Jurnal Teknik Elektro, Vol. 19, No. 1, Maret 2026: 18-24
P-ISSN 1411-870X, E-ISSN 2656-3304

AIR QUALITY PREDICTION USING IOT AND MACHINE LEARNING

Iwan Handoyo Putro
Electrical Engineering Department, Petra Christian University
Siwalankerto 121-131, Surabaya 60236, Indonesia
E-Mail: iwanhp@petra.ac.id

Abstract – Air pollution has become a critical environmental and public health concern, particularly in urban areas where industrial activities and transportation contribute significantly to particulate matter emissions. The emergence of Internet of Things (IoT) technologies has enabled continuous and real-time monitoring of environmental conditions through distributed sensor networks. However, raw sensor data alone is insufficient without intelligent analysis for accurate forecasting and decision-making. This study proposes a machine learning-based approach for air quality prediction using IoT-derived environmental data. The Beijing PM2.5 dataset was utilized to simulate real-world IoT sensor measurements, incorporating meteorological and temporal features. Three machine learning models: Linear Regression, Random Forest, and Gradient Boosting were implemented and evaluated using standard performance metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). Experimental results indicate that the Random Forest model achieved the best performance, with an RMSE of 47.05, and R^2 score of 0.75. In comparison, Gradient Boosting produced an RMSE of 66.27 and R^2 of 0.50, while Linear Regression showed the lowest performance with an RMSE of 80.14 and R^2 of 0.27. These results demonstrate that tree-based ensemble methods, particularly Random Forest, are more effective in capturing the nonlinear relationships present in environmental data. This work highlights the potential of integrating IoT sensing with machine learning models to support accurate air quality prediction and informed environmental management.

Keywords – Internet of Things, Air Quality Prediction, Beijing PM2.5 dataset, Machine Learning

I. INTRODUCTION

Air pollution is one of the most pressing environmental challenges that is facing modern society. It has significant impacts on human health, climate change, and overall quality of life. Fine particulate matter, particularly PM2.5, poses serious health risks due to its ability to penetrate deep into the respiratory system and bloodstream [1]. Rapid urbanization, industrial emissions, and increasing vehicular activity have further exacerbated air quality issues, especially in densely populated cities [2].

Traditional air quality monitoring systems rely on a limited number of stationary monitoring stations, which often provide insufficient spatial and temporal coverage [3]. These systems are typically expensive to deploy and maintain, resulting in delayed or incomplete environmental information [4]. As a result, there is a growing need for scalable, real-time, and cost-effective solutions for monitoring and predicting air pollution levels [5].

The Internet of Things (IoT) has emerged as a promising technology to address these limitations [6]. IoT-based environmental monitoring systems utilize interconnected sensors to continuously collect data such as temperature, humidity, wind conditions, and pollutant concentrations. These sensor networks enable high-resolution data acquisition across wide geographic areas, providing a more comprehensive understanding of environmental conditions [7]. Moreover, IoT systems facilitate real-time data transmission and integration with cloud-based platforms for further analysis [8].

Despite the advantages of IoT-based data collection, the large volume and complexity of sensor data present significant challenges for interpretation and decision-making [9]. Raw sensor measurements alone are insufficient for proactive environmental management without predictive capabilities. This is where machine learning techniques play a crucial role. By analyzing historical and real-time IoT data, machine learning models can identify patterns, capture nonlinear relationships, and generate accurate predictions of air pollution levels [10], [11], [12].

In this study, we propose a machine learning-based framework for air quality prediction using IoT sensor data [13]. Publicly available environmental datasets are utilized to simulate real-world IoT measurements , enabling the evaluation of predictive models without the need for physical sensor deployment. Three machine learning algorithms: Linear Regression (LR), Random Forest (RF), and Gradient Boosting (GB) are implemented and compared to assess their effectiveness in predicting PM2.5 concentrations.

The main contributions of this work are as follows: (1) the development of a data-driven approach for air quality prediction using IoT-based environmental data, (2) a comparative analysis of multiple machine learning models for PM2.5 forecasting, and (3) the demonstration of the effectiveness of ensemble learning methods in improving prediction accuracy. The results of this study provide insights into the integration of IoT and machine learning for intelligent environmental monitoring systems.

