IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 515~523
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp515-523

Air quality prediction using boosting-based machine learning models for sustainable environment

Ahmad Fauzi[1], Maharina[2], Jamaludin Indra[1], Ayu Ratna Juwita[1], Agustia Hananto[2], Euis Nurlaelasari[1]
[1]Informatics Engineering Program, Faculty of Computer Science, Universitas Buana Perjuangan Karawang, Karawang, Indonesia
[2]Information Systems Program, Faculty of Computer Science, Universitas Buana Perjuangan Karawang, Karawang, Indonesia

Article Info

Article history:

Received Oct 2, 2024
Revised Nov 10, 2025
Accepted Dec 15, 2025

Keywords:

AdaBoost classifier
Air quality prediction
LGBM classifier
Machine learning
XGBoost classifier

ABSTRACT

High levels of air pollution are extremely harmful to humans and the environment. They increase the risk of respiratory infections and lung cancer, especially among vulnerable populations. Therefore, developing effective pollution control measures is crucial for mitigating these negative impacts. We need to implement effective methods to predict and manage air quality for the sake of public health and a healthier environment. In recent years, machine learning (ML) methods have been increasingly utilized in air quality prediction due to their ability to analyze datasets and identify complex patterns. However, the reliability and accuracy of air quality prediction models remain a challenge. This study proposes a boosting-based ML model for predicting air quality. We implemented three stages in the proposed method. In the first stage, we conducted data preprocessing and analysis to eliminate noise, remove redundant data, and encode categorical features. In the second stage, we predicted air quality categories by leveraging 25 ML models, dividing them into three distinct categories. The results show that the extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), and adaptive boosting (AdaBoost) models outperform the others in air quality prediction, achieving an accuracy of 99%. Finally, we compared these three models using 10-fold cross-validation to ensure they generalize well in last stage.

Corresponding Author:

Maharina
Information Systems Program, Faculty of Computer Science, Universitas Buana Perjuangan Karawang
H.S. Ronggowaluyo Street, Teluk Jambe, Karawang, Indonesia
Email: maharina@ubpkarawang.ac.id

1. INTRODUCTION

Air pollution refers to the presence of harmful substances in the air that adversely affect health [1]–[3]. These harmful substances can be fine particles, toxic gases, or other chemical compounds suspended in the atmosphere. When inhaled, these substances can cause various health problems, such as irritation of the eyes and throat to chronic respiratory diseases. Additionally, air pollution can have widespread environmental impacts, including damage to plants, animals, and entire ecosystems [4], [5]. Air pollution, which has many harmful effects [6], [7] must be avoided, and therefore effective management measures are necessary. In supporting sustainable urban development, accurate air quality monitoring and prediction technologies play a crucial role [8], [9]. These technologies provide essential guidance for decision-making related to urban environmental management [10], [11].

Several studies on air quality prediction using machine learning (ML) have been conducted using various methods across different locations [12]–[20]. Research by Imam *et al.* [21] in Rabindra and Victoria, India, employed support vector classifier (SVC) and random forest (RF) techniques. This study focused on

pollutants such as $PM_{2.5}$, PM_{10}, NO_2, CO, SO_2, and O_3. During data preprocessing stage, 70% of the data was allocated for training and 30% for testing. The experimental results reported accuracy of 97.98 and 93.29% for Rabindra and Victoria, respectively. Meanwhile, the research in [15], [22] used an 80:20 split ratio for their training and test datasets. Khadom *et al.* [22] proposed the use of multilayer perceptron (MLP) and long short-term memory (LSTM) models to predict air quality in Baghdad, while Janarthanan *et al.* [15] employed LSTM to forecast air quality in India.

Recent study in air quality prediction was conducted by Resti *et al*. [23] in Shanghai, China, utilized ensemble naïve Bayes (NB), decision tree (DT), and RF methods, achieving exceptionally high accuracy (99.89%). Similarly, Livingston *et al.* [24] study in Beijing also manually split the dataset. This study applied fuzzy logic techniques and included a broader range of variables, such as temperature, humidity, and wind speed. However, most of the air quality prediction studies did not implement k-fold validation techniques as shown in Table 1. The use of manual dataset splitting without k-fold validation can compromise the reliability of the results. Although the reported accuracies are promising, the lack of comprehensive validation may affect the model's ability to generalize.

