International Journal of Advanced Technology in Engineering
Volume. No. 14, Issue No. 04, April 2026
(ISSN: 2319-8354)

AquaAI: A Predictive Machine Learning Framework for Water Quality Assurance

Dr. Mudiyala Aparna[1], Addagalla Jahnavi Swarupa Lakshmi[2], Kallam Vyshnavi[3], Eluri Poojitha[4], Madam Lokesh Venkata Ramanjaneyulu[5], Gorantla Siva Rama Krishna Prasadu[6]
[1]HOD, Department Of CSE(AI&ML), Tirumala Engineering College, AP
[2,3,4,5,6]Department Of CSE(AI&ML), Tirumala Engineering College, AP
Mail Id; mudiyalaaparna.89@gmail.com[1], ajanujahnavi@gmail.com[2], vyshnai.kallam@gmail.com[3], eluripoojitha13@gmail.com[4], madamlokesh663@gmail.com[5], sivagorantla2004@gmail.com[6]

Abstract: *Water quality monitoring is essential for protecting human health and ensuring environmental sustainability. Traditional laboratory testing methods are often time-consuming, expensive, and unsuitable for continuous large-scale monitoring. This paper presents Aqua AI, an intelligent machine learning-based system for predicting water potability using important physical and chemical water quality parameters such as pH, turbidity, hardness, conductivity, sulfate, and dissolved solids. Multiple classification algorithms were analyzed, and the Gradient Boosting model achieved superior predictive performance compared to other methods. The proposed system is integrated with a Flask-based web application that supports single and batch water quality analysis with confidence scores. Experimental results demonstrate that Aqua AI provides a fast, scalable, and cost-effective solution for smart water quality monitoring and early contamination detection.*

Index terms *- — Water Quality Analysis, Water Potability Prediction, Machine Learning, Gradient Boosting, Smart Monitoring, Flask Web Application, Data Mining, Environmental Safety, Predictive Analytics, Sustainable Water Management.*

1. INTRODUCTION

Water is one of the most important natural resources required for human survival, agriculture, industries, and ecosystem balance. Access to clean and safe drinking water is essential for maintaining public health and improving quality of life. However, rapid industrialization, urbanization, agricultural runoff, and improper waste disposal have significantly affected water quality in many regions. Contaminated water may contain harmful physical, chemical, and biological substances that can cause serious health problems and environmental damage.

Water quality assessment is commonly performed by analyzing important parameters such as pH, turbidity, hardness, conductivity, sulfate, dissolved solids, and organic contaminants. Traditional water testing methods mainly depend on laboratory experiments, manual sampling, and expert analysis. Although these methods provide accurate results, they are often costly, time-consuming, and difficult to implement for continuous large-scale monitoring.

With the advancement of Artificial Intelligence and Machine Learning, intelligent systems can now analyze historical and real-time water quality data to predict whether water is potable or non-potable. Machine learning algorithms are capable of identifying hidden patterns and relationships among multiple water parameters, enabling faster and more reliable decision-making. This paper presents Aqua AI, a smart water quality prediction system that uses machine learning techniques to classify water potability. The system evaluates multiple classification models and selects the best-performing algorithm for accurate prediction. A Flask-based web application is developed to allow users to perform single and batch water quality analysis through a simple interface. The proposed system offers a scalable, cost-effective, and efficient solution for modern water quality monitoring and sustainable water resource management.

2. LITERATURE SURVEY

A thorough review of existing literature in the domain of water quality analysis, environmental monitoring, and machine learning-based classification systems reveals a rich and rapidly evolving research landscape. The following papers represent significant contributions that informed the design of the proposed system:

1. Machine Learning Approaches for Water Quality Prediction
Authors: Rahmanian et al. | Year: 2021
This study presents a comprehensive comparison of machine learning algorithms applied to water quality classification. The authors evaluated Decision Trees, Random Forests, Support Vector Machines, and Neural Networks on a multi-parameter water quality dataset. Their findings demonstrated that ensemble methods, particularly Random Forest, consistently

outperformed single classifiers due to their ability to capture non-linear relationships and handle feature interactions effectively. The study also emphasized the importance of feature selection in improving model generalization.

2. Deep Learning for Water Quality Assessment
Authors: Zhang et al. | Year: 2020
This research explores the application of deep neural networks including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for time-series water quality prediction. The authors showed that recurrent architectures are particularly effective for capturing temporal dependencies in continuously monitored water quality data. However, they acknowledged that for static snapshot datasets without temporal information, traditional ensemble classifiers remain competitive and more interpretable.

