Safeguarding Public Health Through Comprehensive Water Purity Analysis

Dr.D. Kishore Babu
Computer Science and Engineering
Bapatla Engineering College
Acharya Nagarjuna University
Bapatla, India
(Associate Professor)

Ponakala Neelima
Computer Science and Engineering
Bapatla Engineering College
Acharya Nagarjuna University
Bapatla, India
(Y21ACS543@becbapatla.ac.in)

T.Ramanjaneyulu
Computer Science and Engineering
Bapatla Engineering College
Acharya Nagarjuna University
Bapatla, India
(Y21ACS574@becbapatla.ac.in)

R.Sravani Bai
Computer Science and Engineering
Bapatla Engineering College
Acharya Nagarjuna University
Bapatla, India
(Y21ACS554@becbapatla.ac.in)

T.Siva Krishna
Computer Science and Engineering
Bapatla Engineering College
Acharya Nagarjuna University
Bapatla, India
(Y21ACS577@becbapatla.ac.in)

***Abstract*— Access to clean drinking water is important for public health, because polluted water causes significant risks for human health and the environment too. This study outlines a machine learning approach to find water safety by analyzing nine important water quality parameters such as pH levels , hardness , turbidity , arsenic , chloramine , bacterial presence , lead concentration , nitrate levels , and mercury content. The system incorporates data-preprocessing and data-exploring that ensures numeric data types for attributes without any missing and NaN values to guarantee better performance.The framework utilizes the predictive classification models which are Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine. The base model predictions are aggregated using a meta-model built with Logistic Regression, leveraging ensemble learning techniques and cross-validation to improve predictive accuracy. The meta-model combines the strengths of individual classifiers, enhancing overall performance metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that the meta-model achieves an accuracy of 93% outperforming individual base models.This research contributes to safeguarding public health by providing a reliable and efficient method for evaluating water safety.In further enhancement the practical application of this project, future work will mainly focus on developing a more user-friendly interface that allows water quality monitoring personals to easily input water quality data and receive immediate solutions of water safety after assessment. Additional, exploring the integration of real-time sensor data and cloud-based platforms could enable continuous storing and monitoring of water quality, leading to more timely interventions and improved public health outcomes in water-scarce or pollution-prone areas.**

Keywords—Decision Tree,Ensemble Learning,Logistic Regression,Machine Learning,Random Forest classifier, Support Vector Machine and Water Purity

I. INTRODUCTION

Ensuring safe drinking water is a global challenge that continues to threaten public health and environmental sustainability. Contaminated water often leads to the spreading diseases, environmental degradation and causes difficulty to achieve sustainable development goals. In regions with limited resources, water purity monitoring frequently relies on traditional testing methods that are time-consuming and lacks scalability,adaptability. Technological advancements particularly in artificial intelligence gives an opportunity to fill these gaps by providing solutions that are precise, automated and adaptable to different contexts. Addressing the growing need for efficient water safety evaluation systems, this study explores the potential of machine learning in analyzing water purity parameters to make sure drinking water safety.

This research focuses on developing a framework that integrates machine learning techniques for real-time and accurate water quality assessment. By analyzing nine physical,chemical and biological water quality parameters which are pH levels, hardness, turbidity, arsenic, chloramine, bacterial presence, lead concentration, nitrate levels, and mercury content.The study aims to create a system capable of predicting water safety. Four supervised machine learning models are employed that classifies whether water is safe to use or not which are Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine.Each base model is trained to analyze these parameters independently. Their outputs are refined through a meta-model based on Logistic Regression which uses ensemble learning techniques to combine the strengths of individual classifiers and improve overall prediction performance.

This system has ability to transform conventional approaches to water quality evaluation by offering a scalable, reliable and data-driven solution. By automating the process, the framework significantly improves the efficiency of water safety monitoring thereby contributing to public health and environmental protection efforts. The significance of this study extends beyond water safety; it demonstrates how cutting-edge computational methods can play a vital role in

addressing real-world challenges and advancing societal well-being.

