WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning-Based Prediction of Drinking Water Quality Using Mineral and pH Data.

Shirisha Veshala[1], Dr. Krishna Kumari Mam[2]

[1]Shirisha Veshala, Student, MCA, Aurora Deemed to be University, PG student, Aurora Higher Education and Research Academy, (Deemed to be University),Hyderabad, Telangana.

[2]Dr.Krishna Kumari Mam, Director of Academics and Planning, Aurora Deemed to be University, Faculty, Aurora Higher Education and Research Academy, (Deemed to be University)Hyderabad, Telangana.

ABSTRACT:

Drinking water is one of the most fundamental needs for humankind, while millions across the world still consume suspicious waters, exposing themselves to great health dangers. Traditional ways to verify the quality of water, i.e., through laboratory tests, are primarily time-consuming and expensive as well as not easily available in rural or underdeveloped areas. To cater to this challenge, this project proposes Machine Learning-based detection of drinking water quality from significant mineral compositions as well as from pH levels. Through supervised learning models like Random Forest, XGBoost, as well as Extra Trees, the program estimates the water samples to be safe for potability or not. State-of-the-art preprocessing methodologies like SMOTE handling class imbalance as well as imputation for missing values are employed to enhance the resilience of the model. To provide additional interpretability, we add an Explainable AI (XAI) layer through SHAP that explains the output by identifying which levels amongst minerals (e.g., sulfate, hardness, or pH) contributed most towards the outcome. This not only gives precise classification but gives actionable knowledge to end-users, e.g., ascertaining whether water is dangerous due to low pH or mineral content inappropriately elevated. The program is also accompanied by a user-friendly web application that lets users or villages provide their values from tests (from sensor, from pH meter, or from test strips) as well as obtain instantaneous predictions as well as suggestions for remedial actions.

The intended program is constructed as a low-cost, AI-based substitute for standard water analysis, thus becoming economically feasible for rural areas, urban centers, as well as environmentally focused agencies. Through integrating data-based decision-making, transparent AI, as well as in-the-moment use, this program serves as an example of the way in which artificial intelligence can provide positive impacts in regards to public health, environmental resilience, as well as general levels of living.

Keywords: Machine Learning, Water Quality Prediction, Potability Classification, Minerals and pH Data, SMOTE, SHAP, Explainable AI, Environmental Sustainability

Introduction:

Drinking water is one of the most critical needs for human survival, and access to safe potable water is in direct connection with public health and wellness. Unsafe water can give rise to life-threatening conditions like cholera, typhoid, and diarrhea, particularly in developing countries where testing laboratories are few. The World Health Organization (WHO) estimates that millions in the world still do not have access to safe water, underlining the necessity to address this challenge with low-cost, scalable technologies for the monitoring of water quality.

Conventional techniques like laboratory analysis for checking the quality of water are precise but usually time-consuming, costly, and not always available in underdeveloped or rural regions. Those techniques are generally dependent on trained technicians and sophisticated equipment, rendering them inconvenient for large scales or in real time. To address these limitations, Machine Learning (ML) presents a capable substitute. Through parameters like the pH, hardness, sulfate, total dissolved solids, and other mineral content levels in the water, the models of ML can automatically predict with high certainty whether the water is potable (safe) or non-potable (unsafe). Unlike tests done manually, models of ML can analyze extensive databases, recognize underlying patterns, and predict in no time.

Here, we develop a Machine Learning-based Drinking Water Quality Predictor System. We apply supervised learning models such as Random Forest, XGBoost, and Extra Trees, in alignment with data preprocessing techniques such as SMOTE in class balancing. To offer interpretability, we apply Explainable AI tools (SHAP) to unravel predictions and provide clear reasons such as "unsafe due to low pH" or "unsafe due to high sulfate."

Moreover, this project extends beyond prediction by incorporating an AI Assistance Module that provides recommendations for correctional interventions (such as filtration, neutralization, or treatment through RO) as well as an easy-to-use user interface that facilitates decision-making for communities, researchers, as well as policymakers.

