Journal of Computation Physics and Earth Science
ISSN: 2776-2521 (online)
Volume 5, Number 1, April 2025, Page 31 - 39
https://journal.physan.org/index.php/jocpes/index

PM2.5 Concentration Prediction Model in Jakarta Area Using Random Forest Algorithm

Muhammad Naufal Afif Al Arsy[1] Ahmad Meijlan Yasir[1]
[1]Undergraduate Program in Applied of Instrumentation Meteorology, Climatology Geophysics (STMKG)

Article Info

Article history:

Received March 3, 2025
Revised March 8, 2025
Accepted March 9, 2025

Keywords:

PM2.5 Prediction
Random Forest
Air Quality
Machine Learning, Jakarta,
Hyperparameter Tuning
Environmental Monitoring
Data-Driven Policy

A B S T R A C T

This study predicts PM2.5 concentrations in Jakarta using the Random Forest algorithm with historical air quality data from 2015 to 2024. Hyperparameter tuning was performed to optimize model performance, focusing on parameters such as n_estimators, max_depth, and min_samples_split. The model achieved a Mean Absolute Error (MAE) of 14.44, a Root Mean Square Error (RMSE) of 18.75, and an R^2 Score of 0.61. While the model captured general PM2.5 fluctuation patterns, deviations at certain points indicate room for improvement. Descriptive analysis showed an average PM2.5 concentration of 94.46 µg/m³, with peaks up to 209 µg/m³, exceeding healthy air quality thresholds. The model can be integrated into real-time monitoring systems and support data-driven policies. Future work could incorporate meteorological variables and evaluate longer-term trends to enhance accuracy.

Corresponden Author:

Muhammad Naufal Afif Al Arsy,
Undergraduate Program in Applied of Instrumentation Meteorology, Climatology Geophysics (STMKG)
Tangerang City, Banten, Indonesia
Email: nono.naufal10@gmail.com

1. INTRODUCTION

PM2.5, a fine air particle with a diameter of ≤2.5 µm, is one of the most harmful pollutants to human health because its small size allows it to penetrate into the bloodstreams. Long-term exposure to PM2.5 has been shown to contribute to cardiovascular disease, chronic respiratory distress, and lung cancer. The WHO reports that air pollution, including PM2.5, causes more than seven million premature deaths each year worldwide (Mauboy et al., 2024).

For instance, the average ambient concentration of PM2.5 in Jakarta frequently violates the WHO safe limit of 10 µg/m³, and the significant sources include mobile and stationary emission from transport, industries, and burning of fossils. In addition, meteorological conditions such as temperature, humidity, wind speed, and rainfall also affect the distribution pattern of these pollutants (Goyal & Goyal, 2024; Joharestani et al., 2019). The study found that PM2.5 concentrations tend to increase during the dry season, where low rainfall reduces the atmosphere's ability to deposit particles (Kang et al., 2023).

Recent development of the machine learning has produced better solutions for handling and predicting air quality. Out of all the algorithms used in this study, it is worth mentioning the effectiveness of Random Forest in relation to nonlinear dependencies and texture complexity. This model has been well applied to predict PM2.5 using meteorological variables, satellite data and observation in the region. The findings suggest that Random Forest also has reasonable prediction accuracy compared to other methods, like XGBoost or SVM (Joharestani et al., 2019; Ma et al., 2023; Uzir et al., 2016).

For instance, a study that employed Random Forest technique in China obtained a good performance of up to 0.81 R^2 value in the determination of PM2.5 concentrations (Ma et al., 2023). A similar study conducted in Jakarta also supports the finding that the algorithm holds capability to predict the certain pollutants average concentration through meteorological parameters and air quality data (Joharestani et al., 2019; Mauboy et al.,

2024). Not only are such models useful for achieving high prediction accuracy, but they can also be employed to create warning systems for the public and generate decision-making support for governments (Chen et al., 2023). The focus of this study is to present a PM2.5 concentration prediction model for Jakarta based on the Random Forest algorithm. Based on the above literature review, the following research questions which are the main focuses of this study are identified.

