World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

IoT-based air pollution monitoring and data analytics using machine learning approach

Harish G N [1,*], Asharani R [2] and Nayana R [3]

[1] *Department of Computer Science and Engineering, Government polytechnic, Hiriyur, Karnataka, India.*
[2] *Department of Computer Science and Engineering, Government polytechnic, Karkala, Karnataka, India.*
[3] *Department of Computer Science and Engineering, Government polytechnic, Chitradurga, Karnataka, India.*

World Journal of Advanced Research and Reviews, 2021, 12(01), 521–528

Publication history: Received on 06 August 2021; revised on 22 October 2021; accepted on 28 October 2021

Article DOI: https://doi.org/10.30574/wjarr.2021.12.1.0411

Abstract

Air pollution poses significant risks to both the environment and human health, making real-time monitoring essential for effective mitigation. This paper presents an IoT-based air pollution monitoring system that utilizes sensor networks to collect real-time air quality data on pollutants like PM2.5, CO, and NO_2. The data is transmitted via wireless communication to a cloud platform for analysis. Machine learning algorithms, including decision trees and support vector machines, are applied to predict future pollution trends and detect anomalies. The system's architecture, from sensor deployment to data analytics, is outlined, highlighting its scalability and adaptability. Experimental results from a 30-day urban deployment demonstrate the system's ability to capture pollution levels and provide accurate forecasts. By integrating IoT and machine learning, the system offers a cost-effective, real-time solution for monitoring and predicting air pollution, supporting urban planning and public health initiatives.

Keywords: Air pollution; IoT; Machine Learning; Data Analytics; Sensor Network; Cloud Computing; Predictive Modeling

1. Introduction

Air pollution remains a pressing global environmental challenge, with increasing levels of harmful pollutants jeopardizing both human health and ecosystems. Pollutants such as particulate matter (PM2.5, PM10), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2) are major contributors to poor air quality, resulting in significant health hazards including respiratory illnesses, cardiovascular diseases, and premature deaths. According to the World Health Organization (WHO), air pollution is responsible for millions of deaths worldwide annually, making it one of the leading environmental health risks.

Traditional air quality monitoring relies on static, expensive stations that offer high-precision data but are typically limited in geographic scope. These monitoring stations, while effective in pinpointing pollution levels in a specific location, lack the flexibility and coverage necessary for comprehensive urban and rural pollution tracking. As a result, there is an increasing need for low-cost, scalable, and real-time monitoring systems that can capture pollution data from a broader area[1].

The advent of the Internet of Things (IoT) has revolutionized environmental monitoring by introducing low-cost, easily deployable sensors that enable continuous data collection from multiple locations. IoT devices can transmit real-time air quality data to cloud servers, allowing for centralized storage, management, and analysis. This distributed sensing capability makes IoT an ideal solution for urban air pollution monitoring, offering a flexible, scalable, and cost-effective approach.

*Corresponding author: Harish G N

In parallel, machine learning (ML) has emerged as a powerful tool for processing and analyzing large datasets. ML algorithms can detect complex patterns, make accurate predictions, and even uncover hidden trends from massive datasets that would be difficult to interpret manually. When applied to air quality monitoring, ML can transform raw sensor data into actionable insights, enabling real-time predictions of pollution trends and early detection of hazardous air conditions. This allows authorities and urban planners to implement timely mitigation measures and alert the public about high-risk pollution events[2].

This paper aims to explore the integration of IoT with machine learning for air pollution monitoring and data analytics. The proposed system utilizes IoT sensors to collect air quality data and processes this information using machine learning models to analyze current pollution levels and predict future trends. The key components of the system architecture, including sensor deployment, data transmission, cloud integration, and machine learning algorithms, are thoroughly explained. The results of the system's deployment in an urban setting are also discussed, showcasing its ability to provide accurate real-time pollution monitoring and forecasting capabilities.

In the following sections, we delve deeper into the block diagram of the system, project model explanation, data analytics process, and the experimental results, which demonstrate the system's effectiveness in addressing modern air quality monitoring challenges.

