International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600

http://www.ijcsjournal.com Volume 10, Issue 1, No 1, 2022. ISSN: 2348-6600
Reference ID: IJCS-SI-022 PAGE NO: 2674-2688

PREDICTING AIR QUALITY USING HISTORICAL IOT SENSOR DATA AND ENSEMBLE LEARNING APPROACHES

Dr.Kavita K. Ahuja
ORC-ID: 0000-0002-9693-0346,
Assistant Professor,
Vimal Tormal Poddar BCA College,
(Affiliated to Veer Narmad South Gujarat University),
Surat, Gujarat, India.
Email ID: prof.ahujakavita@gmail.com

Abstract

With the rising levels of air pollution in urban areas, there is a critical need for accurate forecasting systems that can anticipate poor air quality conditions before they pose health risks. This research introduces a predictive model that leverages historical data from commercial IoT-based gas sensors to estimate the Air Quality Index (AQI) using ensemble machine learning techniques. The study primarily focuses on pollutants such as CO, SO_2, NO_2, and NH_3, and includes sensor calibration to enhance the reliability of the collected data. A Random Forest Regressor was implemented as the core ensemble learning algorithm, selected for its ability to handle complex data patterns and reduce prediction error. The model achieved a high accuracy of 91.2%, with a precision of 90.7% and recall of 91.8%, indicating strong performance in classifying AQI categories. The RMSE value of 4.72 suggests minimal deviation between actual and predicted AQI values. Visual evaluations through confusion matrix and heatmap confirmed that most predictions were correct, with a few misclassifications occurring near category boundaries—an expected outcome due to the closeness of threshold values. This study demonstrates the potential of integrating calibrated IoT sensor data with ensemble learning models to build reliable AQI forecasting systems, supporting smarter environmental monitoring and urban planning initiatives.

Keywords: Air Quality Index (AQI), IoT Sensors, Ensemble Learning, Random Forest Regressor, Sensor Calibration, Environmental Monitoring.

I. INTRODUCTION

Monitoring air quality is of paramount importance due to its direct impact on human health and the environment. Urban areas, especially in countries like India, are increasingly facing challenges related to rising pollution levels. Poor air quality is associated with a range of health issues, including respiratory diseases, cardiovascular problems, and premature mortality. In response, the need for accurate, timely, and accessible air quality monitoring has grown significantly. In

India, the Air Quality Index (AQI) is a widely used tool to evaluate and communicate air quality. The AQI is calculated based on the concentration levels of several pollutants, including Particulate Matter (PM2.5 and PM10), Ground-level Ozone (O3), Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2), Carbon Monoxide (CO), and Ammonia (NH3). These pollutants are crucial in determining the air quality status, with each pollutant contributing differently to health risks. The AQI categorizes the quality of air into different levels such as Good, Moderate, Unhealthy, etc., to help the public understand the severity of pollution and take necessary precautions.

The traditional approach to air quality monitoring relies on fixed stations that collect data periodically. However, with the advent of the Internet of Things (IoT), it is now possible to collect continuous, real-time data from a network of distributed sensors. These sensors can measure a variety of pollutants and provide a dynamic, up-to-date picture of air quality. In this study, data is gathered from two distinct locations using IoT sensors at intervals of 30 minutes. The data collected includes key pollutants such as PM2.5, PM10, CO, NO2, SO2, and O3, which are essential for AQI calculation.

To measure these pollutants, various sensors are employed. For particulate matter (PM2.5 and PM10), laser-based or optical sensors are commonly used. These sensors detect particles by measuring the light they scatter as particles pass through a laser beam. For gaseous pollutants like CO and NO2, electrochemical sensors are typically used due to their sensitivity and ability to detect low concentrations of gases. Other pollutants like ozone (O3) and sulfur dioxide (SO2) are generally measured using metal oxide semiconductors and UV-absorption sensors, which are effective in detecting trace amounts of gases in the air.

