

Optimization of the algorithms use ensemble and synthetic minority oversampling technique for air quality classification

Aziz Jihadian Barid¹, Hadiyanto², Adi Wibowo³

¹Master of Information System Study Program, Postgraduate School, Diponegoro University, Semarang, Indonesia

²Department of Chemical Engineering, Faculty of Engineering, Diponegoro University, Semarang, Indonesia

³Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia

Article Info

Article history:

Received Oct 23, 2023

Revised Dec 13, 2023

Accepted Dec 25, 2023

Keywords:

K-fold cross validation

K-nearest neighbors

Random forest

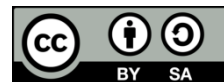
SMOTE

Support vector machine

ABSTRACT

Rapid economic development, industrialization, and urbanization in Indonesia have caused a large increase in air pollution with negative impacts on the environment and public health. The aim of this research is to use machine learning techniques to categorize air quality and generate an air quality index (AQI) using a dataset that includes six prevalent air pollutants. Next steps are preprocessing and data extraction, K-nearest neighbors (KNN) classification, support vector machine (SVM), and random forest (RF) models are implemented. Furthermore, synthetic minority oversampling technique (SMOTE) is incorporated into the ensemble learning process to improve the results. This research uses K-fold cross validation for improve classification accuracy and reduce overfitting. Research findings show that the application of SMOTE brings a significant increase in model accuracy, effectively solving the problem of imbalanced data sets. These insights provide direction for effective air quality monitoring systems and informed decision making in air pollution management.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Aziz Jihadian Barid

Master of Information System Study Program, Postgraduate School, Diponegoro University

Semarang, Indonesia

Email: azizjb100@gmail.com

1. INTRODUCTION

Since the 21st century, Indonesia has experienced significant economic development, industrialization, and urbanization [1], these factors include rapid economic growth, industrial development, and population migration to urban areas. The growth of industry and the mobility of motorized vehicles in big cities are the main causes of air pollutant emissions. Additionally, agricultural practices, forest fires and deforestation also contribute to air pollution. Air pollution has become an increasingly serious issue that impacts the ecological environment and draws worldwide attention [2]. Air quality problems have become a major strategic national concern. Generally, people frequently discuss and analyze six common air pollutants [3], Air quality issues have emerged as an important and crucial national priority. Typically, there are six air pollutants that can be released into the environment often discussed due to human activities (such as agriculture, industry, land burning, fossil fuels and vehicle emissions) as well as natural events (such as forest fires and volcanic eruptions), Pollutants. This has a negative impact on public health and causes harm to plants and animals [4]. Evaluating air quality relies on the air quality index (AQI). Ambient air quality standards (GB 3095-2012) [5] studies show exposure to significant air pollution, when inhaled, is often associated with various diseases, especially respiratory disorders [6], [7]. Under such conditions, experts can utilize computer technology, machine learning, data mining, and other tools to gather precise data from monitoring stations. This data can aid in tackling pollution and forecasting air quality.

Research using classification and prediction methods [8] has been conducted to assist environmental and health experts in determining health risks associated with specific levels of air pollution. One of the technologies that can aid in this classification is machine learning [9], [10]. Machine learning is becoming a key tool. This branch of artificial intelligence relies on algorithms and statistical models in computer systems that can learn automatically. This enables machines to generate predictions or make decisions without requiring explicit programming for these specific tasks [11]. With an emphasis on addressing air pollution problems as a long-term endeavor, the classification of air quality has emerged as a tool capable of preventing damage caused by air pollution [12]. The results are expected to provide maximum contribution to readers by presenting information in a structured and easy to understand manner. Therefore, it is essential to classify air quality in forecasts at the right time, this empowers government departments and the public to implement protective measures and prevent severe pollution incidents [13], [14]. For instance, some factories in Indonesia have been temporarily closed due to air pollution issues [15].

Numerous studies have been conducted on this subject using various classification techniques such as the Backpropagation neural network, support vector machine (SVM), K-nearest neighbors (KNN), Naive Bayes, and decision trees (DT), all of which have proven successful in predicting air quality [16]. Study in the literature has shown that researchers have primarily focused on air quality classification. However, despite reliable prediction results, there is a tendency for models to suffer from overfitting [17], one of the reasons being data imbalance, which impacts model performance [18].

