Water Quality Test Prediction for Concrete Mixing

Kalyan Appala   
Department of EEE  
Amrita Vishwa VidyapeethamCoimbatore, India  
cb.en.u4elc22004@cb.students.amrita.edu

D Sai pavan  
Department of EEE  
Amrita Vishwa VidyapeethamCoimbatore, India  
cb.en.u4elc22011@cb.students.amrita.edu

V Kiran  
Department of EEE  
Amrita Vishwa VidyapeethamCoimbatore, India  
cb.en.u4elc22065@cb.students.amrita.edu

A Samanvay  
Department of EEE  
Amrita Vishwa VidyapeethamCoimbatore, India  
cb.en.u4eee22157@cb.students.amrita.edu

*Abstract*—*Water quality has a major impact on the strength and durability of concrete. Impurities such as chloride, organic carbon, and turbidity can degrade concrete, causing structural failure. Conventional water testing is time-consuming, causing delays in decision-making during construction. This project creates a machine learning model to determine whether water quality is appropriate or inappropriate for concrete mixing.*

*With a dataset that has attributes like pH, Chloride, Organic Carbon, Solids, Sulphate, and Turbidity, models such as XGBoost, ANN, Random Forest, and SVM are trained and tested. Accuracy, precision, recall, F1-score, and confusion matrix are used to measure performance. A PyQt5 GUI allows real-time prediction, enabling rapid, data-based decisions. The system enhances construction efficiency, avoids structural collapse, and saves costs...*

Keywords— XGBoost (Extreme Gradient Boosting), ANN, Random Forest, SVM, Confusion matrix, Accuracy, Precision, Accuracy, Confusion Matrix, Graphical User Interface (GUI)

# Introduction

Concrete is a basic building material used extensively in infrastructure, buildings, and other structural uses. Concrete quality is influenced by the nature of cement, aggregates, and water employed in the mixing process. Of these, water has a significant influence on the strength, durability, and workability of concrete. Poor-quality or contaminated water may result in low bonding strength, higher porosity, and long-term structural deterioration. Factors like pH, chloride level, organic carbon, turbidity, solids, and sulphate content affect the hydration process and general performance of the concrete.

Conventional water quality analysis is based on laboratory measurements, which may be time consuming and unsuitable for real-time decisions on the construction site. Delays in water evaluation can cause the utilization of poor-quality water, resulting in weakened concrete integrity. To overcome this problem, machine learning (ML) methods provide a potential solution by streamlining water quality categorization according to historical data and real-time readings.

This project plans to create an ML-based classification system that will forecast whether water can be used for concrete mixing. Based on past datasets, various models like XGBoost, Artificial Neural Networks (ANN), Random Forest, and Support Vector Machines (SVM) will be developed and tested. Major performance measures including accuracy, precision, recall, F1-score, sensitivity, specificity, and confusion matrix will be utilized in order to evaluate the effectiveness of models.

In order to render the system user-friendly and accessible, a graphical user interface (GUI) will be created using PyQt5. This interface will enable construction professionals to feed water quality parameters and obtain instantaneous classification results. The aim is to enhance decision-making efficacy, mitigate risks arising from poor water quality, and reinforce the longevity of concrete structures. Through the integration of machine learning and real construction requirements, this system provides an economical yet trustworthy way of maintaining high-quality concrete mixing operations.

# Literature survey

# Water quality plays a crucial role in concrete mixing, directly impacting the strength, durability, and overall performance of the final product. Various studies have explored the impact of different water impurities, such as dissolved solids, pH levels, and organic matter, on concrete strength. Brindha et al. (2023) utilized XGBoost for predicting water quality parameters and highlighted its potential in ensuring optimal water selection for concrete mixing [2]. Akshay et al. (2022) further demonstrated that Random Forest and other ensemble learning models effectively classified water samples based on their suitability for construction applications [3]. Their findings emphasize the importance of machine learning in automating the assessment of water quality to enhance the reliability of concrete mixtures.

