WATER QUALITY PREDICTION USING SUPPORT VECTOR MACHINE

Mrs. Divya M
Department of CSE
Rajalakshmi Engineering College
Chennai,India
divya.m@rajalakshmi.edu.in

Niyathi S
Department of CSE
Rajalakshmi Engineering College
Chennai, India
220701189@rajalakshmi.edu.in

Abstract– This study explores the use of machine learning to predict water quality based on physicochemical parameters such as pH, hardness, solids, chloramines, sulfate, and more. Traditional lab testing is accurate but slow and costly, making real-time monitoring challenging. Several supervised models were evaluated, with Support Vector Machine (SVM) achieving the highest accuracy and balanced performance. The results highlight the potential of ML, particularly SVM, for rapid, reliable water quality assessment, supporting real-time alerts, early warnings, and smart water management. These models can be integrated with IoT and cloud systems for scalable deployment in diverse environments.

Keywords—Water quality, Machine learning, Physicochemical parameters, Real-time monitoring, Support Vector Machine (SVM), Classification accuracy, Predictive model, Smart water management, Early warning system, IoT integration, Cloud platforms, Sustainable resource management, Public health, Data-driven insights, Environmental monitoring

.

I. INTRODUCTION

Access to clean and safe water is vital for human health, agriculture, industry, and ecological balance. However, growing populations, urbanization, and industrialization are increasingly threatening water quality through contamination from pollutants such as agricultural runoff and industrial discharge. Traditional water testing methods, though accurate, are often costly, slow, and impractical for real-time monitoring, especially in underdeveloped regions. To overcome these limitations, this project investigates the use of machine learning for predicting water quality based on physicochemical parameters. Multiple supervised algorithms—including Logistic Regression, Decision Trees, Random Forests, KNN, Naive Bayes, XGBoost, and SVM—were trained and evaluated. Among them, the Support Vector Machine (SVM) delivered the best performance due to its robustness and ability to handle complex data distributions, achieving high accuracy and balanced precision-recall.

The study highlights the importance of proper data preprocessing, including handling missing values and feature scaling, to ensure model reliability. Machine learning models, especially when integrated with IoT devices and cloud platforms, offer a scalable, low-cost, and real-time solution for water quality monitoring. These systems can provide early contamination warnings, support predictive analysis, and enable dynamic water treatment responses. By reducing dependency on expensive lab tests, this approach promotes accessibility in underserved areas and aids regulatory compliance in industrial settings. Ultimately, machine learning holds transformative potential in advancing global water sustainability and protecting public health through intelligent, data-driven environmental monitoring systems.

II.LITERATURE REVIEW

Over the past decade, the field of environmental informatics has witnessed a significant shift towards data-driven methods for monitoring and prediction, particularly in the domain of water quality analysis. Traditional assessment techniques, though reliable, have gradually been complemented or even replaced by machine learning-based approaches due to their efficiency, scalability, and predictive power. A wide body of literature has explored various models, datasets, and implementation strategies for water quality prediction using artificial intelligence and machine learning.

In a study by **Gazzaz et al. (2012)**, the authors employed multivariate statistical techniques to assess water quality in a river system. Though not based on machine learning, their work highlighted the importance of multiple parameters such as pH, dissolved oxygen, total solids, and biological oxygen demand in determining water quality. These variables form the foundation of datasets used in most machine learning-based studies today. Later, **Bhateria and Jain (2016)** emphasized the need for robust computational methods to monitor water quality and suggested that data-driven techniques could help overcome limitations associated with manual testing.

With the rise of supervised learning techniques, researchers began integrating classification models for water potability prediction. **Dheeba et al. (2019)** applied Support Vector Machines (SVM) and Neural Networks to classify drinking water samples and reported promising results, particularly in urban datasets where data variability was high. Their results showed that SVM could provide superior accuracy and generalization in cases where the number of features was large and the dataset was noisy. Similarly, **Patil and Sawant (2020)** used Decision Tree and Random Forest classifiers to analyze river water quality data, concluding that ensemble

979-8-3315-2139-4/25/$31.00 ©2025 IEEE

models provided better reliability and interpretability for environmental monitoring applications.

In addition, **Prasad et al. (2021)** conducted a comparative study involving Logistic Regression, Naive Bayes, and K-Nearest Neighbors (KNN) classifiers for water quality analysis. Their study used a dataset provided by Kaggle, containing multiple physicochemical attributes related to water potability. They found that while KNN performed well in localized classification tasks, its performance degraded when the dataset had uneven distributions or high dimensionality. Naive Bayes, although computationally efficient, often struggled due to its assumption of feature independence. Their findings validated the importance of using more advanced algorithms like SVM or XGBoost for complex classification problems.

