Performance Evaluation of Classification Methods Utilizing Resampling Techniques for Water Quality Prediction on Imbalanced Data

Rahmi Fadhilah
Department of Statistics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
rahmifadhilah11072000@gmail.com (corresponding author)

Heri Kuswanto
Department of Statistics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
heri.kuswanto@its.ac.id

Dedy Dwi Prastyo
Department of Statistics, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
dedy-dp@statistika.its.ac.id

Received: 30 April 2025 | Revised: 16 June 2025 | Accepted: 28 June 2025

 DOI: https://doi.org/10.48084/etasr.11832

ABSTRACT

Commonly observed challenges in water quality anomaly detection using Machine Learning (ML) classifiers include unbalanced class distribution and missing data. Classifiers trained on such imbalanced datasets often exhibit biased accuracy, favoring the majority class and neglecting the minority class, while incomplete datasets limit the applicability of more complex models and hinder thorough analysis. This research addresses the handling of incomplete data and class imbalance by proposing a robust framework for an ML-based water quality anomaly detection system using several resampling techniques. A comparative study was conducted on six imputation methods for missing data, including Expectation Maximization (EM) and Multiple Imputation by Chained Equations (MICE), alongside three resampling techniques: Random Under Sampling (RUS), Rapidly Converging Gibbs (RACOG) sampler, and RACOG combined with RUS (RACOG-RUS). These methods were evaluated across three classifiers: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Naïve Bayes (NB). The models were assessed using stratified 5-fold cross-validation and evaluated based on accuracy, Receiver Operating Characteristic Area Under Curve (ROC-AUC), and F1-score. Further experiments incorporated feature selection methods such as Boruta and Mean Decrease Accuracy (MDA) to optimize performance. Results demonstrate that RF combined with RACOG-RUS and EM achieved the highest F1-score of 0.9954, effectively addressing both class imbalance and missing data. Additionally, computational analysis highlights the efficiency of RF when optimized with appropriate hyperparameters.

Keywords-water quality monitoring; machine learning classifier; class imbalance; missing value methods

I. INTRODUCTION

Drinking water is a limited resource globally and a fundamental human right essential for health and survival. Despite significant advancements in water treatment technologies and regulatory frameworks ensuring drinking water safety, the global demand for clean water remains a persistent challenge [1]. The main causes of water contamination include industrial activities, agricultural runoff, urbanization, and environmental changes [2]. These sources introduce pollutants that alter various water quality parameters, rendering water unfit for consumption, such as arsenic, which is particularly concerning [3]. The assessment of water quality is of utmost importance in various sectors, such as in aquaculture systems, where water quality directly affects aquatic organism survival [4]. Additionally, according to the World Health Organization, over two billion people worldwide lack access to safe drinking water, underscoring the importance of water conservation and quality maintenance as a global sustainability goal [5].

Microbial contamination in water can be detected through traditional laboratory techniques or modern approaches such as biosensors and optical methods [6–9]. Conventional monitoring, which involves manual sampling and laboratory testing, is resource-intensive and time-consuming [10]. While accurate, this method is unsuitable for real-time or large-scale monitoring, particularly in low-resource settings.

Recent technological advances have positioned Artificial Intelligence (AI) and Machine Learning (ML) at the core of various applications. These technologies offer potential for real-time water quality prediction and monitoring, surpassing the limitations of traditional techniques in speed and scalability [11]. However, the effective implementation of ML for water potability assessment is complex. Sensor data may be noisy, incomplete, or inconsistent, leading to decreased model reliability and performance [12]. Therefore, explainable AI is essential to build trust among stakeholders and ensure the interpretability of ML model outputs [13].

Authors in [14] proposed several deep learning–based approaches for anomaly detection; however, they often failed to address the issue of class imbalance, where instances of poor water quality are underrepresented. Additionally, imputation methods such as Expectation Maximization (EM) and k-Nearest Neighbors (k-NN) typically overlook the spatial and temporal dependencies inherent in in situ water quality data. Given the heterogeneity of such datasets, normalization or standardization is essential for effective ML processing [15].

Recent studies have also demonstrated the effectiveness of Artificial Neural Networks (ANNs) and feature selection strategies in classification tasks across environmental and agricultural domains. For example, transfer learning with ANN-based ensemble models has been applied to enhance water quality classification in agricultural settings [16]. Multi-Layer Perceptron (MLP) models have shown success in crop disease identification [17], livestock health monitoring [18], and household-level data classification [19]. Moreover, ANN-based feature engineering has been used to improve predictive performance in water quality modeling [20]. These findings underscore the importance of combining feature selection and resampling techniques to strengthen classifier performance in the presence of missing or imbalanced data.

This study addresses key challenges in ML–based water quality modeling by proposing and evaluating three supervised classifiers, Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Naïve Bayes (NB), for predicting drinking water suitability using various water quality parameters.

The main contributions of this experimental study include:

- Evaluating the impact of six missing data imputation methods on the performance of three ML classifiers for drinking water suitability classification.
- Investigating how combinations of imputation and resampling methods affect class imbalance in water quality datasets.
- Exploring the effectiveness of preprocessing techniques, such as RF–based feature selection, in enhancing model robustness to incomplete and imbalanced data.
- Performing systematic hyperparameter optimization of the RF classifier using grid search or similar methods.
- Proposing a hybrid classification framework that integrates imputation, resampling, and feature importance analysis to improve predictive accuracy and model stability.

