Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

AIR QUALITY INDEX FORCASTING USING IOT & MACHINE LEARNING

Mrs. N. Sowjanya Kumari1, Gonnabattula Usha sai kumari2, Imandi Yamini3, Gollivilli Thanmayi4, Guthala Varshini5
1,2,3,4,5 Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

ABSTRACT

Recognizing the paramount importance of air quality for both society and individuals, the integration of machine learning algorithms to forecast forthcoming air quality trends, notably through the Air Quality Index (AQI), emerges as a pivotal endeavor. Therefore, it is strongly advocated to incorporate machine learning techniques such as Random Forest, Support Vector Machine, Genetics, and K-Nearest Neighbors into AQI prognostication endeavors. This amalgamation entails the real-time aggregation of data from Internet of Things (IoT) sensors, enabling continuous environmental surveillance and model refinement as necessitated. Genetic algorithms play a pivotal role in expeditiously identifying optimal solutions within the search space, thereby fostering iterative optimization and gradual model enhancement. The applications of AQI forecasting are diverse, encompassing environmental monitoring, policy formulation, healthcare management, urban planning, and public health protection. This holistic approach not only ensures proactive environmental stewardship but also empowers decision-makers with actionable insights to effectively address contemporary air quality challenges.
Keywords: Extreme Learning Machine (ELM), Air Quality Index (AQI), Genetic Algorithm (GA), Forecasting, Internet of Things (IoT) Integration, Environmental Monitoring, Machine Learning, Optimization.

INTRODUCTION

To effectively oversee air quality in urban areas grappling with pollution, the implementation of a Polluted Air Prediction system[1] is essential. Notable indoor contaminants encompass CO, smoke, CNG, humidity, temperature, and LNG. Given the substantial time individuals spend indoors, ensuring air quality control becomes imperative to mitigate health risks. However, conventional air quality monitoring stations lack portability and cost-effectiveness. To tackle this issue, a sensor-based device harnessing genetic algorithms[4] has been devised.

Air quality profoundly impacts both individual and community well-being, underscoring the integration of machine learning algorithms into Air Quality Index (AQI) forecasting[2]. Employing sophisticated techniques such as Linear Regression, Random Forest, Support Vector Machine, Genetic Algorithms, K-Nearest Neighbors, Logistic Regression[3], and Extreme Learning Machines, alongside real-time data collection through Internet of Things (IoT) sensor networks, facilitates continuous monitoring and predictive modeling of air quality trends.

The primary objective was to engineer a portable, cost-effective device capable of precisely measuring indoor environmental pollutant concentrations. The AQI serves as a metric for short-term air pollution health effects, providing insight into the health implications of local air quality[5]. Widely adopted in developed nations for over three decades, the AQI expeditiously disseminates real-time air quality data, emphasizing the significance of upholding air quality for overall health and well-being.

Linear Regression:
One machine learning technique for determining a relationship between an independent and dependent variable is linear regression. In this case, the variable being predicted is referred to as the dependent variable, and the variable utilized to make the prediction is referred to as the independent variable. The most basic version of the regression function, represented as a linear equation of variables, is said to be provided by linear regression.The regression function's linear nature makes it simple to interpret the parameters.
Random Forest:
utilized to increase prediction accuracy and facilitate group learning. A well-liked ensemble learning technique in machine learning for both regression and classification problems is called Random Forest. During training, it builds a large number of decision trees, from which it outputs the mean prediction (for regression) or mode (for classification) of each tree.
The way Random Forest operates is as follows:
1. Bootstrapping: This technique ensures that every tree in the forest receives slightly different datasets by creating random subsets of the datasets with replacements.
2. Random Feature Selection: A random selection of characteristics is taken into consideration for splitting at each decision tree node. This lessens overfitting and aids in the decorrelation of the forest's trees.
3. Decision Tree Construction: Every tree is expanded to its utmost potential without pruning.
In classification tasks, each tree "votes" for a class; the final output is the mean of all the predictions. This process

