G-Tech : Jurnal Teknologi Terapan
Vol. 9, No. 3, July 2025, p. 1176-1185
E-ISSN: 2623-064X | P-ISSN: 2580-8737

Advanced Machine Learning Techniques for Assessing Water Quality: A Comparative Study Using Ensemble, Neural Networks, and Instance-Based Models

Muhammad Hafiz[1], Johan Iswara[2], Bari Fakhrudin[3], Widitra Nararya Rama[4], Avellino Vincent Juwono[5], Gilang Raka Rayuda Dewa[6]✉

[1,2,3,4,5,6] Department of Computer Science, Sampoerna University, Indonesia

✉ **Corresponding Author:** Gilang Raka Rayuda Dewa (e-mail: gilang.dewa@sampoernauniversity.ac.id)

Article Information

Article History

Received: May 27, 2025
Revised: June 1, 2025
Published: July 02, 2025

***Keywords*:**

ensemble;
instance;
machine learning;
neural networks;
water quality.

ABSTRACT

Access to safe water remains a significant issue, with around 5.8 billion people lacking access to potable water globally. Rapid and accurate identification of water safety is thereby essential to reduce public waterborne diseases. However, conventional laboratory-based testing is typically time-consuming and expensive. On the other hand, machine learning provides time- and cost-effective assessments based on physicochemical properties. Unfortunately, most studies only evaluate a single model type in a small dataset, resulting in limited insight that makes it hard to determine the actual effectiveness of these models. To address this limitation, the present study conducts a comparative analysis of three machine learning paradigms: ensemble-based, neural network-based, and instance-based models. Using a publicly available dataset of 7,999 samples, each model is evaluated using key performance metrics, including accuracy, precision, and confusion matrix analysis. The evaluation results show that the ensemble-based model achieves the highest accuracy of 96.62% and precision of 96.53%, outperforming the neural network-based model, which achieves an accuracy of 94.75% and precision of 70.47%. Additionally, the instance-based model achieves an accuracy of 91.12% and a precision of 83.04%. These results indicate the effectiveness of the ensemble-based model for real-time water quality monitoring.

INTRODUCTION

Ensuring access to clean water has become a considerable concern for researchers regarding human health and well-being. Over 5.8 billion people worldwide lack access to safely managed drinking water, exposing them to significant health risks from chemical contaminants, e.g., heavy metals and pesticides, and biological contaminants, e.g., bacteria and viruses (*State of the World's Drinking Water*, 2022). Waterborne diseases cause approximately 505,000 deaths each year, including over 395,000 children under five, primarily due to unsafe water, poor sanitation, and inadequate hygiene (World Health Organization, 2023). Moreover, significant chemical and biological substances can easily contaminate water, compromising its safety (Ghoochani et al., 2023; Mohialden et al., 2024). On the other hand, manually monitoring water quality using conventional tools is challenging due to the potential for human error and inefficient use of resources (Jha, 2020). Relying on laboratory testing methods is also time-consuming and delays the implementation of proper actions in anticipating public health risks (Nuanmeesri et al., 2024).

Accordingly, to address these challenges, researchers have intensively explored machine learning (ML) techniques (Zhu et al., 2022). By analyzing the historical data patterns of water quality, ML models can immediately predict the water status based on various physicochemical

This is an open access article under the CC BY license

parameters (Nuanmeesri et al., 2024; Zhu et al., 2022). Moreover, ML techniques can identify complex interactions among contaminants that are impossible to detect using conventional statistical methods. However, due to the diversity of machine learning paradigms, selecting the most suitable model for water quality classification is a challenging task. Each model offers various powerful concepts and limitations, depending on the data context, the complexity, and the objective functions. Therefore, a comparative analysis is critical here to evaluate overall performance and trade-offs in computational efficiency under real-world conditions. The corresponding analysis informs the model selection regarding context sensitivity and adaptability to the specific environmental monitoring needs. This study focuses on three machine learning paradigms: Random Forest (RF), a representative of ensemble-based models; Artificial Neural Network (ANN), a representative of neural-network-based models; and K-Nearest Neighbors (KNN), a representative of instance-based models.

