International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

Machine Learning-Based IOT Air Quality and Pollution Detection

K. Siva Krishna[1], Dr Thatavarti Satish[2], Dr. Jyotirmaya Mishra[3]
[1]Research Scholor,Department of Computer Science and Engineering,GIET University, AP, India
[2]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[3]Department of Computer Science and Engineering,GIET University, AP, India
Email Id: drsatishthatavarti@kluniversity.in

Abstract

In India, gas leakage from the different factories harmful to human surveying in the last fifty years is very low. However, there is a lack of prior detection of the chemical gases detection system in the situation raised. So, In this regard, there is a gap identification of chemical gases intensity detection needed. In this work, the main objective is to identify chemical gases intensity and maintain the stream data in the database from different locations. To fill this gap, that is identifying the high-intensity chemical gases from the chemical gas disaster areas. The first step needs to identify the different chemical gases and natural gas compositions. In this regard in this work for design internet-based gases in the air system. So, the sensors MQ2(Ethanol i-Butane Methane Alcohol Gas Sensor Sensor), MQ3(Sensitivity Alcohol Detector), MQ4 (Methane and Natural Gas (CNG)), MQ-5 (LPG GAS SENSOR), MQ-7 (CO Gas Sensor Module Test Carbon Monoxide Detector), MQ-8 (hydrogen Gas Sensor), MQ-9 (carbon monoxide), MQ-135 Sensor(Air Quality Sensor Hazardous Gas Detector) and DHT11 Digital Temperature Humidity Sensors. These sensors are interfacing with the micro-control STM 32 board. It is also called one Pollution identification terminal by using it to pull the sensor stream data from location to centralized data. This stream data transportation is a service to pull the data. For this Data pulling, design an algorithm store into a cloud database. In this research work, design the electronic terminal with a wifi circuit using IoT technologies.Moreover, getting these attributes as data. Data need to apply the preprocessing techniques and extracted feature techniques also. This paper discusses mainly designing the terminal for pollution attributes , cleaning the data, and applying the Machine Learning based Feature extraction techniques.

Keywords:- Internet of things, Data as a Request, Cloud, Sensors, Pollution, Micro Controller,feature extraction and data cleaning

I. INTRODUCTION

The Internet of Things (IoT) is a vast network of autonomous and heterogeneous devices and sensors that send massive amounts of data to monitoring systems that analyze the data and make decisions. Smart Homes, Smart Grid, Public Safety and Environment Monitoring such as weather monitoring and water quality, Medical and Healthcare (Internet of Medical Things IoMT), Industrial Processing such as California (Cf) CoAP framework, Agriculture and Breeding such as Climate-Smart Agriculture (CSA), and connected vehicles (IoCV) are some of the applications that use IoT technology and devices [1, 2]. Nowadays, the use of these applications and their benefits play an important role in improving one's quality of life. As a result, the future of IoT devices, technology, and applications will shape our future. [3].

Over the past 130 times, the world has warmed by around0.85 °C. The situation calculated that between 2030 – 2050, climate change will beget over deaths worldwide and will add between€1.8-3.6 billion per time to health care costs. Air pollution in numerous overpopulated metropolises is far too high for mortal exposure, causing resides to wear face masks or original PPE all time round. In numerous metropolises, people may not indeed be apprehensive of the high situations of dangerous adulterants they expose to at certain times of the day or time. This leads to significant health pitfalls to help — contributors to climate change[4].

In this regard, the Internet of Thing Technology mingles to identify the impact of natural gas on position and mortal beings. Air pollution is the vacuity of substances in the air, which generates an advanced threat to a living being. Machine, real estate, and inorganic husbandry are the major contributors are in air pollution. Still, it can not deny that rainfall and climate change also play a significant part in Air pollution. The proposed System uses gas detectors(MQ135) and a microcontroller(STM 32) to cost the data. Thus, the Air Pollution Index(API) Reporting System can prove to be an effective tool for communication of probable pitfalls[5]. Figure 1 shows the airpolution in india year 2019

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

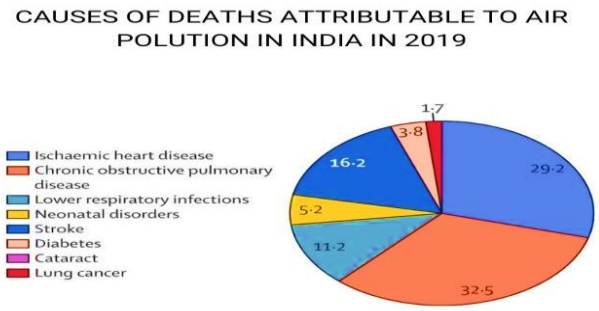

Figure 1.Airpolution in india year 2019

The report also goes into detail about the causes of death. According to the data, lung ailments were the leading cause of death in India as a result of air pollution. Chronic obstructive pulmonary disease (COPD) was responsible for 32.5 percent of all air pollution-related deaths. The two most common causes of death attributable to air pollution were ischaemic heart disease (29.2 percent) and stroke (16.2%)[7].

II. RELATED WORKS

In this literature, the survey finds out the alternative pollution detection system. Different authors present the circuit design and the data pulling techniques. From the last four years, improvements in this area are present in the survey.

[10] and colleagues This study installed and utilised data from 14 IoT emission sensors to construct machine learning forecast models for NO2 pollution concentration. For five months, the researchers used big data analytics technology to retrieve a vast quantity of data gathered in ten seconds. Weather data from the UK meteorology department and traffic data from the transport department were collected and integrated for the matching time and location where the pollution sensors are located. Findings – The results reveal that the hybrid BA-GS-LSSVM beats all existing solo machine learning prediction Models for NO2 pollution. [Sean Mc Grath] et.al. This paper develops an alternative, cheap, IoT-based air quality monitor, which can track air pollution in real-time and transmit the relevant Data rapidly through a low power vast area network. An extensive network of such monitors can generate a vast amount of data, which may then be processed and analyzed in the cloud in real-time and correlated with time of day, month, year, weather, and other factors.

[5] et al. investigated the ion (NCR) and forecasted air quality for the following day using linear regression as a machine learning approach. Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) are the four performance measurements used to evaluate the model (MAPE).

The study then evaluates with a benchmark model and finds significant results.[11] et al. discuss the to keep them all in check, the System will monitor it and provide us with a report on the climate in the area.The report can be accessed from any computer or mobile device.The System includes numerous sensors such as (temperature, humidity, noise, CO, rain gauge).

A rain gauge and geotagging system can substantially benefit the geological survey of the specific area because they are low-cost. Qussai Yaseen et al. discuss how the System is used to mitigate malicious and malfunctioning Internet of things[12]. As a result, the proposed System can improve the efficiency of Sensor systems such as smart cities and lower the risk of hostile IoT devices, particularly in sensitive systems such as military applications that rely on IoT devices. Furthermore, this paper presents a new identification approach for uniquely and worldwide identifying IoT devices wherever they roam. Furthermore, the research provides a novel method for evaluating the reputations of IoT devices and generating proper values based on these reputations. The results reveal that the suggested technique performs admirably in detecting malicious IoT devices and computing very close values to the actual values[13].

