DOI: http://dx.doi.org/10.21111/fij.v9i2.12489
Diterima: 31 Juli 2024 Revisi: 10 September 2024 Terbit: 11 Oktober 2024

Water Quality Identification Using Ensemble Machine Learning and Hybrid Resampling SMOTE-ENN Algorithm

Moch Deny Pratama [1)]*, Rifqi Abdillah [2)], Dina Zatusiva Haq [3)]

Department of Informatics Management, Universitas Negeri Surabaya, Surabaya, Indonesia [1)],
Department of Informatics Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia [2)],
Department of Mathematics, UIN Sunan Ampel Surabaya, Surabaya, Indonesia [3)]
mochpratama@unesa.ac.id [1)], rifqiabdillah@unesa.ac.id [2)]*, zatusivad@gmail.com [3)]*

Abstract

Water is essential for all living organisms, yet only a small fraction is fresh and suitable for consumption. The limited availability of freshwater sources, worsened by pollution, overuse, and climate change, underscores the urgent need for sustainable water management. Traditional water quality identification methods are labour-intensive, slow, and costly. Water quality identification often struggles with data quality, imbalanced datasets, and model interpretability. These challenges lead to inaccuracies, especially in detecting minority classes, which is crucial for identifying pollution. This research explores machine learning (ML) techniques to address the limitations of water quality classification by integrating ensemble learning using LightGBM and hybrid Resampling using SMOTE-ENN. Ensemble learning techniques improve accuracy and robustness by aggregating the strengths of multiple models, effectively handling imbalanced data and reducing overfitting. Hybrid Resampling techniques enhance model sensitivity by generating synthetic minority-class samples and refining datasets through noise reduction. Together, these integrations provide a more reliable framework for water quality identification, enabling timely and accurate. This innovative method offers a robust solution for addressing data imbalance and overfitting, ensuring more effective detection of polluted conditions. This study highlights the importance of advanced ML techniques in improving water quality tasks and underscores LightGBM's effectiveness in handling imbalanced data post-SMOTE-ENN application. This method is known for its superior performance, achieving the highest performance evaluation metrics in water quality classification with accuracy, F1-Score, and increasing the recall value by 3% with values of 94.50%, 94.76% and 93.00%, respectively.

Keywords: *Water Quality, Machine Learning, Imbalanced Data, LightGBM, SMOTE-ENN, Ensemble Learning, Hybrid Resampling.*

Abstrak

Air sangat penting bagi semua organisme hidup, namun hanya sebagian kecil yang segar dan layak untuk dikonsumsi. Terbatasnya ketersediaan sumber air bersih, yang diperburuk oleh polusi, penggunaan berlebihan, dan perubahan iklim, menggarisbawahi kebutuhan mendesak akan pengelolaan air berkelanjutan. Metode identifikasi kualitas air tradisional memerlukan banyak tenaga kerja, lambat, dan mahal. Identifikasi kualitas air sering kali bermasalah dengan kualitas data, kumpulan data yang tidak seimbang, dan kemampuan interpretasi model. Tantangan-tantangan ini menyebabkan ketidakakuratan, terutama dalam mendeteksi kelompok minoritas, yang sangat penting dalam mengidentifikasi polusi. Penelitian ini mengeksplorasi teknik pembelajaran mesin (ML) untuk mengatasi keterbatasan klasifikasi kualitas air dengan mengintegrasikan pembelajaran ensembel menggunakan LightGBM dan pengambilan sampel hybrid menggunakan SMOTE-ENN. Teknik pembelajaran ensemble meningkatkan akurasi dan ketahanan dengan menggabungkan kekuatan beberapa model, menangani data yang tidak seimbang secara efektif, dan mengurangi overfitting. Teknik pengambilan sampel hibrid meningkatkan sensitivitas model dengan menghasilkan sampel kelas minoritas sintetik dan menyempurnakan kumpulan data melalui pengurangan noise. Bersama-sama, integrasi ini memberikan kerangka kerja yang lebih andal untuk identifikasi kualitas air, sehingga memungkinkan dilakukannya identifikasi secara tepat waktu dan akurat. Metode inovatif ini menawarkan solusi yang kuat untuk mengatasi ketidakseimbangan dan overfitting data, sehingga memastikan deteksi kondisi tercemar dengan lebih efektif. Studi ini menyoroti pentingnya teknik ML tingkat lanjut dalam meningkatkan tugas kualitas air dan menggarisbawahi efektivitas LightGBM dalam menangani data yang tidak seimbang pasca penerapan SMOTE-ENN. Metode ini dikenal dengan *kinerjanya yang unggul, mencapai metrik evaluasi kinerja tertinggi dalam klasifikasi kualitas air dengan akurasi, F1-Score, dan meningkatkan nilai recall sebesar 3% dengan nilai masing-masing 94,50%, 94,76% dan 93,00%.*

Kata kunci*:* *Kualitas Air, Pembelajaran Mesin, Data Ketidakseimbangan, LightGBM, SMOTE-ENN, Pembelajaran Ensemble, Pengambilan Sampel Hibrid.*

1. INTRODUCTION

Water covers over two-thirds of Earth's surface, making it a crucial resource for all living organisms. Despite this apparent abundance, only a small fraction is fresh and safe for consumption. Freshwater sources, like rivers, lakes, and aquifers, are limited and unevenly distributed, leading to insufficiency in many regions. Pollution, overuse, and climate change further reduce the availability of clean water, making sustainable management vital for future needs [1]. Water is an essential natural resource importance for humans. It supports basic physiological needs and is essential for sanitation, agriculture, and industry. Socially, it influences public health, cultural practices, and community well-being. Without acceptable water supply, human survival and progress are severely threatened, leading to potential conflicts and humanitarian crises. Currently, over 1.1 billion people lack access to clean drinking water, which poses significant health risks and affects daily life [2]. Water quality in urban environments is influenced by various factors, including industrial discharge, sewage, stormwater runoff, and pollution from vehicles. These contribute to contaminants such as heavy metals, chemicals, and pathogens entering water sources.

