Applied Information System and Management (AISM)
Volume 7, (2) 2024, p. 1–6
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v7i2.38466

Analysis of the Impact of Meteorological Factors on Predicting Air Quality in South Tangerang City using Random Forest Method

Nurchaerani Kadir[1*], Muhammad Faisal[2], Fachrul Kurniawan[3]

***Abstract*—Air pollution has become one of the most significant environmental problems in many cities throughout the world, which can endanger public health and the environment. Understanding the impact of meteorological conditions on air quality is very important to understanding air pollution patterns. This study investigates the influence of meteorological variables on air quality predictions in South Tangerang City, Indonesia, using the Random Forest method. Modeling is carried out by building two scenarios, namely predictions using meteorological variables and predictions without meteorological variables. Prediction performance analysis is measured using MAE, MSE, RMSE, R-square, and accuracy. The accuracy results of the research show that predictions without meteorological variables provide good prediction results with a value of 86.42%, but predictions with meteorological variables have better performance with a value reaching 98.99%. The largest error values from each model were 2.58 MAE, 71.82 MSE, and 8.4747 RMSE obtained in prediction modeling without meteorological variables, while the smallest error values were obtained in prediction modeling using meteorological variables, namely 0.00, 0.01, and 0.0219, respectively, for MAE, MSE, and RMSE. This research contributes to a better understanding of the relationship between meteorology and air pollution and air quality in urban areas and helps develop targeted mitigation strategies to improve air quality and public health, especially in South Tangerang City and the surrounding area.**

***Index Terms*—Air quality, meteorological factors, random forest.**

I. Introduction

Air pollution has become one of the most significant environmental problems in many cities around the world [1]. The city of South Tangerang is one of the cities in Indonesia with the worst air pollution levels since 2023 [2]. Poor air quality is now a major threat to global societies, causing devastating effects on individuals, medical systems, ecosystem health, and economies in both developing and developed countries [3]. Increased transport activity [4], industrial growth, and rapid urbanisation have led to increased emissions of air pollutants, which can endanger public health and the environment [5], [6].

WHO estimates that air pollution is responsible for around 7 million premature deaths per year from ischemic heart disease, stroke, chronic obstructive pulmonary disease, and lung cancer, but also from acute respiratory infections such as pneumonia, which mainly affect children in countries with low and middle economic levels [7].

Climate change is exacerbating pollution and environmental damage due to the effects of widespread globalization [8]. Several factors, including local meteorological conditions, affect air quality [9]. Temperature, humidity, wind direction, wind speed, and precipitation are some of the meteorological factors that can affect the dispersion and transportation of air pollutants in the atmosphere [10], [11]. Meteorological variables such as temperature and wind speed often have the highest correlation with air pollution levels [12].

In addition, air pollutants such as particles (PM2.5 and PM10), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), and ozone (O_3) are essential for setting air quality standards [13]. PM 25 is one type of air pollutant that is a deadly threat because it can cause cardiovascular disease and aggravate asthma [14]. Currently, the ISPU (Air Pollution Standard Index) has become the official standard applied in Indonesia to assess air quality. There are seven key parameters involved in evaluating ISPU, including PM10, PM2.5, NO_2, SO_2, CO, O_3, and HC [15]. This level of pollution can vary significantly according to human activities, such as industrial activities, transportation, and energy use [16].

Reference [17] suggested that meteorological factors and air pollutants have a strong influence on air quality in cities. High temperatures and low wind speeds can inhibit the spread of pollution, but high temperatures and high wind speeds lead to increased particulate concentrations and decreased urban air quality [18].

Received: 18 April 2024; Revised: 25 April 2024; Accepted: 08 Mei 2024.
*Corresponding author
[1]Nurchaerani Kadir, Universitas Islam Negeri Maulana Malik Ibrahim Malang Indonesia, (e-mail: 220605210018@student.uin-malang.ac.id).
[2]Muhammad Faisal, Universitas Islam Negeri Maulana Malik Ibrahim Malang Indonesia, (e-mail: mfaisal@ti.uin-malang.ac.id).
[3]Fachrul Kurniawan, Universitas Islam Negeri Maulana Malik Ibrahim Malang Indonesia, (e-mail: fachrulk@ti.uin-malang.ac.id).