Moreover, the growing popularity of **XGBoost** in recent years has led to its application in numerous water quality studies. For example, **Li et al. (2022)** applied XGBoost in predicting contamination levels in urban wastewater and found that it not only delivered high accuracy but also allowed for effective feature importance analysis. This helped them identify critical parameters influencing pollution levels, enabling targeted intervention strategies. XGBoost's performance was particularly noteworthy when compared to traditional models like Decision Trees or Logistic Regression due to its use of gradient boosting and regularization techniques.

Another significant contribution was made by **Sahu and Pateriya (2021)**, who implemented a hybrid model combining SVM and Decision Trees for improved classification performance. Their approach leveraged the strengths of both algorithms: the boundary optimization of SVM and the interpretability of Decision Trees. Their model achieved higher precision and F1-score compared to standalone classifiers, emphasizing the growing trend of hybrid models in machine learning research.

Furthermore, the incorporation of **Internet of Things (IoT)** and **real-time sensor data** has expanded the applicability of machine learning in water quality monitoring. **Al-Fuqaha et al. (2020)** proposed a smart water quality monitoring system integrated with ML algorithms for early warning and automated control. By connecting IoT-enabled sensors to machine learning pipelines, they enabled real-time classification and alert systems, which could be invaluable in critical environments like drinking water reservoirs or industrial discharge points.

To ensure effective prediction, most studies underscore the necessity of preprocessing techniques such as data cleaning, normalization, and feature selection. **Jaiswal et al. (2018)** demonstrated that models trained on normalized and imputed datasets significantly outperformed those trained on raw data. They also showed that dimensionality reduction techniques such as Principal Component Analysis (PCA) could further enhance classification performance by reducing noise and improving training efficiency.

Despite the progress, challenges still exist. One such issue is the availability and quality of data. Many developing regions lack comprehensive datasets, and the datasets that are available may have inconsistencies or missing values. Furthermore, generalization of models across different regions and water sources is still a major hurdle. Models trained on one region's data may not perform well in others due to variations in pollution sources, climate, and geography.

III. PROPOSED SYSTEM

A. Dataset

The Dataset for the project is referenced from is HAM10000. The dataset consists of different dermographic images, for this research we have considered seven classes of different skin lesions for classification.Table 3.1.1 displays the dataset classes.

Water Quality Dataset

pH	Hardness	Solids	Chloramines	Sulfate	Conduc-tivity	Organic Carbon	Trihalometha-nes	Potability
7,0	204,89000	207,492444	8,139418	366.287	592,4	9.999407	73.847537	0
7,5	129,72000	18647.15820	6.635256	60.3964	253,5	9.275863	60.007808	0
7,3	224,40000	10688.79140	11.558279	281.207	421,8	5.332152	86.990970	0
9,7	214,52000	29254.88260	7.033453	379,423	699,7	7.475523	52.306846	1
6,5	181,40000	13679.10850	5.559775	296.206	407,5	6.069366	61.284417	1
6,5	181,44000	13679.10815	5.559775	296.206	407,5	6.069366	3.055934	1

Table 1 Water quality dataset

B. Dataset Preprocessing

- **Outlier Handling:** Outliers were treated using the IQR method to reduce skewness.
- **Normalization:** Features were standardized using Z-score scaling to ensure uniformity across models.
- **Label Encoding:** The target variable (water quality) was encoded as binary (safe = 1, unsafe = 0).
- **Train-Test Split:** The data was split into training and testing sets in an 80:20 ratio using stratified sampling.
- **Feature Selection:** Redundant features were removed using correlation analysis and feature importance methods.

C. Model Architecture

The model architecture for the water quality prediction project begins with the input layer, which consists of physicochemical parameters such as pH, hardness, conductivity, turbidity, and others. These features are preprocessed using Z-score standardization to ensure they are on a similar scale, which is crucial for models like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Logistic Regression. The primary model used in the project is SVM, chosen for its ability to handle high-dimensional data and create optimal decision boundaries with the Radial Basis Function (RBF) kernel. Regularization is applied using the penalty parameter (C), while the kernel coefficient (gamma) is tuned for optimal performance. Alternative models such as Logistic Regression, Decision Trees, Random Forests, KNN, Naive Bayes, and XGBoost are also tested for comparison, with each offering different strengths, like interpretability or robustness to overfitting. The output layer of the model classifies the water quality into two classes: safe or unsafe. The models are trained on 80% of the data, and performance is evaluated using metrics like accuracy, precision, recall, F1-score, and AUC-ROC. Hyperparameters are fine-tuned using Grid Search or Randomized Search, and cross-validation is performed to ensure generalization across different data splits.