Effective management requires monitoring key indicators like contaminants like heavy metals, and microbial content [3]. Advanced technologies, such as machine learning models, are increasingly used to assess and predict water quality, enabling timely interventions to protect public health and ecosystems [4]. Assessing Water Quality (WQ) traditionally involves labor-intensive processes like manual sampling and laboratory analysis, which are slow, costly, and may not provide real-time insights needed for timely interventions [5]. Intelligent systems leveraging Machine Learning (ML) as part of the Artificial Intelligence (AI) scope offer a promising alternative. ML techniques enable systems to autonomously learn patterns from WQ data, enhancing the accuracy and efficiency of identifications [6].

Several machine learning algorithms used in water quality identification research include Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT) and Support Vector Machine (SVM) [7][8]. These algorithms classify water quality data based on parameters such as pH, dissolved oxygen, and turbidity levels, etc., [9] and based on contaminants like heavy metals and microbial content such as aluminium, barium, bacteria, nitrates, uranium, etc., [10]. In the previous research conducted by [11] using comparison various classification algorithms in water quality classification task. The Random Forest model achieved the highest performance metrics. Specifically, an accuracy of 91.00%, precision was 93.00%, recall was 92.00%, and the F1-score, which balances precision and recall, was 91.00% value. The research conducted by [12] shows perfect recall of 100.00% indicates that the Stochastic Gradient Descent (SGD) method is correctly identifying all the positive instances. However, F1-score of 58.8% suggests that the precision is quite low. This inconsistency usually points to a high number of false positives, which can be a sign of an imbalance dataset problem. The research conducted by [13] using XGBoost resulted in an F1-score of 60.00% and a recall of 65.00%. This indicates that the model has moderate effectiveness in identifying positive instances but struggles with precision. The use of these algorithms aids in effectively detecting pollution and predicting changes in water quality. Machine learning methods appearance with several limitations that impact their effectiveness. Its often require large, well-labeled datasets, which can be difficult to obtain [14]. Scalability is another issue, as these models struggle to efficiently process very large datasets. Its also perform poorly with imbalanced data, often favoring the majority class, which can skew results. These challenges highlight the need for more approaches to improve performance and applicability.

Water quality classification often deals with imbalanced data, where certain quality classes are underrepresented compared to others [15]. This imbalance poses challenges for machine learning models, which may become biased towards predicting the majority class [16]. Techniques such as resampling or employing algorithms designed ensemble methods, can help address this issue. These strategies ensure that models are better equipped to accurately identify minority classes, which are often critical for detecting pollution or unsafe conditions. In imbalanced data scenarios like confusion matrix, provides a detailed breakdown of model predictions is crucial for understanding, especially when sensitivity (recall) is important. Sensitivity, calculated as the ratio of true positives to the sum of true positives and false negatives, measures the model's ability to correctly identify positive cases. This is particularly vital when the minority class is of high importance, such as in medical diagnoses or fraud detection, where missing positive instances can have significant consequences. Unlike accuracy, which can be skewed by the majority class, sensitivity ensures that the model's focus is on capturing the critical minority class [17]. Improving performance identification can be achieved by combining ensemble learning with hybrid Resampling techniques [18].

The LightGBM its a powerful ensemble learning technique, superior in boosting model performance through its innovative approach to gradient boosting. It uses multiple weak learners (decision trees) combined to create a strong predictive model [19]. One of its key strengths is speed, as it employs histogram-based algorithms to enhance computational efficiency and reduce memory usage, making it faster than traditional boosting methods [20]. LightGBM introduces innovations like leaf-wise tree growth, which allows for optimal splits, and Gradient-based One-Side Resampling (GOSS), which reduces the need for data instances in each iteration. These innovations contribute to LightGBM's high performance, providing both speed and accuracy, enhance classification performance by aggregating predictions from multiple models, thereby increasing accuracy and robustness [21]. On the other hand, the hybrid Resampling techniques like SMOTE combined with Edited Nearest Neighbors (SMOTE-ENN), address class imbalance by generating synthetic samples for minority classes and removing noise, improving model sensitivity (recall) [22]. Together, these approaches enhance the ability to detect and predict water quality issues, ensuring more reliable and accurate identifications, especially in datasets with imbalanced class distributions.

Based on research on water quality conducted is urgent due to the increasing threats of pollution, overuse, and climate change, which reduce the availability of clean water. The research gap lies in developing real-time, accurate methods for assessing water quality, as traditional techniques are slow, costly, and labour-intensive. The proposed research aims to innovate by leveraging ensemble machine learning, specifically using LightGBM and SMOTE-ENN, to improve water quality identification. LightGBM provides a powerful ensemble learning framework that enhances model accuracy and efficiency, while SMOTE-ENN addresses data imbalance issues, ensuring better detection of minority classes crucial for identifying pollution and hazardous conditions. This combination offers a more reliable and timely approach to water quality identification, enabling proactive interventions and sustainability.