The use of machine learning algorithms has become the most popular tool and is considered to have succeeded in making predictions with strong and fast levels of prediction against big data [19]. One of the best-performing machine learning techniques in the field of air quality prediction is Random Forest [20], [21]. Random Forest is an interesting alternative to finding out how pollutants and meteorology affect the air quality index in Southern Tangerang City. Random forest is a powerful ensemble algorithm capable of making accurate predictions in complex conditions and managing various types of data [22], [23]. The study conducted in [24] by applying the random forest method to predicting air quality produced very accurate results and had better performance compared to artificial neural networks. Apart from that, [25] in his research applied several machine learning methods to predict air quality, including random forest, decision tree, and deep backpropagation neural network. It was found that the random forest method had the best performance in predicting air quality.

Random forest enables non-linear modelling of the relationship between predictor variables (meteorology and pollutants) and the target variable (air quality index). Therefore, we can find a complex relationship between these variables and analyze variables that have a significant influence on air quality. Thus, the aim of this study is to analyse the influence of meteorology on the air quality index of Southern Tangerang City using the Random Forest method. This research is expected to provide a better understanding of the air pollution patterns in the area.

This research is carried out because good air quality plays a key role in human health, the environment, and natural ecosystems. A better understanding of air quality is expected to provide understanding to the public, local governments, and stakeholders about efforts to manage air quality in Southern Tangerang City. Knowing how much meteorological and polluting factors influence the air quality index, it is expected that more efficient mitigation and management strategies can be developed and implemented. This will encourage the development of more efficient policies and sustainable solutions to ensure that the air in Tangerang City and its surrounding areas remains clean and healthy in the future.

II. Research Method

The research was conducted to analyse and find out to what extent meteorological variables affect air quality in South Tangerang City, one of the areas with the highest levels of air pollution in Indonesia. There are several phases involved in this research, such as data collection, preprocessing, data splitting, and model implementation. The following research stages can be seen in Fig. 1.

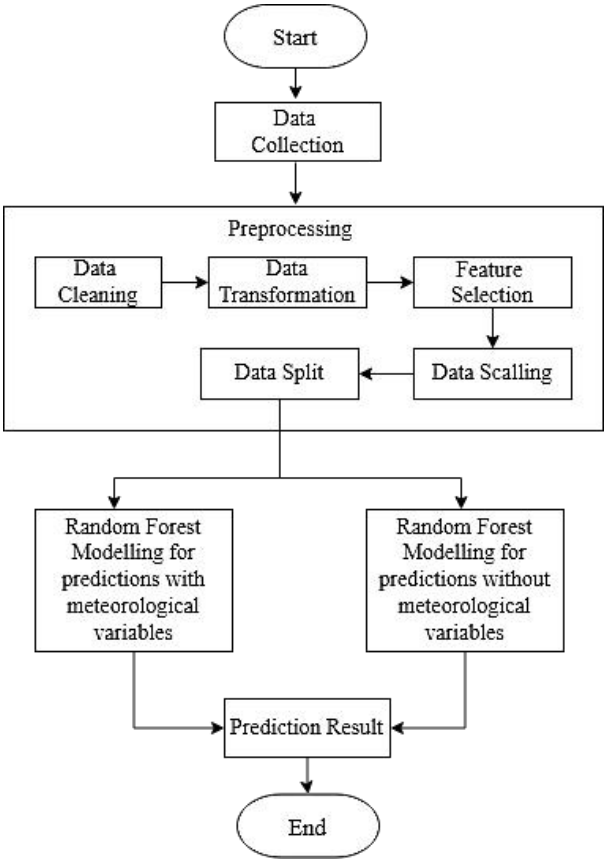

Fig. 1. Research method design.

A. Data Collection

The air pollution data for South Tangerang City used in this study was obtained from the official government website of the Ministry of the Environment (https://ispu.menlhk.go.id). Datasets are monitored every hour for 24 hours a day. As for the monitored variables, PM25, PM10, SO, NO2, O3, CO, and HC.

As for the meteorological data used, it was obtained from the old website (https://dataonline.bmkg.go.id/). The variables used for the weather data consist of temperature, wind direction, wind speed, and humidity. The following characteristics of the dataset can be seen in Fig. 2.

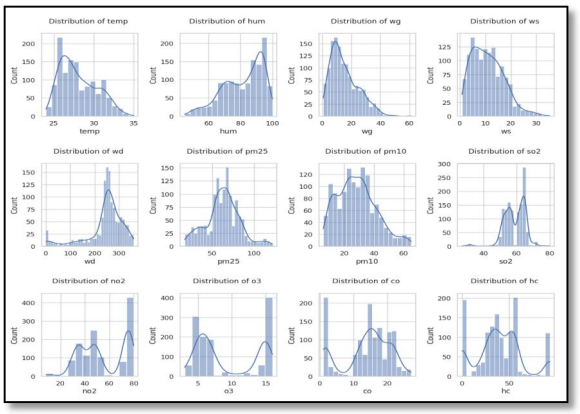

Fig. 2. Distribution of dataset.

Applied Information System and Management (AISM)
Volume 7, (2) 2024, p. 1–6
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v7i2.38466

B. *Preprocessing*

The preprocessing stage is a necessary step to convert data that is initially not optimal to a more qualitative format, so it can be used as an input in testing the proposed method and produce a higher level of accuracy [26]. Here are some steps that are taken in data preprocessing.

1) *Data Cleaning*
   Cleaning data is an important part of data preparation. Data cleaning is an initial step in preprocessing that involves deleting unwanted data, such as data that does not match the format in the data set. The purpose of the data cleaning process is to improve the accuracy of the prediction result.
2) *Data Transformation*
   At the transformation stage, the structure of all data variables is fixed, such as by changing the format of the data to suit the operation to be performed. At this stage, date and time variables are converted to the Datetime format to ensure compatibility with the analysis carried out.
3) *Feature Selection*
   Considering that this study is aimed at evaluating the influence of meteorological variables on air quality, the initial testing was conducted to test the prediction of air quality without involving meteorological input variables. Meanwhile, the subsequent testing involved both meteorological and pollutant input variables in its analysis. Therefore, in the initial test, a selection of the characteristic variables of pollutants included PM25, PM10, SO, NO_2, O_3, CO, and HC.
4) *Data Scaling*
   Before starting the training of the recommended model, the first step is to renormalize the dataset using normalization and standardization methods. The purpose of the standardization is to adjust the data scale to the normal range, i.e. between 0 and 1. To do this, the following equations will be used:

$$\boldsymbol{x_{norm}} = \frac{\boldsymbol{x - x_{min}}}{\boldsymbol{x_{max} - x_{min}}} \quad (1)$$

   where $\boldsymbol{x}$ is the original value of the dataset, and $\boldsymbol{x_{min}}$ and $\boldsymbol{x_{max}}$ are the minimum and maximum values of the data.
5) *Data Splitting*
   After performing data preprocessing, the next step is to divide the dataset into two parts: training data and testing data. Training data will be used to train the suggested model, while testing data will be used to test or evaluate the performance of the trained model.

C. *Implementation Random Forest*

At the Random Forest implementation stage, modelling was carried out to predict air quality in South Tangerang City. Modelling was carried out by building two scenarios, namely air quality predictions with meteorological variables and air quality predictions without meteorological variables.

The prediction model with meteorological variables uses temperature, wind direction, wind speed, humidity, PM25, PM10, SO, NO_2, O_3, CO, and HC as inputs for predictions. Meanwhile, the prediction model without meteorological variables uses PM25, PM10, SO, NO_2, O_3, CO, and HC as input.

Random forest is a machine learning algorithm that can handle regression and classification problems with a high degree of precision and a low probability of overfitting [27], [11]. This algorithm consists of many decision trees, each of which contributes to the prediction results. The advantage of random forest is its ability to handle lost data and prevent overfitting. In addition, this algorithm is extremely efficient in handling large datasets [11], [28]. The following illustration of Random Forest can be seen in Fig. 3.

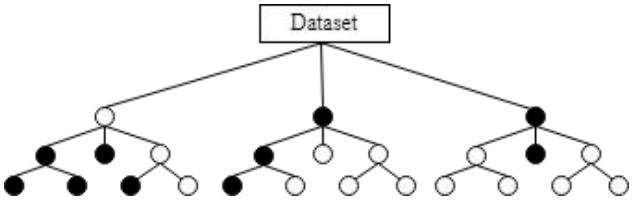

Fig. 3. Random forest ilustration.

III. RESULT

The study uses the Python programming language to use the Random Forest model to analyze predictions of air quality with and without meteorological variables. Here are the results of modeling air quality prediction in South Tangerang City using Random Forest.

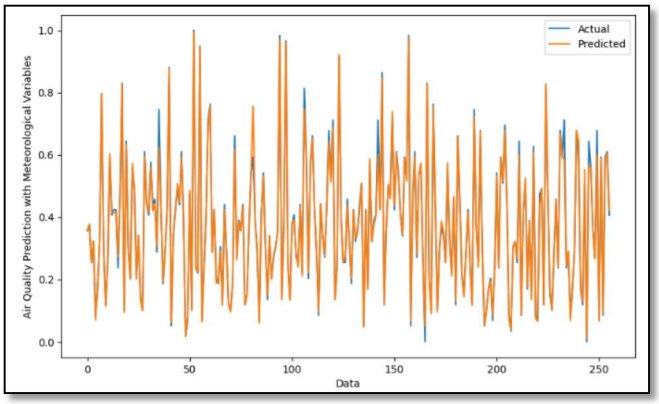

Fig. 4. Air quality prediction modeling with meteorological variables

Based on Fig. 4, it is seen that the data patterns resulting from the air quality prediction have a trend pattern up and down at certain hours.

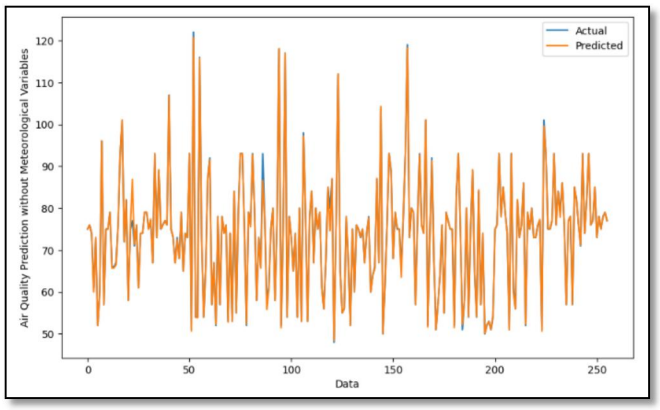

Fig. 5. Air quality prediction modeling without meteorological variables.

As seen in Fig. 5, the data patterns resulting from the air quality prediction form a trend pattern that goes up and down at certain hours. This is influenced by the high pollution during the working hours or the hours of return to work. It's because during those hours there's a lot of industrial activity, and there'll be lots of congestion that produces air-polluting particles.

The accuracy of the predictions obtained can be analyzed by considering the mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and determination coefficient [29]. The following is the equation of the evaluation metrics [16], [30].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_{observed} - x_{predicted}\right| \quad (2)$$

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(x_{observed} - x_{predicted}\right) \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{observed} - x_{predicted}\right)^2} \quad (4)$$

$$R^2 = 1 - \frac{\sum\left(x_{observed} - x_{predicted}\right)^2}{\sum\left(x_{observed} - x_{predicted}\right)^2} \quad (5)$$

The results of air quality predictions with meteorological variables and without meteorological variables are shown in Table 1. Prediction results are assessed based on MAE, MSE, and RMSE values. The dataset used is divided into two parts: 80% is used as training data to train the model, and the remaining 20% is used as test data for validation.

Table 1.
Random Forest Performance Results for Air Quality Prediction

Measurement Metrics	Prediction with Meteorological Variables		Prediction without Meteorological Variables	
	Training	Testing	Training	Testing
MAE	0.00	0.00	0.79	2.58
MSE	0.00	0.01	4.25	71.82
RMSE	0.0089	0.0219	2.0619	8.4747
R^2	0.9983	0.9899	0.9933	0.8642

Based on Table 1, the Random Forest prediction model has the highest correlation on prediction with the meteorological variable by reaching the respective values; training data is 0.9983 and testing data is 0.9899. While predictions without meteorology variables reach R-square, respectively, training data are 0.9933 and testing data are 0.8642.

For the error analysis of Random Forest methods, include MAE, MSE, and RMSE values. Testing data for predictions without meteorological variables has the highest error of the other tests, being 2.58 MAE, 71.82 MSE, and 8.4747 RMSE. Whereas for data training, we have MAE values, MSE, and RMSE of 0.79, 4.25, and 2.0619. As for the minimum error values obtained on training data for the prediction using meteorologic variables, they are 0.00, 0.00, and 0.0089, respectively, for MAE and MSE. For data testing, have values of 0.00, 0.01, and 0.9899.

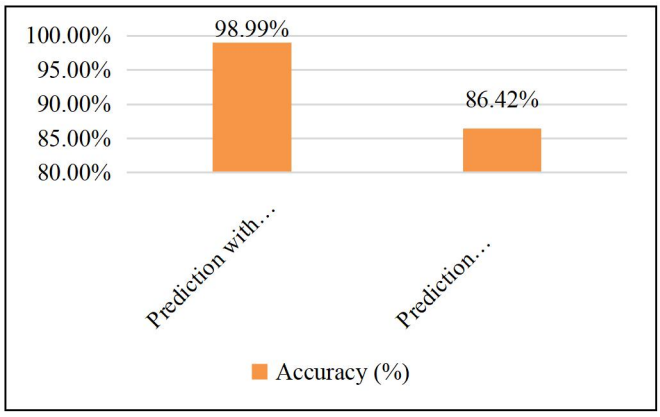

Fig. 6. Comparison of prediction accuracy results.

The accuracy analysis of the proposed method is as shown in Fig. 6. The accuracy results show that predictions with meteorological variables are better than predictions without meteorological variables. The accuracy achieved in testing using meteorological variables was 98.99%, while for testing without meteorological variables, the accuracy value was 86.42%.

It turns out that meteorological factors have a significant influence on the prediction of air quality. Based on the accuracy comparison results in Fig. 6, it proves that Random Forest's performance in predicting air quality with the involvement of weather factors has better performance than predictions without involving weather factors.

The analysis of the Random Forest method in predicting air quality shows that this method is very good for making predictions of air pollution data in South Tangerang City and other cities because it gives good results based on relatively small RMSE results and high accuracy results. The limitation of the amount of data used in the process of modeling air quality predictions affects the degree of accuracy of the prediction results.

Applied Information System and Management (AISM)
Volume 7, (2) 2024, p. 1–6
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v7i2.38466

IV. DISCUSSION

This research aims to analyze the impact of meteorological factors on air quality predictions in South Tangerang City using the Random Forest method. This study found that by including meteorological variables such as temperature, humidity, wind speed, and wind direction, the accuracy of air quality predictions improved significantly compared to models that only used air pollutant concentration data.

The research results show that the random forest model with meteorological variables has the highest accuracy in predicting air quality in South Tangerang City. The accuracy of this model reached 98.99%, much higher than the model that only used air pollutant data (86.42%). This confirms that meteorological factors have a significant influence on air pollutant concentrations and should be considered in air quality prediction models [11]. In addition, there are studies that show that meteorological conditions such as high temperatures, low wind speeds, and low humidity tend to cause increased concentrations of air pollutants [9], [10]. In the study [12], it is said that air temperature and wind speed often show a very close relationship with the level of pollutant concentration in the air.

The advantage of using the Random Forest method in this research is its ability to accurately identify important variables that influence air quality. By including meteorological factors, the model can provide more precise and reliable predictions compared to models that only use air pollutant data [11].

V. CONCLUSION

In this study, the air quality prediction in South Tangerang City, Indonesia was analyzed using daily air pollutant and meteorological data monitored every hour from a website provided by the Indonesian government. The study applied the Random Forest method to model air quality in Tangerang City, Indonesia. The predictive performance analysis is measured using MAE, MSE, RMSE, R-square, and accuracy. The study's accuracy results show that predictions with meteorological variables give a better prediction result with a value of 98.99% compared to predictions without meteorological variables, which have a performance value of 86.42%. As for the largest error values of each model, 2.58 MAE, 71.82 MSE, and 8.4747 RMSE are obtained in the modeling of the prediction without the meteorological variable, whereas the smallest error value is obtained in the pre-modeling using the meteorology variables of 0.00, 0.01, and 0.0219 for MAE, MSE, and RMSE, respectively. In further research, it is recommended to conduct research in urban areas with a high density of population or areas with high industrial activity that may affect future air quality.

