factors is graded according to how well they predict the response variable in order to gain insights into the prediction models. For the purpose of ranking the input variables some of permutation variable values having important values are processed here. VIR provides comprehensive description of how a feature is used by a forecast model and how it influences the model's predictions.

EXPERIMENTAL RESULTS WITH DISCUSSION

Empirical analysis was done and Experimental setup was discussed here for forecasting AQI values based on airborne contaminants. Before evaluating ML models with GA approach dataset having air pollution data is separated into two subsets as training subset having 75% weightage and testing subset having 25% weightage. Python programs executed Google pro cloud environment that has Tesla P50-PCIE-8 GB with Intel(R) processor with 1.99GHz, 16GB RAM and storage space 256 GB SSD. Some Python libraries as Seaborn, NumPy, Scikit-learn, Pandas etc. are considered for data processing. The dataset is then investigated with the goal of determining AQI overall values in relation to those contaminants that significantly contribute to increasing the AQI value. The construction of ML-based AQI proposed approach is explained in methodology that finds effectiveness of AQI forecasting. Classes are split unevenly because of some missing values in the target attribute, AQI Bucket. The imbalanced datasets issue is often ignored by ML models, which might result in subpar classification and prediction performances. Synthetic Minority Oversampling Technique (SMOTE) was used to address data imbalance issue. Instead of making duplicates of already existing items for minority classes, this strategy uses an algorithm that synthesizes new elements from scratch. It works by selecting a point at random from minority class and further calculating k-nearest neighbor distances. Between selected point and its neighbors, freshly made synthetic points were inserted for better results. Here Python module used called imbalanced-learn to develop SMOTE for class imbalance. The AQI level has now been predicted using both the SMOTE and non-SMOTE resampling techniques using four well-known ML models: SVR, RF and ANN with proposed hybrid model RGA. The outcomes of the employed ml models based on precision, accuracy, f1-score and recall in training period was elaborated in table 11 and during testing phase in table 12. While recall is the percentage of relevant examples that have been recovered, precision indicates the percentage of relevant instances that are present in the retrieved instances. The ratio of accurately identified attributes to the entire set of variables is known as accuracy. A weighted average of recall and precision is the F1-score. Be aware that the SVR model had the lowest accuracy, while the RGA model had best accuracy.

Method	Accuracy	F1 score	Recall	Precision	Training time
SVR	80	87	92	88	0.255
ANN	87	90	87	92	0.100
RF	92	90	94	96	0.530
RGA	93	99	88	98	0.101

Table (11) - Comparison of model results in the training set

Precision shows the proportion of appropriate instances that are available in recovered instances, whereas recall shows the percentage of relevant examples that have been recovered. Accuracy is defined as proportion of qualities that were correctly identified to all other variables. F1-score is evaluated as weighted average of precision and recall values. Be mindful that while the RGA model had the best accuracy, the SVR model had the lowest in testing phase also.

Method	Accuracy	F1 score	Recall	Precision	Prediction time
SVR	77	82	89	89	0.030
ANN	85	91	90	91	0.020
RF	90	93	96	97	0.039
RGA	92	95	97	88	0.018

Table (12) - Comparison of various methods outputs in testing set

The forecast models are constructed using 11 inputs (RF, RH, TEMP, SR, WS, WD, AP, CO, O_3, SO_2 and NO_2) as explanatory variables. The following paragraphs go into further depth on the anticipated outcomes for the suggested models. Shrot-term (7-day) forecasts are created using the forecast models. The $PM_{2.5}$ and PM_{10} levels for 7 days of the same season are predicted using seasonal concentrations. The actual and forecasted PM levels for each season are shown in Figure 6, Figure 7 and Figure 8.