II. Literature Review

Research in water quality assessment and purification has evolved considerably over the years, integrating advanced technologies to address critical challenges in ensuring safe drinking water.In 2016, Achio, Kutsanedzie, and Ameko conducted a study titled *"Comparative Analysis of Filtration Techniques for Water Purity"*. This research examined the effectiveness of physical filtration methods for removing water contaminants. While it provided valuable insights into physical purification processes, it failed to address chemical and biological contaminants, leaving a significant gap in understanding comprehensive water filtration. Subsequent studies aimed to overcome this limitation by incorporating chemical analysis into water purification frameworks to improve overall contaminant removal [1].

In 2019, Muniz and Oliveira-Filho published *"Multivariate Statistical Analysis for Water Quality Assessment"*, which focused on the application of statistical methods such as Principal Component Analysis (PCA) and Cluster Analysis to evaluate water quality in various ecosystems. These multivariate approaches highlighted patterns and relationships in water quality data, improving understanding of contamination levels. However, the absence of real-time monitoring technologies limited the applicability of the findings in dynamic water quality scenarios. This limitation prompted future research to integrate IoT and machine learning technologies for real-time data analysis [2].

By 2020, Ahmed et al. introduced emerging technologies for water quality monitoring in their study titled *"Water Quality Monitoring: From Conventional to Emerging Technologies"*. Their research proposed IoT-based systems for real-time water quality monitoring, marking a shift from traditional methods to more automated processes. Despite the benefits of IoT systems, the high costs and infrastructural requirements posed challenges, driving researchers to explore more cost-effective solutions. Machine learning was subsequently adopted to reduce expenses and enhance predictive capabilities, addressing these constraints [3].

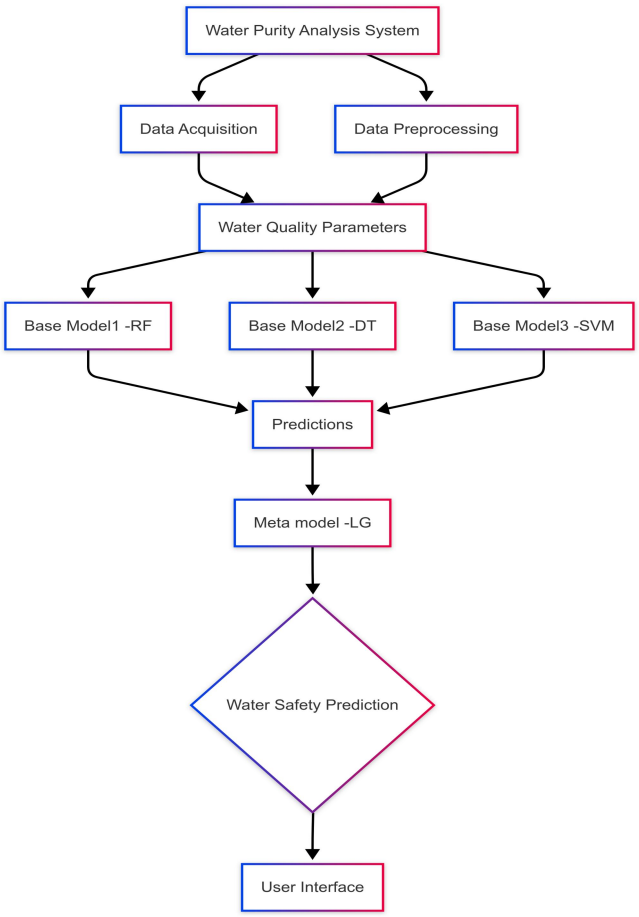
In 2023, Malagi presented *"Water Potability Prediction Using Machine Learning"*, which explored the use of algorithms such as Random Forest and Decision Trees to predict water potability. This study demonstrated the potential of machine learning in improving prediction accuracy but was limited by the use of small datasets, which restricted the generalizability of its results. The need for larger datasets and optimized algorithms was highlighted as a future direction to bolster the reliability and performance of water safety predictions [4].

These research efforts collectively underscore the evolution of water quality assessment techniques, from traditional filtration methods to advanced IoT and machine learning approaches. Building upon the contributions of these studies, the current research integrates multiple machine learning models and ensemble techniques to develop a scalable and automated system for comprehensive water purity analysis. This framework aims to address the limitations of earlier methods, offering a reliable solution for safeguarding public health.