2. RESEARCH METHOD

The research method follows a structured process beginning with the first step is data collection from relevant sources to ensure comprehensive coverage. In the second step, data preprocessing is conducted, which includes cleaning to remove noise and inconsistencies, handling missing values through imputation or removal, and normalizing the data to maintain consistency. After preprocessing, the dataset is split into training and testing subsets. To address class imbalance, hybrid resampling techniques like SMOTE-ENN are applied to the training data, enhancing the model's ability to generalize across classes. The study employs ensemble machine learning, specifically LightGBM combined with SMOTE-ENN, to train models on the balanced data. Model performance is then evaluated using the testing data, with metrics such as accuracy, precision, recall, and F1-score used to determine the most effective approach. This structured process ensures thorough analysis and reliable insights, leveraging the strengths of ensemble learning to achieve superior performance in handling imbalanced datasets. Based on the explanation that has been presented previously, the flow of the research method can be illustrated through the diagram in Figure 1 as follows.

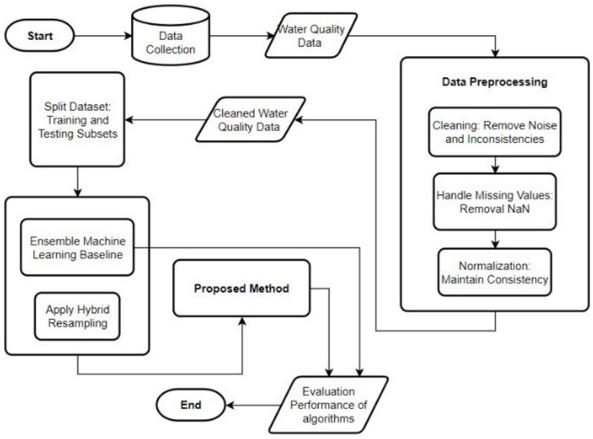
Start
Data Collection
Water Quality Data
Data Preprocessing
Cleaning: Remove Noise and Inconsistencies
Handle Missing Values: Removal NaN
Normalization: Maintain Consistency
Cleaned Water Quality Data
Split Dataset: Training and Testing Subsets
Ensemble Machine Learning Baseline
Apply Hybrid Resampling
Proposed Method
Evaluation Performance of algorithms
End

Figure 1. Research Method Flow

A. Data Collection

Data collection for water quality identification involves gathering samples data from water sources to assess their safety and quality [23]. This process typically includes measuring physical, chemical, and contaminants like heavy metals, and microbial content. Accurate data collection is crucial for identifying pollutants and assessing safety standards.

Table 1. Water Quality Data

Aluminium	barium	bacteria	...	nitrates	uranium	is_safe
1.65	2.85	0.20	...	16.08	0.02	1
2.32	3.31	0.65	...	2.01	0.05	1
1.36	2.96	0.71	...	1.41	0.05	1
0.92	0.20	0.13	...	6.74	0.02	1
1.01	0.58	0.05	...	14.16	0.01	0
...	...	...	...	...	...	...
0.05	1.95	0.00	...	14.29	0.03	1
0.05	0.59	0.00	...	10.27	0.08	1
0.09	0.61	0.00	...	15.92	0.05	1
0.01	2.00	0.00	...	0.00	0.00	1
0.04	0.70	0.00	...	15.92	0.05	1

Table 1. represent of water quality data, indicates a clear distinction between polluted and

neutral (safe) water based on several features. The data includes features such as aluminium, barium, bacteria, nitrates, uranium, etc., with a classification label indicating whether the water is "Polluted" (is_safe = 1) or "Neutral or Safe" (is_safe = 0). Polluted water samples generally have higher concentrations of Aluminium, Barium, Bacteria, and Nitrates, while Uranium levels do not significantly vary between classes. This pattern suggests that Aluminium, Barium, Bacteria, and Nitrates are more indicative of water pollution, making them critical features for classification models. However, the dataset also reflects a challenge of imbalanced classes, as polluted samples (is_safe = 1) are more frequent than neutral samples (is_safe = 0). This imbalance can affect the performance of classification models, making it essential to use techniques like Resampling or specialized algorithms to ensure accurate predictions. Variability in feature values within each class suggests the need for robust models that can handle such diversity in water quality measurements.

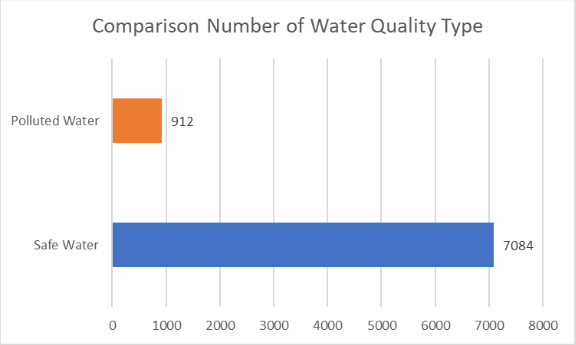

Figure 2. Comparison Water Quality Data

The Figure 2 illustrates a significant imbalance between the number of samples classified as "Safe Water" with value of 7,084 and "Polluted Water" with value of 912. This disparity presents a challenge for modeling, as it can lead to biased predictions favoring the majority class, "Safe Water." To counteract this, techniques like SMOTE-ENN are essential for balancing the training data, ensuring that models can accurately identify "Polluted Water" cases. The imbalance also means that relying solely on accuracy as a performance metric could be misleading, as it might not reflect the model's ability to detect pollutants effectively. Therefore, focusing on metrics like recall for the minority class is crucial to ensure reliable and unbiased model performance in water quality classification.