To the best of the authors' knowledge, no existing study has conducted a direct and head-to-head comparison of the three major machine learning paradigms in water quality classification. Most previous research has focused on individual models or limited pairs, typically based on small datasets with low geographic diversity. An RF study relies on a small dataset, which increases the risk of overfitting and poor generalizability (Kukartsev et al., 2024) . Similarly, another RF-based model does not use the hyperparameter tuning technique, leaving performance optimization largely unexplored (Xiong et al., 2023). In another study, an Artificial Neural Network (ANN) model notes that computational costs increase significantly as the model complexity, e.g., the number of hidden layers or neurons, grows, which may hinder real-time implementation. Also, preprocessing is not conducted, leaving the possibility of data imbalance. On the instance-based side, a K-Nearest Neighbors (KNN) method uses input features that are not correctly scaled, resulting in significant performance degradation. Additionally, this study suggests that the KNN model is highly susceptible to the curse of dimensionality in high-dimensional datasets (Mawarti et al., 2024). Similarly, another study employs KNN models for multi-class water quality classification but does not provide specific accuracy metrics. They instead highlight the model's sensitivity to class imbalance, which can substantially distort classification outcomes (Isık & Akkan, 2025). Additionally, a study directly comparing RF and KNN does not employ standardized data preprocessing across models, which makes it difficult to isolate the source of performance differences. (Sutanto et al., 2024).

In general, most existing studies rely on datasets with limited temporal or geographic diversity, often lacking robust preprocessing, class balancing, or uniform evaluation metrics. Furthermore, there is a scarcity of research that evaluates these three distinct paradigms side by side under the same experimental conditions. Without such comparative analyses, stakeholders and policymakers may struggle to choose the most context-appropriate machine-learning model for water quality monitoring in different environmental settings.

Addressing these research gaps, the current study presents a systematic and comprehensive comparative evaluation of three representative machine learning paradigms: Random Forest (RF), an ensemble-based model; Artificial Neural Network (ANN) as a neural-based model; and K-Nearest Neighbors (KNN) as an instance-based model. Unlike previous work, this study employs a large-scale, well-preprocessed, and balanced dataset that reflects broad geographic diversity. All three models are evaluated using consistent performance metrics, such as accuracy, precision, and a confusion matrix, via identical training-test splits to ensure methodological fairness Moreover, the study employs standardized preprocessing pipelines, including scaling, normalization, and class balancing, that are crucial for reducing model bias and enhancing real-world applicability Accordingly, the study aims to benchmark each model's predictive performance andereby offering practical insights for deploying machine learning in water quality monitoring systems.

RESEARCH METHOD

This section presents the overall procedure for developing and evaluating the water safety classification models. This section covers the dataset description, the preprocessing step, and the detailed configuration of three different machine learning paradigms.

Dataset Description and Preprocessing

This study utilizes the dataset (Smarty, 2021) comprising 7,999 water samples, each characterized by 20 numeric attributes related to water quality. Based on regulatory standards, the target variable uses a binary choice, whether the water sample is safe (1) or unsafe (0) for human consumption. The distribution of the target classes is notably imbalanced, with 912 samples (11.41%) labeled as safe and 7,084 samples (88.59%) labeled as unsafe. This imbalance highlights the crucial need for models that accurately identify the minority class without bias. Each feature corresponds to a specific contaminant or water characteristic, along with an established threshold above which the contaminant is considered hazardous.

The dataset is processed via a series of preprocessing steps to ensure data quality and consistency (Đorđević et al., 2022). The first step involved inspecting the dataset for invalid entries. From inspection, it is known that 'ammonia' and 'is safe' columns contain several non-numeric values represented as '#NUM!'. These entries are replaced with missing values (NaN). The number of affected rows is minimal compared to the overall dataset size; therefore, the rows are removed entirely. Removing a small proportion of incomplete data helps maintain the dataset's integrity without introducing significant bias (Alam et al., 2023).