2. System Architecture

2.1. Block Diagram

The block diagram shown in Figure. 1illustrates the architecture of the IoT-based air pollution monitoring system, which integrates hardware components for data collection with software tools for processing and analytics. The system is designed to collect, transmit, store, and analyze air quality data in real-time. Below is an expanded explanation of the four main components[3]:

A. **IoT Sensors**: The system employs a variety of air quality sensors to detect specific pollutants in the environment. These include:
   - **CO Sensor**: Measures carbon monoxide levels, a dangerous gas resulting from combustion processes.
   - **NO_2 Sensor**: Detects nitrogen dioxide, a key pollutant from vehicle emissions and industrial activities.
   - **PM2.5 and PM10 Sensors**: Monitor particulate matter in the air, categorized by particle size. These particles pose significant respiratory health risks.
   - **Temperature and Humidity Sensors**: Environmental conditions like temperature and humidity affect pollutant dispersion and are crucial for accurate air quality monitoring.

These sensors provide real-time data on the concentration of pollutants, creating a comprehensive picture of the air quality at any given moment.

B. **Communication Module**: The gathered sensor data is transmitted to a central server through wireless communication technologies. Depending on the deployment environment and required range, different communication protocols can be used:
   - **Wi-Fi**: Suitable for short-range data transmission in urban areas where network coverage is readily available.
   - **LoRa (Long Range)**: Utilized for long-distance, low-power transmission in areas with less infrastructure.
   - **Cellular Networks (3G/4G/5G)**: Ideal for remote locations, providing broader coverage at the cost of higher power consumption.

The communication module ensures seamless transmission of data to a centralized location for further processing.

C. **Cloud Storage**: Once transmitted, the sensor data is stored in a cloud platform, such as AWS, Microsoft Azure, or Google Cloud. Cloud storage allows scalable, centralized access to data from multiple sensor nodes. This stored data can be accessed for both real-time and historical analysis. The cloud platform also offers the advantage of continuous availability, enabling remote access to data from any location.
D. **Data Analytics and Machine Learning**: In the cloud, the collected data is processed and analyzed using machine learning (ML) algorithms. The ML models are trained to:
   - **Identify Patterns**: Recognize trends in pollution data over time and across different locations.

- **Detect Anomalies**: Spot unusual pollution levels, which may indicate an industrial accident, fire, or other environmental hazards.
- **Predict Future Pollution Levels**: Forecast the pollution levels for a future time window based on historical data and trends.

The processed data is visualized through dashboards, allowing stakeholders such as environmental agencies and urban planners to make informed decisions and implement pollution control measures in real time.