is known as voting (classification) or averaging (regression). The average of each tree's predictions is used for regression tasks. Random Forest provides several benefits:
Feature importance, robustness to overfitting, high accuracy, and efficiency on large datasets. Random Forest is extensively utilized for tasks like classification, regression, and anomaly detection in a variety of fields, including bio-informatics, finance, and healthcare. When interoperability is critical or the work involves extremely unbalanced data, it cannot perform as well as alternative approaches
Genetic Algorithms(GA):
It is used to enhance prediction performance and optimize model parameters. The concepts of natural selection and genetics serve as the inspiration for genetic algorithms (GAs), which are optimization approaches. GAs canbe used for a variety of machine learning challenges, not just machine learning ones. They are especially useful in feature selection, hyperparameter optimization, and model optimization. Usually, a genetic algorithm operates as follows:
1. Initialization: A population of possible fixes, such as chromosomes or persons, is created at random. Every option shows a potential fix for the issue.
2. Selection: Based on their fitness, or how well they execute the task being optimized (e.g., a model's accuracy), individuals from the population are chosen for reproduction. Rank-based selection, roulette wheel selection, and tournament selection are examples of common selection techniques.
3. Crossover: A combination of chosen individuals (parents) results in offspring. Usually, this is accomplished by combining elements of the parent solutions to produce new ones. Crossover facilitates effective search space exploration by imitating the natural process of reproduction.
4. Mutation: To encourage variety in the population and avoid an early convergence to less-than-ideal solutions, random modifications are added to the progeny solutions. New genetic material enters the population through mutation.

5. Replacement: Some of the population's less fit members are substituted by the offspring solutions. This keeps the caliber of solutions constant over time by ensuring that only the most fit individuals survive from one generation to the next.
6. Termination: Until a certain number of generations have passed or a termination requirement is satisfied, the process is continued.

The capacity to handle non-differentiable and discontinuous objective functions, explore huge solution spaces, and locate global optima in intricate, multimodal search spaces are only a few of the benefits of genetic algorithms. Genetic algorithms can be applied to machine learning problems including neural architecture search, hyper- parameter optimization, and feature selection. When conventional optimization techniques are ineffective and the search space is vast and complicated, they are especially helpful. To obtain good performance, they do, however, require careful parameter adjustment and can be computationally expensive.

Performance metrics: To analyze the performance of a machine learning model we need some metrics. These metrics are statistical criteria that can be used to measure and monitor the performance of a model. As our thesis deals with prediction, we've considered MAE and RMSE as performance metrics.

Mean absolute error (MAE): MAE is the arithmetic average of the difference between the ground truth and the predicted values. It can also be defined as a measure of errors between paired observations expressing the same phenomenon. It tells us how far the predictions differed from the
actual result. Mathematical representation for MAE is given above. Where, yj = Prediction, $\bar{y}j$= True value, N = Total number of data points

R squared (R2): R square performance metric indicates how well predicted values matches actual values. To compute R squared value, we can use the r2_score function of sklearn.metrics.

Root mean square error (RMSE): RMSE is the square root of the average of the squared difference between the target value and the value predicted by the model. It is square root of mean square error (MSE). The implementation is very much similar to MSE. The machine learning models are validated by comparing the performance metrics. The lower the MAE, RMSE and higher the r-squared, the machine learning mode.

LITERATURE SURVEY

Zhongjie Fu, Haiping Lin, Bingqiang Huang, and Jiana Yao "Research on air quality prediction method in Hangzhou based on machine learning" [1]: The quality of air directly affects human health and well-being,and air pollution has emerged as a major topic of current environmental research. This research predicts Hangzhou's air quality using a Bayesian network model. The model uses SO2,NO2, O3, CO, PM2.5, and PM10 as assessment criteria. The model's output is the AQI value, after which the Bayesian network model is created. Lastly, the model is applied to forecast air quality and contrast it with the measured value. In most situations, the predicted number is close to the actual value, and the data demonstrate that air quality predictions have anaccuracy of over 80%.

Mrs. A. Gnana Soundari Mrs. J. Gnana Jeslin, Akshaya A.C "INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING" [2]: We use machine learning to estimate the air quality index of a certain place to forecast the quality of the air in India. The Indian Air Quality Indexis a commonly used metric to show the levels of pollutants (so2, no2, rpm,etc.) across time. We createda gradient decent boosted multi-variable regression problem model to forecast the air quality index based on historical data from prior years and forecasting over a certain upcoming year. By using cost estimation

for our predictive problem, we increase the model's efficiency. When given historical pollutant concentration data, our model will be able to accurately estimate the air quality index for a whole county, a state, or any bounded region. We were able to outperform the conventional regression models in our model by putting the suggested parameter-reducing formulations into practice. Our model can predict the air quality index for the entire country of India with 96% accuracy using the dataset that is currently available. We also use the AHP MCDM technique to determine the order of preference based on how close the ideal solution is.