Subsequently, feature scaling is applied to the dataset. All numeric attributes, except the target variable 'is_safe', are standardized using Standard Scaling, which transforms the data into a mean of zero and a standard deviation of one. Standardizing the features ensures that differences in scale among the attributes do not unfairly influence the model's learning process (Raju et al., 2020).

Finally, the cleaned and scaled features are recombined with the target variable to form the final dataset, which is then saved for use in the modeling stage. These preprocessing steps established a strong and consistent foundation for training the classification models, thereby maximizing their potential effectiveness in predicting water safety outcomes.

Model Implementation

Over the past decade, numerous researchers have investigated the use of machine learning, specifically the RF algorithm, the KNN technique, and the ANN algorithm, in assessing water quality. The RF algorithm constructs multiple decision trees during training and establishes predictions by taking a majority vote among the trees' outputs (Salman et al., 2024). A random subset of the training data is utilized to establish each tree via a bootstrapping technique; thus, the tree only considers a random subset of features at each split. The dataset is split into 70% training and 30% testing subsets using stratified sampling to maintain class distribution.

The Bayesian optimization (Turner et al., 2021) is leveraged to efficiently tune the seven critical Random Forest parameters, including tree count (50-500), depth (3-20), minimum split of 2-10, minimum leaf sample (1-5), feature selection method of sqrt or log2, bootstrap sampling of True or False, and class weighting of balanced or None. The model is applied to optimize the runs with 50 iterations, with 5-fold cross-validation. It automatically finds configurations that maximize accuracy and obtains an optimal model. The Bayesian optimization results in 500 decision trees, a maximum tree depth of 14 levels, and logarithmic feature selection. In addition, splitting nodes and leaves is adjusted with a minimally required sample size of 2 and 1, respectively. Figure 1 summarizes the implementation of the RF algorithm.

A study applying the RF algorithm to 137 groundwater samples achieved 94% accuracy and an F1-score of 0.967 (Xiong et al., 2023). Another study introduces the RF algorithm to 200 river samples, utilizing the Synthetic Minority Over-sampling Technique to balance the data (Kukartsev et al., 2024). This model only exploits three features: pH, conductivity, and nitrate.

The evaluation demonstrates that the model achieves 100% accuracy, highlighting the significance of the RF feature in identifying primary contaminants.

Input:
Dataset $D = \{(x_i, y_i)\}_{i-1}^{n}$, where $x_i \in \mathbb{R}^d, y_i \in \{0,1\}$
Output:
Trained a Random Forest model f'', optimal parameters θ', and evaluation metrics.
Steps:
1. Data Loading
   Load D From 'water quality processed.csv'.
   Extract: $X \in \mathbb{R}^{n \times d}$ (features)
   $y \in \{0,1\}^n$ (labels)
2. Train-Test Split
   Split D into 70% training and 30% test set using stratified sampling on y.
3. Bayesian Optimization Setup
   Define model $f(x;\theta)$ Random Forest.
   Search space H: $n_{\text{colimators}} \in [50,500]$, max_depth ∈ [3,20]
   min_samples_split ∈ [2,10], max_features ∈ $\{\sqrt{d}, \log_2 d\}$
   bootstrap ∈ { True, False }, class_weight ∈ {balanced, None }
4. Model Training
   Use 5-fold CV and accuracy to maximize $\theta^* = \arg\max_{\theta \in H}$ Accuracy $_{\text{CV}}(f(x;\theta))$
5. Model Evaluation
   Apply $f'' = f(x;\theta')$ to test the set
   Compute Accuracy: $\frac{1}{m}\sum_{i-1}^{m} 1(f^*(x_i) = y_i)$, Weighted Precision,
   and Normalized Confusion Matrix
6. Output Results
   Return $\boldsymbol{\theta}^*$, accuracy, precision, and confusion matrix

Figure 1. RF Algorithm Implementation.