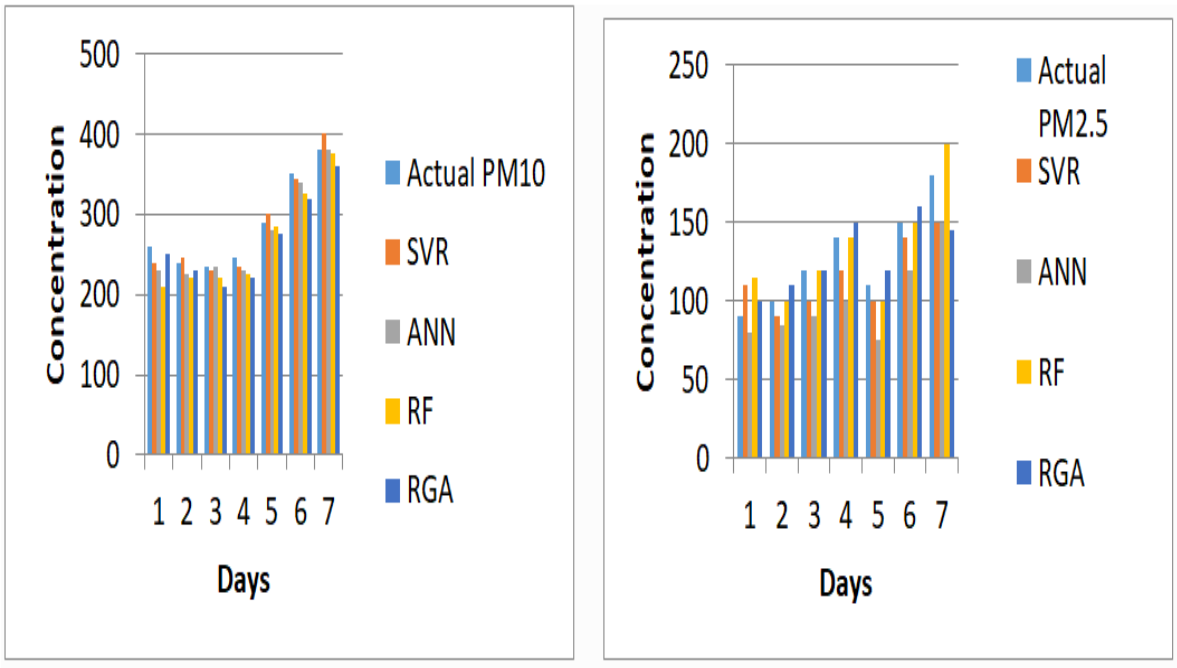

Figure – 6. Forecast of PM_{10} and $PM_{2.5}$ Level (7 days Monsoon Data)