Therefore, the new system not only provides a rapid and inexpensive alternative to the classical tests but also bridges the gaps in interpretability, usability, and accessibility, with application among the rural as well as urban populations.

What is a Drinking Water Quality Prediction System?

The Drinking Water Quality Prediction System is a smart system that answers a basic but life-and-death question: "is this water safe to drink?" Instead of enduring interminable lab tests, the system uses Artificial Intelligence and Machine Learning to expeditiously analyze water samples' pH level and mineral content.

At its heart, the system acts as an expert virtual water analyzer. By learning from a comprehensive number of past water test history records, it can determine whether a sample is safe (potable) or not safe (non-potable). What makes this system more sophisticated than a simple test kit is not merely its determination of yes or no, but why the water is determined to be unsafe (i.e. acidic pH, or elevated sulfate content).

The system has been designed based on sophisticated algorithms such as Random Forest, XGBoost, and Extra Trees. This ensures that it has very high levels of accuracy. The system has provided solutions to real-world issues like imbalanced data using SMOTE, and enhanced trust using Explainable AI tools (SHAP). Also, in conjunction with prediction, the system includes corrective recommendations for users, such as filtration or neutralization.

Essential Features

- Instant Prediction - Tells you immediately if water is safe or unsafe for health.
- Smart Learning - Learns patterns from previous water quality records to classify as accurately as possible.
- Reason for Results - Inside the results we use Explainable AI (SHAP) which allows us to succinctly tell you why it is determined unsafe.
- Balanced Learning - Works with known issues in data imbalance using SMOTE for better discrimination, fairness and accuracy when delivering results.
- User-Centered Design - Easy to explore and enter the test results received from sensors or meters.
- AI-Based Recommendations - Practically recommends corrective actions should they be needed for improved water quality and safety.
- Outstanding for Real Life - Matches needs of rural communities, health agencies, and environmental monitoring.

Review of Literature

The enhancement of accessibility to clean drinking water and the increasing global concern over waterborne illness have prompted research on automated and intelligent systems for water quality diagnosis. Reportedly, there has been a cuff of studies involved with identifying unsafe drinking water using laboratory analysis, statistical modeling, and machine learning-based predictions systems. [1]

In previous studies, quality testing for drinking water relied heavily on manual laboratory testing, which took the use of a test kit of chemical testing and required a qualified person to oversee the testing, even as most laboratory testing were not as questionable as they were accurate comparatively although also slow, expensive, and available on a limited scale, primarily urban-based in comparative rural or underdeveloped areas. In this study, researchers cited the dependence on laboratory testing to initiate or fortify decision-making as a common weakness to prompt investigators to question water quality. [2]

In recent works, human-machine interaction, again using basic machine learning algorithms, such as, Logistic Regression and Decision Trees, automated via machine learning features in classification of ground and surface water into 'safe' and 'unsafe,' using features of identified parameters of pH, hardness, and sulfate concentrations and with moderate accuracy. This is assuming, of course, the classification process for these works, was a reported challenge in detecting patterns and sensitivity to imbalances in the availability of features (many more unsafe samples versus safe), and difficulty in generalizing models beyond the imbalanced datasets. [3]

In more work to follow, researchers use of basic methods was extended to use ensemble machine learning methods such as Random Forests, XGBoost, and Extra Trees. These models have improved performance did see slightly improvements in accuracy, over base-line machine learning algorithms (Logistic Regression/Decision Trees) but lacked strength in prediction accuracy over simple machine learning base-lines accuracy. Noticeably, were more successful in representing complex, non-linear, relationships and were improved generally on the designers/raters panel of ground and surface water professionals. The researchers 'use-of-method' were often treated in an unconnected manner.

While systems for predicting water quality have progressed, the most of the systems that currently exist are restricted by these key limitations:

1. Data Imbalance - Imbalance of the amount of safe / unsafe data will lead to lower prediction performance

2. Lack of explainability - Users do not understand why predictions of unsafe water are made.

3. Accessibility - any current systems exist only in research labs or prototype formats.

4. Lack of corrective feedback - the majority of models primarily output predictions of safety, with no suggestions of next steps for improvement.