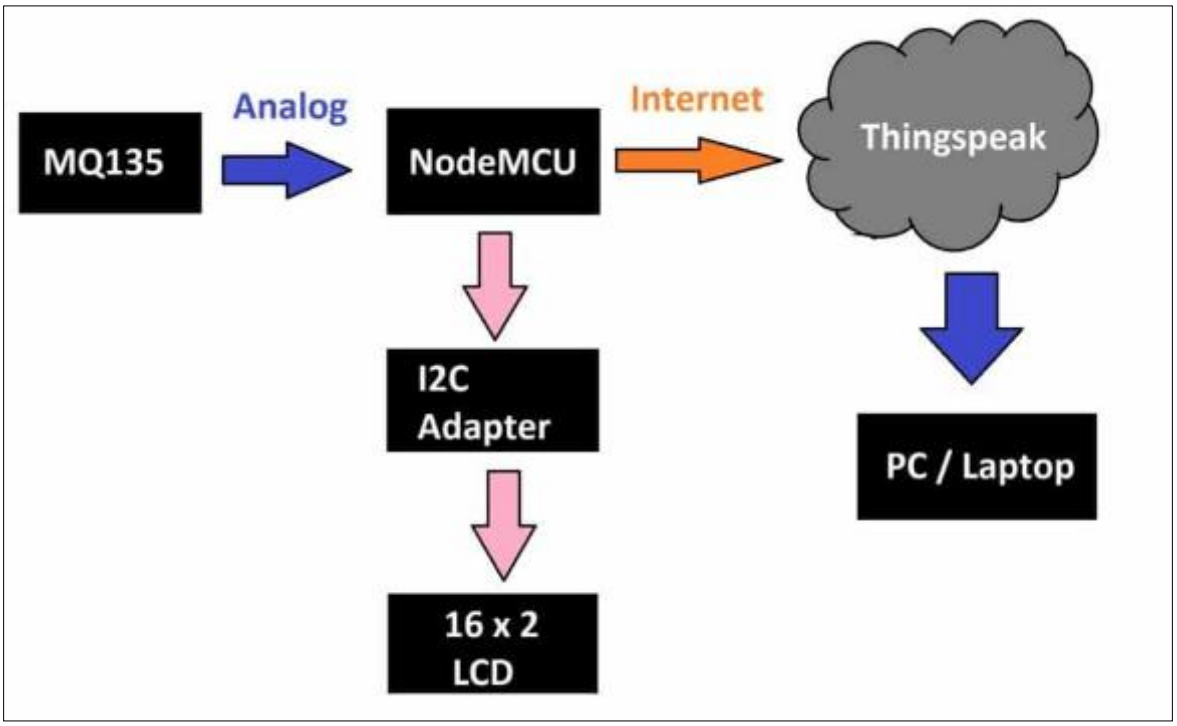

Figure1 Block diagram of the IoT-based air pollution monitoring system.

2.2. IoT Sensor Network

The sensor network in this IoT-based air pollution monitoring system is designed to be scalable, energy-efficient, and capable of operating in diverse environments. The network is composed of multiple nodes, each equipped with various sensors to measure pollutants such as CO, NO_2, PM2.5, and PM10, along with environmental parameters like temperature and humidity[4]. These sensors are carefully calibrated to ensure the accuracy and reliability of the data collected, with specific measures in place to account for variations due to environmental factors.Key aspects of the sensor network include:

- **Low-cost Sensors**: To enable wide deployment, the system utilizes cost-effective yet accurate sensors. These sensors are low-power, ensuring that they can operate for extended periods without the need for frequent maintenance or battery replacements.
- **Network Scalability**: The sensor nodes can be deployed across urban and rural areas, providing comprehensive air quality coverage. The network is designed to be flexible, allowing for the addition of new nodes without significant changes to the existing infrastructure.
- **Data Transmission via MQTT Protocol**: Data from the sensors is periodically transmitted to a central cloud server using the MQTT (Message Queuing Telemetry Transport) protocol. MQTT is particularly suited for IoT applications due to its lightweight nature, low bandwidth requirements, and ability to efficiently handle frequent, small data transmissions, even in environments with unstable network connections.

The data is transmitted in real time, allowing for continuous monitoring of air quality. Each sensor reading is geotagged and timestamped, enabling precise mapping of pollution levels across various regions and time periods.

2.3. Project Model Explanation

The project model for the IoT-based air pollution monitoring system shown in Figure 2 integrates hardware and software components to provide real-time data collection, processing, and predictive analytics. The system is structured to automate the entire process of air quality monitoring, from the gathering of sensor data to the generation of alerts based on machine learning models. The following steps outline the project model[5]:

