Exploratory Data Analysis and Water Potability Classification using Supervised Machine Learning Algorithms

Priya Kamath B.
Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal-576104, Udupi, Karnataka, India
priya.kamath@manipal.edu

Geetanjali Sharma
Department of Computer Science and Engineering, Pimpri Chinchwad College of Engineering, Pune, India
geetu2k3@gmail.com

Anupkumar Bongale
Department of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India
anupkumar.bongale@sitpune.edu.in (corresponding author)

Deepak Dharrao
Department of Computer Science and Engineering, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India
deepak.dharrao@sitpune.edu.in

Modisane Seitshiro
Centre for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa | National Institute for Theoretical and Computational Sciences (NITheCS), South Africa
modisane.seitshiro@nwu.ac.za (corresponding author)

Received: 11 September 2024 | Revised: 10 December 2024, 23 December 2024, and 07 January 2025 | Accepted: 11 January 2025

 DOI: https://doi.org/10.48084/etasr.8904

ABSTRACT

This study investigates the critical task of assessing water potability using supervised machine-learning techniques. The problem statement involves accurately predicting water potability based on chemical and physical parameters, which are crucial for public health and environmental sustainability. Exploratory Data Analysis (EDA) highlighted significant insights into feature distributions and correlations, guiding preprocessing steps and model selection. The Synthetic Minority Oversampling Technique (SMOTE) was applied to mitigate class imbalance, ensuring robust model training. Three classification algorithms, namely Logistic Regression (LR), K-Nearest Neighbors (KNN), and Random Forest (RF), were evaluated, with RF exhibiting superior performance after Optuna hyperparameter tuning, achieving an accuracy of 68%. Based on the performance of RF and KNN, a weighted voting-based ensemble technique achieved an accuracy of 71%. This study emphasizes the importance of leveraging machine learning to support water quality assessment, offering reliable tools for decision-making in public health and environmental management.

Keywords-SMOTE; machine learning; water quality; water potability; random forest; k nearest neighbors; logistic regression

I. INTRODUCTION

Drinking water is scarce in this world. It is a fundamental human right, critical for sustaining life and promoting good health. Although there are substantial improvements in water treatment technologies, with regulatory frameworks around the world to ensure water potability, it remains a major challenge that persists worldwide [1]. Industrial activities, agricultural practices, urbanization, and natural environmental changes are some of the main causes of water contamination [2]. The contamination of water bodies can change some water quality parameters, leading to non-potable water. Water can be contaminated by various factors, with arsenic being one of the contaminants [3]. The assessment of water quality is even essential for aquaculture systems. Researchers are developing Internet-of-Things-based systems to measure water quality. Water contamination can lead to low dissolved oxygen levels, which threatens aquatic life [4, 5]. According to the World Health Organization (WHO), approximately 2 billion people lack access to safe and potable water. Therefore, the conservation and maintenance of drinking water is an essential sustainable goal worldwide [6].

The microbial contamination of water bodies is examined with traditional laboratory-based and molecular detection techniques [7], biosensors and electromechanical approaches [8], optical detection techniques [9], and fluorescence [10]. These techniques play an important role in ensuring good water quality. However, traditional water quality monitoring generally relies on periodic sampling and manual testing, which can be labor-intensive and time-consuming [11]. The field of Artificial Intelligence (AI) and Machine Learning (ML) has made significant progress in the recent era. AI has become part of everyday human activity. ML models are especially suitable for data collection and sensor-based environments to process data effectively [12]. Despite these advances, leveraging the full benefits of AI and ML models for tasks such as water potability assessment is not an easy task. As data continually evolve, the performance of ML models can decrease. If the collected data are incorrect or inaccurate, it can lead to adverse behavior of the ML model, leading to unreliable results [13]. There is a definite need for explainable AI-based solutions that can assist managers and policymakers trust the recommendations obtained from ML-based analyses [14]. To handle these challenges, this study proposes three supervised ML classification models for efficiently classifying if the water is potable or nonpotable for the given set of water quality parameters.

