Sains Malaysiana 54(6)(2025): 1629-1639
http://doi.org/10.17576/jsm-2025-5406-17

SMOTE-PCADBSCAN: A Novel Approach for Addressing Class Imbalance in Water Quality Prediction

(SMOTE-PCADBSCAN: Suatu Pendekatan Baharu untuk Menangani Ketidakseimbangan Kelas dalam Ramalan Kualiti Air)

NORASHIKIN NASARUDDIN[1,2,*], NURULKAMAL MASSERAN[1], WAN MOHD RAZI IDRIS[3] & AHMAD ZIA UL-SAUFIE[4]

[1]Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Kedah Branch, 08400 Merbok, Kedah, Malaysia
[3]Department of Earth Science and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
[4]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia

Received: 12 August 2024/Accepted: 13 March 2025

ABSTRACT

An accurate and trustworthy prediction model is essential for supporting policy decisions in environmental management concerning water quality prediction. Nonetheless, imbalanced datasets are prevalent in this discipline and hinder identifying crucial ecological factors accurately. This study proposed a novel SMOTE-PCADBSCAN model to enhance the categorisation of water quality data by employing three key components: (i) synthetic minority over-sampling technique (SMOTE), (ii) principal component analysis (PCA), and (iii) density-based spatial clustering of applications with noise (DBSCAN). The minority class was initially augmented using SMOTE, which PCA then decreased the dimensionality. Subsequently, DBSCAN was utilised to generate superior-quality synthetic data by detecting and eliminating extraneous data points. A Malaysia-based multi-class water quality dataset was employed to determine the efficiency of this model. Four different versions of the dataset (Original, SMOTE, SMOTE-DBSCAN, and SMOTE-PCADBSCAN) also utilised five classifier types for the analysis process: (i) decision tree, (ii) random forest, (iii) gradient boosting method, (iv) adaptive boosting, and (v) extreme gradient boosting. Although the original datasets exhibited great accuracy, class imbalance occurred when detecting minority classes. Among the datasets, the metric performances of SMOTE-DBSCAN and SMOTE-PCADBSCAN-based synthetic datasets were superior. The highest accuracy and optimal F1 scores were also demonstrated by RF using the SMOTE-PCADBSCAN approach, which presented excellent water quality classification and imbalanced data management. Consequently, the classification accuracy of imbalanced environmental datasets could be enhanced by employing advanced oversampling techniques and ensemble approaches.

Keywords: DBSCAN; imbalanced data; PCA; SMOTE; water quality

ABSTRAK

Model ramalan yang tepat dan boleh dipercayai adalah penting untuk menyokong keputusan dasar dalam pengurusan alam sekitar berkaitan ramalan kualiti air. Walau bagaimanapun, set data yang tidak seimbang sering berlaku dalam disiplin ini dan menghalang pengenalan faktor ekologi yang penting dengan tepat. Penyelidikan ini mencadangkan model SMOTE-PCADBSCAN yang inovatif untuk meningkatkan pengelasan data kualiti air dengan menggunakan tiga komponen utama: (i) teknik pengambilan sampel berlebihan minoriti sintetik (SMOTE), (ii) analisis komponen utama (PCA) dan (iii) pengelompokan ruang berasaskan ketumpatan aplikasi dengan bunyi (DBSCAN). Kelas minoriti pada mulanya ditambah menggunakan SMOTE, yang kemudiannya mengalami pengurangan dimensi oleh PCA. Seterusnya, DBSCAN digunakan untuk menghasilkan data sintetik berkualiti tinggi dengan mengesan dan menghapuskan titik data yang tidak relevan/berlebihan. Set data kualiti air pelbagai kelas dari Malaysia digunakan untuk menentukan keberkesanan model ini. Empat versi dataset yang berbeza (Asal, SMOTE, SMOTE-DBSCAN dan SMOTE-PCADBSCAN) melibatkan lima jenis pengelas untuk proses analisis: (i) pokok keputusan, (ii) hutan rawak, (iii) mesin penggalakan kecerunan, (iv) penggalakan adaptif dan (v) penggalakan kecerunan ekstrem. Walaupun dataset asal menunjukkan ketepatan yang tinggi, ketidakseimbangan kelas berlaku apabila mengesan kelas minoriti. Antara dataset, prestasi metrik dataset sintetik berasaskan SMOTE-DBSCAN dan SMOTE-PCADBSCAN adalah lebih baik. Ketepatan tertinggi dan skor F1 optimum juga ditunjukkan oleh RF menggunakan pendekatan SMOTE-PCADBSCAN yang menunjukkan prestasi cemerlang dalam pengelasan kualiti

air dan pengurusan data tidak seimbang. Oleh itu, ketepatan pengelasan dataset alam sekitar yang tidak seimbang boleh dipertingkatkan dengan menggunakan teknik pengambilan sampel berlebihan lanjutan dan pendekatan ansambel.

Kata kunci: Data tidak seimbang; DBSCAN; kualiti air; PCA; SMOTE

INTRODUCTION

Effective environmental management and policy-making in ecological data analysis rely heavily on accurate water quality prediction. However, the challenge of imbalanced datasets often emerges, particularly in detecting polluted water events, which occur less frequently than acceptable water quality events. This imbalance significantly impacts machine learning models, as accurately identifying minority class occurrences is critical for effective environmental decision-making. Numerous studies have highlighted the challenges posed by imbalanced datasets in water quality prediction and emphasized the need for innovative approaches to enhance classification performance.

In Malaysia, water quality is commonly assessed using the Water Quality Index (WQI), a metric that integrates six essential variables: dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammonia-nitrogen (NH3-N), suspended solids (SS), and pH. These variables are converted into sub-indices (SI) using formulas provided by the Department of the Environment (DOE) and combined through a weighted summation technique to compute the WQI (DOE 2022). The resultant WQI categorizes water quality into three classes: clean (C), slightly polluted (SP), and polluted (P), providing a standardized framework for assessing water bodies.

