Research Article

Air Pollution Monitoring Using IoT and Machine Learning in the Perspective of Bangladesh

Md Monirul Islam[1], Shalah Uddin Perbhez Shakil[1], Nasim Mahmud Nayan[2], Mohammod Abul Kashem[3] and Jia Uddin[4,*]

[1]Daffodil International University, Bangladesh
monir.duet.cse@gmail.com; perbhezshakil.swe@diu.edu.bd
[2]University of Information Technology & Sciences (UITS), Bangladesh
smnoyan670@gmail.com
[3]Dhaka University of Engineering and Technology, Bangladesh
drkashemll@duet.ac.bd
[4]Woosong University, South Korea
jia.uddin@wsu.ac.kr
*Correspondence: jia.uddin@wsu.ac.kr

Received: 28th March 2024; Accepted: 28th June 2024; Published: 1st July 2024

Abstract: Air pollution is a big concern in developing countries due to its negative effects on both human well-being and the environment. Collecting real-time values of air quality is challenging using traditional methods. Because they have limited coverage and may not accurately reflect pollution levels in a specific location. However, Advances in Internet of Things (IoT) technology and Machine Learning (ML) algorithms, can play a vital role in collecting and analyzing large amounts of air quality data, resulting in a more complete and exact knowledge of pollution levels. Throughout this work, based on several air pollutants including sulfur dioxide (SO2), ozone (O3), nitrogen dioxide (NO2), particulate matter (PM) 2.5, particulate matter (PM) 10 and carbon monoxide (CO) across a large urban region, we establish an IoT-based framework to collect real-time data. After collecting the real-time values, we applied two types of machine learning algorithms named regression and classification models including linear regression, decision trees (DT), random forest (RF), K-Nearest Neighbours (KNN), Naive Bayes (NB), and gradient boosting (GB), to analyze the gathered data and estimate pollution levels into good, satisfactory, moderate, poor and very poor. Among the machine learning models, RF outperforms the result. This work and the dataset will be helpful for researchers, environmental practitioners and agencies.

Keywords: ***Air pollutant monitoring; AQI prediction; IoT; Machine Learning; Real-time values***

1. Introduction

Air pollution has a widespread influence in a multitude of places across the Globe, providing a serious environmental and public health risk. Pollutants such as sulfur dioxide (SO2), nitrogen dioxide (NO2), particulate matter (PM), ozone (O3), and carbon monoxide (CO) have been linked to respiratory disorders, cardiovascular disease, and cancer [1].

Traditional methods of measuring air quality often rely on fixed sensors, which might provide restricted or erroneous information about pollution levels in a given region [2]. There are a lot of

Md. Monirul Islam, Shalah Uddin Perbhez Shakil, Nasim Mahmud Nayan, Mohammod Abul Kashem and Jia Uddin, "Air Pollution Monitoring Using IoT and Machine Learning in the Perspective of Bangladesh", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 17-33, Vol. 8, No. 3, 1st July 2024, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2024.03.002, Available: http://aetic.theiaer.org/archive/v8/v8n3/p2.html.

applications of IoT like water quality monitoring, aqua fisheries, and tube well water monitoring [3-6]. However, recent advances in IoT technology and ML algorithms have simplified the accumulation and assessment of copious amounts of air quality data, allowing for a more complete and exact understanding of air pollution levels [7].

To assess the amount of air pollution in a large metropolitan region by using machine learning techniques in conjunction with an IoT-based sensor network is the primary objective of the study [8]. A network of sensors strategically placed across the city collects data on various air pollution components, such as SO2, NO2, CO, PM2.5, and PM10. The collected data is examined using machine learning techniques to predict pollution levels depending on environmental factors [9-10].

This research has several potential implications in the fields of environmental policy, urban development, and public health. Precise and real-time monitoring of air pollution levels can provide decision-makers with important information to help them make decisions that will reduce air pollution and protect public health [11]. Moreover, the comprehensive and elaborate information furnished by the Internet of Things (IoT) sensor network aids urban planners in pinpointing pollution hotspots in the city and formulating focused measures to mitigate the problem.

Overall, this study emphasizes how IoT and machine learning technology may be used to improve public health outcomes and tackle challenging environmental issues. Modern technology combined with reliable data-gathering techniques has the potential to improve our comprehension of and response to the serious problem of air pollution in urban areas around the globe.

key pollutants comprise:

1. Particulate Matter (PM): PM particles can cause respiratory and cardiovascular problems including asthma, bronchitis, lung cancer, and heart disease because of their ability to enter the circulation and penetrate deep into the lungs. It's critical to stress the serious health risks that come with exposure to PM. In addition, PM can aggravate pre-existing medical disorders and irritate the throat, nose, and eyes [12].
2. Sulfur oxide (SOX): When mixed with other airborne substances, to create acid rain SOX can react, which has the potential to be harmful to some buildings, plants, and aquatic environments. Additionally, SOX emissions may lead to particulate matter formation, which can cause lung illnesses and have a negative impact on respiratory health.
3. Nitrogen oxide (NOX): NOx is a pollutant that has been recognized as the cause of respiratory issues, significantly in susceptible segments, including children, the elderly, and those with underlying medical disorders. It also contributes to ground-level ozone creation. Including Asthma, other respiratory conditions potentially be made worse by elevated NOx levels [13].
4. Carbon monoxide (CO): Because CO is colorless and odorless, it can be difficult to detect without specialist equipment, making it a serious threat. Headaches, nausea, and dizziness are some of the symptoms that can arise with excessive CO exposure. It can potentially result in death in extreme cases [14].
5. Ozone (O3): Ozone is a problem as it has no smell and is invisible, making it difficult to detect without sophisticated equipment. High-level ozone exposure can cause a number of health problems, such as respiratory discomfort and the aggravation of pre-existing illnesses. The elevated ozone levels and extended exposure can trigger severe respiratory issues also detrimental impacts on one's health [14].

The dataset is available in Data Mendeley platform[1] and this is the extended version of our previously published work [15]. The primary contributions of this paper as follows:

[1] https://data.mendeley.com/datasets/4r25x9sc7k/1

1. Creating real-time values by using IoT sensors like CO, SO2, NO2, O3, PM 2,5, and PM 10 from three different places in Dhaka and Gazipur districts, Bangladesh. And Arduino boards make it easier to gather data in real-time. This methodology facilitates ongoing surveillance of various air contaminants.
2. Applying two types of machine learning algorithms regression type and classification type. For analyzing the sensor values, we used a linear regression model and classified the quality of the index into five categories good, satisfactory, moderate, poor, and very poor, we used classifier algorithms, named KNN, Logistic regression, Naïve Bayes, random forest, gradient boosting, decision tree etc.
3. A variety of assessment criteria are used by the study to gauge how well the machine learning models are doing. These measures, which include regression metrics, F1 score, precision, recall, and Kappa score, offer a thorough assessment of the efficacy and accuracy of the model's predictions.

To sum up, the focal points of the study are the prospects of IoT and machine learning to address complicated environmental problems and improve the wellness of the public consequences. There is potential for better knowledge and mitigation of the issues linked with global air pollution through cutting-edge technology integration with meticulous data-gathering procedures.