Classification use CNN test results show that estimation accuracy in terms of R2 for PM2.5, PM10, and AQI based on daytime (nighttime) images reaches 76% (83%), 84% (84%), and 76% (74%) [19]. Therefore, there is a need for more appropriate methods and improved performance in air quality classification. SVM method is suggested for air quality classification, given its exceptional performance in accuracy, complexity, and problem-solving capability, as evidenced by research [20]–[22] in these studies, classification performance with high accuracy surpassing 90% is considered excellent performance.

In a study [23] the classification is based on the distance method between KNN to classify a new example. Machine learning has been created and utilized for forecasting the daily concentrations of six prevalent pollutants, it has the capability to automatically discover the optimal “Model + Hyperparameters”, utilizing several algorithms, pollutant concentrations, emission pollutants. The system integrates model analysis data from a knowledge base, serving as its foundation. It incorporates five common machine learning models (multiple linear regression (MLR), multi-layer perceptron (MLP), RF, gradient boosted decision tree (GBDT), support vector regressor (SVR)) and an ensemble model SG. A key innovation of the proposed system lies in its automatic identification of the optimal “Model + Hyperparameters”, achieved through the utilization of random CV parameters or grid search CV [24].

Ensemble methods have also been used in research [25], [26] showing that ensembles can be a solution for classification cases, as they can handle imbalanced or overfitting data. Therefore, the proposed ensemble algorithms can be a solution to imbalanced class issues. Based on conducted tests, ensembles optimize accuracy compared to other single classifiers.

Research has introduced the AQP-EDLMRA technique for automatic air quality estimation. The AQP-EDLMRA approach undergoes experimental evaluation using a series of air quality data. A comprehensive comparison demonstrates that the AQP-EDLMRA technique, employing ensemble voting, has achieved superior forecasting results [27]. Therefore, this research proposes the integration of the KNN, SVM, and random forest (RF) algorithms with ensemble learning using the synthetic minority oversampling technique (SMOTE) and use K-fold cross-validation is employed to reduce overfitting and improve the stability of air quality classification accuracy. This combination is expected to yield an optimal model for effective and efficient air quality maintenance decision-making.

2. METHOD

The methodology developed to improve the previous results [19]. Aims at classifying air quality according to the AQI. Data preprocessing and extraction are required before training the classification model. The air quality classification process used in this study is illustrated in Figure 1.

2.1. Air quality dataset

The research begins with the collection of a dataset, which is obtained from the Kaggle repository in .xlsx format. This dataset comprises around 8,000 rows of data with 6 parameters: pm10, pm2.5, so2, co, o3, and no3. The data is categorized into 5 air quality classes: good, moderate, unhealthy, very unhealthy, and hazardous, as shown in Table 1.

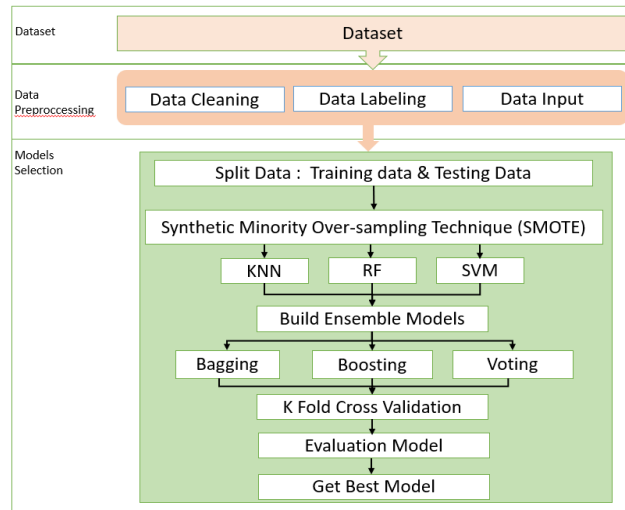


Figure 1. Research method

Table 1. AQI categories

Indeks	Category
1-50	Good
51-100	Moderate
100-199	Unhealthy
200-299	Very unhealthy
>300	Hazardous

2.2. Labeling and processing

Before the model implementation process, the dataset undergoes preprocessing stages. These stages involve removing irrelevant data, cleaning empty data, identifying and removing outliers, and addressing data imbalances. The cleaned data then undergoes transformation for ML modeling, using Scikit-learn [28].

