Explainable AI-Based Water Quality Prediction System

Jayarama Krishna Challagundla[1*], Alla Nithin Reddy[2], Arul Elango[3], Surya Vipparla[4], Manoj Dasari[5] and Rodda Maheswara Reddy[6]
{jayaramakrishnachallagundla@gmail.com[1], nithinreddyalla369@gmail.com[2], arulelango2012@gmail.com[3], suryavipparla5058@gmail.com[4], manojdasari456@gmail.com[5], maheswarareddyrrr5@gmail.com[6] }

Department of Advanced Computer Science and Engineering, VFSTR Deemed to be University, Vadlamudi, Guntur, 522213 andhra Pradesh, India[1, 2, 3, 4, 5, 6]

Abstract. Access to safe and clean drinking water is a global issue, especially in developing countries. Traditional water testing methods are always slow, costly and not available in remote areas. To overcome this limitation, we proposed Explainable AI based Water Quality Prediction System which uses machine learning and explainable models to classify whether the water is drinkable or not. We used a dataset which contains 3000 water samples, each sample contained six core parameters: pH, Turbidity (NTU), Chlorides (mg/L), Dissolved Solids (mg/L), Alkalinity (as $CaCO_3$) and Fluorides (mg/L). While preprocessing we converted the categorical values into numerical values and applied normalization using StandardScaler. Several classification algorithms were developed and evaluated, including Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Gaussian Naïve Bayes and Support Vector Machine (SVM). Furthermore, dimensionality reduction was performed using Linear Discriminant Analysis (LDA), with performance comparisons made before and after reduction. For model interpretability and trust, explainable AI techniques such as LIME and SHAP were utilized to highlight feature importance as well as provide insight into the decision process. The proposed system is highly accurate and interpretable and is therefore suitable for implementation in real-world, resource-constrained environments where accurate and interpretable water quality prediction is key to public health.

Keywords: Water Quality, Explainable AI (XAI), Machine Learning, SHAP, LIME, Potability Prediction, Dimensionality Reduction, Linear Discriminant Analysis (LDA), Classification Algorithms.

1 Introduction

Clean and safe drinking water is a basic requirement for public health and sustainable development. Nevertheless, water pollution is a major issue in most areas of the world, particularly in rural and underdeveloped areas. Conventional water quality analysis methods are usually expensive, time-consuming and laboratory facility-dependent and hence unsuitable for extensive applications. Machine learning has become a useful tool for predicting water quality from chemical parameters in recent years. However, the "black box" character of most ML models restricts user trust and interpretability. This paper introduces an Explainable AI-based Water Quality Prediction System that not only predicts water potability but also provides transparency through explainability tools like LIME and SHAP. By employing a set of 3000 samples against six important parameters, we can develop a model that is interpretable and accurate, which would be perfect to be used in actual, low-resource settings.

ICITSM-Part I 2025, April 28-29, Tiruchengode, India
Copyright © 2025 EAI
DOI 10.4108/eai.28-4-2025.2357836

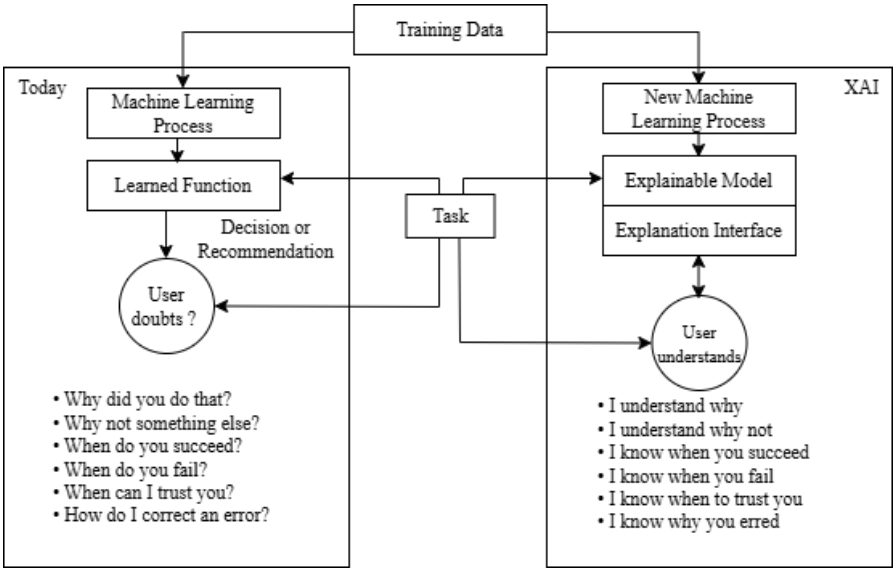

Fig. 1. Comparing Traditional Machine Learning with xAI's Explainable Approach.

