iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84382

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Machine Learning Approach for Water Purity Detection

Y. Hari[1], Tulasi Miriyala[2]
[1]*MCA Student, Department of MCA, AQJ Centre for PG Studies, Gudilova, Anandapuram, Visakhapatnam, Andhra Pradesh-531173*
[2]*Assistant Professor, Department of MCA, AQJ Centre for PG Studies, Gudilova, Anandapuram, Visakhapatnam, Andhra Pradesh-531173*

Abstract: ***Humans rely on resources from the earth, which are valued for their quality. Climate change and its effects on the environment have long been monitored. The product's quality is carefully assessed before to usage. Water, an unavoidable resource, has become increasingly important in terms of quality monitoring due to the influence of different external elements such as industrial effluents, acid rain, and so on. This research presents a system for monitoring water quality that employs statistical quality control techniques and machine learning algorithms to improve the accuracy of classification. Here we are using gradient boosting, Random forest, XG boost, naïve bayes, Ada boost, cat boost algorithms the categorization is concerned with determining whether the water is appropriate for home use or not.***
Keywords: Gradient boosting, Random Forest, XGBoost, Naïve Bayes, Ada boost, cat boost algorithms.

I. INTRODUCTION

Potable water quality forecast is critical to ensuring safety for everyone. It is a vital concern for an individual to live a healthy life. Contaminated water consumption may lead to many different types of illnesses, depending on the data, about 3,575,000 individuals suffer annually from water-based ailments. Determining usable water will be tough in places with few potable water resources. Pharmaceutical dirt was the main contributor of contamination of water throughout the era of industrialization. There are numerous approaches for anticipating water that is safe to consume. Neural networks, gray theory, mathematical modeling, and the concept of chaos are the most applicable methodologies. The statistical procedure outperforms other methods for building ideal models. A neural network outperforms other models in forecasting and research.

The quality of water for drinking is primarily determined by fundamental parameters like pH, hardness, sulfate, organic carbon, turbidity, and a few more. The machine learning algorithms produce strong outcomes for forecasting for the water's purity forecasting. The most widely employed machine learning algorithms for forecasting are artificial neural networks (ANN), convolutional neural networks (CNN), deep neural networks (DNN), random forests (RF), and support vector machines (SVMs).

Detecting the purity of water for consumption is an important step toward preserving the well-being of the public. Polluted water can endanger human health, resulting in waterborne infections and other difficulties. Traditional techniques of executing water quality testing are time-consuming, expensive, and might not offer real-time data. However, ml approaches have grown up as effective methods for addressing this problem by executing the recognition and categorization of water quality data. The application of machine learning techniques facilitates the creation of precise and effective models capable of analyzing vast amounts of water-related data and providing immediate advice.

II. LITERATURE REVIEW

A literature survey on drinking water quality encompasses a broad review of existing research, studies, and reports on various aspects of drinking water, including its sources, contamination, treatment processes, and health impacts.

[1] "Drinking Water Quality Assessment and Prediction Using Machine Learning Techniques" by Li et al. (2018).

The article investigates the use of different machine learning methods, such as random forest, support vector machines, and artificial neural networks, to forecast the purity of water for drinking. The authors describe the significance of feature selection and conduct a thorough review of the efficiency of several approaches.

[2] "Drinking Water Quality Assessment Based on Machine Learning Algorithms" by Zhu et al. (2019).
The research compares the methods of machine learning employed for evaluating the purity of drinking water, such as decision trees, K-nearest neighbors, and logistic regression. The writers explore how preliminary data analysis and choice of features strategies can increase precision for forecasting.

[3] "A Review on Drinking Water Quality Assessment Using Machine Learning Techniques" by Jain et al. (2020).
The present article delivers a brief summary of the various machine learning approaches applied to evaluate water used for drinking quality. It examines the issues of keeping track of water quality, picking features approaches, and the effectiveness of different models such random forest, gradient boosting, and deep learning.

[4] "Water Quality Monitoring and Early Warning System: A Review" by Zhang et al. (2020).
The study addresses how methods based on machine learning can be integrated into surveillance technologies to identify pollution occurrences before on. It investigates the use of techniques like decision trees, random forests, and deep learning to forecast characteristics of water quality and issue appropriate notifications.