2.3. Split data

The subsequent phase includes dividing the dataset into a training set and a testing set with an 80:20 proportion. The model is trained using the data in the training set, and the performance of the model is evaluated using the testing set. The division of data has been modified to meet the specific needs of both training and testing data. Every data point serves as input in the research procedure, utilized for both algorithmic training and testing.

2.4. SMOTE

Is a fairly popular method implemented in dealing with class imbalance in datasets. Aims to improve minority classes by creating case-made exams in minority classes. These synthetic points are added to the data set in the minority class. The artificial generation of data points is different from the multiplication method [29].

2.5. Model implementation

KNN algorithm is based on the distance between data points. It tests data by calculating the distance to training data and selects the KNN to classify the data. The algorithm considers the k data samples closest to the test sample and associates the majority class with the test sample [30].

$$d_i = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2} + (X_i - Y_i)^2 \quad (1)$$

RF is algorithm constructs numerous decision trees randomly to combine prediction results [31]. Decision trees are built by randomly selecting data [32]:

$$Entropy(S) = \sum_{i=1}^n -p_i * \log 2 p_i \quad (2)$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i) \tag{3}$$

SVM stands out as one of the most widely employed supervised learning algorithms, extensively utilized for addressing both classification and regression problems. Nevertheless, its primary application in the realm of machine learning is focused on classification tasks [33].

2.6. Build ensemble model

The prediction method based on ensemble learning is a method of combining various student bases to obtain optimal results. Usually, ensemble learning has better learning abilities compared to just relying on a single algorithm [34]. The algorithms are then combined into ensemble Figure 2, a bagging ensemble shown in Figure 2(a), boosting ensemble shown in Figure 2(b), and a voting ensemble shown in Figure 2(c).

