IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, March 2025
Impact Factor: 7.67

AquaPura

Dr Brinthakumari, Aniket Patil, Vipul Bailkar, Aditya Patil
Department of Computer Engineering
New Horizon Institute of Technology and Management, Thane, India
brinthakumaris@nhitm.ac.in, aniketpatil212@nhitm.ac.in
vipulbailkar212@nhitm.ac.in, adityapatil212@nhitm.ac.in

Abstract: *Water is the most crucial resource of life and it is necessary for the survival of all living creatures including human beings. The survival of business and agriculture depends on freshwater. An essential step in managing freshwater assets is the evaluation of the quality of the water. Before using water for anything, including drinking, chemical spraying (pesticides, etc.), or animal hydration, it is crucial to assess its purity. The ecosystem and the general public's health are directly impacted by water quality. Therefore, analyzing and predicting water quality is necessary for both environmental and human protection. Machine learning can be used to analyze and predict the water quality based on the parameters like PH value, turbidity, hardness, conductivity, dissolved solids in water and other parameters. In this work, the water quality is predicted by giving the concentration of various parameters as input to machine learning algorithms and the water is classified as safe or unsafe for the usage of domestic purposes.*

Keywords: Conductivity, Hardness, Machine Learning, PH, Turbidity, Water Quality

I. INTRODUCTION

Water is an essential resource for all living beings, playing a vital role in daily activities like drinking, cleaning, and agriculture. Access to clean water is a fundamental human right, crucial for maintaining public health. Contaminated water can lead to severe diseases such as cholera, typhoid, dysentery, and hepatitis. The degradation of water bodies affects ecosystems and signals environmental deterioration. Effective management of water resources requires monitoring water quality to ensure safety and sustainability. A well-maintained hydrological system supports a healthy ecosystem and produces high-quality water. Predicting and analyzing water quality is essential to control pollution and safeguard human and environmental health.

Machine learning (ML) offers a powerful approach to analyzing and predicting water quality by processing large datasets efficiently. By utilizing ML algorithms such as k-nearest neighbors (KNN), support vector machine (SVM), decision tree (DT), and random forest (RF), researchers can determine water potability based on key parameters like pH, conductivity, hardness, chloramines, and turbidity. These models assess whether input values fall within safe limits and predict if the water is suitable for consumption. This advanced technology aids in environmental protection, ensures water safety, and helps researchers understand human impacts on ecosystems.

II. LITERATURE REVIEW

[1] Perumal, Bhagavathi, et al. "Water Quality Prediction Based on Hybrid Deep Learning Algorithm." *Advances in Civil Engineering*, vol. 2023, 2023, pp. 1–9. Wiley Online Library

This study integrates LSTM networks with Grey Wolf Optimization (GWO) and Feature Selection Optimization (FSO) to enhance water quality predictions. Using data from the Thamirabarani river basin, the model predicts key water quality parameters such as Dissolved Oxygen (DO), Chemical Oxygen Demand (COD), and ammonia nitrogen (NH_3-N) with high accuracy.

[2] Mahesh, N., et al. "Water Quality Prediction Using LSTM with Combined Normalizer for Efficient Water Management." *Desalination and Water Treatment*, vol. 317, 2024, p. 100183. ResearchGate

ISSN 2581-9429 IJARSCT
DOI: 10.48175/IJARSCT-24523

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, March 2025
Impact Factor: 7.67

This paper presents an LSTM-CN model that combines z-score, interval, and max normalization techniques for better preprocessing of multi-factor water quality data. The model achieves 99.3% accuracy and is optimized to maintain computational efficiency while preserving data characteristics, making it ideal for smart water management applications.

[3] Paneru, Biplov, and BishwashPaneru. "WaterQualityNeT: Prediction of Seasonal Water Quality of Nepal Using Hybrid Deep Learning Models." *arXiv preprint*, arXiv:2409.10898, 17 Sept. 2024. arXiv

This study proposes a hybrid deep learning model combining CNN and LSTM for seasonal water quality prediction in Nepal. The model processes historical datasets and meteorological parameters to forecast water pollution trends, contributing to real-time environmental monitoring and decision-making.