II. EMPIRICAL FRAMEWORK

A. Experimental Setup

The experiments were implemented and simulated using the R programming language, while data analysis was conducted in the Python Anaconda environment using Spyder. This study investigates the effects of four methodological dimensions on anomaly detection in water quality classification: feature selection, class imbalance, missing data, and predictive model choice. A six-directional resampling strategy was employed, encompassing three classifiers, one missing data imputation technique, three resampling methods, feature selection, stratified 5-fold cross-validation, and six evaluation metrics (Figure 1). Experiments were performed on a system equipped with a 12th Generation Intel® Core™ i7-1260P processor (2.10 GHz) and 16 GB of RAM.

The study aims to enhance predictive accuracy and anomaly detection in water quality classification, particularly for imbalanced datasets with missing data collected from the Bengawan Solo region. The parameter settings and methods used are detailed in Table I.

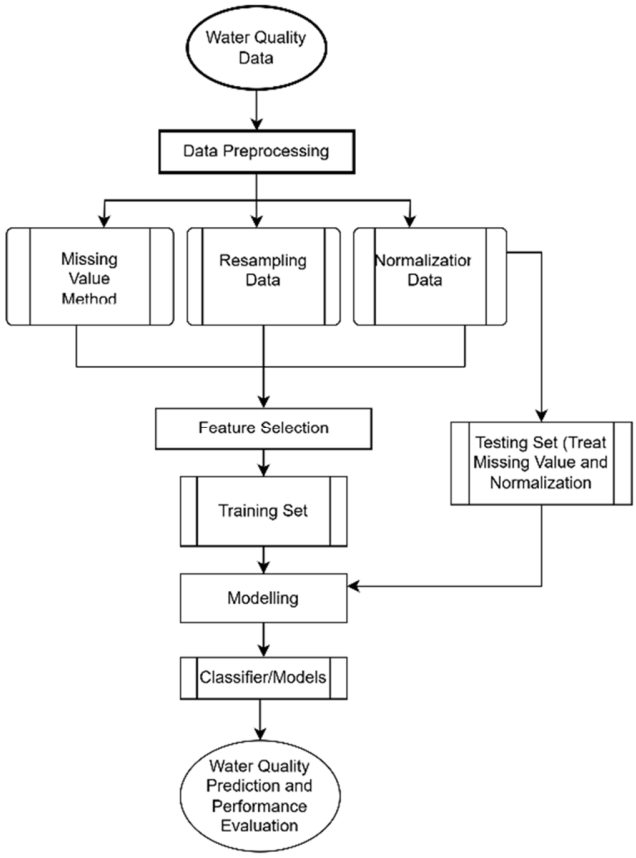

Fig. 1. The experimental evaluation procedure.

B. Datasets

The secondary river water quality data used in this study were obtained from a publicly accessible data portal managed by the Directorate General of Water Resources and the

Bengawan Solo River Basin Organization [21]. The dataset comprises 720 manually collected samples from 30 monitoring stations distributed across the city of Solo, recorded between January 2022 and December 2023. Each sample includes one dependent variable, water quality status, and eight independent variables: pH, Total Dissolved Solids (TDS), Total Suspended Solids (TSS), temperature, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), and nitrate (NO_3).

Water quality status is classified into two categories: 0 (polluted) and 1 (unpolluted). Of the 720 samples, 602 are labeled as polluted and 118 as unpolluted, yielding an approximate class imbalance ratio of 5:1. This significant imbalance indicates a predominance of polluted water samples. Additionally, some values were Missing Completely at Random (MCAR).

TABLE I. EXPERIMENTAL STUDY METHODS

Methodological Category	Methods	Parameter Setting	References
Missing Value	EM	- Iterations: 100 - Convergence threshold: 0.001 - Algorithm: MLE - Data assumption: Multivariate normal distribution	[24, 25]
	Replace with Zero / Listwise Deletion	- Strategy = 'constant' - Fill_value: 0	[19]
	Mean Imputation	- strategy = 'mean'	[19]
	Mode Imputation	- strategy = 'most_frequent'	[19]
	Multiple Imputation by Chained Equations (MICE)	- Default parameter settings	[19]
	RF Imputation (MissForest)	- n_estimators: 10 - criterion: 'entropy'	[21]
Data-level Resampling	Random Under Sampling (RUS)	- Sampling strategy: auto - Reduces the majority class samples to match the minority - Applicable for binary and multiclass problems	[23]
	Rapidly Converging Gibbs (RACOG) Sampler	- Iterations: 50 - Sampling strategy: Bayesian Gibbs sampling - Focuses on the minority class generation - Effectively for highly imbalanced datasets	[21–23]
	Hybrid RACOG-RUS	- Combines RACOG for minority class oversampling and RUS for balancing the majority class - Balances data for improved classifier performance	[22, 23]
Feature Selection	Mean Decrease Gini (MDG)	- Measures feature importance using Gini impurity reduction - Applied within decision tree-based algorithms like RF	[24, 25]
	Mean Decrease Accuracy (MDA)	- Evaluates importance by permuting feature values and calculating impact on accuracy - A higher decrease indicates higher importance	[26]
	Boruta algorithm	-Wrapper-based method using RF -Iteratively removes irrelevant features - Outputs a ranked list of important features	[27, 28]
	Variable Selection Using Random Forest (VSURF)	- ntree: 2000 (the number of trees in the RF for each selection stage). - mtry: √p (number of features randomly selected for each split, with p as the total number of features). - nfor.thres: 50. - nfor.interp: 25. - nfor.pred: 25. - parallel: FALSE	[29]
Classifier	RF	- n_estimators: 100 - ntree: 500 Criterion: Entropy - Max_depth: None - Min_samples_split: 2 - Min_samples_leaf: 1	[21]
	XGBoost	- Objective: Binary: Logistic - Max_depth: 6 - Learning_rate: 0.3 - n_estimators: 100 - Booster: gbtree	[22]
	NB	- Assumes independence between predictors - Continuous variables follow a Gaussian distribution - Output: Probabilistic prediction	[23]
Cross Validation	Stratified 5-fold cross-validation	- N_splits: 5 - Shuffle: True - Random_state: 1	[21]