2. Literature Review

Air pollution is a problem for both developed and developing nations, particularly in the latter cities, where industrial expansion continues [16]. Health issues linked to air pollution have increased faster frequency [17]. Consequently, there has been a lot of interest in raising public awareness of air quality through the monitoring of air pollutants. Air pollution could exert a negative impact on ecosystems and the environment, in addition to endangering human health [18]. Track the amounts of carbon monoxide (CO) and nitrogen dioxide (NO2) using IOT devices in Dhaka, Saini *et al.* [19] aim to compile a data set. Grounded on the outcomes of different sessions and measured values, it displays variance in CO and NO2 emissions. Unfortunately, this approach lacks the utilize of machine learning techniques to forecast air pollution and also ignores important parameters like PM2.5 or PM10.

The researchers proposed forecasting the air quality index utilizing two machine learning algorithms based on a secondary data-set named Central Pollution Control Board, India [20]. Payne *et al.* [21] proposed comparing uncertain output to supervised algorithms like Support Vector Machines, K-nearest Neighbour, and decision trees with unsupervised neural network techniques. Although neural networks perform better overall than these algorithms, they are unable to estimate air pollution levels on an hourly basis due to the use of pre-existing information. Parmar *et al.* [22], compared to other models specifically ANN, Random Forest, Decision Trees, Least Squares Support Vector Machine Model, and Deep Belief Network, require hourly data prediction. However, there are several issues with sensor quality due to manufacturing flaws in the device is an obstacle of that. According to Ali *et al.* [23], machine-learning methods are crucial for correctly estimating the air quality index. Auto-regression, artificial neural networks (ANN), and logistic regression can be used to calculate PM2.5 levels. The article's greatest findings are attributed to ANN. Molinara *et al.* [24] proposed indoor air quality monitoring using a microchip named SENSIPLUS consisting of sensors and machine learning models. Their accuracy is 75.01%.

In [25], the researchers presented an IoT method for internal air quality observation utilizing VOC, aerosol, CO, CO2, and temperature parameters implemented at Hanyang University. In [26], the authors proposed an IoT framework for air quality index monitoring using only three parameters named PM 2.5, CO, and NOx. They also applied only three machine learning models to predict AQI. In [27], this paper proposed an IoT framework for air monitoring using only two sensors named MQ7 and MQ135 for measuring the CO and air quality. However, only using the MQ135 sensor makes it impossible to measure

the actual quality of the air. They also used machine learning for data analysis. In [28], The researchers experimented with an IoT setup for air quality observation. The system consists of LPG, Humidity, Smoke, Temperature, Carbon Monoxide, and PM2.5 and 10. However, they did not analyze the data by utilizing machine learning algorithms. In [29], the author applied machine learning algorithms to forecast the air quality index based on two secondary datasets namely geospatial dataset and urban air pollution data in China. In [30], this paper produced only one factor named PM 2.5 real-time values of air quality in a specific area like 1 km by 1km and 1-hour resolution in Beijing, China. They also applied only one machine learning model namely decision tree. The paper introduced an air quality observation system engaging two parts named IoT and machine learning. They used four sensors for producing PM, CO, NH3, Temperature, and Humidity. For data analysis, they applied the ARIMA model [31]. Table 1 describes the summary of the existing works.

Table 1. Summary of the related works

Ref.	Location	Parameters	ML methods	Key findings	Limitations
[19]	Dhaka, Bangladesh	CO, NO2	--	Dataset compilation with variance in CO and NO2 emissions	Dataset compilation with variance in CO and NO2 emissions
[20]	India	AQI	--	Proposed forecasting AQI using machine learning	Specific techniques and results not detailed
[21]		AQI	SVM, K-nearest Neighbour, Decision Trees, Unsupervised Neural Networks	Neural networks perform better overall	Unable to estimate hourly pollution levels due to reliance on pre-existing data
[22]		AQI	ANN, Random Forest, Decision Trees, LSSVM, Deep Belief Network	Models require hourly data for accuracy	Sensor quality issues due to manufacturing flaws
[23]		PM2.5	Auto-regression, ANN, Logistic Regression	ANN yields best results for PM2.5 estimation	Does not address other pollutants
[24]		AQI	Machine Learning Models	Accuracy of 75.01% for indoor air quality monitoring	Limited to indoor settings
[25]	Hanyang University	VOC, Aerosol, CO, CO2, Temperature		IoT method for internal air quality observation	ML techniques not detailed
[26]		PM2.5, CO, NOx		IoT framework for AQI monitoring	Limited parameters and ML models
[27]		AQI	Machine Learning for Data Analysis	Insufficient for measuring actual air quality with only MQ135 sensor	Limited sensor variety
[28]		LPG, Humidity, Smoke, Temperature, CO, PM2.5, PM10		IoT setup for air quality observation	No machine learning analysis
[29]	China	AQI	Various ML models	Forecasting AQI using geospatial and urban air pollution data	Secondary datasets only
[30]	Beijing, China	PM2.5	Decision Tree	Real-time PM2.5 values in a specific area with 1 km by 1 km and 1-hour resolution	Only one pollutant and one ML model used
[31]		PM, CO, NH3, Temperature, Humidity	ARIMA Model	Integrated IoT and machine learning for air quality observation	Limited to specific parameters and one ML model

3. Methodology

The proposed methodology is illustrated in Figure 1 which visually represents this paper's sequential steps and interactions. It provides a clear and structured overview of how tasks are completed, decisions are made, and information flows within the system.

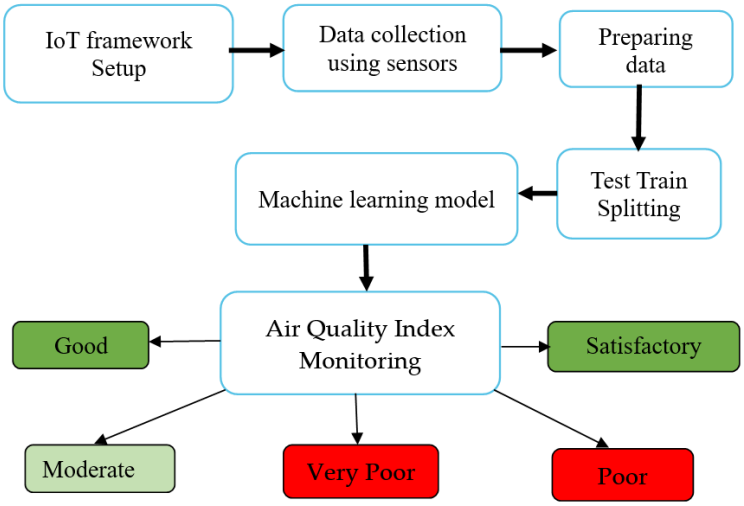

Figure 1. Proposed Methodology

Table 2 provides a detailed summary of the maximum permitted levels of various contaminants in rural and ecological zones throughout different time periods. Permissible amounts are defined using several units of measurement, such as micrograms per cubic meter (μg/m³) for lead (Pb) and particulate matter (PM2.5, PM10), and parts per million (ppm) for gases including SO2, NO2, CO, and O3. These mandated concentration levels are critical in assisting politicians, environmental agencies, and researchers in assessing and regulating air quality in order to protect both public health and the natural environment. Furthermore, they create regulatory benchmarks to maintain acceptable air quality norms in rural and ecological environments.