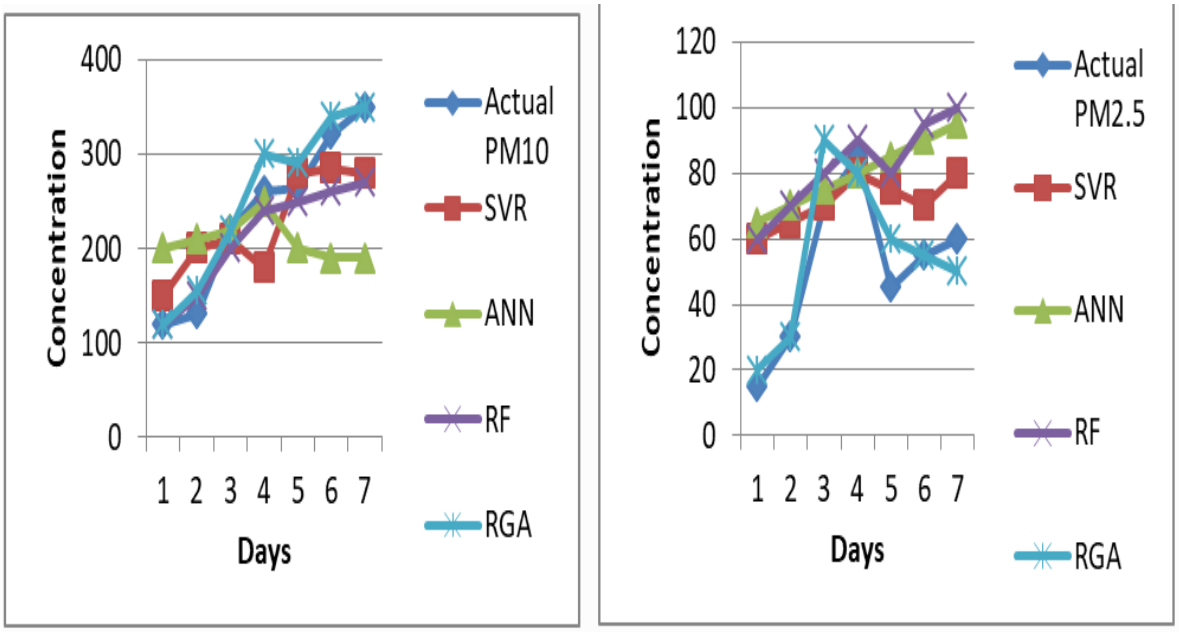

Figure – 7. Forecast of PM_{10} and $PM_{2.5}$ Level (7 days Spring-Summer Data)

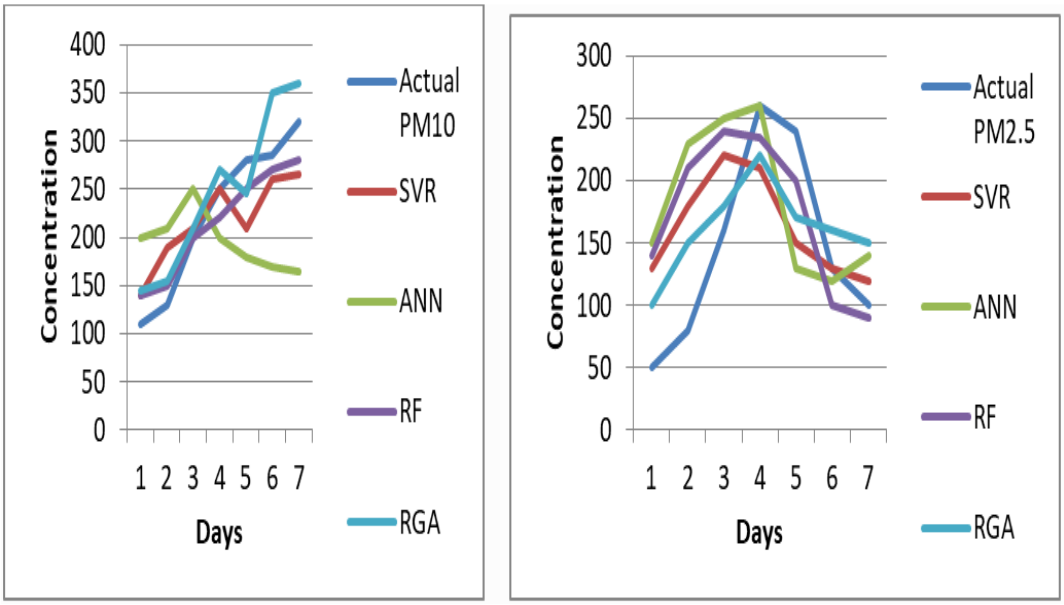

Figure – 8. Forecast of PM_{10} and $PM_{2.5}$ Level (7 days Winter Data)

Predictions of seven days by proposed approach for $PM_{2.5}$ and PM_{10} concentrations using the Monsoon data-set as input are shown in Figure 6. It is observed that RGA performs better as compared to other models for PM level predictions in Monsoons. Forecasting of PM levels in Summer-Spring session is as shown in Figure 7. RGA and RF perform better than other models $PM_{2.5}$ and PM_{10} levels in spring-summer sessions. Figure 8 shows the seven days predictions for PM levels in winter sessions. Note that for all seasonal outcomes RGA and RF exhibit best while ANN is at lowest level.