The remainder of this paper is organized as follows. Section II presents the related work on air quality prediction and IoT-based environmental monitoring. Section III describes the dataset, preprocessing steps, and proposed methodology. Section IV discusses the experimental results and performance evaluation of the implemented machine learning models. Finally, Section V concludes the paper and outlines potential directions for future work.

II. RELATED WORK

Air quality monitoring and prediction have become increasingly important research areas due to the rapid growth of urbanization and industrial activities. The adverse effects of air pollution, particularly fine particulate matter (PM2.5) [14], [15], on human health and environmental sustainability have motivated the development of intelligent monitoring and forecasting systems. In recent years, the integration of Internet of Things (IoT) technologies with data-driven approaches has emerged as a promising solution for addressing these challenges.

Conventional air quality monitoring systems are typically based on a limited number of high-cost monitoring stations equipped with precise analytical instruments. While these systems provide accurate measurements, they suffer from limited spatial coverage and lack real-time responsiveness. To overcome these limitations, IoT-based environmental monitoring systems have been widely investigated [16]. IoT systems utilize distributed sensor networks to collect environmental data such as temperature, humidity, wind speed, and pollutant concentrations in real time. These sensors are interconnected through wireless communication protocols and transmit data to centralized platforms for storage and analysis .

Several studies have demonstrated the effectiveness of IoT-based air quality monitoring systems. For instance, [5] proposed an IoT-enabled pollution monitoring framework that integrates sensor networks with cloud computing to provide real-time data visualization. Similarly, [3] developed a scalable IoT architecture for urban air quality monitoring, highlighting the advantages of low-cost sensors in achieving high spatial resolution. Despite these advancements, many IoT-based systems focus primarily on data acquisition and visualization, with limited emphasis on predictive analytics. As a result, these systems are often reactive rather than proactive, providing information only after pollution levels have already increased.

To address this limitation, researchers have increasingly incorporated machine learning techniques into IoT-based environmental systems. Machine learning models enable the extraction of meaningful patterns from large volumes of sensor data and facilitate the prediction of future air quality conditions. Early approaches in air quality prediction relied on statistical models such as multiple linear regression and autoregressive integrated moving average (ARIMA) [17]. Although these methods are computationally efficient and easy to implement, they often fail to capture the complex nonlinear relationships between meteorological variables and pollutant concentrations.

In recent years, advanced machine learning algorithms have been widely applied to improve prediction accuracy. Tree-based models such as Decision Trees and Random Forests have gained popularity due to their ability to model nonlinear relationships and handle high-dimensional data. For example, [18] demonstrated that regression-based models significantly outperform traditional ML techniques in predicting PM2.5 concentrations. Similarly, ensemble learning methods such as GB have been shown to provide robust performance by combining multiple weak learners into a strong predictive model. These methods are particularly effective in capturing interactions between environmental variables and reducing overfitting.

Deep learning techniques have also been explored for air quality prediction, particularly for time-series data. Models such as artificial neural networks (ANN), convolutional neural networks (CNN), and long short-term memory (LSTM) networks have been used to capture temporal and spatial dependencies in environmental data. LSTM models, in particular, have demonstrated strong performance in forecasting tasks due to their ability to retain long-term dependencies. However, deep learning approaches often require large datasets, extensive computational resources, and careful parameter tuning, which may limit their practicality for real-time IoT applications, especially in resource-constrained environments.

Despite the significant progress in this field, several challenges remain in the integration of IoT and machine learning for air quality prediction. One major challenge is the quality and reliability of sensor data. Low-cost IoT sensors are prone to noise, calibration errors, and missing values, which can negatively impact model performance. Additionally, environmental data is inherently dynamic and influenced by multiple external factors, making accurate prediction a complex task.

Another important limitation identified in the literature is the lack of comprehensive comparative studies. Many existing works focus on a single machine learning model and report its performance without benchmarking against alternative approaches. This makes it difficult to determine the most suitable model for a given application. Furthermore, some studies emphasize system design without adequately evaluating predictive performance, while others focus solely on modeling without considering practical IoT deployment scenarios.