The importance of effective water quality management in Malaysian rivers has been extensively studied. Fitri et al. (2020) analyzed the freshwater quality of the Sungai Kelantan, proposing measures to mitigate pollution levels. Ahmed et al. (2020) assessed heavy metal levels in the Sungai Langat and its water supply chain, offering insights into drinking water safety. Yasin and Karim (2020) introduced a fuzzy weighted multivariate regression analysis to design a novel WQI model aligned with DOE requirements. Studies by Ahmed et al. (2022) and Hashem, Ahmad and Yusuf (2021) focused on pollution sources and river basin management for the Sungai Petani and Sungai Langat, respectively. These findings collectively underscore the need for advanced tools to manage water quality more effectively.

Recent advancements in oversampling techniques have introduced sophisticated methods to address class imbalance in water quality datasets. Hybrid approaches, such as the integration of SMOTE with Tomek Links or Edited Nearest Neighbors, not only balance datasets but also effectively mitigate noise (Dogo et al. 2021). Generative adversarial networks (GANs) have further enhanced oversampling by generating synthetic minority samples in a data-driven manner, improving the overall performance of classification models (Poudevigne-Durance 2024). Wong et al. (2023) introduced a stacked ensemble deep learning model for predicting water quality indices (WQI) from imbalanced datasets, achieving notable improvements in accuracy and robustness. Additionally, Shehab et al. (2023) developed a water quality classification model leveraging raw flush sets, demonstrating the utility of advanced techniques in managing class imbalance. These studies highlight the potential of integrating advanced oversampling methods with predictive models to support environmental decision-making and promote sustainable management practices. As water quality monitoring becomes increasingly critical for environmental health, the adoption of these advanced techniques can enhance the reliability of datasets, providing valuable insights for policy development and resource management. However, despite their promise, these methods often entail significant computational demands, are susceptible to overfitting, and involve complex parameter optimization, underscoring the need for further refinement and accessibility.

Despite these advancements, SMOTE remains widely used due to its simplicity, interpretability, and adaptability across various domains, including water quality. As Taloor et al. (2025) demonstrated, SMOTE significantly improves the performance of machine learning models in environmental studies. However, SMOTE has notable limitations, such as sensitivity to noise and its inability to account for data structure. To overcome these limitations, hybrid methods like RN-SMOTE, which incorporate noise reduction techniques, have been proposed to enhance classification performance (Arafa et al. 2022). This study builds on SMOTE's foundation while addressing its limitations by introducing dimensionality reduction and clustering.

This paper proposes a novel SMOTE-PCADBSCAN model that integrates three components: (i) synthetic minority over-sampling technique (SMOTE) for class balancing, (ii) principal component analysis (PCA) for dimensionality reduction, and (iii) density-based spatial clustering of applications with noise (DBSCAN) for noise identification and removal. By enhancing the quality of synthetic data and reducing class disparities, SMOTE-PCADBSCAN model aims to improve the accuracy and reliability of predictive models in water quality classification.

The subsequent sections of this paper are organized as follows: the Materials and Methods section describes the SMOTE-PCADBSCAN model and study methodology; Results and Discussion presents the comparative performance of classifiers and datasets; and Conclusion summarizes key findings and outlines potential future research directions.

MATERIALS AND METHODS

STUDY AND DESIGN

This study assessed the oversampling impact of the SMOTE-PCADBSCAN model on a multi-class water quality dataset obtained from the Malaysian DOE. The dataset encompassed diverse water quality variables, and the DOE station was tasked with overseeing the quality of all Malaysian water resources. Approximately 5511 recordings between 2018 and 2020 were obtained regarding the Malaysian rivers. The primary dependent variable in this study was WQC. This variable classified the water quality into three classes: (i) C, (ii) SP, and (iii) P. The WQI was also employed to conduct this categorisation, which was further subdivided into various categories based on the ranges of sub-indices established by the DOE. Approximately 14 parameters were finalised as the independent variables in this study to define the WQC based on previous studies. Table 1 tabulates the statistical characteristics of the dataset used in this study.

Approximately 19% (1047/5511), 26.5% (1460/5511), and 54.5% (3004/5511) instances were reported in the dataset corresponding to C, SP, and P classes, respectively. Even though this observation implied a minor class imbalance, clean and dirty water should be accurately anticipated owing to the smaller sample sizes in these categories. Hence, the DOE necessitates efficient water management enforcement and conservation policies through a transparent and discerning process of cleaning and contaminating water.

SMOTE-PCADBSCAN

Considering that Chawla et al. (2002) established the highly effective SMOTE, this study employed the algorithm for the SMOTE-PCADBSCAN model by linearly interpolating between a randomly chosen minority sample and one of its neighbouring samples to produce synthetic datasets (Douzas, Bacao & Last 2018). A random minority sample (x_i) is initially chosen. Another sample (x_j) is then selected from the k nearest neighbours belonging to the minority class as follows:

$$x_{new} = x_i + (x_j - x_i) \times \delta \quad (1)$$

where δ represents a random number from 0 to 1. SMOTE is also more advantageous for oversampling because it can prevent overfitting than other algorithms. For example, random oversampling (ROS) duplicates samples from the minority class. In contrast, the SMOTE builds synthetic instances (Jeatrakul, Wong & Fung 2010). This technique pertains to the noise in the initial dataset and the production of novel samples, resulting in additional dissemination and noise amplification. The disruptive samples (outliers) hinder the improvement of various classifiers when the datasets are oversampled using SMOTE (Cheng et al. 2019). Hence, noise elimination methods are required while implementing SMOTE. This process enhances the efficiency of the classifiers employed for dataset classification by mitigating the SMOTE-related noise or the noise inherent in the original datasets.