Table 2. WHO air quality standards where ppm = parts per million, μg/m³ = micrograms per cubic meter

Pollutant Component	Time avg.	Ecological area	Rural area
CO	8-hr	5 ppm	10 ppm
NO2	Annual	0.020 ppm	0.053 ppm
PM10	Annual	10 μg/m³	20 μg/m³
Pb	Annual	0.5 μg/m³	0.5 μg/m³
PM2.5	Annual	5 μg/m³	10 μg/m³
O3	8-hr	0.080 ppm	0.100 ppm
SO2	24-hr	0.02 ppm	0.05 ppm

3.1. IoT Framework Setup

The MQ135 sensor, MQ131 sensor, MQ9 sensor, PM2.5 and PM10 sensor, and dust sensor are among the commonly utilized sensors for air quality monitoring. MQ-135: Among many air contaminants that may be detected by the MQ135 gas sensor are ammonia, benzene, and carbon dioxide. The MQ135 provides essential data for evaluating the overall quality of the air because of its excellent sensitivity and dependability. The MQ-9: With a focus on identifying dangerous gases like carbon monoxide and methane, the MQ9 sensor improves the system's capacity to recognize and measure other contaminants, allowing for a more thorough examination of the air quality. The connectivity of the MQ-9 is similar resembles that of the MQ-135. As for the dust sensor: Our IoT framework includes a dust sensor whose sole purpose is to gauge the number of airborne particles present. This information is essential for determining the total amount of particle pollution and for building a more thorough picture of the air quality in your immediate

area. The air quality is greatly threatened by particulate matter, and our architecture takes this into account by including PM2.5 and PM10 sensors. A thorough examination of airborne particulate pollution and its possible health effects is made possible by these sensors, which provide exact measurements of small particles and respirable coarse particles, respectively. Our IoT framework is built for dependable performance, simple usage, and seamless integration. The gathered data is sent to a cloud platform or central server so that it can be seen and examined instantly. Users are empowered to make educated decisions, carry out focused interventions, and help create living environments that are healthier and more sustainable with the use of intuitive dashboards and comprehensive reports. Take advantage of our cutting-edge IoT framework right now for unmatched air quality insights and a proactive strategy for pollution control. The complete hardware setup can be observed in Figure 2.

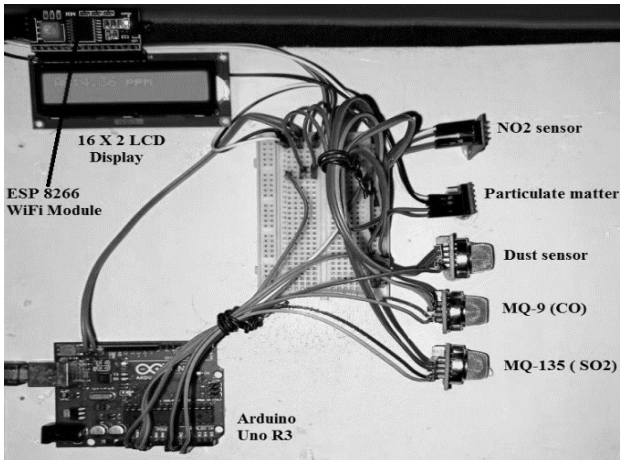

Figure 2. Combined circuits with all the sensors

3.2. Data Collection

IoT devices were employed to gather real-time data on air pollution in Gazipur and Dhaka. Particulate matter 2.5, particulate matter 10, ozone, carbon monoxide, nitrogen dioxide, and other quantities were detected by fitting appropriate sensors into the hardware design. The Arduino IDE code was released to make connecting with Internet of Things devices easier. The Excel data streamer was started, and then the relevant Excel file was accessed. The IoT devices were connected to the PC using the appropriate COM port selected by the Arduino IDE. The "Start Data" option on the data streamer toolbar allowed real-time data to stream into the Excel sheet.

IoT devices were able to continuously monitor and record air pollutants after they pressed the "Record Data" button to begin data collection. The "Stop Recording" button was clicked in order to end data collection. The "Stop Data" feature was used to halt data transmission between the computer and Internet of Things devices. To make analysis and interpretation easier, the collected data were stored in a designated file location. Future studies with decision-making processes on air pollution in Dhaka and Gazipur were made possible by this strategy. Additionally, it guaranteed a thorough evaluation and ongoing observation of air quality measurements. Table 3 shows the sample of the dataset with parameters.

Table 3. Sample of the dataset with parameters.

SO2	NO2	CO	O3	PM2.5	PM10
0.04	0.059	1.2	0.0525	57	73
0.04	0.058	1.2	0.0525	59	71
0.04	0.056	1.2	0.0525	59	70
0.04	0.056	1.2	0.0525	58	70
0.03	0.051	1.2	0.0525	61	69
0.03	0.046	1.1	0.0525	61	70

0.03	0.049	1.1	0.0525	57	66
0.03	0.045	1	0.0525	60	71
0.04	0.047	1.1	0.0525	60	72
0.03	0.047	1.1	0.0525	63	74
0.04	0.045	1	0.04	68	76

3.3. Preparing Data

The dataset is fed into our Python code before the necessary libraries are loaded. The dataset comprises readings from sensors measuring SO2, NO2, CO, O3, PM2.5, and PM10 levels at the current location, alongside their respective air quality ratings. As a result, there are eleven columns in this dataset, and the number of rows varies according to the length of the data record. Lastly, we export this dataset from Excel as a.csv file. Due to certain station data being categorized as NAN or unavailable, the source's data contains more noise than usual. We pre-processed the data to eliminate outliers in order to mitigate this. Compared to consistent and reliable findings, these anomalous data points show greater fluctuation and are mainly caused by broken sensors or transmission faults. To find the top and lower quartile ranges, we employed boundary value analysis (BVA) on the data set to identify these outlier points.

Figure 3 shows correlation heatmap of the collected air quality factors. A correlation heatmap serves as a visual representation illustrating the correlations among multiple variables through a color-coded matrix. It operates like a color chart that delineating the degree of association between various variables. Correlation values typically range from -1 to 1., with 1 signifies a perfect positive correlation, -1 signifies a perfect negative correlation, and 0 implies no correlation between variables.

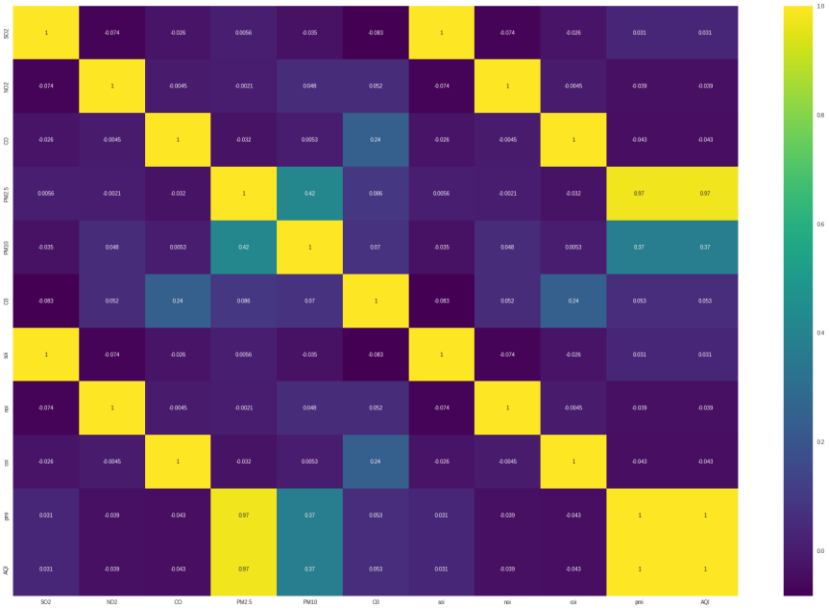

Figure 3. Correlation heatmap of all the sensors.