III. Methodology

The methodology adopted for this research integrates machine learning techniques with water quality data to build a comprehensive water purity analysis system aimed at safeguarding public health.This section outlines the systematic steps taken throughout the project from data collection,exploring, preprocessing to model training and evaluation.The primary objective of the project was to predict whether water is safe or unsafe based on the input parameters such as pH, hardness, turbidity, arsenic, chloramine, bacteria, lead, nitrate, and mercury.

A. System Architecture



The architecture consists of three base models Random Forest, Decision Tree, and Support Vector Machine (SVM) that independently provide predictions, which are further processed by the meta-model, a Logistic Regression ensemble, to generate the final prediction on water safety.The proposed Water Purity Analysis System architecture is composed of three main components: the Input Layer, the Base Models, and the Meta Model. At the core, the system harmonizes ensemble learning principles by integrating predictions from multiple base models into the meta-model for improved accuracy. This modular setup ensures each component independently contributes to the overall prediction process. The Input Layer collects water quality data, which is then analyzed by the Base Models, each functioning autonomously to generate preliminary predictions. These predictions are subsequently aggregated by the Meta Model to produce the final output of "Safe" or

"Not Safe." This architecture facilitates scalability and modularity with each part designed to operate independently while maintaining seamless interaction among the components.

B. Data Collection

The performance and reliability of any machine learning based system heavily depend on the quality and diversity of the dataset used for training and evaluation. For this project, two datasets were employed—**water potability** and **water quality,** which consists of distinct kinds of physical parameters,chemical and biological contaminants that affects the water quality.Those are:

1) Water Potability Dataset

The water potability dataset consists of 3276 entries of different kinds of contaminants that effects purity of water. The major attributes include:

- **pH** - Water's acidity or alkalinity.
- **Hardness**- Calcium and magnesium concentration.
- **Solids** - Total dissolved substances.
- **Chloramines** - Disinfectant compounds.
- **Sulfate** - Sulfate ion levels.
- **Conductivity** - Ionic content indicator.
- **Organic_carbon** - Organic pollution content.
- **Trihalomethanes** - Chlorination byproducts.
- **Turbidity** - Water cloudiness.

Such a dataset provides only basic varities of contaminants and majorly focuses on physical parameters of water.

2) Water Quality Dataset

The second dataset deals chemical and biological contaminants in water. It consists of about 7999 records and includes:

- **Aluminium**- Concentration of aluminium in water.
- **Ammonia**- Level of ammonia, a nitrogen compound.
- **Arsenic**- Toxic element concentration in water.
- **Barium**- Amount of barium ions in water.
- **Cadmium**- Presence of cadmium, a heavy metal.
- **Chloramine**- Disinfectant compounds in treated water.
- **Chromium**- Level of chromium, often from industrial waste.
- **Copper**- Concentration of copper, essential in small amounts but toxic in excess.
- **Fluoride**- Fluoride levels to prevent tooth decay but harmful in excess.
- **Bacteria**- Microbial contamination in water.
- **Viruses** - Presence of viral pathogens in water.
- **Lead**- Lead concentration, harmful even in trace amounts.
- **Nitrates**- Nitrogen compounds from agricultural runoff.
- **Nitrites**- Nitrogenous compounds harmful to health.
- **Mercury**- Toxic heavy metal concentration in water.
- **Perchlorate**- Levels of perchlorate, a pollutant from industrial and military sources.
- **Radium**- Radioactive element concentration in water.
- **Selenium**- Trace element necessary in small amounts but toxic in excess.
- **Silver**- Silver ions used in disinfection, harmful in excess.
- **Uranium**- Concentration of radioactive uranium in water.

The next preprocessing will employ a system in which crop names are attributes are converted into numeric datatypes and missing values,NaN values are replaced with median values order to achieve maximum model performance.

3) Data Source and Preprocessing

The main data sources for the two datasets are publicly available water quality datasets, scientific papers, and water monitoring government websites. Subsequently, diverse preprocessing techniques were applied for:

- The treatment of missing values and NaN values.
- Modified the dataset by adding and removing columns. (New dataset columns are—ph, hardness, turbidity,arsenic, chloramine, bacteria, lead, nitrates, mecury, is_safe)
- Scaler is used to standardize the values.

C. Data Analysis and Visulization

A heatmap based on Pearson correlation coefficients was constructed to analyze the relationships among the features, as shown in Figure 1.The correlation analysis reveals several moderate positive and negative interactions among variables, such as arsenic and chloramine, and mercury and water safety status (is_safe). However, the overall low interaction among features indicates minimal multicollinearity, making these variables suitable candidates for use in supervised learning models without extensive feature elimination. The heatmap provides valuable insights into feature dependencies, aiding in model optimization.