TABLE 1. Summary of the statistical characteristics of the dataset used in this study

No	Variable	Role	Description	Unit	Type
1	WQC	Target	WQC	1 = C; 2 = SP; 3 = P	Categorical
2	Temp	Input	Temperature	°C	Numerical
3	COND	Input	Electrical Conductivity	uS	Numerical
4	SAL	Input	Salinity	ppt	Numerical
5	TURNTU	Input	Turbidity	NTU	Numerical
6	NO_3	Input	Nitrate	mg/L	Numerical
7	PO_4	Input	Phosphate	mg/L	Numerical
8	As	Input	Arsenic	mg/L	Numerical
9	Hg	Input	Mercury	mg/L	Numerical
10	Cd	Input	Cadmium	mg/L	Numerical
11	Cr	Input	Chromium	mg/L	Numerical
12	Pb	Input	Plumbum	mg/L	Numerical
13	Zn	Input	Zinc	mg/L	Numerical
14	OG	Input	Oil & Grease	mg/L	Numerical
15	*E. coli*	Input	*Escherichia coli*	cfu/100 mL	Numerical

Arafa et al. (2022) developed the RN-SMOTE as a pre-processing technique for unbalanced binary data. This method initially applied the SMOTE technique to oversample the training data, which generated noisy synthetic instances in the minority class. The DBSCAN was then utilised to identify and eliminate noise, suggesting that the RN-SMOTE effectively boosts model performance. Ester et al. (1996) created the DBSCAN, which was a clustering method independent of specific parameters (Tran, Drab & Daszykowski 2013). This approach generally clusters data by calculating the density of points within a distance from each point in the dataset. The algorithm can also locate and remove extraneous data, such as random data, to enhance data accuracy (Kumar & Reddy 2016). Moreover, the DBSCAN approach categorises each point in the dataset into three types: (i) core, (ii) border points, and (iii) noise points (outliers) (Dalakleidi et al. 2017). A point is designated as a core point, and a new cluster is formed if the number of points within a neighbourhood distance ε exceeds the minimum criterion. Alternatively, this point is classified as noise if it does not meet the requirements. The cluster is then expanded by including more locations within the ε-neighbourhood in subsequent iterations. Lastly, this process iterates until no additional points can be included, indicating the conclusion of the current clustering process (Ester et al. 1996). The following explanations present a concise summary of the DBSCAN algorithm for the given dataset $D = \{p_i | p_i \in \mathbb{R}, 1 \leq i \leq n\}$ (Ester et al. 1996; Sander et al. 1998):

(1) ε-neighborhood of a point: This term encompasses all points within a given distance ε from p_i and forming the neighbourhood $N_\varepsilon(p_i)$ as follows:

$$N_\varepsilon(p_i) = \{p_j \in D | dist(p_i, p_j) \leq \varepsilon, p_i \neq p_j\} \quad (2)$$

(2) Directly density reachable: This term represents clusters of core points surrounded by border points. The border points are also part of the cluster, which are inside the ε-neighbourhood of a core point. A certain number of points (*MinPts*) in its ε-neighbourhood is necessary for a point to be classified as a core point as follows:

$$|N_\varepsilon(p_j)| \geq MinPts, \text{ then } p_j \text{ is a corepoint} \quad (3)$$

(3) Density reachable: Considering ε and *MinPts*, a point is considered directly density reachable if a sequence of points $\{p_i | 1 \leq i \leq n\}$ is present, where p_n is directly density reachable from p_1. This outcome is attributed to each subsequent point $p_i + 1$ that is directly density reachable from p_i.

(4) Density connected: Given ε and *MinPts*, points p_i and p_j are called densely connected if a point p_o exist such that both p_i and p_j are densely reachable from p_o.

(5) Cluster: Considering ε and *MinPts*, if a point p_i is part of a cluster *C*, then the point p_j also belongs to *C* if it is reachable from p_i with high density. If two points p_i and p_j are part of the same cluster *C*, it implies that they exhibit a high level of connectivity.

(6) Noise: Given ε and *MinPts*, the noise refers to the points in the dataset *C* containing cluster $\{C_i | 1 \leq i \leq k\}$ that are not assigned to any cluster as follows:

$$noise = \{p \in D | \forall i: p \notin C_i\} \quad (4)$$

The DBSCAN algorithm necessitates the incorporation of two parameters (ε and *MinPts*). In contrast, the sorted *k*-distance graph was first proposed as a pioneering and extensively employed method for estimating the parameters of the DBSCAN algorithm (Starczewski, Goetzen & Er 2020). This graphical method involves identifying the *k*-nearest neighbour (KNN) for each point and arranging them in ascending order depending on their distances. The location of the highest degree of curvature on the resulting curve is then identified to determine the value of ε. This study also utilised the silhouette index to ascertain the appropriate *MinPts* value, as it was a widely employed metric for assessing clustering outcomes (Blahova, Horecny & Kostolny 2023). Conversely, a difficulty occurred when using data mining clustering approaches to datasets containing many characteristics (Kanungo et al. 2002).

Previous studies identified the PCA as the most efficient method for reducing data in these situations (Mustakim et al. 2021; Shen et al. 2021; Rahman et al. 2020). Although the dataset used in this study is not highly dimensional, PCA was incorporated into the SMOTE-PCADBSCAN methodology for two primary reasons: (i) to simplify the data structure and optimize the clustering process within DBSCAN by reducing potential redundancies in the feature space, and (ii) to retain the most relevant variance in the data, thereby improving the identification of clusters and mitigating noise in synthetic data. This study subsequently suggested the integration of SMOTE, PCA, and DBSCAN components in the proposed SMOTE-PCADBSCAN model to provide training datasets of superior quality. Initially, the PCA was proposed by Pearson in 1901 and subsequently advanced separately by Hotelling in 1933 and Jolliffe in 1986 (Marsboom et al. 2018). The primary goal of this analysis was to decrease the number of variables while preserving the crucial information (Kavitha & Caroline 2015). Generally, the PCA consists of five stages as follows (Marsboom et al. 2018): 1) Normalising the data by removing the mean from each data value, 2) Calculating the covariance matrix, 3) Determining the eigenvalues and eigenvectors, 4) Selecting components and feature vectors, and 5) Constructing a new dataset.

Figure 1 depicts the process flow of the proposed SMOTE-PCADBSCAN model for this study. The process

begins with the water quality dataset undergoing data preprocessing, where missing values are handled, and variables are standardized to ensure consistency. Next, the dataset is split into training (70%) and testing (30%) groups. SMOTE is applied exclusively to the training data to generate synthetic data for minority classes, addressing class imbalance. Following this, PCA is used to reduce the dimensionality of the synthetic data, simplifying the data structure and retaining the most significant features. The reduced-dimension synthetic data is then processed through DBSCAN, which clusters the data and identifies noisy samples. Noise is removed at this stage, resulting in a cleaned synthetic dataset. The cleaned synthetic data is combined with the original training data to form a comprehensive training set. This final training dataset is used to train machine learning classifiers, while the testing dataset is reserved for model evaluation.