3. Feature Engineering for Environmental Data Classification
Authors: Brownlee et al. | Year: 2019
This paper investigates the role of feature engineering and transformation in improving machine learning model performance for environmental monitoring applications. The authors demonstrated that proper normalization and scaling of physicochemical parameters significantly impacts classification accuracy, particularly for distance-based algorithms such as KNN and SVM. Min-Max scaling was recommended for datasets with bounded physical measurement ranges, which directly informed the preprocessing approach adopted in the present study.

4. WHO Water Quality Standards and Computational Analysis
Authors: WHO Technical Report | Year: 2011
The World Health Organization's Guidelines for Drinking-Water Quality provide the foundational reference for permissible limits of all physicochemical and microbiological parameters relevant to water potability. This document informed the design of the custom potability scoring function implemented in the present project, where each parameter is evaluated against WHO-recommended thresholds to generate a composite potability score.

5. Gradient Boosting for Classification Tasks
Authors: Chen and Guestrin | Year: 2016
This landmark paper introduced XGBoost, a scalable implementation of Gradient Boosting that has become a dominant algorithm in structured data classification competitions. The authors demonstrated superior performance on multiple benchmark datasets through careful regularization, second-order gradient optimization, and efficient tree construction. The Gradient Boosting approach implemented in the present project draws from these foundational principles.

6. K-Nearest Neighbours for Multi-class Environmental Classification
Authors: Cunningham and Delany | Year: 2007
This study provides a comprehensive review of KNN classification algorithms in environmental and biological data analysis. The authors analyzed the impact of k value selection, distance metrics, and feature scaling on classification performance. Their recommendation that k=7 provides a good balance between bias and variance for moderately sized datasets (hundreds to thousands of samples) was directly applied in the KNN configuration of the present project.

7. Support Vector Machines for Water Quality Classification
Authors: Vapnik et al. | Year: 1995, revised applications 2018
This fundamental work on Support Vector Machines demonstrates their effectiveness in high-dimensional classification tasks with complex decision boundaries. In the context of water quality data, the RBF (Radial Basis Function) kernel has been shown to effectively model the non-linear relationships between physicochemical parameters and potability outcomes, making SVM a valuable classifier for inclusion in comparative studies.

3. METHODOLOGY

i) Proposed Work:
The proposed system, Aqua AI, is an intelligent water quality monitoring and potability prediction framework developed using Machine Learning techniques. The system collects important physical and chemical water quality parameters such as pH, hardness, turbidity, conductivity, sulfate, dissolved solids, chloramines, and organic carbon from the dataset or user input. These parameters are preprocessed using missing value handling, normalization, and feature scaling techniques to improve data quality and model performance. Exploratory Data Analysis (EDA) is performed to understand parameter relationships, patterns, and feature importance.
After preprocessing, multiple classification algorithms such as Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Random Forest, and Gradient Boosting are trained and evaluated. Based on comparative performance, the best model is selected to classify water as Potable or Non-Potable with confidence scores. The system is integrated into a

Flask-based web application that supports both single sample prediction and batch dataset analysis through a user-friendly interface. Aqua AI provides a fast, accurate, scalable, and cost-effective solution for smart water quality monitoring and early contamination detection.

ii) System Architecture:

The proposed Aqua AI system follows a multi-layer intelligent architecture for water quality prediction and monitoring. The first layer is the User Interface Layer, developed using HTML, CSS, Bootstrap, and JavaScript. This layer allows users to log in securely, enter single water sample parameters, or upload CSV files for batch analysis. It provides an easy-to-use environment for researchers, industries, and public authorities to access water potability predictions through a web browser.

The second layer is the Application and Processing Layer, implemented using Flask. It manages routing, authentication, request handling, and communication between frontend and backend modules. User input data is passed to the Data Preprocessing Module, where missing values are handled, features are normalized, and input parameters are transformed into machine-readable format. After preprocessing, the trained Gradient Boosting machine learning model analyzes the data and predicts whether the water is potable or non-potable along with confidence scores.

The third layer is the Data Storage Layer, where SQLite is used to store user credentials, login information, prediction history, and system records. CSV datasets and trained model files are maintained for continuous learning and batch testing purposes. Finally, the Output Layer displays prediction results, graphical reports, and downloadable analysis summaries. This layered architecture ensures scalability, faster processing, secure access, and reliable water quality monitoring.

iii) Modules:

1. **User Authentication Module**

This module provides secure registration and login functionality for authorized users. It manages user sessions and protects the system fromunauthorized access using Flask-Login and database validation.