Water quality prediction is a challenging and socially-must-required issue. There are several types of water, such as drinking water, wastewater, groundwater, surface water, seawater, and freshwater [15]. Each of these categories has unique characteristics and chemical compositions that make them relatively distinct. The importance of accurate water classification of water potability cannot be overstated, especially in regions where access to clean water is limited [16]. Ensuring that water is safe for consumption is crucial for public health, and automated classification systems based on ML offer a powerful tool to assist in this process [17-19]. ML algorithms, particularly ensemble methods, are well-suited to handle the multifaceted nature of water quality data, which can include a variety of physical, chemical, and biological parameters [20, 21]. This study proposes a novel ensemble classification model for water potability classification, combining the benefits of the Random Forest (RF) and K-Nearest Neighbors (KNN) classifiers using a weighted voting mechanism to integrate their predictions and improve prediction accuracy and robustness. The proposed approach not only advances the field of water quality monitoring but also demonstrates the broader potential of ML and ensemble techniques in solving critical real-world problems with significant social impact.

Water potability classification is an important research goal. Although the current research showcases the relevance of ML models for water potability classification, it lacks interpretability and suffers from model overfitting. This paper presents efficient water potability classification models with the following contributions:

- Preprocesses the dataset with SMOTE analysis to address the data imbalance.
- Implements three supervised machine learning classification models to efficiently classify water as potable or nonpotable for the given set of water quality parameters.
- By evaluating the performance of the supervised machine learning models on the water quality dataset, the RF classifier is suggested as the most suitable model for the classification task.
- The RF classifier is systematically hyperparameter-optimized with the Optuna framework to further enhance the classification results.
- Proposes a weighted ensemble classifier that combines the strengths of the RF and KNN classifiers.

II. WATER POTABILITY DATASET

The problem considered is to classify the suitability of water as potable or not using ML models. A dataset was obtained from [22], which consists of several water quality parameters, such as pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, and Turbidity. These water quality characteristics are used to classify whether the water is suitable for drinking. Potability is a categorical target variable. If water is potable, then Potability has a value of 1 and, in the case of non-potable, the value is 0. In water quality features, pH measures the water level acidity and alkalinity, ranging from 0 to 14. Hardness Water represents the amount of soap to be precipitated by the water, with recorded values ranging from 47.43 to 323.12 mg/L. Total Dissolved Solids (TDS) are represented by the Solids feature, with a minimum of 320.94 to a maximum of 61227.19 mg/L. Similarly, chloramine and sulfate compounds are represented using the Chloramines and Sulfate features, respectively. Conductivity measures water's capacity to conduct electricity because of dissolved salts and minerals. Organic carbon, with values between 2.2 and 28.3 mg/L, indicates the concentration of carbon-containing compounds derived from natural and synthetic sources. The Trihalomethanes (THMs) feature varies

from 0.73 to 124 µg/L, whereas Turbidity ranges from 1.45 to 6.73 NTU (Nephelometric Turbidity Units). The dataset contains 3,276 rows. Figure 1 shows a snapshot of sample values of the dataset.

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
7.4	204.9	20791.3	7.3	368.5	564.3	10.4	87.0	3.0
3.7	129.4	18630.1	6.6	341.8	592.9	15.2	56.3	4.5
8.1	224.2	19909.5	9.3	329.2	418.6	16.9	66.4	3.1

Fig. 1. Water potability dataset sample.

III. EXPLORATORY DATA ANALYSIS

Figure 2 illustrates the distribution of water samples classified by potability status. Within the dataset, there are 1998 entries classified as 0, representing non-potable water samples, accounting for approximately 61.0% of the dataset. In contrast, there are 1278 entries classified as 1, indicating potable water samples, comprising approximately 39.0% of the dataset. This distribution highlights a significant proportion of non-potable water samples compared to potable ones, underscoring the importance of accurately classifying water quality to ensure safe drinking water standards are met.