The above study is about the terminal designing and pulling the data and analysis used to preprocess the data and predict models representing it. In this paper, there are six sections divided, 1nad 2 are Introduction and relative work. The following sections are section 3 proposed model for circuit designing interfacing mechanism and proposed System architecture for data access from the terminal push into the cloud-based database. Section 4. Preprocessing and feature extraction methods with results and discussions Moreover, Section 5 Conclusion and Future Work.

III. PRAPOSAL MODEL AND DESIGNING AND INTERFACING OF THE POLLUTION DETECTION TERMINAL

In this methodology, there are three steps are there. The first one is interfacing hardware components. In this scenario, first prepare interfacing hardware as semantic diagram represents the interfacing sensors with the microcontroller board is STM32 board. Here each sensor required 3.2v to 5v of the voltage. Moreover, each sensor needs to interface ground connection for neutral and another connection retrieving the sensor data. In this diagram, each sensor represents data pins with PD_x (i.e., x represents the 1,2,3,4,5,67,8….etc) pins; data read from these pins connect with the microcontroller board same PINs.Each sensor

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

connected with VCC and Ground with a similar interface. After interfacing, the Second part is to write the software code for reading the Input and Output interfacing code with the embedded C. Reading the data sends the data from the sensors to the cloud database. Moreover, the third phase is that the service requires a REST API(Application Programme Interface) to push the data to the internet-based database in JSON(JavaScript Object Notation). Figure 2 shows semantic diagram from the STM32 board interfacing with different air quality sensors.

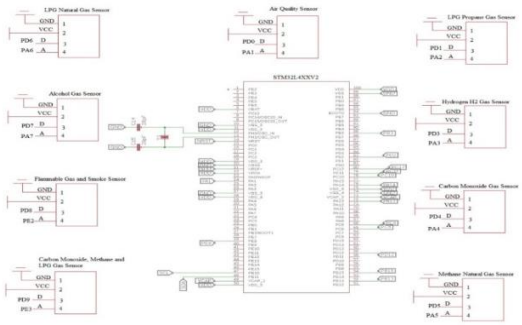

Figure 2.Semantic Diagram from the STM32 board interfacing with different air quality sensors.

Table 1.Different sensor ranges lower to higher value

Sensor Names	**Low Value(ppm)**	**High Value(ppm)**
LPG Natural Gas Sensor	**300**	**10000**
LPG Prapone Gas Sensor	**100**	**10000**
Alcohol Gas Sensor	**10**	**300**
Hydrogen H2 Gas Sensor	**0**	**1000**
Flammable Gas Sensor and Smoke Sensor	**200**	**10000**
Carbon Monoxide Gas Sensor	**20**	**2000**
Carbon monoxide methine	**300**	**10000**
Methane Natural Gas Sensor	**200**	**10000**
Carbon Monoxide, Methane, and LPG Gas Sensor	**10000**	**10000**

Table 1 shows the different sensor ranges lower to higher value.In this architecture, there are three phases. One is to get the data from the terminals and store it into the centralized database, and these databases need to preprocess the data. Then, apply the preprocessing techniques to process the data extract the features best-fitted features. Moreover, here produced the training data set for artificial intelligent models and for Train the models to predict the pollution levels. Figure 3 shows the system architecture for getting the data from and preprocessing techniques

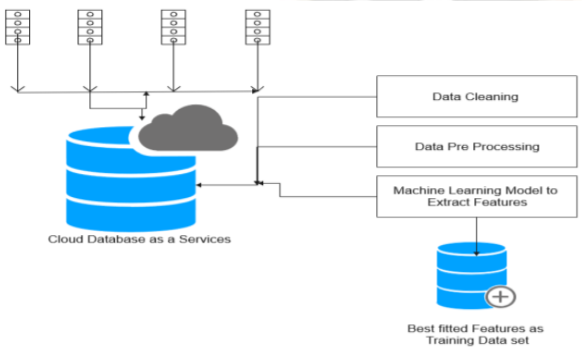

Figure 3. System Architecture for getting the data from and preprocessing techniques

IV. PREPROCESSING AND RESULT ANALYSIS

The efficiency of the supervised machine learning algorithms significantly depends on the accuracy of the data considered. So, it is mandated to preprocess the data before a model analyses it. The most puzzling issue in analytic ML is to identify and remove noise values [14]. In most of the cases, it is identified that the models are majorly deviating from attaining accuracy as the dataset possesses too many null values. These extremely deviating features are known as outliers. In addition, handling missing data values is another big challenge in data preprocessing steps [14].

In general, logical or symbolic learning algorithms are efficient at interpreting categorical data. However, the real-time datasets possess both numerical as well as illustrative data elements [15]. Therefore, it is tough to discretize continuous data and categorize the attributes of symbolic elements. The model's performance can be enhanced by properly identifying behaviours needed in obtaining decision trees and the respective decision rules [15].

In general, real-time data possess several characteristics and features. However, only some of these features are of high priority on which the output function mainly depends. Hence, the Data is primarily analyzed to find the essential features as well as interdependencies among them. Furthermore, it also enables in identifying and removing outdated and redundant data. This step helps decrease the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

dimensionality of the data and makes the learning model faster and more effective [14].

4.1 Data Set Attributes

This section mainly presents the dataset and its corresponding attributes. To extract the dataset, the data source used is Kaggle. The dataset possesses almost 29532 records collected over a period between 2015 and 2020. The dataset consists the information about the pollution criteria that occur in the different cities region every month. The attributes included in the dataset are City, Date, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, LPG, Methane, Carbon Monoxide, AQI, AQI_Bucket. However, the implemented model cannot be directly applied to the dataset, as the Sample dataset consists of numerical and categorical data. Hence, the dataset needs to be generalized and freed from the noise before applying the model to the dataset.Table 2 presents the description of the sample data set and its corresponding attributes.

Table 2.Description of the sample data set and its corresponding attributes

City	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	LPG,	Methane,	Carbon Monoxide	AQI	AQI_Bucket
AHB	38.88	192	2.18	16.28	17.74	53.31	2.18	26.4	12.47	3	6.96	6.37	128	Moderate
Brajra	128.	192	19.8	23.62	38.45	53.31	1.36	15.05	7.28	6.08	4.38	1.22	308	Very Poor
Brajra	128	192	19.8	23.62	38.45	53.31	1.36	15.05	7.28	6.08	4.01	0.86	308	Poor
Brajra	106	17.9	16.8	23.87	34.84	50.44	1.37	14.31	7.71	17.33	3.69	0.98	286	Poor
HYD	51.86	108	4.68	28.92	23.65	22.27	1.2	4.67	27.96	0.99	4.38	1.22	139	Serve
HYD	37.66	78	7.38	27.73	24.53	20.13	1.1	7.1	26.39	0.64	4.01	0.86	107	Moderate
HYD	45	93.9	8.98	33.13	29.96	22.38	1.02	7.21	29.12	0.68	3.69	0.98	105	Very Poor

4.2 Identify and handle the missing values

In real-world data sets, the particular characteristic is left absent because of various reasons like failing to load the information, corrupted data, or Data being incompletely collected. So, one of the crucial challenges any data analyst faces is handling such missing values as the models are trained and tested using these datasets. Let us look at different ways of assigning the missing values [21]. Table 3 shows the data set contains the absent values, and missing values are represented. In the above table, empty cells are represented with a red mark. There are four varied techniques to handle and control the missing fields in the data set. The four approaches are as below.

Table 3.Data set contains the absent values, and missing values are represented

City	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	LPG,	Methane,	Carbon Monoxide	AQI	AQI_Bucket
AHB	38.88	192	2.18	16.28	17.74	53.31	2.18	26.4	12.47	3	6.96	6.37	128	Moderate
Brajra	128.	192		23.62	38.45	53.31	1.36	15.05	7.28	6.08	4.38	1.22	308	Very Poor
Brajra	128	192		23.62	38.45	53.31	1.36	15.05	7.28	6.08	4.01			Poor
Brajra	106	17.9	16.8	23.87	34.84	50.44	1.37	14.31	7.71	17.33	3.69			Poor
HYD	51.86	108	4.68	28.92	23.65	22.27	1.2	4.67	27.96	0.99	4.38	1.22	139	Serve

4.3.Deleting the Rows

Deleting rows is the most widely used technique in processing null values. If any particular feature possesses a missing value, the specific row is deleted, or a column is deleted if it possesses 70-75% of missing values. Before deleting the row or column, it is mandated to confirm that the deletions do not adversely impact results or outputs' accuracy [17].

4.4 Replace with Mean\Mode\Median values

This technique best suits the features containing statistical Data. In this approach, the incomplete data or missing values are replaced with the Data's mean, median, or mode. Instead of removing the entire row or column to handle the missing values, it is identified that this technique is generating better results. The technique commonly used to handle the linear Data is to assess the deviation of neighboring values with it. Another numerical way to handle the missing values is to substitute the three estimates with those above [18]. Table 4 Show the data replacement by using Mean values

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

Table 4.Data replacement by using Mean values

City	Date	PM 2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	LPG,	Methane,	Carbon Monoxide	AQI	AQI_Bucket
Ahmedabad	23-05-2015	38.88	140.2	2.18	16.28	17.74		2.18	26.4	12.47	3	6.96	6.37	128	Moderate
Brajrajnagar	06-12-2018	128.42	192.96	19.81	23.62	38.45	53.31	1.36	15.05	7.28	6.08	5.116667	2.816667	308	Very Poor
Brajrajnagar	06-12-2018	128.42	192.96	19.81	23.62	38.45	53.31	1.36	15.05	7.28	6.08	5.116667	2.816667	308	Very Poor
Brajrajnagar	07-12-2018	106.56	174.97	16.89	23.87	34.84	50.44	1.37	14.31	7.71	17.33	5.116667	2.816667	286	Poor
Hyderabad	20-04-2016	45	93.9	8.98	33.13	29.96	22.38	1.02	7.21	29.12	0.68	3.69	0.98	105	Moderate
Mean			140.2									5.116667	2.816667		

4.5 Features Extracting Models

The sections above discussed the varied ways to handle the noisy data, such as removing the null and missing values and the dataset needed for a predictive model. The different techniques are available in dividing the dataset into training and test datasets. Now, the next step is to build the appropriate features used by a predictive model to make effective and accurate pollution predictions. Data Feature extraction plays a vital role in making accurate predictions. The following section presents the varied techniques of feature extraction on the pollution dataset.

4.6 Principle Components Analysis (PCA)

PCA is an unsupervised linear transformation technique used mostly for feature extraction and dimensionality reduction. It seeks the maximum variance directions in high-dimensional data and projects the data onto a new subspace with the same or less dimensions as the original. Take note of the maximum variance of data in the diagram below. This is represented by PCA1 (first maximum variation) and PC2 (second maximum variation) (2nd maximum variance). [19].

4.7 Algorithm of PCA's:

Feature extraction is the process of translating features into new feature subspaces while maintaining information in the original features.

- A new features subspace is formed by translating a d-dimensional data set into a k-dimensional data set using a projection matrix.The projection matrix is built by picking the K most important eigenvectors.

- Essential eigenvectors are chosen by ranking eigenvalues and calculating their explained variance ratio.
- Eigenvectors and eigenvalues are generated by performing an eigendecomposition of an initial data set's covariance matrix.

Before applying PCA, the data set must be standardized. The PCA steps describe how the PCA technology can be implemented to the pollution dataset to discover the connection between variables and reduce the dataset's complexity.

1. Co-variance Matrix XTX containing estimations of how each variable in X relates to each other variable in X. Understanding how one variable is linked to another is quite useful.
2. Second, eigenvalues and eigenvectors are critical.Eigenvectors represent directions.Consider plotting your data on a multidimensional scatterplot.Then consider each eigenvector as a distinct "direction" in your data scatterplot.Eigenvalues signify magnitude or significance.
3. Greater eigenvalues indicate more essential directions. Finally, we suppose that greater variability in one direction correlates with better explaining the dependent variable's behavior. For example, a high number of variations usually implies a signal, whereas a low number of variations usually indicates noise. Thus, the greater the variability in a specific direction, the more likely we are to identify something significant.

The PCA is a linear dimensionality reduction approach that enables data projection in the direction of highest variance (Monasterio et al., 2009).This method is used to extract useful features from an ECG data collection. The signal segment of a heartbeat, for example, is represented by, as in.

$$\mathbf{y}(k) = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(M) \end{bmatrix} \quad (1)$$

Where M is the number of pollutant samples. As a result, heartbeats are N observations of heartbeats, as shown in the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

following equation. The M X N matrix represents the whole ensemble of heartbeats.

$$\mathbf{Y} = [\mathbf{y}_1\ \mathbf{y}_2 \cdots \mathbf{y}_N] \quad (2)$$

The PCA consists of the following steps:

1. Determine the mean vector.Each pollution attribute's mean vector is determined as

$$\bar{\mathbf{y}} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{y}_i \quad (3)$$

2. Calculate the adjusted mean data.

$$\mathbf{yadj}_i = \mathbf{y}_i - \bar{\mathbf{y}}$$

$$\mathbf{Yadj} = [\mathbf{yadj}_1\ \ \mathbf{yadj}_2\ \ \cdots \mathbf{yadj}_N] \quad (4,5)$$

3. Make a covariance matrix.

$$\mathbf{C} = \frac{1}{M-1}\sum_{i=1}^{M}(\mathbf{y}_i - \bar{\mathbf{y}})^T(\mathbf{y}_i - \bar{\mathbf{y}}) \quad (6)$$

4. Calculate the eigenvectors and eigenvalues of the covariance matrix. The eigenvalues and eigenvectors , correspond to

$$\mathbf{C}\cdot\mathbf{e}_i = \lambda_i \cdot \mathbf{e}_i, \quad i = 1,\ldots N \quad (7)$$

5. Component selection and feature vector formation To begin, the primary component is the eigenvalue with the largest value. The eigenvalues are then arranged by eigenvalues from highest to lowest, resulting in the components being returned in the order of significance.The dimension is then decreased by choosing K-principal components that maintain physiological information.Thus, applying yields the percentage of variance of each eigenvalue.

$$\mathbf{r_K} = \frac{\sum_{i=1}^{K}\lambda_i}{\sum_{i=1}^{N}\lambda_i} \quad (8)$$

Furthermore, we select the principal components whose percentage of variance is higher than the percentage threshold, that is 0.9 or 0.95 as shown. Table 5 Show the training data for PCA algorithm

$$\hat{\mathbf{r}}_\mathbf{K} = (\mathbf{r_K} \geq th) \quad (9)$$

6. Deriving the new data set. The final dataset is obtained by

$$\mathbf{Ypca}(k) = \hat{\mathbf{r}}_\mathbf{K}{}^T\ \mathbf{Yadj}^T \quad \mathbf{(10)}$$

Table 5 .The training data for PCA algorithm

City	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	LPG,	Methane,	Carbon Monoxide	AQI	AQI_Bucket
AHB	38.88	192	2.18	16.28	17.74	53.31	2.18	26.4	12.47	3	6.96	6.37	128	Moderate
Brajra	128.	192	19.8	23.62	38.45	53.31	1.36	15.05	7.28	6.08	4.38	1.22	308	Very Poor
Brajra	128	192	19.8	23.62	38.45	53.31	1.36	15.05	7.28	6.08	4.01	0.86	308	Poor
Brajra	106	17.9	16.8	23.87	34.84	50.44	1.37	14.31	7.71	17.33	3.69	0.98	286	Poor
HYD	51.86	108	4.68	28.92	23.65	22.27	1.2	4.67	27.96	0.99	4.38	1.22	139	Serve
HYD	37.66	78	7.38	27.73	24.53	20.13	1.1	7.1	26.39	0.64	4.01	0.86	107	Moderate
HYD	45	93.9	8.98	33.13	29.96	22.38	1.02	7.21	29.12	0.68	3.69	0.98	105	Very Poor

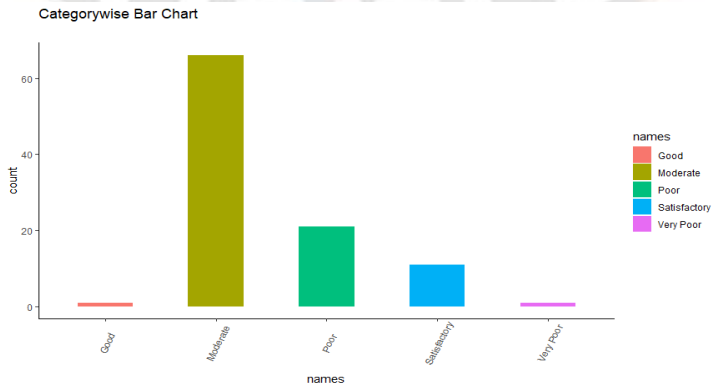

Figure 4. This graph represents categories of the data dependents upon the attributes how many attributes consist in the data sets.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

In the tabel 6 the above algorithm, step by step, is implemented and applied to the pollution data set. Here is an algorithm that calculates the standard deviation. Figure 4 shows the graph represents categories of the data dependents upon the attributes how many attributes consist in the data sets.

Tabel 6. Pollution data set

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	2.09	1.8	1.19	0.95	0.906	0.82	0.71	0.61	0.46	0.42	0.383	0.266	0.06
Proportion of Variance	0.33	0.25	0.10	0.070	0.063	0.051	0.03	0.028	0.016	0.013	0.011	0.005	0.033
Cumulative Proportion	0.33	0.58	0.69	0.76	0.83	0.88	0.92	0.95	0.96	0.98	0.99	0.99	1

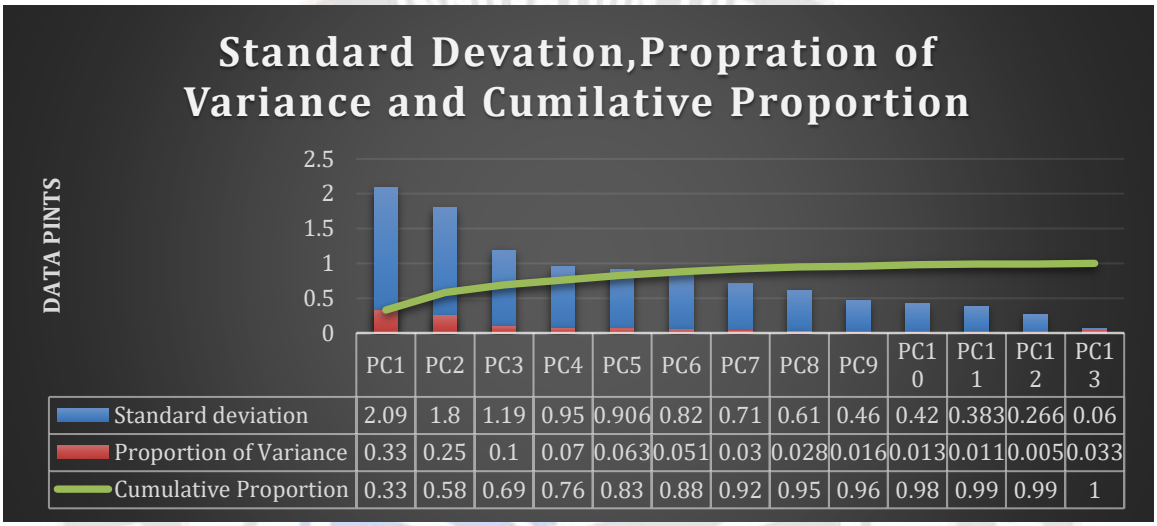

Figure 5: This graph represents categories of the Standard devation. Propration of variance.

Tabel 7 :The below table, step by step, is implemented and applied to the pollution data set. Here is an algorithm that calculates the standard deviation.

