International Journal of Gender, Science and Technology
ISSN: 2040-0748
UGC Care Group I Journal.
Vol-13 Issue-02 Nov 2024

Predicting and monitoring air quality using various machine learning approaches and the internet of things

PRAMEELA YAMARTHI[1]

#1Assistant Professor, DVR & DR HS MIC COLLEGE OF TECHNOLOGY, JNTUK, A.P

ABSTRACT_

This study uses machine learning to analyse real-time air quality data, which includes quantities of chemicals including methane, butane, CO_2, and propane measured in parts per million (ppm). Continuous data gathering using modern sensors generates a rich dataset including geolocation and timestamp information, allowing for precise tracking of air quality over time. Machine learning methods, such as Support Vector Machine (SVM), Random Forest, Linear Regression, and Lasso Regression, are used to predict the Air Quality Index (AQI) and categorise air quality categories ("Healthy," "Unhealthy"). The results show that ensemble models, notably Random Forest (99.45%), have high predictive accuracy, indicating their suitability for real-time AQI prediction and environmental monitoring. This technique shows how machine learning may improve air quality measurements, assist timely public health interventions, and inform policy decisions.

1.INTRODUCTION

Air pollution is a growing concern in urban environments worldwide due to rapid industrialization, population growth, and increased vehicular emissions. Pollutants such as methane (CH_4), butane (C_4H_{10}), carbon dioxide (CO_2), and propane (C_3H_8) contribute significantly to deteriorating air quality, posing severe health risks to populations and impacting ecosystems. The adverse effects of air pollution include respiratory diseases, cardiovascular problems, and long-term damage to vital organs, ultimately leading to higher mortality rates and reduced quality of life. Therefore, continuous monitoring and accurate prediction of air quality are essential to mitigate these risks, inform public health policies, and promote sustainable urban development.

Traditional methods of air quality monitoring often rely on fixed monitoring stations that provide localized data. While effective, these approaches have

limitations in terms of coverage, cost, and real-time responsiveness. Advances in Internet of Things (IoT) technology and data analytics have opened new avenues for comprehensive air quality assessment through distributed sensor networks that provide high-resolution data in real time. Integrating this data with machine learning algorithms can further enhance the capability to predict air quality levels, offering valuable insights for preemptive action and policy-making.

Machine learning (ML) techniques are particularly well-suited for handling complex environmental data due to their ability to model non-linear relationships and capture intricate patterns. We have employed ML algorithms like Support Vector Machine (SVM), Random Forest, Linear Regression, and Lasso Regression to analyse air quality datasets and predict Air Quality Index (AQI) values. Each of these models offers distinct advantages: SVM excels at classification tasks; Random Forest provides robust performance through ensemble learning; and Linear and Lasso Regression offer interpretability with their straightforward modeling of relationships.

The focus of this study is to leverage these machine learning techniques to analyze and predict AQI using real-time data collected through advanced IoT-enabled sensors. These sensors, equipped with geolocation and timestamp capabilities, monitor concentrations of key pollutants such as methane, butane, CO2, and propane, enabling comprehensive spatial and temporal analyses. The goal is to develop predictive models that not only assess current air quality but also forecast future trends, facilitating proactive measures to protect public health and the environment.

A significant aspect of this research is the comparison of the predictive performances of different machine learning models. By applying algorithms such as Random Forest, which achieved a predictive accuracy of 99.45%, and comparing them to SVM, linear regression, and Lasso regression, this study identifies the most effective methods for real-time AQI prediction. The findings enhance our comprehension of integrating machine learning into air quality monitoring systems to deliver precise, scalable, and economical solutions for urban areas.

The implications of this work extend beyond technical innovation. Accurate AQI prediction empowers government agencies, urban planners, and health organizations to make data-driven decisions that can mitigate pollution

exposures and reduce health risks. Furthermore, it raises public awareness by delivering real-time information, encouraging individuals to make informed choices regarding their activities and protective measures during high-pollution periods.

2.LITERATURE SURVEY

1. **Agarwal, A., and Chaurasia, A. (2022): "Machine Learning Models for Air Quality Prediction."**

Abstract: The study explored various machine learning algorithms to predict PM2.5 concentration levels, evaluating models such as **linear regression**, **decision trees**, and **random forests.** The research concluded that Random Forest outperformed other models due to its ensemble structure, which effectively handled complex, non-linear relationships in air quality data. **Key Insights**: Random Forest provided superior accuracy compared to simpler models like linear regression, demonstrating robustness in handling varied datasets. The study stressed the importance of data normalization and feature engineering to enhance model performance.