The K-Nearest Neighbors (KNN**)** technique assigns a dataset based on the majority label among the closest neighbors. This technique relies on simple computations that require minimal training. However, as an instance-based model, the KNN technique is typically sensitive to feature scaling and data imbalance (Juna et al., 2022). The KNN classifier uses the Euclidean distance, also known as the L2 norm, to measure similarity between samples with uniform neighbor weighting. The Euclidean Distance is computed as

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \quad (1)$$

where x indicates a point or vector in n-dimensional space and i indicates the index of the vector component. Feature scaling is applied to ensure equal contribution of all chemical concentrations to the distance metric. Hyperparameter optimization focused on the number of neighbors, e.g., k = 3, 5, 7, with Bayesian search to identify the optimal configuration. Data preparation involved handling missing values through KNN imputation, standardizing data types, and applying random oversampling to address class imbalance. The dataset was divided into 70% training and 30% testing subsets and scaled using StandardScaler, which is essential for distance-based algorithms. Hyperparameter tuning was performed using GridSearchCV (5-fold cross-validation) to optimize the values of k (3, 5, 7) and the Euclidean distance metric. Figure 2 summarizes the KNN algorithm.

A study applies the KNN technique to a large dataset of 7,999 samples, achieving 87.83% accuracy with k = 5 (Mawarti et al., 2024). In addition, the comparative analysis of the RF algorithm and the KNN technique is proposed to provide a holistic understanding of machine learning-based performance (Sutanto et al., 2024). In this study, the RF algorithm achieves an accuracy of 95.75%, outperforming the KNN technique, which achieves an accuracy of 90.19% on the same balanced dataset. While several studies do not evaluate performance under outliers

and environmental noise, the overall performance of the RF algorithm is robust in hierarchical classification tasks.

Input:
Dataset $D = \{(x_i, y_i)\}_{i-1}^{n}$, where $x_i \in \mathbb{R}^d, y_i \in \{0,1\}$
Output:
Trained a KNN model f'', optimal parameters k, and evaluation metrics
Steps:

1. Data Preparation
   Load: D → ReadCSV()
   Clean: Handle missing, convert types, and remove duplicates
   Separate: $X \in \mathbb{R}^{n \times d}, y \in \{0,1\}^n$
   Balance: Apply oversampling → $D_{\text{balanced}} = \{(x_i, y_i)\}_{i=1}^{n^i}$
2. Split & Scale
   Stratified split: $D_{\text{balanced}} \rightarrow D_{\text{train}} \cup D_{\text{test}}$, 70/30
   Scale:
   $\tilde{X}_{\text{train}} = \text{StandardScaler}(X_{\text{train}})$
   $\tilde{X}_{\text{test}} = \text{StandardScaler}(X_{\text{test}})$
3. Train with Grid Search
   Define $K = \{3,5,7\}$, metric = Euclidean
   GridSearchCV (5-fold):
   $f^* = \arg\max_{k \in K} \text{Accuracy}_{\text{CV}}(\text{KNN}_k)$
4. Evaluation
   Predict: $\hat{y} = f^*(\tilde{X}_{\text{test}})$
   Compute metrics:
   Accuracy, Precision, Recall, F1
   Confusion matrix:
   $CM[i,j] = \text{count}(y = i \wedge \hat{y} = j)$

Figure 2. KNN Algorithm Implementation

On the other hand, the ANN utilizes several features scaled via Z-score normalization to enable the model to learn from the data pattern effectively. Each feature x is transformed according to:

$$z = = \frac{x - \mu}{\sigma}, \tag{2}$$

where μ represents the mean and σ the standard deviation of the feature values calculated from the training set. Standardization is essential because neural networks are sensitive to the scale of input variables, and scaling facilitates faster and more stable convergence during training.

The ANN architecture is structured sequentially, consisting of an input layer connected to a hidden layer comprising 16 neurons. Each neuron in the hidden layer utilizes the Rectified Linear Unit (ReLU) activation function. This activation function introduces non-linearity to prevent the vanishing gradient problem that can occur during training. The sigmoid function maps the network's outputs to a probability range of 0 to 1, making it suitable for binary classification tasks, such as determining water safety. Training uses the Stochastic Gradient Descent (SGD) optimization algorithm, configured with a learning rate of 0.01 and a momentum factor of 0.2. The weight update rule with momentum is expressed as

$$v = \gamma v + \eta \nabla L(w), \tag{3}$$

where v represents the momentum term, γ is the momentum coefficient, η is the learning rate, and ∇L(w) is the gradient of the loss function concerning the weights. Incorporating momentum in SGD helps accelerate training and smooth the optimization trajectory, particularly in the presence of noisy gradients. The model is trained over 1,000 epochs with a batch size of 32 using the scaled training data. A final model evaluation is conducted on the testing set to assess the ANN's generalization performance. The summary of ANN implementation is presented in Figure 3.