First, what are the temporal and spatial patterns of PM2.5 concentrations in the Jakarta area based on historical data? Understanding this pattern is important because air pollution in Jakarta often varies depending on time and location, which can be studied using machine learning algorithms, such as Random Forest, which have been shown to be effective in predicting air pollution in various large urban areas(Goyal & Goyal, 2024).

The second problem is how do meteorological parameters such as temperature, humidity, wind speed, and rainfall affect PM2.5 concentrations in Jakarta? Previous research has shown that these factors greatly affect pollutant concentration levels, and understanding these relationships can improve the accuracy of the predictive models built (Joharestani et al., 2019; Ma et al., 2023).

The third problem that needs to be answered is how the Random Forest algorithm can be used to accurately predict PM2.5 concentrations in Jakarta. Random Forest is known for its ability to handle non-linear data and can identify important features that affect air quality, making it a good choice for air pollutant prediction (Kang et al., 2023; Ma et al., 2023).

Finally, how can the prediction results from this model be used to support air quality management policies in Jakarta? Accurate predictive models can provide early warning of increased PM2.5 concentrations, allowing for appropriate mitigation measures to protect public health. This will help the government in formulating air quality management policies based on more accurate data (Chen et al., 2023).

This study aims to analyze the temporal and spatial patterns of PM2.5 concentrations in Jakarta based on historical data. This analysis is important for understanding the distribution of air pollution by time (daily, weekly, or seasonal) and location (by region with high and low pollution levels). Several previous studies have shown that models like Random Forest are very effective in identifying such patterns, especially in large urban areas such as Beijing and Jing-Jin-Ji (Joharestani et al., 2019; Ma et al., 2023).

In addition, this study evaluated the relationship between meteorological parameters, such as temperature, humidity, wind speed, and rainfall, with PM2.5 concentrations in Jakarta. Meteorological factors play an important role in the distribution of air pollutants, where rainfall tends to lower PM2.5 concentrations, while humidity and wind speed can affect their spread (Kang et al., 2023; Ma et al., 2023).

This study also developed a PM2.5 concentration prediction model using the Random Forest algorithm. This model was chosen because of its ability to handle non-linear data and identify important features in complex datasets. The developed model will be tested using historical air quality data and meteorological parameters to ensure prediction accuracy (Chen et al., 2023; Ma et al., 2023). The results of the prediction are expected to be used to support air quality management policies in Jakarta, including providing early warning to the public about deteriorating air conditions (Chen et al., 2023; Kang et al., 2023).

This research has several benefits. For governments, the predicted results can be used to support data-driven policies for air pollution mitigation, such as controlling motor vehicle emissions and reducing industrial activities during periods with high pollution levels (Chen et al., 2023; Kang et al., 2023). For the public, predictive information can increase awareness of health risks and provide reliable data for preventive measures such as mask use or restrictions on outdoor activities (Chen et al., 2023; Joharestani et al., 2019). For academics, this study enriches the scientific literature related to the application of the Random Forest algorithm in air quality prediction and opens up opportunities for further research, such as integration with satellite data for more accurate predictions (Joharestani et al., 2019; Ma et al., 2023). Environmentally speaking, this research helps identify times and locations with high pollution risks, supporting more effective mitigation measures (Chen et al., 2023).

2. LITERATURE REVIEW

2.1 Air Quality Prediction Using Machine Learning Techniques

Air quality prediction has become a critical task for urban management due to the rising concerns about pollution and its effects on human health. Traditional statistical methods, such as linear regression, have been used to predict air quality levels, but they often fail to capture the complex, non-linear relationships found in air pollution data (Grell et al., 2005). In recent years, machine learning (ML) techniques have been widely adopted to improve the accuracy of air quality prediction models. These techniques can handle large datasets, identify complex patterns, and adapt to real-time data, making them suitable for urban air quality management.

2.2 Common Machine Learning Techniques for Air Quality Prediction

Machine learning methods such as Random Forest, Decision Trees, Artificial Neural Networks (ANN), and Support Vector Machines (SVM) have been applied to predict air quality indices (AQI) and

particulate matter (PM) concentrations. Each method has its strengths and limitations in terms of performance, accuracy, and computational complexity.

a. Random Forest

The Random Forest algorithm has been widely used for air quality prediction due to its robustness in handling noisy data and reducing overfitting (Ameer et al., 2019). In their study, (Yu et al., 2016) applied Random Forest for predicting air quality using urban sensing systems and found it to outperform other methods in terms of accuracy. Similarly, (Brokamp et al., 2017) compared regression and Random Forest approaches for estimating particulate matter levels, concluding that Random Forest provided more accurate predictions in urban environments(Liu et al., 2012; Livingston, 2005; Segal, 2003).

b. Support Vector Machines (SVM)

SVM is another popular machine learning algorithm used for air quality prediction(Lu' et al., 2002) utilized SVM for spatiotemporal analysis of air quality data derived from social media and found it to be effective in predicting AQI levels. However, the computational cost of SVM increases significantly with larger datasets, which can limit its scalability in real-time applications.

c. Artificial Neural Networks (ANN)

ANN, especially Deep Learning models, have been applied for time-series analysis in air quality prediction. (Dobrea et al., 2020) used ANN models to forecast air pollutant concentrations, reporting high accuracy but also noting the increased training time and risk of overfitting, particularly with small datasets. (Sharma et al., 2021) also employed ANN for predicting air pollutant levels, emphasizing its capability to model non-linear relationships in pollution data.

d. Decision Trees

Decision Tree algorithms, including Gradient Boosting and its variants, have shown promise for air quality prediction. Decision Trees are easy to interpret and can handle categorical data effectively. (Jamal & Nodehi, 2017) demonstrated the use of Decision Trees for meteorological data analysis and air quality forecasting, highlighting their useful application in real-time decision support systems.

2.3 Fuzzy Time Series and Hybrid Models

Other research approaches have included the use of hybrid and fuzzy models with the view of enhancing the accuracy of the machine learning. Currently, (Alyousifi et al., 2020) developed a model known as Fuzzy Time Series Markov Chain that integrates the fuzzy logic system with probabilistic Markov models for daily air pollution index prediction. This method was also found to reduce significant uncertainty for air quality data better than other conventional techniques as it can offer glimpses into the changes in temporal trends in pollution.

2.4 Big Data and IoT-Based Air Quality Prediction

The key development in the field of air quality prediction is the usage of Internet of Things (IoT) and Big Data technologies. Various IoT sensors spread across the cities can capture current pollution levels, and such data can be analyzed using Machine Learning techniques to predict the current pollution levels. (Ameer et al., 2019) presented a systematic review on machine learning models for air quality estimation in smart urban environments and implemented Apache Spark for distributed computing. Their work also stressed the role of time to process in huge-scale AQM, Random Forest, and Gradient Boosting turned out to be dominant in regards to accuracy and computational time.

Similarly, the study by (Yu et al., 2016) identified that IoT-based urban sensing systems are helpful in predicting air quality. Random forest used by the authors for creating their model proved effective along with the ability of their algorithms in conceiving pollution level their IoT based system could analyze data from multiple sensors demonstrating the capability of IoT for air quality monitoring in real-time.

2.5 Comparative Studies on Machine Learning Models for Air Quality Prediction

Some comparative analysis has been made in order to compare the efficacy of several types of the machine learning algorithms in solving the problem of air quality prediction. For example, (Agarwal, 2015) have compared the accuracy of multiple classifiers as ANN, SVM, Decision Trees; thereby, concluding that although ANN had the highest accuracy, it consumed more energy. Similarly, (Sharma et al., 2021)employed a performance comparison of different machine learning approaches and found that Random Forest was again most accurate and less sensitive to fluctuation.

Ameer et al. (Ameer et al., 2019) proposed a comprehensive analysis of the regression techniques such as Decision Trees, Random Forest, and Gradient Boosting that are employed with different data sets of

smart cities. They concluded that the two best algorithms of the two groups, Random Forest and Gradient Boosting were the best, but it was recommended that other issues like time and size of the data set be considered first while choosing a model.

2.5 Challenges and Future Directions

However, some issues still persist with the machines learning models used in the prediction of air quality. Such issues as how to manage large swathes of missing data, how to be able to process large chunks of data in real time, and how to fuse data from different and dissimilar sources such as weather conditions and traffic flows. Furthermore, issues such as the time taken to train the models like ANN and SVM in big datasets are costly in large applications (Dobrea et al., 2020); (Sharma et al., 2021).

Future studies should therefore aim at identifying better associated algorithms that allow fast and real time processing of massive amount of dataset generated from IoT sensors. At the same time, communication of hybrid models, which use machine learning in combination with prior knowledge from the atmosphere chemistry models, could provide higher accuracy and stabilities in the air quality prediction compared to the application of pure machine learning methods (Alyousifi et al., 2020).

3. METHODOLOGY

In the following part of the paper, the methodology used to compare various machine learning algorithms for measuring air quality is presented. Thus the evaluation system can be divided into several steps, data preprocessing, feature selection, model building, model's performance analysis and comparison with other models. Each step is detailed below.

3.1 Research Design

This study uses a quantitative approach to build a prediction model of PM2.5 concentration in Jakarta based on historical data available at the AQICN.org site from 2015 to 2024. The Random Forest algorithm was chosen as the main model because of its ability to handle complex non-linear relationships between variables and because of its robustness against overfitting. This study aims to understand the temporal pattern of PM2.5 concentration in Jakarta and produce accurate predictions that can support air pollution risk mitigation in the region.

3.2 Data Source

The PM2.5 concentration data used in this study was obtained from AQICN.org, a platform that provides real-time-based air quality data as well as historical data from various air monitoring stations in Jakarta. The data used includes PM2.5 concentrations recorded hourly during the period 2015 to 2024. This data was chosen to illustrate the daily fluctuations and temporal trends of PM2.5 in Jakarta, focusing on patterns that occur throughout the year.

3.3 Data Collection

The PM2.5 concentration data used in this study was obtained from AQICN.org for the period 2015 to 2024. The observation time range used includes hour-time data, allowing for more accurate modeling based on PM2.5 fluctuations that occurred during the period. This data was chosen for its ability to reflect daily and seasonal variability in Jakarta.

3.4 Data Preprocessing

Before permanently storing the measured PM2.5 data in the SQL database at this stage, the collected PM2.5 data is initially analyzed for unwanted empty entries and excessive outliers. In case of gaps, linear interpolation approaches are used, and for the cases where anomalies or outliers are observed, they are detected and deleted from data using applicable methods. This results from adding new temporal features like year, and month of the data, day of the year, days of the month to help identify seasonal variations that may be influential in PM2.5 concentration. Due to the fact that the data is presented in a yearly format with no further divisions as morning, afternoon or night, the differences and patterns are more apparent in yearly, monthly, and seasonal contexts.

3.5 Predictive Model Development

The used algorithm for building a prediction model which is the basis for decision making was the Random Forest Regressor that can deal with non-linear correlations and distinguish intricate patters. The dataset is divided into two parts: Taking for example 70% of the data will be used in training the model while 30% will be used to test the model. This is followed by dividing the data in order to be able to predict PM2.5 using the generated model should be able to generalize data that is not in the training data set. To improve the accuracy of the model, hyperparameter tuning is performed using Grid Search to find the best

combination of hyperparameters, including the number of trees (n_estimators), the maximum depth of the trees (max_depth), and the minimum sample size for node division (min_samples_split).

3.6 Model Variations

The model is evaluated using several metrics to measure the accuracy and accuracy of the model's predictions against the actual data. The metrics used include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 Score. These metrics are used to assess how well the model is at capturing fluctuations that occur in PM2.5 data and to measure the model's ability to predict data that is not present in the training dataset.

3.7 Result Visualization and Analysis

The results of the model for PM2.5 predictions are shown as temporal graph patterns representing daily, monthly, and annual changes of the concentration values. These graphs will be used to determine specific months of the year that have a higher PM2.5 concentration as well as to make observation on seasonal variation.

The analysis of the results each model were analyzed to ensure that it well suits the actual situation that takes place in Jakarta and to determine specific time that PM2.5 is likely to be higher within the day so as to guide the formulation of air pollution control measures.

4. RESULT AND DISCUSSION

4.1 Descriptive Statistics of PM2.5 Data

Based on PM2.5 concentration data from 2015 to 2024, descriptive statistics and average graphs and distribution of PM2.5 concentrations are obtained as follows:

In order to improve the efficiency of AWS maintenance, the primary goal of this article is to build predictive maintenance for AWS based on identifying anomalies utilizing artificial intelligence autoencoders that can speed up the predictive maintenance flow. To be able to design the desired predictive maintenance, the machine learning autoencoder is used. Anomaly detection is used to predict AWS before failure. This research will be designed in Google collab, which is easy using applications and simple features. The main objective of this study is to create predictive maintenance for Automatic Weather Station (AWS) to avoid the failure of instrumentation.

Table 1. Descriptive Statistics of PM2.5 Data

Table 1. Descriptive Statistics of PM2.5 Data

No	*Statistic*	*Date*	*PM2.5 (µg/m³)*
1	Total	3093	3093
2	Average	-	94,46
3	Minimum Values	2015-12-25	7
4	1st Quartile	2018-03-20	72
5	Median (Q2)	2020-05-28	97
6	3rd Quartile	2022-08-31	117
7	Maximum Value	2024-12-13	209
8	Standard Deviation	-	30,17

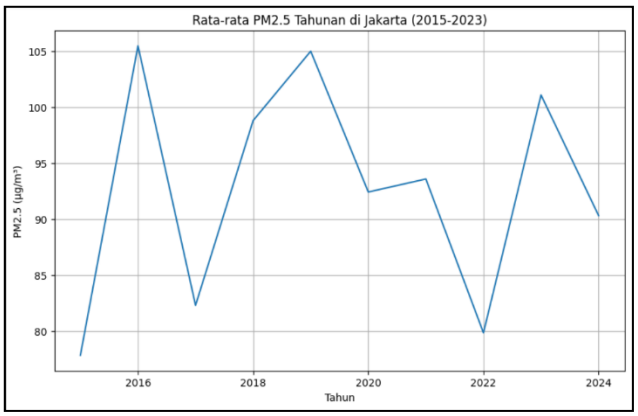

Fig.1. Average Annual PM2.5 in Jakarta (2015-2023)

This statistic shows that PM2.5 concentration has an average of 94.46 μg/m³ with a significant variation, as seen from the standard deviation of 30.17 μg/m³. The maximum value of PM2.5 concentration reached 209 μg/m³, which far exceeded the threshold of healthy air quality.

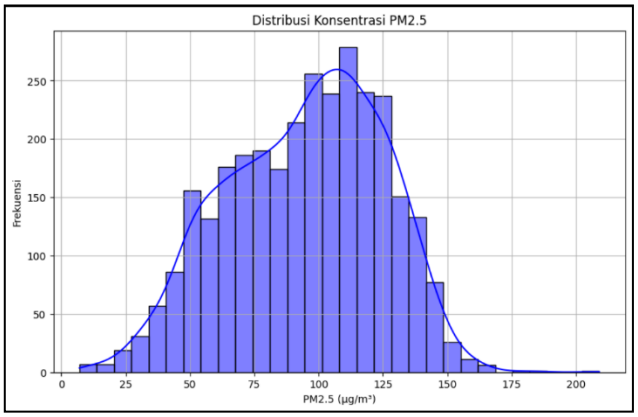

Fig.2. PM2.5 Concentration Distribution

The histogram plot depicts the distribution of PM2.5 concentration levels in the Jakarta region. The distribution exhibits a normal curve-like shape, with a peak around 140-160 μg/m³. This suggests that PM2.5 concentrations in Jakarta tend to be concentrated within this range. Several key observations can be made from the graph:

- Concentration Range: The graph covers a PM2.5 concentration range from approximately 50 μg/m³ to 200 μg/m³, encompassing the observed values in the region.
- Distribution Peak: The distribution peak occurs around 140-160 μg/m³, indicating that this is the most common PM2.5 concentration level observed.
- Distribution Shape: The distribution curve exhibits a relatively symmetric shape, following a normal distribution pattern. This implies that PM2.5 concentrations are evenly distributed across the observed range.
- Frequency: The graph shows the frequency of occurrence for each PM2.5 concentration range. Taller bars correspond to higher frequencies of the respective concentration levels.

The analysis from this study can therefore assist in refining the identification of the nature and spatial pattern of PM2.5 density in the Jakarta region. In extension, the data can be used as reference in the evaluation of the air quality in the said locale as well as to the proper measures that should be implemented.

4.2 Evaluation of Analysis Models

The results of the performance evaluation of the PM2.5 concentration prediction model using Random Forest show the following metrics:

Table 2. Results of Random Forest Model Evaluation

No	*Metrix*	*Value*
1	Mean Absolute Error (MAE)	14.44479525862069
2	Root Mean Square Error (RMSE)	18.746648967790264
3	R^2 Score	0.6082636701026527

This metric shows that the model performs quite well with an R^2 Score of 0.61, which means that the model is able to account for about 61% of the variation in the actual data. However, a fairly high RMSE (18.75) indicates a significant difference between the predicted value and the actual data at some point.

4.3 Comparation of Model Prediction with Actual Data of PM2.5 Concentration

The objective of this research is to create a model for analyzing the dispersion of PM2.5 concentration in the Jakarta area based on Random Forest modeling. For this purpose, we first analyze whether the model

correctly estimates the observed values of PM2.5 concentration at different time intervals. As indicated in the following graph, we have the results of the research.

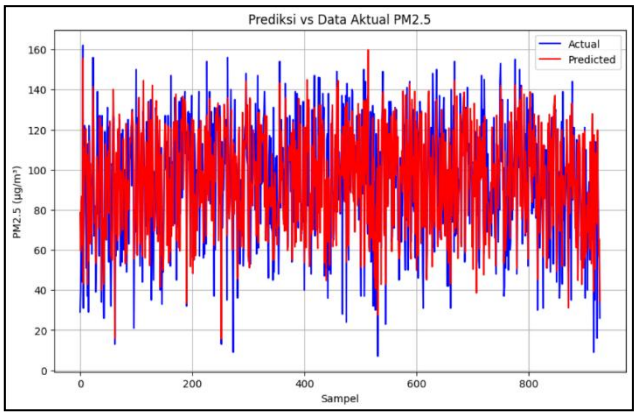

Fig.3. Prediction vs Actual Data PM2.5

In general, there is always a tight correspondence between the changes predicted by the prediction models and the actual change data even if they do not have to be identical. A few observations I made from this chart are the following:

- Fluctuation Pattern Suitability: The adjustment of the prediction model can reflect the approximate fluctuation of PM2.5 concentration, confirming the model's performance in perceiving changing trends.
- Differences in Prediction Values: Thus, the picture of simulation results indicates that, while the patterns of fluctuations are practically identical, there are certain differences in some points of the model's prediction indicators and real data. This implies that it is still possible even to enhance the predictive ability of the models and grow the accuracy of the results.
- Overall Performance: Overall, the model was successful in estimating PM2.5 concentrations but still requires enhancements to minimize the difference to actual data.

4.4 Improved Model Accuracy

To improve the accuracy of the model, hyperparameter tuning is performed using the Grid Search method, which involves the following parameters:

- n_estimators: Number of trees in the model.
- max_depth: Maximum depth of the tree.
- min_samples_split: Minimum sample size for node separation.

This tuning succeeded in reducing the error value of the model, but there are still limitations due to the limitations of the data period obtained and the type of data that other types of data are used as inputs. Further evaluation of long-term trends and the addition of variations in data types such as temperature and humidity can provide a better understanding of model performance and factors affecting PM2.5 concentrations.

5. CONCLUSION

This study aims to predict PM2.5 concentrations in the Jakarta area using the Random Forest model. The results show that this model has a fairly good performance with a Mean Absolute Error (MAE) of 14.44, a Root Mean Square Error (RMSE) of 18.75, and an R^2 Score of 0.61. The R^2 Score value shows that the model is able to explain about 61% of the actual PM2.5 data variation, although there are still significant differences at some prediction points indicated by the high RMSE value. In general, the model successfully follows a pattern of fluctuations in PM2.5 concentrations, although it is not yet completely accurate.

Based on descriptive statistical analysis, PM2.5 concentration data from 2015 to 2024 has an average of 94.46 µg/m³ with significant variations, as seen from the standard deviation of 30.17 µg/m³. The maximum value reaches 209 µg/m³, which is well above the threshold of healthy air quality. Efforts to improve the accuracy of the model have been made through tuning hyperparameters using Grid Search, which helps improve prediction performance. However, limitations still exist because comparisons are not made with data that covers longer periods.

This prediction model has the potential to be implemented in various scenarios, such as real-time air quality monitoring, supporting data-based policymaking for air pollution mitigation, and developing public applications that can provide air quality prediction information to the public. For future studies, it is recommended to use data with a longer period, consider other factors such as meteorological parameters

(temperature, humidity, and wind speed) as well as local pollution emission sources, and compare the performance of this model with other prediction models to obtain more optimal results.

Thus, this research is expected to contribute to air quality monitoring efforts and become the basis for formulating more effective policies to reduce the impact of air pollution in the Jakarta area.

REFERENCE

[1] Agarwal, Amit. (2015). Proceedings on 2015 1st International Conference on Next Generation Computing Technologies (NGCT) : September 4th-5th, 2015, Center for Information Technology, University of Petroleum and Energy Studies, Dehradun. IEEE.

[2] Alyousifi, Y., Othman, M., Sokkalingam, R., Faye, I., & Silva, P. C. L. (2020). Predicting daily air pollution index based on fuzzy time series markov chain model. Symmetry, 12(2). https://doi.org/10.3390/sym12020293

[3] Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities. IEEE Access, 7, 128325–128338. https://doi.org/10.1109/ACCESS.2019.2925082

[4] Brokamp, C., Jandarov, R., Rao, M. B., LeMasters, G., & Ryan, P. (2017). Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmospheric Environment, 151, 1–11. https://doi.org/10.1016/j.atmosenv.2016.11.066

[5] Chen, M. H., Chen, Y. C., Chou, T. Y., & Ning, F. S. (2023). PM2.5 Concentration Prediction Model: A CNN–RF Ensemble Framework. International Journal of Environmental Research and Public Health, 20(5). https://doi.org/10.3390/ijerph20054077