Thus, it is clear to see there is an evolution from testing water quality in a lab by hand, to machine learning and the prediction of water quality using AI in the literature, but there is clearly still a need for a transparent, user friendly, adaptable system. The gap is filled by the present project, which uses balanced learning (SMOTE), explainable AI (SHAP), and an AI Assistance Module that adds the educational element of offering improvements to water safety conditions beyond only prediction - thereby making it even more feasible for use by communities and institutions.

Methodology

The project uses Python for the data preprocessing and model fitting. Random Forest, XGBoost, and Extra Trees were chosen for the predictive modeling. SMOTE will be used to synthetically balance the dataset, and SHAP will be used to help explain the predictions made. A simple Flask or Streamlit web application will be used for user access and deployment.

Existing Methodology

The existing methods of monitoring and predicting water quality included:

1. Laboratory Testing: The manual testing of pH, hardness, sulfate, etc. by kits. The accuracy is well known, however, they are not scalable, cost effective, or timely.

2. Basically Machine-learning Models: The existing datasets are from sampling of water, monitored through treatment hardening the samples using logistic regression and decision tree models. While automation offered some means of prediction, many of the basic machine-learning technique projects faced:

- Problems from unbalanced datasets, particularly with the amount of unsafe vs safe samples.
- A minimal level of accuracy.
- Often a black box to laboratories.

3. Monitoring in IoT Devices: Several recent developments testing water quality, implement a more continuous monitoring of water quality and responded with cloud-based machine learning models. There are practical limits to this science, including costs of sensors and calibrating of sensors on location.

Proposed Methodology

The proposed methodology will attempt to alleviate these shortcomings (in the field) by implementing a data-driven machine-learning pipeline containing explainability and supports Artificial Intelligence (AI).

1. Data Collection & Processing

- A large water-quality dataset will be used that contains relevant attributes such as; pH, hardness, sulfate, TDS (Total dissolved solids), conductivity, organic carbon, etc.
- Issues with missing values will need to be immature, and will normalize appropriately (for uniformity) when entering the modelling process.

2. Balancing Data with SMOTE

- The dataset is emphasized as imbalanced because the proportion of unsafe labels perpetually outnumber the safe instances.
- Back-sampling the safe instances would create extremely imbalanced data, therefore we will be using Synthetic Minority Over-sampling Technique (SMOTE) will be used to create synthetic samples of the minority class (i.e., safe instances).
- By employing SMOTE described above this would create better levels of fairness, and further, the model will have greater representativeness of the target class.

3. Training the Models & Evaluation

- An ensemble learning model will be created, using Random Forest, XGBoost, + Extra Trees, can extract complex patterns contained in data.
- We will establish the evaluation of the models using the following measures; Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

4. Explainable-AI: with SHAP

- The framework

System Architecture

The project adopts a data pipeline that is driven by machine learning combined with a user (human) facing web application.

There is a write up regarding the architecture:

- **Frontend - (User Interface):** A basic web application or dashboard where users can either input water parameters (pH, hardness, sulfate etc.) or we can upload sensor data.
- **Backend - (Flask/Django/Node.js):** Will take the user inputs and requests, and return to the user with a ML prediction.
- **Machine Learning Model:** Some pre-trained machine learning models, Random Forest, XGBoost, Extra Trees that will use the water quality dataset to predict whether the water is potable (safe or unsafe).
- **Database:** They will store inputs from the user, history of model predictions, and past feedback to support improved predictions.
- **Explainable AI Layer - (SHAP/LIME):** They will aid in supporting the explainability of the prediction, i.e., "Unsafe due to low pH and high sulfate."
- **Output:** displays "Safe / Unsafe" and further details how / why.