C. Performance Metrics for Imbalanced Classification

Commonly used metrics for the evaluation of classification models include accuracy, sensitivity, specificity, precision, recall, Balanced Accuracy (BA), the Receiver Operating Characteristic Area Under Curve (ROC-AUC), F1-score, Geometric mean (G-mean), and Matthews Correlation Coefficient (MCC) [22]. However, accuracy can be misleading in imbalanced settings, as it tends to be biased toward the majority class. Therefore, using a combination of metrics provides a more comprehensive and reliable assessment of model performance [23].

In this study, F1-score, BA, and ROC-AUC are selected, as they have demonstrated effectiveness in anomaly detection tasks for water quality classification, particularly in addressing class imbalance issues [23, 30].

$$\mathrm{F1-score} = 2\frac{\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \quad (1)$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \quad (2)$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad (3)$$

BA accounts for both sensitivity and specificity and reduces bias due to class imbalance:

$$\mathrm{BA} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right) \quad (4)$$

The ROC-AUC measures the model's ability to distinguish between the positive and negative classes across all classification thresholds. It is given by:

$$\mathrm{ROC-AUC} = \frac{1+\mathrm{TPR}-\mathrm{FPR}}{2} = \frac{1+\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}-\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}}{2} \quad (5)$$

where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative, TPR = True Positive Rate, and FPR = False Positive Rate.

III. RESULTS AND DISCUSSION

A. Experiment 1: Comparison of Missing Data Methods without Resampling and Feature Selection

Figure 2 illustrates the original dataset before resampling, showcasing the imbalance of the dataset between polluted (class 0) and unpolluted (class 1) water.

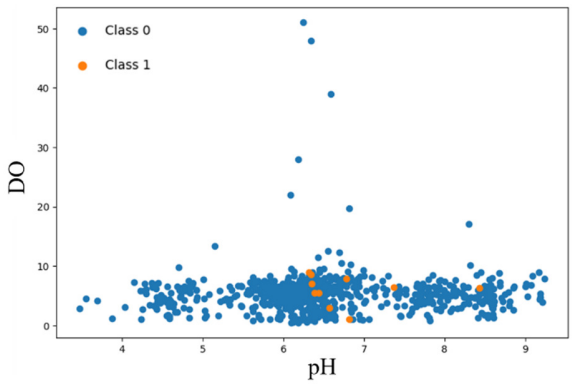

Fig. 2. Data prior to the implementation of resampling techniques.

The comparison of the three classifiers under the six data imputation techniques was initially conducted without applying any resampling methods or feature selection. The evaluation results are summarized in Table II.

Among the methods tested, XGBoost with mean imputation achieved the highest performance, with an F1-score of 0.9690 and BA of 0.5845, followed closely by RF with an F1-score of 0.9464 and BA of 0.5732. Zero imputation yielded the best result for NB, achieving a BA of 0.6429, though its F1-score and ROC-AUC values were comparatively lower. All classifiers produced relatively similar ROC-AUC scores (ranging from 0.5070 to 0.5370), indicating that the choice of imputation method did not significantly impact the model's ability to separate the classes. However, the F1-score varied more substantially, especially for RF and XGBoost, revealing that imputation strategy plays a critical role in the trade-off between precision and recall. Although overall performance was not uniformly high across all settings, both EM and MissForest demonstrated consistent performance across the classifiers, suggesting that they offer more stable handling of missing data.

B. Experiment 2: Comparison of Missing Data Methods with Resampling

The F1-score results of the comparison between the three classifiers under the six data imputation techniques after the implementation of RUS, RACOG, and RACOG-RUS resampling are summarized in Table III.

Among all tested configurations, RF consistently outperformed the other classifiers across various imputation and resampling combinations. For RF, EM imputation combined with RUS achieved an F1-score of 0.8936, while combinations of EM and RACOG yielded scores of 0.9373. Notably, the hybrid RACOG-RUS method paired with EM or MICE imputation produced strong results, achieving F1-scores of 0.9270 and 0.8783, respectively. In contrast, simple imputation methods such as zero-filling led to poorer performance, with F1-scores of 0.5809 for RUS and 0.7185 for RACOG-RUS.

XGBoost showed optimal performance with the EM and RACOG combination, achieving an F1-score of 0.9291. The RACOG-RUS variant also performed similarly well (0.9172). In contrast, simpler imputation strategies, such as zero-fill, resulted in F1-scores of 0.5725 with RUS and 0.9165 with RACOG.