Table 1. Previous research on air quality prediction using ML methods

Researchers	Datasets	Methods	Features	Tasks	Data processed	
					Split_manually?	k-fold_ validation
Livingston *et al.* [24]	Beijing, China	Fuzzy	NO_2, CO, O_3, $PM_{2.5}$, PM_{10}, SO_2, TEMP, PRES, DEWP, RAIN, WD, WSPM	Regression	Manually	No
Khadom *et al*. [22]	Baghdad, Iraq	MLP and LSTM	$PM_{2.5}$	Regression	80:20	No
Janarthanan *et al*. [15]	India	LSTM	CO, SO_2, NO_2, $PM_{2.5}$	Regression	80:20	No
Imam *et al.* [21]	Victoria, India	RF	$PM_{2.5}$, PM_{10}, NO_2, CO, SO_2, O_3	Classification	70:30	No
Imam *et al.* [21]	Rabindra, India	SVC	$PM_{2.5}$, PM_{10}, NO_2, CO, SO_2, O_3	Classification	70:30	No
Resti *et al*. [23]	Shanghai, China	Ensemble NB, DT, RF	Weather factors and atmospheric variables (temperature, sun hour, humidy, wind, total snow, heat, moon illumination, cloud)	Classification	Manually	No

To address the issue, this study proposes a boosting-based ML model for predicting air quality. Boosting algorithms, including adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), have shown promising results in various prediction tasks due to their ability to handle complex data and their robustness against overfitting. The main contribution of this paper lies in the fact that the proposed model not only performs well with manual data splitting but also maintains its high accuracy when cross-validation techniques are applied. The key findings can be summarized as:

- We propose using boosting-based ML models (XGBoost, LGBM, and AdaBoost) to accurately predict air quality.
- Among 25 ML models, performance of XGBoost outperforms the others.
- This method was demonstrated using a flood dataset from Jakarta, Indonesia.
- We utilize the proposed boosting methods (XGBoost, LGBM, and AdaBoost) with 10-fold cross-validation to minimize bias and ensure that the model's performance is generalized.

2. MATERIAL AND METHOD

The methodology of this study consists of several key steps as shown in Figure 1. The initial stage involves data preprocessing and analysis, which includes handling missing values, addressing redundant data, encoding categorical features, and splitting the dataset into training and testing subsets. This process is crucial for ensuring that the data is clean and ready for analysis, ultimately enhancing the quality and reliability of the subsequent ML applications. Following this, a diverse array of ML algorithms is applied to the preprocessed data.

The dataset used in this study contains data on the air pollution standard index collected from five air quality monitoring stations located across the Province of DKI Jakarta, Indonesia. The data spans the period from January 2021 to December 2021. The air quality categories in this dataset are classified into three levels: "good", "moderate", and "unhealthy". These features are summarized in Tables 2 and 3, which provides further details on the structure of the dataset. In this study, categorical features encoding is

applied to prepare the dataset for use in ML algorithms. To achieve this, the scikit-learn package (https://scikit-learn.org/) is utilized [16]. Then the dataset is split into 80% for training and 20% for testing.