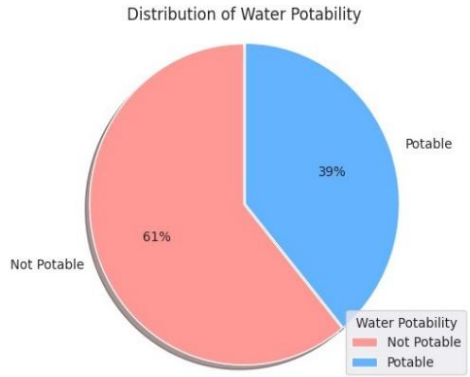

Fig. 2. Distribution of water potability.

The correlation analysis reveals the relationships between Potability and various water quality parameters. Among these attributes, pH and Hardness show negligible negative correlations of -0.0036 and -0.014, respectively, indicating minimal influence on water potability. Solids and Chloramines exhibit positive correlations of 0.034 and 0.024, respectively, suggesting a slight association with Potability. Conversely, Sulfate and Conductivity display negative correlations of -0.024 and -0.0081, respectively, implying a minor impact on Potability. Organic_carbon, Trihalomethanes, and Turbidity show negligible correlations near zero. These findings highlight complex relationships between water quality parameters and Potability, underscoring the nuanced factors that influence the classification of drinking water.

Several steps were taken to address the issue of missing values and class imbalance in the dataset. Initially, it was identified that three attributes, pH (491 missing values), Sulfate (781 missing values), and Trihalomethanes (162 missing values), lacked complete data points. These missing values were managed to ensure data integrity and reliability. Subsequently, the Synthetic Minority Oversampling Technique (SMOTE) was employed to mitigate class imbalance in the dataset using the *imblearn.over_sampling.SMOTE* module. This works by generating synthetic samples for the minority class (Potability = 1) to match the number of samples in the majority class (Potability = 0). In addition, to prepare the data for modeling after handling missing values and applying SMOTE, *MinMaxScaler* was utilized to normalize the features, ensuring that all variables were on a consistent scale for optimal performance of ML algorithms. This scaling step normalized each attribute to a range between 0 and 1, thus reducing potential biases due to different scales among the variables.

IV. MACHINE LEARNING CLASSIFICATION MODELS

A. Logistic Regression (LR) Classifier

LR was the first algorithm used to implement the classification, as it is one of the preferred binary classification models. It models the probability of a binary outcome, i.e., potable or nonpotable, based on input water quality features. LR estimates the probability that a water sample belongs to a specific class. A sigmoid function is used to map the predictions between 0 and 1.

B. K-Nearest Neighbors (KNN) Classifier

The KNN classifier is a simple yet effective algorithm for binary classification tasks. It classifies a sample by a majority vote of its k nearest neighbors in the feature space. In this context, the algorithm calculates the distances between the sample to be classified and all other samples in the training set. The sample is then assigned to the class that is most common among its k nearest neighbors. The KNN classifier is intuitive and adaptable to various datasets, making it suitable for identifying water samples as potable or non-potable based on their measured attributes.

C. Random Forest (RF) Classifier

The RF classifier is a powerful ensemble learning method for binary classification tasks. It constructs multiple decision trees during training and outputs the mode of the classes (potable or non-potable) as the prediction. Each tree in the forest is built using a random subset of features and a random subset of training data, reducing overfitting and improving accuracy. RF can handle large datasets with high dimensionality and is robust against noise and outliers. It also provides insight into feature importance, aiding in understanding the factors that influence the classification of water quality.

D. Fine-tuning Random Forest Classifier

The Optuna library effectively optimized the RF classifier, yielding improved performance with specific parameter settings: $bootstrap = False$, $max_{depth} = 48$, $max_{features} = 'auto'$, $min_{samplesleaf} = 3$, and $min_{samplesSplit} = 5$. After

parameter tuning, 67.32% accuracy was obtained, indicating a substantial improvement over the initial performance of the model and proving that hyperparameter tuning worked well. This leads to a deeper and more complex tree ensemble, which in turn might reduce overfitting and improve generalization. The result reiterates the significance of systematically tuning in improving model accuracy and robustness, allowing a preference stamp on the RF classifier once fine-tuned for accurate prediction of water potability.

E. Proposed Ensemble Classifier

Based on the performance of the KNN and RF classifiers, a weighted voting-based ensemble technique was used to leverage the strengths of both classifiers and further improve classification accuracy. Each classifier's prediction is weighted according to its overall performance (F1-score). The final prediction is based on the class that receives the highest weighted vote, as shown in Algorithm 1.

Algorithm 1: Ensemble Model for Potability Classification using KNN and RF
Data: Training dataset with features X and labels y
Result: Prediction for potability $\hat{y}$ for a new sample x
Step 1: Train individual classifiers:
  Train KNN to obtain $P_{KNN}(x)$;
  Train RF to obtain $P_{RF}(x)$;
Step 2: Compute weights for classifier:
  Calculate F1-scores for KNN ($F1_{KNN}$) and RF ($F1_{RF}$);
  Computer normalized weights:
  $w_{KNN} = \frac{F1_{KNN}}{F1_{KNN}+F1_{RF}}$ and $w_{RF} = \frac{F1_{RF}}{F1_{KNN}+F1_{RF}}$
Step 3: Ensemble prediction:
  For each class $c \in \{nonpotable, potable\}$
   Calculate ensemble probability:
   $P_{Ensemble}(x) = w_{KNN} \cdot P_{KNN}(x) + w_{RF} \cdot P_{RF}(x)$
  Assign predicted class:
  $\hat{y} = arg\ \max_{c \in \{non-potable,\ potable\}} P_{Ensemble}(x)$
Step 4: Output predictions:
  Return $\hat{y}$

The ensemble model integrates predictions from KNN and RF using a weighted voting mechanism. Let $P_{KNN}(x)$ and $P_{RF}(x)$ represent the likelihood of a sample input features x belonging to class c. To balance the influence of these classifiers, their contributions are weighted by their respective F1-scores, $F1_{KNN}$ and $F1_{RF}$. The weights, calculated as $w_{KNN} = \frac{F1_{KNN}}{F1_{KNN}+F1_{RF}}$ and $w_{RF} = \frac{F1_{RF}}{F1_{KNN}+F1_{RF}}$, ensure that classifiers with higher performance metrics have a proportionally greater impact. The ensemble model aggregates these weighted probabilities as:

$$P_{Ensemble}(x) = w_{KNN} \cdot P_{KNN}(x) + w_{RF} \cdot P_{RF}(x)$$

The final classification decision, y, is made by selecting the class c with the highest ensemble probability.

V. RESULTS AND DISCUSSION

A. Software Requirements

The implementation of the water potability classification system was carried out in the Google Colab environment with Python ver. 3.12.3. Google Colab is an ideal platform for running the code, as it offers cloud-based resources. Key libraries required for the implementation include NumPy for numerical operations, Pandas for data manipulation, and visualization libraries such as Seaborn and Matplotlib for data exploration and analysis. Machine learning models were developed using scikit-learn libraries. Additionally, MinMaxScaler was used for feature scaling.

The Optuna library was used to tune the model's parameters. The TPESampler from the Optuna sampler's module enables the use of the Tree-structured Parzen Estimator algorithm.

B. Evaluation Metrics

Accuracy measures the proportion of correctly classified samples (both potable and non-potable) among all samples in the dataset, providing an overall assessment of the model's correctness.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision measures the proportion of correctly predicted potable water samples (true positives) among all samples predicted as potable, quantifying the model's ability to avoid false positives.

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

Recall measures the proportion of correctly predicted potable water samples (true positives) among all actual potable samples, indicating the model's ability to identify all positive instances.