III. EXISTING SYSTEM

Because of a lack of knowledge regarding presentation of data, implementing algorithms that utilize machine learning in the current framework is quite difficult. In current approach, mathematical computations are performed to generate models, which may require a long time and be difficult. To address all of these issues, we employ machine learning programs from the repository of scikit-learn.

IV. PROPOSED SYSTEM

Multiple algorithms based on machine learning have been developed to differentiate pure from non-pure, however none have satisfactorily solve the incorrect identification issues. Stevens Multi Factor Water Quality Monitors are suitable for this kind of use. Furthermore, previous research that established algorithms for evaluating water quality usually does not include the differing quality and quantity of the data. We suggest Random Forest, Gradient Boosting, Gaussian NB, XG Boost, Ada Boost, and Cat Boost. Machine learning is used to carry out classification evaluations.

V. METHODOLOGY

Drinking water purity detection is implemented and analyzed using various machine learning models. The below diagram represents a pictorial overview of the process.
The first step is to collect the dataset and load the dataset which is of prime importance. After loading we are using some preprocessing techniques in order to filter the data for identifying the required fields for model and fetching the splitting data for training (70%) and testing (30%). Building various machine learning models on the dataset and comparing their accuracy to identifying the best model.

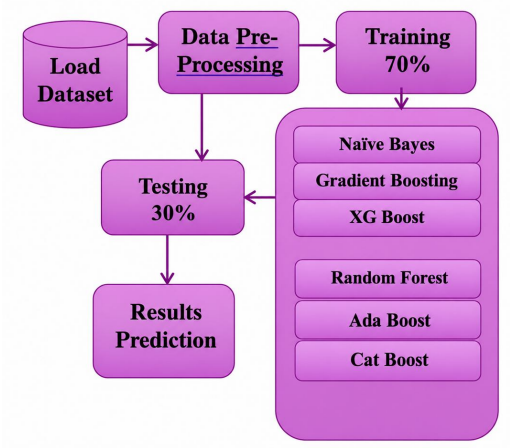


Figure 1: Proposed System Architecture

A. *Dataset*

We will use the water portability dataset received from Kaggle. The dataset consists of 3277 rows and 10 columns. The columns contain the following information

- ph: pH of 1. water (0 to 14).
- Hardness: Capacity of water to precipitate soap in mg/L.
- Solids: Total dissolved solids in ppm.
- Chloramines: Amount of Chloramines in ppm.
- Sulfate: Amount of Sulfates dissolved in mg/L.
- Conductivity: Electrical conductivity of water in µS/cm.
- Organic_carbon: Amount of organic carbon in ppm.
- Trihalomethanes: Amount of Trihalomethanes in µg/L.
- Turbidity: Measure of light emitting property of water in NTU.
- Potability: Indicates if water is safe for human consumption. Potable -1 and not potable -0

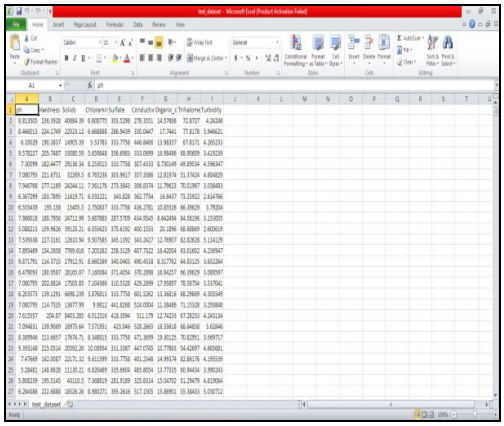


Figure2: Water Portability Dataset Sample

B. *Dataset pre-Processing*

- Null value handling
- Categorical handling
- Scaling

Data pre-processing involves transforming unprocessed data into a format that is understandable and useful. The specified data set for this data preparation module is missing some values, which may lead to inaccurate predictions. Remove missing values using Python. Also, to get correct results, the replicated data should be removed from the data set. Normalization is the process of scaling data to a specific range.

VI. ALGORITHMS

A. *Random Forest*

Random forests are used in supervised machine learning, when the target variable is labelled. Random forests can be used to solve problems involving regression (numeric target variable) and classification (categorical objective variable). Random forests are an ensembles approach, which means that their predictions are combined with those of other models. A decision tree is one of the smallest models in the random forest ensemble.