[4] Farhat, Manel, Ali Khalfallah, and Med Salim Bouhlel. "A Real-Time Court Detection and Tracking System to Tennis Videos." *Journal of Testing and Evaluation*, vol. 47, no. 4, 2019, pp. 1–10. ASTM Internationa

This paper develops a real-time court detection system that tracks players and court boundaries using computer vision techniques. The approach is designed to improve player tracking and shot analysis, making it useful for automated match analytics and sports performance evaluation.

III. METHODOLOGY

Figure 3.1 illustrates the block diagram that explains the various steps used in this study for water potability classification. The ML processes that are used in this study are presented in this figure.

Data Processing and Feature Analysis

The first step in the water quality prediction process involves data exploration and feature analysis. This includes examining the dataset to understand patterns, distributions, and relationships between different water quality parameters, such as pH, turbidity, conductivity, and total dissolved solids. Feature analysis helps in selecting the most relevant parameters for prediction, ensuring that only the most impactful variables contribute to the model. Following this, outlier detection and handling is performed to identify and correct anomalous data points that could distort the model's performance.

Outliers may arise due to errors in data collection, measurement issues, or extreme variations in environmental conditions. Handling these outliers ensures that the data remains clean and reliable for training the machine learning model. Data preprocessing and imputation follow, where missing values are addressed using statistical methods or predictive techniques to maintain data integrity and improve the model's predictive power.

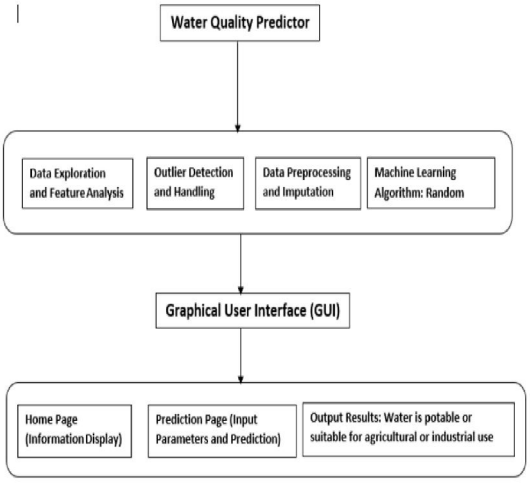

Fig. 3.1 Block Diagram Showing ML Steps Used

ISSN 2581-9429 IJARSCT
DOI: 10.48175/IJARSCT-24523

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, March 2025
Impact Factor: 7.67

Machine Learning Model Implementation

After data preprocessing, a machine learning algorithm is applied to develop a predictive model. The selected algorithm in this case is Random Forest, which is a robust ensemble learning method that operates by constructing multiple decision trees. Random Forest provides high accuracy, handles non-linearity well, and is less prone to overfitting. The model is trained on historical water quality data, learning patterns and relationships to make predictions regarding water potability or suitability for industrial or agricultural use.

Graphical User Interface (GUI) Development

Once the model is developed, a Graphical User Interface (GUI) is designed to facilitate user interaction. The GUI serves as a bridge between the user and the prediction model, allowing for seamless input of water quality parameters and visualization of results. It consists of three main sections:

Home Page (Information Display): This section provides an overview of the system, explaining its functionality and usage. Users can understand how the model works, what parameters are considered, and how to interpret the results.

Prediction Page (Input Parameters and Prediction): Here, users can enter specific water quality parameters, such as pH, conductivity, and turbidity, which serve as inputs to the machine learning model. Once the parameters are entered, the model processes the data and generates a prediction.

Output Results: The final section displays the prediction results, determining whether the water is potable, suitable for industrial use, or fit for agricultural purposes. The results guide users in decision-making regarding water consumption and treatment.

Algorithm Used Random Forest (RF)

The Random Forest algorithm is an ensemble machine learning technique that operates by constructing multiple decision trees during training and merging their outputs for improved accuracy and robustness. It is widely used for both classification and regression tasks. The algorithm works by randomly selecting subsets of data and features to build individual decision trees, reducing overfitting and improving generalization. Each tree in the forest votes for the final prediction, and the majority decision is taken in classification, while the average is used in regression.