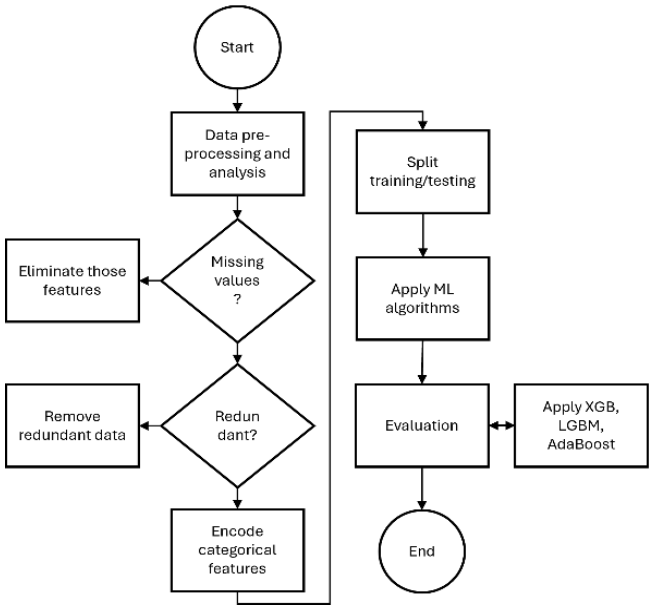

Figure 1. Workflow of the proposed research method

Table 2. Summary of features in the air quality dataset

No	Features	Descriptions	Data type
1	Date	Air quality measurement date	Object
2	PM_{10}	Particulates are one of the parameters measured	Integer
3	$PM_{2.5}$	Particulates are one of the parameters measured	Floating point
4	SO_2	Sulfide (in the form of SO_2) is one of the parameters measured	Integer
5	CO	Carbon monoxide is one of the parameters measured	Integer
6	O_3	Ozone is one of the parameters measured	Integer
7	NO_2	Nitrogen dioxide is one of the parameters measured	Integer
8	Max	The highest measured value of all parameters measured at the same time	Integer
9	Critical	The parameter with the highest measurement results	String
10	Category	Category of results of air pollution standard index calculation	String
11	Locations	Measurement location at the station	String

Table 3. The detailed of dataset

No	Date	PM_{10}	$PM_{2.5}$	SO_2	CO	O_3	NO_2	Max	Critical	Category	Location
1	10/6/2021	66	103.0	66	10	58	35	103	$PM_{2.5}$	Unhealthy	DKI4
2	3/5/2021	65	81.0	54	16	59	30	81	$PM_{2.5}$	Moderate	DKI2
3	6/24/2021	80	119.0	54	19	42	52	119	$PM_{2.5}$	Unhealthy	DKI4
4	9/21/2021	61	99.0	52	11	58	36	99	$PM_{2.5}$	Moderate	DKI4
5	10/21/2021	53	74.0	61	11	57	32	74	$PM_{2.5}$	Moderate	DKI3
6	12/4/2021	50	65.0	45	13	43	16	65	$PM_{2.5}$	Moderate	DKI3
7	11/27/2021	37	56.0	41	10	45	22	56	$PM_{2.5}$	Moderate	DKI4
8	12/2/2021	35	56.0	42	7	40	14	56	$PM_{2.5}$	Moderate	DKI4
9	9/18/2021	57	101.0	53	9	51	22	101	$PM_{2.5}$	Unhealthy	DKI4
…	…	…	…	…	…	…	…	…	…	…	…
346	12/13/2021	53	68.0	44	11	34	23	68	$PM_{2.5}$	Moderate	DKI3

2.1. Data pre-processing

Preprocessing methods play a crucial role in developing accurate ML models [21], [25]. The initial stage of the analysis involves data preprocessing, which is includes handling missing values and removing redundant data. In addressing missing values, rows containing not a number (NaN) values in the $PM_{2.5}$ variable are removed to ensure the integrity of the dataset. Such missing values can lead to misinterpretation and inaccurate predictions, as the model lacks the complete information necessary for making prediction. A visualization of the missing values in the $PM_{2.5}$ feature is presented in Figure 2. Furthermore, redundant data, are eliminated to prevent bias in the analysis.

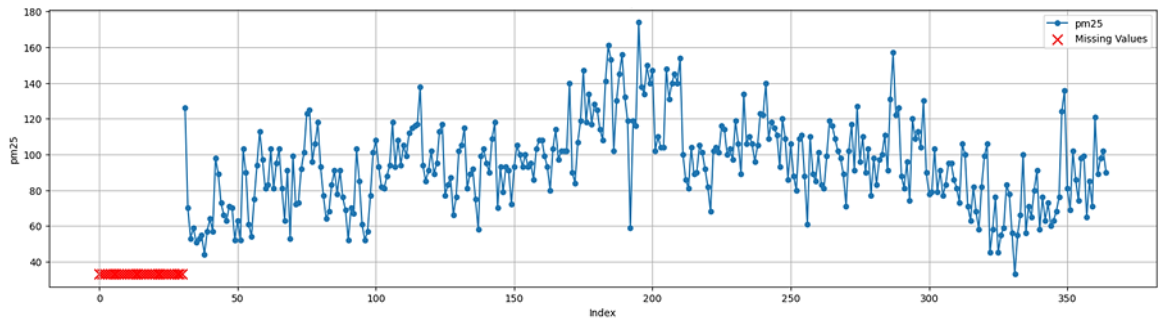

Figure 2. Graph of the time series for the $PM_{2.5}$ feature showing missing values

The presented correlation matrix as shown in Figure 3 offers an analysis of the relationships among various air quality parameters within this study. The heatmap demonstrate the correlation of air quality features, with yellow indicating a positive correlation and dark blue representing a negative correlation. Lighter shades indicate a stronger correlation between the features.