Input:
Dataset $D = \{(x_i, y_i)\}_{i-1}^{n}$, where $x_i \in \mathbb{R}^d, y_i \in \{0,1\}$
Output:
Trained an ANN model f'', optimal parameters k, and evaluation metrics
Steps:

1. Data Preparation:
   Load X, y,
   Stratified split: 70% train, 30% test,
   Apply StandardScaler to $X_{train,} X_{test}$
2. Model Definition
   ANN with input d, one hidden layer (16, RELU), output (1, Sigmoid)
   Loss: RMSE
   Optimizer: SGD (lr = 0.01, momentum 0.2)
3. Training
   Train f for 1000 epoch, bath size = 32
4. Evaluation
   Predict on test set: $f(X_{test})$
   Compute RMSE, Accuracy

Figure 3. ANN Algorithm Implementation

A ANN study investigates the water quality of eight river stations using an ANN algorithm (Isık & Akkan, 2025). The findings show that the ANN algorithm can effectively model the water quality index in well-preprocessed and time-series data. Although the study does not report classification accuracy, it suggests that the performance of the ANN algorithm is highly dependent on the data input, model architecture, and training strategy.

RESULTS AND DISCUSSION

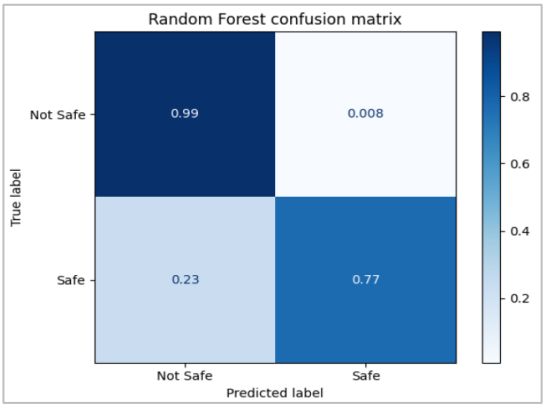

Figure 4. Confusion Matrix of RFAlgorithm

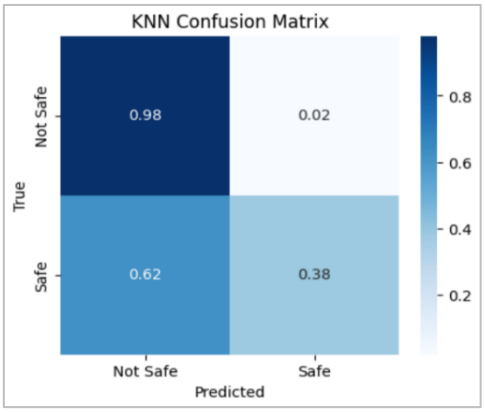

Figure 5. Confusion Matrix of KNN Algorithm

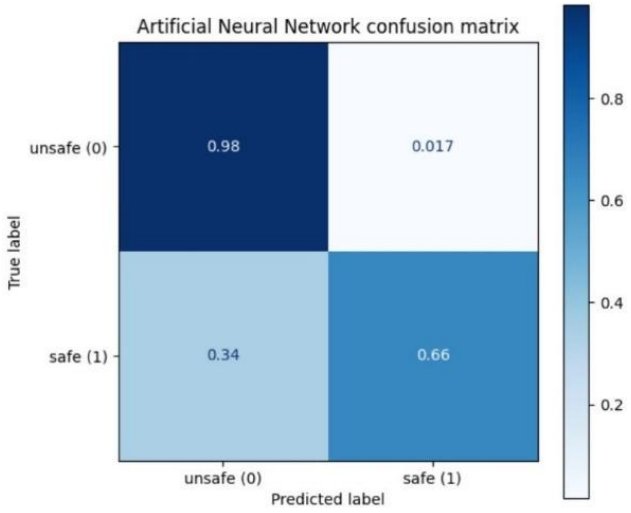

Figure 6. Confusion Matrix of ANN Algorithm

This section analyzes the results of the machine learning models applied to the water quality classification task. The performance of ensemble-based, neural-network, and instance-based models is validated based on several evaluation metrics, including accuracy, precision, and confusion matrix analysis. The confusion matrices for three machine learning paradigms — RF, KNN, and ANN —are shown in Figures 4, 5, and 6, respectively. The RF algorithm predicts water safety with an overall test accuracy of 96.62% and a weighted precision score of 96.53%. The negative rate for correctly identifying unsafe water samples was 99.2%, indicating particularly robust performance for this task. Although the model achieves an actual positive rate of 76.6%, the higher false negative rate of 23.4% indicates that some unsafe samples are misclassified as safe.

The KNN technique demonstrates high reliability in identifying unsafe water samples, correctly classifying 98% of them; in contrast, its performance is considerably lower when predicting safe samples, with only 38% correctly identified. This asymmetry reflects the model's bias toward the dominant class, a common outcome in imbalanced classification tasks (Nurhidayat & Dewi, 2023). As a result, while the model achieves a high overall accuracy of 91.12%, the recall performance for the safe class remains low. The precision of 70.47% indicates that the model can only be trusted to predict safe samples, not applicable for unsafe samples.