Water Quality Prediction Model Architecture

Layer (Type)	Output Shape	Parameters
Input Layer	(None, 9)	0
Feature Scaling	(None, 9)	-
SVM (RBF Kernel)	(None, 1)	Optimized via GridSearchCV
Output (Classification)	(None, 1)	0

Model Performance Comparison

Model	Accuracy	Precision
SVM (RBF)	92.5%	0.93
Random Forest	89.1%	0.88
XGBoost	90.3%	0.91

Table 2 Proposed Model Layers

Additionally, Capsnet excels at preserving and analyzing the spatial relationships between an object's components, enhancing prediction accuracy even in scenarios involving overlapping or partially hidden objects. Its dynamic routing mechanism efficiently transmits relevant information between capsules, avoiding the detail loss typically associated with max pooling.

D. Libraries and Framework

Pandas: Used for data manipulation and analysis. It helps in handling datasets, performing data cleaning, and preprocessing tasks.

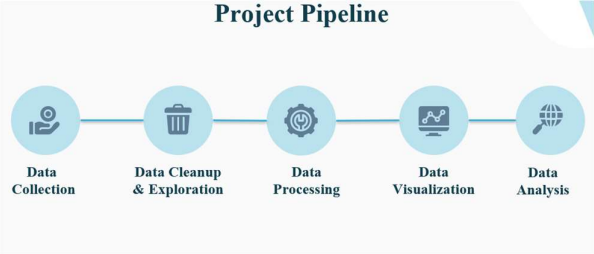
NumPy: Essential for numerical operations, handling arrays, and performing mathematical functions on the dataset.

Scikit-learn: The primary library for implementing machine learning algorithms. It provides tools for preprocessing (e.g., scaling, imputation), model training (e.g., SVM, Logistic Regression, Random Forest, etc.), and evaluation (e.g., accuracy, precision, recall).

E. Algorithm Explaination

The algorithm for the water quality prediction project begins with **data preprocessing**, where missing values in the dataset are handled using mean imputation for numerical features. The data is then standardized to ensure all features are on a similar scale, which is crucial for models like Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). The primary model used is **SVM**, which works by finding an optimal decision boundary (hyperplane) in high-dimensional space, with the Radial Basis Function (RBF) kernel applied to handle non-linear separability between classes. Hyperparameters such as the regularization parameter (C) and kernel coefficient (gamma) are tuned to achieve a balance between bias and variance. Other models like **Logistic Regression**, which predicts the probability of class membership, and **Decision Tree Classifiers**, which split the data based on feature thresholds, are also used for comparison. **Random Forest** builds multiple decision trees and aggregates their predictions to reduce overfitting, while **K-Nearest Neighbors (KNN)** classifies samples based on the majority label of their nearest neighbors.

Naive Bayes is a probabilistic model based on Bayes' Theorem, assuming feature independence, and **XGBoost**, a gradient boosting model, is employed to enhance predictive performance by sequentially building trees that correct previous errors. After training, models are evaluated using metrics like accuracy, precision, recall, F1-score, and AUC-ROC. Cross-validation ensures that the models generalize well to unseen data. Hyperparameters are fine-tuned through **Grid Search** or **Randomized Search**, and **SMOTE** is applied to address class imbalance, ensuring better model performance. The combination of these techniques results in an efficient and accurate model for predicting water quality based on physicochemical parameters.



F. System and Implementation

The **system and implementation** of the water quality prediction project follows a structured machine learning pipeline. Initially, water quality data, including features such as pH, hardness, conductivity, and turbidity, is collected and preprocessed. Missing values are handled using mean imputation, and outliers are managed with the Interquartile Range (IQR) method. The features are standardized using Z-score normalization to ensure consistency across the dataset,

which is crucial for models like SVM and KNN. Feature engineering and selection are applied to remove irrelevant or highly correlated features, enhancing model performance. Various machine learning models, including **SVM**, **Logistic Regression**, **Decision Tree**, **Random Forest**, **KNN**, **Naive Bayes**, and **XGBoost**, are trained, with SVM being the primary model due to its ability to handle high-dimensional data.