# Several machine learning models have been studied for water quality prediction, particularly for industrial applications. Khan and See (2016) explored the use of Support Vector Machine (SVM) for predicting water contamination and found it to be a reliable model for detecting impurities that may affect concrete strength [6]. Min (2011) improved Support Vector Regression (SVR) techniques to enhance predictive accuracy, ensuring better decision-making in selecting water sources for concrete mixing [7]. Pongianann et al. (2024) combined SVM with decision trees to evaluate water quality, leading to more precise classifications that can assist in maintaining the consistency of concrete mixtures *[4].* These models demonstrate the effectiveness of machine learning in ensuring water quality standards for industrial use.

The proposed project aims to develop a predictive model for evaluating water quality specifically for concrete mixing applications. By leveraging models such as XGBoost, Random Forest, SVM, and Artificial Neural Networks (ANN), this system will enable construction professionals to assess water quality in real time, reducing the risks associated with substandard mixtures. Kumar (2002) showed that ANN models outperform traditional methods in complex, non-linear predictions, making them highly suitable for industrial water quality assessment [8]. Haghibi et al. (2018) further refined deep learning models for water quality prediction, improving their robustness against variations in water sources [9]. Additionally, Flores et al. (2023) highlighted regional disparities in water quality monitoring, emphasizing the need for localized datasets to enhance prediction accuracy [5]. The integration of these machine learning models into a real-time assessment system will significantly improve the efficiency and reliability of concrete production processes

# METHOLOGY

A historical dataset containing water quality parameters: pH, Chloride, Organic Carbon, Solids, Sulphate, Turbidity.

1. Data Preprocessing: Handle missing values using imputation techniques. Remove outliers using statistical methods like Z-score or IQR.
2. Feature Scaling & Normalization: Apply Min-Max normalization to scale features between [0,1] for uniformity. Standardize numerical features if required. Since no feature is extremely redundant, the dataset is well-structured for machine learning models.
3. Dataset Splitting: Split the dataset into 80% training and 20% testing to ensure model generalization.
4. Model Selection & Training: Implement XGBoost, Random Forest, Support Vector Machine (SVM), and Artificial Neural Networks (ANN) for prediction. Optimize models using Grid Search for hyperparameter tuning.
5. Model Evaluation: Assess performance of each models using accuracy, precision, recall, F1-score, and specificity.
6. Deployment & Application: Deploy the best-performing model for real-time water quality prediction in concrete mixing. Ensure optimal mix quality and durability based on predicted water parameters.

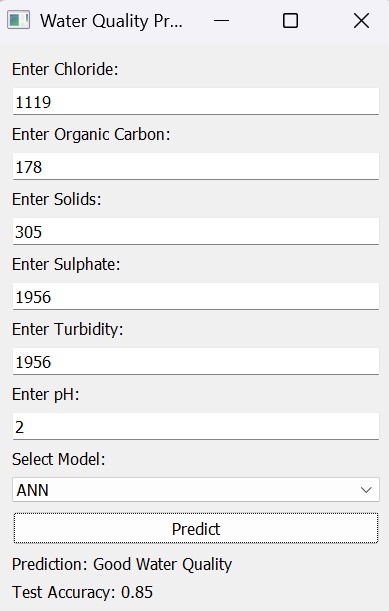


Figure 1: GUI predicted Good Water Quality

# MODELS

### XGBoost (Extreme Gradient Boosting): XGBoost is a decision-tree-based ensemble learning method that uses gradient boosting. It is widely used in machine learning competitions and real-world applications due to its efficiency and predictive power. XGBoost works by sequentially training models, where each new model corrects the errors of the previous ones. It includes built-in regularization, which helps prevent overfitting. Key Characteristics Works well with tabular data and structured datasets. Can handle missing values and imbalanced datasets. Uses tree pruning and parallelization for fast training. Requires careful tuning of hyperparameters for optimal performance.