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

The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both metrics.

$$\text{F1} - \text{score} = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (4)$$

Support refers to the number of actual occurrences of each class in the dataset, providing context to the precision, recall, and F1-score by showing the distribution of the classes. In these formulas, TP (True Positives) is the number of correctly predicted positive observations, TN (True Negatives) is the number of correctly predicted negative observations, FP (False Positives) is the number of incorrectly predicted positive observations, and FN (False Negatives) is the number of positive observations that were incorrectly predicted as negative.

C. Comparative Evaluation

LR showed moderate precision and F1-score but relatively lower recall, suggesting that it may struggle to correctly identify positive instances. KNN exhibited higher precision and F1-score, indicating better performance in correctly

determining positive cases, although its recall was slightly lower than that of RF. RF demonstrated the highest values across all metrics, suggesting that it effectively balances precision and recall, making it the most promising classifier for this task. These findings highlight RF as the preferred choice due to its overall superior performance in this classification context. Tables I, II, and III show the precision, recall, F1-score, and support observed for LR, KNN, and RF, respectively. The ensemble model had the best performance, as shown in Table IV.

TABLE I. PRECISION, RECALL, AND F1-SCORE OF LR

Class Label	Precision	Recall	F1-score	Support
Non-Potable	0.5	0.56	0.53	641
Potable	0.51	0.45	0.48	659

TABLE II. PRECISION, RECALL AND F1-SCORE OF KNN

Class Label	Precision	Recall	F1-score	Support
Non-Potable	0.64	0.85	0.61	641
Potable	0.63	0.68	0.65	659

TABLE III. PRECISION, RECALL, AND F1-SCORE OF RF

Class Label	Precision	Recall	F1-score	Support
Non-Potable	0.67	0.7	0.68	641
Potable	0.69	0.66	0.68	659

TABLE IV. PRECISION, RECALL AND F1-SCORE OF THE ENSEMBLE CLASSIFIER

Class Label	Precision	Recall	F1-score	Support
Non-Potable	0.70	0.78	0.74	641
Potable	0.73	0.75	0.74	659

The average accuracies of LR, KNN, RF, and Ensemble classifiers were 51%, 62%, 68%, and 71% respectively, showing a progressive improvement. LR, with an accuracy of 51%, performs only slightly better than random guessing, indicating that it struggles to classify water potability. It can be clearly understood that such unreliable classifications are unsuitable for real-world applications. KNN, with an accuracy of 62%, is relatively better, suggesting that it has learned some patterns in the data and is making more correct classifications. However, it still leaves significant room for error, which could be problematic if the model is used in sensitive areas such as public health or environmental monitoring. RF was better, with an accuracy of 68%. Finally, the proposed ensemble classifier shows the best performance among the four models. Although not perfect, with 71% accuracy, it still might not be sufficient for high-stakes decision-making, but with more training data, the accuracy can be further improved.

D. Novel Contributions and Knowledge Gaps

This research makes a significant contribution to the field of water potability classification by introducing a weighted ensemble classifier that combines the strengths of two well-established ML algorithms, RF and KNN. Although individual classifiers have been widely used in similar classification tasks, their performance can be limited by inherent biases or overfitting particular patterns in the data. By integrating these classifiers through a weighted voting scheme, the proposed ensemble approach overcomes these limitations, offering a more robust and accurate model for water potability classification and leading to improved generalization and performance on unseen data. While RF excels at capturing complex nonlinear patterns in high-dimensional datasets, KNN contributes through its ability to model local relationships in the feature space. By combining these classifiers with a weighted ensemble, the aim is to overcome the individual limitations of each method, thus improving overall predictive performance.

This study fills a critical gap by addressing both the hyperparameter optimization of the RF model and the imbalance in the dataset. The Optuna framework was used to ensure that the RF model parameters were fine-tuned for maximum predictive accuracy. Furthermore, the dataset used in this research was balanced using SMOTE. By incorporating these advanced techniques, namely hyperparameter optimization, ensemble learning, and data balancing, this study not only enhances the accuracy of water potability classification but also provides a more scalable and reliable solution for real-world applications in water quality monitoring.