Applying Variable importance ranking (VIR)

Understanding how each explanatory variable affected the creation of the forecast model is possible with VIR [47]. Average of all-season values relevance to every value of eleven explanatory value sets (RH, RF, SR, TEMP, WD, WS, CO, AP, O_3, SO_2 and NO_2) for the RGA model for the $PM_{2.5}$ and PM_{10} concentrations is displayed in Figure 9 because it is the model that performs the best. The most crucial input factor for PM prediction according to the VIR analysis of the RGA approach for the $PM_{2.5}$ and PM_{10} data sets is CO followed by NO_2 and atmospheric pressure (AP). The weakest parameters are rainfall (RF), relative humidity (RH), wind direction (WD) and SO_2. Rainfall is the least significant explanatory factor for the $PM_{2.5}$ data set whereas relative humidity is for the PM_{10} data set. VIR analysis for RGA publicized that carbon monoxide (CO) is most important input feature in prediction of PM levels, followed by nitrogen dioxide (NO2) and atmospheric pressure (AP).

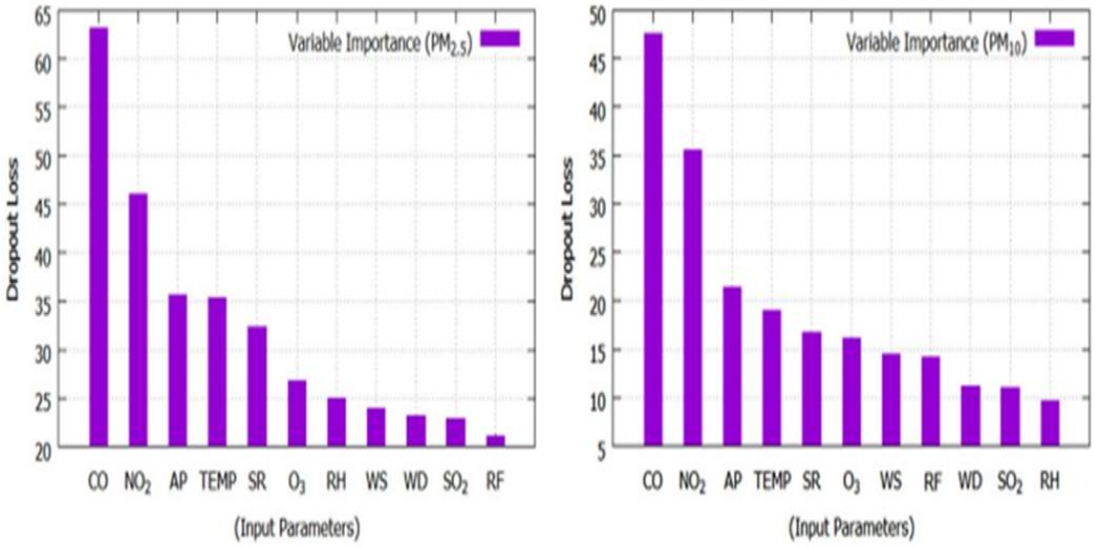

Figure – 9. VIR for Air Pollutants and meteorological parameters

CONCLUSION

Due to unpredictable nature of contaminants, dynamic environment and fluctuation in time and space prediction of air quality is very difficult. A constant monitoring and analysis of air quality is required in developing nations because of serious effects of air pollution on people, plants, animals, climate, environment and historical sites. The AQI prediction in India has drawn little attention by researchers. Here some cities of Punjab's air pollution data of five years were investigated. After filling all NAN dataset values, then resolving outliers afterwards normalization of data values by cleaning initially and then preprocessed. AQI-affecting contaminants are filtered for further research using correlation approach with feature selection method and skewed features are transformed. Outcomes of ML approaches for train-test data sub-sets are discussed with some metrics like F1-Score, precision, accuracy and recall. Proposed hybrid RGA achieved highest accuracy and SVR has lowest accuracy using train-test sets. Further work examined with RGA, SVR, ANN and RF techniques for forecasting PM_{10} and

$PM_{2.5}$ concentrations. Using air pollution parameters and meteorological as input variables forecasting accuracy was improved. With variable importance ranking significance of explanatory factors helpful for creating forecast techniques. Eleven explanatory variables and three seasonal data-sets are computed for prediction and analysis of PM_{10} and $PM_{2.5}$ levels considering seven days forecasting using proposed approach. AQI prediction can be forecasted by using deep learning methods.