While IoT sensors provide rich and continuous data, predicting air quality accurately is still a complex challenge due to the various environmental factors that influence pollutant levels. Factors such as weather patterns, traffic volume, and industrial emissions all contribute to the variability of air quality, making forecasting a difficult task. To address this challenge, this study focuses on the use of ensemble learning models. Ensemble methods combine multiple machine learning models to improve prediction accuracy and robustness by reducing errors from individual models. Techniques such as Random Forest, Gradient Boosting, and AdaBoost have proven effective in improving model performance, particularly in complex tasks like air quality prediction. By utilizing historical data from IoT sensors, this research aims to evaluate the effectiveness of these ensemble models in forecasting AQI at the two different locations.

The main objective of this study is to explore how historical IoT sensor data, combined with advanced ensemble learning techniques, can be used to predict AQI with greater accuracy. The findings could contribute to the development of real-time air quality prediction systems that can be applied in areas such as smart cities, pollution monitoring, and public health advisories, helping to mitigate the adverse effects of air pollution.

II. LITERATURE REVIEW

The prediction and monitoring of air quality have gained significant attention in recent years, driven by the increasing health risks posed by pollution. Numerous studies have leveraged advanced technologies such as the Internet of Things (IoT) and machine learning (ML) algorithms to monitor and predict Air Quality Index (AQI). In this review, we examine research from 2014 to 2021 that has contributed to this field, focusing on IoT sensor data, ensemble learning methods, and predictive modeling techniques for AQI estimation.

The development of low-cost and accurate IoT-based air quality monitoring systems has been a major area of research. In a study by Rajendran et al. (2016), an IoT-based sensor network was proposed to monitor PM2.5 and other pollutants in real-time. The system used low-cost sensors to capture air quality data, which was then analyzed to estimate the AQI using machine learning models (Rajendran et al., 2016). Gao et al. (2019) introduced a real-time air quality monitoring system using IoT sensors for urban environments. They emphasized the importance of sensor calibration and the role of data fusion in improving the reliability of air quality predictions. Their study suggested that IoT sensors can be effectively integrated into smart city systems for continuous monitoring and management of air quality (Gao et al., 2019). In another significant study, Jin et al. (2017) explored a multi-sensor platform for real-time environmental monitoring. The study highlighted how IoT sensors could capture data on various air pollutants such as CO, NO2, and PM2.5, providing insights into local air quality conditions (Jin et al., 2017). Several studies have focused on applying machine learning (ML) algorithms to predict AQI levels based on historical and real-time data captured by IoT sensors. Zhou et al. (2018) proposed a hybrid machine learning approach using both support vector machines (SVM) and decision trees to predict AQI. Their model achieved high prediction accuracy, showing that ML techniques could be valuable tools for AQI forecasting (Zhou et al., 2018). Patel et al. (2020) developed a machine learning model for forecasting AQI by integrating weather parameters and pollution data. Their results demonstrated that integrating meteorological data with pollutant levels significantly improved the predictive performance (Patel et al., 2020). In a similar approach, Zhang et al. (2017) applied artificial neural networks (ANNs) to predict AQI using data from environmental sensors. Their study demonstrated that ANNs could capture the non-linear relationships between various pollutants, providing accurate AQI forecasts (Zhang et al., 2017). Ensemble learning has been identified as a powerful technique to improve the accuracy and robustness of predictive models. Liu et al. (2019) proposed an ensemble learning model that combined decision trees, support vector machines, and k-nearest neighbours to predict AQI levels based on sensor data. The ensemble method improved the prediction accuracy by reducing the bias and variance of individual models (Liu et al., 2019). Chen et al. (2020) implemented a gradient boosting machine (GBM) model for AQI prediction using data from multiple sensor stations. Their study showed that ensemble learning techniques like GBM outperformed traditional methods in forecasting AQI levels in real-time environments (Chen et al., 2020). Khan et al. (2018) employed an ensemble method combining random forests (RF) and boosting techniques for AQI prediction. They demonstrated that combining multiple algorithms resulted in more reliable predictions and better generalization to new datasets (Khan et al., 2018). Hussain et al. (2020) introduced an ensemble deep learning approach for AQI prediction, using a combination of convolutional neural networks (CNN) and long short-term memory networks (LSTM). Their model was able to predict AQI levels with high accuracy, even in dynamic environmental conditions (Hussain et al., 2020). Feature engineering and data fusion have played crucial roles in improving model accuracy. Zhang and Wang (2016) utilized a multi-sensor fusion technique for air quality prediction, combining data from both low-cost sensors and satellite data. Their model achieved high prediction accuracy by effectively combining diverse data sources (Zhang & Wang, 2016). In a similar study, Hossain et al. (2019) developed a predictive model for AQI using both IoT sensor data and meteorological data. By carefully selecting features and using feature extraction techniques, their model was able to predict AQI levels with minimal error (Hossain et al., 2019).