Fig 1 contrasts old-school machine learning (left), where decisions might leave users scratching their heads, with xAI's fresh approach (right). xAI's explainable model and friendly interface help users easily grasp why choices are made, building trust and confidence.

The remaining parts of this paper is organized as follows: Part II describes related work, Part III describes the methodology, Part IV shows the results and analysis and Part V concludes the study with future work.

2 Related works

Decision Trees, XGBoost, SHAP, LIME is used in previous work for interpreting machine learning methods and explainable AI methods applied to water quality forecasting considering accuracy and interpretability of results maintaining public health significance. Alipio [1] refers to some of the machine learning and ensemble learning research, encompasses domains such as cost-sensitive techniques for sentiment analysis and soft voting for a single class. It refers to Think Speak and Scikit-learn for Machine Learning.

Muhammad I'tikaf Khoirul Haq [2] tried to classify water drinkability using machine learning which is Decision Tree and Naïve Bayes. Decision Tree produces 97.22% of right guesses which is much bigger than Naive Bayes 62.53% and therefore the work in classification water quality can be done by using Decision Tree.

Reading this XAI literature in depth, as reviewed by Saša Brdnik [3] there is a lot on the tipping point, especially when we look at the growing interest into healthcare. The roadblocks the critics and better explainability nominees are arguing about now, is how to create a standard process for evaluating these tools, which SHAP, LIME tend to be leading. Future

work the key areas of Ethical concerns and User-centred design onto OPDs that have been presented as opportunities for future works to further extend the research.

Water data of India was analysed by Bura Vijay Kumar [4] but without any scaling. Methods: This study was based on various methods such as CD RennaiScaled (re-mapping scale or re-scaling a pattern), Water Original Data UR, Luler, Min Max, Max Abs, Data RMS (Re-Moving of mean square) Wavelet and Wielkenscaler & L1Columnar. Water drink ability classification, Comparison of xgboost and random forest classifiers. XGBoost performed better with an accuracy of 87.22% as opposed to Random Forest whose accuracy was 61.11%. However, the XGBoost performed well in predicting water quality and handling the assumptions about relationships between the variables, giving good results.

Yusuf Husain [5] describes water pollution and the effects of contamination on health, writer with a quote that receives an approximate 2.5 million deaths/year. Topic It requires using new ways to be analyze and classify the quality of water. Some by literature review and Exploratory data analysis. Mehul Patel [6] envisions machine learning with not the original but rather just 20 features to predict water potability. F1-score of XGBoost: 0.9798 (highest) This study sets an example of typical water quality management where data preprocessing and model performance are emphasized.

In [7], Reem Alnaqeb compares ML models in predicting water potability where the LightGBM got 99.74% accuracy. Taken from an 8000-sample dataset, the discovery emphasized the crucial role of safe drinking water in public health and suggested that future research should be based on expanded data sets. Be it the machine learning prediction models build for water potability [8] by Aditya Kumar Mohapatra with 67% accuracy. The maximum difference has been observed in TDS and pH. Water resource management and public health are studied under the guidance of data-driven analysis.

In explainable AI, Trisna Ari Roshinta [9] does a comparison of LIME and SHAP with LIME excelling for its local explanations and efficiency while SHAP gives the global insights and stability. Depending on requirement whether we need quick results or stable, full blown feature importance analysis. Hasan: Machine learning classifier models for drinking water quality monitoring – comparing of the accuracy [10]. The related parameters that need to be focused are Turbidity, TDS and pH which would be major approach for efficiency and health security for water quality monitoring.

Julliet Kirui [11], used an Indian Government data and trained machine learning models targeting a house-level prediction of drinking water quality. The maximum accuracy which a XGBoost model could provide was 95.12% and it shows how these models can effectively be used in public health water quality monitoring applications.