	PM1	PM2	PM3	PM4	PM5	PM6	PM7	PM8	PM9	PM10	PM11	PM12	PM13
PM10	-0.44	0.06	-0.11	0.07	0.02	-0.09	-0.09	0.14	-0.06	-0.34	0.57	0.56	-0.02
ï..PM2.5	-0.44	0.01	-0.19	0.03	-0.05	-0.16	-0.18	0.22	0.04	-0.13	0.2	-0.78	0.02
AQI	-0.41	0.07	-0.22	0.11	0.01	0.01	-0.14	0.27	-0.53	0.35	-0.49	0.18	-0.01
NO2	-0.41	-0.14	0.26	0.18	0.13	0.04	0.06	-0.06	0.49	0.12	-0.22	0.06	-0.62
NOx	-0.38	-0.22	0.3	0.19	0.08	0.13	0.07	-0.2	0.2	0.06	-0.11	0.03	0.75
O3	-0.17	0.12	-0.61	-0.23	0.33	0.05	0.06	-0.64	0.09	0.05	-0.04	0	0
SO2	-0.14	-0.19	0.27	-0.7	0.33	-0.41	0.27	0.15	-0.11	0.05	-0.02	0.01	0.02
NH3	-0.12	-0.25	-0.24	-0.15	-0.8	-0.24	0.31	-0.1	0.13	0.08	-0.07	0.1	0.01
NO	-0.1	-0.45	0.25	0.03	-0.09	0.26	0.02	-0.43	-0.58	-0.11	0.18	-0.15	-0.24
Toluene	0.05	-0.44	-0.29	-0.15	0.1	0.36	0	0.31	0.11	-0.57	-0.35	0.05	0.02
Benzene	0.1	-0.45	-0.26	0	0.15	0.26	0.04	0.27	0.12	0.6	0.42	0.01	0.02
CO	0.14	-0.38	-0.02	-0.01	0.01	-0.44	-0.76	-0.17	0.1	0.04	-0.08	0.11	0.02

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

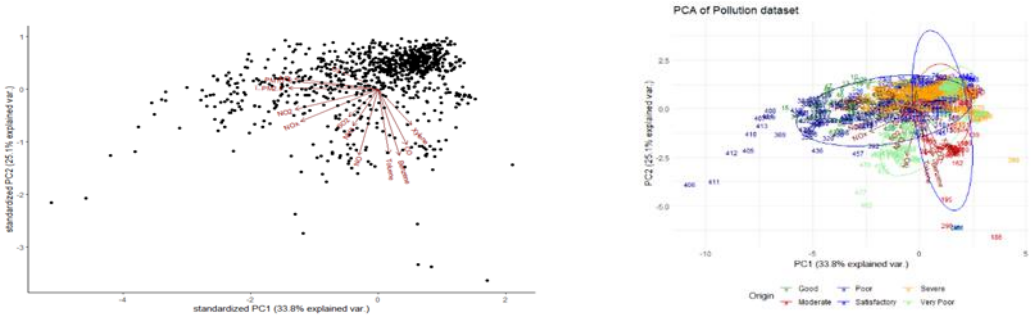

Figure 6.This graph represents the attributes orientation of the overall data and co-related with the labels.

Table 15.Present how attributes are strengthened for algorithms and fit for the algorithm. Tabel 7 :The below table, step by step, is implemented and applied to the pollution data set. Here is an algorithm that calculates the standard deviation. Figure 6 :This graph represents the attributes orientation of the overall data and co-related with the labels. Figure 6 :This graph represents the attributes orientation of the overall data and co-related with the labels. Figure 7 represents the performance of the PCAs for class 0 and class 1 and 2 ,3 .Table 8 represents the dataset generated after completing the preprocessing as well as encoding label classification. In the following table, the categorical data with "poor" or "moderate" or "very poor" encoded as 0 ,1,2,3.

The following confusion matrix and the corresponding performance parameters are generated on applying PCAs to the above data set.

Table 8.Illustrates the resultant Confusion Matrix on applying PCAs

Confusion Matrix	**Precision(%)**	**Recall(%)**	**f1score(%)**
0 Class	90	93	92
1 Class	95	93	93
2 Class	92	94	95
3 Class	93	94	96
Accuracy	93	95	96
macro avg	97	98	96
Weighted avg	95	96	98

From the resultant confusion matrix, it is identified that the dataset possesses 14 features, two classes and achieved a performance of 96%.

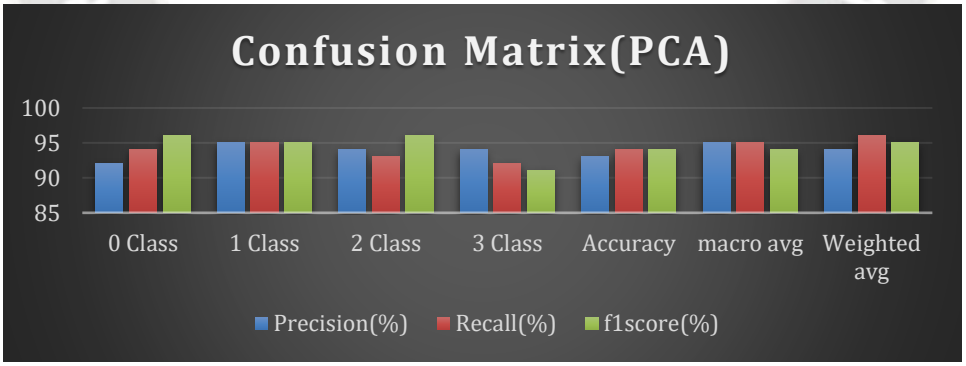

Figure 7.Represents the performance of the PCAs for class 0 and class 1 and 2 ,3 .

4.8 k-nearest Neighborhood(kNN)

ICA is utilized to extract the important signals or data from the data source possessing a mixture of signals. The dataset may possess different types of data, such as sounds or stock markets, or videos. ICA is successfully being used in several applications such as audio signals, biological testing and medical signals. ICA is also utilized in dimensionality reduction as it can preserve and delete a single source. ICA is even used as a filtering system as it is used both in filtering as well as erasing signals [19].

ICA accepts a collection of individual components as input and deletes all the noise there by defining each input correctly. Two features are considered to be independent if their linear yet not linear influence is equal to zero[20].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

Table 9 presents the resultant confusion matrix of ICA. Figure 8 represents the performance of the ICAs for class 0,1 as well as class 2,3.

Table 9.Presents the resultant confusion matrix of ICA.

Confusion Matrix	Precision(%)	Recall(%)	f1score(%)
0 Class	92	94	96
1 Class	95	95	95
2 Class	94	93	96
3 Class	94	92	91
Accuracy	93	94	94
macro avg	95	95	94
Weighted avg	94	96	95

Figure 8 represents the performance of the ICAs for class 0,1 as well as class 2,3.

4.9Linear Discriminant Analysis (LDA)

LDA is a dimensionality reduction technique and uses supervised ML classifiers. This technique achieves classification by maximizing the distance amid each class's mean and minimizing the spread within the class. LDA performs better classification by maximizing each class's reach by projecting the data in a lower-dimensional space. Therefore, LDA uses measures within as well as among categories. [22].

LDA Methodology and Steps

The following work presents the role of dimensionality reduction and feature engineering methods on the ML algorithms' performance in assessing the pollution dataset. Figure 9 describes the approach LDA uses in feature reduction.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

Figure 9.Describes the approach LDA uses in feature reduction

Table 10 .LDAs resultant Confusion Matrix. It shows that it achieved 96.4% performance with respect to the performance metrics considered.