In this study, the machine learning algorithms utilized include decision tree (DT) and ensemble methods such as random forest (RF), gradient boosting machine (GBM), AdaBoost, and XGBoost, all integral to the proposed SMOTE-PCADBSCAN model. Ensemble learning combines multiple base models to enhance predictive accuracy and robustness (Dong et al. 2020). DT algorithms, used for classification and regression, construct decision trees where nodes test attributes and branches represent outcomes, effectively modeling complex decision processes (Abedinia & Seydi 2024). RF, introduced by Breiman (2001), employs multiple decision trees and majority voting to improve predictive performance (Alqahtani et al. 2022). GBM iteratively adds small decision trees to minimize residual errors, boosting overall model accuracy (Sarker 2021). AdaBoost adjusts sample weights to focus on misclassified instances, combining weak learners into a strong classifier (Schapire 1999). XGBoost, a modern gradient boosting method, optimizes performance with enhanced software and hardware implementations (Chen & Guestrin 2016). Classifier performance was evaluated using six metrics: accuracy, sensitivity, specificity, precision, F1 score, and average F1 score.

RESULTS AND DISCUSSION

This section presents the results of applying different classifiers to a multi-class water quality dataset under four scenarios: the original dataset, SMOTE oversampled dataset, SMOTE-DBSCAN dataset, and SMOTE-PCADBSCAN dataset. The classifiers were evaluated using accuracy, sensitivity, specificity,

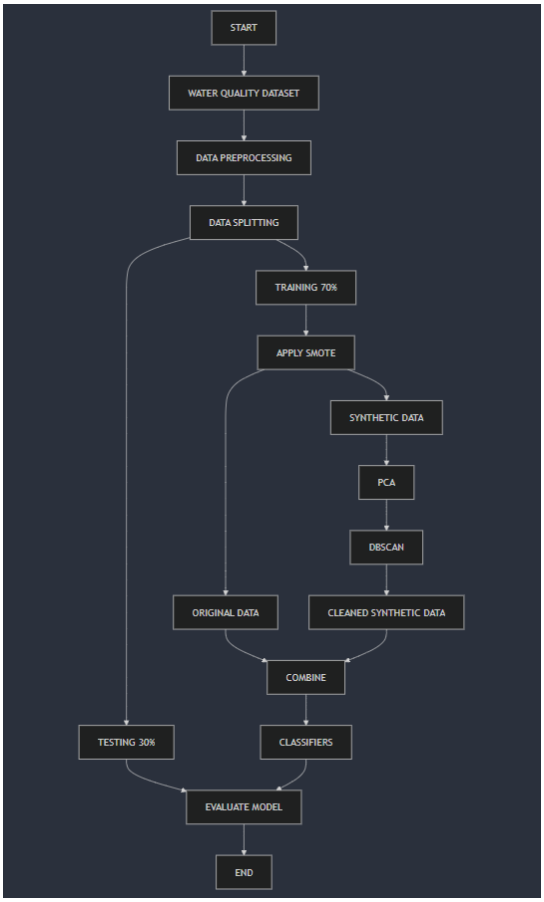

FIGURE 1. The flowchart of the proposed SMOTE-PCADBSCAN model in this study

precision, F1 score, and average F1 score to assess the performance improvements provided by the proposed SMOTE-PCADBSCAN model.

DATA PRE-PROCESSING

In this phase, various measures were taken to replace missing data and to standardise the data in order to achieve optimal classification results. Figure 2 depicts that missing values were observed for five variables: *E. coli*, PO4, WQC, TURNTU, and NO3, with the rate of missing values ranging from 0.0002% to 5.2%. These were imputed using the KNN method with k=5. Min-max normalisation was then performed to standardise the data set and ensure uniform scaling. Figure 3 displays the correlation matrix of the 14 input variables utilised in this investigation. The matrix showed strong correlations between COND and SAL, TURNTU and NO3, TURNTU and Zn, and NO3 and Zn with correlation coefficients of 1.0, 0.99, 0.99, and 1.0, respectively. Meanwhile, SAL, NO3, and Zn were excluded from the dataset to prevent repetition. The revised dataset consisted of 11 input variables for water quality classification, which were then divided into two categories: Training (70%) and Test (30%).

CLASS IMBALANCE

The dataset showed an imbalance in water quality classes, potentially affecting the training process. To address this, four dataset variations were generated: Original, SMOTE, SMOTE-DBSCAN, and SMOTE-PCADBSCAN. The dataset consisted of 5511 samples, with 3861 (70%) used for training and 1650 (30%) for testing. Table 2 summarizes the original dataset distribution, with 739, 2104, and 1018 samples in classes C, SP, and P, respectively.

SMOTE was applied to mitigate the imbalance, increasing the samples for classes C and P to 2217 and 2036, respectively. Further refinement using SMOTE-DBSCAN adjusted these counts to 2151 (C) and 2026 (P). The proposed SMOTE-PCADBSCAN model produced slightly more balanced distributions, with 2195 (C) and 2025 (P) samples, improving class equalization and enhancing model performance. The final test dataset retained 308, 900, and 442 samples for classes C, SP, and P, respectively.

CLASSIFICATION

The classification phase aimed to evaluate the effectiveness of five algorithms—DT, RF, GBM, AdaBoost, and XGBoost—in determining water quality status across various dataset scenarios: Original, SMOTE, SMOTE-DBSCAN, and SMOTE-PCADBSCAN. Performance metrics included accuracy, sensitivity, specificity, precision, F1 score and average F1 score (Table 3).