Pongiannan R.K. [12] proposed the development of machine learning (ML)approach for real-time monitoring and prediction system of water quality using major parameters pH, Turbidity, Hardness etc. The water quality classification based on RF provides the excellent accuracy also because RF which is the ultimate promoter for environmental sustainability as well as public health monitoring. 2 Case study: forecasting water potability El-Bacha Rachid [13] explores predictive models for water potability with the help of machine learning techniques. The work demonstrates the promise of machine learning for making water quality prediction models more robust.

Productivity: Hybrid machine learning models are proposed by J. Biju [14] that integrate multiple algorithms and increase the rate of water quality prediction, which in turn provide greater accuracy and efficiency to monitor unsafe drinking water. V. Sreekumar [15], has shown a survey of different machine learning models for water potability prediction that describes the major significance of machine learning in management systems related to water.

3 Methodology

This section introduces the planned methodology of an Explainable AI-Based Water Quality Forecasting System. The process is divided into five important steps: data preprocessing, model selection, training and optimizing, applying explainable AI methods and performance evaluation. It aims at predictively determining potability in water accurately and defining model decision transparency through explainable AI models like SHAP and LIME.

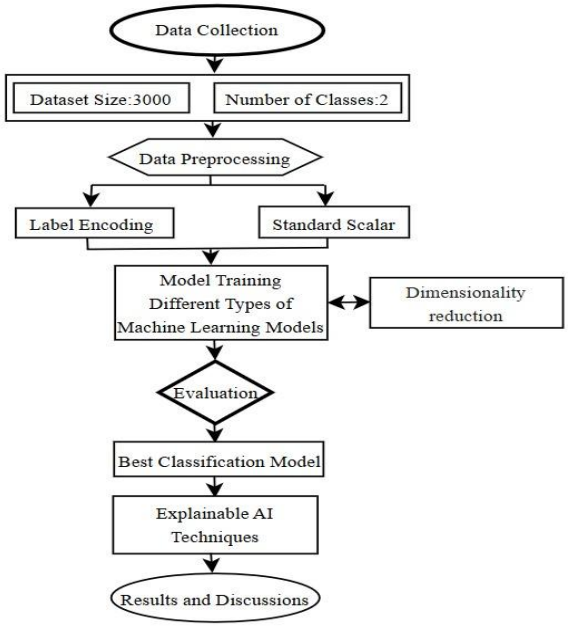

Fig. 2. Flow diagram of the proposed methodology.

Fig 2 illustrates the complete pipeline of the Explainable AI-Based Water Quality Prediction System. It begins with the gathering of six major water quality measurements. Data are preprocessed through techniques such as missing value treatment and normalization. Dimensionality reduction and feature separability are done through optional Linear Discriminant Analysis (LDA). Preprocessed data are employed to train machine learning classifiers. The best classifier is selected based on the performance metric. Finally, explainability techniques such as SHAP and LIME are applied to explain predictions in the model, resulting in an explainable and trustworthy system for water potability prediction.

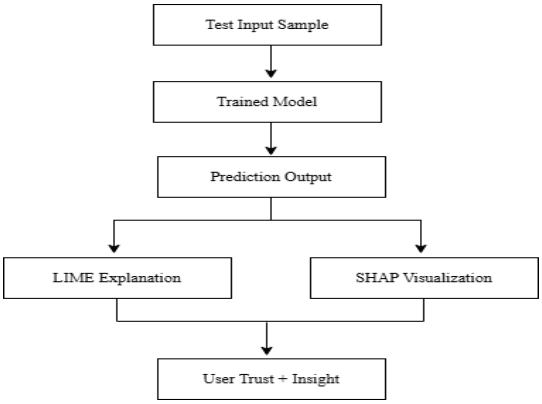

Fig. 3. Model Explainability Subsystem.

Fig 3 focuses on the explainability component of the system. A test input sample is passed through the trained model, producing a prediction on water potability. To explain the decision, LIME and SHAP techniques are used. LIME provides local explanations by approximating the model around a specific prediction, while SHAP offers a global perspective on feature contributions. These insights are visualized and presented to the user, increasing trust and understanding of how the model reaches its decisions.