3.4. Test Train Splitting

The training set is employed to instruct the model, whilst the testing set determines how effectively the model can generalize newly discovered data. When 20–30% of the data is used for testing and 70–80% for training, the best outcomes are obtained. To utilize an 80/20 split for training and testing, one needs to import the train_test_split package from Sci-kit. Machine learning techniques for air pollution regression problems include random forest, decision tree, and linear regression. The logistic regression, decision tree, random forest, and KNN models were selected for classification in order to forecast air pollution.

3.5. Machine Learning Model

Predicting air pollution can be aided by the application of machine learning algorithms. Machine learning, a subdivision of artificial intelligence, facilitates implementations to dependably predict results without the need for custom coding. Machine learning algorithms use available data as input to anticipate new results. With machine learning (ML), a computer program can process large amounts of data and draw conclusions without further guidance.

3.5.1. Linear Regression (LR)

In straightforward terms, LR entails matching a straight line via a scatter plot of data points such that the line most accurately depicts the trend or pattern in the data. The formula 1 of the line is commonly depicted.

$$y = mx + b \quad (1)$$

where, y = dependent or outcome variable, x = independent or predictor variable, m = slope of the line, b = the y-intercept of the line.

3.5.2. Decision Tree (DT)

A DT is a discontinuous, nonlinear structure that serves as a model by making decisions depending on an input set of attribute values. This kind of model is adaptable and can handle both regression and classification problems. It is a member of the supervised learning class of machine learning algorithms. To create judgments depending on the incoming data, decision trees use a sequence of operations [32].

3.5.3. Random Forest (RF)

Several decision trees are combined in RF, an ensemble learning approach in machine learning, to improve prediction accuracy. During training, this approach builds many decision trees and combines their predictions to furnish a more accurate and dependable outcome. A haphazard segment of features and a haphazard sample of the data with replacement (bootstrap sampling) are used to train each tree in the haphazard forest. Either by averaging the forecasts or by choosing the majority vote from all the trees in the forest, the final prediction is obtained for both regression and classification issues. Renowned for its ability to manage high-dimensional data, address missing values, and reduce overfitting risk, Random Forest is a robust and widespread technique for a variety of machine-learning applications [33].

3.5.4. K-Nearest Neighbours (KNN)

KNN is a lazy learning algorithm. It does not construct a clear model while the method is being trained. Rather, it keeps the whole training dataset and uses it to generate predictions when testing. KNN is easy to build, doesn't require any presumptions regarding the distribution of the underlying data, and may be used for regression tasks as well as binary and multiclass classification [34]. However, KNN can provide computational difficulties since it may need a lot of resources, and its effectiveness may depend on variables like the distance metric and K selection. Moreover, it could not function well with big datasets or noisy data.

3.5.5. Gradient Boosting (GB)

A potent machine learning technique called gradient boosting is famous for its capacity to generate precise forecasts by repeatedly improving poor learners into robust predictive models. Gradient boosting, in contrast to some other algorithms, concentrates on improving the performance of the model by reducing mistakes in later rounds. An ensemble of decision trees is progressively constructed, with each additional tree being taught to fix the mistakes caused by the ones that came before it. Gradient boosting may be utilized for both regression and classification tasks and is very skilled at managing a wide range of data formats. In addition to being adaptable in managing outliers and missing values, it typically performs well even with noisy data [35].

3.6. Air Quality Index Monitoring

In this part, we find out the AQI index into five classes named Good, Satisfactory, Moderate, Poor and Very Poor. After validating the experimental work, in Kuril Bishwa Road, air is mostly moderated and Uttara's air is good. And Tongi's air is poor.

4. Result and Discussion

This section showcases the discoveries and results of the investigation. It typically involves statistical analyses, simulations, visual representations, and textual explanations to present the outcomes clearly and systematically.

4.1. Simulation and Calculation of Air Quality Index (AQI)

According to the linear segmentation principle, the equation 2 can be used for calculating the AQI.

$$I_i = \frac{I_{max} - I_{min}}{B_{max} - B_{min}} \times \left(C_p - B_{min}\right) + I_{min} \quad (2)$$

Here, B_{max} and B_{min} are the breakpoints of AQI where B_{max} is greater than or equal to the given concentration and B_{min} is less than or equal to the given concentration, I_{max} is equal to AQI value corresponding to B_{max} , I_{min} is equal to AQI value corresponding to B_{min} , C_p is equal pollutant concentration.

Table 4 and Table 5 exhibit the AQI categories as well as the ranges for pollutants. PM2.5, NO2, O3, and SO2 are all measured in g/m3, and CO is measured in mg/m3.

Table 4. AQI Category Range

Range	AQI Category
0-50	Good
51-100	Satisfactory
101-200	Moderate
201-300	Poor
301-400	Very Poor

Table 5. Air Pollution factors Category Range where CO is in mg m^{-3} and others is in μg $^{m-3}$

Category	PM2.5	NO_2	O_3	CO	SO_2
Good	0-30	0-40	0-50	0-1.0	0-40
Satisfactory	31-60	41-80	51-100	1.1-2	41-80
Moderate	60-90	81-180	101-168	2.1-10	81-380
Poor	91-120	181-280	169-208	10.1-17	381-800
Very Poor	121-250	281-400	209-748	17.1-34	801-1600

There are other methods or formulas for calculating AQI; however, this is one of the more popular or often used forms. Figure 4 displays the AQI range and associated health effects.

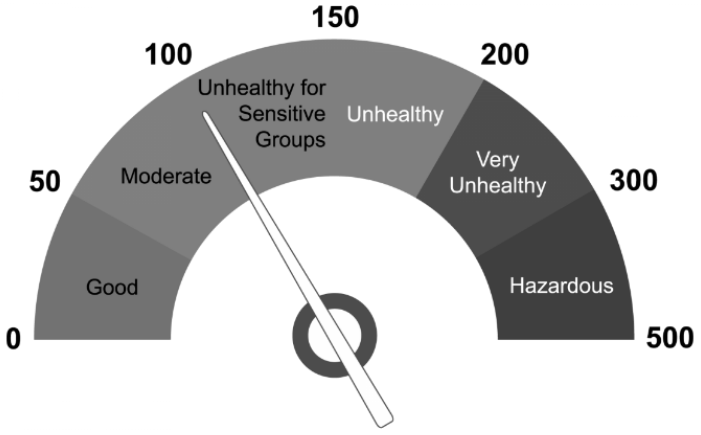

Figure 4. WHO AQI standard level

The data acquired includes many columns of sensor readings from Dhaka's Tongi, Uttara, and Kuril regions. Each column depicts pollutant data, demonstrating the relationship between pollutant concentrations and their corresponding indicators. Algorithm 1 shows an example computation of the Air Quality Index (AQI) for carbon monoxide.