Confusion Matrix	Precision(%)	Recall(%)	f1score(%)
0 Class	96	95	96
1 Class	95	96	96
2 Class	92	91	90
3 Class	90	91	93
Accuracy	96	95	96
macro avg	94	96	96
Weighted avg	96	95	94

Table 10 represents the LDAs resultant Confusion Matrix. It shows that it achieved 96.4% performance with respect to the performance metrics considered.

4.10 Resultant and Inferences

In this section, varied feature extraction methods are compared and analyzed. These models are developed using R programming and R Studio. Data set is extracted from the source, IMD data provider. Predictive models cannot be applied directly on the extracted dataset as it possess raw data. Hence, the feature extraction techniques such as PCA, ICA, and LDA and different dimensionality reduction techniques were applied to the dataset to make the dataset ready and determine the accuracy, f1 score, Recall, specificity, and precision. The following results present the performance of the feature extraction models individually.

4.11 Accuracy of the feature extraction models

In this section, the accuracy of the feature extraction models PCA, ICA, and LDA is calculated and assessed. Accuracy (ACC) is measured by dividing the number of accurate forecasts made by the total number forecasts made. The highest precision is 1.0, while the worst one is 0.0 [26].

$$ACC = \frac{TP+TN}{TP+TN+FN+FP} = \frac{TP+TN}{P+N} \text{-------(1)}$$

True positive (TP): To predict the positive class correctly.

False positive (FP): To predict the positive class incorrectly.

True negative (TN): To predict the negative class correctly.

False negative (FN): To predict the negative class incorrectly.

Table 11 represents the feature extraction algorithms' name and their relative accuracy.

Table 11.The feature extraction algorithms' name and their relative accuracy.

S.no	Algorithm	Accuracy(%)
1	PCA	93.45
2	ICA	94.23
3	LDA	96.18

The above table presents the accuracy of the feature extraction algorithms, PCA, ICA, and LDA. The accuracy can be calculated by considering the number of true positive and true negative predictions the algorithm made divided by the total number of predictions it made. The table proves that LDA outperformed compared to PCA and ICA as it attained 96.4% accuracy whereas, PCA attained 93.45% and ICA attained 94.23% respectively. Figure 10 represents the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

bar chart regarding the performance comparisons of the fourteen feature extraction algorithms

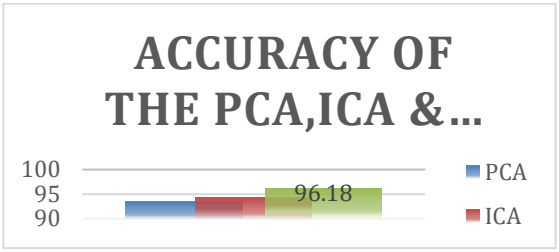

Figure 10 represents the bar chart regarding the performance comparisons of the fourteen feature extraction algorithms

4.12 F1 score

It is nothing but a Correlation coefficient of the data set correlation attributes with the classified data. The following represents the equation of the Mathew correlation [24]. Table 12 presents the performance of the feature extraction models w.r.t F1 Score

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \text{------(1)}$$

- True positive (TP): To predict the positive class correctly.
- False positive (FP): To predict the positive class incorrectly.
- True negative (TN): To predict the negative class correctly.
- False negative (FN): To predict the negative class incorrectly.

Table 12.Performance of the feature extraction models w.r.t F1 Score

F1 Scores	Precision(%)	Recall(%)	f1score(%)
0 Class	92	93	96
1 Class	93	94	95
2 Class	91	92	93
3 Class	90	89	91
macro avg	94	96	96
Weighted avg	93	94	95

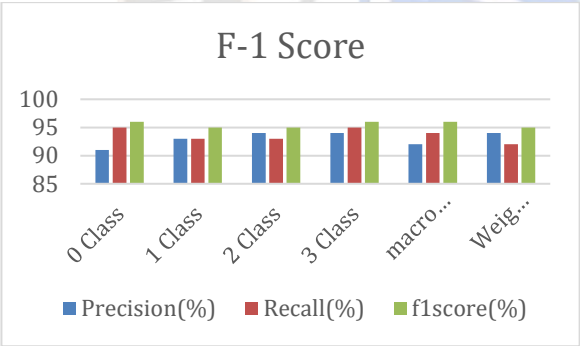

Figure 11.Line Chart show the performance of PCA, ICA, and LDA w.r.t F1 score

The above chart shows the F1 correlation scores of the 0s class and 1s class, a macro average of the attributes, and the weighted average of the features. The experimental results showed that LDA is the best among the three algorithms, as it succeeded in building a solid correlation among the attributes.

4.13 Recall and Precision

The amount of accurate positive forecast determined by Recall (SN) divides by a total of positive predictions. The REC or true positively rate (TPR) is also known as memory. The highest sensitivity is 1.0, while the worst sensitivity is 0.0. Table 13 presents the performance of the feature extraction models Recall. Figure 12 Bar chart shows the Weighed average, macro average, and class labels regarding the Recall performances.

Table 13.Performance of the feature extraction models Recall.

	Recall		
Confusion Matrix	**PCA(%)**	**ICA(%)**	**LDA(%)**
0 Class	93	94	96
1 Class	93	93	95
2 Class	91	90	91
3 Class	94	92	96
macro avg	95	95	96
Weighted avg	92	94	95

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

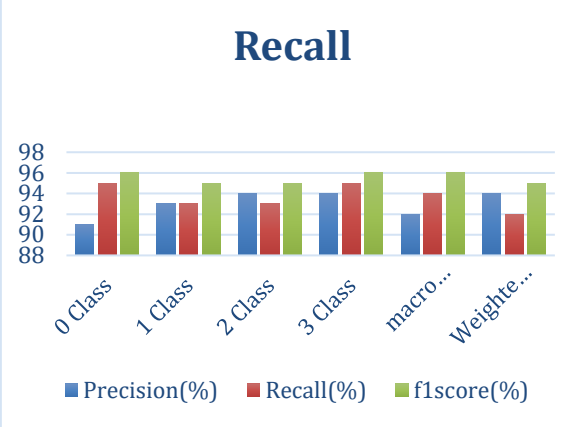

Figure 12.Bar chart shows the Weighed average, macro average, and class labels regarding the Recall performances.

Precision (PREC) tests the correct sum of successful predictions divided by overall optimistic predictions. The PPV is also known as the positive predictive value. The maximum accuracy is 1.0, the minimum accuracy of 0.0.

Table 14 shows the Precision performances on the class labels' attributes and average values in comparing these three algorithms

Table 14 .Precision performances on the class labels' attributes and average values in comparing these three algorithms.