REFERENCES

[1] Abdurrahman, Muhammad Isa, Sukalpaa Chaki, and Gaurav Saini. "*Stubble burning: Effects on health & environment, regulations and management practices*," Environmental Advances 2 (2020)

[2] Singh, Jabrinder, Naveen Singhal, Shailey Singhal, Madhu Sharma, Shilpi Agarwal, and Shefali Arora. "*Environmental implications of rice and wheat stubble burning in north-western states of India*," In Advances in Health and Environment safety, pp. 47-55. Springer, Singapore, 2018

[3] Pöschl, Ulrich. "*Atmospheric aerosols: composition, transformation, climate and health effects*," Angewandte Chemie International Edition 44, no. 46 (2005): 7520-7540.

[4] Kam, Winnie, Kalam Cheung, Nancy Daher, and Constantinos Sioutas. "*Particulate matter (PM) concentrations in underground and ground-level rail systems of the Los Angeles Metro*," Atmospheric Environment 45, no. 8 (2011): 1506-1516.

[5] Shaughnessy, William J., Mohan M. Venigalla, and David Trump. "*Health effects of ambient levels of respirable particulate matter (PM) on healthy, young-adult population*," Atmospheric environment 123 (2015): 102-111.

[6] Feng, Shaolong, Dan Gao, Fen Liao, Furong Zhou, and Xinming Wang. "*The health effects of ambient PM2. 5 and potential mechanisms*," Ecotoxicology and environmental safety 128 (2016): 67-74.

[7] Shereen, Muhammad Adnan, Suliman Khan, Abeer Kazmi, Nadia Bashir, and Rabeea Siddique. "*COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses*," Journal of advanced research 24 (2020): 91-98.

[8] Ali, Muhammad Ubaid, Siyi Lin, Balal Yousaf, Qumber Abbas, Mehr Ahmed Mujtaba Munir, Audil Rashid, Chunmiao Zheng, Xingxing Kuang, and Ming Hung Wong. "*Pollution characteristics, mechanism of toxicity and health effects of the ultrafine particles in the indoor environment: Current status and future perspectives*," Critical Reviews in Environmental Science and Technology 52, no. 3 (2022): 436-473.

[9] Nagar, Pavan K., and Mukesh Sharma. "*A hybrid model to improve WRF-Chem performance for crop burning emissions of PM2. 5 and secondary aerosols in North India*," Urban Climate 41 (2022): 101084.

[10] Kaur, Hardeep, and Manvendra Singh. "*An Assessment of Environmental Pollution and Policy Initiatives in Punjab, India: A Review*," (2022)

[11] Tiwari, Rajesh Kumar, and Tapan Kumar Dey. "*A Novel Approach for Analysis of Air Quality Index Before and After Covid-19 Using Machine Learning*," In Applications of Artificial Intelligence and Machine Learning, pp. 139-149. Springer, Singapore, 2022

[12] Jaiswal, Harsh, Aman Singh, Akhilesh Chauhan, and Dicksha Sharma. "*Air Quality Index Prediction using Machine Learning*," (2022)

[13] Kumar, Parmod. "*Energy Generation by Use of Crop Stubble in Punjab*," In Climate Change Challenge (3C) and Social-Economic-Ecological Interface-Building, pp. 507-518. Springer, Cham, 2016

[14] Bellinger, Colin, Mohomed Shazan Mohomed Jabbar, Osmar Zaïane, and Alvaro Osornio-Vargas. "*A systematic review of data mining and machine learning for air pollution epidemiology*," BMC public health 17, no. 1 (2017): 1-19.

[15] Zalakeviciute, Rasa, Yves Rybarczyk, Jesús López-Villada, and Maria Valeria Diaz Suarez. "*Quantifying decade-long effects of fuel and traffic regulations on urban ambient PM2. 5 pollution in a mid-size South American city*," Atmospheric Pollution Research 9, no. 1 (2018): 66-75.