B. *Ada Boost Algorithm*

Ada Boost, short for Adaptive Boosting, is a popular ensemble learning algorithm primarily used for classification tasks. Boosting is employed to minimize skew and variation in supervised learning. It is based on the premise that learners develop progressively. Save for the first, each succeeding learning is derived by earlier created scholars. Simply said, weak learners are transformed into powerful ones. The AdaBoost method operates on the same idea as boosting, with a little variation.

C. XG Boost Algorithm:

XG Boost, which stands for "Extreme Gradient Boosting," has emerged as one of the more prevalent and extensively utilized machine learning algorithms because of its ability to handle large datasets and achieve cutting-edge efficiency for a variety of tasks involving machine learning such as regression as well as classification.

D. Navie Bayes:

A naive Bayes classifier is a statistical machine learning model designed for classification tasks. The theory of Bayes is crucial to the classifier's operation. Using Bayes' theorem, we may calculate the likelihood of A occurring given that B has occurred. B is the proof, while A is the supposition. The assumption here is that predictors and features are independent.

VII. RESULTS

A. Home Page

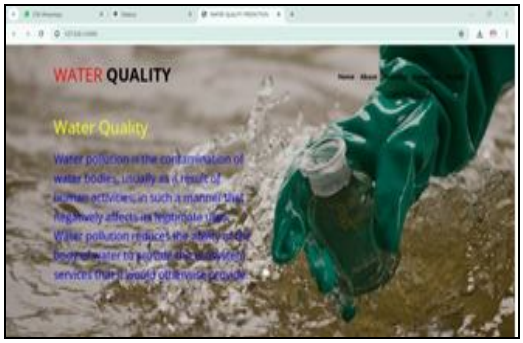


Figure 3: Water quality home page

Figure 4: Random forest accuracy

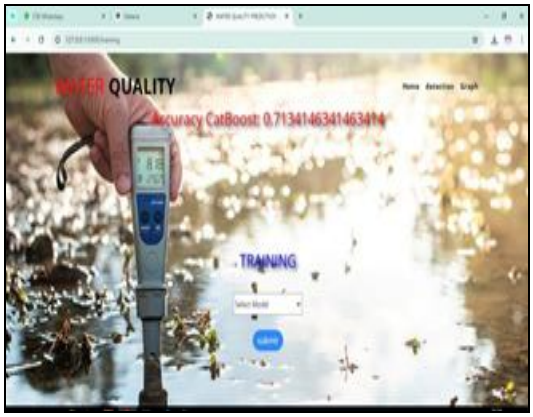


Figure 5: Cat Boost accuracy

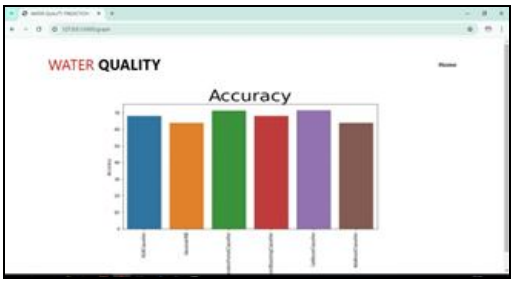


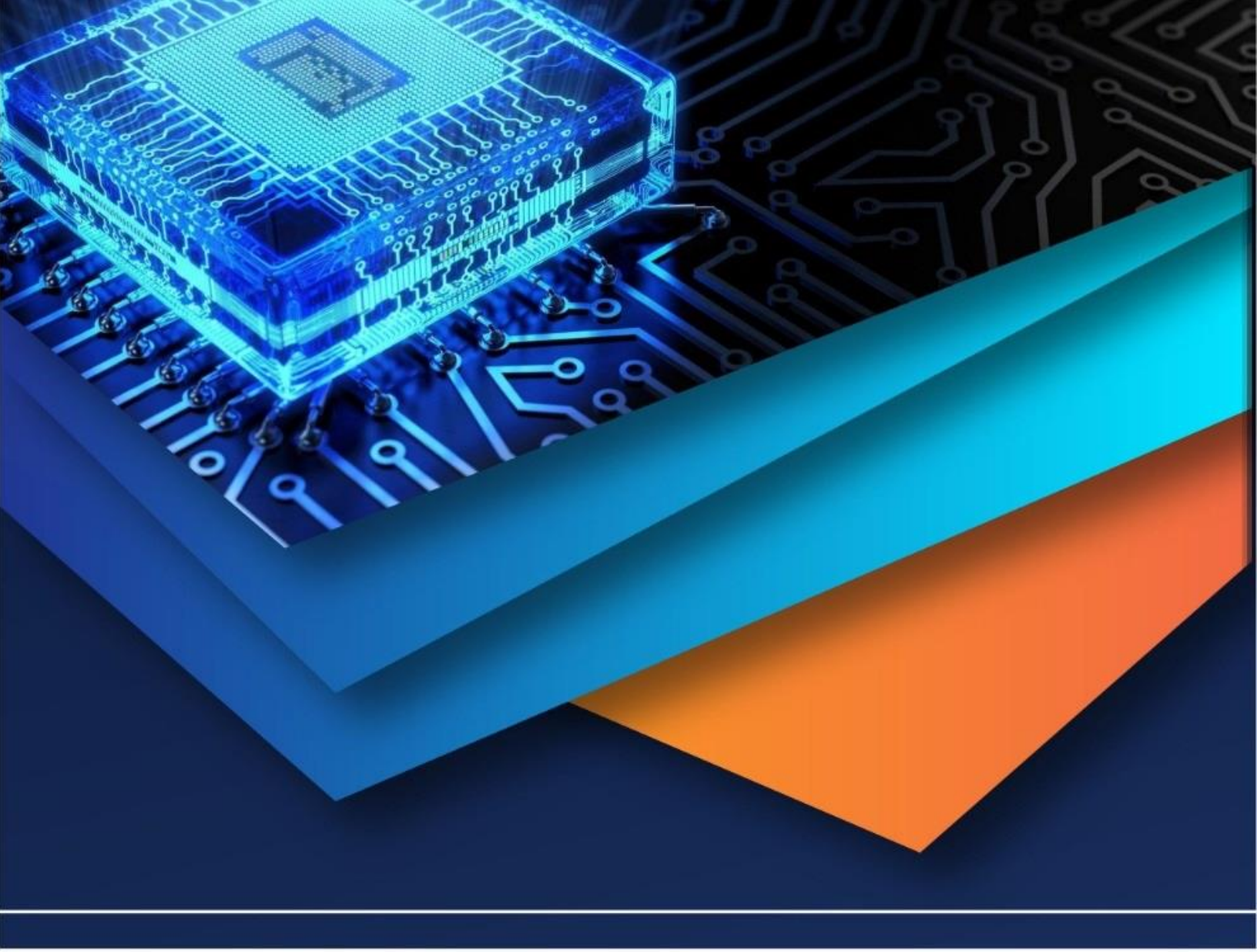
Figure 6: Accuracy Graph

VIII. CONCLUSION

