

KNN-MVO-SMOTE Algorithm for Air Quality Imbalanced Data Classification

Muhammad Miftahur Rizky¹, Muhammad Itqan Mazdadi², Muliadi Muliadi³,
Mohammad Reza Faisal⁴, Fatma Indriani⁵, Hasri Akbar Awal Rozaq⁶, Oktay Yıldız⁷

^{1,2,3,4,5}Faculty of Mathematics and Natural Science, Department of Computer Science, Lambung Mangkurat University, Kalimantan, Indonesia

⁶Graduate School of Informatics, Department of Computer Science, Gazi University, Ankara, Türkiye

⁷Faculty of Engineering, Department of Computer Engineering, Gazi University, Ankara, Türkiye

Article Info

Article history:

Received Jul 16, 2025

Revised Nov 11, 2025

Accepted Dec 22, 2025

Keywords:

Accuracy

Data Imbalance

Environmental Analysis

Machine Learning

Optimization

ABSTRACT

Air pollution represents a critical global challenge requiring accurate monitoring and prediction systems for effective public health protection. This study proposes an enhanced machine learning approach for air quality classification by integrating Multi-Verse Optimization (MVO) with Synthetic Minority Over-sampling Technique (SMOTE) in the K-Nearest Neighbors (KNN) algorithm. The research addresses two fundamental challenges in air quality prediction: class imbalance in datasets and optimal parameter selection. Using air quality data from DKI Jakarta monitoring stations in 2021, the study systematically evaluated four KNN variants: baseline KNN, KNN with SMOTE, KNN with MVO, and the proposed KNN-MVO-SMOTE combination. The comprehensive optimization process evaluated 504 parameter combinations through 5-fold cross-validation, identifying optimal configurations with $k=9$, distance weighting, and Manhattan distance metric. SMOTE effectively balanced the originally imbalanced dataset (165 BAIK, 344 SEDANG, 265 TIDAK SEHAT samples) to 344 samples each class using 7 neighbors and minority sampling strategy. The proposed KNN-MVO-SMOTE method achieved superior performance with 89.41% accuracy, 88.70% precision, and 91.07% recall, significantly outperforming previous approaches including standalone KNN (88.4%), Support Vector Regression (88.0%), Random Forest (83.0%), and Naive Bayes (81.0%). The exceptional recall performance is particularly crucial for detecting unhealthy air conditions, making this approach highly reliable for environmental monitoring applications and supporting effective air quality management systems.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Muhammad Itqan Mazdadi,
Faculty of Mathematics and Natural Science
Department of Computer Science
Lambung Mangkurat University, Kalimantan, Indonesia
Email: mazdadi@ulm.ac.id

1. INTRODUCTION

Air pollution, a complex global issue, is defined by geographical and temporal variability [1]. These factors influence the composition and sources of pollution, creating challenges in their effective monitoring and management. According to [2], the impact of air pollution is significant on people's lives and can threaten human survival. Recent studies indicate that air pollution causes approximately 7 million premature deaths annually worldwide, making accurate air quality

prediction crucial for public health protection [3]. The government has set the air quality index (AQI) standard as a method to monitor air quality daily, focusing on five significant pollutants, namely SO₂, NO₂, CO, O₃, and PM₁₀ [4], [5]. However, traditional monitoring approaches often struggle with the inherent complexity and non-linear relationships present in atmospheric data, particularly when dealing with imbalanced datasets where certain air quality categories are significantly underrepresented [6]. An approach capable of processing and analyzing large volumes of data is essential to interpret and utilize AQI data effectively. This highlights the need for technological advancements such as artificial intelligence (AI), particularly machine learning (ML), which is the cutting-edge layer within AI, dedicated to pattern recognition and predictive analytics.

Machine learning offers innovative solutions that enable large-scale data analysis to identify patterns and make accurate predictions about air quality [7], [8]. By combining mathematical models and sophisticated algorithms, machine learning plays a vital role in the air quality determination process, managing big data, and recognizing patterns [9]. These algorithms' adaptive capabilities can expand our understanding and facilitate more informed data-driven decision-making. Several previous studies have demonstrated that machine learning can effectively identify air quality. Research [10] stated that machine learning algorithms could support air quality assessment with high accuracy. Further research [11] shows that machine learning is very effective in classifying air quality, with an accuracy rate of 90.56%. Despite these promising results, most existing studies suffer from methodological limitations, including inadequate handling of class imbalance, limited hyperparameter optimization, and insufficient validation strategies that may lead to overly optimistic performance estimates. This underscores machine learning's efficacy in air quality classification. One of the popular algorithms in this field is the K-Nearest Neighbors (KNN) algorithm, which is a model that utilizes a predefined number of nearest neighbors (K value) to classify new data points based on their similarity to training data [12]–[14].

However, the effectiveness of KNN is highly dependent on the quality and balance of the data used, which makes addressing data imbalances crucial [15]. Air quality datasets commonly exhibit severe class imbalance, where "Good" air quality instances are often rare compared to "Moderate" or "Unhealthy" categories, leading to biased models that perform poorly on minority classes. Research has shown that data balancing can improve classification accuracy by about 3% [16]. The Synthetic Minority Over-sampling Technique (SMOTE) is an effective method for this purpose, synthesizing new samples from the minority class to balance the dataset [17]–[19]. Additionally, optimizing the K value in KNN is essential to avoid biased results, with Multi-Verse Optimization (MVO) providing a novel method to achieve this by optimizing based on the inflation rate values of different 'universes' in the algorithm [20], [21]. While MVO has shown promise in various optimization tasks, its effectiveness in the specific context of air quality classification with imbalanced data remains understudied.