VI. CONCLUSION AND FUTURE WORK

This study investigated the classification of water potability using LR, KNN, and RF algorithms, the latter two showing relatively better classification accuracy. As a result of fine-tuning, the RF model reached an accuracy of 68%, demonstrating its potential to accurately predict potable water. The dataset was useful in the interpretation of importance and observations for the feature set, which is necessary to bridge toward a thoroughly operational ML paradigm. When applied to solve complex classification problems relevant to water quality assessment, the performance of the RF classifier was found to be the best. It should be noted that further fine-tuning model parameters or experimenting with different algorithms might improve performance, but these results demonstrate potential efficacy. Although the accuracy of the algorithm improved, there is still much work to improve it further. Additionally, developing more ML models will be considered in the future. Future research could explore additional feature engineering strategies and evaluate the model's performance across diverse geographical and temporal datasets, further validating its applicability in real-world scenarios.

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. Shete, A. Bongale, and A. Bongale, "Internet of Things based Messaging Protocols for Aquaculture Applications - A Bibliometric Analysis and Review," *Library Philosophy and Practice (e-journal)*, Apr. 2021.

[6] 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.

[7] 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, 2023, Art. no. e1622, https://doi.org/10.1002/wat2.1622.

[8] 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, Sep. 2016, https://doi.org/10.1109/JSEN.2016.2570423.

[9] 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.

[10] 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.

[11] 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, Jan. 2023, Art. no. 3534, https://doi.org/10.3390/rs15143534.

[12] 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, https://doi.org/10.1016/j.heliyon.2023.e18170.

[13] K. Gunasekaran and S. Boopathi, "Artificial Intelligence in Water Treatments and Water Resource Assessments," in *Artificial Intelligence Applications in Water Treatment and Water Resource Management*, IGI Global Scientific Publishing, 2023, pp. 71–98.

[14] 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.

[15] M. Zhu *et al.*, "A review of the application of machine learning in water quality evaluation," *Eco-Environment & Health*, vol. 1, no. 2, pp. 107–116, Jun. 2022, https://doi.org/10.1016/j.eehl.2022.06.001.

[16] M. M. M. Syeed, M. S. Hossain, M. R. Karim, M. F. Uddin, M. Hasan, and R. H. Khan, "Surface water quality profiling using the water quality index, pollution index and statistical methods: A critical review," *Environmental and Sustainability Indicators*, vol. 18, Jun. 2023, Art. no. 100247, https://doi.org/10.1016/j.indic.2023.100247.

[17] N. Nasir *et al.*, "Water quality classification using machine learning algorithms," *Journal of Water Process Engineering*, vol. 48, Aug. 2022, Art. no. 102920, https://doi.org/10.1016/j.jwpe.2022.102920.

[18] T. Yan, S.-L. Shen, and A. Zhou, "Indices and models of surface water quality assessment: Review and perspectives," *Environmental Pollution*, vol. 308, Sep. 2022, Art. no. 119611, https://doi.org/10.1016/j.envpol.2022.119611.

[19] M. G. Uddin, S. Nash, A. Rahman, and A. I. Olbert, "Performance analysis of the water quality index model for predicting water state using machine learning techniques," *Process Safety and Environmental Protection*, vol. 169, pp. 808–828, Jan. 2023, https://doi.org/10.1016/j.psep.2022.11.073.

[20] J. Park, W. H. Lee, K. T. Kim, C. Y. Park, S. Lee, and T. Y. Heo, "Interpretation of ensemble learning to predict water quality using explainable artificial intelligence," *Science of The Total Environment*, vol. 832, Aug. 2022, Art. no. 155070, https://doi.org/10.1016/j.scitotenv.2022.155070.

[21] S. I. Abba *et al.*, "Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index," *Environmental Science and Pollution Research*, vol. 27, no. 33, pp. 41524–41539, Nov. 2020, https://doi.org/10.1007/s11356-020-09689-x.

[22] A. Kadiwal, "Water Quality." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/adityakadiwal/water-potability.