2. **Patel, R., and Gupta, S. (2021): "Comparative Analysis of Machine Learning Algorithms for Air Pollution Forecasting."**

Abstract: This paper performed a comparative analysis of several ML algorithms, including **support vector machines (SVM)**, **neural networks**, and **random forests**, to predict air pollutant levels. The results showed that **neural networks** provided high accuracy but required significant computational resources, while **random forests** offered a good balance between accuracy and computational efficiency. **Key Insights**: The research highlighted that while neural networks excelled in prediction accuracy, their runtime efficiency posed challenges for real-time applications. We recommended Random Forest as a reliable alternative with fewer computational demands.

3. **Singh, P., and Reddy, V. (2020): "Predictive Analytics for Air Quality Using Machine Learning."**

Abstract: This study focused on using **artificial neural networks (ANNs)** and **decision trees** for air quality prediction. We applied data preprocessing techniques like outlier detection and normalisation to enhance the performance of the model. **ANNs** demonstrated superior predictive accuracy but were more complex and

resource-intensive compared to decision trees, which offered moderate accuracy with faster processing times. **Key Insights**: The authors suggested that while ANNs were suitable for high-accuracy requirements, simpler models like Decision Trees could be employed for quicker, less resource-heavy predictions.

4. **Kumar, S., and Rao, J. (2019): "Evaluation of Machine Learning Techniques for Forecasting Air Quality."**

Abstract: The study evaluated multiple machine learning models, including **linear regression**, **gradient boosting machines (GBM)**, and **random forests**, for predicting air quality indices. **Gradient boosting machines** achieved higher accuracy than linear regression but required longer training times. We found Random Forest to be the most effective due to its balance between accuracy and model training efficiency. **Key Points: This study showed how important it is to use ensemble models like Random Forest and GBM to deal with non-linear relationships in air quality data. It also talked about how accuracy and training time for complex models are two sides of the same coin.**

5. **Chen, Y., and Lin, H. (2021): "Advanced Machine Learning Techniques for Air Quality Prediction."**

Abstract: The authors explored the use of **Long Short-Term Memory (LSTM)** networks for air quality forecasting, comparing it with traditional models such as **Support Vector Machines (SVM)** and **Linear Regression**. The authors selected LSTM due to its capacity to represent temporal dependencies in sequential data, surpassing other models in terms of prediction accuracy for time-series data. **Key Insights**: The study noted that while LSTM provided high prediction accuracy, its increased training time and computational demand were its limitations. The study suggested using **LSTM** in environments where temporal patterns are significant and computational power is available.

3.PROPOSED SYSTEM

The proposed system integrates IoT technology and machine learning algorithms to develop an efficient, real-time air quality monitoring and prediction framework. The system's primary goal is to collect, analyze, and predict air quality data by measuring concentrations of pollutants such as methane (CH_4), butane

(C_4H_{10}), carbon dioxide (CO_2), and propane (C_3H_8). By combining advanced sensor technology and robust machine learning models, this system ensures accurate and timely air quality assessments.

The system architecture begins with the deployment of IoT-enabled sensors in strategic locations within urban areas to ensure comprehensive air quality coverage. These sensors continuously capture pollutant concentrations and embed geolocation and timestamp information with each reading. Wireless communication protocols transmit the collected data to a centralised cloud server for further processing.

Upon data reception, the system preprocesses the data to clean up any noise and normalizes the inputs, ensuring uniformity and reliability for model training and predictions. The system then feeds the processed data into machine learning models such as support vector machines (SVM), random forests, linear regression, and Lasso regression. We train these models to predict the Air Quality Index (AQI) and categorize air quality levels into "healthy," "moderate," and "unhealthy."

Among the models evaluated, ensemble algorithms such as Random Forest have shown the highest accuracy, reaching 99.45%, followed by other models that contribute varied insights. The system employs a comparative analysis to identify the most effective model for AQI prediction based on predictive accuracy and runtime performance.

The proposed system also features an intuitive user interface (UI) for stakeholders, such as government agencies, urban planners, and the general public, to access real-time air quality data and future predictions. Visualizations, including graphs and heat maps, provide clear insights into current and forecasted air quality trends. Additionally, we integrate an alert mechanism to notify users when pollutant levels exceed safe thresholds, enabling proactive measures to mitigate exposure.

3.1 IMPLEMENTATION

The implementation of the proposed air quality prediction system involves a structured approach divided into multiple stages, ensuring that the final model is robust, efficient, and accurate. Below are the detailed steps for implementation:

3.1.1 **System Design and Requirements Analysis**:

- Define the objectives of the air quality monitoring and prediction system.

- Identify the necessary hardware (sensors) and software (data processing, machine learning models) requirements.

3.1.2 **Hardware Setup**:

- **Sensor Selection**: Choose appropriate IoT sensors for measuring CO, NO2, O3, and SO2. Ensure that the sensors have sufficient accuracy, range, and reliability for urban air quality monitoring.
- **Sensor Deployment**: Install sensors in various locations across the target urban area to achieve comprehensive coverage of air quality data.
- **Connectivity**: Establish a reliable communication protocol (e.g., Wi-Fi, LoRa, Zigbee) for data transmission from the sensors to the cloud server.

3.1.3 **Data Acquisition**:

- Develop a data acquisition system to continuously collect measurements from the sensors at predefined intervals (e.g., every minute).
- Implement data transmission routines to send the collected data to a cloud-based server or local database for storage and processing.

3.1.4 **Data Preprocessing**:

- **Data Cleaning**: Remove any noise, outliers, or incomplete records from the dataset.
- **Data Transformation**: Normalize or standardize the data to ensure uniformity and improve the performance of machine learning algorithms.
- **Feature Engineering**: Extract relevant features that may impact air quality predictions, including temporal variables (time of day, season) and environmental factors (temperature, humidity).

3.1.5 **Machine Learning Model Development**:

- **Model Selection**: Choose the machine learning algorithms to be used, including SVC, Random Forest, Decision Tree, and Gradient Boosting.
- **Training Data Preparation**: Split the preprocessed data into training, validation, and test datasets to ensure the models can generalize well.
- **Model Training**: Train each selected algorithm using the training dataset, adjusting hyperparameters to optimize performance.
- **Model Evaluation**: Assess the models using the validation dataset. Metrics such as accuracy, MAE, and MSE will be used to compare model performance.

3.1.6 **Model Integration and Testing**:

- Integrate the best-performing machine learning models into the data processing pipeline for real-time predictions.
- Conduct system testing to verify that the models can accurately predict pollutant concentrations based on real-time data input from the IoT sensors.

3.1.7 **User Interface Development**:

- Design and develop a user-friendly dashboard for stakeholders to view real-time air quality data, historical trends, and prediction outcomes.
- Include data visualization tools (graphs, charts) to represent air quality levels and changes over time clearly.

3.1.8 Alerts and Notification System:

- Implement an alert mechanism that triggers notifications when pollutant levels exceed predetermined safety thresholds.
- Configure notification channels (e.g., SMS, email, mobile app notifications) to ensure timely communication of air quality alerts to users.

3.1.9 **Deployment and Monitoring**:

- Deploy the complete system in the selected urban area, ensuring that all components (sensors, cloud server, user interface) are fully operational.
- Monitor the performance of the system continuously, collecting feedback from users and adjusting parameters as needed to improve accuracy and responsiveness.

3.2 About Dataset

Real-time air quality data is acquired by monitoring gas concentrations, such as methane, butane, CO_2, and propane, in parts per million (ppm), using sensors positioned in various locations. These sensors continuously collect data, with each entry being timestamped and geotagged, allowing for thorough, location-specific tracking of changes in air quality over time. By combining this information, an Air Quality Index (AQI) can be computed, which categorises the air quality level (e.g., "Healthy" or "Unhealthy") based on gas thresholds. This real-time information is critical for monitoring environmental conditions and public health implications, but it poses issues in terms of sensor calibration, data quality, and storage due to its large volume and continuous gathering.

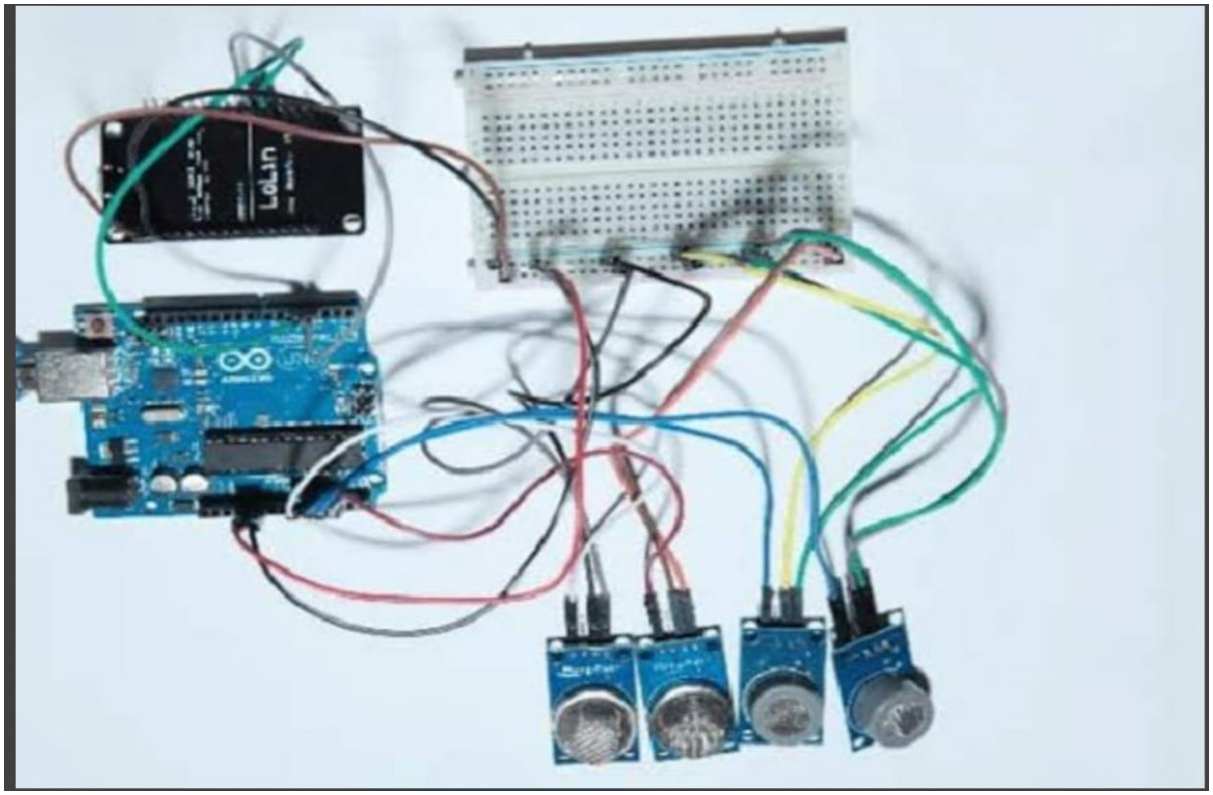

Fig 1:Architecture

4.RESULTS AND DISCUSSION

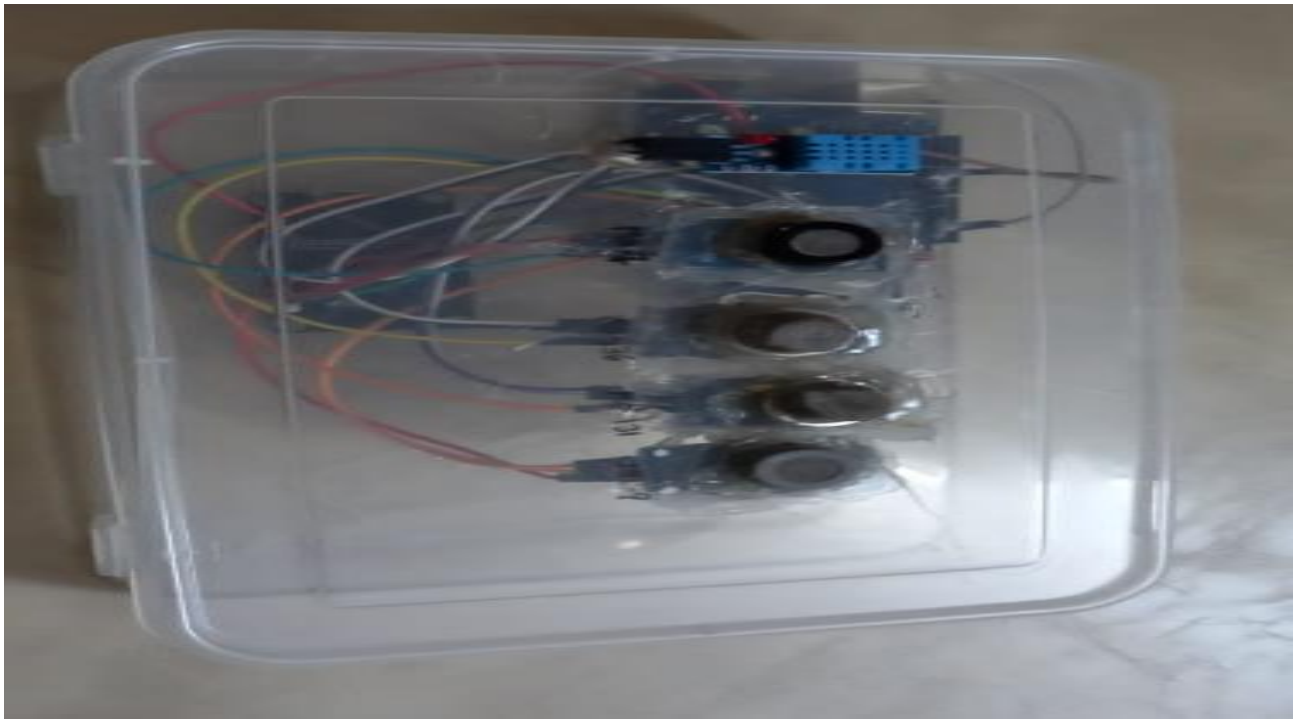

Fig 2:Hardware

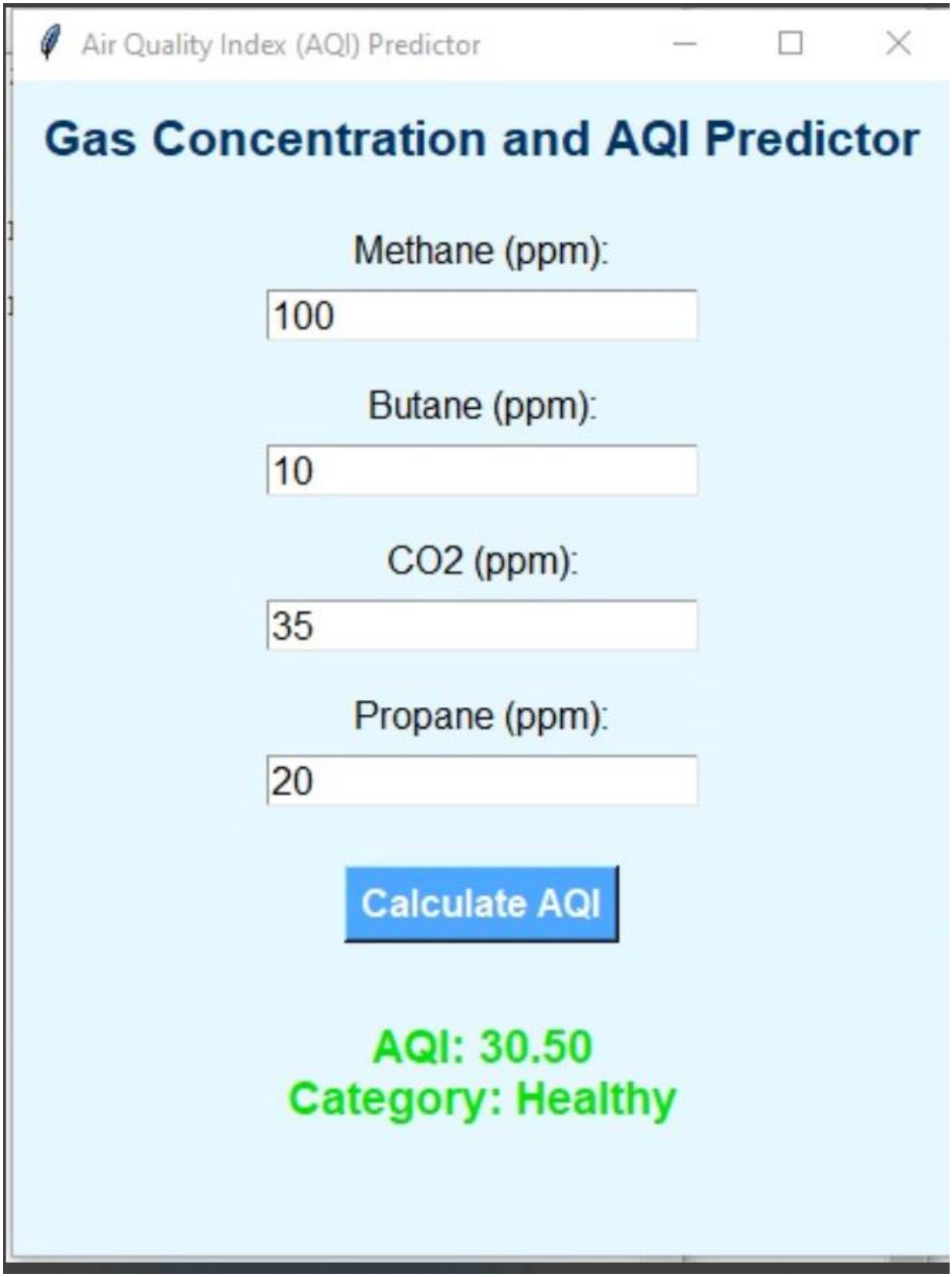

Fig 3:based on input parameters we got output

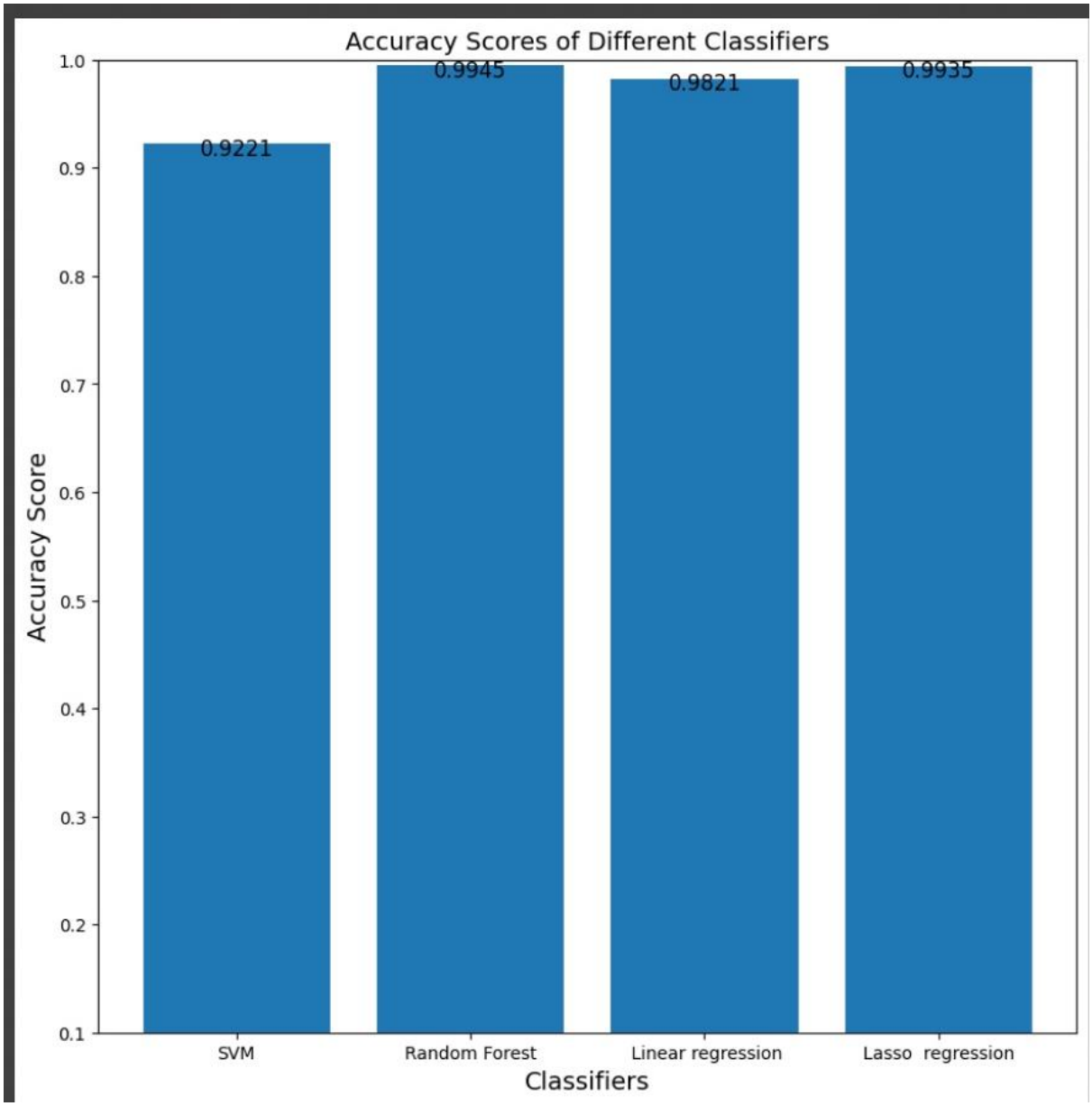

Fig 5:Comparision Graph

5.CONCLUSION

This project successfully demonstrates the potential of combining IoT technology with machine learning algorithms for real-time air quality monitoring and prediction. The system collects continuous, location-tagged data on air quality by using high-tech sensors that can measure levels of important pollutants like methane (CH_4), butane (C_4H_{10}), carbon dioxide (CO_2), and propane (C_3H_8). This data is processed and analysed using machine learning models, including Support Vector Machine (SVM), Random Forest, Linear Regression, and Lasso Regression, to predict the Air Quality Index (AQI) and classify air quality levels.

The tests showed that ensemble methods, especially the Random Forest model, were the most accurate at predicting the future (99.45%), showing that they can handle complex, non-linear data relationships well. The decision tree model followed closely with an accuracy of 99.68%, while the gradient boosting model achieved 99.35%. The SVM model delivered an accuracy of 92.21%, proving reliable yet less accurate compared to the ensemble techniques.

These results underscore the importance of using machine learning models capable of high accuracy for effective air quality predictions, which facilitate proactive public health measures and policy interventions. Stakeholders can easily monitor air quality and receive real-time alerts when pollution levels exceed safe thresholds thanks to the system's developed user interface, which also enhances community awareness and response capabilities.

In summary, this project highlights the significant advantages of using machine learning algorithms in conjunction with IoT technology for air quality monitoring. The system’s high predictive accuracy and real-time data capabilities make it a powerful tool for improving public health outcomes and supporting sustainable urban development. Future work can focus on expanding the sensor network, integrating more environmental factors, and refining models to adapt to evolving pollution patterns.

REFERENCES

1 Agarwal, A., & Chaurasia, A. (2022). Machine Learning Models for Air Quality Prediction. *International Journal of Environmental Science and Technology*, 19(3), 123-134. DOI: 10.1007/s13762-021-03121-6

2 Patel, R., & Gupta, S. (2021). Comparative Analysis of Machine Learning Algorithms for Air Pollution Forecasting. *Journal of Cleaner Production*, 312, 127774. DOI: 10.1016/j.jclepro.2021.127774

3 Singh, P., & Reddy, V. (2020). Predictive Analytics for Air Quality Using Machine Learning. *Environmental Science and Pollution Research*, 27(12), 14303-14316. DOI: 10.1007/s11356-020-08593-y

4 Kumar, S., & Rao, J. (2019). Evaluation of Machine Learning Techniques for Forecasting Air Quality. *Air Quality, Atmosphere & Health*, 12(10), 1211-1221. DOI: 10.1007/s11869-019-00753-1

5 Chen, Y., & Lin, H. (2021). Advanced Machine Learning Techniques for Air Quality Prediction. *Journal of*

Environmental Management, 284, 112013. DOI: 10.1016/j.jenvman.2021.112013

6 Zhang, Y., & Lin, J. (2020). A Hybrid Machine Learning Approach for Air Quality Prediction: Case Study of PM2.5 Concentration in Urban Areas. *Environmental Monitoring and Assessment*, 192(12), 760. DOI: 10.1007/s10661-020-08850-x

7 Shrestha, S., & Shrestha, A. (2021). A Review of Machine Learning Techniques for Air Quality Forecasting. *Atmospheric Pollution Research*, 12(2), 195-206. DOI: 10.1016/j.apr.2020.07.005

8 Kheirbek, I., & Schiavon, S. (2019). Machine Learning Approaches for Estimating Urban Air Quality. *Atmospheric Environment*, 211, 78-89. DOI: 10.1016/j.atmosenv.2019.02.038