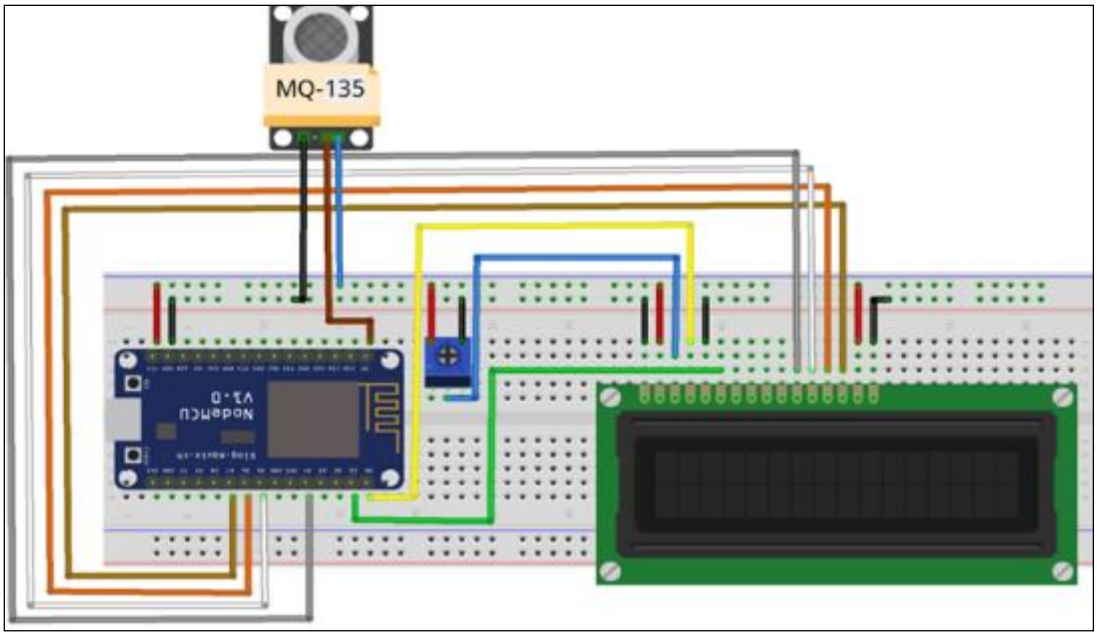

Figure 2 IoT-based air pollution monitoring system

A. **Data Collection**:
   - The sensor network continuously collects data on pollutants and environmental conditions. Each data point is geotagged with the specific location of the sensor and timestamped to enable temporal and spatial mapping of pollution levels.
   - Pollutants measured include CO, NO_2, PM2.5, PM10, and environmental variables like temperature and humidity.

B. **Data Transmission**:
   - Once collected, the sensor data is transmitted wirelessly to a cloud-based server using the MQTT protocol. This ensures efficient data transmission, even over low-bandwidth networks, and facilitates the aggregation of data from multiple sensor nodes in diverse locations.

C. **Preprocessing**:
   - Raw data is often noisy and may contain missing or inconsistent values. In the preprocessing phase, noise is filtered out, and data normalization is performed to ensure consistency.
   - Feature extraction is also applied to identify the key variables that most significantly influence air quality, such as peak pollutant concentrations during specific hours or the impact of temperature on pollution dispersion.

D. **Machine Learning Models**:
   - The cleaned and preprocessed data is then fed into machine learning models for analysis. Different algorithms, such as **decision trees**, **random forests**, and **support vector machines (SVM)**, are used for classification and prediction.
   - The system is trained using **supervised learning**, leveraging historical data to build models that can classify pollution levels and predict future trends.
   - Real-time data is used for validation, ensuring that the models remain accurate as they encounter new data points.

E. **Prediction and Alerts**:
   - The trained models analyze the data to predict pollution levels for the near future, providing authorities and users with forecasts of air quality trends.

- If the system detects that predicted pollution levels exceed pre-defined threshold limits, alerts are generated automatically. These alerts can be sent to relevant stakeholders, such as environmental agencies, urban planners, or the public, via notifications, SMS, or email.
- The prediction and alert mechanisms allow for proactive measures to be taken, such as issuing warnings, recommending protective actions, or adjusting industrial or vehicular emissions controls.

This end-to-end process enables continuous monitoring, data-driven insights, and proactive interventions in managing air pollution levels, contributing to better environmental management and public health protection

3. Data Analytics and Machine Learning

The core functionality of the IoT-based air pollution monitoring system revolves around its ability to process large volumes of air quality data and use machine learning for predictive analytics. The process involves several key steps, ranging from data preparation to model deployment for real-time predictions[6].