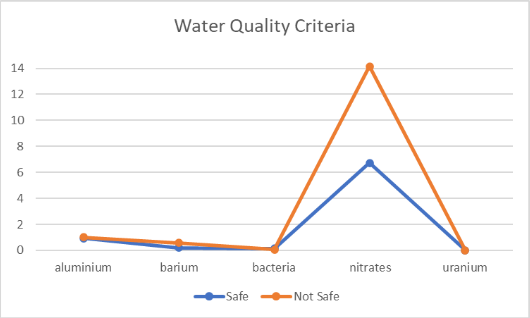

Figure 3. Visualization Quality Criteria

Figure 3 illustrates water quality criteria comparing "Safe" and "Not Safe" levels across various contaminants, such as: aluminum, barium, bacteria, nitrates, and uranium. The "Not Safe" category peaks significantly with nitrates, indicating a higher risk level compared to other contaminants. In contrast, aluminum, barium, and uranium show minimal or no distinction between safe and unsafe levels. The "Safe" category remains consistently low across all contaminants, highlighting potential concerns with water quality safety, particularly for nitrates and bacteria. This suggests that targeted interventions may be necessary to address these specific issues.

B. Preprocessing Data

Data preprocessing is essential for preparing a dataset for effective machine learning model training. It begins with cleaning, which involves removing noise and inconsistencies from the data, such as correcting errors and unifying different formats. This ensures the data is accurate and reliable. Next, handling missing values addresses gaps in the dataset, either by imputing missing values with estimates like the mean or median, or by removing affected records altogether to prevent skewed results. Finally, normalization scales the data to a consistent range or distribution, which helps ensure that all features contribute equally to the model. This step is crucial for algorithms that are sensitive to feature scale. Together, these preprocessing steps improve the quality of the dataset, leading to more reliable and accurate model performance [24].

C. Ensemble Learning Algorithm

This research conducted by ensemble learning algorithm called LightGBM (Light Gradient Boosting Machine) is an efficient and scalable gradient boosting framework developed by Microsoft, known for its speed and performance. It is optimized to handle large datasets and high-dimensional data, making it a popular choice in machine learning applications. LightGBM's key strength lies in its speed, achieved through the use of histogram-based algorithms that improve computational efficiency and reduce memory usage. Its scalability allows for handling massive amounts of data with support for parallel learning. A notable innovation is its leaf-wise tree growth, which results in more optimal splits compared to traditional level-wise growth. The workflow of this method represent in Formula (1) as follows [21].

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \qquad (1)$$

Where the symbol based on formula (1) represent $F(x)$ is the final prediction, α_t is is the weight of the $t = th$ with th is number of iteration model. $h_t(x)$ is the $t = th$ weak learner (decision tree), and T is the total number of trees.

LightGBM also introduces Gradient-based One-Side Resampling (GOSS), focuses on samples with large gradients which reduces data instances per iteration. Exclusive Feature Bundling (EFB), which efficiently handles high-dimensional data by reduces the number of features by bundling mutually exclusive features. These features contribute to its high performance, enabling fast and accurate model training process [21]. Additionally, this supports various data types, including continuous, categorical, and missing values, and offers extensive customization options through its hyperparameters.

D. Hybrid Resampling

This research also conducted hybrid Resampling by SMOTE-ENN, which combines Synthetic Minority Over-Resampling Technique (SMOTE) with Edited Nearest Neighbors (ENN), for handling imbalanced data. Each component works in SMOTE to synthetic sample creation for each minority class sample, its selects k nearest neighbors and generates synthetic samples along the line segments joining the sample and its neighbors. The workflow of this synthetic sample's method represent in Formula (2) as follows [25].

$$x_{new} = x_i + \beta \times (x_{nn} - x_i) \qquad (2)$$

Where the symbol based on formula (2) represent, each component works in ENN to noise removal, removes samples whose class differs from the majority of its k nearest neighbors, reducing noise and cleaning the data. Then combination process to apply SMOTE to Generate synthetic samples for the minority class, and apply ENN to remove noisy samples from the combined data (original + synthetic). This hybrid approach improves model performance by balancing the data and enhancing sensitivity [25].

E. Performance Evaluation

Confusion matrix was applied to provides a comprehensive performance evaluation of classification task [26]. The metrics are defined: True Positive (TP): This indicates that a water sample is classified as polluted, and the classifier correctly identifies it as polluted. True Negative (TN): This occurs when a water sample is classified as clean, and the classifier accurately predicts it to be clean. False Positive (FP): In this case, a clean water sample is incorrectly classified as polluted by the classifier. False Negative (FN): This represents a scenario where a polluted water sample is incorrectly classified as clean by the classifier.

True Label	Not Safe (0)	True Positive (TP)	False Positive (FP)
	Safe (1)	False Negative (FN)	True Negative (TN)
		Not Safe (0)	Safe (1)
		Predicted Label	

Figure 4. Confusion Matrix Scheme

The Figure 4 illustrates confusion matrix scheme, that an accuracy is used to measures the overall correctness of the classifier. Precision is used to evaluates the proportion of correctly identified polluted samples among all samples classified as polluted. Sensitivity (recall) is used to assesses the classifier's ability to correctly identify polluted samples. F1-Score is used to the harmonic mean of precision and recall, providing a balance between the two, shown in Formula (3), (4) and (5) respectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

$$Sensitivity\ (recall) = \frac{TP}{TP+FN} \qquad (4)$$

$$F1 - Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \qquad (5)$$

These metrics are essential for determining the effectiveness of the classification model in identifying water quality accurately, helping ensure safe and clean water management practices.