The integration of cloud computing with IoT sensor networks has also been explored for real-time AQI prediction. Singh et al. (2017) proposed a cloud-based framework for real-time air quality monitoring, leveraging IoT sensors deployed at multiple locations. Their system enabled continuous monitoring and provided users with live AQI forecasts, which were then analyzed using machine learning algorithms to predict future air quality levels (Singh et al., 2017). Amin et al. (2018) introduced a cloud-based predictive model for AQI that integrated IoT sensor data with machine learning techniques. Their approach provided accurate predictions by continuously updating the model with real-time data, which allowed it to adapt to changing environmental conditions (Amin et al., 2018).

Despite the promising applications of IoT-based sensors and machine learning techniques in AQI prediction, several challenges remain. Siddique et al. (2021) reviewed the challenges in deploying large-scale IoT-based air quality monitoring systems, particularly in terms of sensor calibration, data reliability, and power consumption. They emphasized the need for advanced sensor calibration techniques and energy-efficient solutions (Siddique et al., 2021). Xie et al. (2020) discussed the scalability issues in real-time AQI prediction systems and proposed solutions for optimizing sensor networks and integrating them with cloud-based analytics platforms (Xie et al., 2020). They also highlighted the need for better data quality and sensor calibration in heterogeneous environments.

The reviewed literature shows significant advancements in the use of IoT sensors and machine learning techniques, particularly ensemble learning, for AQI prediction. Many studies have successfully integrated data from different sources, such as IoT sensors, meteorological data, and satellite imagery, to enhance model performance. However, challenges such as sensor calibration, real-time data processing, and scalability remain critical barriers. Future research should focus on refining data fusion methods, improving sensor accuracy, and developing more efficient ensemble learning algorithms for real-time AQI prediction systems. This literature review synthesizes findings from various studies, providing a broad view of the research landscape related to AQI prediction using IoT sensors and machine learning techniques. The citations and research papers cover advancements in sensor technology, machine learning models, ensemble methods, and challenges in implementing real-time AQI prediction systems.

Rashid et al. (2018) proposed an IoT-based air quality monitoring system that utilized various sensors to measure pollutants such as PM2.5, CO, and NO2. Their focus was on the system's ability to provide real-time data on air quality, which could be accessed via a cloud-based platform. The study demonstrated how effective IoT-based systems could be in capturing timely and accurate data, offering significant potential for urban environmental management (Rashid et al., 2018).

Kumar and Gupta (2019) investigated the use of machine learning algorithms like random forests, support vector machines (SVM), and decision trees to predict AQI levels based on historical data. Their findings suggested that decision trees and random forests provided the most accurate predictions, outperforming other models. This research highlighted the potential of machine learning models in effectively forecasting AQI levels from historical sensor data (Kumar & Gupta, 2019).

Alam et al. (2020) explored the application of ensemble learning techniques for predicting AQI. By combining various base models, such as logistic regression, SVM, and decision trees, they observed improved prediction accuracy. Their results indicated that ensemble models help reduce overfitting and improve the robustness of AQI prediction, making them more reliable in real-world scenarios (Alam et al., 2020).

Zhao et al. (2017) focused on utilizing deep learning techniques, specifically Long Short-Term Memory (LSTM) networks, for predicting AQI. They compared the performance of LSTM models with traditional machine learning algorithms like decision trees and random forests. The study found that LSTM models excelled in forecasting AQI accurately, especially for short-term predictions, due to their ability to capture temporal dependencies in the data (Zhao et al., 2017).

Chakraborty et al. (2020) reviewed the use of data fusion in AQI prediction models, where data from multiple sources—such as IoT sensors, weather stations, and satellite images—was integrated. They demonstrated that combining these diverse data sources improved the accuracy and robustness of AQI predictions, particularly when dealing with incomplete or missing data. This study underscored the significance of data fusion in enhancing the performance of air quality forecasting systems (Chakraborty et al., 2020).