[16] Sharma, Nidhi, Shweta Taneja, Vaishali Sagar, and Arshita Bhatt. "*Forecasting air pollution load in Delhi using data analysis tools*," Procedia computer science 132 (2018): 1077-1085.

[17] Sweileh, Waleed M., Samah W. Al-Jabi, Sa'ed H. Zyoud, and Ansam F. Sawalha. "*Outdoor air pollution and respiratory health: a bibliometric analysis of publications in peer-reviewed journals (1900–2017)*," Multidisciplinary Respiratory Medicine 13, no. 1 (2018): 1-12.

[18] Dua, Radhika Dua, Divyam Madaan Madaan, Prerana Mukherjee Mukherjee, and Brejesh Lall Lall. "*Real time attention based bidirectional long short-term memory networks for air pollution forecasting*," In *2019* IEEE fifth international conference on Big Data computing service and applications (BigDataService), pp. 151-158. IEEE, 2019

[19] Kumar, K., and B. P. Pande. "*Air pollution prediction with machine learning: a case study of Indian cities*," International Journal of Environmental Science and Technology (2022): 1-16.

[20] Mahalingam, Usha, Kirthiga Elangovan, Himanshu Dobhal, Chocko Valliappa, Sindhu Shrestha, and Giriprasad Kedam. "*A machine learning model for air quality prediction for smart*

cities," In 2019 International conference on wireless communications signal processing and networking (WiSPNET), pp. 452-457. IEEE, 2019

[21] Singh, Anish, Raja Kumar, and Nitasha Hasteer. "*Comparative Analysis of Classification Models for Predicting Quality of Air*," In *2020 IEEE* 5th International Conference on Computing Communication and Automation *(ICCCA)*, pp. 7-11. IEEE, 2020

[22] Castelli, Mauro, Fabiana Martins Clemente, Aleš Popovič, Sara Silva, and Leonardo Vanneschi. "*A machine learning approach to predict air quality in California*," Complexity 2020 (2020)

[23] Bamrah, Sumneet Kaur, K. R. Saiharshith, and K. S. Gayathri. "*Application of Random Forests for Air quality estimation in India by adopting terrain features*," In *2020* 4th International Conference on Computer, Communication and Signal Processing (ICCCSP), pp. 1-6. IEEE, 2020

[24] Kumar, Saurabh, Shweta Mishra, and Sunil Kumar Singh. "*A machine learning-based model to estimate PM2. 5 concentration levels in Delhi's atmosphere*," Heliyon 6, no. 11 (2020): e05618.

[25] Harishkumar, K. S., K. M. Yogesh, and Ibrahim Gad. "*Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models*," Procedia Computer Science 171 (2020): 2057-2066.

[26] Liang, Yun-Chia, Yona Maimury, Angela Hsiang-Ling Chen, and Josue Rodolfo Cuevas Juarez. "*Machine learning-based prediction of air quality*," Applied Sciences 10, no. 24 (2020): 9151.

[27] Madan, Tanisha, Shrddha Sagar, and Deepali Virmani. "*Air quality prediction using machine learning algorithms–a review*," In *2020* 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN*)*, pp. 140-145. IEEE, 2020

[28] Madhuri, R., S. Sistla, and K. Srinivasa Raju. "*Application of machine learning algorithms for flood susceptibility assessment and risk management*," Journal of Water and Climate Change 12, no. 6 (2021): 2608-2623.

[29] Monisri, P. R., R. K. Vikas, N. K. Rohit, M. C. Varma, and B. N. Chaithanya. "*Prediction and analysis of air quality using machine learning*," Int J Adv Sci Technol 29, no. 5 (2020): 6934-6943.

[30] Patil, Miss Priti Ashok, and Ms Ashwini Salunkhe. "*Comparative analysis of construction cost estimation using artificial neural networks*," J Xidian Univ 14 (2020): 1287-305.