NB yielded the lowest overall performance among the three classifiers. However, its best result was achieved with the EM and RACOG combination (0.8780), followed by RACOG-RUS (0.8540). Other notable combinations include MICE and RUS (0.7141), while simpler imputation strategies such as zero-fill and mean imputation underperformed with the RACOG-RUS method, scoring 0.8137 and 0.8235, respectively. Mode imputation delivered slightly better results with RACOG-RUS, reaching an F1-score of 0.8336.

These findings align with existing literature, reinforcing the importance of using sophisticated imputation and resampling

techniques in imbalanced classification tasks such as water quality anomaly detection.

C. Experiment 3: Comparison of Missing Data Methods with Resampling and Feature Selection

The F1-score results of the comparison between the three classifiers under the six data imputation techniques, after the implementation of resampling and feature selection methods MDA, MDG, Boruta, and VSURF, are summarized in Table III. Additionally, the RF hyperparameters have been optimized via grid search, resulting in: $n_{estimator} = 200$, $max_{depth} = 15$, $min_{samplesSplit} = 2$, $min_{samplesLeaf} = 1$, $max_{features} = sqrt$, and $bootstrap = true$.

According to the results in Table IV, the highest performance was achieved using RF optimized with grid search, EM imputation, RACOG resampling, and MDA feature selection, resulting in an F1-score of 0.9954. This configuration significantly improved both model stability and accuracy in the presence of missing values and class imbalance. XGBoost also performed well, achieving a maximum F1-score of 0.9870 with zero-filling imputation, RACOQ resampling, and VSURF feature selection. NB, while stable, was consistently underperformed, with a highest recorded F1-score of 0.8935 with zero-filling imputation, RACOQ resampling, and MDA feature selection.

TABLE II. RESULTS FOR EVALUATING CLASSIFIERS WITH MISSING VALUES METHODS

Missing Value Methods	RF			XGBoost			NB		
	BA	F1-score	ROC-AUC	BA	F1-score	ROC-AUC	BA	F1-score	ROC-AUC
Filling with a value (value=0)	0.5535	0.9070	0.5229	0.5690	0.9380	0.5327	**0.6429**	0.4859	0.5133
Filling with a value (value=mean)	**0.5732**	**0.9464**	0.5194	**0.5845**	**0.9690**	0.5348	0.5147	0.5295	0.5130
Filling with a value (value=mode)	0.5704	0.9408	0.5233	0.5288	0.9577	0.5208	0.5943	0.4887	0.5136
EM	0.5633	0.9267	0.5276	0.5746	0.9492	0.5326	0.5866	0.4732	0.5130
MICE	0.5584	0.9169	0.5072	0.5676	0.9352	0.5273	0.5950	0.4901	0.5125
MissForest	0.5669	0.9338	0.5255	0.5746	0.9492	0.5369	0.5915	0.4830	0.5136

TABLE III. F1-SCORE RESULTS FOR EVALUATING CLASSIFIER WITH MISSING VALUES AND RESAMPLING

Missing Value Methods	RF			XGBoost			NB		
	RUS	RACOG	RACOG- RUS	RUS	RACOG	RACOG-RUS	RUS	RACOG	RACOG-RUS
Filling with a value (value=0)	0.5809	0.8619	0.7185	0.5725	0.9151	0.9165	0.6549	0.5099	0.8137
Filling with a value (value=mean)	0.7387	0.8774	0.8190	0.5704	0.9208	0.9164	0.5514	0.5028	0.8235
Filling with a value (value=mode)	0.6823	0.8647	0.8180	0.5704	0.8908	0.5162	0.6049	0.5099	0.8336
EM	**0.8936**	**0.9373**	**0.9270**	**0.8361**	**0.9291**	**0.9172**	**0.8090**	**0.8780**	**0.8540**
MICE	0.833	0.8961	0.8783	0.5725	0.8951	0.8865	0.7141	0.8028	0.7445
MissForest	0.6774	0.8549	0.8676	0.5725	0.8851	0.8964	0.7049	0.6451	0.7137