3.1 Data Collection

A dataset comprising 3000 water samples was prepared for this study. Each sample contained six key physicochemical features relevant to water quality assessment:

- pH - Indicates the acidity or alkalinity of the water.
- Turbidity (NTU) - Measures the cloudiness caused by suspended solids.
- Chlorides (mg/L) - It is the content of chlorine derived compounds.
- Dissolved Solids (mg/L) - Refers to the number of suspended particles.
- Alkalinity (as $CaCO_3$) (mg/L) - Is the water's buffering capacity.
- Fluorides (mg/L) - They can impact health at higher levels water potability.

The target variable is binary, pointing out if the water is potable (1) or not potable (0).

3.2 Data Preprocessing

Careful preprocessing is critical to increase the predictive validity and accuracy of AI models. The following preprocessing methods were implemented in the obtained water quality dataset:

3.2.1 Handling Missing Values

Missing values samples were removed to preserve the integrity and coherence of the dataset.

3.2.2 Feature Scaling

All the features (pH, Turbidity, Chlorides, Dissolved Solids, Alkalinity and Fluorides) were normalized to a unified range through Min-Max normalization within the [0,1] range to enhance learning efficiency.

3.2.3 Outlier Removal

Major feature outliers were handled utilizing the IQR method to minimize the impact of outlier details regarding the training process.

3.2.4 Train-Test Split

Data split into 80% train and 20% test to get an impartial verdict of the model performance.

3.3 Model Selection: Learning Water Quality Patterns

3.3.1 Random Forest Classifier

A strong ensemble model that builds many decision trees and merges them to improve accurate and does not overfit. It is also suitable for nonlinear relationships as well as robust to noisy data.

3.3.2 Logistic Regression

A simple linear model but a very effective binary classifier, which can be used to find the impact of one feature on potability.

3.3.3 K-Nearest Neighbors (KNN)

A non-parametric method which categorizes the water samples based on likeness with characteristics of the 'k' nearest points. It is efficient and easy when there are non-linear decision boundaries.

3.3.4 Decision Tree Classifier

Data partitioning through tree model feature threshold values to generate human-readable rules for potability prediction. It's great for interpreting decision paths. and feature importance.

3.3.5 Gaussian Naïve Bayes

A probabilistic classifier that forecasts feature independence and is of a Gaussian distribution. It is a very efficient and scalable solution for real-time calculations.

3.3.6 SVM

A robust model that determines the optimal hyperplane to differentiate potable and non-potable samples. It is capable of handling high-dimensional data and when there is distinct margin separation.

All the models were compared to determine which one has predictive power on the water quality data set. The combination of various algorithms allows for a high-dimensional analysis of the data and it indicates to us which models best capture the underlying patterns between water quality parameters.

3.4 Dimensionality Reduction

For better classification performance and easier visualization, Linear Discriminant Analysis (LDA) was employed to decrease dimensionality. LDA projects the features into a lower-dimensional space with maximum class separability.

- LDA enabled identification of latent class distributions and eliminated noise caused by redundant features.
- Model performance was tested prior to and subsequent to the use of LDA to assess its impact.

Better interpretability of the model and small improvements in performance were revealed by the results, especially for algorithms to be impacted by the complexity of the feature space.

3.5 Explainable AI: Improving Transparency and Trust

In order to explain the model's predictions and understanding of feature contributions, we used explainability methods as described below:

3.5.1 Explainability Using SHAP

SHAP values were calculated to provide each feature's contribution towards specific predictions, global and local explanation. For further enhancing model interpretability, SHapley Additive exPlanations (SHAP) was used on the Logistic Regression model. SHAP provides a single feature contribution score through approximating each feature's contribution to the model output using game theory principles.

3.5.2 Mathematical Formulation of SHAP

SHAP distributes the importance value to each feature by dividing the prediction equally among the features. It is cooperative game theory-based and the explanation model is provided by:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \quad (1)$$

Where $f(x)$ is the model's output for instance x, ϕ_0 is the expected value of the model output and ϕ_i represents the contribution of feature i.

Each ϕ_i is computed as:

$$\phi_i = \sum_{s \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (2)$$

Here, S is any feature subset other than i, F is the universe of all features and f_S is the model learned from feature subset S. This definition ensures a fair contribution attribution of every feature to the global prediction.