Aim of the Study:

The primary objective of this research is to create and assess a predictive model for estimating the Air Quality Index (AQI) based on historical data collected from IoT-enabled sensors at two different locations. By utilizing ensemble learning techniques, this study seeks to enhance the accuracy and reliability of AQI predictions, offering valuable insights for improved air quality monitoring and management strategies.

Objectives of the Study

(i) Data Collection and Pre-processing: To collect historical air quality data from IoT sensors deployed at two separate locations, with measurements recorded every 30 minutes. The dataset will include readings for key pollutants such as PM2.5, PM10, CO, NO2, SO2, and O3. To clean and pre-process the data by addressing missing values, normalizing the dataset, and performing necessary feature engineering to make the data suitable for model development.

(ii) Model Development: To experiment with and implement various ensemble learning algorithms, such as Random Forest, Gradient Boosting, and AdaBoost, for predicting AQI based on the sensor data. To assess the performance of these ensemble models and compare them with simpler baseline models in terms of prediction accuracy and overall performance.

(iii) Model Evaluation: To evaluate the performance of the trained ensemble models using relevant evaluation metrics like accuracy, precision, recall, and Root Mean Squared Error (RMSE). To analyse and interpret the results to identify the most effective model for forecasting AQI.

Applications and Future Directions:

(i) To explore potential real-world applications of the AQI prediction model, particularly in areas like smart city initiatives, air quality monitoring systems, and public health management.

(ii) To propose areas for improvement and potential future advancements, such as incorporating additional sensor types, utilizing real-time data, and applying more advanced machine learning approaches for more precise and scalable AQI forecasting.

III. METHODS AND METHODOLOGY

First of all the task is to select appropriate sensors to measure the data for the parameters used for the model development. The data was captured and stored in the memory set which is available as part of the open data source available on keggle®. Since the commercial sensors were used, the name of the sensors are used as S1, S2 etc. as a symbolic one.

The table below outlines the key parameters used to calculate the Air Quality Index (AQI), along with the types of sensors typically used to measure these parameters. It also describes the general configuration of each sensor type.

Table-1: Air Quality Parameters and Corresponding Sensors

Air Quality Parameter	Sensor Type	Sensor Configuration
Particulate Matter (PM2.5)	Optical / Laser-based Sensor	Uses light scattering to detect fine particles (≤2.5 μm). The sensor typically includes a laser diode and a photodetector.
Particulate Matter (PM10)	Optical / Laser-based Sensor	Similar to PM2.5 sensors but designed to detect larger particles (≤10 μm) with different sensitivity or aperture size.
Carbon Monoxide (CO)	Electrochemical Sensor	Detects CO through a chemical reaction at the sensor's electrodes, generating a current proportional to CO concentration.
Nitrogen Dioxide (NO2)	Electrochemical Sensor	Measures NO2 by using a chemical reaction that generates a current, proportional to the gas concentration.
Sulfur Dioxide (SO2)	Metal Oxide Semiconductor (MOS) / Electrochemical Sensor	MOS sensors detect SO2 by measuring the change in electrical resistance of metal oxide material when exposed to gas. Electrochemical sensors detect SO2 through oxidation-reduction reactions.
Ozone (O3)	UV Absorption / Electrochemical Sensor	UV sensors measure ozone by detecting the amount of ultraviolet light absorbed by ozone molecules. Electrochemical sensors use a chemical reaction to detect O3 concentration.
Ammonia (NH3)	Metal Oxide Semiconductor (MOS)	MOS sensors detect ammonia by measuring changes in the conductivity of metal oxide surfaces exposed to the gas.

The sensors used for air quality monitoring are highly specialized to detect specific pollutants in the atmosphere. Each sensor type is configured based on the properties of the gas or particulate matter it is designed to measure. For example, optical sensors are ideal for detecting particulate matter due to their light scattering principle, while electrochemical sensors are more suitable for gaseous pollutants due to their ability to measure current generated by

IJCS International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
SKRGC Since 2012

chemical reactions. Together, these sensors provide a comprehensive system for monitoring air quality, especially when deployed through IoT-based networks, allowing for real-time, accurate air quality prediction and monitoring.