TABLE IV. F1-SCORE RESULTS FOR EVALUATING CLASSIFIER WITH MISSING VALUES AND RESAMPLING AND FEATURE SELECTION

Imputation Methods	Feature Selection Methods	RF			XGBoost			NB		
		RUS	RACOG	RACOG-RUS	RUS	RACOG	RACOG-RUS	RUS	RACOG	RACOG-RUS
Filling with a value (value=0)	MDA	0.9125	0.9395	0.9372	0.9283	0.9631	0.9652	0.8896	0.8935	0.8884
	MDG	0.9030	0.9384	0.9371	0.9275	0.9850	0.9647	0.8881	0.8929	0.8873
	Boruta	0.9135	0.9374	0.9373	0.9268	0.9860	0.9638	0.8878	0.8925	0.8868
	VSURF	0.9138	0.9381	0.9374	0.9272	0.9870	0.9651	0.8882	0.8930	0.8875
Filling with a value (value=mean)	MDA	0.9032	0.9382	0.9497	0.9261	0.9855	0.9635	0.8875	0.8928	0.8871
	MDG	0.9130	0.9379	0.9491	0.9256	0.9865	0.9629	0.8871	0.8926	0.8867
	Boruta	0.9127	0.9377	0.9489	0.9248	0.9860	0.9621	0.8868	0.8923	0.8865
	VSURF	0.9128	0.9378	0.9492	0.9250	0.9867	0.9624	0.8870	0.8927	0.8869
Filling with a value (value=mode)	MDA	0.9122	0.9375	0.9486	0.9138	0.9858	0.9619	0.8864	0.8921	0.8862
	MDG	0.9020	0.9374	0.9483	0.9139	0.9845	0.9615	0.8862	0.8920	0.8859
	Boruta	0.9021	0.9373	0.9485	0.9042	0.9850	0.9617	0.8863	0.8922	0.8860
	VSURF	0.9123	0.9376	0.9487	0.9124	0.9865	0.9620	0.8865	0.8924	0.8861
MICE	MDA	0.9119	0.9497	0.9482	0.9138	0.9856	0.9612	0.8861	0.8919	0.8858
	MDG	0.9117	0.9491	0.9480	0.9034	0.9848	0.9608	0.8858	0.8917	0.8857
	Boruta	0.9018	0.9489	0.9481	0.8937	0.9852	0.9611	0.8860	0.8920	0.8859
	VSURF	0.9119	0.9492	0.9484	0.8940	0.9860	0.9614	0.8862	0.8922	0.8861
MissForest	MDA	0.9116	0.9370	0.9479	0.9032	0.9753	0.9606	0.8857	0.8916	0.8856
	MDG	0.9014	0.9368	0.9477	0.8827	0.9843	0.9602	0.8855	0.8915	0.8854
	Boruta	0.9115	0.9369	0.9478	0.8730	0.9851	0.9605	0.8856	0.8917	0.8855
	VSURF	0.9016	0.9370	0.9479	0.8733	0.9854	0.9607	0.8858	0.8918	0.8857
EM	MDA	0.9118	**0.9954**	0.9513	0.9136	0.9371	**0.9513**	0.8859	0.8722	**0.8922**
	MDG	0.9016	0.9769	0.9501	0.9128	0.9846	0.9503	0.8855	0.8917	0.8856
	Boruta	0.9117	0.9272	0.9495	0.9232	0.985	0.9206	0.8856	0.8919	0.8856
	VSURF	0.912	0.9573	0.9504	0.9135	0.9852	0.9511	0.886	0.8923	0.8858

These findings are consistent with previous research. For instance, authors in [27] highlighted the superior stability and accuracy of RF under conditions of missing data and class imbalance, while authors in [28] emphasized RF's resilience to class imbalance. Furthermore, authors in [29] observed that simple imputation methods, such as mean or mode imputation, are generally less effective than EM, which offers superior performance in handling incomplete and imbalanced data.

D. Statistical Test

Given that the dataset did not meet the assumptions of normality and homogeneity of variance, which are essential conditions for parametric tests, a nonparametric statistical analysis was conducted. Specifically, the Friedman test was chosen to compare the classifiers' performance across multiple settings, as recommended in [29]. Where the Friedman test indicated statistically significant differences, a Nemenyi post hoc test was applied. This test determines whether the average ranks of classifiers differ by more than the Critical Distance (CD), which depends on the number of classifiers, datasets, performance metrics, and the significance level ($\alpha = 0.05$) [29]. The Friedman test statistic is calculated as:

$$X_F^2 = \frac{12D}{K(K+1)} \sum_{j=1}^{K} \left(AvegR_j - \frac{K(K+1)}{2} \right)^2 \quad (6)$$

$$AvegR_j = \frac{1}{D} \sum_{i=1}^{D} r_i^j$$

where K is the number of classifiers, D is the number of datasets, and r_i^j is the rank of the j^{th} classifier on the i^{th} dataset. The CD for the Nemenyi test is given by:

$$CD = q_\alpha \sqrt{\frac{K(K+1)}{6D}} \quad (7)$$

where q_α is the critical value obtained from the Studentized range statistic with a significance of $\alpha = 0.05$. The statistical test was performed in each experiment independently, and the results are summarized in Table V.

TABLE V. RESULTS FROM ALL THE EXPERIMENTS ON THE FRIEDMAN TEST

Experiment	Friedman Test Statistic	P-Value
Missing Value Imputation Method + Classifier	43.42	$2.33 \cdot 10^{-5}$
Missing Value Imputation + Resampling + Classifier	61.78	$9.87 \cdot 10^{-7}$
Missing Value Imputation + Resampling + Feature Selection + Classifier	73.12	$1.35 \cdot 10^{-8}$

For the first experiment, employing only imputation techniques, the Friedman test yielded a statistic of 43.42 with a p-value of $2.33 \cdot 10^{-5}$, indicating significant performance differences. The Nemenyi post hoc test identified RF as the top-performing classifier, followed by XGBoost and NB. Figure 3 illustrates the average performance rankings, showing RF's significant superiority, with its rank nearly exceeding the CD threshold. For the second experiment, the Friedman test produced a statistic of 61.78 with a p-value of $9.87 \cdot 10^{-7}$, confirming significant differences among classifiers. According to the Nemenyi test, RF again ranked highest, especially when paired with RACOG-RUS. Figure 3 shows the average rankings, with a CD of 1.15, highlighting RF's consistent superiority. For the third experiment, the Friedman test yielded a test statistic of 73.12 and a p-value of $1.35 \cdot 10^{-8}$, strongly indicating significant differences. Nemenyi's post hoc analysis reaffirmed RF as the most effective classifier. Although XGBoost improved in performance when paired with Boruta, it remained second to RF. Figure 4 displays the average ranks, with a CD of 1.25, and RF leading in every configuration.