[6] Dobrea, M., Badicu, A., Barbu, M., Subea, O., Balanescu, M., Suciu, G., Birdici, A., Orza, O., & Dobre, C. (2020). Machine Learning algorithms for air pollutants forecasting. 2020 IEEE 26th International Symposium for Design and Technology in Electronic Packaging, SIITME 2020 - Conference Proceedings, 109–113. https://doi.org/10.1109/SIITME50350.2020.9292238

[7] Goyal, K., & Goyal, S. (2024). Predicting PM2.5 Air Quality Using Random Forest Regression Enhanced with Polynomial Features. International IEEE Conference Proceedings, IS, 2024. https://doi.org/10.1109/IS61756.2024.10705219

[8] Grell, G. A., Peckham, S. E., Schmitz, R., McKeen, S. A., Frost, G., Skamarock, W. C., & Eder, B. (2005). Fully coupled “online” chemistry within the WRF model. Atmospheric Environment, 39(37), 6957–6975. https://doi.org/10.1016/j.atmosenv.2005.04.027

[9] Jamal, A., & Nodehi, R. N. (2017). A R T I C L E I N F O R M A T I O N PREDICTING AIR QUALITY INDEX BASED ON METEOROLOGI-CAL DATA: A COMPARISON OF REGRESSION ANALYSIS, ARTI-FICIAL NEURAL NETWORKS AND DECISION TREE. In Journal of Air Pollution and Health (Vol. 2, Issue 1). http://japh.tums.ac.ir

[10] Joharestani, M. Z., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere, 10(7). https://doi.org/10.3390/atmos10070373

[11] Kang, J., Zou, X., Tan, J., Li, J., & Karimian, H. (2023). Short-Term PM2.5 Concentration Changes Prediction: A Comparison of Meteorological and Historical Data. Sustainability (Switzerland), 15(14). https://doi.org/10.3390/su151411408

[12] Liu, Y., Wang, Y., & Zhang, J. (2012). New Machine Learning Algorithm: Random Forest. In LNCS (Vol. 7473).

[13] Livingston, F. (2005). Implementation of Breiman’s Random Forest Machine Learning Algorithm. In ECE591Q Machine Learning Journal Paper. Fall.

[14] Lu’, W., Wan&, W., Tileungl, A. Y., Lo’, S.-M., Kyued, R. K., Xu1, Z., & Fan’, H. (2002). Air Pollutant Parameter Forecasting Using Support Vector Machines.

[15] Ma, X., Chen, T., Ge, R., Xv, F., Cui, C., & Li, J. (2023). Prediction of PM2.5 Concentration Using Spatiotemporal Data with Machine Learning Models. Atmosphere, 14(10). https://doi.org/10.3390/atmos14101517

[16] Mauboy, L. M., Raihan Abhirama, M., Salsabila, S., & Kurniawan, R. (2024). Perbandingan Klasifikasi PM2.5 di Daerah Khusus Jakarta Algoritma C5.0, Random Forest, dan SVM. Seminar Nasional Sains Data, 2024.

[17] Segal, M. R. (2003). UCSF Recent Work Title Machine Learning Benchmarks and Random Forest Regression Publication Date Machine Learning Benchmarks and Random Forest Regression.

[18] Sharma, M., Jain, S., Mittal, S., & Sheikh, T. H. (2021). Forecasting and prediction of air pollutants concentrates using machine learning techniques: The case of India. IOP Conference Series: Materials Science and Engineering, 1022(1). https://doi.org/10.1088/1757-899X/1022/1/012123

[19] Uzir, N., Raman, S., Banerjee, S., & Nishant Uzir Sunil R, R. S. (2016). Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets Experimenting XGBoost Algorithm for Prediction and Classifi cation of Different Datasets. International Journal of Control Theory and Applications, 9. https://www.researchgate.net/publication/318132203

[20] Yu, R., Yang, Y., Yang, L., Han, G., & Move, O. A. (2016). RAQ–A random forest approach for predicting air quality in urban sensing systems. Sensors (Switzerland), 16(1). https://doi.org/10.3390/s16010086