To address these gaps, this study presents a systematic evaluation of KNN variants for air quality classification, specifically comparing: (1) baseline KNN, (2) KNN with SMOTE for handling class imbalance, (3) KNN with MVO for parameter optimization, and (4) the proposed KNN-MVO-SMOTE combining both techniques. Integrating SMOTE with MVO in a KNN setting creates a potentially robust framework to address two significant obstacles in air quality classification: data imbalance and parameter optimization. Our comprehensive experimental design employs stratified cross-validation and multiple evaluation metrics to ensure robust and generalizable results, addressing the methodological limitations identified in previous studies. By tailoring this approach, this research aims to harness the power of ML in addressing the complex variability of air pollution, expanding our ability to manage its impact on the environment and public health. The main contributions of this work include: (1) a systematic comparison of KNN variants for air quality classification, (2) empirical evidence of the individual and combined effects of SMOTE and MVO optimization, and (3) a robust evaluation framework that can guide future research in environmental machine learning applications. As such, this research opens up new potential in using ML for environmental analysis. It significantly contributes to the global fight against air pollution, demonstrating the importance of technological innovation in addressing contemporary environmental issues.

2. RESEARCH METHOD

The research utilizes a machine learning strategy, particularly emphasizing supervised learning, and applies the classification method using the KNN algorithm. This research highlights the optimization of the K value through the MVO approach and implements SMOTE to address class imbalance issues. The stages in this research can be observed through the flowchart in Figure 1.

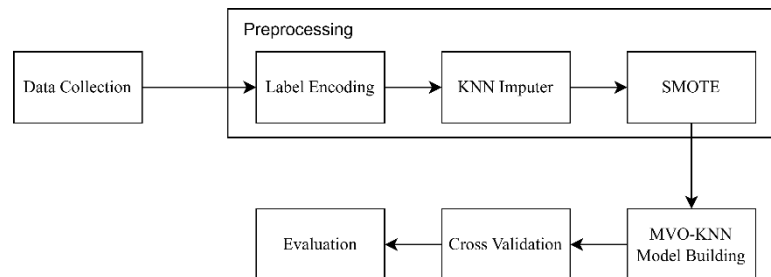


Figure 1. Flow of Research

2.1. Data Collection

This study utilizes comprehensive air quality monitoring data from the Air Pollution Standard Index (ISPU - Indeks Standar Pencemaran Udara) in DKI Jakarta Province for the year 2021. The dataset was sourced from the official Indonesian environmental monitoring system and made publicly available through the Kaggle platform to ensure data accessibility and reproducibility [20]. The complete dataset can be accessed via <https://www.kaggle.com/datasets/yunaperdana/indeks-standar-pencemaran-udara-spku-dkijakarta>. This dataset represents a significant improvement over previous studies by incorporating multi-location monitoring data, providing a more comprehensive view of Jakarta's air quality patterns.

The dataset comprises 775 daily air quality measurements collected from five strategically distributed monitoring stations across Jakarta: Lubang Buaya (East Jakarta), Kebon Jeruk (West Jakarta), Kelapa Gading (North Jakarta), Bundaran HI (Central Jakarta), and Jagakarsa (South Jakarta). This multi-station approach ensures geographical diversity and captures the spatial variability of air pollution across different urban zones, including residential, commercial, and mixed-use areas. The dataset includes 11 distinct features encompassing temporal information, pollutant concentrations, and categorical classifications, as detailed in Table 1.

Table 1. Dataset's Information

Features	Data Types	Descriptions
Date	Date Time	attribute
pm10	numeric	attribute
pm25	numeric	attribute
so2	numeric	attribute
co	numeric	attribute
o3	numeric	attribute
no2	numeric	attribute
max	numeric	attribute
critical	numeric	attribute
location	numeric	attribute
category	numeric	label

The selection of these monitoring stations follows the Indonesian National Air Quality Monitoring Network standards, ensuring data quality and consistency. Each station employs standardized measurement protocols and calibrated equipment, providing reliable and comparable measurements across all locations. The temporal coverage of one full year (2021) captures seasonal variations and different meteorological conditions, essential for developing robust classification models. This air quality data analysis will be conducted using machine learning techniques, with

the dataset including essential attributes such as the date and time of data collection, PM10 and PM2.5 particle concentrations indicating particle size in micrometers, as well as levels of gaseous pollutants such as sulfur dioxide (SO₂), carbon monoxide (CO), ozone (O₃), and nitrogen dioxide (NO₂). Other attributes include the maximum value of the pollutant concentration (max), the pollutant with the highest concentration in a given period (critical), the numerical classification of the air quality (category), and the location of the numerically represented data collection (location). This analytical approach will utilize time and attribute data to reveal patterns, trends, and relationships in air quality over time and location.

2.2. Data Preprocessing

Data preprocessing is crucial before proceeding to subsequent phases. The gathered data must undergo this initial process, which includes several steps such as label encoding, managing missing values, balancing datasets, and segmenting data for training and testing purposes. All data preprocessing and model training tasks were performed on Google Colab using Python 3.10, a robust and adaptable computational platform.

2.2.1. Label Encoding

In machine learning practice, some models cannot process class labels in text form, so converting such labels to a numerical format becomes an important step. A commonly used technique in this transformation is Label Encoding, which effectively converts text labels into quantitative identifiers, allowing machines to read and run models more smoothly [22]. By applying Label Encoding, each class label is assigned a unique integer value, ensuring that categorical data can be interpreted by machine learning algorithms [23]. As a practical example of the application of Label Encoding, in size classification, the label 'small' is changed to 1, 'medium' to 2, and 'large' to 3, thus simplifying the representation of categorical data into a format that machines can efficiently process and understand [24]. This process not only facilitates the accuracy and efficiency of the model's predictions [25].