The ROC curve was applied to visualization of the performance of different machine learning models on a binary classification task. It shows the trade-off between the TPR and FPR for each model at various classification thresholds. True positive rate (TPR), also known as recall, represents the proportion of actual positive cases that were correctly identified by the model. False positive rate (FPR) is the proportion of negative observations that were incorrectly classified as positive. An ideal ROC curve would be a line that goes straight up the left side of the ROC space and then across the top. This would indicate a model with 100% TPR and 0% FPR, which means it perfectly classified all positive and negative cases.

3. RESULTS AND DICUSSION

The performance of various classification algorithms in ensemble learning, specifically focusing on tree-based methods, is crucial for understanding their effectiveness in different scenarios. Here, we provide a detailed analysis of classifiers based on their accuracy, F1-score, and sensitivity (recall) evaluation metrics, highlighting their strengths and weaknesses. Tree-based ensemble methods, such as Random Forest, Gradient Boosting, and LightGBM, are often highly effective due to their ability to handle complex datasets and model interactions between features.

Table 2. Performance Evaluation Research Results

Algorithm	Accuracy (%)	F1-Score (%)	Sensitivity Recall (%)
Random Forest	96.06	95.80	85.32
AdaBoost	92.75	92.12	76.36
Gradient Boosting	95.88	95.64	85.64
Bagging	96.88	96.75	89.21
Extra Trees	93.13	92.29	75.07
XGBoost	96.88	96.79	90.29
LightGBM	**97.00**	**96.92**	**90.57**

The Table 2 represent Random Forest classifier demonstrates robust performance with an accuracy of 96.06% and an F1-Score of 95.80%, although its recall of 85.32% suggests it is slightly less sensitive to positive instances. AdaBoost, with an accuracy of 92.75% and an F1-Score of 92.12%, performs well but has a lower recall of 76.36%, indicating it might not be the best choice for highly imbalanced datasets. Gradient Boosting also shows strong performance with an accuracy of 95.88% and an F1-Score of 95.64%, coupled with a higher recall of 85.64% compared to AdaBoost. Bagging achieves the highest accuracy among the classifiers at 96.88%, along with an F1-Score of 96.75% and a recall of 89.21%, highlighting its balanced performance. Extra Trees, with an accuracy of 93.13% and an F1-Score of 92.29%, has the lowest recall of 75.07% among tree-based methods, indicating a potential weakness in identifying positive instances. The XGBoost, achieving an accuracy of 96.88% and an F1-Score of 96.79%, along with a high recall of 90.29%, shows its effectiveness in both balanced and imbalanced datasets. Lastly, LightGBM outperforms all with an accuracy of 97.00%, an F1-Score of 96.92%, and the highest recall of 90.57%, making it the most reliable choice across different metrics.

In summary, while Bagging and LightGBM exhibit the highest accuracy, classifiers like XGBoost and LightGBM also maintain high F1-Scores and recall values, making them more suitable for imbalanced datasets. The analysis highlights the importance of selecting classifiers based on specific performance metrics adapted to the application's needs.

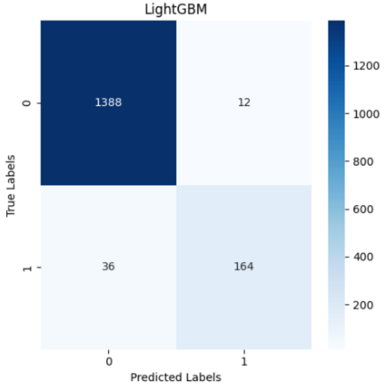

Figure 5. Confusion Matrix LightGBM

The confusion matrix represent in Figure 5 is the performance of the model result from a LightGBM on a classification task. The rows represent the actual labels, and the columns represent the predicted labels. The diagonal cells show the number of correct predictions, and the off-diagonal cells show the number of incorrect predictions. The model shows is not very good at identifying positive cases (those with a label of "0"). It has high precision (most of the times it predicted a label of "0" it was correct), but very low recall (it only identified a small fraction of the actual positive cases). This suggests that the model might be biased towards predicting the negative class.

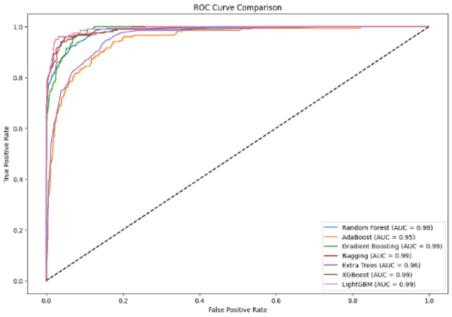

Figure 6. ROC Curve Comparison Results

The ROC curve in Figure 6 is a visualization of the performance of different machine learning models. Random Forest (AUC 0.90), this model's performance is good, but not quite as good as the models mentioned above. It has an AUC of 0.90. XGBoost (AUC 0.93), this model's ROC curve departs from the diagonal line sooner than most of the other models, which indicates it has a good balance between TPR and FPR. It has an Area Under the Curve (AUC) of 0.93, which is a good score. LightGBM (AUC 0.99), this model's ROC curve comes closest to the ideal ROC curve, which means it has the best performance among the models shown, that has an AUC of 0.99, which is a very good score.