Figure 3. Correlation heatmap of the air quality dataset

2.2. Prediction

In second stage, we utilize several ML including AdaBoost [14], XGBoost [26], LGBM [17], RF [26], DT [26], extra trees, bagging, stochastic gradient descent (SGD) classifier, logistic regression, linear discriminant analysis, nearest centroid, label propagation, label spreading, perceptron, linear SVC [26], passive aggressive classifier, calibrated classifier cross-validation (CV), ridge classifier [26], ridge classifier CV [26], SVC [26], Bernoulli naïve Bayes (BernoulliNB), quadratic discriminant analysis, Gaussian naïve Bayes (GaussianNB), k-nearest neighbors (KNN) [26], and dummy classifier. In third stage, we leverage three models: XGBoost, LightGBM, and AdaBoost. These models are selected to evaluate and compare classification performance across different learning approaches.

2.3. Cross-validation technique

In the third stage, we implemented 10-fold cross-validation. This technique divides the dataset into 10 subsets (folds), where each subset takes turns serving as the test set while the remaining 9 subsets are used as the training set. This process is repeated 10 times, ensuring that each subset acts as the test set exactly once. The primary goal of this method is to reduce bias in the model and improve the model's generalization to new data.

2.4. Evaluation model

In order to evaluate the proposed method for water quality prediction, we utilized accuracy and F1-score metrics, as presented in Table 4. TP refer to true positive, FP for false positive, FN for false

negative, and TN for true negative. These metrics provide a comprehensive assessment of the model's performance, allowing us to determine its effectiveness in correctly predicting air quality.

Table 4. Evaluation model equations

Metrics	Equations
Accuracy	(TP+TN)/(TP+TN+FP+FN)
F1-score	2TP/(2TP+FP+FN)

3. RESULTS AND DISCUSSION

Table 5 highlights the performance of various ML algorithms for air quality prediction, focusing on their accuracy, F1-score, and inference time. AdaBoost, XGBoost, LGBM, RF, and DT achieved the highest performance with accuracies and F1-scores of 0.99 and 0.98, respectively. Notably, AdaBoost, XGBoost, and LGBM delivered identical results in both accuracy and F1-score (0.99) as shown in Figure 4, while maintaining inference times under 0.35 seconds.

Table 5. Performance of air quality prediction using ML algorithms

Models	Accuracy	F1-score	Inference time (seconds)
AdaBoost	0.99	0.99	0.35
XGBoost	0.99	0.99	0.30
LGBM	0.99	0.99	0.26
RF	0.98	0.98	0.41
DT	0.98	0.98	0.02
Extra trees	0.97	0.97	0.32
Bagging classifier	0.96	0.97	0.07
SGD classifier	0.95	0.96	0.03
Logistic regression	0.94	0.95	0.06
Linear discriminant analysis	0.93	0.94	0.03
Nearest centroid	0.84	0.87	0.06
Label propagation	0.81	0.82	0.03
Label spreading	0.81	0.82	0.03
Perceptron	0.98	0.98	0.05
LinearSVC	0.98	0.98	0.03
Passive aggressive classifier	0.97	0.97	0.02
Calibrated classifier CV	0.97	0.97	0.10
Ridge classifier	0.94	0.94	0.04
Ridge classifier CV	0.94	0.94	0.04
SVC	0.94	0.94	0.06
BernoulliNB	0.93	0.93	0.02
Quadratic discriminant analysis	0.92	0.92	0.04
GaussianNB	0.90	0.90	0.03
KNN	0.86	0.86	0.03
Dummy classifier	0.63	0.49	0.02