2.2.2. KNNImputer

The K-Nearest Neighbors Imputation (KNNImputer) method represents one of the most innovative data imputation techniques, adopting the well-known K-Nearest Neighbors (KNN) approach to address the problem of missing values in a dataset. Specifically, this method identifies the K nearest neighbors with complete data for each sample missing one or more values. Once the most similar neighbors are identified, the missing values are calculated and imputed by considering and analyzing the existing values of these neighbors. Thus, KNNImputer enables the maintenance of dataset integrity by filling in data gaps based on pre-existing patterns and relationships, ensuring that imputation is performed consistently and logically according to the existing data structure [23]. This makes KNNImputer an invaluable solution for handling missing values, especially in cases where complex data patterns and relationships between variables play a crucial role in predictive analysis and statistical modeling.

2.2.3. Synthetic Minority Oversampling Technique

The Synthetic Minority Oversampling Technique (SMOTE) represents a significant evolution in oversampling techniques, introducing a more sophisticated and efficient strategy for addressing the problem of class imbalance in datasets. Smartly, the SMOTE method performs a k-nearest neighbors search (i.e., finding k-nearest neighbors) for each sample in the minority class. This process is then followed by creating synthetic data samples, which are adjusted based on the desired proportion between minority samples and randomly selected k-nearest neighbors [24]. Uniquely, the SMOTE algorithm does not simply increase the number of minority samples randomly but creates new samples logically integrated with the existing data structure, based on Euclidean distance calculations, whose formula follows.

$$dist = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \dots + (X_n - Y_n)^2} \quad (1)$$

This allows SMOTE to prevent overfitting problems often arising from traditional oversampling approaches. Based on research [24], the general formula used by the SMOTE algorithm can be explained through the notation $SMOTE(X, N, k)$, where X represents minority class samples, N indicates the percentage of synthetic samples that will be created, and k refers to the number of nearest neighbors that identified for each minority sample. This confirms SMOTE's effectiveness in navigating the challenges of imbalanced datasets, making it one of the most respected and widely adopted oversampling techniques in the data science and machine learning communities.

2.2.4. Cross Validation

To ensure reliable and unbiased evaluation of the proposed KNN variants, this study employs stratified k-fold cross-validation as the primary evaluation methodology. Stratified k-fold cross-validation extends the standard k-fold approach by ensuring that the relative class frequencies are approximately preserved in each fold [26]. This technique is particularly crucial for air quality classification tasks, where class imbalance is inherent to the problem domain. The stratified approach prevents the creation of folds that lack representation from minority classes, which is essential when dealing with imbalanced environmental datasets [27].

The implementation utilizes 5-fold stratified cross-validation, where the dataset is partitioned into 5 folds while maintaining the original class distribution in each partition. In each iteration, 4 folds serve as training data while the remaining fold functions as validation data, with this process repeated 5 times [28]. To prevent data leakage, SMOTE oversampling is applied exclusively within the cross-validation loop, generating synthetic samples only from the training folds while keeping validation folds untouched. The following presents a simulation of cross-validation demonstrating how data is divided and evaluated in each fold to ensure robust and unbiased model evaluation.

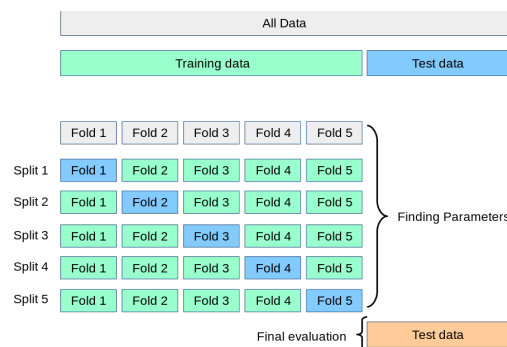


Figure 2. Cross Validation Simulation

The final performance metrics represent mean values across all folds, accompanied by standard deviations to quantify variance, enabling fair comparison between the four KNN variants under identical evaluation conditions. This cross-validation framework provides more reliable estimates than single train-test splits, particularly important when assessing model confidence intervals [29].

2.3. MVO-KNN

K-Nearest Neighbors (KNN) is a supervised learning method used to classify [8]. KNN is a learning model that works by checking the K value set on the test data and looking for nearest neighbors based on the training data, so it is often referred to as a "Lazy" algorithm. One of the obstacles in KNN classification is finding a K value that only depends on the nearest neighbors [30]. KNN works by looking for the closest groups in the training data and grouping the testing data into groups with the highest similarity [31]. The Euclidean distance method is often used to calculate the distance between training data and testing data, and this method can be used to find the distance between testing data and training data according to equation 1 [32]. After calculating the distance between the new vector and all training data vectors, the next step is to take the K

nearest neighbors. Classification is then determined based on these points [33]. Then, KNN will later be combined with MVO to refine the K value.

Multi-Verse Optimization (MVO) is an optimization method introduced by Seyedali Mirjalili in 2015. MVO differentiates itself from other methods with its mechanism for merging, moving, or expanding solutions during optimization. This technique can increase the K value in KNN by navigating the multiverse space to find the optimal solution. With solid adaptability and exploration capabilities, MVO provides a practical approach to optimization [20]. The main goal of MVO is to find a global optimal solution for complex optimization problems [34]. This algorithm is inspired by three concepts: Black Hole, White Hole, and Wormhole [35]. This concept is grounded in the theory that matter in the universe is modeled by transitioning from a white hole to a black hole via a wormhole [36].