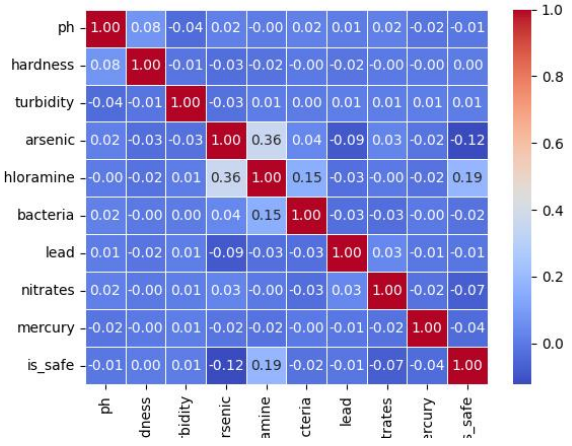


Figure 1: feature Correlation Heatmap of Water Quality datset

Figure 2 illustrates the distributions of the input features used used for water quality prediction like (ph, hardness, turbidity, arsenic, chloramine, bacteria, lead, nitrates, mercury, is_safe).Most features exhibit a bell-shaped or moderately skewed distribution. These histograms help assess the necessity for scaling or normalization prior to model training.

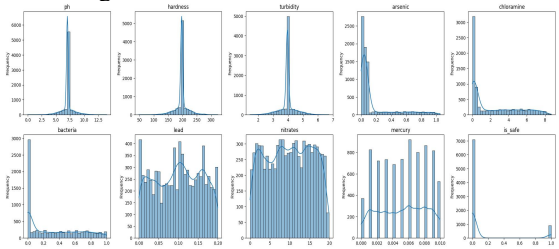


Figure 2

D. Model Development

The system for safeguarding public health through comprehensive water purity analysis leverages machine learning models to predict water safety. Input features include pH, hardness, turbidity, arsenic, chloramine, bacteria, lead, nitrate, and mercury—critical indicators of water quality. The architecture consists of two layers for in model developing: base models and a meta-model. The base models including Random Forest, Decision Tree, and Support Vector Machine (SVM), are individually trained and validated using the input features to generate preliminary safety predictions. These predictions serve as inputs to the meta-model which is Logistic Regression, which performs ensemble learning to produce final predictions about whether the water is safe or unsafe for consumption. This layered design ensures high accuracy by combining the strengths of multiple algorithms while managing feature dependencies effectively.

IV. Evaluation Metrics

The assessment of performance for machine learning model was based on a given evaluation metric founded on the type of task being performed by a model: classification model.

A. Classification Evaluation

For the classification models used in evaluating crop recommendations, several measures were considered:

- Accuracy: it is defined as the ratio of the correctly predicted crop labels to the total number of predictions made. This measure essentially gives an overall view of how good a model is performing descriptive-wise.
- Precision and Recall: Precision measures the exactness of the classifier predicting an instance of a specific crop while recall measures its ability to identify all relevant instances of the given crop. This is viable while analyzing the performance of individual classes in a multi-class classifier.
- F1-Score: F1 is a measure for the harmonic mean of precision and recall, thus giving one measure by making a balance between both of them. This is especially useful in the case of a highly imbalanced dataset.
- Confusion Matrix: Regarding visualization of the model performance, a confusion matrix was generated for appreciation of understandings of misclassification tendencies across different crop categories.

V. Results and Discussions

1) Performance Metrics of Base models and Meta-Model

A variety of the machine-learning models applied in water purity analysis system performance were measured through some metrics like test accuracy, precision, recall, and F1-score. The results of the evaluation are listed in Table 1 below.

Table 1

Model	Accuracy	Precision	Recall	F1-Score
Meta Model (Logistic regression)	93.11	0.93	0.93	0.93
Random Forest	93.03	0.93	0.93	0.93
SVM	92.85	0.94	0.93	0.93
Decision Tree	91.32	0.91	0.91	0.91

Table 1 offers a comparative evaluation of the performance metrics for various machine learning models. The Meta Model emerges as the top performer with the highest accuracy of 93.11%, demonstrating balanced precision (0.93), recall (0.93), and F1-score (0.93). Its ability to aggregate predictions from base models ensures robust and consistent classification results.