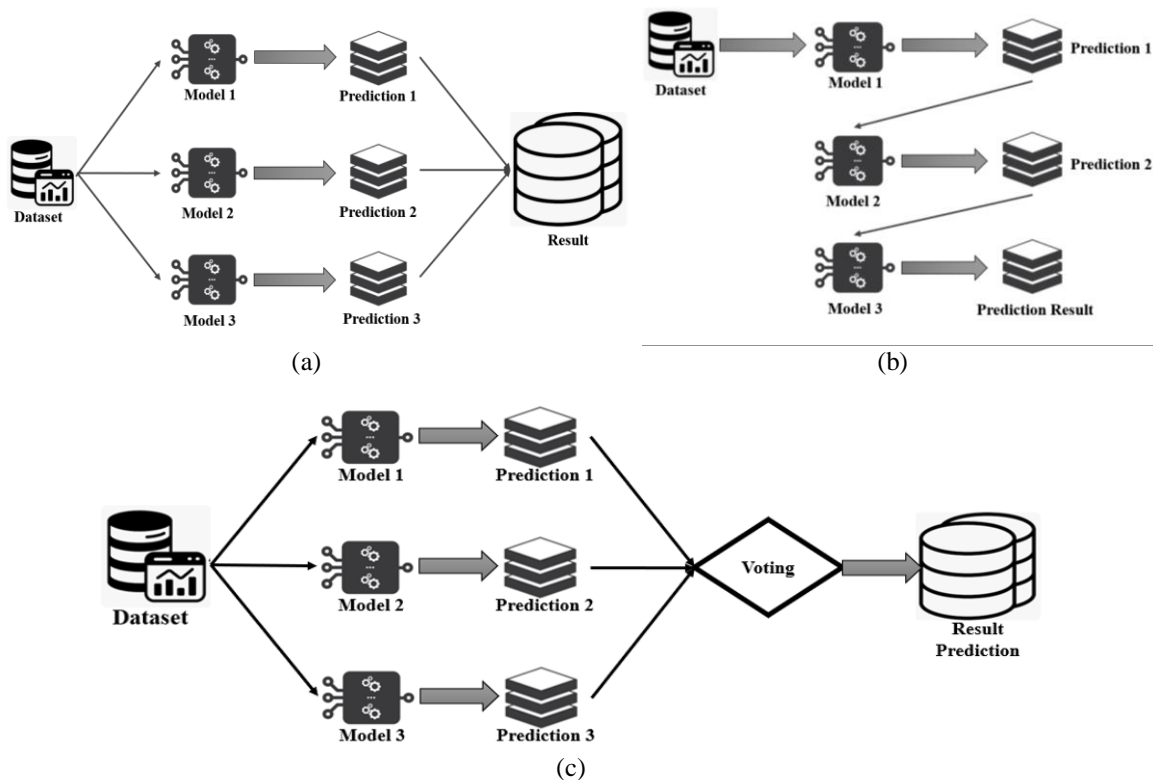


Figure 2. Ensemble model; (a) ensemble bagging, (b) ensemble boosting, and (c) ensemble voting [35]

2.7. K-fold cross validation

Cross validation is an additional method of data mining techniques that aims to obtain maximum accuracy. This method is often referred to as k-fold cross validation where k times are tried for one model with the same parameters [36]. This technique is a validation method that developed from the split validation model, where validation involves measuring training errors by testing data using test data or test data.

3. RESULTS AND DISCUSSION

In this research, there are approximately 8,527 datasets presented in Table 2, divided into an 80:20 ratio, where 80% is used as training data and 20% as testing data. In preprocessing, data imbalance is identified in the air quality dataset, as presented in Table 3. Dealing with imbalanced datasets constitutes a notable challenge in this research. To overcome this problem, the study employs the SMOTE technique to balance the samples. Table 2 contains an air quality dataset from January to December 2021, consisting of 8527 rows and 6 variables: pm10, pm2.5, SO₂, CO, O₃, and NO₂. With this amount of data, it provides a sufficiently large sample for statistical analysis and modeling.

Table 2. Dataset air quality

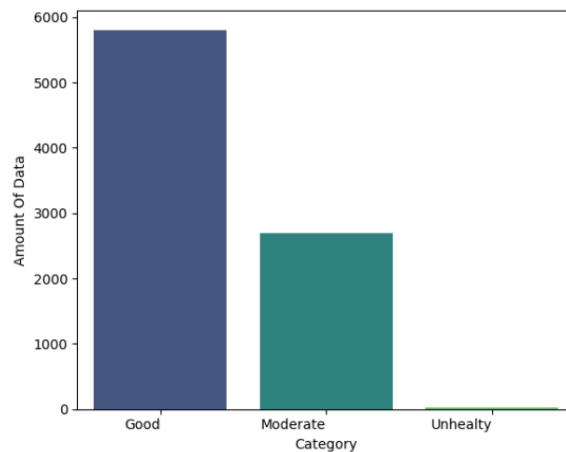
Date	Time	PM 10	PM 2.5	SO ₂	CO	O ₃	NO ₂	Category
01-01-21	01.00	9	30	21	23	36	3	Good
02-01-21	01.00	10	32	21	23	36	3	Good
03-01-21	01.00	10	70	21	24	37	3	Moderate
04-01-21	01.00	11	113	21	24	39	3	Unhealthy
05-01-21	01.00	10	32	21	23	36	3	Good

3.1. SMOTE

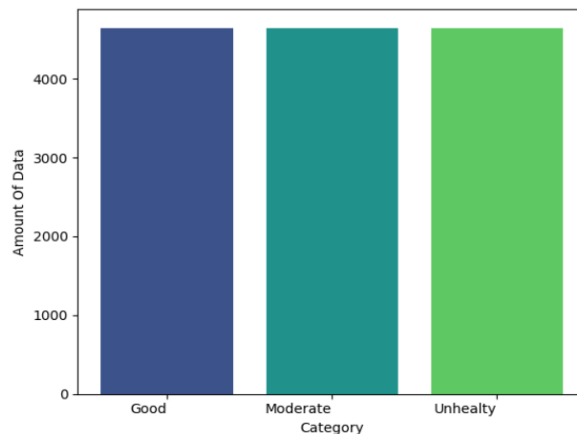
Table 3 shows that the air quality dataset exhibits significant imbalances in the number of samples between air quality categories shown in Figure 3. Particularly, the ‘good’ category has a much larger number of samples than ‘moderate’ and ‘unhealthy,’ as shown in Figure 3(a). Such missing data can lead to incorrect predictive models and degrade model performance [37]. After applying the SMOTE technique, the dataset has undergone oversampling for minority categories, creating duplicate data to ensure a balanced number of samples in each category, as presented in Figure 3(b).