The DT algorithm achieved an accuracy of 63.21% and an average F1 score of 74.61% on the original dataset. However, its sensitivity for classes C (49.03%) and P (23.98%) was notably low, despite high specificity values for both classes. When applied to oversampled datasets, the accuracy declined to 52.79%, but sensitivity for class P improved significantly to 91.63%, while sensitivity for SP decreased to 31.33%. This outcome highlights a trade-off between accuracy and sensitivity, underscoring the limitations of DT in effectively handling imbalanced data.

RF demonstrated superior performance, with the highest accuracy of 73.45% on the original dataset. Oversampling improved sensitivity and specificity for minority classes, with SMOTE-DBSCAN and SMOTE-PCADBSCAN yielding higher average F1 scores (84.07% and 84.14%, respectively). The SMOTE-PCADBSCAN model achieved the best balance across metrics, improving sensitivity for classes C (75.65%) and P (64.48%).

GBM achieved a maximum accuracy of 72.24% and high precision for class C (81.50%) on the original dataset. However, it exhibited lower sensitivity for classes P (54.98%) and SP (60.80%), indicating challenges in handling class imbalance. When SMOTE and SMOTE-DBSCAN were applied, average F1 scores improved to 83.42% and 83.57%, respectively, although accuracy slightly decreased to 71.21% and 71.45%. These oversampling techniques enhanced sensitivity for class P (67.87%) and specificity for SP (72% and 72.40%). The SMOTE-PCADBSCAN model produced the lowest accuracy (70.73%) but achieved balanced performance with an average F1 score of 83.08%, showing improved sensitivity and specificity for classes C and SP. While the original dataset yielded the highest accuracy for GBM, the SMOTE-DBSCAN dataset demonstrated a more equitable performance across all metrics, addressing class imbalance more effectively.

AdaBoost achieved the highest accuracy (68.91%) and precision (84.34%) for class C on the original dataset. However, it struggled with class imbalance, showing poor sensitivity for class P (40.05%) and moderate specificity for SP (46.13%). Accuracy slightly decreased with SMOTE (66.55%) and SMOTE-DBSCAN (66.79%), but both datasets demonstrated higher average F1 scores (81.29% and 81.27%, respectively), indicating better handling of class imbalance. These oversampling methods improved sensitivity for P (68.30% and 66.52%) and specificity for SP (73.60% and 71.73%). The SMOTE-PCADBSCAN model offered the best balance, achieving an accuracy of 67.58% and the highest average F1 score (81.77%). It also enhanced sensitivity for P (69.23%) and maintained balanced specificity across all classes. While AdaBoost delivered the highest accuracy on the original dataset, the SMOTE-PCADBSCAN model provided a more equitable and robust performance, effectively addressing class imbalance.

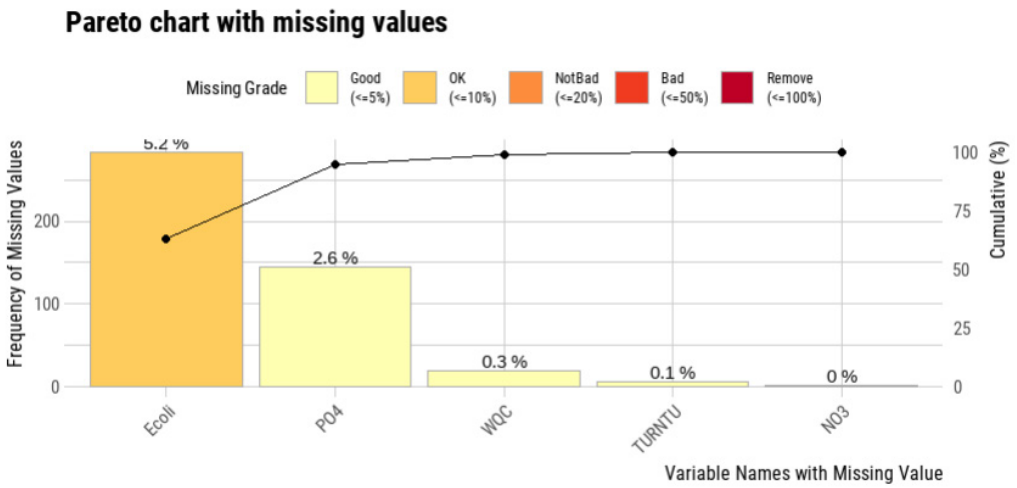

FIGURE 2. The input variables with missing values

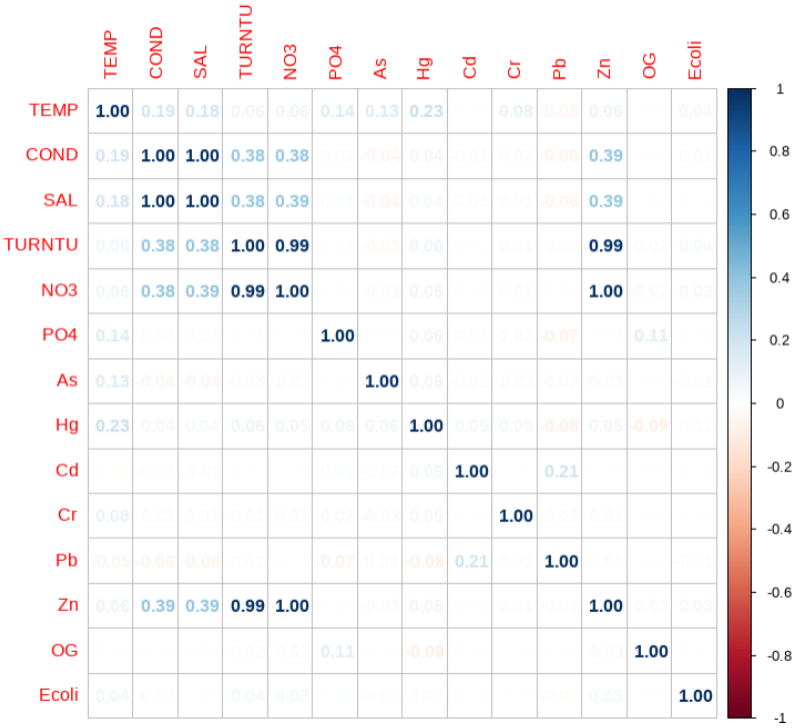

FIGURE 3. The correlations between input variables

TABLE 2. Summary of the training and testing dataset samples

Dataset	Scenarios	C	SP	P
Training	Original	739	2104	1018
	SMOTE	2217	2104	2036
	SMOTE-DBSCAN	2151	2104	2026
	SMOTE-PCADBSCAN	2195	2104	2025
Testing	Original	308	900	442