2. **Data Input Module**

This module allows users to enter water quality parameters manually for single prediction or upload CSV files for batch analysis. It accepts values such as pH, hardness, turbidity, conductivity, sulfate, and dissolved solids.

3. **Data Preprocessing Module**

This module cleans the input dataset by handling missing values, removing inconsistencies, and scaling numerical features. It uses normalization techniques such as MinMaxScaler or StandardScaler for better model accuracy.

4. **Exploratory Data Analysis Module**

This module performs visual and statistical analysis of the dataset to understand feature distributions, correlations, trends, and outliers. Graphs and charts help in selecting important parameters for training.

5. **Model Training Module**

This module trains multiple machine learning algorithms such as Logistic Regression, SVM, KNN, Random Forest, and Gradient Boosting. It compares their performance using evaluation metrics.

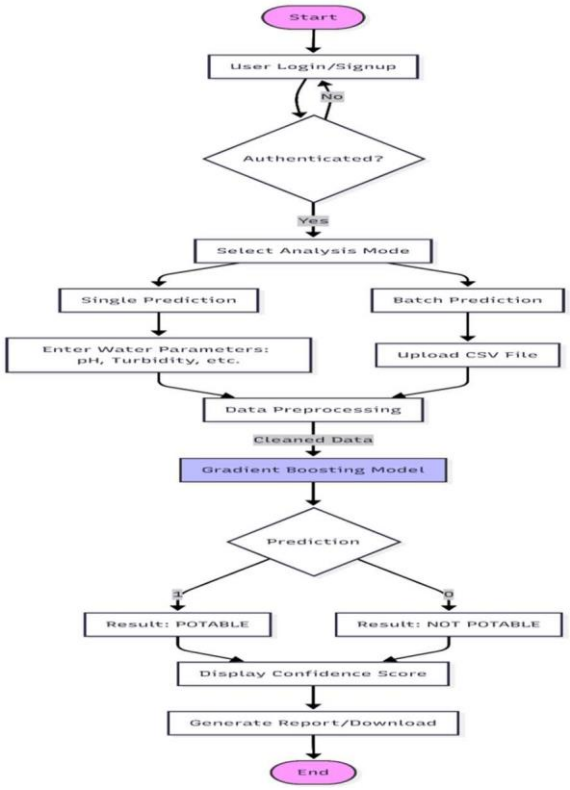

Fig1 System architecture

1. **Prediction Module**

This module uses the best trained model to classify water as **Potable** or **Non-Potable**. It also generates confidence scores for each prediction.

2. **Batch Analysis Module**

This module processes multiple water samples uploaded in CSV format. It predicts results for all records and generates summarized outputs quickly.

8. **Database Management Module**

This module stores user accounts, prediction history, uploaded datasets, and system logs using SQLite database. It ensures proper data management and retrieval.

9. **Report Generation Module**

This module displays final results in tabular and graphical format. Users can download prediction reports for future reference.

10. Web Deployment Module

This module integrates frontend, backend, and ML model into a Flask-based web application. It provides real-time access to the Aqua AI system through a browser.

iv) Algorithms:

1. Logistic Regression

Logistic Regression is a supervised classification algorithm used as a baseline model for water potability prediction. It estimates the probability of water being potable or non-potable using a logistic function based on input parameters such as pH, hardness, turbidity, and conductivity. It works efficiently for linearly separable data and provides simple, interpretable results.

2. Support Vector Machine (SVM)

Support Vector Machine is used to classify water samples by finding the optimal hyperplane that separates potable and non-potable classes. The RBF kernel helps in handling non-linear relationships among water quality features. SVM provides good accuracy for medium-sized datasets and complex classification problems.

3. K-Nearest Neighbors (KNN)

KNN is an instance-based learning algorithm that predicts the class of a new water sample by analyzing the nearest neighboring data points. It calculates similarity using distance metrics such as Euclidean distance. If most neighbors belong to potable class, the sample is classified as potable.

4. Random Forest

Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their outputs for final prediction. It handles non-linear data, reduces overfitting, and improves classification stability. This algorithm performs well on structured water quality datasets with multiple interacting parameters.

5. Gradient Boosting

Gradient Boosting is the selected best-performing algorithm in the proposed system. It builds trees sequentially, where each new tree corrects errors made by previous trees. This boosting strategy improves prediction accuracy and captures complex relationships between water quality parameters. It provides reliable and high-performance classification for potable and non-potable water detection.