1. *Random Forest:* Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. Each tree is trained on a random subset of the data, and the final prediction is determined by averaging (for regression) or majority voting (for classification). Key Characteristics: Handles high-dimensional data and non-linear relationships well. Provides feature importance scores, making it interpretable. Reduces variance by combining multiple decision trees. May require many trees to achieve high accuracy, which increases computational cost.
2. *Support Vector Machine (SVM):* SVM is a supervised learning algorithm that classifies data by finding the optimal hyperplane that maximizes the margin between different classes. It is effective in both linear and non-linear classification tasks, especially when using kernel functions. Key Characteristics: Works well with small to medium-sized datasets. Uses kernel functions (such as radial basis function and polynomial) to model complex relationships. Sensitive to hyperparameter choices, particularly the regularization parameter (C) and kernel parameters. Computationally expensive for large datasets due to support vector computations.
3. *Artificial Neural Network (ANN):* ANN is a deep learning model inspired by the human brain, consisting of multiple layers of interconnected neurons. It learns patterns from data through backpropagation and gradient descent. ANNs are widely used for tasks involving complex relationships and high-dimensional data. Key Characteristics: Can capture non-linear and hierarchical relationships in data Requires a large dataset to perform well. Computationally intensive, especially with deep architectures. Can be prone to overfitting, requiring techniques such as dropout and batch normalization for regularization.

# Result

A blue squares with white text

AI-generated content may be incorrect.

Figure 2:Confusion Matrix of ANN

A diagram of a confusion matrix

AI-generated content may be incorrect.

Figure 3:Confusion Matrix of SVM

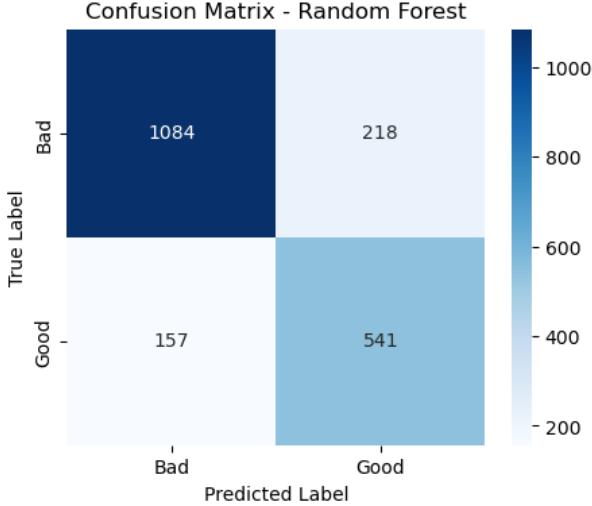


Figure 4:Confusion Matrix of Random Forest

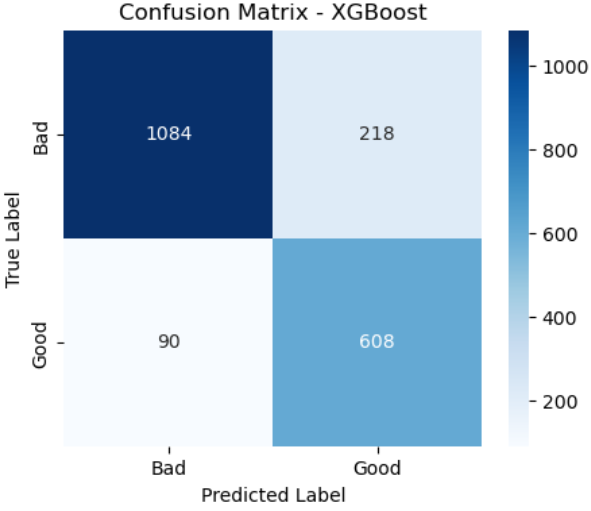


Figure 5:Confusion Matrix of XGBoost

The table below summarizes the performance of XGBoost, Random Forest, SVM, and ANN models based on various evaluation metrics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall (Sensitivity)** | **F1-Score** | **Specificity** |
| **XGBoost** | 0.8460 | 0.7361 | 0.8711 | 0.7979 | 0.8303 |
| **Random Forest** | 0.8190 | 0.6656 | 0.6218 | 0.6430 | 0.8326 |
| **SVM** | 0.8485 | 0.7377 | 0.8782 | 0.8018 | 0.8326 |
| **ANN** | 0.8485 | 0.7377 | 0.8782 | 0.8018 | 0.8326 |

SVM and ANN Perform the Best: Both models achieve the highest accuracy (84.85%). Their Recall (87.82%) is the highest, indicating strong ability to detect "Bad" water quality cases. Their F1-score (0.8018) suggests a good balance between Precision and Recall. XGBoost Performs Well: Accuracy: 84.60% (better than Random Forest but lower than SVM and ANN). Precision and Recall are moderate, making it a balanced choice. Random Forest Underperforms: Lowest Accuracy (81.90%) and Recall (62.18%). It might not be the best choice if Recall is a priority (detecting "bad" water cases). Using PyQt5 the GUI for the model is created. Each model is saved as a .pkl file for the GUI and loaded into the GUI. ANN can be used in .pkl file, So it is saved in .keras format.