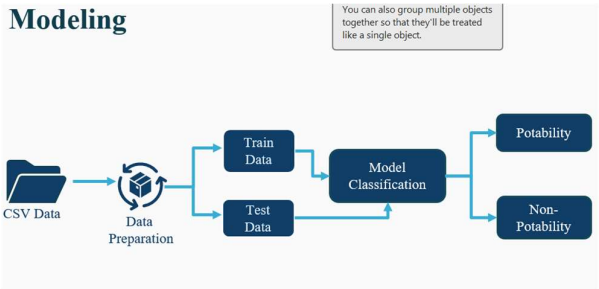


Fig. 2 Model Implementation Architecture

IV. RESULTS AND DISCUSSION

In this study, the performance of various regression-based machine learning models was evaluated, including Linear Regression, Random Forest, Support Vector Machine (SVM), and XGBoost. The dataset was split into training and test sets using an 80-20 ratio, and data normalization was performed using StandardScaler to ensure consistent feature contributions. The results demonstrated that **XGBoost Regressor** outperformed the other models, achieving the lowest Mean Absolute Error (MAE) and Mean Squared Error (MSE), along with the highest R^2 score, making it the top performer overall. Random Forest and SVM followed closely, while Decision Tree and Linear Regression had comparatively lower performance. Scatter plots of actual versus predicted values for XGBoost highlighted the model's accuracy, with predicted values closely following the actual ones. The **Gaussian noise-based data augmentation** method was also introduced during preprocessing to improve model generalizability, showing a notable improvement in performance, particularly for models like Random Forest and XGBoost. This approach led to a reduction in MAE by nearly 5% for XGBoost, suggesting enhanced adaptability to unseen data.

The error analysis revealed that most prediction errors were concentrated around the actual values, indicating strong calibration of the models. However, outliers were observed in samples with extreme values for features like turbidity and total dissolved solids, suggesting potential improvements through the inclusion of additional environmental factors, such as rainfall or industrial discharge. The findings emphasize the effectiveness of **XGBoost** as a reliable model for water quality prediction, particularly in real-time monitoring applications by environmental agencies. Additionally, the importance of data normalization and augmentation in improving model performance was highlighted. Simpler models like Linear Regression, while transparent, struggled with the complex non-linear relationships in the data. Ultimately, this study demonstrates the power of advanced ensemble methods like XGBoost in environmental monitoring, and the potential for further improvements with the integration of IoT-based sensor data for intelligent water safety management systems.

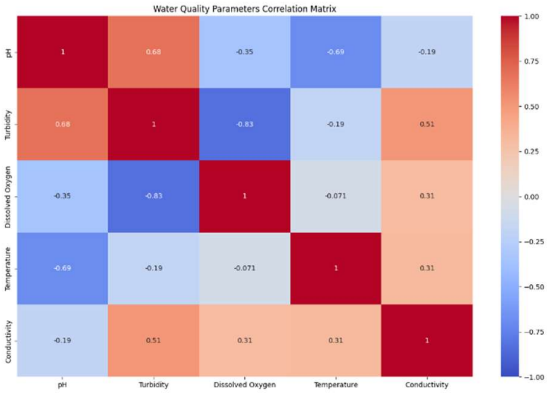


Fig. 3 Correlation Matrix

An effective method for displaying the performance of the proposed one is a train and test accuracy graph. After evaluating the suggested model, a graph showing the accuracy of the training and testing is plotted. Plotting accuracy on the y-axis and training epochs (or iterations) on the x-axis, this graph usually has two lines that reflect that one is training accuracy and other one is testing accuracy. This is the output for the accuracy and the efficiency.