Table 3. Imbalance dataset

No	Category	Dataset before SMOTE	Dataset after SMOTE
1	Good	5,807	4,646
2	Moderate	2,695	4,646
3	Unhealthy	25	4,646



(a)



(b)

Figure 3. Displays the dataset count, with (a) representing the dataset before SMOTE and (b) representing the dataset after SMOTE

3.2. Correlation

The relationship between variables in the analyzed dataset is assessed using pearson’s correlation. Figure 4 illustrates a positive correlation between PM2.5 and PM10 variables, with a correlation coefficient of 0.6. When the PM2.5 concentration increases, there is a tendency for the PM10 concentration to increase as well, and vice versa. This indicates a positive correlation between small PM2.5 particles and larger PM10 particles in this dataset. An increase in PM2.5 concentration is not always a direct cause of increased PM10 concentration, and vice versa.

3.3. Model evaluation

This model evaluation involves the utilization of a confusion matrix, a powerful tool that provides a detailed breakdown of the classification results. The confusion matrix allows us to analyze the true positive, true negative, false positive, and false negative predictions made by the models. By these metrics, we gain a deeper understanding of the accuracy, precision, recall, and F1-score of each algorithm.

3.4. Model evaluation using K-fold cross validation

Validation showcases a comprehensive evaluation of overall accuracy and performance through a 5-fold cross-validation, measured in terms of accuracy. Table 4 provides insight into the effectiveness of the classification model during training for each fold. The accompanying graph illustrates the model’s accuracy in both training and validation, serving as a means to address and mitigate potential overfitting.

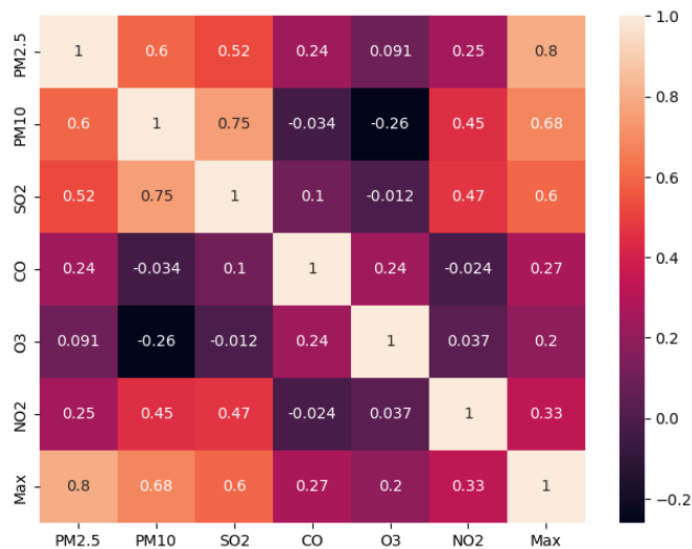


Figure 4. Correlation between variables

Table 4. K-fold cross validation

Cross validation	Accuracy					
	KNN	SVM	RF	Bagging	Boosting	Voting
Fold 1	0.9807	0.9039	0.9947	0.9941	0.9947	0.9871
Fold 2	0.9877	0.8968	0.9994	0.9994	0.9994	0.9930
Fold 3	0.9871	0.8979	0.9982	0.9971	0.9988	0.9918
Fold 4	0.9801	0.8997	0.9971	0.9947	0.9959	0.9871
Fold 5	0.9865	0.8921	0.9977	0.9977	0.9982	0.9874

3.5. Model comparison

There is a comparison of accuracy between the dataset with imbalanced data and the dataset with oversampling using SMOTE, resulting in a significant increase in accuracy, as shown in Table 5 and visualized in Figure 5. Table 5 and Figure 5 demonstrate that oversampling using SMOTE significantly improves model accuracy. The importance of proper data preprocessing, especially for imbalanced datasets, can enhance performance compared to imbalanced data [38].