Random Forest is resistant to noise and missing data, making it highly reliable for real-world applications like water quality prediction. It efficiently handles large datasets with high-dimensional features and can identify feature importance, aiding in better decision-making. The algorithm is computationally more expensive than single decision trees but provides superior performance. Additionally, it is less sensitive to outliers and does not require extensive parameter tuning, making it a versatile and powerful machine learning model.

IV. PARAMETERS

The dataset "aquapura.csv" containing water quality metrics for 16776 various water bodies has been considered from **Central Ground Water Board (CGWB)** for this study. The ideal values for the water quality parameters are as shown in Table 4.1

TABLE 4.1 VALUES OF WATER QUALITY METRICES FOR SAFE WATER

PARAMETER	RANGE
pH	6.5 - 8.5 (Optimal for drinking water)
EC (µS/cm)	<400 microS/cm (Electrical conductivity, indicates salinity)
CO_3 (mg/L)	Varies, contributes to alkalinity
HCO_3 (mg/L)	Helps maintain pH stability in water
Cl (mg/L)	<250 mg/L (High levels affect taste and corrosion)
F (mg/L)	0.5 - 1.5 mg/L (Excess causes fluorosis)
SO_4 (mg/L)	3 – 30 mg/L (High levels cause laxative effects)
NO_3 (mg/L)	<10 mg/L (Excess leads to health risks, especially for infants)
PO_4 (mg/L)	<0.1 mg/L (Excess leads to algal blooms)
Total Hardness	60 - 120 mg/L (Indicates mineral content in water)
Ca (mg/L)	20 - 100 mg/L (Essential for health, causes scaling)
Mg (mg/L)	10 - 50 mg/L (High levels cause hardness)
Na (mg/L)	<200 mg/L (Affects taste and salinity)
K (mg/L)	1 - 10 mg/L (Essential for metabolism)
SiO2	5 - 50 mg/L (Essential for diatom growth)

ISSN 2581-9429 IJARSCT
DOI: 10.48175/IJARSCT-24523

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, March 2025
Impact Factor: 7.67

- **pH** – Indicates the acidity or alkalinity of water; an optimal range is 6.5-8.5 for drinking water.
- **EC (Electrical Conductivity)** – Measures the ability of water to conduct electricity, which depends on dissolved ions.
- **CO_3 (Carbonate, mg/L)** – Represents carbonate ion concentration, affecting water alkalinity and buffering capacity.
- **HCO_3 (Bicarbonate, mg/L)** – Helps regulate pH and alkalinity in water, influencing its buffering ability.
- **Cl (Chloride, mg/L)** – High chloride levels can indicate contamination from sewage or industrial waste.
- **F (Fluoride, mg/L)** – Essential in small amounts for dental health but harmful in excess.
- **SO_4 (Sulfate, mg/L)** – Naturally occurring; high levels may cause a laxative effect and affect water taste.
- **NO_3 (Nitrate, mg/L)** – High levels can indicate agricultural runoff and pose health risks, especially for infants.
- **PO_4 (Phosphate, mg/L)** – Supports aquatic life but excessive levels can lead to algal blooms.
- **Total Hardness** – Sum of calcium and magnesium; affects water scaling properties in pipes and appliances.
- **Ca (Calcium, mg/L)** – Essential for human health but contributes to water hardness.
- **Mg (Magnesium, mg/L)** – Another contributor to water hardness; necessary for human metabolism.
- **Na (Sodium, mg/L)** – Affects water taste and is a key factor in salinity.
- **K (Potassium, mg/L)** – Essential nutrient but present in low concentrations in natural water sources.
- **SiO2 (Silicon Dioxide) -** Essential for diatom growth, biomineralization, and influences water quality by affecting scaling in industrial systems.