In MVO, these three elements are presented mathematically to design optimal techniques by describing the process of teleportation and exchange of objects between universes through wormholes and white/black holes. The principle of universe inflation is applied in MVO by considering the inflation rate as an indicator of solution suitability [37]. According to [38], the inflation rate influences the probability of the formation of black and white holes, and there is a possibility that wormholes move objects randomly toward the best universe, regardless of the inflation rate. The following is the formula for MVO.

$$U = \begin{matrix} X_1^1 & \dots & X_1^d \\ \dots & \dots & \dots \\ X_n^1 & \dots & X_n^d \end{matrix} \quad (2)$$

$$X_i^j = \begin{cases} X_k^j r_1 < NI(U_i) \\ X_k^j r_1 \geq NI(U_i) \end{cases} \quad (3)$$

In Equation 2, where d is the number of parameters (variables), and n is the number of universes as candidate solutions. In Equation 3, X_i^j is the parameter representation of universe i , with ui as the universe and $NI(U_i)$ as the normalized inflation rate. r_1 is a random variable taken from the range $[0,1]$, X_k^j while indicating the parameter j of the universe k chosen using the Roulette Wheel Mechanism [39].

MVO involves calculating two parameters to regulate how often and to what extent the solution is changed: Wormhole Existence Probability (*WEP*) in equation 4 and Traveling Distance Rate (*TDR*) in equation 5. *WEP* and *TDR* are coefficients that regulate the probability of the existence of wormholes in the universe. These two coefficients need to be increased linearly during the iteration to emphasize exploitation in the optimization process. The degree of travel distance also plays a role in determining how far objects can be teleported by wormholes around the best universes that have been discovered so far. In contrast to *WEP*, *TDR* is improved during iterations to improve more precise local searches around the best universe found [20].

$$WEP = a + t * \left(\frac{b-a}{T}\right) \quad (4)$$

$$TDR = 1 - \frac{t^{\frac{1}{P}}}{T^{\frac{1}{P}}} \quad (5)$$

By using the equation 6 mathematical model, wormholes can connect to the best universe. In this equation, X_j represents the best parameter value in the current universe j , while *TDR* and *WEP* are variable coefficients. lb_j and ub_j denote the lower and upper bounds of universe j , while r_2 , r_3 , and r_4 are random.

$$X_i^j = \begin{cases} (r_2 < WEP) \begin{cases} (r_3 < 0.5) X_j + TDR \times ((ub_j - lb_j) \times r_4 + lb_j) \\ (r_3 > 0.5) X_j - TDR \times ((ub_j - lb_j) \times r_4 + lb_j) \end{cases} \\ (r_2 > WEP) X_i^j \end{cases} \quad (6)$$

2.4. Evaluation

A confusion matrix is an evaluation method used in classification to measure the performance of a model by comparing the class predictions produced by the model with the actual

class of the data [38]. The confusion matrix provides more detailed information about the performance of a classification model than a single metric, such as accuracy, because it describes the model's ability to classify each class separately. Therefore, the confusion matrix is an important evaluative tool for pinpointing the strengths and weaknesses of classification models [39]. The confusion matrix consists of four primary cells: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

Table 2. Confusion Matrix

Actual Class	Prediction Class	
	Class=Yes	Class=No
Class=Yes	TP	FN
Class=No	FP	TN

2.4.1. Accuracy

Accuracy is a measure that shows how precisely the model classifies all data samples, calculated as the ratio between the number of correct predictions (True Positive and True Negative) and the total number of samples. Accuracy can be calculated using the following equation.

$$Acc = \frac{TP+TN}{TP+FP+TN+FN} \quad (7)$$

2.4.2. Precision

Precision quantifies the accuracy of positive predictions, determined by the proportion of True Positives relative to the total positive predictions (True Positive plus False Positive). It can be calculated using the equation below.

$$Prec = \frac{TP}{TP+FP} \quad (8)$$

2.4.3. Recall

Recall is also known as sensitivity, measuring how many positive classes are successfully predicted correctly by the model, calculated as the ratio between True Positive and the number of all positive classes (True Positive and False Negative). Recall can be calculated using the following equation.

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

2.4.4. F1-Score

F1-score is a composite metric that combines precision and recall. It provides the overall performance of a classification model by considering the balance between the two, calculated as the harmonic mean of precision and recall. F-Measure can be calculated in the equation as follows.

$$F1 - Score = 2 * \frac{Prec*Recall}{Prec+Recall} \quad (10)$$

3. RESULTS AND DISCUSSION

In this section, the research results are explained, and a comprehensive discussion is given. The results can be presented in figures, graphs, tables, and other formats that make the reader understand easily [2, 5]. The discussion can be divided into several sub-chapters.

3.1. Preprocessing Result

Several attributes, including tanggal (date), stasiun (station), critical, and max, were omitted due to their limited relevance to the focused analysis of air pollution variables. The kategori feature, originally containing BAIK (good), SEDANG (moderate), and TIDAK SEHAT (unhealthy) values, was converted into binary values 0, 1, and 2 to simplify the model's computational demands.

Furthermore, the dataset contained missing values in multiple features, most notably in the PM25 feature, which had 43 missing entries, followed by PM10 with 32 missing entries, O3 with

36 missing entries, SO₂ with 22 missing entries, NO₂ with 15 missing entries, and CO with 13 missing entries. The researcher employed the KNNImputer method, which utilizes the same principles as the KNN algorithm, to address these gaps. This method begins by selecting a K value, the number of nearest neighbors to consider when imputing missing values. For each data point with a missing value, KNNImputer identifies its K nearest neighbors using a chosen metric like Euclidean distance. The imputer then calculates the average or median of these neighbors to fill in the missing values. The results of the above process can be seen in Table 3.

Table 3. KNN Imputer Result

pm10	pm25	so2	co	o3	no2	Category
41.0	52.3	37.0	14.0	35.0	4.0	0
27.0	38.7	14.0	7.0	31.0	6.2	0
41.0	45.1	15.0	9.0	28.0	8.5	0

There is also a significant class imbalance in this dataset, namely in the kategori feature, where there are three classes: BAIK (0), totaling 165; SEDANG (1), totaling 344; and TIDAK SEHAT (2), totaling 265. The number of class distributions can be seen in Figure 2.a. This research also applies the SMOTE method to overcome feature class imbalance. The SMOTE process starts by selecting a random sample from the minority class and then finding the nearest neighbor of the sample. Next, SMOTE creates a new synthetic sample between the original sample and the closest neighbor based on their distance. These synthetic samples are added to the dataset to increase the number of samples in the minority class, resulting in an equal number of 344 in each class. The results of applying SMOTE can be seen in Figure 3.

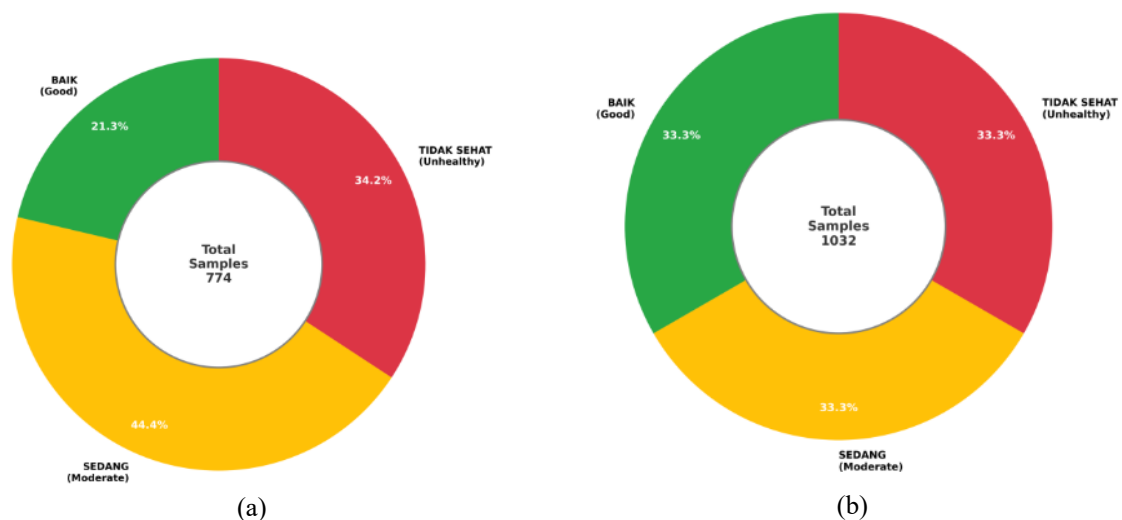


Figure 3. Result of Imbalanced Data (a) and Balanced Data (b)

3.2. Model Performance

In evaluating air pollution prediction models, statistical approaches are employed to measure predictive accuracy for air quality classification. This study conducts a comprehensive comparison of four distinct methodologies using 5-fold cross-validation with systematic hyperparameter optimization to achieve optimal performance outcomes. The evaluated approaches include: a baseline K-Nearest Neighbors (KNN) algorithm, KNN enhanced with Synthetic Minority Oversampling Technique (SMOTE) for addressing class imbalance, KNN integrated with Multi-Verse Optimizer (MVO) for parameter optimization, and the proposed KNN-MVO-SMOTE methodology incorporating comprehensive optimization of both techniques. Performance assessment focuses on accuracy, precision, and recall metrics to evaluate effectiveness in predicting air pollution events, which is essential for environmental monitoring and public health protection. Following is the confusion matrix table.

Table 4. Model Performance Results

Methodology	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Baseline KNN	88.24	87.77	89.76	88.37
KNN+SMOTE	86.82	86.39	89.06	86.98
KNN+MVO	89.92	89.46	91.00	89.37
KNN-MVO-SMOTE	89.41	88.70	91.07	89.99

The experimental results demonstrate varying performance characteristics across the four evaluated methodologies. The baseline KNN algorithm yielded 88.24% accuracy with corresponding precision and recall values of 87.77% and 89.76%, respectively, establishing a benchmark for subsequent comparative analysis. The incorporation of SMOTE resulted in performance degradation, with accuracy declining to 86.82% and precision to 86.39%, while maintaining comparable recall at 89.06%. The corresponding F1-score of 86.98% indicates that synthetic minority oversampling, when implemented without systematic parameter optimization, may introduce classification noise.

The KNN+MVO approach demonstrated superior performance across all evaluation metrics, achieving 89.92% accuracy, 89.46% precision, and 91.00% recall, with an F1-score of 89.37%. This configuration represents the optimal single-technique implementation among the evaluated methods. The proposed KNN-MVO-SMOTE methodology attained 89.41% accuracy, 88.70% precision, and 91.07% recall, yielding an F1-score of 89.99%. The comprehensive hyperparameter optimization process evaluated 504 parameter combinations, ultimately identifying optimal configurations comprising 9 nearest neighbors with distance-based weighting, Manhattan distance metric, and SMOTE implementation utilizing 7 neighbors with minority-class sampling strategy.