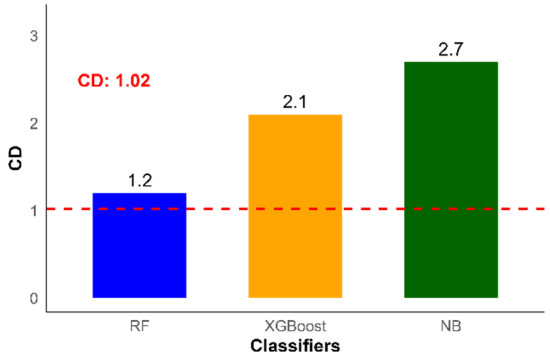

Fig. 3. Average ranks and CD diagram for classifier performance with missing value imputation methods.

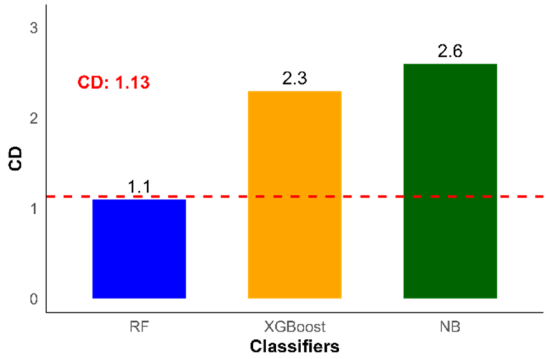

Fig. 4. Average ranks and CD diagram for classifier performance with missing value imputation and resampling methods.

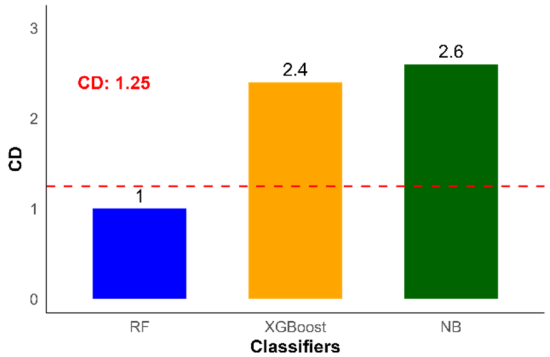

Fig. 5. Average ranks and CD diagram for classifier performance with missing value imputation, resampling, and feature selection methods.

E. Computational Complexity

This research also investigated the computational complexity of the models, with a primary focus on time complexity, which refers to the execution time required to process and optimize learning tasks. To evaluate time complexity, execution times were measured for each algorithm across two hardware platforms: i) a local PC with Intel® Core™ i7-1260P CPU @ 2.10 GHz, 16 GB RAM, and ii) an AWS EC2 with 4 CPU cores and 16 GB RAM. The results are summarized in Table VI.

TABLE VI. COMPARATIVE ANALYSIS OF MODEL RUNTIMES

Model	Local PC (s)	AWS EC2 Instance (s)	Difference (s)	Difference (%)
NB	0.25	0.15	0.1	40.00
RF	58.32	37.41	20.91	35.90
XGBoost	89.56	58.24	31.32	35.00
RUS	15.48	9.87	5.61	36.20
RACOG	78.91	50.62	28.29	35.80
RACOG-RUS	102.35	65.98	36.37	35.50

The results show that NB had the shortest execution time, with an improvement of 0.10 seconds (40.00%) on the AWS EC2 instance compared to the local PC. This makes NB the most computationally efficient among the evaluated classifiers. In contrast, RF and XGBoost required significantly more time, with the EC2 instance executing 35.90% and 35.00% faster, respectively, than the PC. This highlights the scalability benefits of cloud infrastructure for more complex models.

As expected, resampling techniques introduced additional computational overhead due to the complexity of handling imbalanced data. Nevertheless, their performance also improved on the AWS EC2 instance, showing execution time reductions of 36.20%, 35.80%, and 35.50%, respectively.

These results emphasize the trade-off between computational cost and model performance. Although advanced techniques such as RACOG-RUS improve classification outcomes, they come at the expense of higher runtime. However, deploying such models on cloud infrastructure like AWS EC2 can mitigate this cost significantly by accelerating execution without sacrificing accuracy.

F. Limitation of Study

This study presents a comparative analysis of RF, XGBoost, and NB classifiers on imbalanced datasets, integrating systematic strategies for handling missing values and class imbalance through imputation and resampling. However, several limitations should be noted:

- Model sensitivity: Algorithms like NB are highly sensitive to class imbalance, often yielding suboptimal results when applied to unseen test data. Although RF and XGBoost demonstrate greater resilience, their performance still benefits significantly from appropriate resampling techniques.
- Computational cost: Feature selection techniques such as Boruta demonstrated improved model performance but incurred considerable computational cost, particularly on large datasets. This highlights the need for more efficient optimization strategies for scalability.
- Hyperparameter optimization: While this study uses grid search to optimize RF, default parameters were applied in many cases. Further tuning, particularly for tree-based and ensemble methods, could yield even better predictive performance [31, 32].
- Problem scope: The study is confined to binary classification tasks with imbalanced data. As such, the conclusions may not generalize to multiclass classification problems or datasets with varied distribution patterns. Future work should extend the evaluation to more complex classification tasks and data settings.

G. Validity Threats

This section outlines potential threats to the validity of the experimental findings, aiming to ensure that conclusions are supported by the data and that observed outcomes reflect the real effects of the methods applied [33].