[31] Chhapariya, Koushikey, Anil Kumar, and Priyadarshi Upadhyay. "*A fuzzy machine learning approach for identification of paddy stubble burnt fields*," Spatial Information Research 29, no. 3 (2021): 319-329.

[32] Sanjeev, Dyuthi. "*Implementation of machine learning algorithms for analysis and prediction of air quality*," International Journal of Engineering Research & Technology (IJERT) 10, no. 3 (2021): 533-538.

[33] Arif, M., K. K. Alghamdi, S. A. Sahel, S. O. Alosaimi, M. E. Alsahaft, M. A. Alharthi, and M. Arif. "*Role of machine learning algorithms in forest fire management: a literature review*," J Robotics Autom 5, no. 1 (2021): 212-226.

[34] Barthwal, Anurag, Debopam Acharya, and Divya Lohani. "*Prediction and analysis of particulate matter (PM2. 5 and PM10) concentrations using machine learning techniques*," Journal of Ambient Intelligence and Humanized Computing (2021): 1-16.

[35] Kaur, Rajveer, and Puneeta Pandey. "*Air pollution, climate change, and human health in Indian cities: a brief review*," Frontiers in Sustainable Cities 3 (2021): 705131.

[36] Pardasani, Raghav. "*Rethinking Rice Cultivation: A Multiple Regression Analysis of Factors Influencing the Prevalence of Stubble Burning in Punjab*,", 2021

[37] Keil, Alwin, P. P. Krishnapriya, Archisman Mitra, Mangi L. Jat, Harminder S. Sidhu, Vijesh V. Krishna, and Priya Shyamsundar. "*Changing agricultural stubble burning practices in the Indo-Gangetic plains: is the Happy Seeder a profitable alternative?*," International Journal of Agricultural Sustainability 19, no. 2 (2021): 128-151.

[38] Sangwan, V., and S. Deswal. "*PM 2.5 modelling during paddy stubble burning months using artificial intelligence techniques*," Journal of Achievements in Materials and Manufacturing Engineering 110, no. 1 (2022)

[39] Pant, Alka, Sanjay Sharma, Mamta Bansal, and Mandeep Narang. "*Comparative Analysis of Supervised Machine Learning Techniques for AQI Prediction*," In 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA), pp. 1-4. IEEE, 2022

[40] Aruna, Ch, B. Sai Charitha, G. Mahitha, Ch Bhuvana, D. Bhargavi, and D. Lakshmi Mounika. "*Agri-Stubble Aggregation and Disposal*," Volume 5, Issue 6, June 2022

[41] Bhawan, Parivesh, and East Arjun Nagar. "*Central Pollution Control Board*," (2020)

[42] Kumar, Rajesh, Suresh K. Sharma, J. S. Thakur, P. V. M. Lakshmi, M. K. Sharma, and T. Singh. "*Association of air pollution and mortality in the Ludhiana city of India: a time-series study*," Indian journal of public health 54, no. 2 (2010): 98.

[43] Gupta, Shreekant, Shalini Saksena, and Omer F. Baris. "*Environmental enforcement and compliance in developing countries: Evidence from India*," World Development 117 (2019): 313-327.

[44] Liu, Hui, and Chao Chen. "*Prediction of outdoor PM2. 5 concentrations based on a three-stage hybrid neural network model*," Atmospheric Pollution Research 11, no. 3 (2020): 469-481.

[45] Biancofiore, Fabio, Marcella Busilacchio, Marco Verdecchia, Barbara Tomassetti, Eleonora Aruffo, Sebastiano Bianco, Sinibaldo Di Tommaso, Carlo Colangeli, Gianluigi Rosatelli, and Piero Di Carlo. "*Recursive neural network model for analysis and forecast of PM10 and PM2. 5*," Atmospheric Pollution Research 8, no. 4 (2017): 652-659.

[46] Mirjalili, Seyedali. "*Genetic algorithm*," In Evolutionary algorithms and neural networks, pp. 43-55. Springer, Cham, 2019

[47] Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre. "*Correlation and variable importance in random forests*," Statistics and Computing 27, no. 3 (2017): 659-678 .