Zhongjie Fu, Haiping Lin, Bingqiang Huang, and Jiana Yao "Research on air quality prediction method in Hangzhou based on machine learning" [1]: The quality of air directly affects human health and well-being,and air pollution has emerged as a major topic of current environmental research. This research predicts Hangzhou's air quality using a Bayesian network model. The model uses SO2,NO2, O3, CO, PM2.5, and PM10 as assessment criteria. The model's output is the AQI value, after which the Bayesian network model is created. Lastly, the model is applied to forecast air quality and contrast it with the measured value. In most situations,the predicted number is close to the actual value, and the data demonstrate that air quality predictions have anaccuracy of over 80%.

Mrs. A. Gnana Soundari Mrs. J. Gnana Jeslin, Akshaya A.C "INDIAN AIR QUALITY PREDICTION AND ANALYSIS USING MACHINE LEARNING" [2]: We use machine learning to estimate the air quality index of a certain place in order to forecast the quality of the air in India. The Indian Air Quality Indexis a commonly used metric to show the levels of pollutants (so2, no2, rspm, spm, etc.) across time. We createda gradient decent boosted multi-variable regression problem model to forecast the air quality index based on historical data from prior years and forecasting over a certain upcoming year. By using cost estimation for our predictive problem, we increase the model's efficiency. When given historical pollutant concentration data, our model will be able to accurately estimate the air quality index for a whole county, a state, or any bounded region. We were able to outperform the conventional regression models in our model by putting the suggested parameter-reducing formulations into practice. Our model can predict the air quality index for the entire country of India with 96% accuracy using the dataset that is currently available. We also use the AHP MCDM technique to determine the order of preference based on how close the ideal solution is.

"Aditya C R(et al.2018)" [3]: It used machine algorithms to forecast and identify the amount of PM2.5 concentration based on a dataset that included the atmospheric conditions in a certain city. Additionally, they forecasted the PM2.5 concentration on a specific day. Using the Logistic Regression technique, they first determine if the air is contaminated or not. Afterward, they use the Auto Regression algorithm to forecast the value of PM2.5 in the future based on historical data.

"Guangyuan Pan (Member, IEEE) in 2023" [4]: For the general public and society at large, air quality has traditionally been one of the most significant environmental concerns. Macro-level examination of future trends in air quality benefits from the use of machine learning techniques for Air Quality Index (AQI) prediction. It can be difficult to get a reliable prediction result when employing a single machine learning model to predict air quality under different AQI fluctuation trends. A genetic algorithm-based improved extreme learning machine (GA-KELM) prediction approach is improved to

successfully tackle this issue. To create the kernel matrix, which takes the role of the hidden layer's output matrix, a kernel approach is first presented. A genetic algorithm is then used to optimize the number of hidden nodes and layers of the kernel limit learning machine to address the problem with the conventional limit learning machine, which is that the number of hidden nodes and the random generation of thresholds and weights lead to the degradation of the network learning ability. The root mean square error, weights,

and thresholds are utilized to define the fitness function. Lastly, the model's output weights are calculated using the least squares approach. Genetic algorithms can iteratively optimize a model until they identify the best solution in the search space, hence improving the model's performance. Based on the basic data collected for a long-term air quality forecast at a monitoring point in a Chinese city, the optimized kernel extreme learning machine is applied to predict air quality (SO2, NO2, PM10, CO, O3, PM2.5 concentration and AQI) with comparative experiments based on SVM (Support Vector Machines), DBN-BP (Deep Belief Networks with Back- Propagation), and CMAQ (Community Multiscale Air Quality). This is done to verify the predictive ability of GA- KELM. The findings indicate that the suggested model learns more quickly and produces predictions with higher accuracy.

"HeidarMaleki (et al.2019)"[5]: For the stations Naderi, Havashenasi, MohiteZist, and Behdasht in Ahvaz, Iran -the most polluted city in the world predicted the hourly concentration values for the ambient air pollutants NO2, SO2, PM10, PM2.5, CO, and O3. Additionally, they have estimated and computed the Air Quality Health Index (AQHI) and Air Quality Index (AQI) for the four Ahvaz air quality monitoring sites that were previously mentioned. They predicted the hourly concentration of air pollutants and the two air quality indices, AQI and AQHI, between August 2009 and August 2010 using an Artificial Neural Network (ANN) machine learning method. Time, date, temperature, air pollution concentration, and meteorological parameters are among the inputs used by ANN algorithms.