Algorithm 1. Carbon Monoxide Sub-Index measurement

```
def co_sub_index(s1):
  if s1 <= 1:
    return s1 * 50 / 40
  elif s1 <= 2:
    return s1 + (s1 - 1) * 50 / 40
  elif s1 <= 10:
    return 100 + (s1 - 2) * 100 / 8
  elif s1 <= 17:
    return 200 + (s1 - 10) * 100 / 7
  elif s1 <= 34:
    return 300 + (s1 - 17) * 100 / 17
  elif s1 > 34:
return 400 + (s1 - 34) * 100 / 17
  else:
    return 0;
```

Algorithm 2. AQI_Range Calculation

```
def AQI7level(s1):
  if s1 <= 50:
    return "Good"
  elif s1 <= 100:
    return "Satisfactory"
  elif s1 <= 200:
    return "Moderate"
  elif s1 <= 300:
    return "Poor"
  elif s1 <= 400:
    return "Very Poor"
  else:
    return np.NaN
```

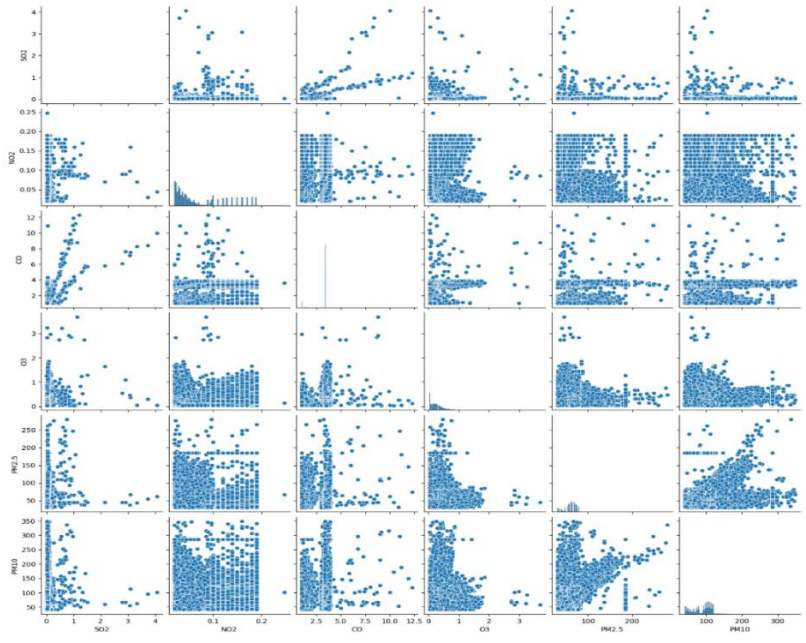

Figure 5. Pair plot for all sensors

Using the same computation approach as for carbon monoxide (CO) AQI, analogous calculations were used to determine the individual AQI values for sulfur dioxide (SO2), nitrogen dioxide (NO2), ozone (O3), particulate matter with a diameter of 2.5 micrometers or smaller (PM2.5), and particulate matter with a diameter of 10 micrometers or smaller. The highest pollution value recorded at a certain site is used to calculate the AQI. The AQI is calculated using the greatest sub-index among all contaminants found.

To determine the AQI range, we used an if-else statement that defined the range based on the WHO AQI standard. As a result, the final dataset was adjusted to include the AQI range. Algorithm 2 outlines the procedure and outcomes of calculating the AQI range.

A pair plot of the collected sensor data is shown in Figure 5, which aims to highlight the relationships between various pairs of values acquired from various sensors. By displaying patterns, trends, and connections between the measured variables, this plot helps the viewer visually examine the data. Understanding the interconnectedness of data from various sensors may be gained by examining the location and distribution of dots on the plot. But a thorough comprehension requires having access to the precise number or fine-grained details of the sensor data.

4.2. Evaluation

We partition the dataset into response and predictor variables for the application of machine learning models. Within regression analysis, the Atmospheric Quality Index (AQI) serves as the response variable, while in classification tasks, AQI functions as a predictor with AQI_Range as the target variable. Subsequently, the dataset undergoes a split into training and testing sets prior to model implementation.

Figure 6 depicts the AQI values based on their range. We used numerous assessment measures to assess the effectiveness of the regression model, including root mean square error (RMSE), mean absolute error (MAE), r2 score, and R-squared. A low RMSE combined with a high R-squared value indicates that the model is accurate. Furthermore, the classification model's assessment criteria include the precision score, recall score, F1 score, and Kappa score. A high Kappa value suggests that the model is highly accurate.

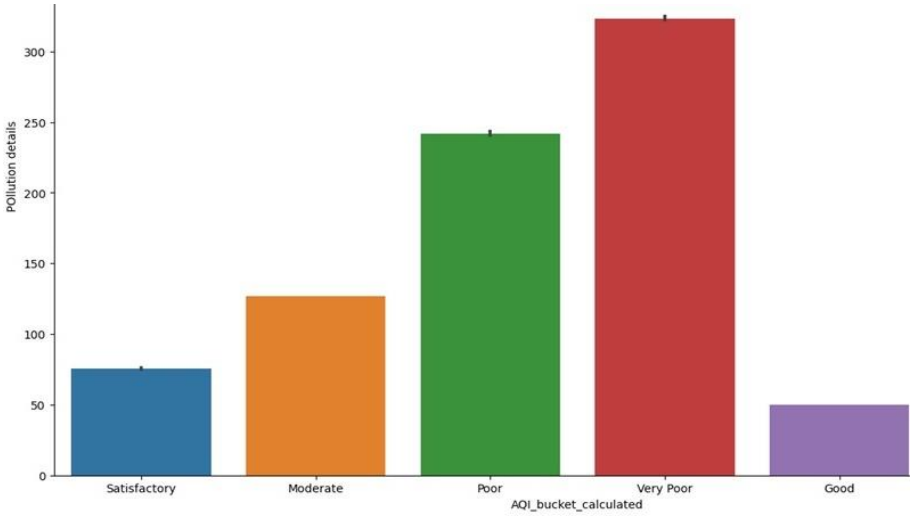

Figure 6. AQI values according to AQI Range

Additionally, we assessed the precision of the classifier model by calculating accuracy scores for all the classification models utilized in the investigation. With this model, it is feasible to manually input the predictor values and obtain the predicted atmospheric quality range. As indicated by the findings depicted in Figure 7, it is apparent that the decision tree model tends overfitting, whereas the Random Forest model demonstrates superior performance as a regression model for forecasting purposes.

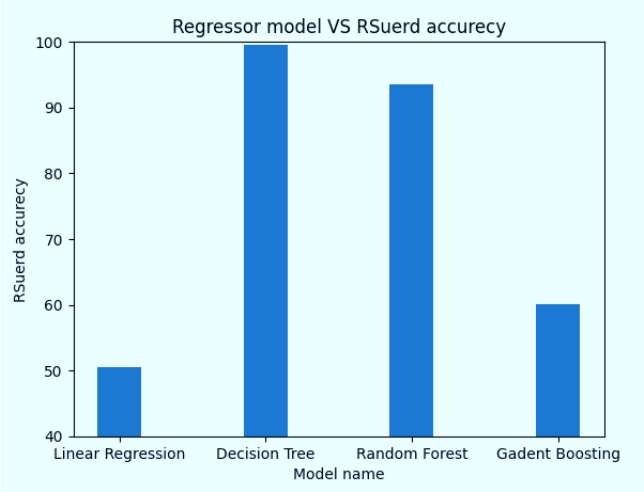

Figure 7. R squared values of the evaluated model

Figure 8 shows the accuracy of all the classification models. From the Figure 8, it is cleared that Random Forest outperforms the accuracy.