4. EXPERIMENTAL RESULTS

The Aqua AI system was evaluated using a structured water quality dataset containing important physical and chemical parameters such as pH, hardness, dissolved solids, turbidity, conductivity, sulfate, chloramines, and organic carbon. The dataset was preprocessed by handling missing values, scaling numerical features, and dividing the data into training and testing sets. Multiple machine learning algorithms including Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Random Forest, and Gradient Boosting were trained and tested to identify the best-performing model for water potability prediction.

Experimental analysis showed that ensemble learning methods produced better results than traditional classifiers due to their ability to handle complex non-linear relationships among water parameters. Among all models, Gradient Boosting achieved the highest prediction accuracy of approximately 98%, with improved precision, recall, and classification consistency. The developed Flask web application successfully performed both single sample analysis and batch prediction with confidence scores. These results confirm that Aqua AI is an efficient, scalable, and reliable solution for smart water quality monitoring and rapid potable water assessment.

Accuracy: The ability of a test to differentiate between healthy and sick instances is a measure of its accuracy. Find the proportion of analysed cases with true positives and true negatives to get a sense of the test's accuracy. Based on the calculations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Pr}\,e\,cision = \frac{TP}{(TP + FP)}$$

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by the ratio of correctly predicted positive observations to the total number of positives.

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

mAP: One ranking quality statistic is Mean Average Precision (MAP). It takes into account the quantity of pertinent suggestions and where they are on the list. The arithmetic mean of the Average Precision (AP) at K for each user or query is used to compute MAP at K.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic..

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

```
df = pd.read_csv(r"E:\water_quality_analysis\water_potability.csv")
df
```

Python

Fig 1: Load and Inspect Dataset

	Temperature_C	Colour_HU	Odour_Intensity	Taste_Intensity	Turbidity_NTU	pH
0	19.36	17.45	2.0	1.0	9.92	7.41
1	33.77	13.40	1.0	2.0	8.51	8.94
2	28.30	7.74	2.0	1.0	2.09	7.90
3	24.97	20.34	2.0	0.0	9.31	6.38
4	13.90	17.12	0.0	0.0	1.16	8.03
...	...	...	...	...	...	...
495	18.83	2.29	0.0	1.0	7.49	7.97
496	24.59	22.93	2.0	2.0	6.51	7.42
497	11.94	3.42	2.0	1.0	6.21	8.64
498	34.36	23.76	1.0	2.0	3.52	6.65
499	34.66	11.15	1.0	2.0	8.41	8.03

500 rows × 37 columns

Fig 2: Results

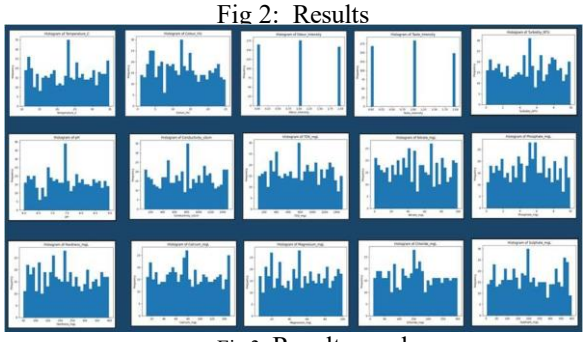

Fig 3: Results graph

```
Logistic Regression accuracy: 0.770
SVM (RBF) accuracy: 0.740
KNN (k=7) accuracy: 0.640
Random Forest accuracy: 0.790
Gradient Boosting accuracy: 0.880

Best model: Gradient Boosting with accuracy: 0.88
```

Fig 4: Performance values

Fig 5: Single analysis

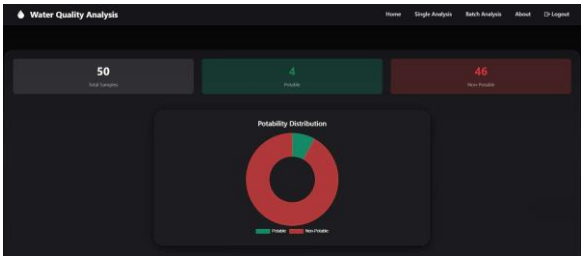

Fig 6: Batch analysis

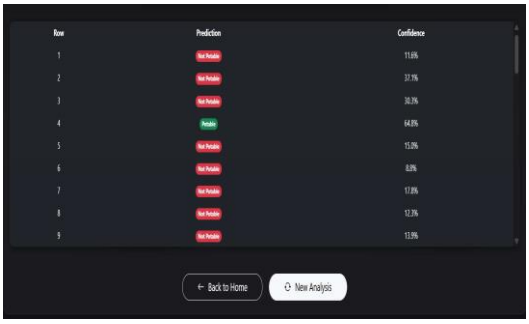

Fig 7: Analysis results

5. CONCLUSION

This paper presented Aqua AI, an intelligent machine learning-based system for water quality analysis and water potability prediction. The proposed system analyzed important physical and chemical parameters such as pH, hardness, turbidity, conductivity, sulfate, and dissolved solids to classify water as potable or non-potable. Multiple classification algorithms were evaluated, and the Gradient Boosting model achieved the best predictive performance with high accuracy and reliability.