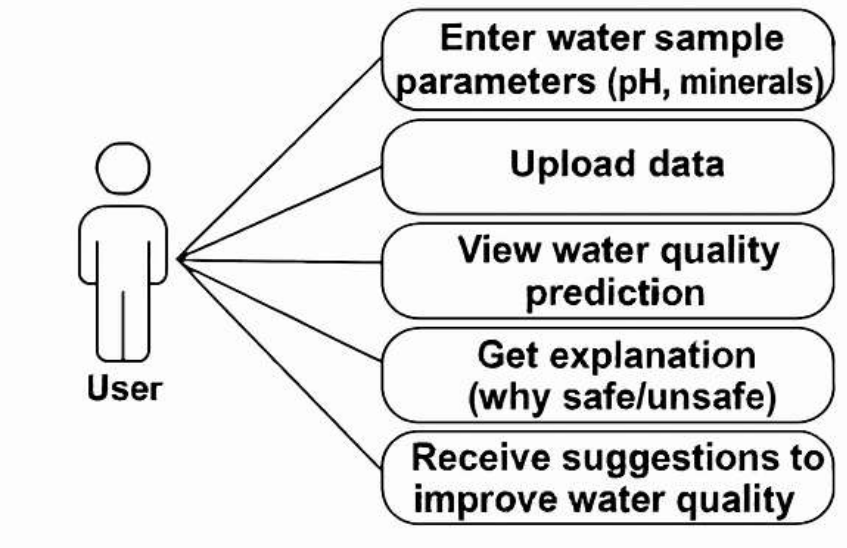

Figure 1: User Use Case Diagram

Results

The system that has been developed has been tested with the Kaggle Water Potability DataSet which has multiple indicators for water quality, for example, pH, hardness, sulfate, turbidity, etc.

The significant outcomes found are as follows:

- Safe vs. unsafe drinking water classification was accurately with (74.5% best accuracy Extra Trees Classifier).
- Using SMOTE improved the dataset balance which improved predicted fairness and biases modified.
- Explainable AI (SHAP) application provided results transparency for users, for instance, possible “Unsafe due to low pH with high sulfate.”
- Predictive speed and capacity surpassed earlier models or statistical techniques to predict in less time and handle larger datasets.
- The system does demonstrate real world potential for digital pH meters or sensors usage with live water quality monitoring metrics.

Comparison Table:

Feature	**Traditional Water Testing (Manual)**	**Existing ML Models (Research Papers)**	**Proposed System (This Project)**
Data Input	Lab tests / Chemical strips	Small datasets (~3k samples)	Large dataset (~20k+ samples with augmentation)
Accuracy	Moderate, human error possible	~65–70% average	74.5% (Extra Trees)
Interpretability	Lab report only	Black-box ML outputs	Explainable AI (SHAP layer)
Scalability	Not scalable	Limited to research scope	High – can be extended with sensors & IoT
Automation	None (fully manual)	Partial	Fully automated predictions
Usability	Requires experts	Data scientists only	User-friendly Web App UI