REFERENCES

[1] S. Calo, F. Bistaffa, A. Jonsson, V. Gómez, and M. Viana, "Spatial air quality prediction in urban areas via message passing," *Eng. Appl. Artif. Intell.*, vol. 133, part B, Jul. 2024, Art. no. 108191, doi: 10.1016/j.engappai.2024.108191.

[2] C. M. Annur, "10 Cities with the Worst Air Pollution in Indonesia 2023," Databoks.katadata.co.id. https://databoks.katadata.co.id/datapublish/2024/01/17/10-kota-dengan-polusi-udara-terburuk-2023-tangsel-teratas (accessed May 5, 2024).

[3] L. Liang and P. Gong, "Urban and air pollution: a multi-city study of long-term effects of urban landscape patterns on air quality trends," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, 2020, doi: 10.1038/s41598-020-74524-9.

[4] C. Lee, "Impacts of urban form on air quality in metropolitan areas in the United States," *Comput. Environ. Urban Syst.*, vol. 77, 2019, Art. no. 101362, doi: 10.1016/j.compenvurbsys.2019.101362.

[5] S. Gunasekar, G. Joselin Retna Kumar, and G. Pius Agbulu, "Air quality predictions in urban areas using hybrid ARIMA and metaheuristic LSTM," *Comput. Syst. Sci. Eng.*, vol. 43, no. 3, pp. 1271–1284, 2022, doi: 10.32604/csse.2022.024303.

[6] J. Persis and A. Ben Amar, "Predictive modeling and analysis of air quality – Visualizing before and during COVID-19 scenarios," *J. Environ. Manage.*, vol. 327, Feb. 2023, Art. no. 116911, doi: 10.1016/j.jenvman.2022.116911.

[7] WHO, "Air pollution: The invisible health threat," Who.int. https://www.who.int/news-room/feature-stories/detail/air-pollution--the-invisible-health-threat (accessed May 5, 2024)..

[8] E. X. Neo et al., "Towards integrated air pollution monitoring and health impact assessment using federated learning: A systematic review," *Front. Public Heal.*, vol. 10, no. May, pp. 1–19, 2022, doi: 10.3389/fpubh.2022.851553.

[9] R. Janarthanan, P. Partheeban, K. Somasundaram, and P. Navin Elamparithi, "A deep learning approach for prediction of air quality index in a metropolitan city," *Sustain. Cities Soc.*, vol. 67, Apr. 2021, Art. no. 102720, doi: 10.1016/j.scs.2021.102720.

[10] S. K. Grange and D. C. Carslaw, "Using meteorological normalisation to detect interventions in air quality time series," *Sci. Total Environ.*, vol. 653, pp. 578–588, 2019, doi: 10.1016/j.scitotenv.2018.10.344.

[11] Y. Liu, P. Wang, Y. Li, L. Wen, and X. Deng, "Air quality prediction models based on meteorological factors and real-time data of industrial waste gas," *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, 2022, doi: 10.1038/s41598-022-13579-2.

[12] R. T. McNider and A. Pour-Biazar, "Meteorological modeling relevant to mesoscale and regional air quality applications: a review," *J. Air Waste Manag. Assoc.*, vol. 70, no. 1, pp. 2–43, 2020, doi: 10.1080/10962247.2019.1694602.

[13] J. Zhang and S. Li, "Air quality index forecast in beijing based on CNN-LSTM multi-model," *Chemosphere*, vol. 308, Dec. 2022, Art. no. 136180, doi: https://doi.org/10.1016/j.chemosphere.2022.136180.

[14] R. Murugan and N. Palanichamy, "Smart city air quality prediction using machine learning," *Proc. - 5th Int. Conf. Intell. Comput. Control Syst. ICICCS 2021*, pp. 1048–1054, 2021, doi: 10.1109/ICICCS51141.2021.9432074.