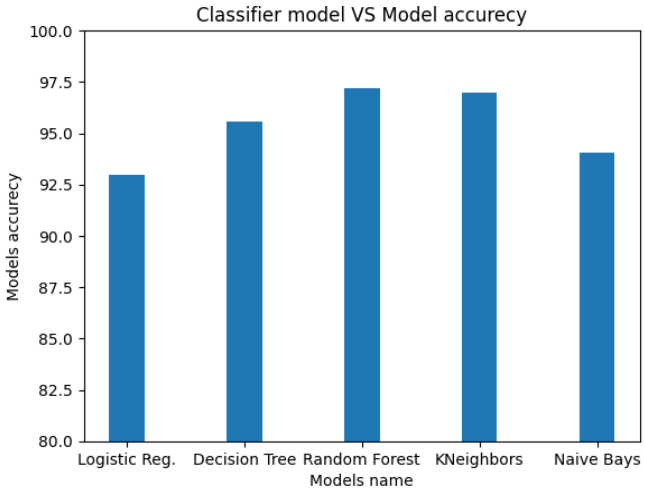

Figure 8. Accuracy score of the model

Table 6 illustrates the regression result of the models. A moderate level of performance was observed with linear regression, resulting in a Mean Absolute Error (MAE) of 9.43 and a Root Mean Squared Error (RMSE) of 20.04. The model's R-squared (R2) value was recorded at 0.505, elucidating approximately 50.5% of the variance. The Decision Tree Regressor exhibited more pronounced errors, displaying an MAE of 11.76 and an RMSE of 25.34. However, it attained a notably high R2 score of 0.995, signifying an excellent fit to the dataset.

Table 6. Regression results for ML model

Model	RMSE	MAE	R Squared
LR	20.04	9.43	0.505
DTR	25.34	11.76	0.995
RFR	18.73	9.00	0.930
GB	17.86	8.34	0.60

Random Forest Regressor (RFR) outperforms by achieving the value of MAE of 9.00 and an RMSE of 18.73. The model takes into account almost 93 percent of the variation with the R2 value of 0.930. Gradient boosting was the technique that performed the best and had the fewest mistakes (MAE of 8.34 and RMSE of 17.86). Its R2 score of 0.60 indicates that the model can elucidate close to 60 percent of the variance.

Significantly, Gradient Boosting exhibited the lowest errors and the most favourable R2 value among the models evaluated, demonstrating its superiority in addressing the regression problem. These findings underscore the effectiveness of gradient boosting in regression analysis.

Table 7. Evaluation Metrics score table

Model	Precision (%)	Recall (%)	F1 (%)	Kappa (%)	Accuracy (%)
LoR	93.0	93.0	93.0	20.6	93.0
DTC	95.4	95.4	95.4	67.1	95.4
RFC	97.2	97.2	97.2	78.7	97.2
KNN	97.0	97.0	97.0	76.0	97.0
NB	94.0	94.0	94.0	65.0	94.0

Table 7 showcases assessment metric scores for different models, encompassing Precision score, Recall score, F1 score, and Kappa score. The Random Forest Classifiers displayed the highest scores across all metrics, highlighting robust performance in classification tasks. The Random Forest classifier (RFC) beats the performance results by achieving an accuracy of 97.2%, a Precision of 97.2%, an F1 of 0.972%, a Recall of 0.972%, and a Kappa of 0.787%. The most accurate algorithm is RFC, which outperforms all others. These findings offer valuable insights into the efficacy of various algorithms, suggesting that RFC could be particularly suitable for the classification task at hand. When selecting appropriate algorithms for similar tasks in the future, researchers and practitioners may find these results to be informative and beneficial.

4.3. Prediction

Figure 9 illustrates a visual depiction of the observed and forecasted Atmospheric Quality Index (AQI) employing the decision tree regression model. Likewise, AQI prognostications can be generated utilizing all the other regression models.

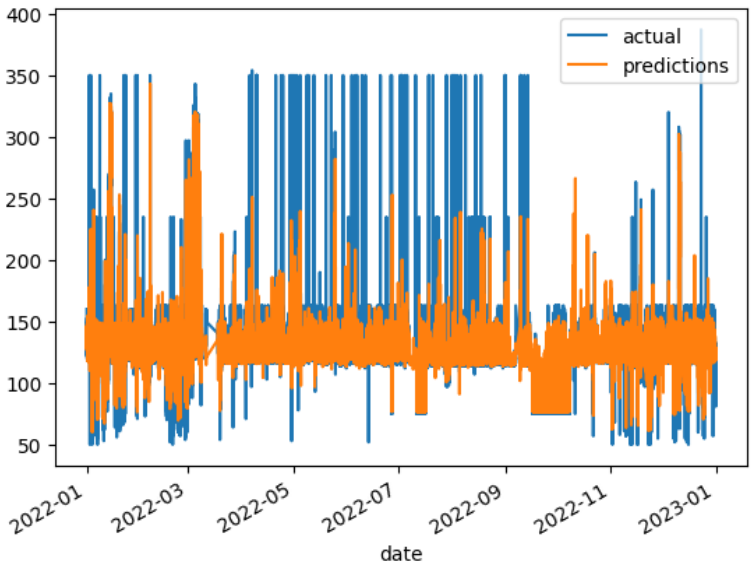

Figure 9. Observed value vs. forecasted value of AQI

4.4. Comparison with Existing Works

This section presents a comparison between our research and previous studies on air pollution detection and forecasting. It is essential to underscore that our methodology utilizes distinct datasets compared to earlier investigations. Unlike other research endeavours relying on secondary data, our study exclusively utilizes real-time data. Additionally, our original dataset has not been previously employed for similar purposes. Our study distinguishes itself through its distinctive dataset and approach, despite the presence of several notable studies focusing on air pollution detection utilizing IoT devices, and prediction

employing machine learning or deep learning methodologies. To provide a comprehensive overview, we present a comparative analysis in Table 8.

Table 8. A Comparison with previous works

Paper	ML\DL Used	Method	Accuracy
[29]	ML	IoT and Machine Learning	
[45]	DL	IoT and Deep Learning	
[46]		IoT Based Proposed System	
[47]		IoT Based Proposed System	
[28]	ML	IoT and Machine Learning	
[27]	ML	IoT and Machine Learning	90%
Our Work	ML	IoT and Machine Learning	97.2%

5. Conclusion and Future works

This study involved the real-time gathering of air pollutant data, including particulate matter 10, sulfur dioxide, ozone, nitrogen dioxide, particulate matter 2.5, and carbon monoxide, using IoT devices installed in three locations in Dhaka and Gazipur, Bangladesh. The dataset was then processed using machine learning methods such as regressors and classification models. Notably, the Random Forest classification model performed exceptionally well, with an accuracy of 97.2%. In regression analysis, MAE, RMSE, and R-squared were measured. Gradient Boosting outperformed linear regression in terms of R-squared and MAE. The findings highlight the potential of IoT and machine intelligence in providing complete information on air pollution levels. This research has the potential to inform the establishment of effective policies and activities to reduce the negative effects of air pollution on the environment and public health.

Acknowledgment

This research is funded by Woosong University Academic Research 2024.

References

[1] Tanisha Madan, Shrddha Sagar and Deepali Virmani, "Air Quality Prediction using Machine Learning Algorithms –A Review", In *Proceedings of the 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 19-20 December 2020, Greater Noida, India, pp. 140-145, Published by IEEE, DOI: 10.1109/ICACCCN51052.2020.9362912, Available: https://ieeexplore.ieee.org/abstract/document/9362912.