	Precision		
Confusion Matrix	**PCA(%)**	**ICA(%)**	**LDA(%)**
0 Class	91	95	96
1 Class	93	93	95
2 Class	94	93	95
3 Class	94	95	96
macro avg	92	94	96
Weighted avg	94	92	95

Here, identify the LDA Feature extraction mechanism as the Precision best performance among the other two algorithms. Algorithms Performances are the PCA, ICA, and LDA 93%,94.5%& and 96.4% of the LDA Algorithm.

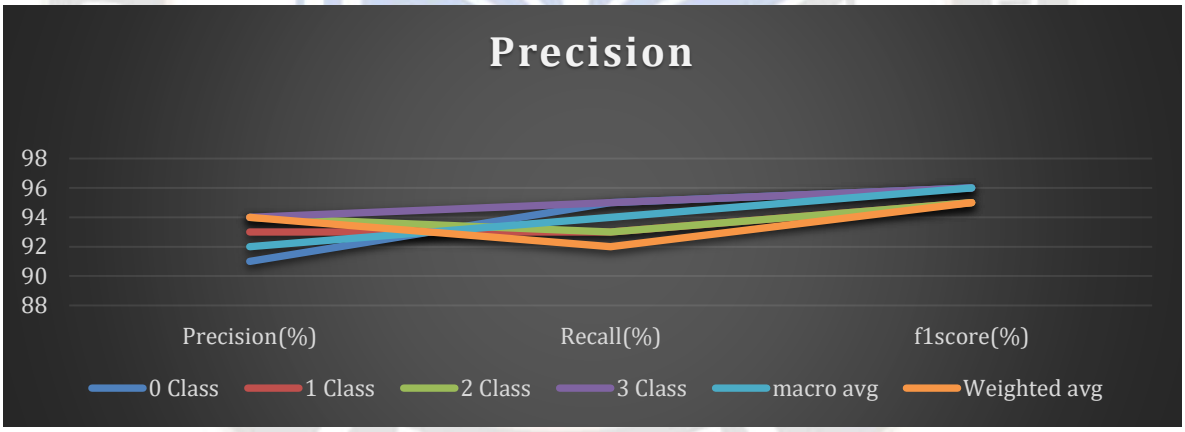

Figure 13. Precision performances of the class labels, macro average, and weighted average.

	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3
F1 Score	87	89	94	97	88	90	92	88	96
Precision	86	94	91	96	91	93	91	93	97
Recall	86	91	91	97	93	86	87	88	89
Accuracy	89	87	96	89	95	88	92	88	95

Figure 13 Line chart represents the precision performances of the class labels, macro average, and weighted average .Table 15 In below table represent how attributes are strengthened for algorithms and fit for the algorithm. Figure 14 Line chart represents how attributes are strengthened for algorithms and fit for algorithm

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 2s
DOI: https://doi.org/10.17762/ijritcc.v11i2s.6036
***Article Received**: 15 November 2022 **Revised**: 19 December 2022 **Accepted**: 04 January 2023*

Table 15.Present how attributes are strengthened for algorithms and fit for the algorithm.

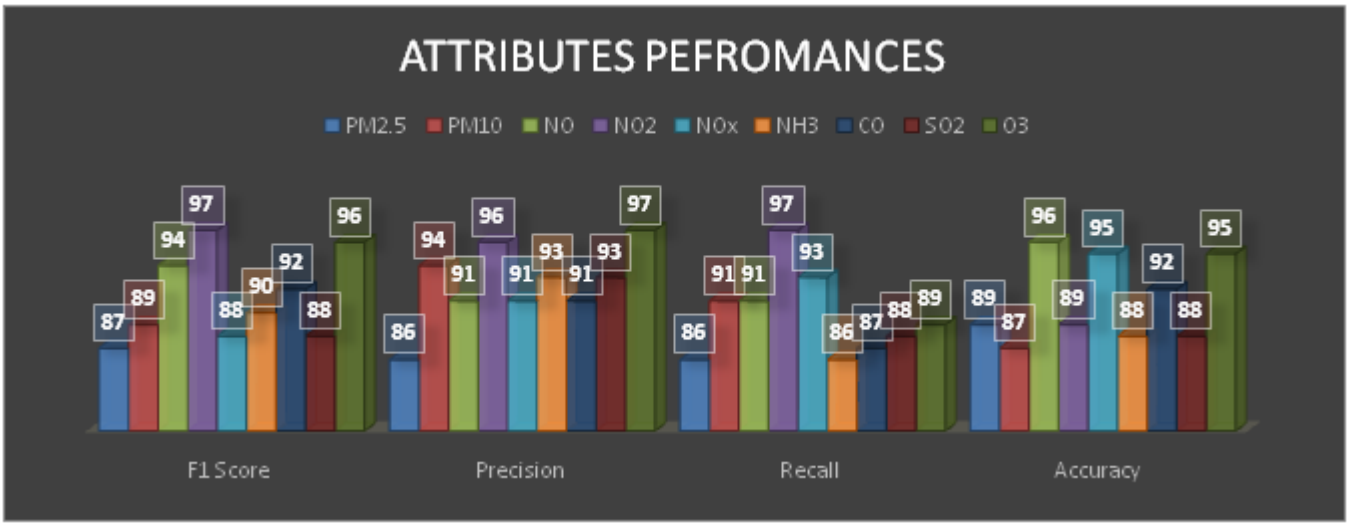

Figure 14. Line chart represents how attributes are strengthened for algorithms and fit for algorithm

V. CONCLUSION

This main paper's objective is to get the data from the IoT-based terminal like pollution detection and data cleaning and feature extraction mechanisms to get the futuristic data. This futuristic Data is used for the future enhancement of this System. In this regard, the design terminal represents the sensing the pollution attributed data and weather information sensing the numerical values post to the cloud-based database. This data need to apply the preprocessing techniques for missing values. After this stage, the exact best-fitted feature and features strengths need to be calculated. In this situation, machine learning-based feature extraction classification models PCA, ICA, and LDA extract the exact features and calculate the strengths of the correlations through the confusion matrix and calculate the accuracy, precision, and Recall. Accuracy of the PCA 93.45%, ICA 94.18%, and LDA 96.45%. LDA gives the best results for best-fitted attributes. These results are maintained a database used for the prediction of AQI_bucket for future enhancement.

Funding Not applicable

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

Research involving human participants and/or animals Not applicable.

Informed consent Not applicable

REFERENCES

[1]. Y. W. Lee, "A stochastic model of particulate matters with AI-enabled technique-based IoT gas detectors for air quality assessment," *Microelectron. Eng.*, vol. 229, p. 111346, 2020, doi: 10.1016/j.mee.2020.111346.

[2]. R. Kumar, P. Kumar, and Y. Kumar, "Time Series Data Prediction using IoT and Machine Learning Technique," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 373–381, 2020, doi: 10.1016/j.procs.2020.03.240.