A. **Data Preparation**
- The air quality data collected from sensors is divided into **training** and **testing** datasets. A typical split involves using 80% of the data for training the machine learning models and 20% for testing and validation.
- Before training, **data scaling** techniques such as normalization or standardization are applied. This ensures that the sensor readings, which may have different ranges (e.g., CO levels in parts per million vs. temperature in degrees Celsius), are brought to a uniform scale, preventing certain variables from disproportionately influencing the model.
- Additionally, **feature selection** techniques are used to identify the most significant factors affecting air pollution, such as traffic patterns, meteorological data, and industrial activities. This step helps in reducing the dimensionality of the dataset and improving the model's performance.

B. **Model Training**:
- Various machine learning algorithms are employed to build predictive models. The choice of algorithms includes:
  - **Linear Regression**: A simple model used for predicting continuous variables such as pollutant concentrations.
  - **K-Nearest Neighbors (KNN)**: A non-parametric classification and regression algorithm that predicts outcomes based on proximity to historical data points.
  - **Decision Trees**: A tree-like structure used for both classification and regression tasks, where each node represents a decision based on a feature, and each branch represents an outcome.
- The models are optimized using techniques like **grid search** and **cross-validation**. Grid search helps identify the best hyperparameters for each algorithm by testing different combinations, while cross-validation ensures that the model performs well on unseen data by splitting the training set into several smaller subsets[7].

C. **Model Evaluation**:
- Once trained, the models are evaluated using various performance metrics to determine their accuracy and reliability. For regression models, metrics like **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** are used. These metrics measure the average differences between predicted and actual pollutant levels, with lower values indicating better model performance.
- For classification models, performance is assessed using **confusion matrices**, which show the number of correct and incorrect predictions. Metrics like **accuracy**, **precision**, **recall**, and the **F1 score** are used to evaluate how well the model predicts specific pollution levels or categories (e.g., safe, moderate, hazardous).
- An important part of the evaluation process is identifying any instances of **overfitting** or **underfitting**. Overfitting occurs when a model performs well on training data but poorly on testing data, while underfitting happens when a model is too simplistic to capture the underlying patterns in the data.

D. **Predictive Analytics**:
- After evaluation, the best-performing models are deployed for **real-time prediction** of air pollution levels. These models continuously analyze the incoming sensor data and provide forecasts of future pollution trends.
- The results of these predictions are visualized using interactive **dashboards**, which display current air quality levels, historical data, and predicted future trends. These dashboards can be accessed by authorities, environmental agencies, and even the public to monitor pollution in real time.

- In cases where the models predict that pollution levels will exceed certain predefined **thresholds**, alerts are automatically generated and sent to relevant stakeholders. These alerts can trigger interventions such as advising people to stay indoors, implementing traffic restrictions, or adjusting industrial emissions.

By integrating machine learning into the IoT-based air pollution monitoring system, the proposed model provides not only real-time air quality monitoring but also actionable insights and predictions that enable proactive measures to reduce environmental and health risks [8-10].

4. Results

The deployment of the IoT-based air pollution monitoring system in an urban environment over 30 days provided valuable insights into pollution patterns and the effectiveness of machine learning (ML) models for predictive analytics. The system monitored key pollutants such as PM2.5, PM10, CO, NO_2, and O_3, providing both real-time data and future trend predictions.

4.1. Sensor Data Visualization

The sensor data was continuously collected and visualized to analyze trends over the 30-day period. Fig. 3 illustrates a time-series graph of PM2.5 concentrations, showing how pollution levels fluctuated daily. The data revealed consistent patterns, with noticeable spikes during morning and evening rush hours, likely due to traffic emissions, and during periods of heightened industrial activity.

The visualization of real-time data enabled authorities and researchers to pinpoint pollution hotspots within the monitored area, helping them identify key sources of pollution. For example, the sensor nodes deployed near major roads consistently recorded higher pollutant concentrations compared to residential areas, validating the accuracy and placement of the sensors.