TABLE 3. Summary of the classifier performance results

Classifier		Accuracy	Average F1-score	Sensitivity			Specificity			Precision			F1-score		
				C	SP	P	C	SP	P	C	SP	P	C	SP	P
DT	Original	0.6321	0.74606	0.49026	0.8733	0.23982	0.94784	0.3453	0.96192	0.68326	0.6155	0.69739	0.91808	0.46126	0.85883
	SMOTE	0.5279	0.74371	0.5974	0.3133	0.9163	0.9292	0.8693	0.5149	0.6595	0.7421	0.4087	0.91928	0.64554	0.66631
	SMOTE-DBSCAN	0.5279	0.74371	0.5974	0.3133	0.9163	0.9292	0.8693	0.5149	0.6595	0.7421	0.4087	0.91928	0.64554	0.66631
	SMOTE-PCADBSCAN	0.5279	0.74371	0.5974	0.3133	0.9163	0.9292	0.8693	0.5149	0.6595	0.7421	0.4087	0.91928	0.64554	0.66631
RF	Original	0.7345	0.83141	0.6494	0.8744	0.509	0.9762	0.5707	0.9305	0.8621	0.7096	0.7282	0.94928	0.66305	0.88191
	SMOTE	0.7255	0.83993	0.763	0.7467	0.6561	0.9411	0.708	0.8717	0.7484	0.7542	0.6517	0.94324	0.70378	0.87277
	SMOTE-DBSCAN	0.7273	0.84067	0.7565	0.7533	0.6538	0.9426	0.7053	0.8742	0.7516	0.7542	0.6553	0.94333	0.70486	0.87381
	SMOTE-PCADBSCAN	0.7297	0.84137	0.7565	0.7622	0.6448	0.9434	0.6987	0.8808	0.754	0.7522	0.6643	0.94372	0.7043	0.87608
GBM	Original	0.7224	0.82936	0.6721	0.8244	0.5498	0.965	0.608	0.9031	0.815	0.7162	0.675	0.94595	0.66862	0.8735
	SMOTE	0.7121	0.83419	0.7597	0.7122	0.6787	0.9344	0.72	0.8535	0.7267	0.7532	0.6289	0.93933	0.69722	0.86602
	SMOTE-DBSCAN	0.7145	0.83568	0.763	0.7156	0.6787	0.9367	0.724	0.8518	0.7344	0.7568	0.6263	0.94087	0.7011	0.86507
	SMOTE-PCADBSCAN	0.7073	0.83084	0.7565	0.7133	0.6606	0.9389	0.7093	0.8485	0.7397	0.7465	0.6147	0.94135	0.69091	0.86026
AdaBoost	Original	0.6891	0.79397	0.5422	0.8811	0.4005	0.9769	0.4613	0.9354	0.8434	0.6625	0.6941	0.93844	0.57523	0.86823
	SMOTE	0.6655	0.81287	0.7565	0.6256	0.683	0.9106	0.736	0.806	0.6601	0.7398	0.563	0.92611	0.67358	0.83893
	SMOTE-DBSCAN	0.6679	0.81266	0.7565	0.6389	0.6652	0.9121	0.7173	0.8195	0.6638	0.7306	0.5742	0.92692	0.66708	0.84399
	SMOTE-PCADBSCAN	0.6758	0.81774	0.7468	0.6433	0.6923	0.9098	0.732	0.8237	0.6553	0.7423	0.5896	0.92465	0.67778	0.85079
XGBoost	Original	0.7176	0.82519	0.6558	0.8278	0.5362	0.9665	0.592	0.9048	0.8178	0.7088	0.6733	0.94499	0.65827	0.87231
	SMOTE	0.7061	0.82724	0.7208	0.7378	0.6312	0.9449	0.6733	0.8626	0.75	0.7305	0.672	0.94065	0.6774	0.86366
	SMOTE-DBSCAN	0.7121	0.83119	0.7403	0.7433	0.629	0.9463	0.684	0.8626	0.76	0.7384	0.6261	0.94354	0.68675	0.8633
	SMOTE-PCADBSCAN	0.7061	0.82785	0.7273	0.7344	0.6335	0.9441	0.68	0.8593	0.7492	0.7336	0.6222	0.94096	0.68045	0.86213

XGBoost achieved the highest accuracy (71.76%) on the original dataset but showed lower sensitivity for classes C (65.58%) and P (53.62%). Notably, it exhibited exceptional specificity for class C (96.65%). Applying SMOTE resulted in a slight decrease in accuracy (70.61%) but improved the average F1 score (82.72%), along with increased sensitivity for C (72.08%) and P (63.12%). The highest average F1 score (83.12%) was obtained with the SMOTE-DBSCAN dataset, which offered the most balanced performance, improving sensitivity (C: 74.03%, SP: 74.33%), precision (C: 76%, SP: 73.84%), and specificity (C: 94.63%, P: 86.26%). The SMOTE-PCADBSCAN model demonstrated comparable accuracy (70.61%) and a slightly lower average F1 score (82.79%) but maintained balanced metrics across all classes.

While the performance of SMOTE-PCADBSCAN did not significantly improve metrics for models like XGBoost and GBM, it notably enhanced the performance of the RF model. Specifically, SMOTE-PCADBSCAN increased the average F1 score for RF to 84.14%, the highest among all combinations, and improved sensitivity and specificity for minority classes (C and P). This highlights the importance of selecting the appropriate model-oversampling combination, as SMOTE-PCADBSCAN is particularly effective with RF for addressing class imbalance.

Overall, while the original dataset achieved the highest accuracy, the synthetic datasets, particularly SMOTE-DBSCAN and SMOTE-PCADBSCAN, provided more robust and balanced performance. These findings highlight their effectiveness in addressing class imbalance and improving classification results.