[2] Rajib Saha, SNM Azizul Hoque, MMR Manu and Aminul Hoque, "Monitoring air quality of Dhaka using IoT: Effects of COVID-19", In *Proceedings of the 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, 5-7 January 2021, Dhaka, Bangladesh, pp. 715–721, Published by IEEE, DOI: 10.1109/ICREST51555.2021.9331026, Available: https://ieeexplore.ieee.org/abstract/document/9331026.

[3] Md Monirul Islam, Jahid Hasan Rony, Md Nasim Akhtar, Shalah Uddin Perbhez Shakil and Jia Uddin, "Water monitoring using internet of things", In *Proceedings of the 7th Internet of Things for Smart Environments 2022*, 26-30 June 2022, Porto, Portugal, pp. 59–69, Published by Springer, DOI: 10.1007/978-3-030-68452-511, Available: https://link.springer.com/chapter/10.1007/978-3-031-09729-4_4.

[4] Md Monirul Islam, Jia Uddin, Mohammod Abul Kashem, Fazly Rabbi and Md Waliul Hasnat, "Design and implementation of an IoT system for predicting aqua fisheries using Arduino and KNN", In *Proceeding of the 12th International Conference in Intelligent Human Computer Interaction (IHCI)*, 24–26 November 2020, Daegu, South Korea, pp. 108–118, Published by Springer Nature, DOI: 10.1007/978-3-030-68452-5_11, Available: https://link.springer.com/chapter/10.1007/978-3-030-68452-5_11.

[5] Md Monirul Islam, Mohammad Abul Kashem and Jia Uddin, "An internet of things framework for real-time aquatic environment monitoring using an Arduino and sensors", *International Journal of Electrical and Computer*

Engineering, ISSN 2088-8708, vol. 12, no.1, pp. 826-833, 2022, Published by IAES, DOI: 10.11591/ijece.v12i1, Available: https://ijece.iaescore.com/index.php/IJECE/article/view/24327/15391.

[6] Jahid Hasan Rony, Nazmul Karim, MD Abdur Rouf, Md Monirul Islam, Jia Uddin *et al.*, "A cost-effective IoT model for a smart sewerage management system using sensors", *J - Multidisciplinary Scientific Journal*, vol. 4, no. 3, pp. 356–366, 2021, Published by MDPI, ISSN 2571-8800, DOI**:** 10.3390/j4030027, Available: https://www.mdpi.com/2571-8800/4/3/27.

[7] Yash Mehta, MM Manohara Pai, Sanoop Mallissery and Shwetanshu Singh, "Cloud enabled air quality detection, analysis and prediction-a smart city application for smart health", In *Proceeding of the 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, 15-16 March 2016, Muscat, Oman, pp. 1–7, Published by IEEE, DOI**:** 10.1109/ICBDSC.2016.7460380, Available: https://ieeexplore.ieee.org/abstract/document/7460380.

[8] Jiangshe Zhang and Weifu Ding, "Prediction of air pollutants concentration based on an extreme learning machine: the case of Hong Kong", *International journal of environmental research and public health*, vol. 14 no. 2, pp. 114, 2017, Publisher MDPI, ISSN: 1661-7827, DOI**:** 10.3390/ijerph14020114, Available: https://www.mdpi.com/1660-4601/14/2/114.

[9] Moolchand Sharma, Samyak Jain, Sidhant Mittal and Tariq Hussain Sheikh, "Forecasting and prediction of air pollutants concentrates using machine learning techniques: The case of India", In *Proceeding of the IOP Conference Series: Materials Science and Engineering*, 24 October 2020, Rajpura, India, vol. 012123, pp. 1-22, IOP Publishing, DOI**:** 10.1088/1757-899X/1022/1/012123, Available: https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012123/meta.

[10] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu and Gang Xie, "Air quality prediction: Big data and machine learning approaches", *International Journal of Environmental Science Development*, Singapore, ISSN: 2010-0264, vol. 9, no. 1, pp. 8–16, 2018, DOI: 10.1109/ICREST51555.2021.9331026, Available: https://ieeexplore.ieee.org/abstract/document/9331026.

[11] Peter Kim Streatfield and Zunaid Ahsan Karar, "Population challenges for Bangladesh in the coming decades", *Journal of health, population, and nutrition*, Publisher BMC, USA, ISSN 2072-1315, vol. 26, no. 3, pp. 261, 2008, PMID: 18831223, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2740702/.

[12] Xiang Li, Ling Peng, Yuan Hu, Jing Shao and Tianhe Chi, "Deep learning architecture for air quality predictions", *Environmental Science and Pollution Research*, Publisher Springer, USA, ISSN 0944-1344, vol. 23 pp. 22408–22417, 2016, DOI**:** 10.1007/s11356-016-7812-9, Available: https://link.springer.com/article/10.1007/s11356-016-7812-9.

[13] Tara L Greaver, Timothy J Sullivan, Jeffrey D Herrick, Mary C Barber, Jill S Baron *et al.*, "Ecological effects of nitrogen and sulfur air pollution in the us: what do we know?", *Frontiers in Ecology and the Environment*, Publisher Frontiers, USA, ISSN 2296-701X, vol. 10, no. 7, pp. 365–372, 2012, DOI**:** 10.1890/110049, Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/110049.

[14] Venkat Rao Pasupuleti, Pavan Kalyan and Hari Kiran Reddy, "Air quality prediction of data log by machine learning", In *Proceeding of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 06-07 March 2020, Coimbatore, India, pp. 1395–1399, Published by IEEE, DOI**:** 10.1109/ICACCS48705.2020.9074431, Available: https://ieeexplore.ieee.org/abstract/document/9074431.

[15] Shalah Uddin Perbhez Shakil, Mohammod Abul Kashem, Md Monirul Islam, Nasim Mahmud Nayan and Jia Uddin, "Investigation of Air Effluence Using IoT and Machine Learning", In *Proceedings of the International Conference for Emerging Technologies in Computing*, 16-17 August 2023, Southend-on-Sea, UK, pp. 183-202, Published by Springer Nature, DOI: 10.1007/978-3-031-50215-6_12, Available: https://link.springer.com/chapter/10.1007/978-3-031-50215-6_12.

[16] S Jeya and L Sankari, "Air pollution prediction by deep learning model", In *Proceeding of the 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 06-07 March 2020, Coimbatore, India, pp. 736–741, IEEE, DOI**:** 10.1109/ICICCS48265.2020.9120932, Available: https://ieeexplore.ieee.org/document/9074431.

[17] Fettah Eren and Serefnur Ozturk, "Evaluation of the effect of air pollution on cognitive functions, cognitive decline, and dementia", *Annals of Indian Academy of Neurology*, ISSN: 1998-354, vol. 25, no. l1, pp. S9-S14, 2022, DOI**:** 10.4103/aian.aian_453_22, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9540830/.

[18] Marius Dobrea, Andreea Bădicu, Marina Barbu, Oana Subea, Mihaela Bălănescu, Geroge Suciu, Andrei Bîrdici, Oana Orza and Ciprian Dobre, "Machine learning algorithms for air pollutants forecasting", In *Proceeding of the IEEE 26th International Symposium for Design and Technology in Electronic Packaging (SIITME)*, 2020, Pitesti, Romania,

pp. 109–113, Published by IEEE, DOI: 10.1109/SIITME50350.2020.9292238, Available: https://ieeexplore.ieee.org/document/9292238.