Table 5. Comparison evaluated model

No	Algorithm	Accuracy before SMOTE	Accuracy after SMOTE
1	KNN	83.56%	98.44%
2	SVM	85.15%	89.81%
3	RF	94.35%	99.74%
4	Ensemble bagging	94.66%	99.66%
5	Ensemble boosting	94.23%	99.74%
6	Ensemble voting	88.06%	98.77%

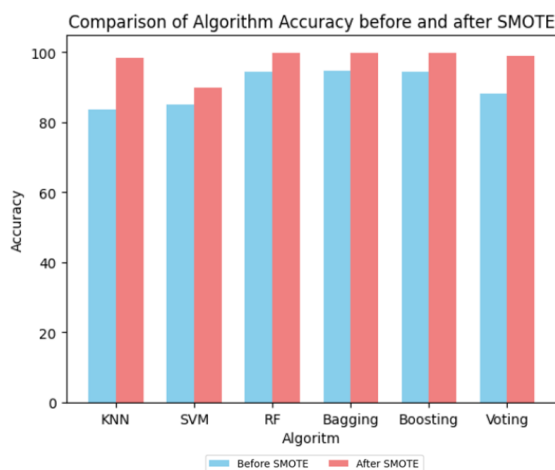


Figure 5. Model evaluation visualization

4. CONCLUSION

This paper employs machine learning techniques to classify air quality. Researchers also utilize a dataset collected from Kaggle. At the outset of this study, dataset imbalances are observed, and the SMOTE technique is applied to address this issue, balancing the dataset. We combine KNN, SVM, and RF algorithms with three types of ensemble methods: bagging, boosting, and voting. The algorithms are implemented using Python to achieve the best results that can deliver best performance. The primary goal of this research is to examine the optimized aspects of machine learning algorithms targeted for air quality classification. Experimental results show that ensemble performance is superior to single algorithms, and classification accuracy significantly improves when the dataset is balanced using the SMOTE technique compared to an imbalanced dataset. Therefore, this research can serve as a reference and provide accurate information for quick and responsive decision-making in air pollution management. For method development, an emphasis can be placed on adding a segmentation step in the preprocessing process, focusing on air quality features. The findings from this study can be used as a basis for the development of a more efficient and accurate air quality monitoring system, contributing positively to decision-making related to air quality management and improvement.

ACKNOWLEDGEMENTS




The data used in preparing this article was obtained from Air Quality in Yogyakarta, Indonesia (<https://www.kaggle.com/datasets/adhang/air-quality-in-yogyakarta-indonesia-2021>).

REFERENCES




- [1] E. Puspitawati, "Indonesian industrialization and industrial policy: Peer learning from china's experiences," in Associate, Research, 2021.
- [2] Q. Wang *et al.*, "Contribution of regional transport to the black carbon aerosol during winter haze period in Beijing," *Atmospheric Environment*, vol. 132, pp. 11–18, 2016, doi: 10.1016/j.atmosenv.2016.02.031.
- [3] H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," *Applied Sciences (Switzerland)*, vol. 9, no. 19, Oct. 2019, doi: 10.3390/app9194069.
- [4] I. Gutiérrez-Avila *et al.*, "Prediction of daily mean and one-hour maximum PM_{2.5} concentrations and applications in Central Mexico using satellite-based machine-learning models," *Journal of Exposure Science and Environmental Epidemiology*, vol. 32, no. 6, pp. 917–925, Nov. 2022, doi: 10.1038/s41370-022-00471-4.
- [5] China Ministry of Environmental Protection, "Ambient air quality standards," China Environmental Science Press, 1970.