PERFORMANCE EVALUATION

Accurately identifying C and P classes is essential in water quality classification to treat polluted rivers promptly. Given that this classification system consisted of three classes (C, SP, and P), noteworthy outcomes were reported when various algorithms (DT, RF, GBM, AdaBoost, and XGBoost) evaluated different training datasets (Original, SMOTE, SMOTE-DBSCAN, and SMOTE-PCADBSCAN). Out of these options, the most effective strategy for increasing classification performance was to combine the RF algorithm with the proposed SMOTE-PCADBSCAN model in this study. Meanwhile, the DT algorithm demonstrated simplicity and interpretability. Nevertheless, the proper management of intricate water quality when accounting for various sampling methodologies using DT must be further assessed. Even though a good accuracy of 63.21% with the original dataset was produced with the original dataset using DT in this study, its performance decreased dramatically to 52.79% when applied to the SMOTE, SMOTE-DBSCAN, and SMOTE-PCADBSCAN datasets. This outcome was attributed to the insufficient consideration of imbalanced data distribution by DT, leading to lower identification performance of C and P.

This study demonstrated that RF was the optimal algorithm compared to other models across all datasets. The finding was concluded due to its capability to achieve the maximum accuracy of 73.45% on the original dataset. When SMOTE, SMOTE-DBSCAN, and SMOTE-PCADBSCAN were applied, higher system performances were also observed. These observations were due to the PCA and DBSCAN components boosting the capability of RF to handle imbalanced data efficiently. Consequently, the resilience and reliability of RF in recognising C and P were showcased through the maximum accuracy and average F1 score levels. The equitable precision and recall variables denoted the effectiveness of RF with the proposed SMOTE-PCADBSCAN model in this study. Exceptional specificity and sensitivity values in C and P were reported using this model, promoting dependable water quality monitoring and prompt intervention in contaminated rivers. Thus, a thorough and detailed comprehension of the data was verified owing to three factors: (i) RF algorithm, (ii) ensemble approach incorporating the decisions of several trees, and (iii) SMOTE-PCADBSCAN. These factors contributed to the most optimal classification accuracy and ensured the precise identification of crucial water quality categories. Eventually, efficient environmental management and pollution control could be realised using the proposed study model.

CONCLUSIONS

This study successfully classified water quality using a labelled, multi-class dataset with the help of various machine learning classifiers. Three methods - SMOTE, SMOTE-DBSCAN, and SMOTE-PCADBSCAN - were employed to address class imbalance in the dataset. A systematic performance comparison of five classifiers (DT, RF, GBM, AdaBoost, and XGBoost) was conducted across four dataset versions, demonstrating improvements in classification accuracy. Missing values were addressed through KNN imputation, while strongly correlated variables were removed during the pre-processing stage to eliminate redundancy and improve the dataset's reliability. The oversampling techniques, particularly SMOTE-based methods, proved beneficial for minority classes, leading to more accurate water quality assessments.

Although the original dataset often achieved the highest accuracy, its inability to handle class imbalance effectively for minority classes was evident. The synthetic datasets generated using SMOTE-DBSCAN and SMOTE-PCADBSCAN demonstrated a more balanced and robust performance across key metrics, including accuracy and F1 score. Notably, the RF model paired with the proposed SMOTE-PCADBSCAN approach significantly improved performance, making it an effective tool for addressing class imbalance in environmental datasets. The RF-SMOTE-PCADBSCAN combination was found to be a reliable method for promptly and accurately identifying clean and polluted rivers, contributing to improved environmental management and public health outcomes.

While this study demonstrated the effectiveness of the proposed SMOTE-PCADBSCAN method, several limitations should be acknowledged. First, the study focused on improving the SMOTE algorithm and did not incorporate recent data imbalance techniques for comparison due to time constraints. This limits the breadth of the analysis and the generalizability of the findings to other advanced oversampling methods. Second, the evaluation was conducted on a single dataset, which may not fully represent the diversity of environmental data. Future studies should validate the model using datasets from different geographical regions or with varying class distributions. Lastly, while the SMOTE-PCADBSCAN model demonstrated enhanced performance for RF, the improvements for other classifiers were less pronounced, suggesting the need for further refinement and exploration of hybrid methods to enhance compatibility with other machine learning algorithms.

In summary, integrating data pre-processing and advanced oversampling techniques with powerful machine learning algorithms can address the challenges posed by imbalanced datasets. The RF-SMOTE-PCADBSCAN model developed in this study proved effective for water quality classification and shows potential for broader applications in environmental studies. Further research should focus on testing these methods on larger, more diverse datasets and refining oversampling techniques to extend their applicability in environmental data science.

ACKNOWLEDGEMENTS

We would like to sincerely thank the Department of Environment for providing the important water quality data for this study. Their support and access to comprehensive data were invaluable to the successful completion of this study. We would also like to thank all the individuals and institutions who indirectly contributed to this study.

REFERENCES

Abedinia, A. & Seydi, V. 2024. Building semi-supervised decision trees with semi-cart algorithm. *International Journal of Machine Learning and Cybernetics* 15: 4493-4510.

Ahmed, M.F., Mokhtar, M.B., Lim, C.K. & Majid, N.A. 2022. Identification of water pollution sources for better Langat River basin management in Malaysia. *Water* 14(12): 1904.

Ahmed, M.F., Mokhtar, M.B., Alam, L., Mohamed, C.A.R. & Ta, G.C. 2020. Investigating the status of cadmium, chromium and lead in the drinking water supply chain to ensure drinking water quality in Malaysia. *Water* 12(10): 2653.

Alqahtani, A., Shah, M.I., Aldrees, A. & Javed, M.F. 2022. Comparative assessment of individual and ensemble machine learning models for efficient analysis of river water quality. *Sustainability* 14(3): 1183.

Arafa, A., El-Fishawy, N., Badawy, M. & Radad, M. 2022. RN-SMOTE: Reduced noise SMOTE based on DBSCAN for enhancing imbalanced data classification. *Journal of King Saud University - Computer and Information Sciences* 34(8): 5059-5074.

