International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Hybrid Modelling Approaches Validation for Innovative Water Quality Prediction using N-Fold Oversampling

Shilpa Agnihotri Pandey [1], Dr. Prajeet Sharma [2]

[1]Research Scholar, Department of CS&E, LNCT University, Bhopal, MP 462022 India
2 Assistant Professor, Department of CS&E, LNCT University, Bhopal, MP 462022 , India
Email: [1] shilpa.agnihotri28@gmail.com [2,] prajeets@lnctu.ac.in

Abstract

This research explores the validations of hybrid machine learning and deep learning methodologies for predicting water quality, utilizing a water_potability dataset characterized by class imbalance and comprising 3,276 samples with essential physicochemical attributes. The study introduces composite models that integrate traditional classifiers—Support Vector Machine (SVM), Decision Tree (DT), and Gradient Boosting (XGBoost), with Convolutional Neural Networks (CNN), aiming to harness their complementary capabilities for enhanced classification performance. To mitigate the effects of data imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is applied, thereby improving model resilience and generalization. Comparative evaluation reveals that the SVM-CNN hybrid model achieves the highest predictive accuracy at 81.20%, surpassing the XGBoost-CNN and DT-CNN configurations, which attain 78.90% and 75.60%, respectively. These outcomes underscore the potential of hybrid ML-DL architectures in modeling complex nonlinear interactions and advancing the reliability of water potability assessments in environmental informatics.

Keywords: Water Quality Prediction, Water Potability, Hybrid Model approach, Support Vector Machine (SVM, Decision Tree, Gradient Boost, CNN, Deep learning

1. Introduction

Water quality is a fundamental determinant of human health, ecological balance, and environmental sustainability. With the rapid growth of populations and the intensification of industrial activities, ensuring the preservation and enhancement of water quality has become an increasingly complex challenge. Water bodies worldwide are exposed to diverse pollutants—chemical, biological, and physical—that significantly alter their composition and safety. Consequently, the demand for precise and dependable water quality prediction methods has become more critical than ever. Predictive modeling serves as a vital tool for evaluating potential risks, informing policy development, and supporting effective water resource management. Nonetheless, conventional modeling techniques often struggle to represent the intricate, nonlinear dynamics of aquatic systems, highlighting the necessity for advanced and data-driven approaches such as machine learning (ML) and artificial intelligence (AI).

The Need for Enhanced Predictive Accuracy: Traditional water quality prediction models, while valuable, often struggle to account for the intricate interplay between numerous environmental variables. These models typically rely on linear relationships and may not adequately handle the variability and uncertainty present in real-world water systems. Figure 1 depicts a conceptual diagram which outlines the fundamental framework for describing importance of water quality prediction (WQP), emphasizing its vital role in environmental monitoring and public health protection.

Received: November 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

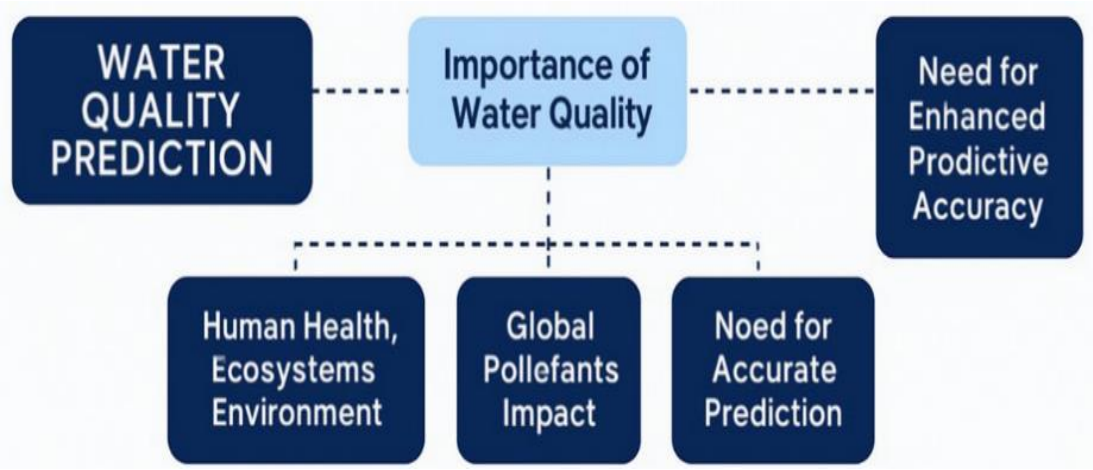


Figure 1. Importance of Water Quality prediction problem

Hybrid Modeling Approaches: To address these challenges, researchers have increasingly turned to hybrid modeling approaches. A hybrid model combines two or more different modeling techniques to leverage the strengths of each while mitigating their individual weaknesses. In the context of water quality prediction, hybrid models often integrate traditional statistical methods with advanced machine learning techniques. This combination allows for the modeling of both linear and nonlinear relationships, improving the overall predictive accuracy. For example, a hybrid model might combine a time series analysis method, such as the autoregressive integrated moving average (ARIMA), with a machine learning algorithm like a neural network. The time series component can effectively model temporal dependencies, while the neural network can capture complex, nonlinear relationships.