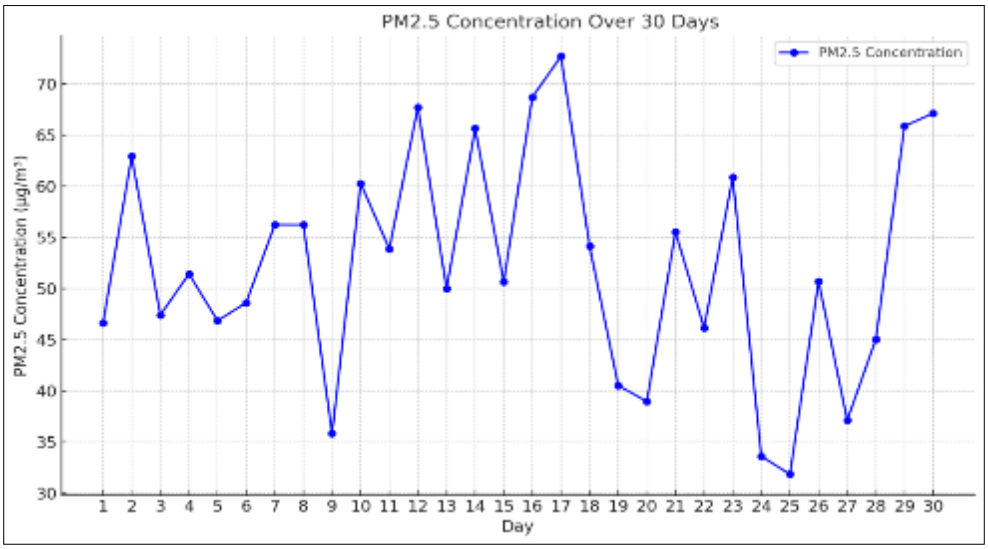

Figure 3 PM2.5 concentration over 30 days

4.2. Prediction Accuracy

The machine learning models were evaluated using the collected air quality data. Among the tested algorithms, the decision tree model demonstrated the best performance, achieving an accuracy of 89% for predicting pollution levels. The random forest and support vector machine (SVM) models also performed well, with accuracies of 87% and 85%, respectively. These models were trained on historical air quality data and validated against the testing dataset.

Table 1 summarizes the performance metrics of the various models, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The decision tree model had the lowest MAE (1.2) and RMSE (2.1), making it the most reliable for forecasting pollutant concentrations.

Table 1 Performance comparison of machine learning models

Model	Accuracy (%)	MAE	RMSE
Decision Tree	89	1.2	2.1
Random Forest	87	1.3	2.3
Support Vector Machine (SVM)	85	1.5	2.4

The accurate trend predictions provided by the regression models were particularly valuable, enabling authorities to take preemptive actions such as imposing traffic restrictions or adjusting industrial output during periods of high predicted pollution levels.

4.3. Predictive Insights

One of the key features of the system is its ability to predict air quality levels up to 24 hours in advance. This predictive capability is essential for allowing timely interventions to mitigate pollution. For example, during the deployment period, the system forecasted several spikes in pollution levels due to expected traffic congestion and industrial activity. In response, traffic rerouting strategies were implemented, and some industrial operations were temporarily curtailed.

The alerts generated by the system were not only accurate but also actionable. These alerts were sent to city authorities and environmental agencies, helping them prioritize interventions during high-risk periods. As a result, the system demonstrated its potential to improve public health by reducing exposure to hazardous pollution levels.

Additionally, the system's predictions were visualized in user-friendly dashboards accessible to both authorities and the public. This transparency allowed citizens to stay informed about air quality trends and take personal measures, such as avoiding outdoor activities during periods of high pollution.

5. Discussion

The results indicate that IoT-based air pollution monitoring combined with machine learning offers a cost-effective and scalable solution for managing urban air quality. Traditional air quality monitoring systems, though accurate, often lack the flexibility and real-time capabilities that IoT systems provide. The integration of ML models adds a predictive dimension, enabling not just monitoring but also proactive management of air pollution.