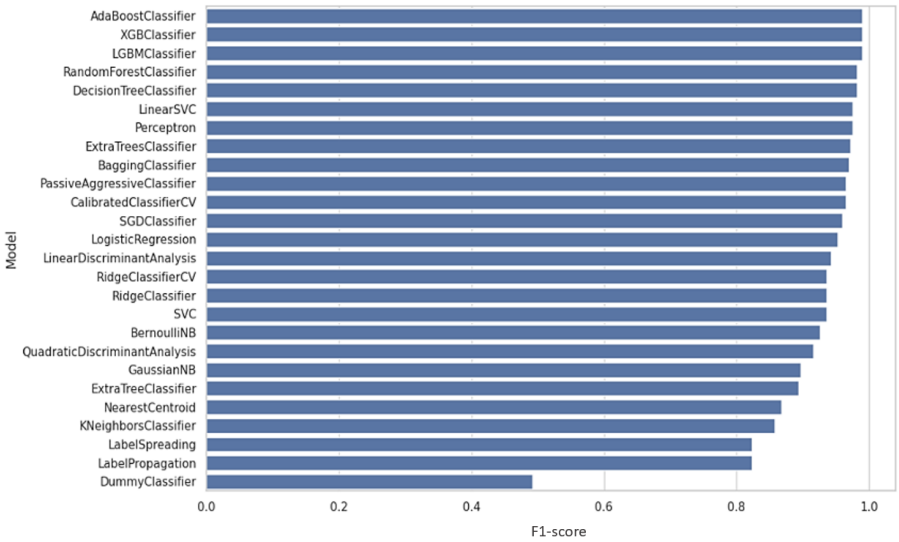

Figure 4. F1-score performance of the models

From the 25 ML approaches that were conducted, three of the best models were selected for further evaluation. In this third stage, AdaBoost, XGBoost, and LGBM models were trained using 10-fold cross-validation to ensure more stable model performance, and assess the generalization of the models to the dataset. Table 6 show the performance comparison of the three models. Among them, XGBoost stands out with the highest average accuracy of 0.9821.

Table 6. Performance comparison of air quality prediction models: AdaBoost, XGBoost, and LGBM classifier

Models	Mean accuracy	Standard deviation	Min accuracy	Max accuracy
AdaBoost	0.9702	0.0327	0.9091	1.0000
XGBoost	0.9821	0.0239	0.9294	1.0000
LGBM	0.9791	0.0272	0.9091	1.0000

Figure 5 shows the visualization of air quality prediction results from three models in stage 3. Figure 5(a) illustrates the distribution of these scores, highlighting how each model performs across different validation folds. Meanwhile, Figure 5(b) show a comparative visualization of the performance metrics obtained from the three models. Figure 5(c) illustrates the importance of features based on the output from three ML models. It is evident that the $PM_{2.5}$ feature holds the highest importance across all three models. Those visualization aids in understanding how the model learns from the dataset and makes accurate predictions by highlighting which features have the most influence on the model's decision-making process.