Moreover, scalability and real-time processing remain open challenges. As IoT networks generate large volumes of data continuously, efficient data processing and model deployment become critical. Edge computing and lightweight machine learning models have been proposed as potential solutions; however, their integration with air quality prediction systems is still an active area of research. These challenges highlight the need for approaches that balance prediction accuracy, computational efficiency, and practical applicability.

In this context, the present study contributes to the existing literature by providing a comparative analysis of multiple machine learning models for air quality prediction using IoT-derived environmental data. Unlike many previous studies that focus on a single model, this work evaluates LR, RF and GB techniques under the same experimental conditions. The use of a publicly available dataset ensures reproducibility and allows for fair comparison with existing studies. The results provide insights into the effectiveness of ensemble learning methods in capturing nonlinear relationships and improving prediction performance.

Furthermore, this study adopts a practical perspective by simulating IoT-based data collection using real-world environmental datasets. This approach bridges the gap between theoretical model development and real-world IoT deployment. By combining IoT-based sensing concepts with machine learning-based prediction, this work contributes to the development of intelligent environmental monitoring systems capable of supporting proactive decision-making and sustainable urban management.

III. METHODOLOGY

This section describes the dataset, preprocessing steps, and machine learning models used for air quality prediction. The overall methodology aims to simulate an IoT-based

environmental monitoring system by utilizing real-world sensor data and applying machine learning techniques for predictive analysis.

The dataset used in this study is the Beijing PM2.5 dataset obtained from the UCI Machine Learning Repository. This dataset contains hourly air quality and meteorological observations collected from Jan 1st 2010 to Dec 31st 2014. It includes measurements of PM2.5 concentration along with environmental variables such as temperature, dew point, atmospheric pressure, wind direction, wind speed, and precipitation indicators.

The PM2.5 concentration is considered the target variable in this study, while the remaining meteorological and temporal attributes are used as input features. The dataset contains missing values in several attributes, which are handled during the preprocessing stage. The use of this dataset allows the simulation of IoT-based environmental sensor data, as it reflects real-world measurements typically collected by distributed sensing devices.

Data preprocessing is a critical step to ensure the quality and reliability of the machine learning models. First, rows with missing PM2.5 values were removed to maintain consistency in the target variable. For the input features, missing values were handled using median imputation for numerical variables and most frequent value imputation for categorical variables.

A datetime variable was constructed from the year, month, day, and hour attributes to preserve temporal information. Additional time-based features such as day of the week and month index were extracted to enhance the predictive capability of the models.

Categorical features, particularly wind direction, were transformed using one-hot encoding to convert them into a numerical format suitable for machine learning algorithms. Numerical features were standardized using feature scaling to ensure uniformity and improve model convergence.

Three machine learning models were implemented in this study to evaluate their effectiveness in air quality prediction:

1. Linear Regression is used as a baseline model due to its simplicity and interpretability. It assumes a linear relationship between input features and the target variable. Although it is computationally efficient, it may not capture complex nonlinear patterns present in environmental data.
2. Random Forest is an ensemble learning method based on multiple decision trees. It improves prediction accuracy by aggregating the outputs of several trees and reducing overfitting. This model is particularly effective in handling nonlinear relationships and interactions between variables.
3. GB is another ensemble technique that builds models sequentially, where each new model attempts to correct the errors of the previous one. It is known for achieving high predictive performance, especially in structured datasets, by optimizing the loss function iteratively.

The dataset was divided into training and testing subsets to evaluate model performance. A chronological split was applied, where the earlier portion of the dataset was used for training and the later portion for testing, ensuring a realistic time-series prediction scenario.

Each model was implemented using a pipeline that integrates preprocessing and model training steps. This approach ensures consistency and prevents data leakage during model evaluation.

Model performance was evaluated using standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). MAE measures the average absolute difference between actual and predicted values. MSE quantifies the average squared deviation between actual and predicted values. RMSE penalizes larger errors more significantly, and R^2 indicates the proportion of variance in the target variable explained by the model.

The overall processing workflow of the proposed system consists of the following steps:

1. Data acquisition from the dataset simulating IoT sensor measurements,
2. Data cleaning and preprocessing, including handling missing values and feature engineering,
3. Feature transformation through encoding and normalization,
4. Model training using LR, RF, and GB,
5. Model evaluation and comparison using performance metrics.

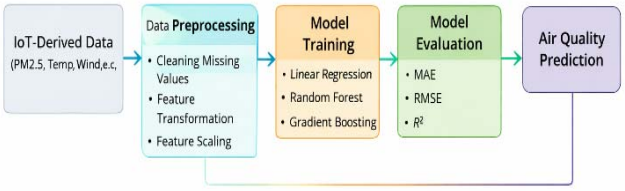

Figure 1. Proposed Methodology for Air Quality Prediction.

Figure 1 overviews the proposed methodology for air quality prediction. The framework consists of IoT-based data acquisition, data preprocessing (cleaning, feature transformation, and scaling), machine learning model training (LR, RF, and GB), model evaluation using standard metrics, and final air quality prediction.

This workflow represents a typical IoT-based air quality prediction pipeline, where environmental data collected from sensors is processed and analyzed using machine learning techniques to generate predictive insights.

IV. DISCUSSION

This section presents the experimental results of the implemented machine learning models and provides a detailed analysis of their performance in predicting PM2.5 concentrations using IoT-derived environmental data.

The performance of the three models: LR, RF and GB were evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). The results indicate that the RF model achieved the best overall performance, with the lowest prediction error and highest explanatory power among the evaluated models.

Specifically, RF obtained an RMSE of 47.05 and an R^2 score of 0.75, outperforming both GB and LR. In comparison, GB achieved an RMSE of 66.27 and an R^2 of 0.50, while LR produced the highest error with an RMSE of 80.14 and the lowest R^2 score of 0.27. These results demonstrate that ensemble-based methods are more effective than linear models in capturing the complex relationships within environmental data.

Table 1 shows the results of MAE calculation conducted using JupyterNotebook and Python script. It indicates that RF model

achieved highest performance by producing lowest MAE score (31.21). On the other hand, LR model perform poorly by getting 56.91 of MAE score.

Table 1. Performance Calculation Result

Model	MAE	RMSE	R^2
RF	31.21	47.05	0.75
GB	45.60	66.27	0.50
LR	56.91	80.14	0.27

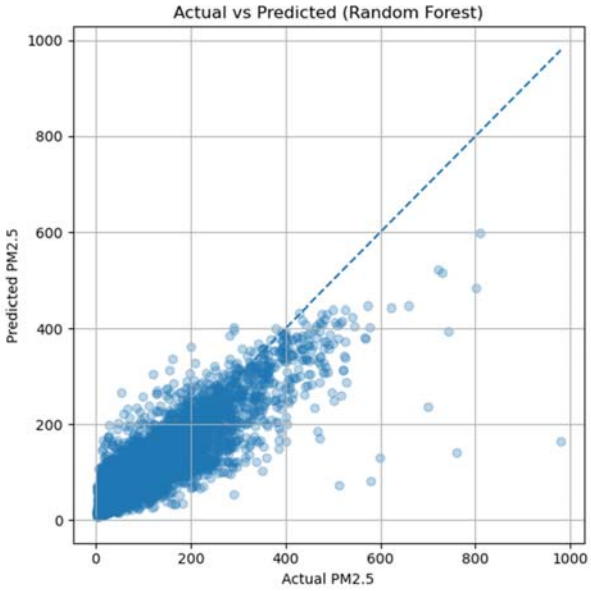

Figure 2. Actual vs Predicted PM2.5 Values Using RF Model

The scatter plot in Figure 2 illustrates the relationship between the actual and predicted PM2.5 values using the RF model. Each point represents a single observation, where the x-axis corresponds to the actual PM2.5 concentration and the y-axis represents the predicted value. The dashed diagonal line indicates the ideal case where predicted values perfectly match the actual values.

As shown in the figure, a large number of data points are distributed close to the diagonal line, particularly in the lower to moderate PM2.5 range (0–300), indicating that the model achieves good prediction accuracy within this range. This suggests that the RF model is effective in capturing the general trends and relationships in the dataset.

However, as the PM2.5 values increase beyond approximately 400, the spread of the data points becomes more pronounced, with several predictions deviating significantly from the ideal line. This indicates that the model tends to underestimate or exhibit higher variability in predictions for extreme pollution levels. Such behavior may be attributed to the limited number of high-value samples in the dataset or the increased complexity of environmental conditions during severe pollution events.

Overall, the figure confirms that the RF model provides reliable predictions for most observations, while highlighting some limitations in handling extreme values. This aligns with the quantitative results, where RF achieved the lowest prediction error and highest R^2 score among the evaluated models.

The superior performance of RF can be attributed to its ability to model nonlinear interactions between meteorological variables and PM2.5 concentration. Environmental data is inherently complex and influenced by multiple interdependent factors such as temperature, wind conditions, and atmospheric pressure. LR, which assumes a linear relationship, is unable to capture these interactions effectively, resulting in higher prediction errors.

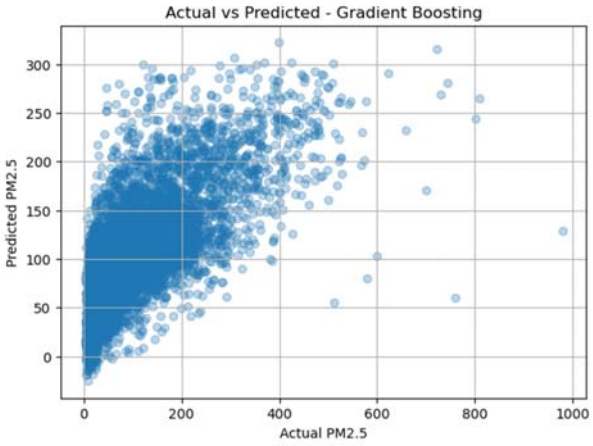

Figure 3. Actual vs Predicted PM2.5 Values Using GB Model

The scatter plot in Figure 3 presents the relationship between the actual and predicted PM2.5 values using the GB model. Each point represents an individual observation, with the x-axis indicating the actual PM2.5 concentration and the y-axis showing the corresponding predicted value.

As observed in the figure, the data points exhibit a wider dispersion compared to the RF model, indicating lower prediction accuracy. While the model captures the general increasing trend between actual and predicted values, many points deviate significantly from the ideal diagonal alignment, particularly in the mid to high PM2.5 ranges.

In the lower concentration range (0–200), the model demonstrates moderate predictive capability, although the spread of points suggests noticeable prediction errors. As the PM2.5 values increase beyond 300, the predictions become more scattered and less consistent, with several outliers indicating substantial underestimation or overestimation. This behavior suggests that the GB model struggles to generalize effectively across the full range of pollution levels in the dataset.

Overall, the figure highlights the limitations of the GB model in this study, particularly in handling extreme values and maintaining consistent prediction accuracy. These observations are consistent with the quantitative results, where GB exhibited higher RMSE and lower R^2 compared to the RF model.

GB, although theoretically powerful, showed lower performance in this study. This may be due to suboptimal hyperparameter settings or sensitivity to noise in the dataset. Unlike RF, which reduces variance through averaging, GB builds models sequentially and may be more prone to overfitting or underfitting if not properly tuned. Additionally, the presence of missing values and variability in sensor data may further impact its performance.

The scatter plot in Figure 4 illustrates the relationship between the actual and predicted PM2.5 values using the LR model. Each point represents an observation, with the x-axis corresponding to the actual PM2.5 concentration and the y-axis indicating the predicted value.

As shown in the figure, the data points are widely dispersed and deviate significantly from the ideal diagonal trend, indicating poor prediction accuracy. Although a general positive relationship between actual and predicted values can be observed, the model fails to accurately capture the variability of PM2.5 concentrations across different ranges.

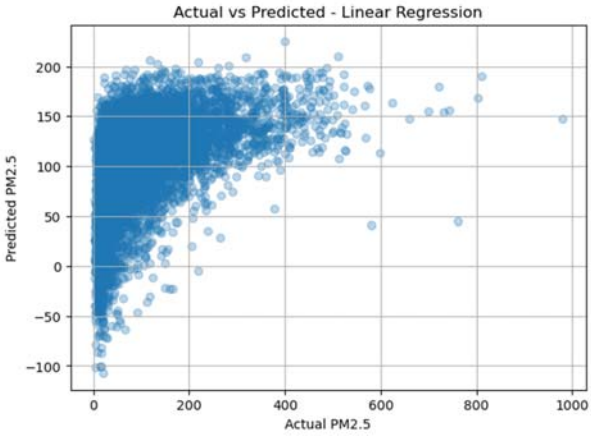

Figure 4. Actual vs Predicted PM2.5 Values Using LR Model

In particular, the model exhibits substantial errors in both low and high concentration ranges. At lower PM2.5 levels, there is a large spread of predictions, including several negative predicted values, which are not physically meaningful in this context. At higher PM2.5 levels, the model tends to underestimate the actual values, as indicated by the clustering of predictions below the expected trend. This suggests that LR is unable to model the nonlinear relationships present in environmental data effectively.Overall, as the weakest model in this experiment, the figure highlights the limitations of LR for air quality prediction, demonstrating its inability to capture complex patterns in the dataset. These observations are consistent with the quantitative results, where LR achieved the highest prediction error and lowest R^2 score among the evaluated models.

The graphical analysis of model predictions further supports the quantitative results. The comparison plots show that predictions from the RF model are more closely aligned with the actual PM2.5 values, with data points distributed near the ideal diagonal line. In contrast, LR exhibits a wider spread of prediction errors, indicating its inability to accurately capture variations in pollution levels.

The error comparison charts, including RMSE and MSE, clearly illustrate the performance differences among the models. RF consistently achieves lower error values, confirming its robustness and reliability for this prediction task. The R^2 comparison further highlights that RF explains a significantly higher proportion of variance in the dataset compared to the other models.

Figure 5 presents the comparison of Root Mean Squared Error (RMSE) values for the three evaluated machine learning models: LR, GB, and RF. RMSE is used as a performance metric to measure the average magnitude of prediction error, where lower values indicate better model performance.

As shown in the figure, the RF model achieves the lowest RMSE value (approximately 47), indicating the highest prediction accuracy among the models. In contrast, GB exhibits a higher RMSE value (around 66), reflecting moderate predictive performance. LR demonstrates the poorest performance, with the highest RMSE value (approximately 80), indicating significant prediction errors.

The results clearly highlight the superiority of ensemble-based methods, particularly RF, in handling complex and nonlinear relationships in environmental data. The substantial gap in RMSE between RF and LR further emphasizes the limitations of linear models for air quality prediction tasks.

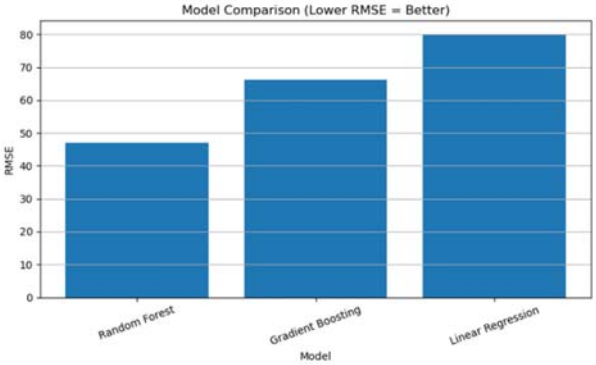

Figure 5. Comparison of RMSE Values for Different Machine Learning Models.

Overall, the figure confirms that RF is the most effective model for predicting PM2.5 concentrations in this study, aligning with the results observed in other evaluation metrics such as MSE and R^2.

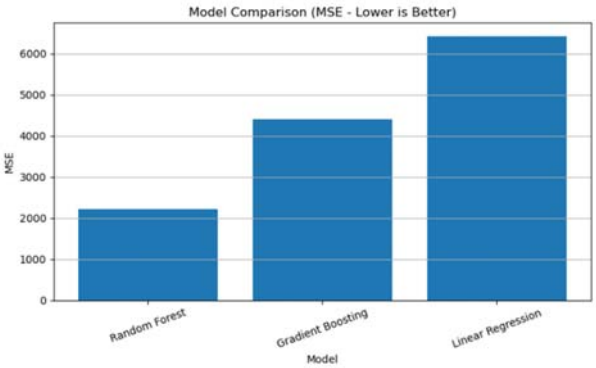

Figure 6. Comparison of Mean Squared Error (MSE) for Different Machine Learning Models.

Figure 6 illustrates the comparison of Mean Squared Error (MSE) values for the three machine learning models evaluated in this study: LR, GB, and RF. MSE measures the average squared difference between actual and predicted values, where lower values indicate better predictive performance.

As shown in the figure, the RF model achieves the lowest MSE value (approximately 2200), indicating the highest prediction accuracy among the evaluated models. GB demonstrates moderate performance with an MSE of around 4400, while LR exhibits the highest error, with an MSE exceeding 6400.

The substantial difference in MSE values highlights the effectiveness of ensemble learning methods in modeling complex environmental data. RF, in particular, benefits from its ability to reduce variance through aggregation of multiple decision trees, resulting in more stable and accurate predictions. In contrast, LR fails to capture nonlinear relationships, leading to significantly higher prediction errors.

Overall, the figure reinforces the findings from the RMSE analysis, confirming that RF provides the most reliable performance for PM2.5 prediction in this study.

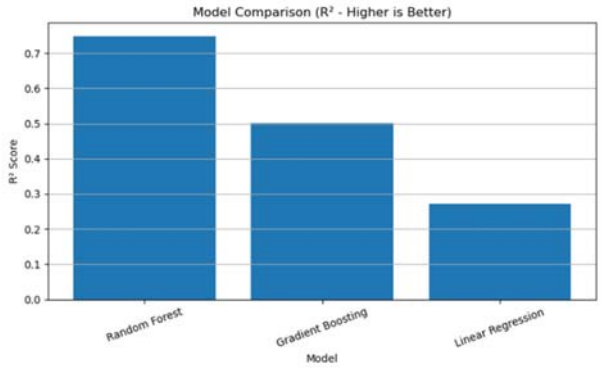

Figure 7. Comparison of R^2 scores for Different Machine Learning Models.

Figure 7 presents the comparison of the coefficient of determination (R^2) for the three evaluated machine learning models. The R^2 metric measures the proportion of variance in the target variable that is explained by the model, where higher values indicate better predictive performance.

As illustrated in the figure, the RF model achieves the highest R^2 score (approximately 0.75), indicating that it explains a substantial portion of the variability in PM2.5 concentrations. GB demonstrates moderate performance with an R^2 value of around 0.50, while LR shows the lowest explanatory power, with an R^2 score of approximately 0.27.

The clear difference in R^2 values highlights the advantage of ensemble-based learning methods in modeling complex environmental data. RF, through its aggregation of multiple decision trees, effectively captures nonlinear relationships and interactions among meteorological variables. In contrast, LR is limited by its assumption of linearity, resulting in significantly lower performance.

Overall, the figure confirms that RF provides the most accurate and reliable predictions in this study, as it achieves the highest level of variance explanation among the evaluated models. These findings are consistent with the results obtained from error-based metrics such as RMSE and MSE.

The results of this study have important implications for the design of IoT-based environmental monitoring systems. While IoT sensors enable real-time data collection, the integration of machine learning models is essential for transforming raw data into actionable insights. The findings suggest that ensemble learning methods, in particular RF, are well-suited for deployment in such systems due to their robustness and ability to handle noisy sensor data.

Moreover, the use of publicly available datasets to simulate IoT sensor data demonstrates that effective predictive models can be developed without requiring costly hardware deployment. This approach can accelerate the development of intelligent air quality monitoring solutions, especially in resource-constrained environments.

Despite the promising results, this study has several limitations. First, the dataset used represents a single geographic location, which may limit the generalizability of the findings to other regions with different environmental conditions. Second, only three ML models were evaluated, and further improvements may be achieved by exploring additional models or advanced deep learning techniques.

Future work may include the use of multi-site datasets, integration with real-time IoT systems, and optimization of model hyperparameters. Additionally, deploying lightweight models on edge devices could enable real-time prediction and decision-making in practical IoT environments.