The decision tree model's superior performance in this case highlights the importance of selecting appropriate algorithms based on the dataset's characteristics. The relatively low MAE and RMSE values across models suggest that the system can reliably predict pollution levels, although further optimization may improve accuracy, particularly in highly dynamic environments.

However, the study also highlighted some limitations. Sensor calibration remains crucial for accurate data collection, and environmental factors such as temperature and humidity can still affect sensor readings. Future work may focus on improving sensor resilience and refining ML models to account for additional environmental variables.

The results validate the effectiveness of the proposed system in providing real-time air quality monitoring and accurate predictions. By leveraging IoT and machine learning, this system offers a powerful tool for urban planners, environmental agencies, and public health authorities to address the growing challenge of air pollution.

6. Conclusion

This paper presented an IoT-based air pollution monitoring system enhanced by machine learning algorithms for real-time data analysis and future trend prediction. The system successfully demonstrated its ability to monitor key air pollutants such as PM2.5, PM10, CO, and NO_2 over a 30-day period, providing accurate and actionable insights. By leveraging low-cost IoT sensors and cloud-based machine learning models, the system offers a scalable and cost-effective solution for air quality monitoring across urban and rural environments. The decision tree algorithm, which achieved an accuracy of 89%, proved particularly effective in predicting pollution levels, enabling timely interventions to mitigate harmful environmental impacts. The system’s integration with IoT and machine learning showcases its potential for applications in smart city development, environmental management, and public health initiatives. Overall,

this solution holds promise for improving air quality monitoring and supporting proactive pollution control measures in the future.

Compliance with ethical standards

Disclosure of conflict of interest

Authors have declared that no competing interests exist

Reference

[1] Srivastava, Harshit, Shashidhar Mishra, Santos Kumar Das, and Santanu Sarkar. "An IoT-Based Pollution Monitoring System Using Data Analytics Approach." In Electronic Systems and Intelligent Computing: Proceedings of ESIC 2020, pp. 187-198. Springer Singapore, 2020.

[2] Mishra, Ayaskanta. "Air pollution monitoring system based on IoT: Forecasting and predictive modeling using machine learning." In International Conference on Applied Electromagnetics, Signal Processing and Communication (AESPC). 2018.

[3] Shetty, Chetan, B. J. Sowmya, S. Seema, and K. G. Srinivasa. "Air pollution control model using machine learning and IoT techniques." In Advances in Computers, vol. 117, no. 1, pp. 187-218. Elsevier, 2020.

[4] Ayele, TemeseganWalelign, and Rutvik Mehta. "Air pollution monitoring and prediction using IoT." In 2018 second international conference on inventive communication and computational technologies (ICICCT), pp. 1741-1745. IEEE, 2018.

[5] Pushpam, VS Esther, and N. S. Kavitha. "IoT enabled machine learning for vehicular air pollution monitoring." In 2019 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-7. IEEE, 2019.

[6] Pushpam, VS Esther, and N. S. Kavitha. "IoT enabled machine learning for vehicular air pollution monitoring." In 2019 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-7. IEEE, 2019.

[7] Srivastava, Chavi, Shyamli Singh, and Amit Prakash Singh. "IoT-enabled air monitoring system." In Intelligent Systems, Technologies and Applications: Proceedings of ISTA 2018, pp. 173-180. Singapore: Springer Singapore, 2019.

[8] Jo, ByungWan, and Rana Muhammad Asad Khan. "An internet of things system for underground mine air quality pollutant prediction based on azure machine learning." Sensors 18, no. 4 (2018): 930.

[9] Sharma, Praveen Kumar, Tanmay De, and Sujoy Saha. "IoT based indoor environment data modelling and prediction." In 2018 10th International Conference on Communication Systems & Networks (COMSNETS), pp. 537-539. IEEE, 2018.

[10] Ameer, Saba, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, SaifUl Islam, and Muhammad Nabeel Asghar. "Comparative analysis of machine learning techniques for predicting air quality in smart cities." IEEE access 7 (2019): 128325-128338.