The ANN model is evaluated after 1000 training epochs. The performance is measured using several key metrics, including accuracy, Root Mean Squared Error (RMSE), precision, and confusion matrix analysis. The model achieves a testing accuracy of 94.75%, a testing RMSE of 0.2133, and a precision score of 0.8304. These results demonstrate that the ANN successfully learned to classify water samples based on their safety properties. However, further analysis reveals specific strengths and weaknesses in the model's performance. Table 2 summarizes the classification performance metrics for the RF, ANN, and KNN models, including confusion matrix proportions, accuracy, and precision values.

Table 2. Summary of Model Performance Metrics

Evaluation Metrics	*Random Forest*	*ANN*	*KNN*
Confusion Matrix (normalized)	True Unsafe: 0.99 False Safe: 0.008 False Unsafe: 0.23 True Safe: 0.77	True Unsafe: 0.98 False Safe: 0.018 False Unsafe: 0.32 True Safe: 0.68	True Unsafe: 0.98 False Safe: 0.02 False Unsafe: 0.62 True Safe: 0.38
Accuracy(%)	96.62	94.75	91.12
Precision (%)	96.53	83.04	70.47

The results presented in Table 2 highlight the strengths and limitations of each classification model across key performance metrics. The Random Forest classifier achieved the highest accuracy of 96.62% and precision of 96.53%, rendering it the most reliable model for classifying water. Its ability to correctly identify 99% of unsafe samples while maintaining a relatively strong 77% actual positive rate for safe samples demonstrates its balanced performance, even under conditions of class imbalance. In contrast, while computationally simpler, the KNN technique shows a drop in both accuracy, at 91.12%, and precision, at 70.47%. The confusion matrix indicates that the KKN only yields high performance in true negatives; however, it cannot accurately identify safe water, with only 38% of safe samples being correctly classified. This performance gap is likely a consequence of the model's sensitivity to class imbalance and reliance on distance metrics, which can be skewed when one class dominates.

The ANN algorithm achieved a middle-ground performance with 94.75% accuracy and a precision of 83.04%. Its confusion matrix indicates strong detection of unsafe water, with an accuracy of 98%, and moderate success with safe samples, with an accuracy of 68%. This indicates that the overall performance of the ANN algorithm is higher than the KNN technique; however, it still tends to favor the majority class. RMSE as a loss function provided further insight into model learning dynamics, revealing minor overfitting tendencies after extended training. In the end, Random Forest outperforms both ANN and KNN in most metrics and shows the most consistent behavior across both classes. Its ensemble nature and robustness to noise make it particularly suited

to this classification task. In addition to its high accuracy and precision, the high reliability performance of RF on the imbalanced dataset can be attributed to its robust ensemble nature in handling class distribution disparities. Unlike typical conventional methods, RF builds multiple decision trees on random subsets of the data and aggregates their predictions, introducing diversity and reducing the likelihood of bias toward the dominant class.