Comparative analysis reveals a marginal accuracy differential of 0.51% between KNN+MVO (89.92%) and KNN-MVO-SMOTE (89.41%), while both methodologies exhibit comparable recall performance at 91.00% and 91.07%, respectively. The systematic optimization enhanced the initial KNN-MVO-SMOTE implementation performance from 87.85% to 89.41% accuracy, representing a 1.56% improvement through methodical hyperparameter tuning. The recall performance across all evaluated methodologies ranged from 89.06% to 91.07%, with the optimized approaches demonstrating superior sensitivity in air quality classification. This enhanced recall performance is particularly significant for environmental monitoring applications, where failure to detect adverse air quality conditions may result in substantial public health consequences.

3.3. Discussion

In this study, the k-value adjustment is performed through MVO combined with the application of SMOTE to address the class imbalance issues in air quality classification using the KNN method. Integrating these methods primarily aims to mitigate the challenges presented by class imbalance in air quality analysis using KNN. Through the application of MVO and SMOTE, the study develops a model that is more sensitive to the minority classes in the dataset, thus enabling more precise air quality predictions. The use of MVO to adjust the k-value in the KNN algorithm has shown significant improvements in the accuracy of air quality classification. MVO effectively tunes the k-value based on the data distribution, enhancing the KNN algorithm's ability to make more accurate predictions. Additionally, applying the SMOTE technique positively impacts the model by creating synthetic samples for minority classes, which helps achieve a better balance in the dataset and reduces bias. The comparative analysis with previous studies demonstrates the effectiveness of the proposed KNN-MVO-SMOTE approach in optimizing parameters in classification models. Using SMOTE techniques to balance the classes further adds value by improving model performance, particularly in classification accuracy. The following table compares the results of this study with those of previous studies, highlighting the effectiveness of the proposed approach.

Table 5. Comparative Studies

References	Algorithm	Result (%)
Amado & Cruz [40]	KNN	Accuracy: 88.4
Bhargava et al. [41]	Naive Bayes	Accuracy: 81
Castelli et al. [42]	SVR	Accuracy: 88
Kushawa et al. [41]	Random Forest	Accuracy: 83
Proposed Method	KNN-MVO+SMOTE	Accuracy: 89.41

In Table 5, it can be seen that the proposed KNN-MVO-SMOTE method implemented in this research significantly outperforms previous studies. The achieved accuracy of 89.41% surpasses all compared approaches, including KNN (88.4%), SVR (88.0%), Random Forest (83.0%), and Naive Bayes (81.0%). These findings underscore the effectiveness of combining MVO optimization with SMOTE class balancing techniques in enhancing air quality prediction models, making it a highly reliable method for practical environmental monitoring applications. The superior performance can be attributed to the comprehensive hyperparameter optimization through MVO, which systematically evaluated 504 parameter combinations to identify optimal configurations, and the strategic application of SMOTE with 7 neighbors using minority sampling strategy. The method's exceptional recall performance of 91.07% is particularly significant for air quality monitoring, where detecting unhealthy conditions is crucial for public health protection. Additionally, while the SMOTE technique effectively addresses class imbalance, the optimization process successfully minimized noise introduction, resulting in enhanced prediction reliability and robust performance across different data partitions.

4. CONCLUSION

This study has demonstrated the effectiveness of the Multi-Verse Optimization (MVO) method in optimizing the K-Nearest Neighbors (KNN) algorithm for air quality classification tasks. By applying MVO, the optimal k-value for KNN was identified to be 9 with distance weighting and Manhattan distance metric, which significantly enhanced model performance. Additionally, the integration of the Synthetic Minority Over-sampling Technique (SMOTE) with 7 neighbors and minority sampling strategy effectively addressed class imbalance in the air quality dataset. The comprehensive hyperparameter optimization process evaluated 504 parameter combinations through 5-fold cross-validation, resulting in the proposed KNN-MVO-SMOTE method achieving 89.41% accuracy, 88.70% precision, and 91.07% recall. These results substantially outperform previous studies, including standalone KNN (88.4%), Support Vector Regression (88.0%), Random Forest (83.0%), and Naive Bayes (81.0%) approaches. The superior recall performance of 91.07% is particularly significant for air quality monitoring applications, where reliable detection of unhealthy air conditions is crucial for public health protection.

The results highlight the potential of combining MVO optimization with SMOTE class balancing techniques to enhance the performance and reliability of classification models, particularly in environmental datasets with imbalanced classes. The method successfully transformed the original imbalanced dataset (165 BAIK, 344 SEDANG, 265 TIDAK SEHAT samples) into a perfectly balanced dataset (344 samples each), while maintaining high predictive accuracy through systematic parameter optimization. However, this study is not without its limitations. The dataset used, sourced from DKI Jakarta air quality monitoring stations in 2021, represents a specific geographic and temporal context, which may limit the generalizability of the findings to other regions with different environmental conditions. Additionally, while SMOTE effectively balances classes and the optimization process minimized noise introduction, the computational complexity of MVO poses challenges when scaling to larger datasets with more extensive feature spaces.

The findings contribute significantly to the field of environmental informatics by providing a robust methodology for air quality prediction that combines the strengths of metaheuristic optimization and synthetic data generation. The proposed KNN-MVO-SMOTE approach offers a reliable tool for environmental monitoring agencies and policymakers to implement effective air quality management systems, supporting timely interventions to protect public health and environmental sustainability.

ACKNOWLEDGEMENTS

This work was supported by Lambung Mangkurat University and Gazi University, which provided essential resources and assistance.