Blahova, L., Horecny, J. & Kostolny, J. 2023. Segmentation of MRI images using clustering algorithms. *IEEE International Conference on Information and Digital Technologies, IDT 2023*. pp. 169-178.

Breiman, L. 2001. Random forests. *Machine Learning* 45: 5-32.

Chawla, N.V., Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16: 321-357.

Chen, T. & Guestrin, C. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 785-794.

Cheng, K., Zhang, C., Yu, H., Yang, X., Zou, H. & Gao, S. 2019. Grouped SMOTE with noise filtering mechanism for classifying imbalanced data. *IEEE Access* 7: 170668-170681.

Dalakleidi, K., Zarkogianni, K., Thanopoulou, A. & Nikita, K. 2017. Comparative assessment of statistical and machine learning techniques towards estimating the risk of developing type 2 diabetes and cardiovascular complications. *Expert Systems* 34(6): e12214.

Department of Environment Malaysia. 2022. Laporan Kualiti Alam Sekeliling 2022. Putrajaya: Jabatan Alam Sekitar Malaysia

Dogo, E.M., Nwulu, N.I., Twala, B. & Aigbavboa, C. 2021. Accessing imbalance learning using dynamic selection approach in water quality anomaly detection. *Symmetry* 13(5): 818.

Dong, X., Yu, Z., Cao, W., Shi, Y. & Ma, Q. 2020. A survey on ensemble learning. *Frontiers of Computer Science* 14(2): 241-258.

Douzas, G., Bacao, F. & Last, F. 2018. Improving imbalanced learning through a heuristic oversampling method based on K-means and SMOTE. *Information Sciences* 465: 1-20.

Ester, M., Kriegel, H.P., Sander, J. & Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD* 96(34): 226-231.

Fitri, A., Maulud, K.N.A., Pratiwi, D., Phelia, A., Rossi, F. & Zuhairi, N.Z. 2020. Trend of water quality status in Kelantan River downstream, Peninsular Malaysia. *Jurnal Rekayasa Sipil (JRS-Unand)* 16(3): 178-184.

Hashem, A.O.A., Ahmad, W.A.A.W. & Yusuf, S.Y. 2021. Water quality status of Sungai Petani River, Kedah, Malaysia. *IOP Conference Series: Earth and Environmental Science* 646(1): 012028.

Jeatrakul, P., Wong, K.W. & Fung, C.C. 2010. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. *Proceedings of the 17th International Conference on Neural Information Processing (ICONIP 2010)*, Part II, Sydney, Australia, pp. 152–159. Springer.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatco, C.D., Silverman, R. & Wu, A.Y. 2002. An efficient *k*-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Learning* 24(7): 881-892.

Kavitha, R.J. & Caroline, B.E. 2015. Hybrid cryptographic technique for heterogeneous wireless sensor networks. *2015 International Conference on Communication and Signal Processing, ICCSP 2015*. pp. 1016-1020.

Kumar, K.M. & Reddy, A.R.M. 2016. A fast DBSCAN clustering algorithm by accelerating neighbor searching using groups method. *Pattern Recognition* 58: 39-48.

Marsboom, C., Vrebos, D., Staes, J. & Meire, P. 2018. Using dimension reduction PCA to identify ecosystem service bundles. *Ecological Indicators* 87: 209-260.

Mustakim, E., Rahmi, M.R., Mundzir, S.T., Rizaldi, Okfalisa & Maita, I. 2021. Comparison of DBSCAN and PCA-DBSCAN Algorithm for Grouping Earthquake Area. *In: Proceedings of the 2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, Taiz, Yemen, pp. 1–5.

Poudevigne-Durance, T. 2024. Generative adversarial networks for the synthesis of unbalanced irregular time series. Doctoral dissertation, Cardiff University (Unpublished).

Rahman, M.A., Hossain, M.F., Hossain, M. & Ahmmed, R. 2020. Employing PCA and T-statistical approach for feature extraction and classification of emotion from multichannel EEG signal. *Egyptian Informatics Journal* 21(1): 23-35.

Sander, J., Ester, M., Kriegel, H.P. & Xu, X. 1998. Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery* 2: 169-194.

Sarker, I.H. 2021. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science* 2(3): 160.

Schapire, R.E. 1999. A brief introduction to boosting. *IJCAI International Joint Conference on Artificial Intelligence* 99(999): 1401-1406.

Shehab, S.A., Darwish, A., Hassanien, A.E. & Scientific Research Group in Egypt. 2023. Water quality classification model with small features and class imbalance based on fuzzy rough sets. *Environment, Development and Sustainability* 27: 1401-1419.

Shen, X., Hu, H., Li, X. & Li, S. 2021. Study on PCA-SAFT imaging using leaky Rayleigh waves. *Measurement* 170: 108708.

Starczewski, A., Goetzen, P. & Er, M.J. 2020. A new method for automatic determining of the DBSCAN parameters. *Journal of Artificial Intelligence and Soft Computing Research* 10(3): 209-221.

Taloor, A.K., Sambyal, S., Sharma, R., Dev, S., Shastri, S. & Kumar, R. 2025. Advanced hydrogeochemical facies classification: A comparative analysis of Machine Learning models with SMOTE in the Tawi basin. *Physics and Chemistry of the Earth, Parts A/B/C* 137: 103785.

Tran, T.N., Drab, K. & Daszykowski, M. 2013. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems* 120: 92-96.

Wong, W.Y., Hasikin, K., Khairuddin, M., Salwa, A., Razak, S.A., Hizaddin, H.F., Mokhtar, M.I. & Azizan, M.M. 2023. A stacked ensemble deep learning approach for imbalanced multi-class water quality index prediction. *Comput. Mater. Contin.* 76(2): 1361-1384.

Yasin, M.I. & Karim, S.A.A. 2020. A new fuzzy weighted multivariate regression to predict water quality index at Perak Rivers. In S. Karim, E. Kadir & A. Nasution (Eds.), *Optimization Based Model Using Fuzzy and Other Statistical Techniques Towards Environmental Sustainability* (pp. 1–27). Singapore: Springer.

*Corresponding author; email: norashikin116@uitm.edu.my