[3]. S. M. Grath, C. Flanagan, L. Zeng, and C. O'Leary, "IoT Personal Air Quality Monitor," 2020 31st Irish Signals Syst. Conf. ISSC 2020, 2020, doi: 10.1109/ISSC49989.2020.9180199.

[4]. H. Balogun, H. Alaka, and C. N. Egwim, "Boruta-grid-search least square support vector machine for NO2 pollution prediction using big data analytics and IoT emission sensors," Appl. Comput. Informatics, no. 2, 2021, doi: 10.1108/ACI-04-2021-0092.

[5]. J. Parmar, T. Nagda, P. Palav, and H. Lopes, "IOT Based Weather Intelligence," 2018 Int. Conf. Smart City Emerg. Technol. ICSCET 2018, pp. 1–4, 2018, doi: 10.1109/ICSCET.2018.8537382.

[6]. M. Maruthi, L. Rao, V. Y. Raghava, and S. S. R. I. Ramya, "ISSN : 0731-6755," vol. XII, no. Xi, pp. 1162–1165, 2019.

[7]. L. Campanile, P. Cantiello, M. Iacono, R. Lotito, F. Marulli, and M. Mastroianni, "Applying Machine Learning to Weather and Pollution Data Analysis for a Better Management of Local Areas: The Case of Napoli, Italy," pp. 354–363, 2021, doi: 10.5220/0010540003540363.

[8]. M. Dhanalakshmi and V. Radha, "A Survey paper on Vehicles Emitting Air Quality and Prevention of Air Pollution by using IoT Along with Machine Learning Approaches," vol. 12, no. 11, pp. 5950–5962, 2021.

[9]. Y. Barhate, R. Borse, N. Adkar, and G. Bagul, "Plant Watering and Monitoring System using IoT and Cloud

Computing," Int. Jouranl Sci. Dev. Res., vol. 5, no. 4, pp. 157–162, 2020.

[10]. F. A. Almalki et al., "Green IoT for Eco-Friendly and Sustainable Smart Cities: Future Directions and Opportunities," Mob. Networks Appl., 2021, doi: 10.1007/s11036-021-01790-w.

[11]. S. M. Bante, S. R. Karale, and G. K. Awari, "A systematic review on real time exhaust gas sensing system for on board sensing of harmful gases in IC engine," IOP Conf. Ser. Mater. Sci. Eng., vol. 1170, no. 1, p. 012012, 2021, doi: 10.1088/1757-899x/1170/1/012012.

[12]. C. Dhule, R. Agrawal, S. Dorle, and B. Vidhale, "Study of Design of IoT based Digital Board for Real Time Data Delivery on National Highway," Proc. 6th Int. Conf. Inven. Comput. Technol. ICICT 2021, pp. 195–198, 2021, doi: 10.1109/ICICT50816.2021.9358560.

[13]. S. H. Haji and A. B. Sallow, "IoT for Smart Environment Monitoring Based on Python: A Review," Asian J. Res. Comput. Sci., vol. 9, no. 1, pp. 57–70, 2021, doi: 10.9734/ajrcos/2021/v9i130215.

[14]. Q. Yaseen and Y. Jararweh, "Building an Intelligent Global IoT Reputation and Malicious Devices Detecting System," J. Netw. Syst. Manag., vol. 29, no. 4, pp. 1–17, 2021, doi: 10.1007/s10922-021-09611-x.

[15]. A. Hussain et al., "Waste management and prediction of air pollutants using IoT and machine learning approach," Energies, vol. 13, no. 15, 2020, doi: 10.3390/en13153930.

[16]. J. Mabrouki, M. Azrour, D. Dhiba, Y. Farhaoui, and S. El Hajjaji, "IoT-based data logger for weather monitoring using arduino-based wireless sensor networks with remote graphical application and alerts," Big Data Min. Anal., vol. 4, no. 1, pp. 25–32, 2021, doi: 10.26599/BDMA.2020.9020018.

[17]. A. Thorat, "Indoor Air Quality Monitoring System," Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 5, pp. 1345–1353, 2021, doi: 10.22214/ijraset.2021.34545.

[18]. S. Majumdar, M. M. Subhani, B. Roullier, A. Anjum, and R. Zhu, "Congestion prediction for smart sustainable cities using IoT and machine learning approaches," Sustain. Cities Soc., vol. 64, no. March 2020, p. 102500, 2021, doi: 10.1016/j.scs.2020.102500.

[19]. S. Jha, L. Nkenyereye, G. P. Joshi, and E. Yang, "Mitigating and monitoring smart city using internet of things," Comput. Mater. Contin., vol. 65, no. 2, pp. 1059–1079, 2020, doi: 10.32604/cmc.2020.011754.

[20]. A. S. Moursi, N. El-Fishawy, S. Djahel, and M. A. Shouman, "An IoT enabled system for enhanced air quality monitoring and prediction on the edge," Complex Intell. Syst., no. 0123456789, 2021, doi: 10.1007/s40747-021-00476-w.

[21]. A. Parashar, "IoT based automated weather report generation and prediction using machine learning," 2019 2nd Int. Conf. Intell. Commun. Comput. Tech. ICCT 2019, pp. 339–344, 2019, doi: 10.1109/ICCT46177.2019.8968782.

[22]. M. M. Rathore, A. Paul, W. H. Hong, H. C. Seo, I. Awan, and S. Saeed, "Exploiting IoT and big data analytics: Defining Smart Digital City using real-time urban data," Sustain. Cities Soc., vol. 40, pp. 600–610, 2018, doi: 10.1016/j.scs.2017.12.022.

[23]. M. Selvakumar, V. Prasannakumari, S. Geetha, and S. Muthulakshmi, "Validation of Point Source Models for Determining Industrial Pollution and Integrating with IOT for vulnerability management," IOP Conf. Ser. Mater. Sci. Eng., vol. 1055, no. 1, p. 012022, 2021, doi: 10.1088/1757-899x/1055/1/012022.

[24]. S. Pandya et al., "Pollution weather prediction system: Smart outdoor pollution monitoring and prediction for healthy breathing and living," Sensors (Switzerland), vol. 20, no. 18, pp. 1–25, 2020, doi: 10.3390/s20185448.

[25]. N. Shahid, M. A. Shah, A. Khan, C. Maple, and G. Jeon, "Towards greener smart cities and road traffic forecasting using air pollution data," Sustain. Cities Soc., vol. 72, no. November 2020, p. 103062, 2021, doi: 10.1016/j.scs.2021.103062.

[26]. R. Singh, "Raspberry Pi and Cloud Computing," pp. 702–707, 2020.

[27]. M. Deopa, "Smart Framework for Pollution Monitoring and Reporting System using IOT," Int. J. Res. Appl. Sci. Eng. Technol., vol. 8, no. 7, pp. 355–358, 2020, doi: 10.22214/ijraset.2020.7058.

[28]. J. Cruz, "Ubicomm." M. K. Gour and A. K. Tiwari, "A Study of Weather Forecasting in IOT Technology," no. 1, pp. 6–9, 2022.