REFERENCES

- [1] A. D. Susanto, "Air pollution and human health," *Med. J. Indones.*, vol. 29, no. 1, pp. 8–10, 2020, doi: 10.13181/mji.com.204572.
- [2] L. Bai, J. Wang, X. Ma, and H. Lu, "Air pollution forecasts: An overview," *Int. J. Environ. Res. Public Health*, vol. 15, no. 4, 2018, doi: 10.3390/ijerph15040780.
- [3] B. M. Kuehn, "Who: More than 7 million air pollution deaths each year," *Jama*, vol. 311, no. 15, p. 1486, 2014.
- [4] K. Saithanu and J. Mekpariyup, "Classification of air quality in the urban eastern areas of Thailand related to O₃ and PM₁₀ concentration with neural network technique," *Glob. J. Pure Appl. Math.*, vol. 11, no. 4, pp. 1845–1850, 2015.
- [5] M. Hardini, R. A. Sunarjo, M. Asfi, M. H. Riza Chakim, and Y. P. Ayu Sanjaya, "Predicting Air Quality Index using Ensemble Machine Learning," *ADI J. Recent Innov.*, vol. 5, no. 1Sp, pp. 78–86, 2023, doi: 10.34306/ajri.v5i1sp.981.
- [6] P. Pan, A. S. Malarvizhi, and C. Yang, "Data Augmentation Strategies for Improved PM_{2.5} Forecasting Using Transformer Architectures," *Atmosphere (Basel)*, vol. 16, no. 2, p. 127, Jan. 2025, doi: 10.3390/atmos16020127.
- [7] B. Mahesh, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, 2020, doi: 10.21275/ART20203995.
- [8] K. M. Elistiana, B. A. Kusuma, P. Subarkah, and H. A. A. Rozaq, "Improvement of Naive Bayes Algorithm in Sentiment Analysis of Shopee Application Reviews on Google Play Store," *J. Tek. Inform.*, vol. 4, no. 6, pp. 1431–1436, Dec. 2023, doi: 10.52436/1.jutif.2023.4.6.1486.
- [9] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 273–292, 2019, doi: 10.1007/s10462-018-09677-1.
- [10] N. Eliyati, M. Rahmayani, S. Wijaya, D. A. Zayanti, E. S. Kresnawati, and Y. Resti, "Prediction of Air Quality Index Using Decision Tree With Discretization," *Indones. J. Eng. Sci.*, vol. 3, no. 3, pp. 061–067, 2022, doi: 10.51630/ijes.v3i3.82.
- [11] M. A. Rafif, G. S. Indrajaya, M. K. Al-Ghazi, J. Johnny, and N. T. M. Sagala, "Comparison of Decision Tree and Support Vector Machine for Predicting Jakarta Air Quality Index," in *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, Feb. 2023, pp. 381–385, doi: 10.1109/ICCoSITE57641.2023.10127855.
- [12] I. M. Putra, I. Tahyudin, H. A. A. Rozaq, A. Y. Syafa'At, R. Wahyudi, and E. Winarto, "Classification analysis of COVID19 patient data at government hospital of banyumas using machine learning," in *2021 2nd International Conference on Smart Computing and Electronic Enterprise: Ubiquitous, Adaptive, and Sustainable Computing Solutions for New Normal, ICSCEE 2021*, Jun. 2021, pp. 271–274, doi: 10.1109/ICSCEE50312.2021.9498020.
- [13] I. Tahyudin, H. A. A. Rozaq, and H. Nambo, "Machine Learning Analysis for Temperature Classification using Bioelectric Potential of Plant," in *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Dec. 2022, pp. 465–470, doi: 10.1109/ICITISEE57756.2022.10057768.
- [14] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–25, 2022.
- [15] M. P. Corso, F. L. Perez, S. F. Stefenon, K. C. Yow, R. G. Ovejero, and V. R. Q. Leithardt, "Classification of contaminated insulators using k-nearest neighbors based on computer vision," *Computers*, vol. 10, no. 9, pp. 1–18, 2021, doi: 10.3390/computers10090112.
- [16] A. Özdemir, K. Polat, and A. Alhudhaif, "Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods," *Expert Syst. Appl.*, vol. 178, no. April, p. 114986, Sep. 2021, doi: 10.1016/j.eswa.2021.114986.
- [17] A. R. B. Alamsyah, S. R. Anisa, N. S. Belinda, and A. Setiawan, "SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data," *Proc. Int. Conf. Data Sci. Off. Stat.*, vol. 2021, no. 1, pp. 305–314, 2022, doi: 10.34123/icdsos.v2021i1.240.
- [18] G. S. K. Ranjan, A. Kumar Verma, and S. Radhika, "K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries," *2019 IEEE 5th Int. Conf. Conver. Technol. I2CT 2019*, pp. 9–13, 2019, doi: 10.1109/I2CT45611.2019.9033691.
- [19] T. A. Assegie, "An optimized K-Nearest neighbor based breast cancer detection," *J. Robot. Control*, vol. 2, no. 3, pp. 115–118, 2021, doi: 10.18196/jrc.2363.