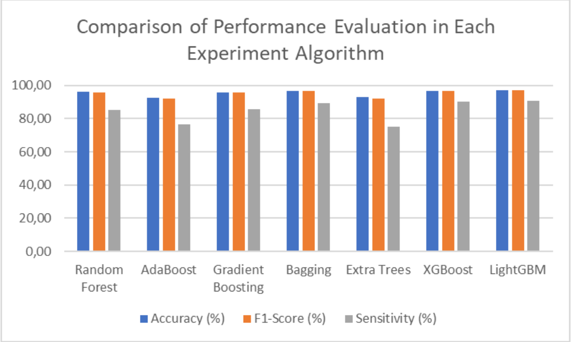

Figure 7. Comparison Performance Results

Figure 7 represent the comparison of performance evaluation in each experiment algorithm. Overall, the chart shows that the LightGBM algorithm has the highest accuracy 97.00%, F1-score 96.92%, and sensitivity 90.57 % compared to the other algorithms in this experiment. This suggests that the LightGBM model performed the best at correctly classifying in this research experiment.

The application of SMOTE-ENN, a hybrid approach combining Synthetic Minority Over-Resampling Technique (SMOTE) and Edited Nearest Neighbors (ENN) represent in Table 3, aims to enhance the sensitivity of classifiers by addressing class imbalance. This analysis evaluates the performance of various classifiers post-Resampling.

Table 3. Performance Evaluation Research Results with Hybrid Resampling Method (SMOTE_ENN)

Algorithm	Accuracy (%)	F1-Score (%)	Sensitivity Recall (%)
Random Forest	94.25	94.46	90.93
AdaBoost	87.63	88.72	83.71
Gradient Boosting	91.19	91.95	91.54
Bagging	93.63	94.03	93.14
Extra Trees	90.81	91.33	85.75
XGBoost	94.13	94.37	91.29
LightGBM	**94.50**	**94.76**	**93.00**

Table 3 represents performance evaluation results, showing that the Random Forest demonstrates superior performance, achieving an accuracy of 94.25% and an F1-Score of 94.46%, with a recall of 90.93%, reflecting its ability to effectively capture minority class patterns. AdaBoost exhibits an accuracy of 87.63% and an F1-Score of 88.72%, with a recall of 83.71%. This suggests that while it benefits from resampling, it may still struggle with highly imbalanced scenarios. Gradient Boosting achieves an accuracy of 91.19% and a strong F1-Score of 91.95%, coupled with a recall of 91.54%, indicating robust performance in detecting positive instances. Bagging presents impressive results with an accuracy of 93.63% and an F1-Score of 94.03%, along with a recall of 93.14%, demonstrating a well-rounded ability to handle imbalanced data. Extra Trees, with an accuracy of 90.81% and an F1-Score of 91.33%, has a recall of 85.75%, showing moderate sensitivity improvements. The XGBoost, with an accuracy of 94.13% and an F1-Score of 94.37%, achieves a recall of 91.29%, reflecting strong performance in sensitivity. LightGBM emerges as the top performer with an accuracy of 94.50% and an F1-Score of 94.76%, along with a recall of 93.00%, underscoring its effectiveness in handling imbalanced data after SMOTE-ENN application.

In summary, applying the Hybrid Resampling Method (SMOTE-ENN) generally leads to improved recall across all algorithms, making them more effective in detecting positive instances in imbalanced datasets. However, this improvement in recall often comes at the cost of a slight decrease in accuracy and F1-Score. The trade-off highlights the importance of selecting the appropriate metrics based on the specific needs of the application, particularly when dealing with imbalanced data where recall is crucial. The proposed method hybrid Resampling SMOTE-ENN significantly enhances sensitivity across LightGBM and Random Forest, with LightGBM emerges as the best performer, showing the most notable improvements generally demonstrates enhanced sensitivity to minority class, making it a valuable approach in scenarios where capturing positive instances is crucial.

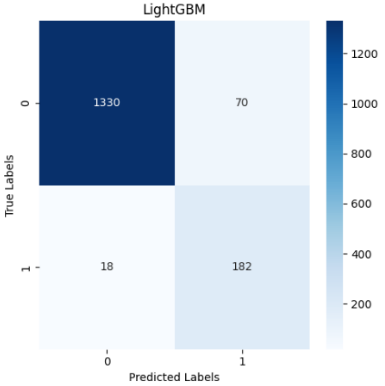

Figure 8. Confusion Matrix LightGBM-(SMOTE-ENN)

The confusion matrix illustrates in Figure 8 shown the performance of the LightGBM-SMOTE-ENN on a classification task. It shows how well the model classified instances into two categories, labeled 0 and 1. The darker blue squares indicate a higher number of correctly classified instances, while lighter areas represent misclassifications. In this case, the model performed very well in predicting class 0, with only a small number of instances incorrectly classified as class 1.

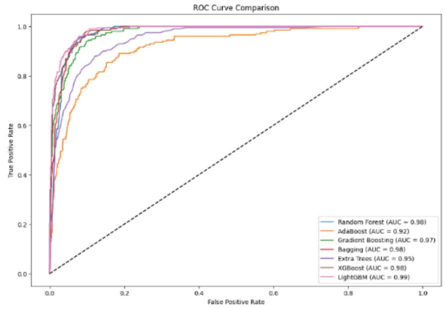

Figure 9. ROC Curve Comparison Results + SMOTE-ENN

Figure 9 visualize the ROC curve shows that the LightGBM model with SMOTE-ENN demonstrates exceptional performance, achieving an AUC (Area Under the Curve) of 0.99, which is very close to the ideal ROC curve. This indicates that the model has an excellent ability to discriminate between positive and negative classes, with a high true positive rate and low false positive rate across various classification thresholds. This superior performance is likely attributed to the combination of the LightGBM algorithm's efficiency and the SMOTE-ENN technique's effectiveness in addressing class imbalance.