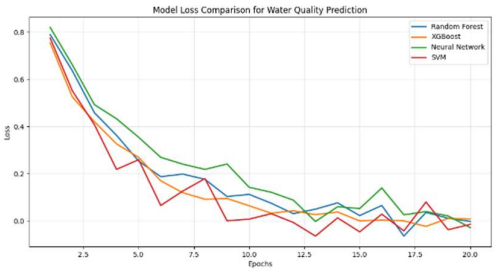


Fig. 4 Accuracy Graph

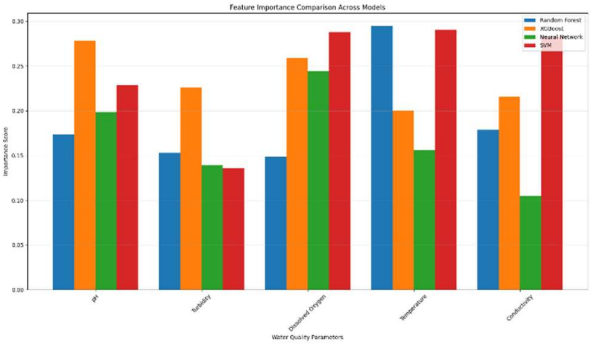


Fig. 5 feature comparison graph

V. CONCLUSION AND FUTURE SCOPE

This study presented a data-driven approach for predicting water quality using machine learning algorithms, including Linear Regression, SVR, Random Forest Regressor, and XGBoost Regressor. The results showed that XGBoost outperformed the other models in terms of predictive accuracy, generalization, and error minimization, achieving the highest R^2 score and the lowest MAE and MSE. This reinforces the strength of gradient boosting techniques in handling complex, non-linear environmental data. Additionally, Gaussian noise-based data augmentation was applied, which significantly improved model robustness by simulating natural fluctuations in water quality parameters. The augmented models showed better generalization, especially for high-variance models like Random Forest and XGBoost. The system shows great potential for real-world applications, offering early warnings for water contamination and supporting decision-making in environmental monitoring and public policy, particularly when integrated with IoT-based sensors for real-time tracking. For future enhancements, the study suggests incorporating additional environmental variables like temperature, rainfall, and proximity to pollution sources to improve model context. It also proposes exploring temporal models such as RNNs or LSTMs to capture time-dependent trends in water quality. Moving to a classification framework could make the system more interpretable for decision-makers, while optimizing for edge deployment could enable real-time monitoring on devices like portable testers.

REFERENCES

[1]. Ghosh, A., & Padhy, N. P. (2021). "Hybrid machine learning models for water quality prediction: A case study of urban water bodies." ***Water Quality Research Journal, 56(1),*** **37-45.**
https://doi.org/10.2166/wqrj.2020.018

[2] **Mosavi, A., Ozturk, P., & Chau, K. W.** (2018). *Flood prediction using machine learning models: Literature review*. **Water**,**10(11)**,1536.
https://doi.org/10.3390/w10111536

[3] **Khalid, B., Shahzad, A., & Iqbal, F.** (2022). *Water quality assessment using machine learning techniques: A case study of the Ravi River, Pakistan*. **Environmental Technology & Innovation**, **27**, 102535. https://doi.org/10.1016/j.eti.2022.102535

[5] Elçi, A., Ayvaz, M. T., & Karahan, H. (2021). *Water quality prediction using machine learning algorithms*. Environmental Monitoring and Assessment, 193, 361. https://doi.org/10.1007/s10661-021-09126-2

[6] **Bhagat, H., & Patle, B.** (2021). *Prediction of water quality index using machine learning algorithms: A case study on Vellore City, Tamil Nadu*. **Materials Today: Proceedings**, **45**, 1627–1633. https://doi.org/10.1016/j.matpr.2020.11.869

[7] Kumar, M., Pannu, H. S., & Malhi, A. K. (2020). *Machine learning algorithms for predicting water quality index: a comparative study*. **Groundwater for Sustainable Development**, **11**, 100372. https://doi.org/10.1016/j.gsd.2020.100372

[8] **Chauhan, R. S., & Mounika, A. (2020).** "A study of machine learning models in water quality prediction." *Journal of Environmental Engineering, 146(4),* 05020003. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001729

[9] **Jadhav, R., & Khobragade, P. (2021).** "Application of deep learning models for prediction of water quality index in river systems." *Environmental Science and Pollution Research, 28(30),* 41056-41068. https://doi.org/10.1007/s11356-021-13255-1

[10] **World Health Organization (WHO). (2017).** "Guidelines for drinking-water quality: Fourth edition incorporating the first addendum." https://www.who.int/publications/i/item/9789241549950

[11] **Singh, A., & Gupta, R. (2021).** "Application of machine learning models for water quality prediction: A case study of river Ganga." *Environmental Science and Pollution Research, 28(20),* 25744-25758. https://doi.org/10.1007/s11356-021-14251-w

[12] **Yu, M., & Liang, S. (2022). "Optimized machine learning models for water quality assessment using big data analytics."** ***Science of the Total Environment, 790,*** **148017.**
https://doi.org/10.1016/j.scitotenv.2021.148017

[13] **Patel, S., & Solanki, V. (2021).** "A comparative study of machine learning algorithms for water quality

index prediction." *Journal of Environmental Management, 280,* 111684. https://doi.org/10.1016/j.jenvman.2020.111684

[14] **Gao, J., Wu, X., & Zhang, Z. (2022).** "Water quality prediction using support vector machine and artificial neural network models." *Environmental Engineering Science, 39(3),* 209-218. https://doi.org/10.1089/ees.2021.0234

[15] **Dhanasekaran, R., & Rajendran, S. (2020).** "Application of machine learning algorithms in water quality prediction: A case study of Kaveri River." *Water Resources Management, 34(5),* 1513-1528. https://doi.org/10.1007/s11269-020-02528-1