EXISTING SYSTEM

The merit of the air AQI is predicted with the help of methods in the current system from basic statistical models to more complex machine learning algorithms to assess contaminants such as sulfur dioxide(so2) nitrogen dioxide(no2) particulates(pm 25) and (pm10) and ozone(O3). Traditional air merit testing systems frequently rely on centralized monitoring stations outfitted with specialized sensors usually these stations send data to central databases so that researchers and ecological organizations can examine it the current system makes frequent use of neural networks,(svm) support vector machines, and linear regression although these approaches can yield plausible forecasts in some scenarios they frequently encounter constraints when addressing the intricacy and fluctuations of air quality. It is feasible that the current system relies on batch-wise processing of past data rather than real-time tracking abilities the capacity of the system to react quickly to abrupt changes in the merit of air or ecological circumstances is hampered by this restriction incapacity to instantly adjust to changing environmental conditions. Conventional patterns might not be able to react swiftly to abrupt changes in the merit of air like contamination spikes or meteorological fluctuations. While some AQI prediction may be possible with the current system its accuracy flexibility and real-time abilities are frequently lacking underscoring the need for sophisticated and integrated methods.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

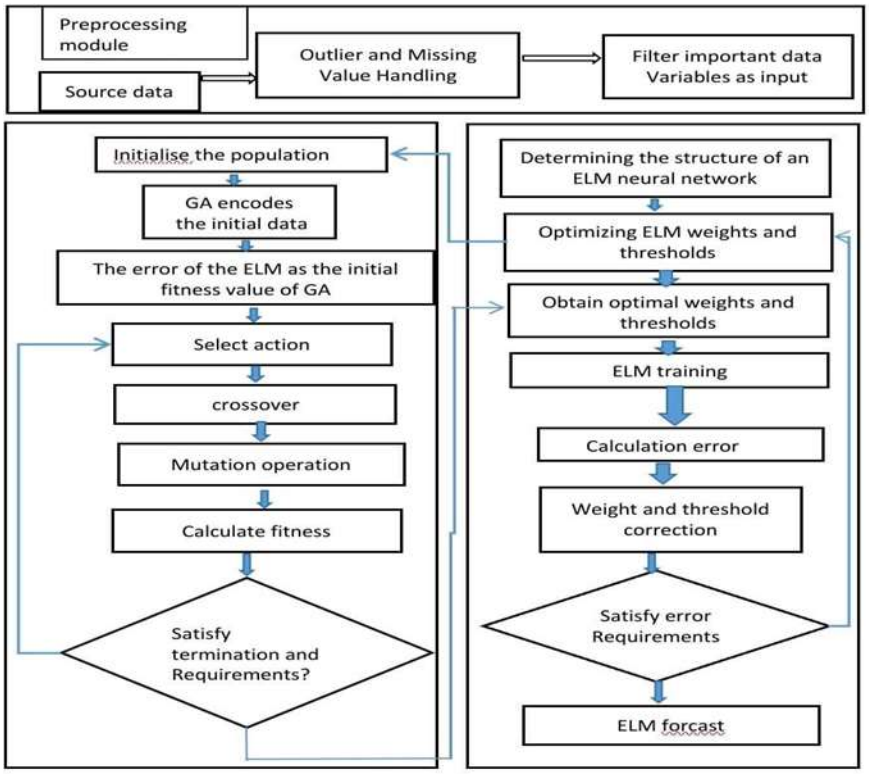

FIG1:EXISTING SYSTEM ARCHITECTURE

DISADVANTAGES

It is feasible that the current system relies on batch-wise processing of past data rather than real-time monitoring abilities. The capacity of the system to react quickly to abrupt changes in air quality or environmental circumstances is hampered by this restriction incapacity to instantly adjust to changing environmental conditions. Conventional models might not be able to react swiftly to abrupt changes in air quality like pollution spikes or meteorological fluctuations. While some AQI prediction may be possible with the current system its accuracy flexibility and real- time abilities are frequently lacking underscoring the need for sophisticated and integrated methods

PROPOSED SCHEME

The contemplated project integrates sensors which are components of IOT and are used to monitor the real-time quality of air and the collection of data undergoes early compilation to handle irrelevant values and extract relevant features. ML algorithms are qualified with the help of previous information for the purpose of forecasting air quality index system accomplishment is evaluated using metrics like accuracy. The iterative optimization process of genetic algorithms gradually improves the accomplishment of the model by finding optimal solutions this method's applications extend for the protection of public health monitoring, environment managing health care urban planning, and policy-making by integrating modern ML methods with IOT techniques. The projected method provides a comprehensive method to calculate the AQI and analyze the forthcoming quality of air phases benefiting society and promoting environmental sustainability.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

ADVANTAGES

IoT devices that deliver continuously obtained data can continuously gather sources from multiple sources, offering real- time updates on the state of the air quality data accuracy by taking into consideration a variety of variables including pollution weather, and geographic features. Machine learning algorithms can evaluate vast volumes of data gathered from Internet of things sensors forecasts from the air quality index are therefore more exact in allocating funds for measures to control pollution accurate projections of the air quality index.