- Internal validity: To mitigate internal validity threats, all experiments were conducted on the same hardware. Consistency was ensured through standardized implementation in RStudio, utilizing reliable libraries such as caret, randomForest, and Boruta. The data was normalized to allow fair comparison across features.
- Reliability: Reproducibility was strengthened by specifying model parameters (Table I), using stratified 5-fold cross-validation, and conducting grid search for hyperparameter tuning of RF. These methods ensure that model evaluations are consistent and statistically robust.

In summary, while this study adopts rigorous methodological practices, further validation across varied datasets and model configurations is necessary to confirm the broader applicability of the findings.

IV. CONCLUSION

This study investigates the impact of missing data, imputation strategies, resampling techniques, and feature selection techniques on the classification performance of Machine Learning (ML) models for water quality anomaly detection. The integration of these preprocessing techniques significantly improves model performance, particularly in addressing class imbalance. Three classifiers were evaluated: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Naïve Bayes (NB).

Among the models tested, RF consistently outperformed the others, achieving the highest F1-score of 0.9954 when combined with Expectation Maximization (EM) for imputation and Rapidly Converging Gibbs (RACOG) sampler for resampling, and Mean Decrease Accuracy (MDA) for feature selection. In contrast, NB demonstrated the weakest performance, with a maximum F1-score of 0.8935 under zero-filling imputation, RACOQ resampling, and MDA feature selection.

Although classification performance was generally lower on the original imbalanced dataset, the results confirm that advanced preprocessing, particularly the use of EM and Multiple Imputation by Chained Equations (MICE) imputation methods alongside RACOG-based resampling, substantially enhances model robustness. RF, in particular, proved to be a highly reliable choice for such tasks.

Future work will focus on further enhancing RF performance through advanced hyperparameter tuning, permutation-based feature importance refinement, and optimized resampling configurations. These improvements aim to deliver higher classification accuracy and better computational efficiency than existing machine learning approaches for imbalanced water quality datasets.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from the Indonesian Ministry of Education, Culture, Research, and Technology through the PMDSU Research Grant No. 038/E5/PG.02.00.PL/2024.

REFERENCES

[1] P. Jeffrey, Z. Yang, and S. J. Judd, "The status of potable water reuse implementation," *Water Research*, vol. 214, May 2022, Art. no. 118198, https://doi.org/10.1016/j.watres.2022.118198.

[2] N. Morin-Crini *et al.*, "Worldwide cases of water pollution by emerging contaminants: a review," *Environmental Chemistry Letters*, vol. 20, no. 4, pp. 2311–2338, Aug. 2022, https://doi.org/10.1007/s10311-022-01447-4.

[3] N. U. H. Shar, G. Q. Shar, A. R. Shar, S. M. Wassan, Z. Q. Bhatti, and A. Ali, "Health Risk Assessment of Arsenic in the Drinking Water of Upper Sindh, Pakistan," *Engineering, Technology & Applied Science Research*, vol. 11, no. 5, pp. 7558–7563, Oct. 2021, https://doi.org/10.48084/etasr.4336.

[4] R. P. Shete, A. M. Bongale, and D. Dharrao, "IoT-enabled effective real-time water quality monitoring method for aquaculture," *MethodsX*, vol. 13, Dec. 2024, Art. no. 102906, https://doi.org/10.1016/j.mex.2024.102906.

[5] R. K. Mishra, "Fresh Water availability and Its Global challenge," *British Journal of Multidisciplinary and Advanced Studies*, vol. 4, no. 3, pp. 1–78, May 2023, https://doi.org/10.37745/bjmas.2022.0208.

[6] H. Gunter, C. Bradley, D. M. Hannah, S. Manaseki-Holland, R. Stevens, and K. Khamis, "Advances in quantifying microbial contamination in potable water: Potential of fluorescence-based sensor technology," *WIREs Water*, vol. 10, no. 1, Jan. 2023, https://doi.org/10.1002/wat2.1622.

[7] W. Yang, X. Wei, and S. Choi, "A Dual-Channel, Interference-Free, Bacteria-Based Biosensor for Highly Sensitive Water Quality Monitoring," *IEEE Sensors Journal*, vol. 16, no. 24, pp. 8672–8677, Dec. 2016, https://doi.org/10.1109/jsen.2016.2570423.

[8] B. Mizaikoff, "Infrared optical sensors for water quality monitoring," *Water Science and Technology*, vol. 47, no. 2, pp. 35–42, Jan. 2003, https://doi.org/10.2166/wst.2003.0079.

[9] T. Maqbool *et al.*, "Exploring the relative changes in dissolved organic matter for assessing the water quality of full-scale drinking water treatment plants using a fluorescence ratio approach," *Water Research*, vol. 183, Sep. 2020, Art. no. 116125, https://doi.org/10.1016/j.watres.2020.116125.

[10] G. E. Adjovu, H. Stephen, D. James, and S. Ahmad, "Measurement of Total Dissolved Solids and Total Suspended Solids in Water Systems: A Review of the Issues, Conventional, and Remote Sensing Techniques," *Remote Sensing*, vol. 15, no. 14, Jul. 2023, Art. no. 3534, https://doi.org/10.3390/rs15143534.

[11] E. K. Nti *et al.*, "Water pollution control and revitalization using advanced technologies: Uncovering artificial intelligence options towards environmental health protection, sustainability and water security," *Heliyon*, vol. 9, no. 7, Jul. 2023, Art. no. e18170, https://doi.org/10.1016/j.heliyon.2023.e18170.

[12] K. Gunasekaran and S. Boopathi, "Artificial Intelligence in Water Treatments and Water Resource Assessments," in *Advances in Environmental Engineering and Green Technologies*, IGI Global, 2023, pp. 71–98.