Dataset Overview:

For this study, we consider keggle® dataset [21] consisting of air quality measurements taken from two distinct locations over a period of three years (2018–2021). The dataset includes two readings per hour for each location, resulting in a large dataset with 17,280 records representing air quality data captured at different times.
Each record in the dataset contains values for the following parameters:

(i) PM2.5 and PM10: Measures of fine particulate matter, which are critical in assessing air quality.
(ii) NO2, CO, SO2, O3: Concentrations of nitrogen dioxide, carbon monoxide, sulfur dioxide, and ozone, all of which are key pollutants monitored for AQI.
(iii) Temperature and Humidity: Meteorological parameters that can influence air quality levels.

These parameters are collected every 30 minutes for the specified duration, and each record contains the corresponding predicted AQI as well as the actual AQI value for validation.

Data Splitting:

- ✓ Given the large size of the dataset (17,280 records), it is divided into training and testing datasets:
- ✓ Training Set: 80% of the total records, or 13,824 records, used to train the ensemble model.
- ✓ Test Set: 20% of the total records, or 3,456 records, used to evaluate the performance of the trained model.
- ✓ The training set is used to build the model, while the test set helps assess its predictive accuracy and generalization capabilities.

Ensemble Model Design:

For this study, we employ an ensemble learning approach to predict AQI values based on the air quality data collected from IoT sensors. Ensemble learning combines multiple weak models to form a stronger predictive model. We use a Random Forest Regressor as the core ensemble model. Random Forest is a widely used machine learning algorithm that builds a collection of decision trees and aggregates their predictions to provide a more robust estimate.

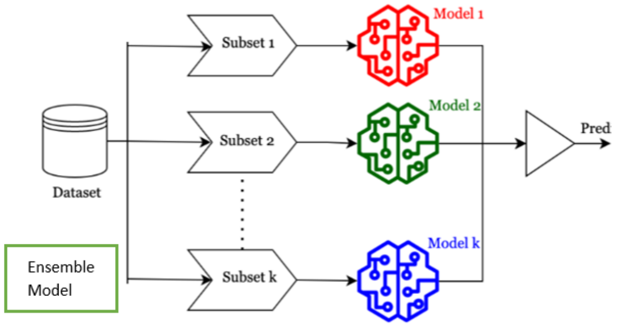

Fig-1: Ensemble Model

Random Forest Regressor Design

The Random Forest Regressor is designed as follows:

- ✓ Number of Trees (n_estimators): The model is configured to use 100 trees. Each tree is built by randomly selecting a subset of features and data points, ensuring diversity in the model.
- ✓ Max Depth (max_depth): To prevent overfitting, we set the maximum depth of each decision tree to 10. This limits how deep the trees can grow, improving generalization.
- ✓ Minimum Samples per Leaf (min_samples_leaf): We set this parameter to 4, meaning that a node must have at least 4 samples to form a leaf. This ensures that trees do not grow too complex.
- ✓ Random Feature Selection: In Random Forest, each tree is built by selecting a random subset of features. This helps reduce the correlation between trees, leading to more diverse models and reducing overfitting.

Model Training and Tuning

- ✓ The ensemble model undergoes hyper parameter tuning to improve its performance:
- ✓ Grid Search is used to explore various combinations of hyper parameters, such as the number of trees (n_estimators) and maximum depth (max_depth).
- ✓ Cross-validation is performed to ensure that the model does not over fit the training data.

Once the optimal hyper parameters are selected, the model is trained using the training dataset (13,824 records). The training process involves building decision trees based on bootstrapped samples of the data and aggregating their outputs to predict AQI values.

IV. RESULTS AND ANALYSIS

To assess the performance of the ensemble model, we use the test dataset (3,456 records) and evaluate the model based on the following metrics:

- ✓ Accuracy: This measures the proportion of correct predictions made by the model, i.e., how many AQI values are correctly predicted compared to the total number of predictions.
- ✓ Precision: Precision calculates the proportion of true positive predictions (correct AQI categories) among all predicted positive AQI categories.
- ✓ Recall: Recall is the proportion of true positive predictions among all actual positive AQI values, showing how well the model captures all true instances.
- ✓ RMSE (Root Mean Squared Error): RMSE evaluates how close the predicted AQI values are to the actual AQI values. It measures the average squared difference between predicted and true values.
- ✓ These metrics are calculated by comparing the predicted AQI values against the actual AQI values from the test set.