Furthermore, the RF algorithm's ability to sample features and data points during training enables it to be more resilient to noise and overlapping class regions, which are common challenges in imbalanced classification tasks. This characteristic has been consistently validated in prior studies, where RF outperformed other algorithms like SVM, Naïve Bayes, and KNN under similar imbalanced conditions. Therefore, the superior performance of the RF algorithm in this study is empirical and aligns with its theoretical strengths in managing data imbalance. While ANN demonstrated potential with deeper feature representation, and KNN offered interpretability, the results show that in the case of class imbalance and classification risk, Random Forest presents the most practical and accurate solution.

CONCLUSION

This study explored the application of three machine learning paradigms: Random Forest (RF) as an example of an ensemble model, Artificial Neural Network (ANN) as an example of a neural-network-based model, and K-Nearest Neighbors (KNN) as an instance-based model, to classify water quality based on chemical and biological indicators. A real-world dataset sample is preprocessed through a systematic pipeline, including data cleaning, feature scaling, and model-specific tuning. Each model is evaluated using multiple metrics, including accuracy, precision, and confusion matrix analysis, to provide a comprehensive view of its predictive performance.

The evaluation results indicate that while all three models show promise in classifying water safety, they vary in their ability to handle class imbalance and accurately identify both safe and unsafe samples. The Random Forest model outperforms other algorithms, achieving the highest overall accuracy of 96.62% and precision of 96.53%, with a strong balance between class predictions. The Artificial Neural Network also demonstrates solid performance in learning generalized patterns, achieving an accuracy of 94.75%. However, it struggles to classify the minority class, as reflected in its lower precision of 70.47%. Meanwhile, the K-Nearest Neighbors model, although practical and straightforward, falls short in accurately detecting safe samples when the dataset is imbalanced, yielding an accuracy of 91.12% and a precision of 83.04%. Overall, these findings show the effectiveness of the ensemble-based approach for reliable real-time water quality monitoring. The comparative analysis indicates that both performance metrics and key factors, including interpretability, computational cost, and sensitivity to data imbalance, influence model selection for water safety prediction.

To enhance classification performance, especially for minority classes, future research could explore the integration of hybrid model architectures, such as combining decision trees with neural networks or ensemble techniques like stacking and boosting, to leverage the strengths of different algorithms. Additionally, implementing cost-sensitive learning or resampling techniques (e.g., SMOTE or ADASYN) can help address data imbalance more effectively. Researchers may also consider incorporating domain-specific knowledge into the feature engineering process or leveraging spatiotemporal data to capture more nuanced patterns in water safety. Finally, evaluating model robustness under varying environmental conditions and deploying explainable AI (XAI) tools can contribute to more trustworthy and actionable predictions.

ACKNOWLEDGEMENT

The authors would like to thank their colleagues at Sampoerna University for moral support, encouragement, and valuable insights throughout this research.

REFERENCES

Alam, S., Ayub, M. S., Arora, S., & Khan, M. A. (2023). An Investigation of The Imputation Techniques for Missing Values in Ordinal Data Enhancing Clustering and Classification

Analysis Validity. *Decision Analytics Journal*, *9*, 100341. https://doi.org/10.1016/j.dajour.2023.100341

Đorđević, A., Milovanović, M., Milojković, M., & Spasić, M. (2022). Data Analysis of Environmental Conditions Influencing The Work of Laboratory Equipment and Application of Machine Learning Models for Classification of Poor Conditions. *Facta Universitatis, Series: Automatic Control and Robotics*, *1*(3), 159. https://doi.org/10.22190/FUACR221118013D