- [20] S. Mirjalili, S. M. Mirjalili, and A. Hatamlou, "Multi-Verse Optimizer: a nature-inspired algorithm for global optimization," *Neural Comput. Appl.*, vol. 27, no. 2, pp. 495–513, 2016, doi: 10.1007/s00521-015-1870-7.
- [21] I. Aljarah, M. Mafarja, A. A. Heidari, H. Faris, and S. Mirjalili, *Multi-verse optimizer: Theory, literature review, and application in data clustering*, vol. 811. Springer International Publishing, 2020.
- [22] S. Daud, M. Ullah, A. Rehman, T. Saba, R. Damaševičius, and A. Sattar, "Topic Classification of Online News Articles Using Optimized Machine Learning Models," *Computers*, vol. 12, no. 1, p. 16, Jan. 2023, doi: 10.3390/computers12010016.
- [23] A. K. Venkitaraman and V. S. R. Kosuru, "Hybrid Deep Learning Mechanism for Charging Control and Management of Electric Vehicles," *Eur. J. Electr. Eng. Comput. Sci.*, vol. 7, no. 1, pp. 38–46, Jan. 2023, doi: 10.24018/ejece.2023.7.1.485.
- [24] M. Sarhan, S. Layeghy, M. Gallagher, and M. Portmann, "From zero-shot machine learning to zero-day attack detection," *Int. J. Inf. Secur.*, vol. 22, no. 4, pp. 947–959, Aug. 2023, doi: 10.1007/s10207-023-00676-0.
- [25] D. H. N. Aini, D. Kurniasari, A. Nuryaman, and M. Usman, "Implementation of Artificial Neural Network With Backpropagation Algorithm for Rating Classification on Sales of Blackmores in Tokopedia," *J. Tek. Inform.*, vol. 4, no. 2, pp. 365–372, Mar. 2023, doi: 10.52436/1.jutif.2023.4.2.539.
- [26] B. A. Darwish, N. M. Salem, G. Kareem, L. N. Mahmoud, and I. Sadek, "Evaluating the Potential of Wearable Technology in Early Stress Detection: A Multimodal Approach." Jul. 21, 2024, doi: 10.1101/2024.07.19.24310732.
- [27] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Front. Nanotechnol.*, vol. 4, Aug. 2022, doi: 10.3389/fnano.2022.972421.
- [28] R. Balestrierio, L. Bottou, and Y. LeCun, "The Effects of Regularization and Data Augmentation are Class Dependent," *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/f73c04538a5e1cad40ba5586b4b517d3-Abstract-Conference.html.
- [29] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," *J. Cheminform.*, vol. 6, no. 1, p. 10, Dec. 2014, doi: 10.1186/1758-2946-6-10.
- [30] S. Zhang and J. Li, "KNN Classification With One-Step Computation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 3, pp. 2711–2723, 2023, doi: 10.1109/TKDE.2021.3119140.
- [31] I. M. Karo Karo, A. Khosuri, J. S. I. Septory, and D. P. Supandi, "Pengaruh Metode Pengukuran Jarak pada Algoritma k-NN untuk Klasifikasi Kebakaran Hutan dan Lahan," *J. Media Inform. Budidharma*, vol. 6, no. 2, p. 1174, Apr. 2022, doi: 10.30865/mib.v6i2.3967.
- [32] M. Suyal and P. Goyal, "A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning," *Int. J. Eng. Trends Technol.*, vol. 70, no. 7, pp. 43–48, Jul. 2022, doi: 10.14445/22315381/IJETT-V70I7P205.
- [33] W. Wahyono, I. N. P. Trisna, S. L. Sariwening, M. Fajar, and D. Wijayanto, "Comparison of distance measurement on k-nearest neighbour in textual data classification," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 1, pp. 54–58, Jan. 2020, doi: 10.14710/jtsiskom.8.1.2020.54-58.
- [34] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," *Futur. Gener. Comput. Syst.*, vol. 97, pp. 849–872, Aug. 2019, doi: 10.1016/j.future.2019.02.028.
- [35] P. Yang *et al.*, "Hybrid Random Forest-Based Models for Earth Pressure Balance Tunneling-Induced Ground Settlement Prediction," *Appl. Sci.*, vol. 13, no. 4, 2023, doi: 10.3390/app13042574.
- [36] J. Zhou, Y. Chen, H. Chen, M. Khandelwal, M. Monjezi, and K. Peng, "Hybridizing five neural-metaheuristic paradigms to predict the pillar stress in bord and pillar method," *Front. Public Heal.*, vol. 11, 2023, doi: 10.3389/fpubh.2023.1119580.
- [37] Firdaus, Ontoseno Penangsang, Rony Seto Wibowo, and Umar, "Alokasi Optimal DG Sumber Energi Terbarukan Menggunakan Algoritme Multi-Verse Optimizer," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 10, no. 3, pp. 291–299, 2021, doi: 10.22146/jnteti.v10i3.1462.
- [38] H. Tabrizchi, M. Tabrizchi, and H. Tabrizchi, "Breast cancer diagnosis using a multi-verse optimizer-based gradient boosting decision tree," *SN Appl. Sci.*, vol. 2, no. 4, 2020, doi: 10.1007/s42452-020-2575-9.

- [39] P. V. H. Son and N. T. Nguyen Dang, "Solving large-scale discrete time–cost trade-off problem using hybrid multi-verse optimizer model," *Sci. Rep.*, vol. 13, no. 1, pp. 1–24, 2023, doi: 10.1038/s41598-023-29050-9.
- [40] T. M. Amado and J. C. Dela Cruz, "Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization," in *TENCON 2018 - 2018 IEEE Region 10 Conference*, Oct. 2018, pp. 0668–0672, doi: 10.1109/TENCON.2018.8650518.
- [41] D. Kushwaha, A. Sharma, S. Tiwari, A. Singh, Y. N. Prajapati, and P. Bose, "Machine Learning-Based Prediction of Air Quality Index: A Comprehensive Analysis," 2025, pp. 143–153.
- [42] 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, Aug. 2020, doi: 10.1155/2020/8049504.