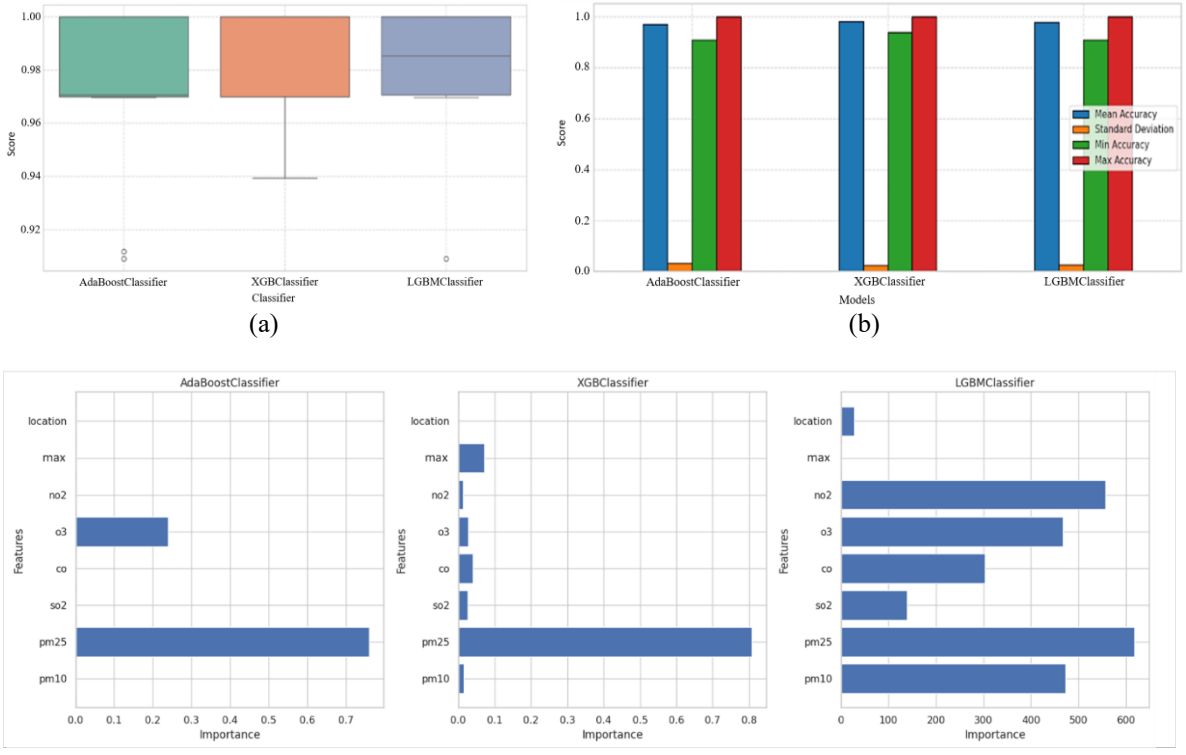

Figure 5. Visualization of air quality prediction results from three models in stage 3 of (a) 10-fold cross-validation score distribution, (b) performance metrics comparison, and (c) feature importance based on the three ML models' outputs

Table 7 show that our proposed method, XGBoost, outperforms other methods in terms of the F1-score. While Resti *et al.* [23] achieved a slightly higher accuracy of 99.89% using ensemble methods (NB, DT, and RF), their F1-score was significantly lower at 79.70%. In contrast, our method demonstrated a balanced performance, with both an F1-score and accuracy of 99.00%, indicating superior predictive

capability and consistency. This highlights the effectiveness of our approach in achieving both high accuracy and F1-score, surpassing the performance of previous studies.

Table 7. Comparison of our method with previous work

Authors	Datasets	Methods	Accuracy	F1-score
Resti *et al.* [23]	Shanghai, China	Ensemble NB, DT, RF	99.89	79.70
Imam *et al.* [21]	Rabindra, India	SVC	97.98	96.00
Imam *et al.* [21]	Victoria, India	RF	93.29	93.00
This work	Jakarta, Indonesia	XGBoost	99.00	99.00

4. CONCLUSION

This study presents a method for accurately predicting air quality using ML techniques. To achieve optimal performance, we propose a three-stage approach: i) data preprocessing is conducted; ii) predictions are made using 25 ML models on the preprocessed data; and iii) the best classifiers are further evaluated in the third stage using a 10-fold cross-validation process. This step ensures that the models generalize their performance well across the entire dataset. The experiments demonstrate that XGBoost has the most stable performance. Additionally, the important features of the dataset that significantly impact the model's ability to predict air quality are visualized. By identifying these key features, we can better appreciate the underlying mechanisms of the model. Although this study shows promising results, there are some limitations, one of which is the relatively small dataset size that fails to capture the full range of real-world conditions. To address this issue, future research should focus on collecting a larger dataset and incorporating more diverse features or variables. This approach will help improve the model's accuracy and reliability in making predictions. Ultimately, ML enable accurate predictions of air quality, thereby facilitating urban planning for sustainable development.

FUNDING INFORMATION

This research was supported by Universitas Buana Perjuangan Karawang, Indonesia.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ahmad Fauzi	✓	✓		✓	✓	✓			✓	✓	✓	✓		
Maharina	✓	✓	✓		✓		✓		✓	✓	✓			✓
Jamaludin Indra			✓	✓	✓	✓		✓	✓	✓	✓		✓	
Ayu Ratna Juwita				✓	✓			✓		✓			✓	✓
Agustia Hananto			✓	✓		✓		✓		✓			✓	
Euis Nurlaelasari				✓	✓	✓	✓			✓			✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [M], upon reasonable request.

REFERENCES

[1] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: a review," *Frontiers in Public Health*, vol. 8, 2020, doi: 10.3389/fpubh.2020.00014.

[2] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in California," *Complexity*, vol. 2020, pp. 1–23, 2020, doi: 10.1155/2020/8049504.
[3] M. Mele and C. Magazzino, "Pollution, economic growth, and COVID-19 deaths in India: a machine learning evidence," *Environmental Science and Pollution Research*, vol. 28, no. 3, pp. 2669–2677, 2021, doi: 10.1007/s11356-020-10689-0.
[4] Q. Liao, M. Zhu, L. Wu, X. Pan, X. Tang, and Z. Wang, "Deep learning for air quality forecasts: a review," *Current Pollution Reports*, vol. 6, no. 4, pp. 399–409, 2020, doi: 10.1007/s40726-020-00159-z.
[5] X. Liu, D. Lu, A. Zhang, Q. Liu, and G. Jiang, "Data-driven machine learning in environmental pollution: gains and problems," *Environmental Science & Technology*, vol. 56, no. 4, pp. 2124–2133, 2022, doi: 10.1021/acs.est.1c06157.
[6] V. V. Tran, D. Park, and Y.-C. Lee, "Indoor air pollution, related human diseases, and recent trends in the control and improvement of indoor air quality," *International Journal of Environmental Research and Public Health*, vol. 17, no. 8, 2020, doi: 10.3390/ijerph17082927.
[7] M. C. Turner *et al.*, "Outdoor air pollution and cancer: an overview of the current evidence and public health recommendations," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 6, pp. 460–479, 2020, doi: 10.3322/caac.21632.
[8] L. Zhang, P. Liu, L. Zhao, G. Wang, W. Zhang, and J. Liu, "Air quality predictions with a semi-supervised bidirectional LSTM neural network," *Atmospheric Pollution Research*, vol. 12, no. 1, pp. 328–339, 2021, doi: 10.1016/j.apr.2020.09.003.
[9] D. Seng, Q. Zhang, X. Zhang, G. Chen, and X. Chen, "Spatiotemporal prediction of air quality based on LSTM neural network," *Alexandria Engineering Journal*, vol. 60, no. 2, pp. 2021–2032, 2021, doi: 10.1016/j.aej.2020.12.009.
[10] L. Fu, J. Li, and Y. Chen, "An innovative decision making method for air quality monitoring based on big data-assisted artificial intelligence technique," *Journal of Innovation & Knowledge*, vol. 8, no. 2, 2023, doi: 10.1016/j.jik.2022.100294.
[11] A. C. O'Regan and M. M. Nyhan, "Towards sustainable and net-zero cities: a review of environmental modelling and monitoring tools for optimizing emissions reduction strategies for improved air quality in urban areas," *Environmental Research*, vol. 231, 2023, doi: 10.1016/j.envres.2023.116242.
[12] V. Gugnani and R. K. Singh, "Analysis of deep learning approaches for air pollution prediction," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 6031–6049, 2022, doi: 10.1007/s11042-021-11734-x.
[13] M. Lee *et al.*, "Forecasting air quality in Taiwan by using machine learning," *Scientific Reports*, vol. 10, no. 1, 2020, doi: 10.1038/s41598-020-61151-7.
[14] T. Y. Deo and A. Sanju, "Data imputation and comparison of custom ensemble models with existing libraries like XGBoost, CATBoost, AdaBoost and Scikit learn for predictive equipment failure," *Materials Today: Proceedings*, vol. 72, pp. 1596–1604, 2023, doi: 10.1016/j.matpr.2022.09.410.
[15] R. Janarthanan, P. Partheeban, K. Somasundaram, and P. N. Elamparithi, "A deep learning approach for prediction of air quality index in a metropolitan city," *Sustainable Cities and Society*, vol. 67, 2021, doi: 10.1016/j.scs.2021.102720.
[16] X. Lu *et al.*, "Feature fusion improves performance and interpretability of machine learning models in identifying soil pollution of potentially contaminated sites," *Ecotoxicology and Environmental Safety*, vol. 259, 2023, doi: 10.1016/j.ecoenv.2023.115052.
[17] X. Guo *et al.*, "Critical role of climate factors for groundwater potential mapping in arid regions: insights from random forest, XGBoost, and LightGBM algorithms," *Journal of Hydrology*, vol. 621, 2023, doi: 10.1016/j.jhydrol.2023.129599.
[18] K. K. Meena, D. Bairwa, and A. Agarwal, "A machine learning approach for unraveling the influence of air quality awareness on travel behavior," *Decision Analytics Journal*, vol. 11, 2024, doi: 10.1016/j.dajour.2024.100459.
[19] S. K. Natarajan, P. Shanmurthy, D. Arockiam, B. Balusamy, and S. Selvarajan, "Optimized machine learning model for air quality index prediction in major cities in India," *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-54807-1.
[20] S. A. Aram *et al.*, "Machine learning-based prediction of air quality index and air quality grade: a comparative analysis," *International Journal of Environmental Science and Technology*, vol. 21, no. 2, pp. 1345–1360, 2024, doi: 10.1007/s13762-023-05016-2.
[21] M. Imam, S. Adam, S. Dev, and N. Nesa, "Air quality monitoring using statistical learning models for sustainable environment," *Intelligent Systems with Applications*, vol. 22, 2024, doi: 10.1016/j.iswa.2024.200333.
[22] A. A. Khadom, S. Albawi, A. J. Abboud, H. B. Mahood, and Q. Hassan, "Predicting air quality index and fine particulate matter levels in Bagdad city using advanced machine learning and deep learning techniques," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 262, 2024, doi: 10.1016/j.jastp.2024.106312.
[23] Y. Resti *et al.*, "Ensemble of naive Bayes, decision tree, and random forest to predict air quality," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 3, pp. 3039–3051, 2024, doi: 10.11591/ijai.v13.i3.pp3039-3051.
[24] S. J. Livingston, S. D. Kanmani, A. S. Ebenezer, D. Sam, and A. Joshi, "An ensembled method for air quality monitoring and control using machine learning," *Measurement: Sensors*, vol. 30, 2023, doi: 10.1016/j.measen.2023.100914.
[25] C. Yu, J. Tan, Y. Cheng, and X. Mi, "Data analysis and preprocessing techniques for air quality prediction: a survey," *Stochastic Environmental Research and Risk Assessment*, vol. 38, no. 6, pp. 2095–2117, 2024, doi: 10.1007/s00477-024-02693-4.
[26] S. Mondal, A. S. Adhikary, A. Dutta, R. Bhardwaj, and S. Dey, "Utilizing machine learning for air pollution prediction, comprehensive impact assessment, and effective solutions in Kolkata, India," *Results in Earth Sciences*, vol. 2, 2024, doi: 10.1016/j.rines.2024.100030.