3.6 LIME

LIME was employed to produce local explanations of about specific predictions by approximating the model with an interpretable, simpler surrogate model.

3.6.1 Feature Importance Visualization

Graphical feature influence representations (summary plots and dependency plots) were created to enable decision-makers and domain experts to comprehend model behavior.

3.6.2 Explainability Using LIME

To enhance transparency, the Local Interpretable Model-agnostic Explanations (LIME) framework was applied to the Logistic Regression model. LIME output is supplied in below fig 5 and 4.

3.6.3 Mathematical Formulation of LIME

LIME estimates the actual complex model f through a more conveniently explainable simple model g within a neighborhood around a specific example instance x. The intention here is to achieve the lowest for the objective function below

$$\underset{g \in G}{arg\,min}\ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{3}$$

Here, f is the complex model, g is an interpretable model (e.g., linear), G is the class of such models, $\mathcal{L}$ is a loss function that measures how closely g approximates f near x, π_x is a proximity function that gives weights to instances near x and Ω (g) is a regularization term that ensures g remains simple and interpretable.

4 Results and Evaluation

The proposed system was evaluated using multiple machine learning classifiers to predict water potability. Each model was tested both before and after applying Linear Discriminant Analysis (LDA) for dimensionality reduction. The comparison of classification accuracy is presented in Table 1.

4.1 Performance Evaluation

To assess the effectiveness and credibility of our system, we employed standard metrics:

4.2 Accuracy

Estimates the ratio of correctly predicted potable and non-potable samples. The blend of accurate forecasts and open explanations makes our system a dependable tool for water rural quality monitoring and decision-making urban environments.

Table 1. Accuracy Comparison of Models Before and After Dimensionality Reduction.

Algorithm	Accuracy	After LDA
Random Forest Classifier	0.93	0.87
KNN (Minkowski)	0.86	0.906
KNN (Manhattan)	0.85	0.905
KNN (Euclidean)	0.86	0.906
Decision Tree (Entropy)	0.80	0.81
Decision Tree (Gini)	0.81	0.81
Gaussian Naive Bayes	0.97	0.91
Support Vector Machine (SVM)	0.90	0.94
Logistic Regression	0.94	0.945

Table 1 depicts classification accuracies of varying machine learning algorithms prior to application of dimension reduction. KNeighborsClassifier and Logistic Regression exhibit improved post-reduction accuracy, with Logistic Regression with the highest score (0.945).

4.3 Performance Insights

From the results, it is evident that:

- The Gaussian Naive Bayes was the most accurate before dimensionality reduction (97%).
- Logistic Regression performed best after dimensionality reduction, with precision of 94.5%.
- K-Nearest Neighbors (KNN) algorithms, especially with Minkowski and Euclidean distances, showed improved accuracy following LDA, over the 90% threshold.
- Most models showed either a slight improvement or stable performance after LDA, suggesting that dimensionality reduction retained essential feature characteristics.

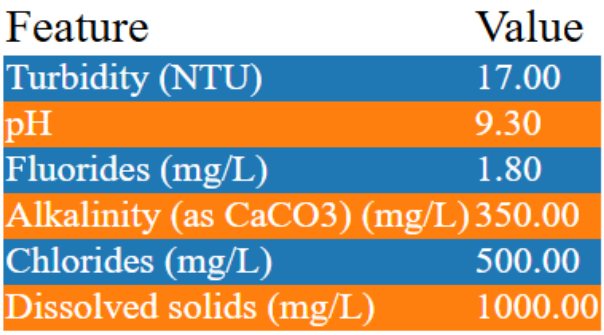

Fig. 4. Feature values used for LIME interpretation.

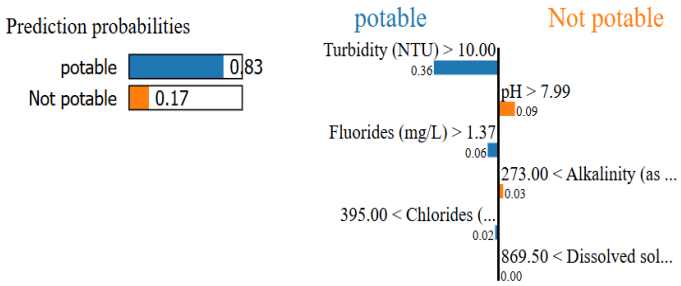

Fig. 5. LIME explanation showing feature contributions for classification.