[19] Rakesh Kumar Saini, Hemraj Saini and Surjeet Singh, "Air pollution quality monitoring system using internet of things for smart cities", *Turkish Journal of Computer and Mathematics Education (TURCOMAT*), ISSN: 3048-4855, vol. 11, no. 2, pp. 1077–1092, 2020. DOI: 10.17762/turcomat.v11i2.12542, Available: https://www.turcomat.org/index.php/turkbilmat/article/view/12542.

[20] U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha and G. Kedam, "A Machine Learning Model for Air Quality Prediction for Smart Cities", In *Proceeding of the International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 21-23 March 2019, Chennai, India, pp. 452-457, Published by IEEE, DOI: 10.1109/WiSPNET45539.2019.9032734, , Available: https://ieeexplore.ieee.org/abstract/document/9032734.

[21] Devon C Payne-Sturges, Melanie A Marty, Frederica Perera, Mark D Miller, Maureen Swanson *et al.*, "Healthy air, healthy brains: advancing air pollution policy to protect children's health", *American journal of public health*, American Public Health Association, USA, ISSN: 0090-0036, vol. 109, no. 4, pp. 550–554, 2019, DOI: 10.2105/AJPH.2018.304902, Available: https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2018.304902.

[22] Gagan Parmar, Sagar Lakhani and Manju K Chattopadhyay, "An IoT based low cost air pollution monitoring system", In *Proceeding of the 2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE)*, 27-29 October 2017, Bhopal, India, pp. 524–528, Published by IEEE, DOI: 10.1109/RISE.2017.8378212, Available: https://ieeexplore.ieee.org/abstract/document/8378212.

[23] H Ali, JK Soe and Steven R Weller, "A real-time ambient air quality monitoring wireless sensor network for schools in smart cities", In *Proceeding of the IEEE first international smart cities conference (ISC2)*, 25-28 October 2015, Guadalajara, Mexico, pp. 1–6, Published by IEEE, DOI: 10.1109/ISC2.2015.7366163, Available: https://ieeexplore.ieee.org/abstract/document/7366163.

[24] Mario Molinara, Marco Ferdinandi, Gianni Cerro, Luigi Ferrigno and Ettore Massera, "An end to end indoor air monitoring system based on machine learning and sensiplus platform", IEEE Access, ISSN: 2169-3536, Published by IEEE, USA, vol. 8, pp. 72204–72215, 2020, DOI: 10.1109/ACCESS.2020.2987756, Available: https://ieeexplore.ieee.org/abstract/document/9064782.

[25] JunHo Jo, ByungWan Jo, JungHoon Kim, SungJun Kim and WoonYong Han, "Development of an IoT-based indoor air quality monitoring platform", *Journal of Sensors*, Published by Hindawi, UK, ISSN 1687-7268, vol. 2020, pp. 1–14, 2020, DOI: 10.1155/2020/8749764, Available: https://www.hindawi.com/journals/js/2020/8749764/.

[26] B.R. Varshitha Chandra, Pooja G. Nair, Risha Irshad Khan and B.S. Mahalakshmi, "Air Quality Monitoring System Using Machine Learning and IoT", In *Proceeding of the Congress on Intelligent Systems (CIS 2020)*, 5-6 September 2020, New Delhi, India, ISSN 214-565, vol. 1, Springer Nature, DOI: 10.1007/978-981-33-6981-8_54, Available: https://doi.org/10.1007/978-981-33-6981-8_54.

[27] Kinnera Bharath Kumar Sai, Somula Rama Subbareddy and Ashish Kumar Luhach, "IOT based air quality monitoring system using MQ135 and MQ7 with machine learning analysis", *Scalable Computing: Practice and Experience*, Published by University De Vest, Romania, ISSN: 1895-1767, vol. 20, no. 4, pp. 599-606, 2019, DOI: 10.12694/scpe.v20i4.1561, Available: https://www.scpe.org/index.php/scpe/article/view/1561.

[28] Harsh Gupta, Dhananjay Bhardwaj, Himanshu Agrawal, Vinay Anand Tikkiwal and Arun Kumar, "An IoT Based Air Pollution Monitoring System for Smart Cities", In *Proceeding of the IEEE International Conference on Sustainable Energy Technologies and Systems (ICSETS)*, 26 February-01 March, Bhubaneswar, India, 2019, pp. 173-177, DOI: 10.1109/ICSETS.2019.8744949, Available: https://ieeexplore.ieee.org/abstract/document/8744949.

[29] Zhiyuan Wu, Yue Wang and Lin Zhang, "Msstn: Multi-scale spatial temporal network for air pollution prediction", In *Proceeding of the 2019 IEEE International Conference on Big Data (Big Data)*, 9-12 December 2020, Los Angeles, USA, pp. 1547–1556. Published by IEEE, DOI: 10.1109/BigData47090.2019.9005574, Available: https://ieeexplore.ieee.org/abstract/document/9005574.

[30] Bu Zhao, "Urban air pollution mapping using fleet vehicles as mobile monitors and machine learning", *Environmental science technology*, American Chemical Society, USA, ISSN: 0013-936X, vol. 55, no. 8, pp. 5579–5588, 2021, DOI: 10.1021/acs.est.0c08034, Available: https://pubs.acs.org/doi/abs/10.1021/acs.est.0c08034.

[31] Mohammed Rakib, "IoT based air pollution monitoring prediction system", In *Proceeding of the International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 26-27 February 2022, Chittagong, Bangladesh, pp. 184–189, Published by IEEE, DOI: 10.1109/ICISET54810.2022.9775871, Available: https://ieeexplore.ieee.org/abstract/document/9775871.

[32] Shabnom Mustary, Mohammod Abul Kashem, Md Nurul Islam Khan, Faruq Ahmed Jewel, Md Monirul Islam *et al.*, "Leach based wsn classification using supervised machine learning algorithm", In *Proceeding of the International Conference on Computer Communication and Informatics (ICCCI)*, 27-29 January 2021, Coimbatore, India, pp. 1–5, IEEE, DOI: 10.1109/ICCCI50826.2021.9457001, Available: https://ieeexplore.ieee.org/abstract/document/9457001.

[33] Md Monirul Islam, Mohammod Abul Kashem and Jia Uddin, "Fish survival prediction in an aquatic environment using random forest model", *International Journal of Artificial Intelligence (IJ-AI)*, Published by IAES, Indonesia, ISSN 2089-4872, vol. 10, no. 3, pp. 614-622, 2021. DOI: 10.11591/ijai.v10.i3.pp614-622, Available: https://ijai.iaescore.com/index.php/IJAI/article/view/20963.

[34] Mahmudul Alam, Md Monirul Islam, Md Rokunojjaman, Sharmin Akter, Md Belal Hossain *et al.*, "Electrocardiogram signal analysis based on statistical approaches using k-nearest neighbor", In *Proceeding of the International Conference on Bangabandhu and Digital Bangladesh*, 31 December 2021, United International University, Dhaka, Bangladesh, pp. 148–160, Springer Nature, DOI**:** 10.1007/978-3-031-17181-9_12, Available: https://link.springer.com/chapter/10.1007/978-3-031-17181-9_12.

[35] Agboeze Jude and Jia Uddin, "Explainable Software Defects Classification Using SMOTE and Machine Learning", *Annals of Emerging Technologies in Computing (AETiC)*, Published by IAER, UK, vol. 8, no. 1, pp. 35-49, 1st January 2024, DOI: 10.33166/AETiC.2024.01.004, Available: http://aetic.theiaer.org/archive/v8/v8n1/p4.pdf.