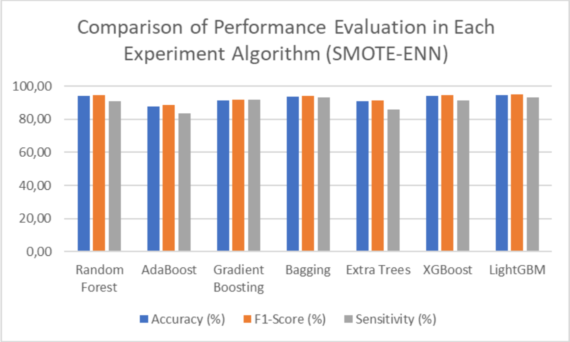

Figure 10. Comparison Performance Results + SMOTE-ENN

Figure 10 represent a performance comparison of several algorithms, including LightGBM, when applied with SMOTE-ENN. It measures accuracy, F1-score, and sensitivity. LightGBM demonstrates the highest performance across all three metrics, significantly outperforming the other algorithms. This indicates that LightGBM, combined with SMOTE-ENN, effectively addresses the class imbalance issue in the dataset and achieves a superior balance between precision and recall.

Table 4 represent the previous research results demonstrate that the proposed method outperforms several established algorithms in terms of classification performance. With an accuracy of 94.50%, an F1-score of 94.76%, and a recall of 93.00%, the proposed method superior in both precision and sensitivity. In comparison, the Random Forest achieves strong metrics with an accuracy of 91.00%, F1-score of 91.00%, and recall of 92.00%, indicating effective performance but still slightly lower than the proposed method. Stochastic Gradient Descent shows a perfect recall of 100.00%, yet its F1-score is only 58.8%, suggesting issues with precision or imbalance problem. The Modified Random Forest and XGBoost have even lower scores, with accuracy and recall metrics that highlight potential weaknesses in handling certain data complexities. The proposed method achieves outstanding performance across all evaluated metrics, demonstrating its effectiveness and robustness in handling classification tasks. With an accuracy of 94.50%, an F1-Score of 94.76%, and a recall of 93.00%, it excels in making accurate predictions, balancing precision and recall, and capturing the majority of positive instances. This superior performance likely stems from several advancements.

Table 4. Comparison of Performance Evaluation with Previous Research Results

Algorithm	Accuracy (%)	F1-Score (%)	Sensitiviy Recall (%)
Random Forest [11]	91.00	91.00	92.00
Stochastic Gradient Descent [12]	**-**	58.8	100.00
Modified Random Forest [27]	77.68	71.00	69.00
XGBoost [13]	-	60.00	65.00
Proposed Method	**94.50**	**94.76**	**93.00**

Research discussion in this study is LightGBM while highly effective, has several limitations to consider. It can consume significant memory, especially with very large datasets, which may be a concern for some applications. The algorithm is sensitive to hyperparameter tuning, requiring careful adjustments to achieve optimal performance. Although it supports categorical features, improper handling can lead to suboptimal results. Additionally, like other ensemble methods, LightGBM often lacks interpretability, making it difficult to understand the underlying decision-making process. Lastly, it may not perform as well on smaller datasets due to its complexity. These factors should be weighed when deciding to use LightGBM for a specific task.

For the future research could benefit from exploring diverse data sources, conducted feature reduction techniques considering that this method consumes significant memory, such as Principal Component Analysis (PCA) [26]. Hyperparameter tuning optimization can be considered [28], and also feature selection algorithms [29], should be examined to enhance models interpretability, efficiency and robustness. Integrating real-time monitoring systems with IoT devices could also facilitate continuous data collection and immediate analysis [30], allowing for timely responses to water quality issues. These approaches have the potential to further advance the effectiveness and applicability of machine learning models in water quality management.

4. CONCLUSION

This research proposed highlights the effectiveness of machine learning classifiers in water quality identification, particularly in addressing class imbalance through advanced techniques. LightGBM consistently demonstrates superior performance, achieving the highest accuracy, F1-Score, and sensitivity recall, making it the most reliable choice across metrics. The implementation of SMOTE-ENN significantly enhances classifier sensitivity, showing notable improvements. These results underscore the importance of selecting appropriate classifiers and using innovative Resampling methods to ensure robust performance, especially in imbalanced datasets. This approach enables more accurate and timely water quality identifications, crucial for effective monitoring and management. LightGBM emerges as the top performer, with an accuracy of 94.50%, an F1-score of 94.76%, and a recall of 93.00%. These results demonstrate its effectiveness in handling imbalanced data, especially after applying SMOTE-ENN.

5. REFERENCE

[1] N. Nasir *et al.*, “Water quality classification using machine learning algorithms,” *J. Water Process Eng.*, vol. 48, p. 102920, 2022.

[2] S. Chidiac, P. El Najjar, N. Ouaini, Y. El Rayess, and D. El Azzi, “A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives,” *Rev. Environ. Sci. Bio/Technology*, vol. 22, no. 2, pp. 349–395, 2023.

[3] P. Vasistha and R. Ganguly, “Water quality assessment of natural lakes and its importance: An overview,” *Mater. Today Proc.*, vol. 32, pp. 544–552, 2020.

[4] S. Zhong *et al.*, “Machine learning: new ideas and tools in environmental science and engineering,” *Environ. Sci. Technol.*, vol. 55, no. 19, pp. 12741–12754, 2021.

[5] A. C. C. Fortes, P. R. G. Barrocas, and D. C. Kligerman, “Water quality indices: Construction, potential, and limitations,” *Ecol. Indic.*, vol. 157, p. 111187, 2023.