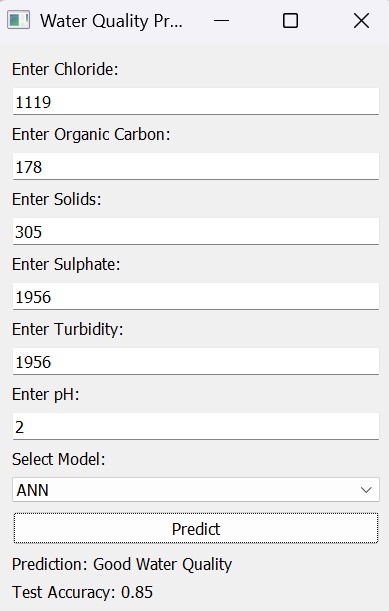


Figure 6:GUI Predicted Bad Water

# Conclusion

The comparison of different machine learning models for water quality prediction emphasizes the advantages and disadvantages of each method. On the basis of accuracy, recall, and overall performance, SVM and ANN models are the best options for this classification task. Their high recall values indicate that they are extremely efficient in identifying cases of poor water quality, which makes them especially valuable in scenarios where failing to detect such cases could have serious implications.

Though XGBoost performs satisfactorily and maintains a trade-off between precision and recall, it is slightly inferior to SVM and ANN in recall. However, it is still a good option, particularly in cases where computational efficiency and interpretability are paramount.

Conversely, Random Forest possesses the lowest accuracy and recall, possibly rendering it less ideal for high-sensitivity applications for distinguishing poor water quality conditions. But it might still prove valuable for ensemble methods or when working with small datasets where simplicity in the model is necessary.

Finally, the best model is determined based on the requirements of the application. In case high recall and accuracy are a priority, SVM and ANN are the preferred options. If efficiency and performance are to be balanced, XGBoost is a viable option. Future research could include more fine-tuning of these models, adding additional features, and trying them on bigger datasets in order to maximize overall predictive performance.

##### References

1. W. S. Davidson, "Water quality monitoring," Conference Record Southcon, Orlando, FL, USA, 1994
2. D. Brindha, V. Puli, B. K. S. NVSS, V. S. Mittakandala and G. D. Nanneboina, "Water Quality Analysis and Prediction using Machine Learning," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023
3. R. Akshay, G. Tarun, P. U. Kiran, K. D. Devi and M. Vidhyalakshmi, "Water-Quality-Analysis using Machine Learning," 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2022
4. Pongiannan R.K., Akash Sivakumar, Shivain Kohli, Poornima P.U., Lourde Jame S., Pemila. M, "Water Quality Evaluation using Machine Learning", 2024 9th International Conference on Communication and Electronics Systems (ICCES),
5. Ricardo Rojas Flores, Marcela Jamett Domínguez, "Contributions to Reduce the Gap on Water Quality Analysis in Chile and Latin America: State of the Art", 2023 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), pp.1-6, 2023
6. Y. Khan and C. S. See, "Predicting and analyzing water quality using Machine Learning: A comprehensive model," 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), Farmingdale, NY, USA, 2016
7. Osim Kumar Pal, "The Quality of Drinkable Water using Machine Learning Techniques", International Journal of Advanced Engineering Research and Science, 2022.
8. Amir Hamzeh Haghibi, Ali Heidar Nasrolahi and Abbas Parsaie, "Water quality prediction using machine learning", Journal of Water quality research, 2018
9. Zhang, Z. (Ed.). (2019). Improved Adam Optimizer for Deep Neural Networks. 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS).