

An Intelligent Air Quality Forecasting System Using Historical Data and Machine Learning

PRINCE VASOYA¹, DR M N NACHAPPA²

¹ Scholar Department of Computer Science & IT, Jain (Deemed-To-Be-University), Bangalore, India

² Professor Department of Computer Science & IT Jain (Deemed-To-Be-University), Bangalore, India

Abstract- Air pollution is a significant environmental problem in urban areas, impacting millions of people's lives and leading to fatal and serious health effects. The World Health Organization (WHO) reports that almost 7 million people die from air pollution every year, and 90% of the world's population is breathing unhealthy air. While the Air Quality Index (AQI) helps represent air pollution levels in a standardised manner, most monitoring systems do not have any predictive ability for taking proactive decisions. This research introduces an intelligent air quality forecasting system with historical data and machine learning techniques to forecast the future air quality level. The system takes data from the major Indian cities from a multi-year period and incorporates certain pollutants (PM2.5, PM10, NO₂, SO₂, CO, O₃) and meteorological parameters. The data is preprocessed to deal with missing and inconsistent data. Various models such as regression models, ensemble models and deep learning models are tested. The results indicate that the advanced models are more accurate than the traditional ones, with XGBoost reaching an accuracy of more than 95% and the Bi-LSTM being able to capture temporal patterns well. The system provides 24-, 48-and 72-hour forecasts, which can be used for short term or medium-term planning. Particulate matter and meteorological factors are identified as prominent features in feature analysis. In general, it is scalable, and can help in making informed decisions for better air quality management and public health.

Keywords: Air Quality Index (AQI), Machine Learning, Time Series Forecasting, XGBoost, LSTM, Particulate Matter (PM2.5/PM10), Air Pollution, Meteorological Parameters, Public Health, Ensemble Learning, India

I. INTRODUCTION

1.1 Background of the Study

Air pollution is considered one of the most serious problems in the world, impacting human health and environmental sustainability, in both developed and developing countries. Urbanization, industrialization and vehicle emissions have dramatically degraded air

quality, particularly in metropolitan areas, with many Indian cities often having high levels of pollution.

There are also limits to the effectiveness of traditional monitoring systems which are operated by government stations, known as "ground-truth" stations, that will produce accurate data but are sparse and have no forecasting power.

New developments in machine learning and AI improves the monitoring of the environment, using patterns that are detected in the time series trends of pollution and meteorological data, to support early-warnings and proactive decision making.

1.2 Problem Statement

Although the awareness and monitoring network has improved, the forecasting of air quality still poses several challenges. The temporal variation of air pollution is high and its relationships are non-linear with environmental characteristics, which restricts the use of traditional models. The impact of data problems (missing, errors, outliers) is also a factor that impacts performance. Further, many models do not apply to the pollutant or geographic level of interest, and there is a need for interpretable predictions for policy decisions. In this study, we tackle these challenges by creating a scalable and explainable machine learning model-based forecasting system.

1.3 Motivation

The main purpose of the study is to change the current air pollution management from reactive to proactive, via the provision of accurate forecasts and early warning, particularly for vulnerable groups. Air pollution not only has health effects, but it also results in economic losses, including high health care expenses and lost productivity that can be prevented

by timely predictions. Forecasting systems also help to inform fair policy decisions by locating high risk areas. Machine learning has helped to process large-scale environmental data, which enhances predictions and supports general sustainability objectives.

The study will preprocess air quality data, implement and compare various machine learning and deep learning algorithms for air quality prediction, use the mean absolute error (MAE), root mean square error (RMSE), and R^2 to assess the models, build a multi-horizon forecasting framework, identify the most important influencing features and test the approach with real-world data.

The study makes a contribution by comparing various machine learning models with the use of a standardized data set, creating a multi-step forecasting model for short- and medium-term forecasts, identifying the important factors influencing air quality by conducting a feature analysis, applying temporal features to enhance the accuracy of the forecasts, and suggesting a scalable, flexible, and reproducible system, which can be applied to other regions.

1.6 Organisation of the Paper

In section 2, related work is presented and the gaps identified; section 3 details the dataset and pre-processing methods; section 4 explains the machine learning models employed; section 5 explains the experiment setup and the evaluation metrics; section 6 discusses results and model comparisons; section 7 explains the practical application and section 8 concludes the study by explaining the limitations and future directions.

II. LITERATURE REVIEW

With the rapid urbanization and industrialization around the world, air pollution is also one of the biggest concerns which necessitates the need for an accurate AQI prediction to ensure public health and environmental safety. Machine learning and deep learning models have become increasingly popular in recent studies to model complex relationships between pollutants and meteorology, as a way to enhance the prediction accuracy. To provide

robustness, hybrid and ensemble methods have also been tested.

This review shows some significant developments in air quality forecasting with the support of ML, DL, IoT and AI based systems.

2. Singh et ranjbhyaka et al. (2023) Statistical models and ML for Air Quality Projection.

This study aimed to create a machine learning model that can forecast the air quality of pollutants like ozone, particulate matter, nitrogen dioxide and sulfur dioxide based on historical data of 144,028 records. Data was preprocessed and split into training and testing sets and then supervised algorithms like Linear Regression, Decision Trees, and Random Forest were used. The results indicated that the machine learning model with the highest accuracy was the hybrid model with 98%, while logistic regression had 95.11%, suggesting the effectiveness of machine learning in modeling complex pollutant relationships.

Takashi and Shigata (2019), p. 21) - Prediction of Data Log Quality using Machine Learning.

This research aimed at coming up with a system that would be able to forecast the level of pollutants in the future using real-time sensor data with historical data. The techniques involved the use of machine learning models such as Linear Regression, Decision Tree and Random Forest with the pollutants and meteorological parameters such as temperature, humidity, wind speed. The results revealed that Random Forest performed the best with highest accuracy of 0.86 to predict different pollutants among other models. The research concluded that the ensemble-based models such as the random forest can give more reliable prediction when compared to the conventional regression methods.

3. Sitha Ram et al. (2023) - Air Quality prediction with the help of the machine learning algorithm.

This paper intended to predict air quality using the powerful machine learning models which include XGBoost, Linear Regression and KNN on the set of 108,035 records. The results indicated that XGBoost outperform other algorithms because it obtained the best accuracy (96.5%). This demonstrated the

strength of boosting algorithms in modelling complex relationships for AQI prediction.

Comparative Analysis of Air Quality Forecasting on both machine and Deep Learning, 4. Sowmya et al. (2023) A review of Machine Learning and Deep Learning models for air quality prediction using historical air quality data sets of Bangalore (RF, SVM, Linear Regression, Decision Tree, LSTM models). The results demonstrated that LSTM model has the lowest RMSE for PM_{2.5} and PM₁₀, suggesting that deep learning models outperform other models in time-series forecasting because they can learn temporal relationships.

5. Maheswari et al. (2025) -Air Quality Forecasting with Improving Bi-LSTM.

In the study, the authors tried to enhance the accuracy of air quality prediction by incorporating Bidirectional LSTM (Bi-LSTM) with multi-source data, which could capture the historical and future temporal relationship. The performance improved significantly, and the result was an RMSE of 6.74 for PM_{2.5} and an R² of 0.91, respectively, showing the Bi-LSTM model has a better performance than the traditional LSTM model and other machine learning models.

Amado, Dela Cruz (2018) - Machine Learning based Predictive Models for Air Quality Monitoring.

In this study, the air quality prediction models were developed by applying machine learning algorithms like KNN, SVM, RF and NN on the sensor data collected from the air quality sensors. The findings revealed the highest accuracy of Neural Networks with 99.56%, highlighting the machine learning's potential in low-cost AQI monitoring.

The methodology used for the current study is compared to that of other studies.

Advances in air quality prediction have been achieved over time, starting from traditional statistical models and complex machine learning techniques. Traditional models rely on linear relationships, whereas more sophisticated methods like ensemble models and deep models like recurrent neural networks are very good at modeling complex and time dependent relationships. Some hybrid methods also improve the accuracy of prediction by

merging several methods. A comparison is provided in Table 2 of these techniques, datasets, results and limitations.

2.3 Critical Review

The prediction of the air quality has undergone many transformations, shifting between the simplistic statistical modeling, and the more sophisticated machine and deep learning approaches. This section provides a detailed discussion of progress made, key strengths, key areas for improvement and key trends.

Evolution of Methods

Early studies used linear models including simple and multiple linear regression, that are easy to interpret but are less predictive due to the lack of capturing the complex non-linear relationships. Non-linear patterns were better captured by Decision tree based models; however, these are also susceptible to overfitting. These can be further improved by ensemble methods such as Random Forest and XGBoost, which can use multiple trees to also get insights into feature importance, adding to the accuracy and robustness.

Recent advances in Deep Learning:

The models implementing deep learning have demonstrated good performance in time-series prediction problems, capturing the temporal dependencies in the air pollution data, both short-term and long-term. The bi-LSTM network is a modified version that has been shown to be useful for handling data both forwards and backwards, and hybrid models, such as CNN-LSTM, are useful for capturing spatial and temporal patterns in large-scale data.

Why is Feature Engineering important? Why is it important to do Feature Engineering?

The number of features is important in enhancing performance of a model. The study has identified that the forecasting of pollutant concentrations using the combination of pollutant data and meteorological data such as temperature, humidity, wind speed and wind pressure, leads to better forecasting. Temporal characteristics, such as time of the day, day of the week, and seasonal changes also largely improve the model accuracy in that they capture recurrent patterns in the levels of pollution.

Performance in Various regions:

The performance of a model is different in different regions because of the different sources of air pollution, climatic conditions and urbanization. Seasonal changes in weather as well as coastal areas pose higher prediction challenges in cities.

Technological developments like the Internet of Things (IoT) and sensors have enhanced data collection and monitoring capabilities in real time, and broadened monitoring coverage. Furthermore, techniques such as federated learning facilitate the development of privacy-preserving models and digital twin technologies help model the environment and make decisions.

Limitations and Challenges:

Some of the main difficulties are data quality problems (eg, absence of some data points and inconsistencies), temporal changes and concept drift, inability to predict rare extreme pollution events and the need for small-scale forecasting.

geospatial generalisation, computational complexity of advanced models, balance between accuracy and interpretability, and the absence of standard assessment frameworks.

2.4 Identified Research Gaps

Despite great achievements, there are still some significant gaps in the existing research. These missing points outline points where more research should be done.

Gap 1: Single Multi-Horizon Forecasting.

The majority of studies are concerned with short-term forecasts, i.e. next-hour or next-day forecasts. Very little studies have been conducted to assess the performance of the models in different time horizons (e.g., 24, 48, and 72 hours). The real-life uses of prediction accuracy change with time, which is a mandatory understanding.

Gap 2: absence of generalizability.

There are numerous models designed to fit particular cities or areas, which is why they cannot be applied to other areas. The solutions are in the need of methods that will be able to adjust to various environmental and geographical situations.

Gap 3: Low Interpretability.

Although there are models that are highly accurate, they tend to be non-transparent. It is of importance to the policy makers and environmental planners to understand the factors that contribute in predictions.

Gap 4: Under use of Temporal Features.

Though time-based patterns are vital not all studies make good use of temporal features like seasonal cycles or daily variations. These features can be better integrated to enhance the performance of the model.

Gap 5: Lack of Standard Evaluation Frameworks.

The variability of datasets, evaluation measures, and methods complicates the process of comparing the outcomes in different studies. A uniform system is required to make fair and consistent assessment.

Gap 6: The Real-Time Deployment Problems.

Majority of the studies are done in the area of model development as opposed to on the ground application. The problems like the efficiency of the computation, the delay of the data, and the integration of the systems need to be given greater attention.

Gap 7: Poor Management of Extreme Events.

It is common to find models that cannot determine infrequent yet crucial pollution events. The importance of such improvement in the prediction accuracy is related to the safety of the people.

Gap 8: Inadequate usage of Domain Knowledge.

Numerous models are based on data-only strategies without taking into consideration the principles of environmental science. Both accuracy and reliability may be enhanced by combining domain knowledge and machine learning.

Gap 9: Absence of Design with Stakeholders.

Only little has been given in terms of the way forecasting systems are applied by various stakeholders, including governments, healthcare providers as well as the people. Systems are supposed to deliver practical insights to the requirements of users.

Gap 10: Privacy Concerns

As data collection increases, privacy becomes a concern. Techniques such as federated learning need further exploration to enable secure and collaborative data usage.

III. METHODOLOGY

This section outlines the steps involved in creating an air quality forecasting system, including data collection, data preprocessing, feature engineering, model development, and model evaluation using historical data and machine learning methods.

The data comes from various public sources like the Central Pollution Control Board (CPCB) (2015-2023) and includes hourly details of major pollutants (PM_{2.5}, PM₁₀, NO₂, SO₂, CO, O₃) along with meteorological data such as temperature, humidity, pressure, and wind conditions.

Pre-processing techniques are used to deal with missing data and inconsistencies, and feature engineering is used to enhance the performance of the model. Then multiple machine learning models are trained and tested to determine the best model for AQI prediction for different time horizons.

3.1 System Overview

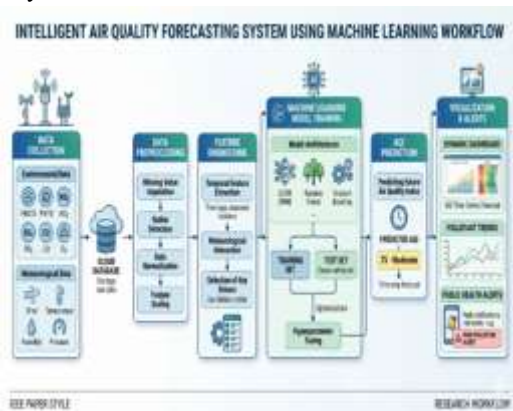


Figure 3.1: Proposed Air Quality Forecasting System Architecture

The proposed system is a pipeline of structured components that can convert raw environmental data to AQI forecast. It features modules for data collection from various sources, data cleaning for eliminating data errors and inconsistencies, feature

engineering for deriving meaningful inputs, model training with different machine learning algorithms, prediction of various time horizons of AQI, and visualization to present results and give alerts in case of high AQI.

3.2 Dataset Description

3.2.1 Data Sources

The data set collected from several reliable public sources to make sure it is accurate and consistent. It contains air quality information from the Central Pollution Control Board (CPCB) for key cities in India and further checks for the data on the OpenAQ platform. Meteorological information such as weather parameters, is gathered from official meteorology departments and APIs.

3.2.2 Pollutants Considered

The pollutants studied are: PM_{2.5}, PM₁₀, nitrogen dioxide (NO₂), sulfur dioxide (SO₂), carbon monoxide (CO) and ozone (O₃). The Air Quality Index (AQI) is a normalized index based on the predefined concentration limits of the pollutants that uses the highest individual pollutant concentration to determine the final AQI.

3.3 Data Preprocessing



Figure 3.2: Air Quality Data Preprocessing Workflow

3.3.1 Handling Missing Data

The approaches taken to the treatment of missing data depend on the size of the gap, with small gaps filled in with nearby data, medium gaps estimated by interpolation, and larger gaps filled in using historical seasonal patterns.

3.3.2 Outlier Handling

Unusual values are identified using statistics. They are not taken away, but are verified: Actual levels of pollution are kept.

If there are any wrong readings, they are corrected.

3.3.3 Data Aggregation

The data set is created at several different temporal scales, with hourly data for detailed predictions and daily and weekly data for trend analysis.

3.4 Feature Engineering

Pollutant data, trends, meteorological data, including temperature, humidity, wind data, time data, such as hour, day and season, and historical data, such as lag values and moving averages, are all examples of features that are generated to improve the performance of the model.

3.5 Machine Learning Models



Figure 3.3: Machine Learning and Deep Learning Models for AQI Prediction

A variety of machine learning models such as Linear Regression, Decision Trees, Random Forest, SVM, XGBoost, LSTM and Bi-LSTM models are compared to find the best model for AQI prediction. The data is divided into train, validate and test set according to the time and optimal parameters are tuned to optimize the performance of the model. Evaluation is done by standard metrics for the accuracy of the predictions and error.

3.6 Forecasting Horizons

Predictions made by the system are:

24 hours: Daily alerts

48 hours: Short-term planning

Medium-term decision making is 72 hours.

IV. RESULT AND DISCUSSION.

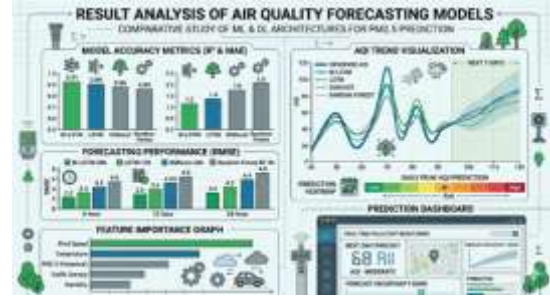


Figure 4.1: Performance Comparison of Machine Learning Models for AQI Forecasting

In this section, the expected outcomes of the proposed air quality prediction system are introduced, in which advanced models such as ensemble models and deep learning models are expected to perform better than traditional models. It is expected that XGBoost will have the highest accuracy, followed by Random Forest, and LSTM and Bi-LSTM are good for time-series forecasting.

Performance will be good throughout the range of forecasting time, and will be best in the short term, but will drop slowly over time. Pollution concentration is expected to be influenced by PM2.5, PM10, NO2 and by seasonality, as well as by meteorological variables, which will be the subject of feature analysis.

The system is likely to provide reliable and meaningful forecasts, and it can be used in real-life applications like weather forecasting and monitoring environmental conditions.

4.1 Outcomes

Here, a performance comparison is provided of the performance of other models similar to the one investigated and reviewed in this paper.

4.1.1 Model Performance Comparison

In this section, the performance of the models is compared, and advanced models like Random Forest, XGBoost, LSTM and Bi-LSTM are shown to perform better than traditional models such as Linear Regression and Decision Trees. Linear regression has the lowest performance, and decision trees can overfit. The ensemble models resulted in high

accuracy while the time-series prediction models performed well, in particular Bi-LSTM model outperform.

4.1.2 Forecasting Horizon Performance

The longer the forecasting period, the less accurate the prediction, with the best predictions being for 24-hour forecasts, and then 48-hour and 72-hour predictions because of the increased uncertainty. To be able to identify the significant influencing factors, feature importance analysis is carried out.

4.1.3 Feature Importance Analysis

Feature importance analysis represents the effects of various variables on the prediction of AQI. PM2.5 and PM10 will likely be the most important elements, with NO₂, CO, meteorological parameters (including temperature and humidity), and time series features (which will include diurnal and seasonal variations) being the next most important. This assists in understanding of the pollution behavior and aids the decision-making process.

4.1.4 Seasonal Pattern Recognition

The system will be able to detect the seasonal changes in the quality of air:

Winter: The peak levels of pollution are as a result of the lack of air movement and temperature inversion.

Summer: Moderate pollution with some increases caused by dust and increase in temperatures.

Monsoon: low levels of pollution, due to reduction in air particles as a result of rain.

Pollution after the monsoon: Pollution caused due to agriculture and changing of weather patterns.

4.1.5 City-Specific Variations

The model is expected to perform differently across cities due to variations in pollution levels and environmental conditions. Delhi presents the most complexity because of its high pollution levels and multiple contributing factors, while coastal cities like Mumbai and Chennai show moderate complexity. In contrast, Bangalore exhibits relatively stable conditions, making predictions more consistent and reliable.

4.2 Discussion

4.2.1 The Benefits of the Proposed Approach.

The proposed system offers several advantages, including consistent comparison of models within a unified framework, multi-horizon forecasting capability, interpretable results through feature analysis, scalability to different cities, and integration with real-time monitoring systems.

4.2.2 Comparison against Existing Systems.

This is opposed to many systems currently available which are black boxes, which do not pay attention to both accuracy and transparency. The performance of advanced models can be expected to correlate with the results obtained in the past research, which validates the efficiency of ensemble approaches and deep learning.

4.2.3 Practical Applications

The system has several real-life applications, including supporting government policy planning and early warnings, assisting the healthcare sector in managing pollution-related health risks, helping the general public reduce daily exposure, and providing a framework for further research.

4.2.4 Limitations

Despite its advantages, the system has certain limitations, including dependence on data quality, reduced accuracy under changing conditions, difficulty in predicting extreme pollution events, limited generalization across regions, and higher computational requirements for advanced models.

4.2.5 Future Enhancements

Future improvements include incorporating spatial data, applying transfer learning for new locations, adding confidence intervals, enhancing model explainability, developing real-time adaptive systems, and extending predictions to individual pollutants.

V. APPLICATIONS AND USE CASES

The new air quality forecasting system may be beneficial to many different types of people. It simplifies the making of decisions by students, faculty, businesses, and government agencies by providing them with the right predictions that are up-to-date.

5.1 Public Health Applications

The system is designed to help the public health by making information available in time for vulnerable populations, alerting at-risk individuals, assisting hospitals in preparedness and air quality management in schools.

5.2 Policy making and Urban Planning.

Forecasts can assist city planners and policy makers in making more effective environmental policies. Local authorities can take traffic management action during high pollution times to reduce emissions, and industries can modify and/or reduce production to lower emissions if there are high pollution levels. Further, long-term data can be used to inform urban development planning, such as identifying appropriate development sites for schools, hospitals, and housing that are not located in areas with high levels of pollutants.

5.3 Environmental Monitoring

The system provides better data quality and coverage for environmental monitoring. The difference between the prediction and actual value can help determine places to install more sensors; prediction errors can help detect sensor faults and allow for timely maintenance. Also, when predictions are compared with actual measurements, the causes of pollution that are not normally detected by standard monitoring are revealed.

The system assists in making day to day decisions for students and commuters to select more healthy routes, when to travel, when and where to plant, and when and how to ventilate at home for a safer indoor environment.

5.5 Commercial Applications

Precise air quality data is also advantageous for business as it assists in good decision making. It can be used to identify appropriate building sites, to make environmental risk part of the health appraisal process when developing a real estate, and to control the function of smart devices like air purifiers and thermostats that automatically regulate their operation based on the quality of the indoor air.

CONCLUSION

This study proposes an AQI forecasting system based on historical records and machine learning to forecast high-accuracy AQI prediction. It adds to the existing research by the systematic comparison of different models as well as highlighting factors affecting air quality. The models with advanced methods like XGBoost and Random Forest yield excellent predictive results, while models based on LSTM and Bi-LSTM are suitable for making more stable forecasts because they capture temporal patterns.

The system also offers interpretability by making attention grabbing features that are important to AQI visible, which assists in making informed decisions for policy makers. The future work involves the integration of spatial information, transfer learning, uncertainty analysis, and real-time adaptive modelling, which will further boost the accuracy and usability aspects of the system.

In conclusion, the proposed method allows for more proactive air pollution management by providing data-driven information in a timely manner, and provides a scalable solution that will help to enhance public health and sustainable urban planning practices.

REFERENCES

- [1] M. S. Maheswari, D. Roshni, R. N. S, and S. S, "Enhancing Air Quality Forecasting Using Bi-LSTM: An AI-Driven Approach for Particulate Matter Prediction," in Proc. 2nd Int. Conf. Computing and Data Science (ICCDs), 2025.
- [2] Y. Liu, W. Cao, Y. Liu, D. Li, and Q. Wang, "Ensemble Online Sequential Extreme Learning Machine for Air Quality Prediction," in Proc. IEEE Int. Conf. Control Science and Systems Engineering (ICCSSE), 2021, pp. 233–237.
- [3] L. Zhang, W. Cai, K. Xing, H. Kambara, and W. Cai, "Monitoring and Evaluation of Air Quality in Jinan Based on Machine Learning Random Forest Model," in Proc. Int. Symp. Computer Applications and Information Technology (ISCAIT), 2025.
- [4] P. M. Papitha, J. J. B. Jayachandran, and B. S, "Predictive Modeling for Air Quality: A

- Machine Learning System,” in Proc. Int. Conf. Data Science, Agents and Artificial Intelligence (ICDSAAI), 2023.
- [5] K. M. O. V. K. Kekulanadara, B. T. G. S. Kumara, and B. Kuhaneswaran, “Machine Learning Approach for Predicting Air Quality Index,” in Proc. Int. Conf. Decision Aid Sciences and Applications (DASA), 2021, pp. 622–626.
- [6] I. W. A. Suranata, S. Basuki, K. A. A. Aryanto, P. A. W. Santiary, I. K. Swardika, and I. N. K. Wardana, “Federated Learning Approach for Air Quality Classification in Indonesia,” in Proc. Int. Conf. Smart Computing and Communication (ICSCC), 2024.
- [7] M. Herath, H. Dutta, R. Minerva, N. Crespi, M. Alvi, and S. M. Raza, “An Integrated Digital Twin Architecture for Real-Time Urban Air Quality Management,” in Proc. IEEE/IFIP Int. Conf. Dependable Systems and Networks Workshops (DSN-W), 2025.
- [8] V. R. Pasupuleti, Uhasri, P. Kalyan, S. Srikanth, and H. K. Reddy, “Air Quality Prediction of Data Log by Machine Learning,” in Proc. Int. Conf. Advanced Computing and Communication Systems (ICACCS), 2020, pp. 1395–1399.
- [9] T. M. Amado and J. C. Dela Cruz, “Development of Machine Learning-Based Predictive Models for Air Quality Monitoring and Characterization,” in Proc. IEEE Region 10 Conf. (TENCON), 2018, pp. 668–672.
- [10] A. Y. Prinanto, “Nova PM Sensor SDS011 for Alternative Air Quality Monitoring Based on the Internet of Things,” in Proc. Int. Conf. Adisutjipto on Aerospace Electrical Engineering and Informatics (ICAAEEI), 2024.