V. RESULTS

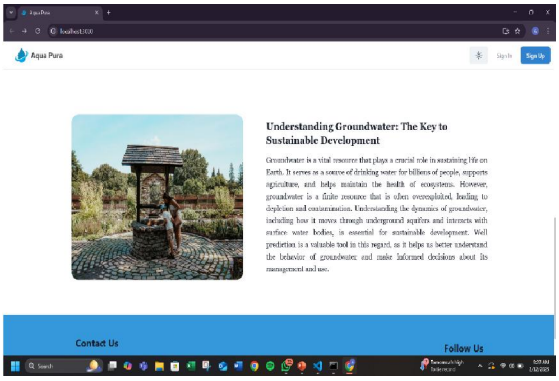

Fig 5.1 Login Page

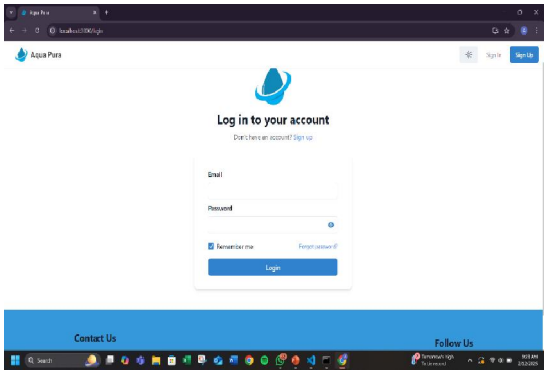

Fig 5.2 Login Panel

DOI: 10.48175/IJARSCT-24523

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, March 2025
Impact Factor: 7.67

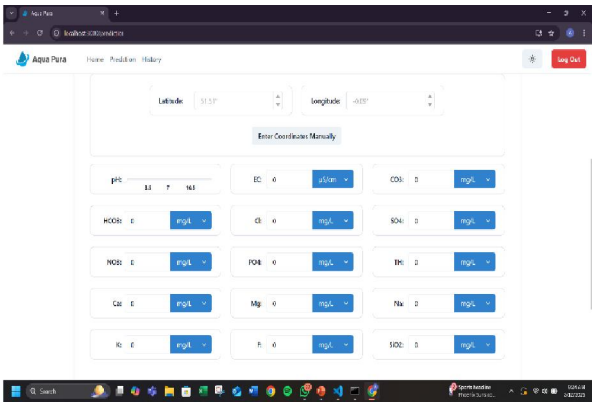

Fig 5.3 Parameters Page

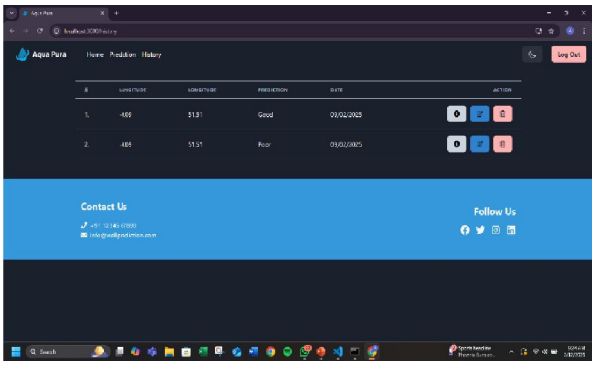

Fig 5.4 History Page

VI. CONCLUSION

This paper presents a comparative study of ML models used in classifying water as potable or not. This study provides an overview of the four ML approaches applied to the water potability dataset and an evaluation of their effectiveness. Different models each displayed specific advantages and disadvantages and no particular model was considered completely superior over the other. Nevertheless, the RF model demonstrated promising results prior to oversampling i.e., on imbalanced dataset, and the DT demonstrated promising results following oversampling.