BIOGRAPHIES OF AUTHORS

Dr. Ahmad Fauzi holds a bachelor's degree in informatics (S.Kom.) from STMIK ROSMA Karawang Indonesia, a master's degree in informatics (M.Kom.) from STTI Benarif Indonesia Jakarta, and a doctor of Information Technology from Gunadarma University Jakarta. Currently active as a lecturer in the undergraduate informatics program at the Universitas Buana Perjuangan Karawang. His research interest in artificial intelligence, digital image processing, expert systems, and data science. He can be contacted at email: afauzi@ubpkarawang.ac.id.

Maharina holds a master information system degree from the School of Informatics and Computer Management LIKMI, Indonesia in 2020. She received her B.Sc. (Computer Science) from Singaperbangsa Karawang University, Indonesia in 2013. She is a researcher and lecturer in the Information Systems Program within the Faculty of Computer Science at Universitas Buana Perjuangan Karawang, Indonesia. Her research interests include machine learning, deep learning, natural language processing, and Kansei engineering. She can be contacted at email: maharina@ubpkarawang.ac.id.

Jamaludin Indra is a faculty member at Universitas Buana Perjuangan Karawang, Indonesia, in the Informatics Engineering Program. Holding a master's degree in Computer Science, he specializes in research areas such as computer vision, internet of things (IoT), deep learning, and machine learning. With a strong academic background, contributions have been made to the advancement of knowledge in these fields through various research projects and publications. The focus is on bridging the gap between theoretical concepts and practical applications, particularly in solving real-world problems using advanced technologies. He can be contacted at email: jamaludin.indra@ubpkarawang.ac.id.

Ayu Ratna Juwita received her bachelor's degree from the Faculty of Computer Science, Singaperbangsa Karawang University in 2015, and her master's degree from Budi Luhur University, Jakarta, Indonesia, in 2018. She is pursuing her doctoral degree at Budi Luhur University, Jakarta, Indonesia. She is currently an active lecturer at Universitas Buana Perjuangan Karawang, Indonesia. In addition to her academic activities, she is also active in research and development in the field of information technology, specifically focusing on web-based application development, system design, software engineering, image processing, and data analysis. She can be contacted at email: ayurj@ubpkarawang.ac.id.

Agustia Hananto earned his bachelor's degree at Sony Sugema College in 2016. He received his master's degree in Computer Science (M.Kom) in 2021 at Budi Luhur University, Jakarta. In 2022, he started his doctoral studies at Asia E University until now. Currently, he is a lecturer in the Information Systems Program at the Faculty of Computer Science, Universitas Buana Perjuangan Karawang, Indonesia. His research includes machine learning, data mining, computer vision, and information systems. Some of his papers have been published in international journals and conferences from September 2019 to December 2023. He can be contacted at email: agustia.hananto@ubpkarawang.ac.id.

Euis Nurlaelasari holds a bachelor of IT and a master degree in Information System. She is a lecturer specializing in informatics engineering and teaches programming languages such as Python and Java. Her expertise includes project management, data mining, and Kansei engineering. She is passionate about guiding students in project management and software development. Her research areas of interest focus on practical applications of data science and technology management. She can be contacted at email: euis.nurlaelasari@ubpkarawang.ac.id.