AQI makes it easier to allocate resources for contamination control programs in best likely way which results in a more efficient and cost-effective use of resources. Health impact assessment machine learning algorithms can evaluate the possible health effects of air pollution on certain populations by combining AQI estimations with health data this allows for focused activities and health recommendations planning for environmental policy machine learning-based long-term analysis of air quality data can help design effective environmental laws and regulations that lower pollution levels and safeguard public health public awareness by combining machine learning-based insights with the internet of things- enabled air quality monitoring systems.

It may be possible to increase public awareness of the significance of air quality and to promote behavioral changes that will reduce pollution emissions. Classification of AQI categories categorizing AQI into different levels good moderate poor etc provides actionable insights for stakeholders and the general public to understand and answer to varying air quality conditions model diversity employing a range of machine learning models like linear regression decision tree etc ensures thorough evaluation of different methods for AQI prediction enhancing the robustness and authenticity of the project. Evaluation metrics The abstract mentions the evaluation of models using metrics like accuracy, precision, recall, and f1 score indicating a rigorous assessment process to validate the effectiveness of the predictive models.

System Block Diagram

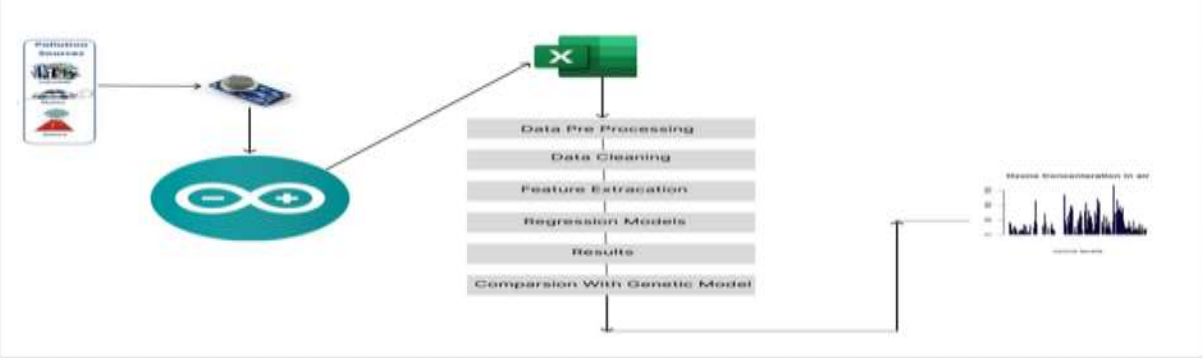

Figure 1: Proposed System Architecture

OUTPUT SCREENS

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

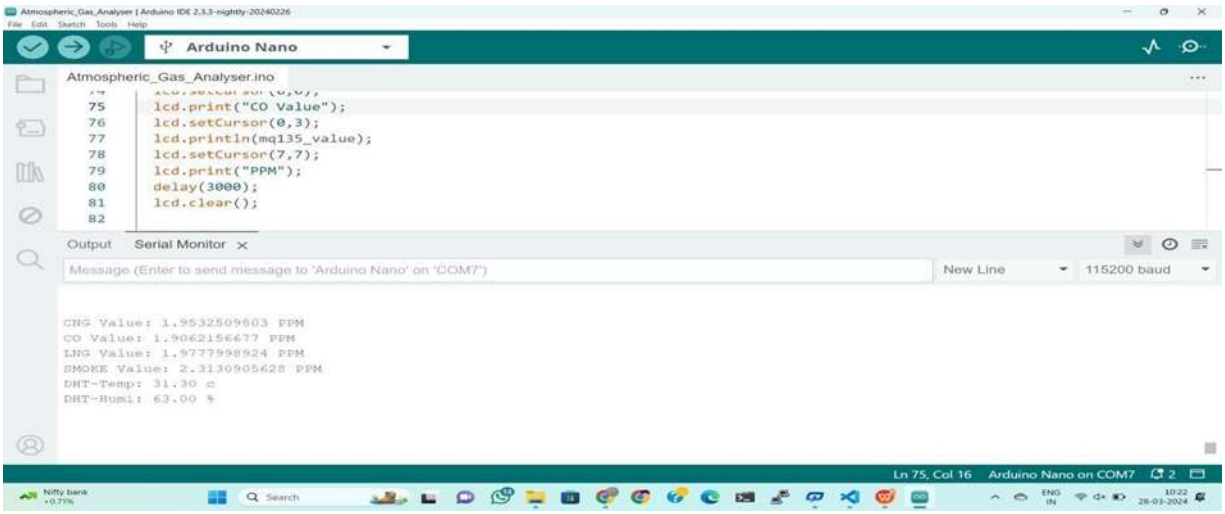

Figure 3: Sensor detected values of various gases

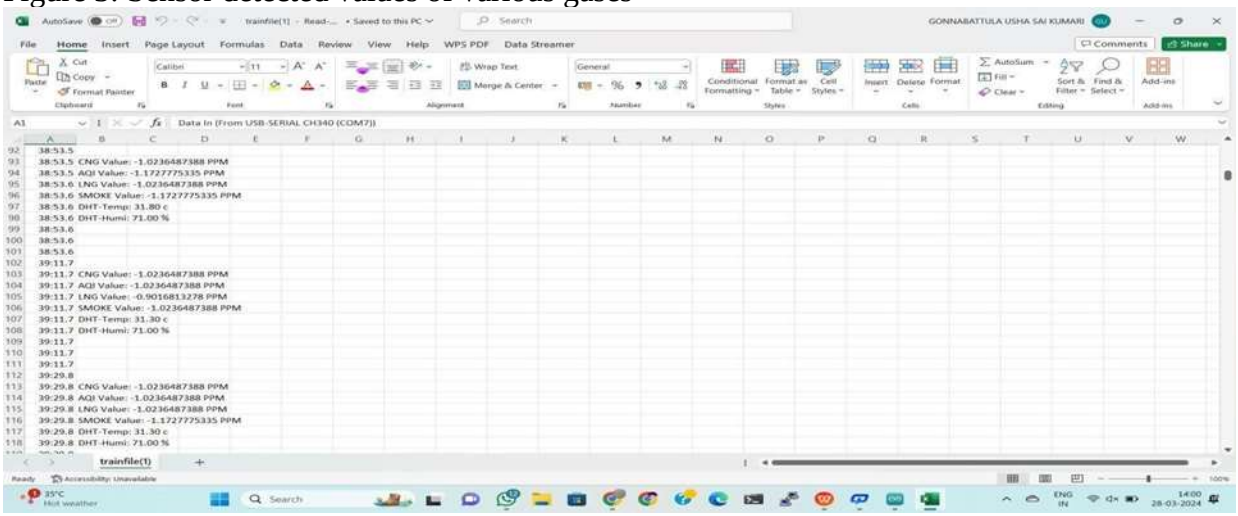
Figure 4: Values extraction by MsExcel

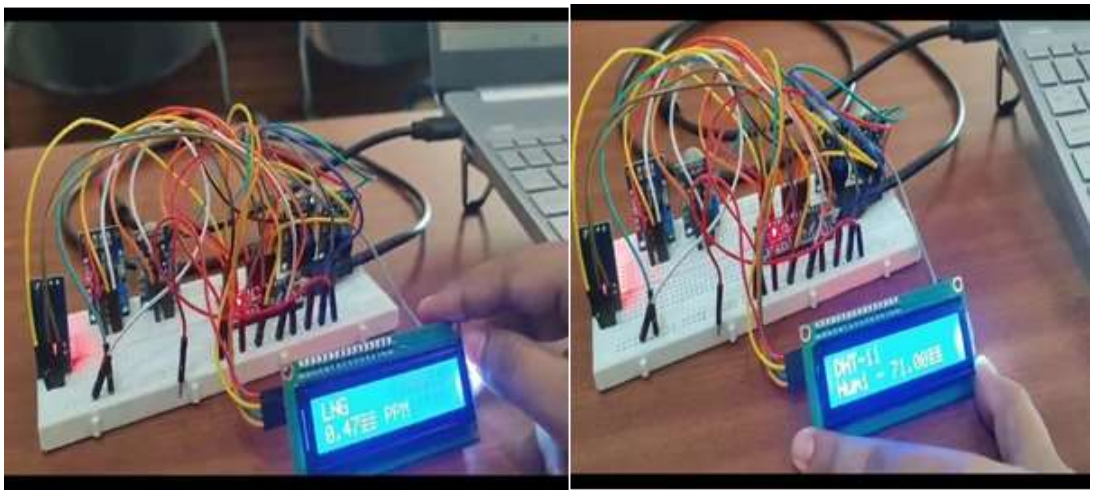
Figure 5: Values detected by various sensors and displayed on lcd display(16*2)