Ghoochani, S., Khorram, M., & Nazemi, N. (2023). *Uncovering Top-Tier Machine Learning Classifier for Drinking Water Quality Detection.* Preprints.org. https://doi.org/10.20944/preprints202308.1636.v1

Isık, H., & Akkan, T. (2025). Water Quality Assessment with Artificial Neural Network Models: Performance Comparison Between SMN, MLP and PS-ANN Methodologies. *Arabian Journal for Science and Engineering*, *50*(1), 369–387. https://doi.org/10.1007/s13369-024-09238-5

Jha, B. K. (2020). Cloud-Based Smart Water Quality Monitoring System using IoT Sensors and Machine Learning. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3), 3403–3409. https://doi.org/10.30534/ijatcse/2020/141932020

Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2022). Water Quality Prediction Using KNN Imputer and Multilayer Perceptron. *Water*, *14*(17), 2592. https://doi.org/10.3390/w14172592

Kukartsev, V., Orlov, V., Semenova, E., & Rozhkova, A. (2024). Optimizing Water Quality Classification Using Random Forest and Machine Learning. *BIO Web of Conferences*, *130*, 03007. https://doi.org/10.1051/bioconf/202413003007

Mawarti, G. P. A. H., Rachmawanto, E. H., & Yaacob, N. B. M. (2024). Water Quality Classification Using K-Nearest Neighbor Algorithm. *International Journal of Engineering Computing Advanced Research*, *1*(1), 34–42.

Mohialden, Y. M., Hussien, N. M., & Salman, S. A. (2024). Automated Water Quality Assessment Using Big Data Analytics. *Mesopotamian Journal of Big Data*, *2024*, 211–222. https://doi.org/10.58496/MJBD/2024/015

Nuanmeesri, S., Tharasawatpipat, C., & Poomhiran, L. (2024). Transfer Learning Artificial Neural Network-based Ensemble Voting of Water Quality Classification for Different Types of Farming. *Engineering, Technology & Applied Science Research*, *14*(4), 15384–15392. https://doi.org/10.48084/etasr.7855

Nurhidayat, R., & Dewi, K. E. (2023). Penerapan Algoritma K-Nearest Neighbor dan Fitur Ekstraksi N-Gram Dalam Analisis Sentimen Berbasis Aspek. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, *12*(1), 91–100. https://doi.org/10.34010/komputa.v12i1.9458

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study The Influence of Normalization/Transformation Process on the Accuracy of Supervised Classification. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 729–735. https://doi.org/10.1109/ICSSIT48917.2020.9214160

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, *2024*, 69–79. https://doi.org/10.58496/BJML/2024/007

Smarty, P. (2021). *Water quality* (Version 9.12) [Dataset]. https://www.kaggle.com/datasets/mssmartypants/water-quality

State of the World's Drinking Water: An Urgent Call to Action to Accelerate Progress on Ensuring Safe Drinking Water for All (1st ed). (2022). World Health Organization.

Sutanto, T., Aditya, M. R., Budiman, H., Noor Ridha, M. R., Syapotro, U., & Azijah, N. (2024). Comparison of Logistic Regression, Random Forest, SVM, KNN Algorithmfor Water Quality Classification Based on Contaminant Parameters. *INTI Journal*, *2022*(1). https://doi.org/10.61453/jods.v2023no48

Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z., & Guyon, I. (2021). *Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2104.10201

World Health Organization. (2023). Drinking-water. *Drinking-Water*. https://www.who.int/news-room/fact-sheets/detail/drinking-water

Xiong, Y., Zhang, T., Sun, X., Yuan, W., Gao, M., Wu, J., & Han, Z. (2023). Groundwater Quality Assessment Based on the Random Forest Water Quality Index—Taking Karamay City as an Example. *Sustainability*, *15*(19), 14477. https://doi.org/10.3390/su151914477

Zhu, M., Wang, J., Yang, X., Zhang, Y., Zhang, L., Ren, H., Wu, B., & Ye, L. (2022). A review of the application of machine learning in water quality evaluation. *Eco-Environment & Health*, *1*(2), 107–116. https://doi.org/10.1016/j.eehl.2022.06.001