- [6] T. H. Bhat, G. Jiawen, and H. Farzaneh, "Air pollution health risk assessment (Ap-hra), principles and applications," *International Journal of Environmental Research and Public Health*, vol. 18, no. 4, pp. 1–29, 2021. doi: 10.3390/ijerph18041935.
- [7] A. van Donkelaar *et al.*, "Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application," *Environ Health Perspect*, vol. 118, no. 6, pp. 847–855, Jun. 2010, doi: 10.1289/ehp.0901623.
- [8] P. A. Rani and D. V. Sampathkumar, "A novel artificial intelligence algorithm for predicting air quality by analysing the pollutant levels in air quality data in tamilnadu," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 5, Sep. 2023, doi: 10.1016/j.prime.2023.100234.
- [9] Y. Li, Z. Sha, A. Tang, K. Goulding, and X. Liu, "The application of machine learning to air pollution research: A bibliometric analysis," *Ecotoxicology and Environmental Safety*, vol. 257, Jun. 2023, doi: 10.1016/j.ecoenv.2023.114911.
- [10] E. Gladkova and L. Saychenko, "Applying machine learning techniques in air quality prediction," in *Transportation Research Procedia*, Elsevier B.V., 2022, pp. 1999–2006, doi: 10.1016/j.trpro.2022.06.222.
- [11] Y. S. Su, Y. D. Lin, and T. Q. Liu, "Applying machine learning technologies to explore students' learning features and performance prediction," *Frontiers in Neuroscience*, vol. 16, Frontiers Media S.A., Dec. 22, 2022, doi: 10.3389/fnins.2022.1018005.
- [12] Z. Yang and J. Wang, "A new air quality monitoring and early warning system: air quality assessment and air pollutant concentration prediction," *Environ Res*, vol. 158, pp. 105–117, 2017, doi: 10.1016/j.envres.2017.06.002.
- [13] X. Wang, M. Wang, X. Liu, Y. Mao, Y. Chen, and S. Dai, "Surveillance-image-based outdoor air quality monitoring," *Environmental Science and Ecotechnology*, p. 100319, Sep. 2023, doi: 10.1016/j.ese.2023.100319.
- [14] S. D. Vito, G. D. Francia, E. Esposito, S. Ferlito, F. Formisano, and E. Massera, "Adaptive machine learning strategies for network calibration of IoT smart air quality monitoring devices," *Pattern Recognit Lett*, vol. 136, pp. 264–271, Aug. 2020, doi: 10.1016/j.patrec.2020.04.032.
- [15] Asia News Network, "Indonesia shuts down factories, to spray mist to reduce air pollution," <https://asianews.network/indonesia-shuts-down-factories-to-spray-mist-to-reduce-air-pollution/>
- [16] Y. Zhao, G. Deng, L. Zhang, N. Di, X. Jiang, and Z. Li, "Based investigate of beehive sound to detect air pollutants by machine learning," *Ecol Information*, vol. 61, Mar. 2021, doi: 10.1016/j.ecoinf.2021.101246.
- [17] F. Biancofiore *et al.*, "Recursive neural network model for analysis and forecast of PM10 and PM2.5," *Atmospheric Pollution Research*, vol. 8, no. 4, pp. 652–659, Jul. 2017, doi: 10.1016/j.apr.2016.12.014.
- [18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.
- [19] P. Y. Kow, I. W. Hsia, L. C. Chang, and F. J. Chang, "Real-time image-based air quality estimation by deep learning neural networks," *Journal of Environmental Management*, vol. 307, Apr. 2022, doi: 10.1016/j.jenvman.2022.114560.
- [20] N. S. A. Yasmin, N. A. Wahab, A. N. Anuar, and M. Bob, "Performance comparison of SVM and ANN for aerobic granular sludge," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 8, no. 4, pp. 1392–1401, Dec. 2019, doi: 10.11591/eei.v8i4.1605.
- [21] A. S. Handayani, S. Soim, T. E. Agusdi, and N. L. Husni, "Air quality classification using support vector machine," *Computer Engineering and Applications*, vol. 10, no. 1, 2021, doi: 10.18495/comengapp.v10i1.350.
- [22] L. C. Wu *et al.*, "Detection of american football head impacts using biomechanical features and support vector machine classification," *Scientific Reports*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/s41598-017-17864-3.
- [23] E. G. Dragomir, "03-108 seria matematică-informatică-fizică air quality index prediction using K-nearest neighbor technique," *Bulletin of PG University of Ploiesti, Series Mathematics, Informatics, Physics, LXII 1.2010*, pp. 103-108, 2010.
- [24] H. Ke *et al.*, "Development and application of an automated air quality forecasting system based on machine learning," *Science of the Total Environment*, vol. 806, Feb. 2022, doi: 10.1016/j.scitotenv.2021.151204.
- [25] A. K. Biswas, R. Seethalakshmi, P. Mariappan, and D. Bhattacharjee, "An ensemble learning model for predicting the intention to quit among employees using classification algorithms," *Decision Analytics Journal*, vol. 9, Dec. 2023, doi: 10.1016/j.dajour.2023.100335.
- [26] G. Ngo, R. Beard, and R. Chandra, "Evolutionary bagging for ensemble learning," *Neurocomputing*, vol. 510, pp. 1–14, Oct. 2022, doi: 10.1016/j.neucom.2022.08.055.
- [27] S. Sivanesh, G. Mani, S. Venkatraman, and R. Nandhini, "Air quality prediction using ensemble voting based deep learning with mud ring algorithm for intelligent transportation systems," *Global NEST Journal*, Apr. 2023, doi: 10.30955/gnj.004810.
- [28] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
- [29] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Applied Soft Computing*, vol. 143, Elsevier Ltd, Aug. 01, 2023, doi: 10.1016/j.asoc.2023.110415.
- [30] P. Arora, N. Periwal, Y. Goyal, V. Sood, and B. Kaur, "iIL13Pred: improved prediction of IL-13 inducing peptides using popular machine learning classifiers," *BMC Bioinformatics*, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s12859-023-05248-6.
- [31] C. Feng, Y. Tian, X. Gong, X. Que, and W. Wang, "MCS-RF: mobile crowdsensing-based air quality estimation with random forest," *International Journal of Distributed Sensor Networks*, vol. 14, no. 10, Oct. 2018, doi: 10.1177/1550147718804702.
- [32] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, Jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.
- [33] S. Akhter and J. H. Miller, "BaPreS: a software tool for predicting bacteriocins using an optimal set of features," *BMC Bioinformatics*, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s12859-023-05330-z.
- [34] J. Wu, L. Kong, Z. Cheng, Y. Yang, and H. Zuo, "RUL prediction for lithium batteries using a novel ensemble learning method," *Energy Reports*, vol. 8, pp. 313–326, Nov. 2022, doi: 10.1016/j.egyrs.2022.10.298.
- [35] Joseph Rocca, "Ensemble methods: bagging, boosting and stacking," towardsdatascience.com.
- [36] A. M. P. Chacón, I. S. Ramírez, and F. P. G. Márquez, "K-nearest neighbour and K-fold cross-validation used in wind turbines for false alarm detection," *Sustainable Futures*, vol. 6, Dec. 2023, doi: 10.1016/j.sfr.2023.100132.
- [37] K. Singh, U. K. Lihore, and N. Agrawal, "Survey on different tumour detection methods from MR images," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 5, pp. 589-594, 2017.
- [38] S. P. Potharaju, M. Sreedevi, V. K. Ande, and R. K. Tirandasu, "Data mining approach for accelerating the classification accuracy of cardiocography," *Clin Epidemiol Glob Health*, vol. 7, no. 2, pp. 160–164, Jun. 2019, doi: 10.1016/j.cegh.2018.03.004.