[15] M. Kusnandar, "Republic of Indonesia Government Regulations," Regul. Minist. Environ. For. Repub. Indones. Number 14 2020 Concern. Air Pollut. Stand. Index, pp. 1–16, 2020.

[16] D. Kothandaraman et al., "Intelligent forecasting of air quality and pollution prediction using machine learning," *Adsorpt. Sci. Technol.*, vol. 2022, pp. 1-15, 2022, doi: 10.1155/2022/5086622.

[17] K. Gu, J. Qiao, and W. Lin, "Recurrent air quality predictor based on meteorology- and pollution-related factors," *IEEE Trans. Ind. Informatics*, vol. 14, no. 9, pp. 3946–3955, 2018, doi: 10.1109/TII.2018.2793950.

[18] T. Wang et al., "Prediction of the impact of meteorological conditions on air quality during the 2022 beijing winter olympics," *Sustainability*, vol. 14, no. 8, pp. 1-13, Apr. 2022, doi: 10.3390/su14084574.

[19] H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," *Appl. Sci.*, vol. 9, no. 19, 2019, Art no. 4069, doi: 10.3390/app9194069.

[20] A. Pant, S. Sharma, and K. Pant, "Evaluation of machine learning algorithms for air quality index (AQI) prediction," *J. Reliab. Stat. Stud.*, vol. 16, no. 2, pp. 229–242, 2023, doi: 10.13052/jrss0974-8024.1621.

[21] A. Akanksha, N. Maurya, M. Jain, and S. Arya, "Prediction and analysis of air pollution using machine learning algorithms," *3rd Int. Conf. Intell. Technol. CONIT*, pp. 1–6, 2023, doi: 10.1109/CONIT59222.2023.10205615.

[22] M. V. V. S. Subrahmanyam, P. V. V. S. D. Nagendrudu, and T. V Ramana, "Comparison of effective machine learning technique for air quality forecast BT," *Proceedings of the 2nd International Conference on Cognitive and Intelligent Computing*, pp. 157–164, 2023.

[23] K. B. Soni, "Credit card fraud detection using machine learning approach," *Appl. Inf. Syst. Manag.*, vol. 4, no. 2, pp. 71–76, 2021, doi: 10.15408/aism.v4i2.20570.

[24] H. Altinçöp and A. B. Oktay, "Air pollution forecasting with random forest time series analysis," *2018 Int. Conf. Artif. Intell. Data Process. IDAP 2018*, pp. 8–12, 2019, doi: 10.1109/IDAP.2018.8620768.

[25] S. Li, X. Deng, and B. Tang, "Using machine learning methods for prediction of air quality in wuling mountain area in china," in *2021 International Conference on Electronic Information Technology and Smart Agriculture (ICEITSA)*, 2021, pp. 426–430, doi: 10.1109/ICEITSA54226.2021.00087.

[26] N. N. Maltare and S. Vahora, "Air quality index prediction using machine learning for ahmedabad city," *Digit. Chem. Eng.*, vol. 7, 2023, doi: 10.1016/j.dche.2023.100093.

[27] M. Mihirani, L. Yasakethu, and S. Balasooriya, "Machine learning-based air pollution prediction model," *IEEE IAS Glob. Conf. Emerg. Technol. GlobConET*, no. 2, pp. 1–6, 2023, doi: 10.1109/GlobConET56651.2023.10150203.

[28] A. Choudhary et al., "Evaluating air quality and criteria pollutants prediction disparities by data mining along a stretch of urban-rural agglomeration includes coal-mine belts and thermal power plants," *Front. Environ. Sci.*, vol. 11, no. 2, pp. 1-22, Nov. 2023, doi: 10.3389/fenvs.2023.1132159.

[29] E. Gladkova and L. Saychenko, "Applying machine learning techniques in air quality prediction," *Transp. Res. Procedia*, vol. 63, pp. 1999–2006, 2022, doi: 10.1016/j.trpro.2022.06.222.

[30] M. Madhuri, G. H. Samyama Gunjal, and S. Kamalapurkar, "Air pollution prediction using machine learning supervised learning approach," *Int. J. Sci. Technol. Res.*, vol. 9, no. 4, pp. 118–123, 2020.