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

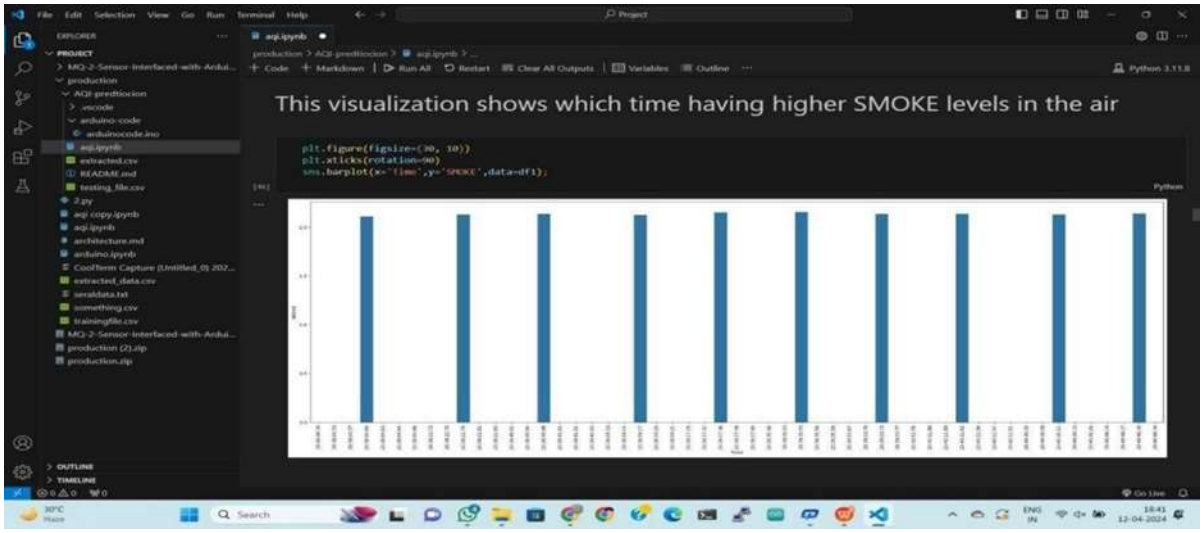

Figure 6: Graph Showing Higher Levels of Smoke In Air

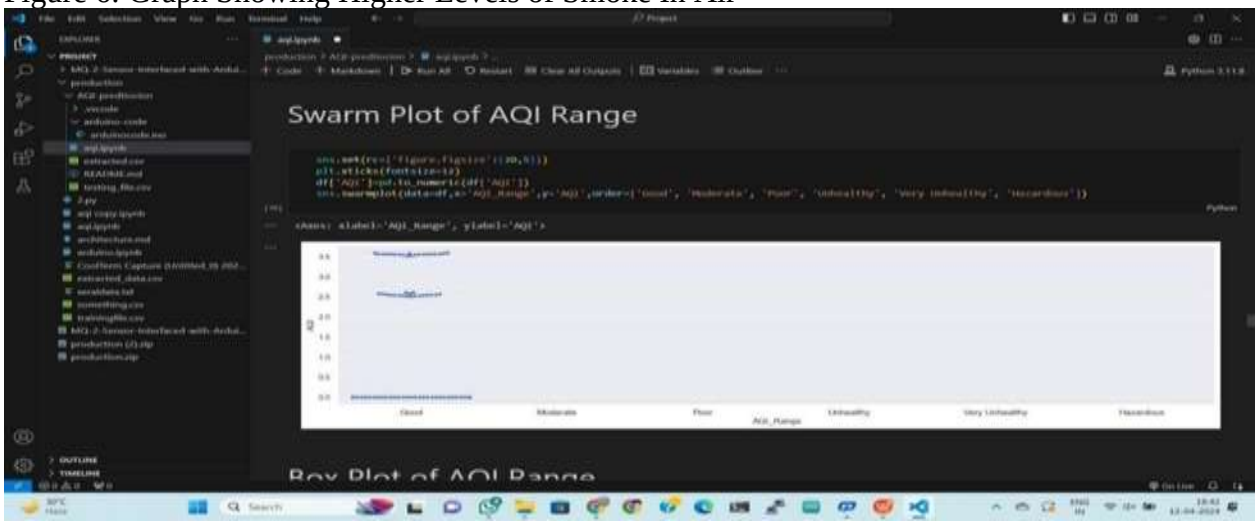

Figure 7: Swarm Plot of AQI Range

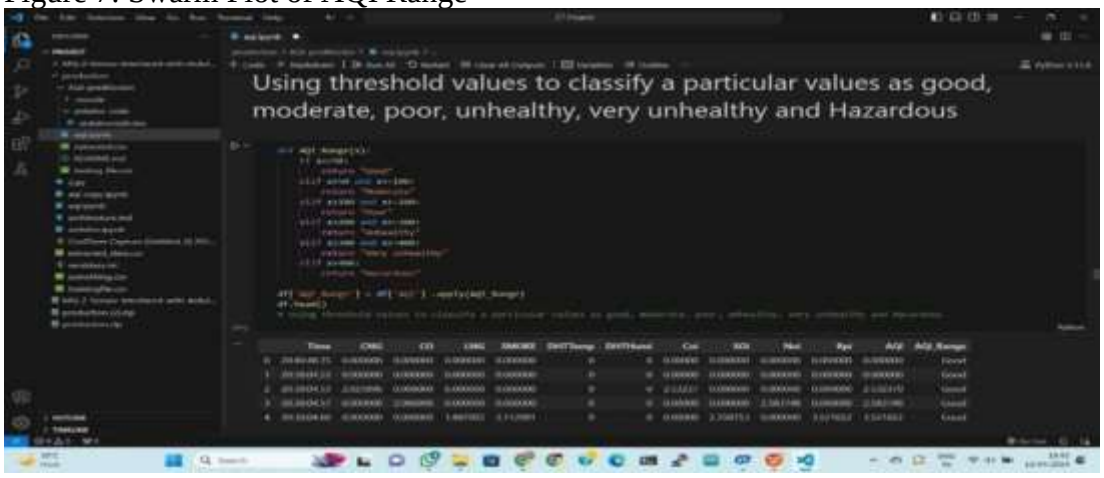

Figure 8: AQI Range Based on Different Gases Values at Different Time Intervals

CONCLUSION

Metropolitan contamination of the air is a significant reason that impacts the prosperity of individuals also the climate fostering a coordinated warning and checking framework is an imperative move toward resolving this issue this framework can assist with the dynamic interaction and give early warning of contamination occasions it

can likewise use information investigation air and sensor innovations to give continuous reports on air quality.

It has been observed that it is feasible to gather process and display data on the quality of air using inexpensive sensors and open-source platforms by putting our idea into action we can now estimate the quality of air ranges with a respectable degree of precise thanks for a combination of ml algorithms opening the door for preventative works for lower impurity risks and also increase the quality of air standards.

Although the current regime offers useful instruments for tracking and predicting air quality there is still room for improvement and the further development we can increase the efficacy and usefulness of the system in tackling the intricate obstacles connected with city pollution of air by adding extra sensors improving data analytics methods also involving investors in citizen science experiments.