In summary, studying weather forecasts using Predicting usable water is critical for environmental sustainability and pollution avoidance. Clean drinking water is essential for maintaining good public health. Drinking water from secure sources ensures that the water is portable. It becomes harder to make reliable predictions about drinking water. The perfect learning algorithm is required to avoid prediction errors. An intelligent model based on five different machine learning algorithms can be used to forecast the portability of drinking water using ten standard criteria like as pH, hardness, organic carbon, and other variables.

REFERENCES

[1] S. Geetha, S. Gouthami. Internet of Things Enabled Real Time Water Quality Monitoring System ,2017.
[2] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan and José García-Nieto. Efficient Water Quality Prediction Using Supervised Machine Learning, 2019.
[3] Ashwini K, D. Diviya, J.Janice Vedha, M. Deva Priya.Intelligent Model For Predicting Water Quality.
[4] A.N.Prasad, K. A. Mamun, F. R. Islam, H. Haqva.Smart Water Quality Monitoring System, 2015.
[5] Hadi Mohammed, Ibrahim A. Hameed, Razak Seidu.Machine Learning: Based Detection of Water Contamination in Water Distribution systems,2018.
[6] Priya Singh,Pankaj Deep Kaur.Review on Data Mining Techniques for Prediction of Water Quality,2017.
[7] Manish Kumar Jha,Rajni Kumari Sah,M.S.Rashmitha,Rupam Sinha,B.Sujatha.Smart Water Monitoring System for Real-Time Water Quality and Usage Monitoring,2018 .
[8] Water QualityMonitoring System using IoT and Machine Learning, in Proceedings of the IEEE International Conference onResearch in Intelligent and Computing in Engineering, pp.1-5, 2018.
[9] Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi, Abbas Parsaie. Water quality prediction using machine learning methods, 2018.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY