BAREKENG: Journal of Mathematics and Its Applications
March 2026 Volume 20 Issue 1 Page 0785–0796
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss1pp0785-0796

COMPARISON OF XGBOOST AND RANDOM FOREST METHODS IN PREDICTING AIR POLLUTION LEVELS

Akas Yekti Pulih Asih [1], Firman Yudianto [2], Puguh Triwinanto [3], Rachman Sinatriya Marjianto [4*], Teguh Herlambang [5], Hamzah Arof [6]

[1] *Department of Public Health, Faculty of Health, Universitas Nahdlatul Ulama Surabaya*
[2,5] *Department of Information Systems, Universitas Nahdlatul Ulama Surabaya*
Jln. Jemursari No 51-57, Surabaya, 60237, Indonesia

[3] *National Research and Innovation Agency (BRIN)*
Jln. MH. Thamrin No. 8, BJ. Habibie Tower, Jakarta, 10340, Indonesia

[4] *Department of Engineering, Faculty of Vocational, Universitas Airlangga*
Jln. Dharmawangsa Dalam Selatan No 28-30 Surabaya, 60286, Indonesia

[6] *Department of Electrical Engineering, University of Malaya*
Kuala Lumpur, 50603, Malaysia

*Corresponding author's e-mail: * **rachmansinatriya@vokasi.unair.ac.id***

Article Info

Article History:

Received: 28th May 2025
Revised: 4th July 2025
Accepted: 4th August 2025
Available online: 24th November 2025

Keywords:

Pollution;
Prediction;
Random forest;
XGBoost.

ABSTRACT

Air is one of the elements needed by living things, including humans, to survive. The air quality in an area also affects the health and quality of human life and its surrounding environment. However, with the current phenomenon, the influence of the increasing number and mobility of humans actually degrades air quality, caused by the pollutants produced. For further impacts, poor air quality can reduce human life expectancy. Big cities in Indonesia, such as Surabaya, also experience the same thing due to the lack of public awareness of air pollution. The biggest contributors to air pollution are motor vehicles and industrial activities that emit carbon monoxide (CO), nitrogen oxides (NO), ozone (O3), and other particles (PM10). This condition is addressed by the Surabaya City Government by installing air condition measuring devices at points considered prone to pollution. This device works to measure urban air conditions daily and provides data that can be utilized to establish strategic policies. By utilizing the data, in this research, we implemented two prediction methods from machine learning technology, namely XG Boost and Random Forest. In accordance with the objective of this research, both methods will be compared for accuracy in predicting air pollution levels in Surabaya based on Ozon (O3) substance within the period of January 1, 2020, to December 31, 2020. Both of them have a similarity in that they implement tree-ensemble based, which are appropriate for handling non-linear data. The XG Boost method managed to achieve the best error value of 0.0510, and the Random Forest method reached the best error value of 0.0468.

How to cite this article:

A. Y. P. Asih, F. Yudianto, P. Triwinanto, R. S. Marjianto, T. Herlambang, and H. Arof, "COMPARISON OF XGBOOST AND RANDOM FOREST METHODS IN PREDICTING AIR POLLUTION LEVELS", *BAREKENG: J. Math. & App.*, vol. 20, no. 1, pp. 0785-0796, Mar, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Air is an important element for living things and is a reference parameter for cleanliness and quality of life in a city. The air that humans need is air containing oxygen (O2). The availability of clean air is very important and is needed as an indicator of air quality measurement in a city. Clean and quality air can have a positive impact on people's lives in the city. With the increasing number and level of people's mobility, especially in big cities, it also affects the quality and condition of the air inside. Three main problems in metropolitan cities that are unfinished issues are population density, many motorized vehicles, and unplanned industrial development [1]. Surabaya, the capital city of East Java, is known as the second most populous metropolitan city in Indonesia after Greater Jakarta [2]. Air quality is a major concern, particularly in urban areas where traffic is very intense [3]. All human activities that produce exhaust emissions and other types of pollutants have the potential to decline the quality of the air (environmental degradation) if not balanced with a properly supported environment.

The impact of this condition is certainly not limited only to numbers or statistics, but also has a direct impact on the health of the population. Air pollution is one of the most significant environmental challenges of our time [4]. Primary air pollutants are represented by oxides of nitrogen, carbon monoxide (CO), sulfur dioxide (SO_2), volatile organic compounds (VOCs), and carbonaceous and non-carbonaceous primary particles [5]. Indonesia has recorded the highest number of premature deaths (over 50,000) associated with air pollution among countries in Southeast Asia [6]. For the other example, as impact of air pollution on transportation includes traffic congestion and air flight disruption [7].

Air pollution models have played a pivotal role in furthering scientific understanding and supporting policy [8]. Several policy changes could help reduce the deleterious components of the exposome and minimize their effect on respiratory health [9]. The application of air pollution control in the regions refers to the Regulation of the State Minister of the Environment is in accordance with Number 12 of 2010 [10], and also refers to the Regional Regulation of Surabaya City Government Number 3 of 2008 about regulations and management of air pollution, states how emissions from moving sources, non-moving sources, and other sources of interference. Various efforts have been made by the Surabaya City Government to reduce the figure of the air pollution rate. Some of them are the realization of a cooperation agreement with Kitakyushu city in Japan [11], and installing measuring devices that have a high concentration of activities. All these efforts are made to move towards a smart city. A smart city is a city that implements technology to handle multiple fields in an integrated and sustainable way [12].

Various advances in technology have assisted human life. From the invention of modern tools, which can be used to filter air, as used in China. An air filter with nanofiber membrane technology works by eliminating particulate matter, harmful gases, and other air pollutants from the air that is inhaled by residents [13]. Also, information technology in the form of Artificial Intelligence (AI) really helps the Surabaya city government in making decisions.

Machine Learning (ML), as a subset of Artificial Intelligence (AI), has grown rapidly in recent years in the context of data analysis and computing, which typically enables applications to function in an intelligent way [14]. Machine learning is also known as a powerful alternative method to analyze time series data, especially when the data is nonlinear [15]. In recent years, ML has grown significantly in terms of application in various domains [16]. This study used two algorithms, that is, Extreme Gradient Boosted (XG Boost) and Random Forest. XG Boost is an algorithm or engineering implementation which is developed from the Gradient Boosted Decision Tree (GBDT) [17]. XG Boost was first proposed by Chen in 2014 [18].

Random Forest (RF) was one of the ML models of the ensemble [19]. It is an innovation based on bagged decision trees, which allows split-variable randomization [20]. This method is combined from many trees forming a forest that is used to analyze and make decisions [21]. Modeling using Random Forest Regression is considered to provide better performance when compared to that using only one decision tree. Random Forest is also quite compatible with handling missing values and is able to produce results with a minimum of error. In accordance with the objective of this research, both methods will be compared for accuracy in predicting air pollution levels in Surabaya.

From the previous research, in 2021, T Madan et al. conducted research about the prediction of air quality in Avd. Francia Station [22]. The result is the Mean Squared Error (MSE) value of CO concentration is 0.53, the MSE value of NO concentration is 29.517, and the MSE value of NO_2 is 14.85. Still in 2021, Lu J et al. have conducted research about the prediction of air quality of several cities in China using $PM_{2.5}$ parameter [23]. The

result is that Random Forest showed suitable performance in both time and space (R^2 = 0.88, RMSE = 11.94 $\mu g/m^3$, BIAS = 0.30 $\mu g/m^3$), which can meet the requirements of air pollution monitoring in urban areas.

J Ma et al. in 2020, conducted research on air pollution prediction in Shanghai, China, using the XG Boost method [24]. The research that has been done successfully and produces renewal that combines the XG Boost method with Weather Research Forecasting-coupled with Chemistry (WRF-Chem) model. The result is that XG Boost successfully achieves a higher accuracy of $PM_{2.5}$ concentration than the WRF-Chem model. Also in 2020, X Ma et al. conducted research about prediction on outdoor air temperature and humidity using the XG Boost method. The result shows the satisfactory ability of XG Boost [25]. For application purposes, the XG Boost method will be integrated into a microcontroller to reduce the cost of implementing energy management. In 2019, TV Vu et al. conducted research about air quality in Beijing using the Random Forest method [26]. Random Forest successfully produced a deviation value between observed and predicted values of $PM_{2.5}$ in the range of 0.4% - 7.9% with an average of 1.5%.

2. RESEARCH METHODS

2.1 Dataset

The dataset used in this research was obtained from the Surabaya City air condition data from 01/01/2020 to 31/12/2020. The data used in this study came from daily measurements recorded by a measuring device installed at an air pollution spot. The dataset contains the names of several pollutants. After the dataset was obtained, then analyzed statistically to get insight into the real condition. Before in-depth analysis, the data was cleaned of missing values, outliers, and other noise. This is all to simplify comprehension and improve the quality of the data to be tested, and to ease the decision of the machine learning model used. An overview of the research data can be seen in Table 1, and the flow of research methodology in Fig. 1 below.

Table 1. Dataset

Date	Particulate Matter	Sulfur Dioxide	Carbon Monoxide	Ozone	Nitrogen Dioxide	Other Pollutant	RESULT
01/01/2020	30	20	10	32	9	32	GOOD
02/01/2020	27	22	12	29	8	29	GOOD
03/01/2020	39	22	14	32	10	39	GOOD
04/01/2020	34	22	14	38	10	38	GOOD
05/01/2020	35	22	12	31	9	35	GOOD
06/01/2020	46	23	16	32	9	46	GOOD
07/01/2020	37	23	26	33	11	37	GOOD
08/01/2020	41	26	20	30	11	41	GOOD
09/01/2020	52	23	29	24	12	52	AVERAGE
10/01/2020	24	24	18	25	8	25	GOOD
...	...	...	...	...	...	...	...
31/12/2020	18	13	6	24	3	24	GOOD

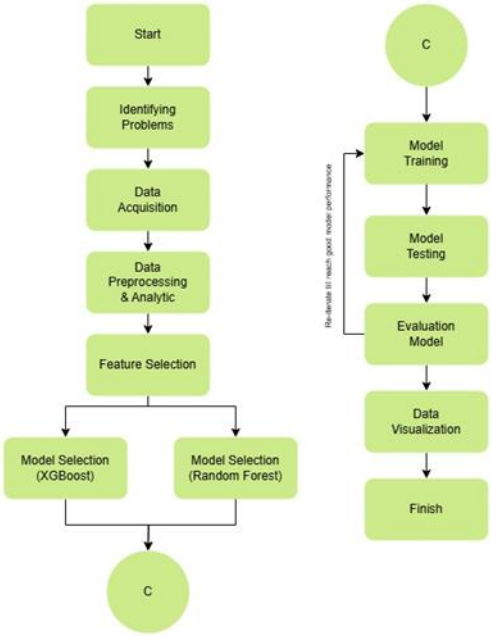

Figure 1. Research Methodology

2.2 Exploratory Data Analysis

This research presents a case study on the prediction of air pollution rates based on Ozone (O3) concentration in the city of Surabaya. The dataset processed consists of 1830 rows and 9 columns. Data is taken in the time span between 01/01/2020 to 31/01/2020. All data types in the dataset are numeric. The data is processed to find anomalies such as missing values, user input errors, etc. Below in Fig. 2 is the condition of the data from its original source.

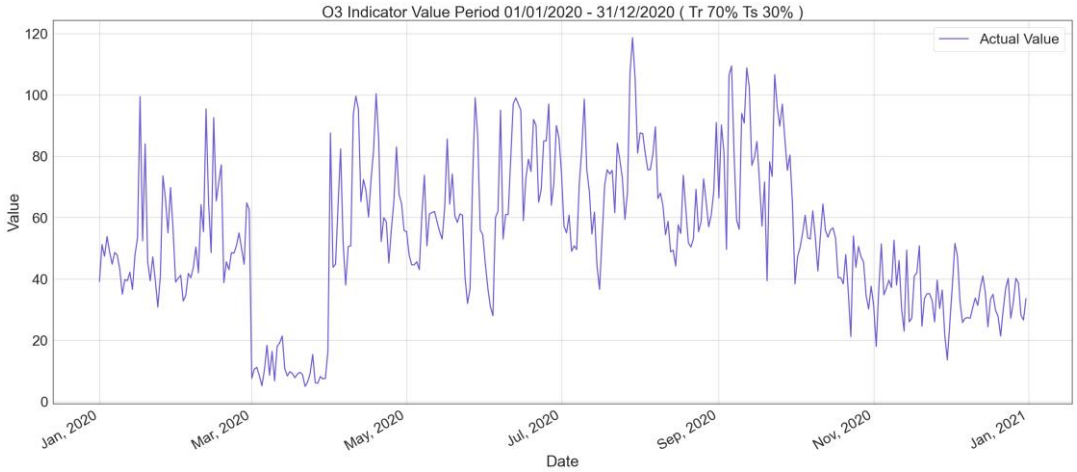

Figure 2. Plot of O3 Indicator Value

From the dataset, a statistical analysis can be drawn in the form of a measurement of central tendency, such as Table 2 below.

Table 2. Measure of Central Tendency

	Particulate Matter	Sulfur Dioxide	Carbon Monoxide	Ozon	Nitrogen Dioxide	Other Pollutant
mean	48.63	22.84	19.5	52.93	19.08	57.19
min	3.00	1.00	3.00	1.00	0.00	1.00
max	111.00	112.00	99.00	191.00	213.00	191.00
std	17.42	15.7	17.66	29.94	27.18	30.78

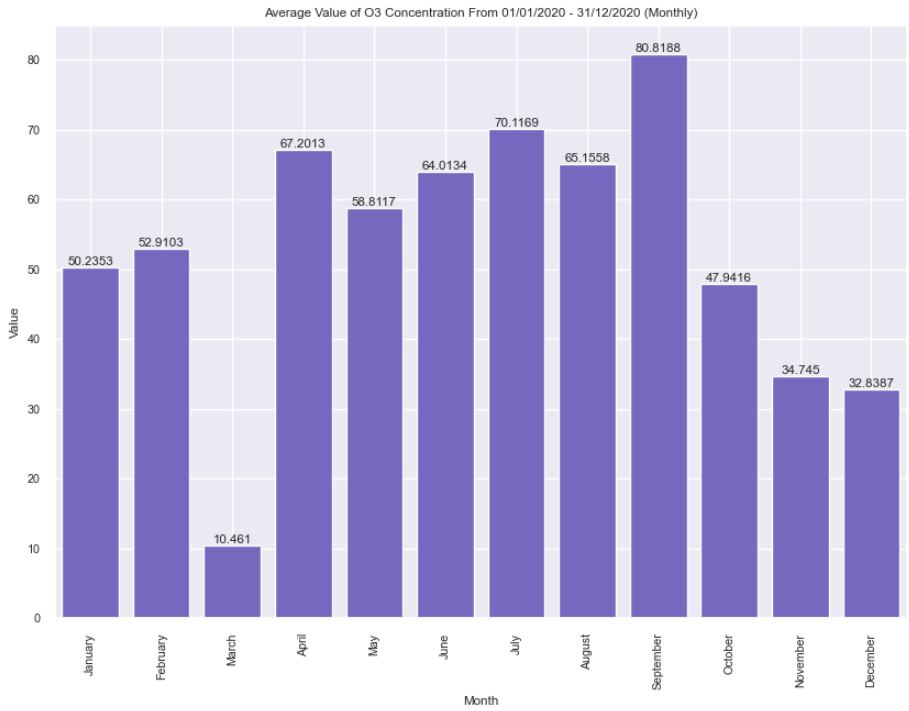

Figure 3. Plot Average Value of O3 from 01/01/2020 – 31/12/2020

From the plot in Fig. 3, it can be known that the average value of the parameter O3 concentration is the highest in September 2020 and the lowest in March 2020. After March 2020, a significant increase occurs in April 2020 and moves dynamically until it touches the highest value in September 2020, and drops back significantly in October and continues until December 2020.

After that, the data is refined by normalizing the values. The function of this normalization is to equalize the range of values between 0-1. Below in Eq. (1) is the function to normalize data, namely the Min Max Scaler.

$$x^1 = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{1}$$

with x^1 is a normalized result, $\min(x)$ is the minimum value of the attribute and $\max(x)$ is the maximum value of the attribute.

2.3 Feature Selection

In many prediction cases with numerical data, the way to determine the features used as independent and dependent variables is to use analysis techniques based on the Pearson product-moment correlation. Below in Eq. (2) is the mathematical function of the Pearson product-moment, and the result of the correlation analysis is shown in Fig. 4 below.

$$\boldsymbol{r_{xy}} = \frac{\boldsymbol{N\sum XY - (X)(Y)}}{\sqrt{\boldsymbol{N\sum X^2 - \sum X^2\, N\sum Y^2 - \sum Y^2}}} \tag{2}$$

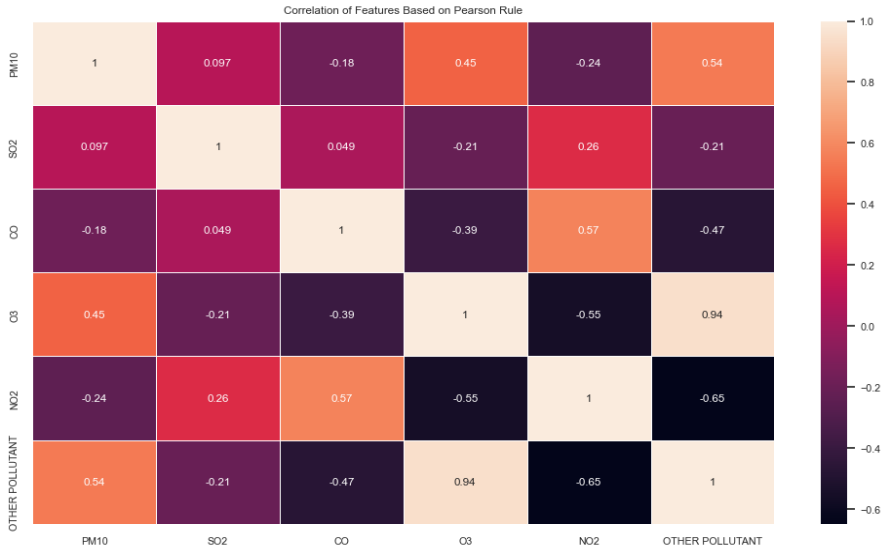

Figure 4. Correlation Plot

From the correlation plot in Fig. 4 above, it is shown that variable 'Other Pollutant' has a correlation value of 0.94, and variable 'PM10' has a correlation value of 0.45. These two variables have a positive correlation, but the variable 'Other Pollutant' has a higher correlation value and is close to 1. Whereas, variable 'Particulate Matter (PM10)' also has a positive correlation value, but includes a low positive category. In selecting independent variables, it is important to pay attention to positive and strong correlation values with the aim of maximizing predictive value [27]. Both of them were selected to be independent variables that have an effect on Ozone (O3) as the dependent variable.

2.4 XG Boost Prediction Model

XG Boost first shaped multiple models called Classification and Regression Trees (CART). These models are applied to predict the data set, and then integrate these trees as a new model. The model will continue to be iteratively improved, and a new tree model generated in each iteration will fit the residual of the previous tree [28]. Below is a function of the model XG Boost in Eq. (3).

$$\hat{y}_i = \varphi(x_i) = \sum_{t=1}^{T} f_t(x_i) \tag{3}$$

From Eq. (3), it can be explained that x_i is the feature of the sample, and $f_t(x_i)$ uses the *t*-th tree to predict the *i*-th sample. Adding the results together, the final predicted value $\hat{y}_i$ and true label is y_i. Then, for Eq. (4) below is an objective function.

$$Obj = \sum_{i=1}^{n} (y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega\,(f_k) \tag{4}$$

where the first term $\sum_{i=1}^{n}(y_i, \hat{y}_i)$ is the loss function and $\sum_{k=1}^{K} \Omega\,(f_k)$ is the regular item to control the complexity and prevent overfitting.

2.5 Random Forest Prediction Model

Random Forest is a learning algorithm that is based on an ensemble of trees [29]. The Random Forest consists of a set of decision trees which taken randomly from a subset of the training set. Random Forest needs more processing time, but has better accuracy than other ML algorithms [30]. The formula of the Random Forest tree is in Eq. (5) as follows:

$$\mathrm{N} = \{(x_1, y_1), (x_2, y_2), (x_n, y_n)\}. \tag{5}$$

This combines with data flows into a Random Forest K-like formulation in Eq. (6).

$$\mathrm{K} = \left\{\left(k_1(x)\right), \left(k_2(x)\right), \left(k_j(x)\right)\right\}, \tag{6}$$

where j is the number in the universe of trees. The utility is calculated using the following formula in Eq. (7).

$$\mathrm{U} = \left\{d_{i_1}, d_{i_2}, d_{i_m}\right\}, \tag{7}$$

where m is the number in the universe of variables.

$$K_j(x) = k\left(\frac{x}{d_i}\right). \tag{8}$$

2.6 Evaluation Model

At this stage, the model that has been trained and tested is calculated for accuracy based on the resulting error value. Accuracy is used as a parameter to compare the measurement result by the model and the actual value obtained before [31]. This study uses the Root Mean Square Error (RMSE) as a method to calculate the error value resulting from the model. One of the main advantages of using RMSE is to assign a higher weightage (as it contains a square) to larger errors [32]. The function of the Root Mean Square Error (RMSE) is in Eq. (9) as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(X_i - F_i)^2}{n}} \tag{9}$$

3. RESULTS AND DISCUSSION

The following are parameters of XG Boost and Random Forest used to build a prediction model and a data splitting table.

Table 3. Data Splitting

Splitting Percentage	Number of Data Training	Number of Data Testing
70% : 30%	1281	549
75% : 25%	1372	458
80% : 20%	1464	366
85% : 15%	1555	275
90% : 10%	1647	183

Table 4. XG Boost Model Parameter

n estimator	max depth	learning rate
100	10	0.03

Table 5. Random Forest Model Parameter

n estimator	max depth	n jobs
100	10	1

From Table 4 and Table 5, it can be seen that both methods are used. The main difference between the XGBoost method and Random Forest lies in the learning rate parameter, which helps reduce the loss function and leads to more optimal prediction results. In addition to the parameter optimization treatment applied to the XG Boost and Random Forest methods, it also refers to the percentage level of data splitting. This will be explained in detail in the next section.

3.1 Simulation Result

In this research was conducted implementation of two algorithms from machine learning, namely XG Boost and Random Forest to forecast air pollution using the Python programming language and comparison was made based on the method and the difference in the composition of training data and testing data as shown

in below, where Fig. 5 is the simulation result of the XG Boost and Random Forest algorithms with 70% training data and 30% testing data.

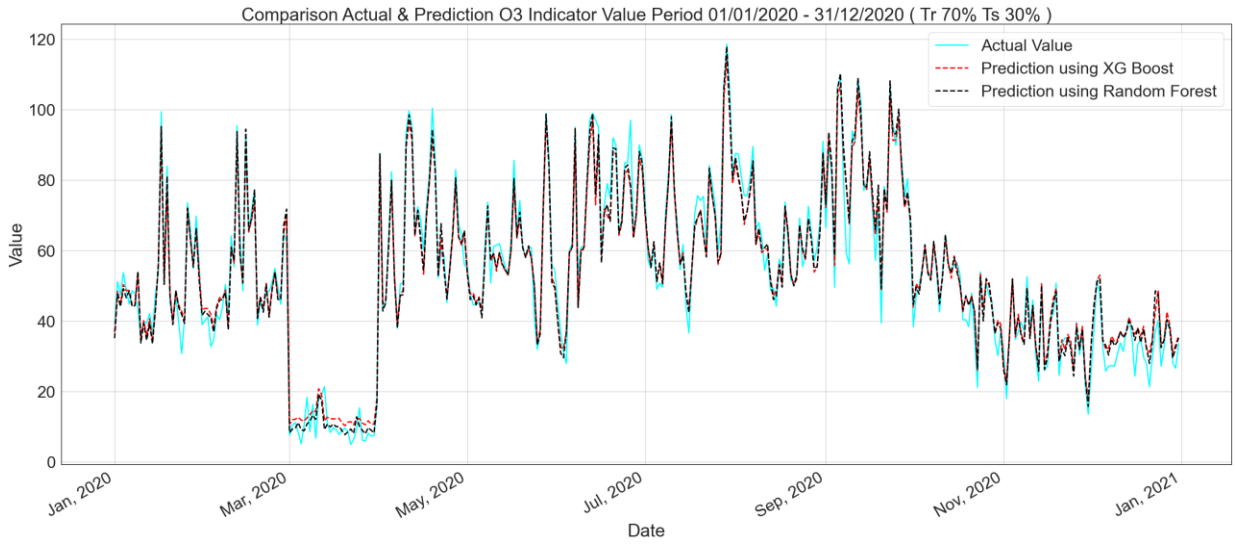

Figure 5. First Simulation Plot of XG Boost and Random Forest (70% Training Data and 30% Testing Data)

The result of the first simulation, presented in Fig. 5, was conducted by using 70% training data and 30% testing data, and both methods achieved prediction results close to the actual value shown by the cyan line. The prediction results of the XG Boost method produced an RMSE value of 0.0510, shown by the red line. Also, the Random Forest method managed to produce prediction results close to the actual value shown by the black line. The prediction results of the Random Forest method produced an RMSE value of 0.0469. In this first simulation, the Random Forest method produced a better RMSE value than the XG Boost method with a slight difference of 0.0041.

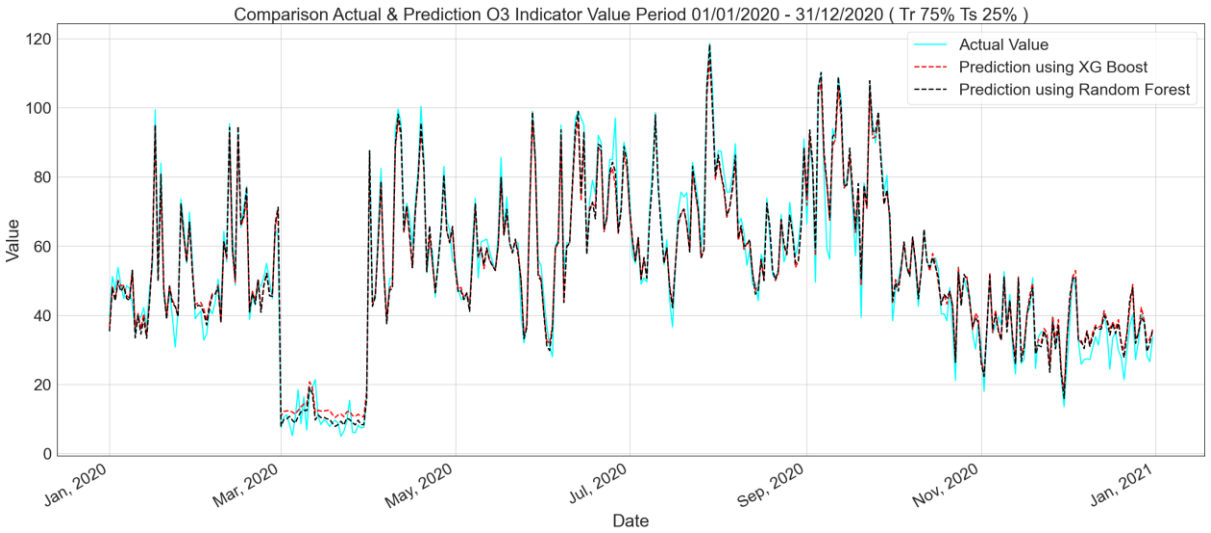

Figure 6. Second Simulation Plot of XG Boost and Random Forest (75% Training Data and 25% Testing Data)

The result of the second simulation, shown in Fig. 6, was conducted by using 75% training data and 25% testing data, and both methods achieved prediction results close to the actual value shown by the cyan line. The prediction results of the XG Boost method produced an RMSE value of 0.0515, shown by the red line. Also, the Random Forest method managed to produce prediction results close to the actual value shown by the black line. The prediction results of the Random Forest method produced an RMSE value of 0.0468. In this second simulation, there was a slight increase in error value in the XG Boost method, while the RF method experienced a slight decrease in error value.

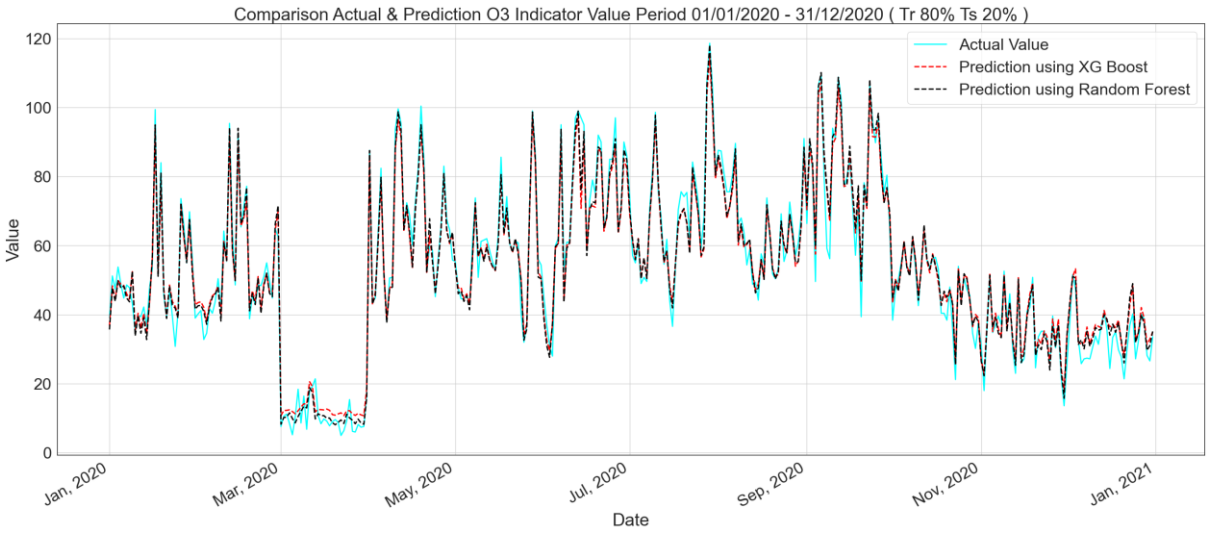

Figure 7. Third Simulation Plot of XG Boost and Random Forest (80% Training Data and 20% Testing Data)

The result of the third simulation, illustrated in Fig. 7, was conducted by using 80% training data and 20% testing data, and both methods achieved prediction results close to the actual value shown by the cyan line. The prediction results of the XG Boost method produced an RMSE value of 0.0528, shown by the red line. Also, the Random Forest method managed to produce prediction results close to the actual value shown by the black line. The prediction results of the Random Forest method produced an RMSE value of 0.0481.

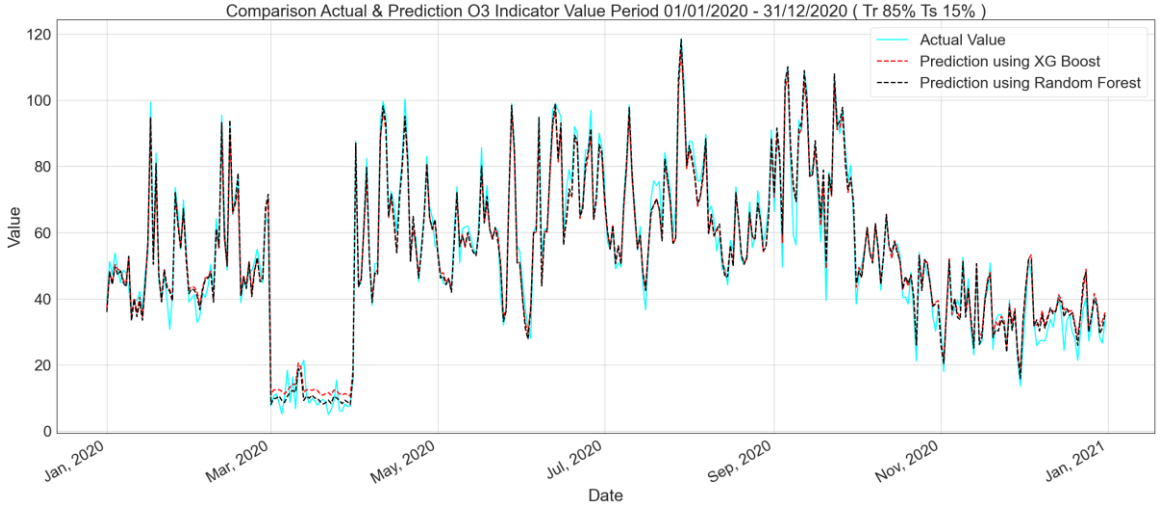

Figure 8. Fourth Simulation Plot of XG Boost and Random Forest (85% Training Data and 15% Testing Data)