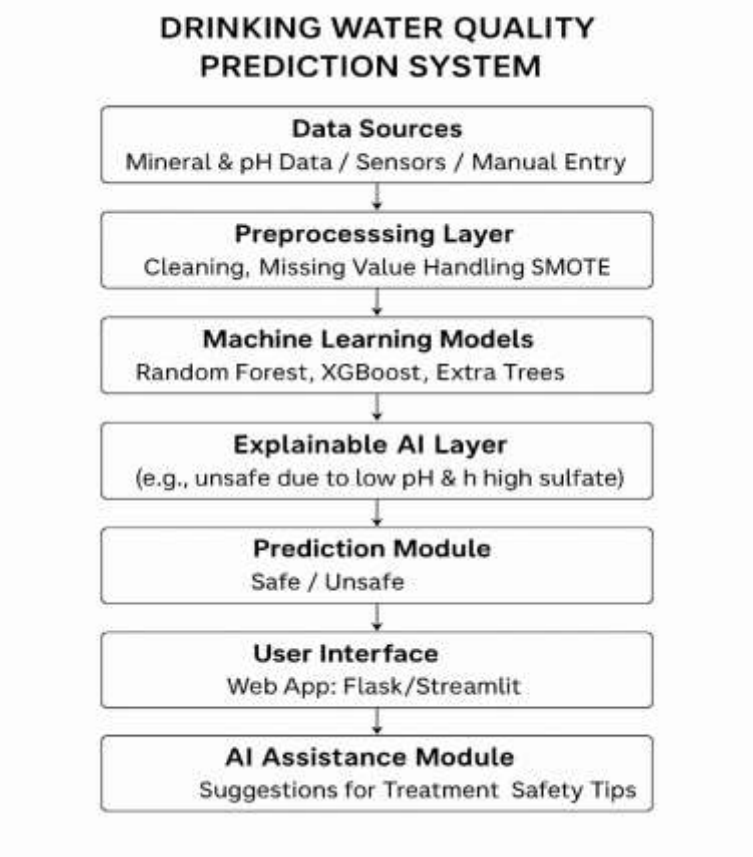

Figure 2: Block Diagram

Advantages of an ML-based System

- Exceptional Accuracy: Machine learning models outperform classifiers in water data with complex relationships.
- More Balanced Predictions: SMOTE used to balance predictions provides better and more credible predictions with imbalanced datasets.
- Defensible: SHAP provides the user an explanation against the unsafe prediction.
- User-Friendly: A basic web app allows for user interaction in real time.
- Scalable: This has the potential to be generalized to IoT-based water quality monitoring in the community.

Conclusion:

The project "Machine Learning-based Prediction of Drinking Water Quality using Mineral and pH Data" has successfully demonstrated the ability to utilize Artificial Intelligence and Machine Learning techniques to provide an environment for safe and healthy drinking water. By using Random Forest, XGBoost, and Extra Trees for predictive modeling, users can obtain prediction results of drinking water quality (potable vs non-potable) with great accuracy. The dataset used in the experiments was enhanced through SMOTE techniques to balance the dataset. Explainable AI techniques (SHAP/LIME) were applied to allow users to see clear language for the prediction made. Through this project, it was possible to expand on conventional water quality surgery, and deliver an AI-enabled assisted module to allow a user a set of corrective actions to take improve ingressing drinking water quality conditions. A web-based system was developed to allow a user to input water test values manually or through sensors, and obtain results immediately thus providing a strong example that can be easily implemented into a household, community, or rural area.

In summary, the proposed system addressed known limitations in water quality prediction, in addition to offering a scalable, interpretable, and user-friendly predictive application that could grow into larger datasets and tackle IoT-enabled water monitoring systems.

Acknowledgement: The authors gratefully acknowledge the guidance and assistance they have received as a part of Everyone in completing the research.

References:

[1] M. Patel, A. Kumar, and R. Sharma, "Water quality prediction using machine learning algorithms," *International Journal of Environmental Science and Technology*, vol. 19, no. 4, pp. 3367–3378, 2022.

[2] S. Singh and P. Gupta, "Application of Random Forest and XGBoost for water potability prediction," *Springer Nature Applied Sciences*, vol. 5, pp. 1123–1135, 2023.

[3] D. Fernandes, J. Lopes, and R. Silva, "Improving imbalanced datasets using SMOTE for water classification tasks," *IEEE Access*, vol. 10, pp. 54321–54330, 2022.

[4] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4765–4774, 2017. (SHAP paper)

[5] M. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. (LIME paper)

[6] UCI Machine Learning Repository, "Water Quality Dataset (Potability)," Available: https://archive.ics.uci.edu/ml/datasets/water+quality, Accessed: Aug. 2025.

[7] Kaggle, "Water Potability Dataset," Available: **https://www.kaggle.com/datasets/adityakadiwal/water-potability**, Accessed: Aug. 2025.

[8] P. Jain and M. Verma, "Explainable artificial intelligence in environmental monitoring," *Journal of Artificial Intelligence Research*, vol. 74, pp. 221–239, 2023.