Fig 4 and 5 present one example of interpretability using LIME. The example was classified as drinkable with 83% confidence. The key positive factors contributing to classification were Turbidity (>10), Fluorides (>1.37) and Chlorides (>395). High pH, Alkalinity and Dissolved Solids had a negative impact on classification.

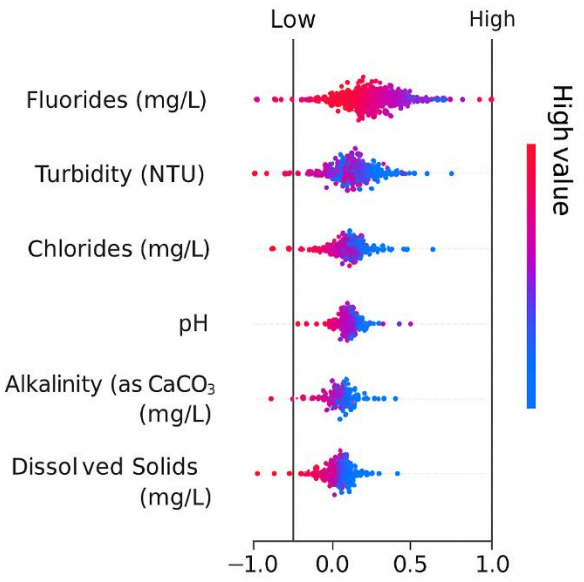

Fig. 6. SHAP summary plot of global feature importance for all predictions.

Fig 6 is a SHAP summary plot that shows global importance of all features for each prediction. Turbidity, Fluorides and Chlorides were the most positively influencing the model output for potability classification. The presence of Alkalinity and Dissolved Solids at high levels negatively affected the classification outcome.

Adding SHAP explanations strengthens LIME by providing global and local interpretability, making it easier to increase trust and transparency in model decisions.

5 Future Work

The above explainable AI-based water quality prediction system has numerous avenues through which it can be extended. One of the most promising ones may be to use deep learning-based models such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) to identify complex patterns from high-dimensional or temporal water quality data that can in turn result in better classification accuracy. The second path is deployment of the model on real applications on Internet of Things (IoT) devices on cloud or edge computing. This would be real-time and ongoing monitoring of water quality in rural and urban settings. From an interpretability standpoint, an ensemble strategy that employs more than one explainability method (e.g., LIME and SHAP) can yield stronger and more complete insights, particularly in high-stakes application domains where model explainability is necessary. Aside from that, integration of user or expert input into the system within the framework of an active learning loop can help the model build and improve itself on a continuous basis in the long run. Finally, we can add seasonally changing and geographically distributed water quality samples to the dataset to further generalize and make the system robust and allow more efficient largescale deployment in heterogeneous environments.

6 Conclusion

This study hypothesized an Explainable AI-based Water Quality Prediction System that could forecast water as potable or not potable using physicochemical characteristics. A 3000-sample dataset with six water quality parameters of significant impact was experimented with various machine learning classifiers like Random Forest, Logistic Regression, K-Nearest Neighbors, Decision Tree, Gaussian Naïve Bayes and Support Vector Machine. Use of Linear Discriminant Analysis (LDA) enhanced classification performance and interpretability of features. Additionally, explainable AI methods like SHAP and LIME were employed to enhance model interpretability by emphasizing feature contributions to point and overall predictions. The suggested system achieves high accuracy without compromising interpretability and is therefore deployable to real-world resource-constrained settings. The system is data-driven and flexible in decision-making for water safety and holds promise for application in scalable and reliable environmental monitoring.

References

[1] M. I. Alipio, "Data-driven IoT-based Water Quality Monitoring and Potability Classification System in Rural Areas," 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea (South), 2020, pp. 634-639, doi: 10.1109/ICTC49870.2020.9289505.

[2] M. I. Khoirul Haq, F. Dwi Ramadhan, F. Az-Zahra, L. Kurniawati and A. Helen, "Classification of Water Potability Using Machine Learning Algorithms," *2021 International Conference on Artificial Intelligence and Big Data Analytics*, Bandung, Indonesia, 2021, pp. 1-5, doi: 10.1109/ICAIBDA53487.2021.9689727.