As shown in Fig. 8, the fourth simulation was conducted by using 85% training data and 15% testing data, and both methods achieved prediction results close to the actual value shown by the cyan line. The prediction results of the XG Boost method produced an RMSE value of 0.0514, shown by the red line. Also, the Random Forest method managed to produce prediction results close to the actual value shown by the black line. The prediction results of the Random Forest method produced an RMSE value of 0.0471.

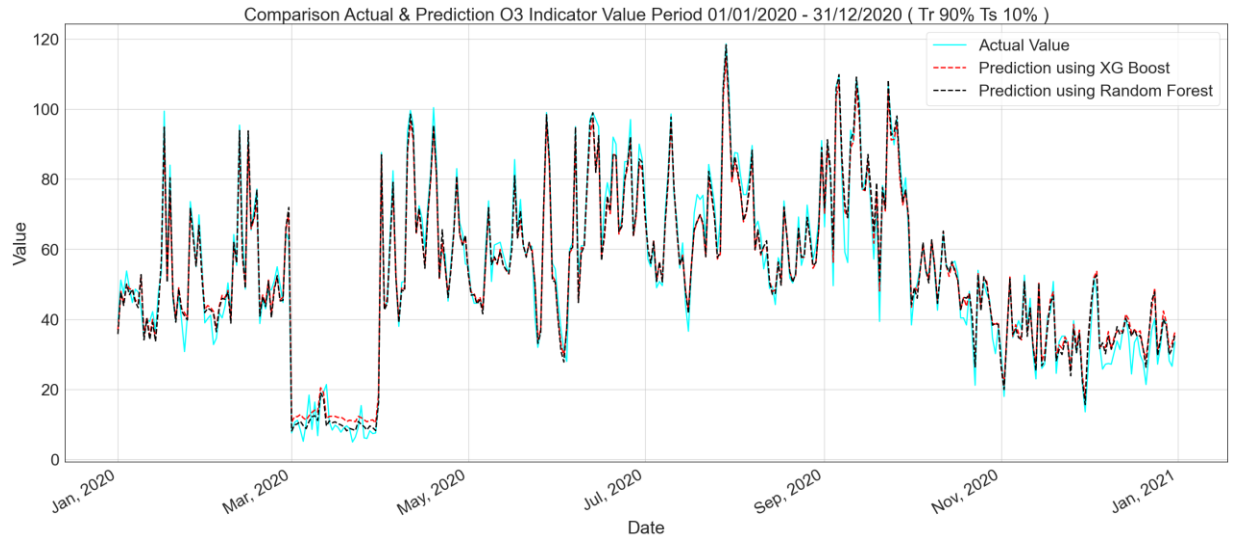

Figure 9. Fifth Simulation Plot of XG Boost and Random Forest (90% Training Data and 10% Testing Data)

The result of the fifth simulation, shown in Fig. 9, was conducted by using 90% training data and 10% testing data, and both methods achieved prediction results close to the actual value shown by the cyan line. The prediction results of the XG Boost method produced an RMSE value of 0.0556, shown by the red line. Also, the Random Forest method managed to produce prediction results close to the actual value shown by the black line. The prediction results of the Random Forest method produced an RMSE value of 0.0490.

From the first simulation until the fifth simulation, the XG Boost method produced an RMSE value in the range of approximately 0.5% and the Random Forest method produced an RMSE value of approximately 0.4%, not until reaching 0.5% or more consistently. The following is a recapitulation table of the simulation results of the XG Boost method and the Random Forest method, respectively, shown below.

Table 6. RMSE Comparison Value of XG Boost and Random Forest

Composition of Training Data and Testing Data	RMSE Value of XGBoost	RMSE Value of Random Forest
70% : 30%	0.0510	0.0469
75% : 25%	0.0515	0.0468
80% : 20%	0.0528	0.0481
85% : 15%	0.0514	0.0471
90% : 10%	0.0556	0.0490

The simulation results generated by the XG Boost method and the Random Forest method of the first simulation through the fifth simulation are presented in Table 6. On the results of the simulations with the 70% training data and 30% testing data, the XG Boost successfully produced the lowest RMSE values of all simulations. The XG Boost method produced the best RMSE value of 0.0510. From each simulation, it is shown that the XG Boost method consistently produces RMSE values in the range 0.0510 to 0.0556. It means that the overall simulation processed by the XG Boost method is approaching the actual value.

On Random Forest method, with 75% data training and 25% data testing, successfully produced the lowest RMSE value of all simulations. The Random Forest method produced the best RMSE value of 0.0468. From each simulation, it could be seen that the Random Forest method consistently produces an RMSE value in the range 0.0468 to 0.0490. Both of them consistently produce simulation results that approach the actual value. This shows that in this research, the Random Forest method shows better performance than the XG Boost method, although at each stage of the simulation, it can be seen that both methods produce good values below 1% and are possible to be applied by the city government or related stakeholders as decision decision-supporting application.

4. CONCLUSION

Based on the results of the simulations conducted, it can be concluded that the results of the first to third simulations using the XG Boost method managed to get the best prediction error value (RMSE) of

0.0510 in the first simulation, with a split of 70% training data and 30% testing data. While the Random Forest method managed to get the best prediction error value (RMSE) of 0.0469, also in the first simulation. By this study, also known that parameters from each method have significant results when performing prediction. Tuning of parameters has been an important part to notice, besides other factors. These results prove that the XG Boost and Random Forest methods provide good and consistent prediction results, so both methods have fulfilled the objective of this study and can be recommended for further study by using another optimizing default parameter or using an evolutionary algorithm like Genetic Algorithm or Particle Swarm Optimization to improve parameters for better results. In addition, the implication of this study enhances the academic sphere by delineating the efficacy outcomes of XG Boost and the Random Forest method, and another further purpose is to assist the local government in providing information on proper air conditioning to the public.

Author Contributions

Akas Yekti Pulih Asih: Conceptualization, Methodology, Writing-Original Draft, Software, Validation. Firman Yudianto: Data Curation, Resources, Draft Preparation. Puguh Triwinanto: Formal Analysis, Validation, Rachman Sinatriya Marjianto: Software, Visualization. Teguh Herlambang: Validation, Writing-Review and Editing. Hamzah Arof: Validation, Writing-Review and Editing. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgment

Many thanks to LPPM Universitas Nahdlatul Ulama Surabaya (UNUSA) for supporting and facilitating the writing of this publication.

Declarations

The authors declare no conflicts of interest regarding this research.

REFERENCES

[1] S. Suhardono *et al*., "CHANGES IN THE DISTRIBUTION OF AIR POLLUTANTS (CARBON MONOXIDE) DURING THE CONTROL OF THE COVID-19 PANDEMIC IN JAKARTA, SURABAYA, AND YOGYAKARTA, INDONESIA," J. Ecol. Eng., vol. 24, no. 4, 202. doi: https://doi.org/10.12911/22998993/159508

[2] B. Wardhani and V. Dugis, "GREENING SURABAYA: THE CITY'S ROLE IN SHAPING ENVIRONMENTAL DIPLOMACY," Bandung, vol. 7, no. 2, pp. 236–258, 2020. doi: https://doi.org/10.1163/21983534-00702005

[3] B. Angelevska, V. Atanasova, and I. Andreevski, "URBAN AIR QUALITY GUIDANCE BASED ON MEASURES CATEGORIZATION IN ROAD TRANSPORT," Civil Eng. J., vol. 7, no. 2, pp. 253–267, 2021. doi: https://doi.org/10.28991/cej-2021-03091651

[4] M. N. Anwar *et al*., "EMERGING CHALLENGES OF AIR POLLUTION AND PARTICULATE MATTER IN CHINA, INDIA, AND PAKISTAN AND MITIGATING SOLUTIONS," J. Hazard. Mater., vol. 416, p. 125851, 2021. doi: https://doi.org/10.1016/j.jhazmat.2021.125851

[5] G.-P. Bălă, R.-M. Râjnoveanu, E. Tudorache, R. Motișan, and C. Oancea, "AIR POLLUTION EXPOSURE—THE (IN)VISIBLE RISK FACTOR FOR RESPIRATORY DISEASES," *Environmental Science and Pollution Research*, vol. 28, no. 16, pp. 19615–19628, Mar. 2021. doi: https://doi.org/10.1007/s11356-021-13208-x

[6] N. Karin, G. Darmawan, and T. Hendrawati, "ENHANCING $PM_{2.5}$ PREDICTION IN KEMAYORAN DISTRICT, DKI JAKARTA USING DEEP BILSTM METHOD," *BAREKENG JURNAL ILMU MATEMATIKA DAN TERAPAN*, vol. 19, no. 1, pp. 185–198, Jan. 2025. doi: https://doi.org/10.30598/barekengvol19iss1pp185-198

[7] J. Yang, Y. Tian, and C. H. Wu, "AIR QUALITY PREDICTION AND RANKING ASSESSMENT BASED ON BOOTSTRAP-XGBOOST ALGORITHM AND ORDINAL CLASSIFICATION MODELS," *Atmosphere*, vol. 15, no. 8, p. 925, 2024. doi: https://doi.org/10.3390/atmos15080925

[8] R. S. Sokhi *et al*., "ADVANCES IN AIR QUALITY RESEARCH–CURRENT AND EMERGING CHALLENGES," *Atmos. Chem. Phys. Discuss.*, pp. 1–133, 2021. doi: https://doi.org/10.5194/acp-22-4615-2022

[9] I. Eguiluz-Gracia *et al*., "THE NEED FOR CLEAN AIR: THE WAY AIR POLLUTION AND CLIMATE CHANGE AFFECT ALLERGIC RHINITIS AND ASTHMA," *Allergy*, vol. 75, no. 9, pp. 2170–2184, 2020. doi: https://doi.org/10.1111/all.14177

[10] A. Feberina, A. W. E. Mulyadi, and R. H. Haryanti, "COLLABORATIVE GOVERNANCE IN SOLVING AIR POLLUTION PROBLEMS IN INDONESIA: A SYSTEMATIC LITERATURE REVIEW," in *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 905, no. 1, p. 012097, Nov. 2021. doi: https://doi.org/10.1088/1755-1315/905/1/012097

[11] N. M. N. Fitriana and C. W. Rubiyanto, "THE IMPACT OF SISTER CITY SURABAYA–KITAKYUSHU COOPERATION ON ENVIRONMENTAL DEVELOPMENT IN SURABAYA," *J. Paradiplomacy City Netw.*, vol. 1, no. 1, pp. 27–38, 2022. doi: https://doi.org/10.18196/jpcn.v1i1.15

[12] M. Sukarno and S. A. G. Putri, "SMART ENVIRONMENT PLANNING FOR SMART CITY BASED ON REGIONAL MEDIUM-TERM DEVELOPMENT PLAN SURABAYA CITY 2021–2026," in *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 1105, no. 1, p. 012023, Dec. 2022. doi: https://doi.org/10.1088/1755-1315/1105/1/012023

[13] Y. Deng *et al.*, "MULTI-HIERARCHICAL NANOFIBER MEMBRANE WITH CURVED-RIBBON STRUCTURE FABRICATED BY GREEN ELECTROSPINNING FOR EFFICIENT, BREATHABLE AND SUSTAINABLE AIR FILTRATION," *J. Membr. Sci.*, vol. 660, p. 120857, 2022. doi: https://doi.org/10.1016/j.memsci.2022.120857

[14] A. P. Ratnasari, B. Susetyo, and K. A. Notodiputro, "COMPARISON OF DOUBLE RANDOM FOREST AND LONG SHORT-TERM MEMORY METHODS FOR ANALYZING ECONOMIC INDICATOR DATA," *Barekeng J. Ilmu Mat. Terap.*, vol. 17, no. 2, pp. 0757–0766, Jun. 2023, doi: https://doi.org/10.30598/barekengvol17iss2pp0757-0766

[15] J. U. Hansen and P. Quinon, "THE IMPORTANCE OF EXPERT KNOWLEDGE IN BIG DATA AND MACHINE LEARNING," *Synthese*, vol. 201, no. 2, p. 35, 2023. doi: https://doi.org/10.1007/s11229-023-04041-5