V. CONCLUSION

This study presented a machine learning-based approach for air quality prediction using IoT-derived environmental data. By utilizing the Beijing PM2.5 dataset to simulate real-world sensor measurements, three models: LR, RF, and GBwere implemented and evaluated. The experimental results demonstrated that the RF model achieved the best performance, with the lowest prediction error and highest R^2 score, indicating its effectiveness in capturing nonlinear relationships in environmental data.

The findings highlight the importance of integrating machine learning techniques with IoT-based monitoring systems to enable accurate and proactive air quality prediction. Compared to traditional linear models, ensemble learning methods provide improved robustness and predictive capability, making them suitable for real-world environmental applications.

Despite these promising results, this study is limited by the use of a single-location dataset and a restricted set of machine learning models. Future work may focus on incorporating multi-site datasets, exploring advanced deep learning approaches, and deploying lightweight predictive models in real-time IoT environments. Such advancements can further enhance the development of intelligent and scalable air quality monitoring systems.

REFERENCES

[1] C. A. Pope, 3rd and D. W. Dockery, "Health effects of fine particulate air pollution: lines that connect," *J Air Waste Manag Assoc,* vol. 56, no. 6, pp. 709-42, Jun 2006.

[2] F. Karagulian *et al.*, "Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level," *Atmospheric Environment,* vol. 120, pp. 475-483, 2015.

[3] E. G. Snyder *et al.*, "The changing paradigm of air pollution monitoring," *Environ Sci Technol,* vol. 47, no. 20, pp. 11369-77, Oct 15 2013.

[4] N. Castell *et al.*, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?," *Environ Int,* vol. 99, pp. 293-302, Feb 2017.

[5] A. R. Al-Ali, I. Zualkernan, and F. Aloul, "A Mobile GPRS-Sensors Array for Air Pollution Monitoring," *IEEE Sensors Journal,* vol. 10, no. 10, pp. 1666-1671, 2010.

[6] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Computer Networks,* vol. 54, no. 15, pp. 2787-2805, 2010.

[7] P. Asha *et al.*, "IoT enabled environmental toxicology for air pollution monitoring using AI techniques," *Environ Res,* vol. 205, p. 112574, Apr 1 2022.

[8] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems,* vol. 29, no. 7, pp. 1645-1660, 2013.

[9] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications,* vol. 19, no. 2, pp. 171-209, 2014.

[10] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM,* vol. 55, no. 10, pp. 78-87, 2012.

[11] E. Gladkova and L. Saychenko, "Applying machine learning techniques in air quality prediction," *Transportation Research Procedia,* vol. 63, pp. 1999-2006, 2022.

[12] S. M. S. D. Malleswari and T. K. Mohana, "Air pollution monitoring system using IoT devices: Review," *Materials Today: Proceedings,* vol. 51, pp. 1147-1150, 2022.

[13] S. Chen. (2017, 7 April). *Beijing PM2.5*. Available: https://archive.ics.uci.edu/dataset/381/beijing+pm2+5+data

[14] H. Alrashidi, F. N. Sibai, A. Abonamah, M. Alrashidi, and A. Alsaber, "PM2.5: Air Quality Index Prediction Using Machine Learning: Evidence from Kuwait's Air Quality Monitoring Stations," *Sustainability,* vol. 17, no. 20, 2025.

[15] A. Makhdoomi, M. Sarkhosh, and S. Ziaei, "PM(2.5) concentration prediction using machine learning algorithms: an approach to virtual monitoring stations," *Sci Rep,* vol. 15, no. 1, p. 8076, Mar 8 2025.

[16] S. M. Popescu *et al.*, "Artificial intelligence and IoT driven technologies for environmental pollution monitoring and management," *Frontiers in Environmental Science,* vol. 12, 2024.

[17] G. Mani, J. K. Viswanadhapalli, and A. A. Stonier, "Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models," *Journal of Engineering Research,* vol. 10, no. 2, pp. 179-194, 2022.

[18] H. K. S. Doreswamy, Yogesh KM, Ibrahim Gad,, "Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models," *Procedia Computer Science,* vol. 171, pp. 2057-2066, 2020.