[3] S. Brdnik and B. Šumak, "Current Trends, Challenges and Techniques in XAI Field; A Tertiary Study of XAI Research," 2024 47th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia, 2024, pp. 2032-2038, doi: 10.1109/MIPRO60963.2024.10569528.

[4] Akilandeswari, B. V. Kumar, S. Padmakala, T. Tirumalaikumari, R. Kesavan and M. J. Carmel Mary Belinda, "Comparative Analysis of XGBoost and Random Forest for Predicting Water Potability," 2023 2nd International Conference on Automation, Computing and Renewable

Systems (ICACRS), Pudukkottai, India, 2023, pp. 1656-1659, doi: 10.1109/ICACRS58579.2023.10405315.
[5] H. Yusuf, S. Alhaddad, S. Yusuf and N. Hewahi, "Classification of Water Potability Using Machine Learning Algorithms," 2022 International Conference on Data Analytics for Business and Industry (ICDABI), Sakhir, Bahrain, 2022, pp. 454-458, doi: 10.1109/ICDABI56818.2022.10041667.
[6] M. Patel, "Predicting Water Potability Using Machine Learning: A Comparative Analysis of Classification Algorithms," 2024 IEEE International Conference on Energy Internet (ICEI), Zhuhai, China, 2024, pp. 631-639, doi: 10.1109/ICEI63732.2024.10917239.
[7] R. Alnaqeb, F. Alrashdi, K. Alketbi and H. Ismail, "Machine Learning-based Water Potability Prediction," 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/AICCSA56895.2022.10017579.
[8] K. Mohapatra, P. Kumar, S. Samantara, J. J. Jena, A. Dash and S. Darshana, "Water Quality Analysis and Prediction of Potability," 2024 IEEE 4th International Conference on Applied Electromagnetics, Signal Processing, \& Communication (AESPC), Bhubaneswar, India, 2024, pp. 1-6, doi: 10.1109/AESPC63931.2024.10872207.
[9] T. A. Roshinta and S. Gábor, "A Comparative Study of LIME and SHAP for Enhancing Trustworthiness and Efficiency in Explainable AI Systems," 2024 IEEE International Conference on Computing (ICOCO), Kuala Lumpur, Malaysia, 2024, pp. 134-139, doi: 10.1109/ICOCO62848.2024.10928183.
[10] N. Hasan and K. M. Alhammadi, "Quality Monitoring of Abu Dhabi Drinking Water Using Machine Learning Classifiers," 2021 14th International Conference on Developments in eSystems Engineering (DeSE), Sharjah, United Arab Emirates, 2021, pp. 1-6, doi: 10.1109/DeSE54285.2021.9719373.
[11] J. Kirui, "Machine Learning Models for Drinking Water Quality Classification," 2024 International Conference on Control, Automation and Diagnosis (ICCAD), Paris, France, 2024, pp. 1-5, doi: 10.1109/ICCAD60883.2024.10553712.
[12] P. R.K., A. Sivakumar, S. Kohli, P. P.U., L. J. S. and P. M, "Water Quality Evaluation using Machine Learning," 2024 9th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2024, pp. 754-758, doi: 10.1109/ICCES63552.2024.10859485.I.A.Guitart, G. Hervet, S.Gelper, Competitive advertising strategies for programmatic television. Journal of the Academy of Marketing Science, 48, (2020)753-775. https://doi.org/10.1007/s11747-019-00691-5
[13] El-Bacha Rachid, Salhi Abderrahim, Abderrafia Hafid, Rabi Souad,Predicting water potability using a machine learning approach,Environmental Challenges,Volume 19,2025,101131,ISSN 2667-0100,https://doi.org/10.1016/j.envc.2025.101131.
[14] J. Biju, C. Badgujar and A. Poulose, "Hybrid Horizons: Advancing Water Potability Prediction Through Hybrid Machine Learning," *2024 Fifteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, Budapest, Hungary, 2024, pp. 175-180, doi: 10.1109/ICUFN61752.2024.10625242.
[15] V. Sreekumar, F. Ihsan, S. Reghuram and S. Sarath, "A Detailed Analysis of Machine Learning Models to Predict Water Potability," *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10725826.