VII. FUTURE SCOPE

The future scope of Water Quality Prediction using the Random Forest algorithm is vast, with potential enhancements in accuracy, real-time monitoring, and scalability. Hybrid models combining Random Forest with Deep Learning can improve prediction precision. IoT sensors can enable real-time data collection, while edge computing can facilitate on-site analysis. Geographic expansion with GIS integration will help assess water quality across regions. Cloud deployment can enhance accessibility, and AI-driven recommendations can suggest water treatment solutions. Collaborations with government agencies can aid in pollution control and early warning systems. User-friendly web and mobile applications can provide real-time alerts to the public. Climate change impact analysis can help predict contamination trends. Automated corrective measures can improve water purification processes. Overall, integrating advanced AI and real-time analytics will revolutionize water quality monitoring

ISSN 2581-9429 IJARSCT
DOI: 10.48175/IJARSCT-24523

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, March 2025
Impact Factor: 7.67

REFERENCES

[1] Mosavi, F. S. Hosseini, B. Choubin, M. Goodarzi and A. A. Dineva, "Groundwater Salinity Susceptibility Mapping Using Classifier Ensemble and Bayesian Machine Learning Models," in IEEE Access, 145564-145576, vol.8, pp. 10.1109/ACCESS.2020.3014908.2020,doi:

[2] O. Al-Sulttani, M. Al-Mukhtar, A. B. Roomi, A. A. Farooque, K. M. Khedher and Z. M. Yaseen, "Proposition of New Ensemble Data- Intelligence Models for Surface Water Quality Prediction," in IEEE Access, vol. 9, pp. 108527-108541, 2021, doi:10.1109/ACCESS.2021.3100490.

[3] Hongfang Lu, Xin Ma, "Hybrid decision tree-based machine learning models for short-term water quality prediction," Chemosphere, 249, 2020, 126169, ISSN 0045-6535, https://doi.org/10.1016/j.chemosphere.2020.126169. Volume

[4] Hadi Mohammed, Hoese Michel Tornyeviadzi, Razak Seidu, "Emulating process-based water quality modelling in water source reservoirs using machine learning," Journal of Hydrology, Volume 609, 2022, https://doi.org/10.1016/j.jhydrol.2022.127675.

[5] Botao Chen, Xi Mu, Peng Chen, Biao Wang, Jaewan Choi, Honglyun Park, Sheng Xu, Yanlan Wu, Hui Yang, "Machine leaming-based inversion of water quality parameters in typical reach of the urban river by UAV multispectral data," Ecological Indicators, Volume 133, 2021, https://doi.org/10.1016/j.ecolind.2021.108434.

[6] Hye Won Lee, Min Kim, Hee Won Son, Bachyun Min, Jung Hyun Choi, "Machine-learning-based water quality management of river with serial impoundments in the Republic of Korea," Journal of Hydrology: Regional Studies, Volume 41, 2022, https://doi.org/10.1016/j.ejrh.2022.101069.

[7] Mashud Rana, Ashfaqur Rahman, Joel Dabrowski, Stuart Arnold, John McCulloch, Bruno Pais, "Machine learning approach to investigate the influence of water quality on aquatic livestock in freshwater ponds, "Biosystems Engineering, Volume 208, 2021, https://doi.org/10.1016/j.biosystemseng.2021.05.017.

[8] Ali El Bilali, Abdeslam Taleb, "Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment," Journal of the Saudi Society of Agricultural Sciences, Volume 19, Issue7, 2020, https://doi.org/10.1016/j.jssas.2020.08.001.

[9] Luka Grbčić, Siniša Družeta, Goran Mauša, Tomislav Lipić, Darija Vukić Lušić, Marta Alvir, Ivana Lučin, Ante Sikirica, Davor Davidović, Vanja Travaš, Daniela Kalafatovic, Kristina Pikelj, Hana Fajković, Toni Holjević, Lado Kranjčević, "Coastal water quality prediction based on machine learning with featureinterpretationandspatio-temporal analysis, "Environmental Modelling & Software, Volume 155, 2022, https://doi.org/10.1016/j.envsoft.2022.105458.

[10] Kangyang Chen, Hexia Chen Chuanlong Zhou, Yichao Huang, Xiangyang Qi, Ruqin Shen, Fengrui Liu, Min Zuo, Xinyi Zou, Jinfeng Wang, Yan Zhang, Da Chen, Xingguo Chen, Yongfeng Deng,Hongqiang Ren, "Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data," Water Research, Volume2020,https://doi.org/10.1016/j.watres.2019.115454.

ISSN 2581-9429 IJARSCT
DOI: 10.48175/IJARSCT-24523