[13] E. Parimbelli, T. M. Buonocore, G. Nicora, W. Michalowski, S. Wilk, and R. Bellazzi, "Why did AI get this one wrong? — Tree-based explanations of machine learning model predictions," *Artificial Intelligence in Medicine*, vol. 135, Jan. 2023, Art. no. 102471, https://doi.org/10.1016/j.artmed.2022.102471.

[14] E. M. Dogo, N. I. Nwulu, B. Twala, and C. O. Aigbavboa, "Empirical Comparison of Approaches for Mitigating Effects of Class Imbalances in Water Quality Anomaly Detection," *IEEE Access*, vol. 8, pp. 218015–218036, 2020, https://doi.org/10.1109/access.2020.3038658.

[15] N. H. A. Malek, W. F. Wan Yaacob, S. A. Md Nasir, and N. Shaadan, "Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques," *Water*, vol. 14, no. 7, Mar. 2022, Art. no. 1067, https://doi.org/10.3390/w14071067.

[16] S. Nuanmeesri, C. Tharasawatpipat, and L. Poomhiran, "Transfer Learning Artificial Neural Network-based Ensemble Voting of Water Quality Classification for Different Types of Farming," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15384–15392, Aug. 2024, https://doi.org/10.48084/etasr.7855.

[17] S. Nuanmeesri, L. Poomhiran, P. Kadmateekarun, and S. Chopvitayakun, "Improving the Water Quality Classification Model for Various Farms Using Features Based on Artificial Neural Network," *TEM Journal*, pp. 2144–2156, Nov. 2023, https://doi.org/10.18421/tem124-25.

[18] S. Nuanmeesri and W. Sriurai, "Multi-Layer Perceptron Neural Network Model Development for Chili Pepper Disease Diagnosis Using Filter and Wrapper Feature Selection Methods," *Engineering, Technology & Applied Science Research*, vol. 11, no. 5, pp. 7714–7719, Oct. 2021, https://doi.org/10.48084/etasr.4383.

[19] S. Nuanmeesri and W. Sriurai, "Thai Water Buffalo Disease Analysis with the Application of Feature Selection Technique and Multi-Layer Perceptron Neural Network," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 6907–6911, Apr. 2021, https://doi.org/10.48084/etasr.4049.

[20] S. Nuanmeesri, "Feature Selection for Analyzing Data Errors Toward Development of Household Big Data at the Sub-District Level Using Multi-Layer Perceptron Neural Network," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 16, no. 05, pp. 121–138, Mar. 2022, https://doi.org/10.3991/ijim.v16i05.22523.

[21] *Sistem Informasi Hidrologi & Kualitas Air*. (2022), Balai Besar Wilayah Sungai Bengawan Solo. [Online]. Available: https://hidrologi.bbws-bsolo.net/kualitasair.

[22] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, https://doi.org/10.1109/tkde.2008.239.

[23] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27–38, Jan. 2009, https://doi.org/10.1016/j.patrec.2008.08.010.

[24] F. M. Shrive, H. Stuart, H. Quan, and W. A. Ghali, "Dealing with missing data in a multi-question depression scale: a comparison of imputation methods," *BMC Medical Research Methodology*, vol. 6, no. 1, Dec. 2006, https://doi.org/10.1186/1471-2288-6-57.

[25] F. Mouret, A. Hippert-Ferrer, F. Pascal, and J.-Y. Tourneret, "A Robust and Flexible EM Algorithm for Mixtures of Elliptical Distributions with Missing Data," *IEEE Transactions on Signal Processing*, vol. 71, pp. 1669–1682, 2023, https://doi.org/10.1109/tsp.2023.3267994.

[26] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, https://doi.org/10.1007/s11749-016-0481-7.

[27] M. Sandri and P. Zuccolotto, "Variable Selection Using Random Forests," in *Studies in Classification, Data Analysis, and Knowledge Organization*, Berlin, Heidelberg, pp. 263–270, https://doi.org/10.1007/3-540-35978-8_30.

[28] H. Toutenburg, "Rubin, D.B.: Multiple imputation for nonresponse in surveys," *Statistical Papers*, vol. 31, no. 1, Dec. 1990, https://doi.org/10.1007/bf02924688.

[29] Janez Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[30] Y. Liu, Y. Zhou, S. Wen, and C. Tang, "A Strategy on Selecting Performance Metrics for Classifier Evaluation," *International Journal of Mobile Computing and Multimedia Communications*, vol. 6, no. 4, pp. 20–35, Oct. 2014, https://doi.org/10.4018/ijmcmc.2014100102.

[31] J. A Ilemobayo *et al.*, "Hyperparameter Tuning in Machine Learning: A Comprehensive Review," *Journal of Engineering Research and Reports*, vol. 26, no. 6, pp. 388–395, Jun. 2024, https://doi.org/10.9734/jerr/2024/v26i61188.

[32] N. Zhu, C. Zhu, L. Zhou, Y. Zhu, and X. Zhang, "Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm," *Applied Sciences*, vol. 12, no. 20, Oct. 2022, Art. no. 10456, https://doi.org/10.3390/app122010456.

[33] T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse, "An Empirical Study of Learning from Imbalanced Data Using Random Forest," in *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*, Patras, Greece, Oct. 2007, pp. 310–317, https://doi.org/10.1109/ictai.2007.46.