BIOGRAPHIES OF AUTHORS

Aziz Jihadian Barid    is a student of the Master of Information Systems study program at Diponegoro University Class of 2022, received a bachelor's degree in 2017. Has been a sinta-indexed publication. He can be contacted at email: azizjb100@gmail.com.



Prof. Dr. Ir. Hadiyanto S.T., M.Sc., IPU    received an Ir degree in Chemical Engineering from Diponegoro University (1998). Continuing his Masters-S3 education at Wageningen University, the Netherlands in the field of Biotechnology. He is currently Deputy Dean II of the Postgraduate School, Diponegoro University is an outstanding lecturer. Published research results in the international proceedings of the e3s web conference Volume 73, 2018. Due to his achievements and perseverance he received the highest academic title as Professor at the Faculty of Engineering and was inaugurated on November 30 2017. Has received an h-index of 28 in Scopus. And h-index 35 on Google Scholar. He can be contacted at email: hadiyanto@lecturer.undip.ac.id and hadiyanto@che.undip.ac.id.



Dr. Eng Adi Wibowo, S.Si, M.Kom    he served as secretary of the Department of Informatics at Diponegoro University. He is also a lecturer at the Faculty of Science and Mathematics. His expertise is in the fields of DNA Nanotechnology, Robotics, Artificial Intelligence. It has got an h-index of 13 in Scopus. A number of more than 90 international publications. He can be contacted at email: bowo.adi@live.undip.ac.id.