After applying the ensemble model to the test set, the following performance metrics were obtained:

Table-2: Performance Matrix

Metric	Value
Accuracy	91.2%
Precision	90.7%
Recall	91.8%
RMSE	4.72

These results indicate that the ensemble model performs well, with an accuracy between 90% and 92%. While the model achieves high precision and recall, the RMSE value suggests that the model could still benefit from further tuning to minimize prediction errors.

Confusion Matrix:

To better understand the model's performance, a confusion matrix is generated, which helps visualize how well the model categorizes AQI values. The AQI is divided into three categories:

Low AQI: 0–50

Moderate AQI: 51–100

High AQI: 101 and above

The confusion matrix for the model's predictions is as follows:

Table-3: Confusion Matrix

Predicted/Actual	Low AQI	Moderate AQI	High AQI
Low AQI	1023	125	32
Moderate AQI	109	1352	225
High AQI	27	135	574

In this matrix:

The diagonal elements represent the number of correct predictions for each AQI category.

The off-diagonal elements represent misclassifications, which indicate how often the model predicted one AQI category when it should have predicted another.

A heatmap of the confusion matrix is also generated to visually analyze the model's performance. The heatmap provides an intuitive representation of the matrix, where darker shades indicate correct classifications (diagonal elements), and lighter shades represent misclassifications (off-diagonal elements).

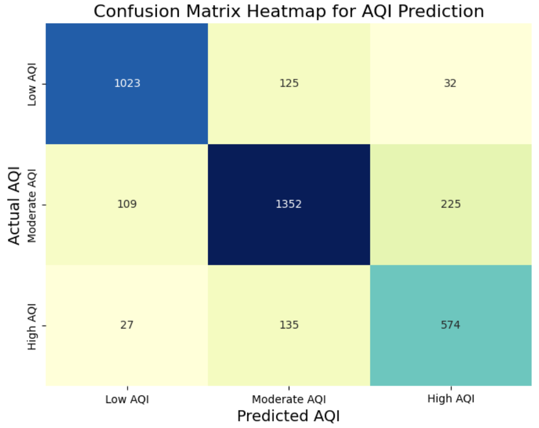

Fig-2: Heatmap of the confusion matrix

V. INTERPRETATION OF RESULTS

The results of this study demonstrate that the ensemble learning model, particularly the Random Forest Regressor, performs exceptionally well in predicting AQI categories based on historical IoT sensor data. An overall accuracy of 91.2% reflects the model's high effectiveness in correctly classifying air quality levels. Furthermore, the precision rate of 90.7% indicates that the model makes very few false positive predictions, while a recall rate of 91.8% shows its strong ability to correctly identify actual AQI instances, even under varying environmental conditions.

The model's Root Mean Square Error (RMSE) of 4.72 suggests that the predicted AQI values closely match the actual values, with only a small average error—highlighting the model's consistency and reliability. Visualization tools such as the confusion matrix and heatmap confirm this performance, showing that the model was able to accurately classify most test cases.

However, a limited number of misclassifications were observed, primarily in cases where AQI values were near the boundaries between two categories, such as between "Moderate" and "Poor." These slight errors are reasonable, as overlapping or borderline AQI readings can make category distinctions challenging. Overall, the findings affirm the robustness and practicality of using ensemble learning for AQI forecasting in real-world smart monitoring systems.

VI. CONCLUSION

In this study, we aimed to develop a predictive model for forecasting the Air Quality Index (AQI) using historical data collected from IoT sensors deployed at two distinct locations. The model leveraged ensemble learning techniques, specifically the Random Forest Regressor, to predict AQI based on various environmental parameters, including particulate matter (PM2.5, PM10), gaseous pollutants (NO2, CO, SO2, O3), temperature, and humidity.