[16] P. Kumar and M. Sharma, "DATA, MACHINE LEARNING, AND HUMAN DOMAIN EXPERTS: NONE IS BETTER THAN THEIR COLLABORATION," *Int. J. Hum.–Comput. Interact.*, vol. 38, no. 14, pp. 1307–1320, 2022. doi: https://doi.org/10.1080/10447318.2021.2002040

[17] H. Ke, S. Gong, J. He, L. Zhang, and J. Mo, "A HYBRID XGBOOST-SMOTE MODEL FOR OPTIMIZATION OF OPERATIONAL AIR QUALITY NUMERICAL MODEL FORECASTS," *Front. Environ. Sci.*, vol. 10, p. 1007530, 2022. doi: https://doi.org/10.3389/fenvs.2022.1007530

[18] B. Zhang, Y. Zhang, and X. Jiang, "FEATURE SELECTION FOR GLOBAL TROPOSPHERIC OZONE PREDICTION BASED ON THE BO-XGBOOST-RFE ALGORITHM," *Sci. Rep.*, vol. 12, p. 9244, 2022. doi: https://doi.org/10.1038/s41598-022-13498-2

[19] N. Palanichamy, S. C. Haw, S. Subramanian, R. Murugan, and K. Govindasamy, "MACHINE LEARNING METHODS TO PREDICT PARTICULATE MATTER PM2.5," *F1000Research*, vol. 11, 2022. doi: https://doi.org/10.12688/f1000research.73166.1

[20] L. Gaur *et al.*, "DISPOSITION OF YOUTH IN PREDICTING SUSTAINABLE DEVELOPMENT GOALS USING THE NEURO-FUZZY AND RANDOM FOREST ALGORITHMS," *Human-Centric Comput. Inf. Sci.*, vol. 11, 2021. doi: https://doi.org/10.22967/HCIS.2021.11.024

[21] D. K. Dalimunthe and R. B. F. Hakim, "APPLICATION OF RANDOM FOREST ALGORITHM ON WATCH PRICE PREDICTION SYSTEM USING FRAMEWORK FLASK," *Barekeng J. Ilmu Mat. Terap.*, vol. 17, no. 1, pp. 0171–0184, Apr. 2023. doi: https://doi.org/10.30598/barekengvol17iss1pp0171-0184.

[22] T. Madan, S. Sagar, and D. Virmani, "AIR QUALITY PREDICTION USING MACHINE LEARNING ALGORITHMS—A REVIEW," in *2020 2nd Int. Conf. Adv. Comput., Commun. Control Netw. (ICACCCN)*, 2020, pp. 140–145. doi: https://doi.org/10.1109/ICACCCN51052.2020.9362912

[23] J. Lu *et al.*, "ESTIMATION OF MONTHLY 1 KM RESOLUTION PM2.5 CONCENTRATIONS USING A RANDOM FOREST MODEL OVER '2+26' CITIES, CHINA," *Urban Climate*, vol. 35, p. 100734, 2021. doi: https://doi.org/10.1016/j.uclim.2020.100734

[24] J. Ma, Z. Yu, Y. Qu, J. Xu, and Y. Cao, "APPLICATION OF THE XGBOOST MACHINE LEARNING METHOD IN PM2.5 PREDICTION: A CASE STUDY OF SHANGHAI," *Aerosol Air Qual. Res.*, vol. 20, no. 1, pp. 128–138, 2020. doi: https://doi.org/10.4209/aaqr.2019.08.0408

[25] X. Ma, C. Fang, and J. Ji, "PREDICTION OF OUTDOOR AIR TEMPERATURE AND HUMIDITY USING XGBOOST," in *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 427, no. 1, p. 012013, 2020. doi: https://doi.org/10.1088/1755-1315/427/1/012013

[26] T. V. Vu *et al.*, "ASSESSING THE IMPACT OF CLEAN AIR ACTION ON AIR QUALITY TRENDS IN BEIJING USING A MACHINE LEARNING TECHNIQUE," *Atmos. Chem. Phys.*, vol. 19, no. 17, pp. 11303–11314, 2019. doi: https://doi.org/10.5194/acp-19-11303-2019

[27] M. Attallah, "PEARSON'S CORRELATION UNDER THE SCOPE: ASSESSMENT OF THE EFFICIENCY OF PEARSON'S CORRELATION TO SELECT PREDICTOR VARIABLES FOR LINEAR MODELS," *arXiv (Cornell University)*, Sep. 2024

[28] Liu, B., Tan, X., Jin, Y., Yu, W., & Li, C. (2021). APPLICATION OF RR-XGBOOST COMBINED MODEL IN DATA CALIBRATION OF MICRO AIR QUALITY DETECTOR. *Scientific Reports*, *11*(1), 15662. doi: https://doi.org/10.1038/s41598-021-95027-1

[29] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, "PREDICTION OF COVID-19 CONFIRMED, DEATH, AND CURED CASES IN INDIA USING RANDOM FOREST MODEL," *Big Data Min. Anal.*, vol. 4, no. 2, pp. 116–123, 2021. doi: https://doi.org/10.26599/BDMA.2020.9020016

[30] A. Nyangarika *et al.*, "ENERGY STABILITY AND DECARBONIZATION IN DEVELOPING COUNTRIES: RANDOM FOREST APPROACH FOR FORECASTING OF CRUDE OIL TRADE FLOWS AND MACRO INDICATORS," *Front. Environ. Sci.*, vol. 10, p. 1031343, 2022. doi: https://doi.org/10.3389/fenvs.2022.1031343

[31] D. Novita, T. Herlambang, V. Asy'ari, A. Alimudin, and H. Arof, "COMPARISON OF K-NEAREST NEIGHBOR AND NEURAL NETWORK FOR PREDICTION OF INTERNATIONAL VISITORS IN EAST JAVA," *Barekeng J. Ilmu Mat. Terap.*, vol. 18, no. 3, pp. 2057–2070, Jul. 2024, doi: https://doi.org/10.30598/barekengvol18iss3pp2057-2070.

[32] M. A. Khan, M. I. Shah, M. F. Javed, M. I. Khan, S. Rasheed, M. A. El-Shorbagy, E. R. El-Zahar, and M. Y. Malik, "APPLICATION OF RANDOM FOREST FOR MODELLING OF SURFACE SALINITY," Ain Shams Engineering Journal, vol. 13, no. 4, p. 101635, Jun. 2022. doi: https://doi.org/10.1016/j.asej.2021.11.004