[6] N. T. Anh, N. T. Nhan, B. Schmalz, and T. Le Luu, “Influences of key factors on river water quality in urban and rural areas: A review,” *Case Stud. Chem. Environ. Eng.*, p. 100424, 2023.

[7] M. G. Uddin, S. Nash, M. T. M. Diganta, A. Rahman, and A. I. Olbert, “Robust machine learning algorithms for predicting coastal water quality index,” *J. Environ. Manage.*, vol. 321, p. 115923, 2022.

[8] M. Zhu *et al.*, “A review of the application of machine learning in water quality evaluation,” *Eco-Environment Heal.*, vol. 1, no. 2, pp. 107–

116, 2022.

[9] N. Wagle, T. D. Acharya, and D. H. Lee, “Comprehensive Review on Application of Machine Learning Algorithms for Water Quality Parameter Estimation Using Remote Sensing Data.,” *Sensors Mater.*, vol. 32, 2020.

[10] M. Shakerkhatibi, M. Mosaferi, M. Pourakbar, M. Ahmadnejad, N. Safavi, and F. Banitorab, “Comprehensive investigation of groundwater quality in the north-west of Iran: Physicochemical and heavy metal analysis,” *Groundw. Sustain. Dev.*, vol. 8, pp. 156–168, 2019.

[11] M. A. Rahu, A. F. Chandio, K. Aurangzeb, S. Karim, M. Alhussein, and M. S. Anwar, “Towards design of Internet of Things and machine learning-enabled frameworks for analysis and prediction of water quality,” *IEEE Access*, 2023.

[12] S. Kaddoura, “Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability,” *Sustainability*, vol. 14, no. 18. 2022. doi: 10.3390/su141811478.

[13] X. Zhou, C. Liu, A. Akbar, Y. Xue, and Y. Zhou, “Spectral and Spatial Feature Integrated Ensemble Learning Method for Grading Urban River Network Water Quality,” *Remote Sensing*, vol. 13, no. 22. 2021. doi: 10.3390/rs13224591.

[14] M.-A. Katsara, W. Branicki, S. Walsh, M. Kayser, M. Nothnagel, and V. Consortium, “Evaluation of supervised machine-learning methods for predicting appearance traits from DNA,” *Forensic Sci. Int. Genet.*, vol. 53, p. 102507, 2021.

[15] M. Bourel *et al.*, “Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters,” *Water Res.*, vol. 202, p. 117450, 2021.

[16] K. Chen *et al.*, “Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data,” *Water Res.*, vol. 171, p. 115454, 2020, doi: https://doi.org/10.1016/j.watres.2019.115454.

[17] I. D. Mienye and Y. Sun, “Performance analysis of cost-sensitive learning methods with application to imbalanced medical data,” *Informatics Med. Unlocked*, vol. 25, p. 100690, 2021.

[18] T. Kimura, “CUSTOMER CHURN PREDICTION WITH HYBRID RESAMPLING AND ENSEMBLE LEARNING.,” *J. Manag. Inf. Decis. Sci.*, vol. 25, no. 1, 2022.

[19] I. K. Nti, A. Zaman, O. Nyarko-Boateng, A. F. Adekoya, and F. Keyeremeh, “A predictive analytics model for crop suitability and productivity with tree-based ensemble learning,” *Decis. Anal. J.*, vol. 8, p. 100311, 2023, doi: https://doi.org/10.1016/j.dajour.2023.100311.

[20] M. Tang *et al.*, “An Improved LightGBM Algorithm for Online Fault Detection of Wind Turbine Gearboxes,” *Energies*, vol. 13, no. 4. 2020. doi: 10.3390/en13040807.

[21] I. D. Mienye and Y. Sun, “A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects,” *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[22] U. Ependi, A. F. Rochim, and A. Wibowo, “A Hybrid Sampling Approach for Improving the Classification of Imbalanced Data Using ROS and NCL Methods,” *Int. J. Intell. Eng. Syst.*, vol. 16, no. 3, pp. 345–361, 2023.

[23] “Water Quality Data.” [Online]. Available: https://www.kaggle.com/datasets/mssmartypants/water-quality/data

[24] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques,” *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, 2022.

[25] Z. Xu, D. Shen, T. Nie, and Y. Kou, “A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data,” *J. Biomed. Inform.*, vol. 107, p. 103465, 2020.

[26] D. C. R. Novitasari, A. N. Ramadanti, and D. Z. Haq, “Enhancing Covid-19 Diagnosis: Glrlm Texture Analysis And Kelm For Lung X-Ray Classification,” *Fountain Informatics J.*, vol. 9, no. 1, 2024.

[27] W. Y. Wong *et al.*, “Water quality index using modified random forest technique: assessing novel input features,” *C. Model. Eng. Sci.*, vol. 132, no. 3, pp. 1011–1038, 2022.

[28] M. Moeini, “Hyperparameter tuning of supervised bagging ensemble machine learning model using Bayesian optimization for estimating stormwater quality,” *Sustain. Water Resour. Manag.*, vol. 10, no. 2, p. 83, 2024.

[29] T. H. Pham and B. Raahemi, “Bio-inspired feature selection algorithms with their applications: a systematic literature review,” *IEEE Access*, vol. 11, pp. 43733–43758, 2023.

[30] K. Zovko, L. Šerić, T. Perković, H. Belani, and P. Šolić, “IoT and health monitoring wearable devices as enabling technologies for sustainable enhancement of life quality in smart environments,” *J. Clean. Prod.*, vol. 413, p. 137506, 2023.