Our design acts as the alert of how vital to advance observation of the surroundings and robust programs for the public for cross-disciplinary collaboration technological invention and data-driven methods. For the sense behind the current and upcoming lifetimes, all can able to work toward cleaner healthier, and more renewable city settings by investing in research growth and accomplishment of quality of air observing systems.

REFERENCES

[1] Research on air quality prediction method in Hangzhou based on machine learning To cite this article: Zhongjie Fu et al 2021 J. Phys.: Conf. Ser. 2010 012011 Zhongjie Fu1*, Haiping Lin1, Bingqiang Huang2, and Jiana Yao2 1 College of Information Engineering, Hangzhou Vocational & Technical College, Hangzhou,Zhejiang, 310018, China 2 Department of Robotics Engineering, Zhejiang University of Science and Technology, Hangzhou, Zhejiang, 310023, China.

[2] International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue) © Research India Publications. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 Research India Publications.

[3] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu, “Detection and Prediction of Air Pollution using Machine Learning Models”, International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018.

[4] Air Quality Index Forecasting via Genetic Algorithm-Based Improved Extreme Learning Machine CHUNHAO LIU, (Graduate Student Member, IEEE), GUANGYUAN PAN, (Member, IEEE), DONGMING SONG, AND HAO WEI School of Automation and Electrical Engineering, Linyi University, Linyi 276000, China VOLUME 11, 2023.

[5] Heidar Malek, Armin Sorooshian, Gholamreza Goudarzi, Zeynab Baboli, Yaser Tahmasebi Birgani, Mojtaba Rahmati, “Air pollution prediction by using an artifcial neural network model”, Clean Technologies and Environmental Policy, (2019) 21:1341–1352.