The Random Forest model closely follows with an accuracy of 93.03%, showcasing commendable precision, recall, and F1-score (all at 0.93), emphasizing its effectiveness in capturing intricate feature interactions.

Support Vector Machine (SVM) achieves an accuracy of 92.85%, excelling in precision (0.94), which could be advantageous in tasks demanding high precision alongside reliable recall and F1-score (both at 0.93).

Finally,the Decision Tree model records an accuracy of 91.32%, with metrics—precision, recall, and F1-score—remaining consistent at 0.91. While it slightly underperforms in comparison to ensemble methods, it remains a straightforward and interpretable choice for certain use cases.

These results underscore the efficiency of machine learning models in evaluating water purity based on critical input features. The final selection of a model depends on specific criteria such as interpretability, computational resources, and the desired trade-off between precision and recall, depending on the application's priority for public health safety.

VI. REFERENCES

[1] Nirmala Malagi, "*Water portability prediction using machine learning*" , IRJMETS, e-ISSN:2582-5208, volume5(2023), pgno:2779-2782.

[2] Heming Gao,Yuru Li, Handong Lu, Shuqi Zhu, *"Water Portability Analysis and Prediction"*, AMMSAC , volume16(2022).

[3] Jatin, *"Water Quality Prediction using Machine Learning"*, October 02, 2021.

[4] Aldhyani, T.H., Al-Yaari, M., Alkahtani, H. and Maashi, M., *" Water quality prediction using artificial intelligence algorithms"*, Applied Bionics and Biomechanics, 2020.

[5] R.Kashefipour & R.Falconer (2002), *"Water quality prediction using artificial neural networks"*, Journal of Environmental Management, 65(2), 185-195.

[6] Samir patel, Khushi Shah, Sakshi Vaghela, Mohmmadali Aglodiya, Rashmi Bhattad, *"Water Portability Prediction Using Machine Learning"*, Research square, 2023.

[7] P.Y.Julien (2002), "River morphology and water quality prediction", Water Resources Research, 38(6), 1-10.

[8] M.Motagh & A.Soltanian (2017), *"Groundwater quality prediction using machine learning techniques",* Journal of Hydrology, 550, 108-118.

[9] S.K.Dey (2014), *"Water quality prediction using decision trees"*, Journal of Environmental Monitoring, 16(3), 789-798.

[10] P. Wu (2007), *"Water quality prediction using genetic algorithms",* Journal of Water Resources Planning and Management, 133(4), 345-352.

[11] Achio, Kutsanedzie, Ameko (2016), *"Comparative Analysis of Filtration Techniques for Water Purity"*.

[12] Muniz, Oliveira-Filho (2019,*"Multivariate Statistical Analysis for Water Quality Assessment"*.

[13] Ahmed et al (2020), *"Water Quality Monitoring: From Conventional to Emerging Technologies"*.

[14] K.Sreelatha, A.Nirmala Jyothsna, M.Saraswathi, P.Anusha, A. Anantha Lakshmi, *"Analysis of Water Quality"*, IJCRT, e-ISSN: 2320-2882, Volume 10 (2022).

[15] M.Anbuchezhian, R.Venkataraman, V.Kumuthavalli, *"Water Quality Analysis and Prediction using Machine Learning Algorithms"*, JETIR, e-ISSN: 2349-5162, Volume 5 (2018), pgno: 1966-1972.

[16] M.J.Pawari, S.M.Gavande, *"Assessment of Water Quality Parameters: A Review"*, IJSR, e-ISSN: 2319-7064, Volume 4 (2015), pgno: 6716-6722.

[17] *"AI for clean water: efficient water quality prediction leveraging machine learning"*, *IWA Publishing*, 2024

[18] "*Water quality prediction using machine learning methods"* Water Quality Research Journal, Vol. 53, No. 1, pp. 3-13, 2018

[19] *"Reliable water quality prediction and parametric analysis using Explainable Artificial Intelligence",* Scientific Reports, Vol. 14, 2024.

[20] *"Drinking water potability prediction using machine learning",* Water Practice & Technology, Vol. 18, No. 12, pp. 3004-3020, 2023.