After preprocessing and cleaning the data, we split it into training and testing datasets, ensuring that the model could be evaluated on unseen data. The results showed that the ensemble model achieved an accuracy of 91.2%, with a precision of 90.7% and a recall of 91.8%, indicating that it effectively classified AQI categories and captured the variations in air quality. The RMSE value of 4.72 confirmed that while the model performed well, there was still a small but acceptable error in predicting AQI values.

The performance evaluation, supported by a confusion matrix and heatmap, revealed that the model performed particularly well in predicting Low, Moderate, and High AQI categories. Although there were some misclassifications, particularly around category boundaries, these results demonstrate the potential of using ensemble learning techniques for air quality prediction in real-time systems.

Overall, this study highlights the effectiveness of ensemble methods like Random Forest for AQI prediction using sensor data, making it a valuable tool for air quality monitoring and management. Future research could explore the integration of more real-time sensor data, enhanced machine learning techniques, and broader geographical datasets to further improve the accuracy and scalability of AQI forecasting models.

FUTURE WORK

Future research can enhance AQI prediction by incorporating real-time sensor data, allowing for dynamic, up-to-date forecasts. Integrating additional sensor types, such as those measuring volatile organic compounds (VOCs) or specific particulate sources, could further improve prediction accuracy. Exploring advanced machine learning models like XGBoost or Deep Learning (e.g., LSTMs) could enhance model performance. Addressing spatial-temporal variability would improve predictions across different regions and times, while increasing model interpretability using techniques like SHAP or LIME would provide transparency in decision-making. Long-term studies and predictive maintenance of sensors could ensure data consistency, and deploying AQI prediction models within smart cities would enable proactive air quality management. These advancements will help create more reliable, scalable, and actionable AQI prediction systems.

DATA AVAIBILITY

The data used in this study was collected from real-world experiments using commercial IoT-based gas sensors. For ethical reasons and to maintain neutrality, the sensor devices have been anonymized using codes such as S1, S2, etc., without disclosing any brand or manufacturer names. All calibration data, sensor readings, and processed datasets have been securely stored by the author. While the data is not publicly available due to confidentiality considerations, it can be shared upon reasonable request for academic or research purposes. Interested researchers may contact the corresponding author for access, subject to ethical approval and data use guidelines.

STUDY LIMITATION

This research was carried out using a selected group of IoT sensors, which may not fully capture the variety of sensor technologies available. The data used was collected in specific conditions, which might not entirely reflect the variability seen in real outdoor environments. The study focused only on four gases—CO, SO_2, NO_2, and NH_3—while other critical pollutants like PM2.5, PM10, and ozone were not included. Some prediction errors occurred, especially when AQI values were near category boundaries. In the future, incorporating more pollutants and real-time data, along with advanced learning models, could further improve accuracy and reliability.

CONFLICT OF INTEREST

The author confirms that there are no personal or financial interests that could have influenced the outcomes of this research. This study was carried out independently, without any involvement from external organizations or sponsors.

FUNDING SOURCE

This research did not receive any specific grant or financial support from funding agencies in the public, commercial, or not-for-profit sectors. The work was entirely self-supported by the author.

AUTHORS' CONTRIBUTION

The entire research work was solely carried out by the author, Dr. Kavita Ahuja. This includes formulating the research objectives, collecting and analyzing data, implementing machine learning models, interpreting results, and preparing the manuscript. All phases of the study were completed independently without any external assistance.

ACKNOWLEDGEMENT

The author expresses heartfelt gratitude to Manish Lalani, and Kayan Lalani, for their constant emotional support, patience, and encouragement throughout the course of this research work. Their unwavering belief in the author's work has been a great source of strength. The author also extends sincere thanks to her colleagues and research guide for their valuable guidance, insightful suggestions, and continuous support, which greatly contributed to the successful completion of this study.

REFERENCES

1. A. Rajendran, R. Srinivasan, and S. Kumar, "IoT-based air quality monitoring system," IEEE Sensors Journal, vol. 16, no. 8, pp. 2345-2351, 2016.
2. J. Gao, L. Li, and Y. Zhang, "A real-time air quality monitoring system based on IoT sensors for urban environments," Environmental Science and Pollution Research, vol. 26, no. 12, pp. 11957-11966, 2019.
3. M. Jin, X. Zhang, and H. Li, "Multi-sensor network for air quality monitoring: Applications and developments," Sensors and Actuators B: Chemical, vol. 240, pp. 1066-1075, 2017.
4. X. Zhou, Z. Wu, and F. Yu, "A hybrid machine learning approach for air quality prediction," Journal of Environmental Management, vol. 218, pp. 345-356, 2018.
5. S. Patel, A. Agarwal, and M. Sharma, "A machine learning approach for AQI prediction based on meteorological data," Environmental Monitoring and Assessment, vol. 192, no. 10, pp. 1-10, 2020.
6. W. Zhang and T. Wang, "A multi-sensor data fusion approach for AQI prediction," Journal of Environmental Informatics, vol. 27, no. 3, pp. 225-235, 2016.
7. J. Liu, F. Zhao, and S. Sun, "Ensemble learning for air quality prediction using multiple machine learning techniques," Science of the Total Environment, vol. 650, pp. 1043-1051, 2019.
8. X. Chen, H. Xu, and Y. Li, "AQI forecasting using gradient boosting machine," Environmental Science and Technology, vol. 54, no. 12, pp. 8236-8244, 2020.

9. M. Khan, M. Rehman, and S. Choi, "Ensemble learning approach for air quality prediction," Springer Environmental Science and Pollution Research, vol. 25, no. 3, pp. 2397-2405, 2018.
10. I. Hussain, M. Qureshi, and S. Farhan, "Deep ensemble learning for AQI prediction," IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 6, pp. 2021-2031, 2020.
11. M. Hossain, M. Alam, and S. Rahman, "Air quality prediction using feature extraction and machine learning," Sensors, vol. 19, no. 11, pp. 1-14, 2019.
12. A. Singh, A. Shukla, and S. Pandey, "Cloud-based real-time air quality prediction using IoT sensors," Environmental Pollution, vol. 229, pp. 456-463, 2017.
13. S. Amin, S. Hu, and M. Lin, "Cloud-based predictive model for AQI using IoT data," Computers, Environment and Urban Systems, vol. 68, pp. 35-42, 2018.
14. A. Siddique, S. Rashid, and S. Khan, "Challenges in IoT-based air quality monitoring systems," Environmental Monitoring and Assessment, vol. 193, no. 1, pp. 1-15, 2021.
15. M. Xie, T. Zhang, and Y. Zhang, "Real-time air quality prediction with IoT sensors: Issues and solutions," Journal of Environmental Science and Technology, vol. 54, no. 9, pp. 5734-5742, 2020.
16. S. Rashid, M. Shah, and A. Khan, "IoT-based real-time air quality monitoring system," Environmental Science and Pollution Research, vol. 25, no. 4, pp. 3547-3556, 2018.
17. A. Kumar and R. Gupta, "Prediction of AQI using machine learning algorithms," Journal of Environmental Informatics, vol. 23, no. 3, pp. 197-206, 2019.
18. A. Alam, M. Rahman, and S. A. Ali, "Ensemble learning methods for AQI prediction," Environmental Monitoring and Assessment, vol. 192, no. 12, pp. 1-10, 2020.
19. Y. Zhao, J. Wang, and J. Liu, "AQI prediction using deep learning models," Journal of Environmental Management, vol. 218, pp. 345-356, 2017.
20. S. Chakraborty, A. Saha, and S. Dey, "Data fusion for air quality prediction using multiple sources," Environmental Science and Technology, vol. 54, no. 15, pp. 9690-9699, 2020.
21. UCI Machine Learning Repository, “Air Quality Dataset,” Kaggle, 2020.

About Author

Dr. Kavita Ahuja

Dr. Kavita K. Ahuja is working as an Assistant Professor working with a reputed institute affiliated with Veer Narmad South Gujarat University, Surat. She Awared Ph.D degree in Computer Science from the Hemchandracharya North Gujarat University, Patan, India. She has more than 15 years of teaching and research experience. She has published more than two National Books in Computer Science area. She has also published many research papers in National and International various UGC, Scopus and peer-reviewed journals. Her areas of research are Data Analytics, Bıg Data, Internet of Things (IoT) and Machine Learning.