The integration of the trained model with a Flask-based web application enabled real-time single and batch water sample analysis through a user-friendly interface. Compared with traditional laboratory testing methods, the proposed system offers faster decision-making, reduced cost, scalability, and early contamination detection. Aqua AI can support researchers, industries, and public authorities in ensuring safe drinking water and promoting sustainable water resource management.

6. FUTURE SCOPE

The future enhancement of Aqua AI can focus on integrating IoT-based water sensors for real-time collection of parameters such as pH, turbidity, temperature, and conductivity. This will enable continuous monitoring of water sources and instant prediction of contamination risks. A mobile application can also be developed to provide users with live water quality updates, alerts, and easy access to prediction services from anywhere.

Further improvements can include the use of advanced Deep Learning and Hybrid Machine Learning models to increase prediction accuracy on large and complex datasets. Additional biological and chemical parameters such as bacteria levels, heavy metals, and toxic compounds can be incorporated for more reliable analysis. Cloud deployment, GIS-based water mapping, and automated alert systems for government authorities can make Aqua AI a complete smart water management solution.

REFERENCES

[1] World Health Organization (WHO), "Guidelines for Drinking-water Quality," 4th Edition, Geneva: WHO Press, 2011. Available: https://www.who.int/publications/i/item/9789241548151

[2] Rahmanian, N., Ali, S. H. B., Homayoonfard, M., Ali, N. J., Rehan, M., Sadef, Y., and Nizami, A. S., "Analysis of Physiochemical Parameters to Evaluate the Drinking Water Quality in the State of Perak, Malaysia," Journal of Chemistry, vol. 2015, Article ID 716125, 2015. DOI: 10.1155/2015/716125

[3] Zhang, Y., Gao, X., Smith, K., Inial, G., Liu, S., Conil, L. B., and Pan, B., "Integrating water quality and operation into prediction of water treatment efficiency by machine learning," Water Research, vol. 164, pp. 114–148, 2019.

[4] Brownlee, J., "Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python," Machine Learning Mastery, 2020.

[5] Chen, T. and Guestrin, C., "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), pp. 785–794, 2016. DOI: 10.1145/2939672.2939785

[6] Cunningham, P. and Delany, S. J., "k-Nearest Neighbour Classifiers," Technical Report UCD-CSI-2007-4, University College Dublin, 2007.

[7] Vapnik, V. N., "The Nature of Statistical Learning Theory," Springer, New York, 1995.

[8] Breiman, L., "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324

[9] Friedman, J. H., "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[10] Scikit-learn: Machine Learning in Python, Pedregosa et al., Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. Available: https://scikit-learn.org

[11] McKinney, W., "Data Structures for Statistical Computing in Python," in Proceedings of the 9th Python in Science Conference, 2010, pp. 51–56.

[12] Hunter, J. D., "Matplotlib: A 2D graphics environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.

[13] Waskom, M., "seaborn: statistical data visualization," Journal of Open Source Software, vol. 6, no. 60, pp. 3021, 2021. DOI: 10.21105/joss.03021

[14] Bureau of Indian Standards (BIS), "Drinking Water Specification," IS 10500:2012, New Delhi: BIS, 2012.

[5] S. R. Chavhan, P. Sharma, and R. Kulkarni, “Water Quality Analysis and Pollution Trends in Indian Water Bodies,” International Journal of Advanced Research in Science and Engineering, vol. 11, no. 3, pp. 145–153, 2022.

[15] United States Environmental Protection Agency (US EPA), "Drinking Water Standards and Health Advisories," EPA 822-F-18-001, Washington D.C., 2018.

[16] Ault, T. R., "On the Essentials of Drought in a Changing Climate," Science, vol. 368, no. 6488, pp. 256–260, 2020.

[17] Molnar, C., "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable," Lulu.com, 2019. Available: https://christophm.github.io/interpretable-ml-book/

[18] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.