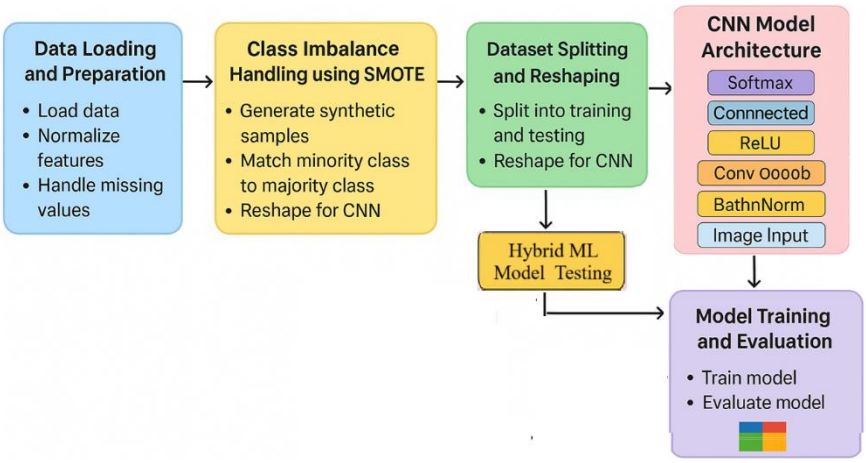


Figure 2. Architecture of the Hybrid ML based water Quality Prediction Models

Advantages of Hybrid Models: Hybrid models offer several advantages over traditional approaches. First and foremost, they can handle a broader range of data types and relationships. By integrating different modeling techniques, hybrid models can more accurately reflect the multifaceted nature of water systems. This results in predictions that are not only more accurate but also more robust across different scenarios and datasets. Additionally, hybrid models can be tailored to specific contexts, allowing for the inclusion of domain-specific knowledge. For instance, in areas where certain pollutants are of particular concern, the model can be adjusted to focus more on those variables, thereby enhancing its relevance and accuracy.

Another significant advantage of hybrid models is their ability to improve predictive performance in the face of data limitations. In many cases, water quality data may be sparse, noisy, or incomplete. Hybrid models can compensate for these issues by using different techniques to fill in gaps, smooth out noise, and make the most of available data. This is particularly important in developing regions where

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

comprehensive water quality monitoring systems may not be in place. By improving the accuracy of predictions, hybrid models can help policymakers and stakeholders make better-informed decisions, ultimately leading to more effective water management practices.

Implementing Hybrid Models: Implementing hybrid models for water quality prediction involves several steps, starting with data collection and preprocessing. High-quality, relevant data is essential for building accurate models. This data typically includes information on various water quality parameters, such as pH, dissolved oxygen, temperature, and the presence of specific contaminants. Environmental factors, such as weather conditions, land use patterns, and industrial activities, are also crucial inputs. Once the data is collected, it must be preprocessed to address issues like missing values, outliers, and noise. This step often involves techniques such as normalization, imputation, and data transformation. Following data preprocessing, the next step is to select the appropriate modeling techniques to be combined in the hybrid model. This selection depends on the specific characteristics of the data and the goals of the prediction. For example, if the primary objective is to model temporal patterns, time series analysis techniques may be chosen. If the data exhibits complex, nonlinear relationships, machine learning algorithms like neural networks or support vector machines may be included. The chosen techniques are then integrated into a single hybrid model, often through techniques like ensemble learning, where multiple models are trained and their predictions are combined to produce a final output.

Challenges of Hybrid Approaches: While hybrid models offer significant advantages, they also come with challenges. One of the primary challenges is the complexity of model design and implementation. Integrating different modeling techniques requires careful consideration of how they will interact and complement each other. This often involves extensive experimentation and tuning to achieve the desired balance between accuracy and generalization. Additionally, hybrid models can be computationally intensive, requiring significant processing power and time, especially when dealing with large datasets.

Consideration is the interpretability of hybrid models. As these models become more complex, understanding the underlying processes and relationships can become more difficult. This can be a barrier to their adoption, particularly in settings where transparency and accountability are important. To address this, researchers are exploring ways to improve the interpretability of hybrid models, such as by using techniques that provide insights into the contribution of different variables to the final prediction. Maximizing water quality prediction precision through innovative hybrid modeling approaches represents a significant advancement in the field of environmental science. By combining the strengths of different modeling techniques, hybrid models offer improved accuracy, robustness, and relevance, making them an invaluable tool for water resource management. While challenges remain, ongoing research and development in this area hold great promise for the future, offering new opportunities to protect and preserve our vital water resources.

2. Literature Review

Recent research on water quality and potability prediction demonstrates a clear transition toward more data-driven, hybrid, and uncertainty-aware modelling frameworks. Traditional machine-learning classifiers and ensemble pipelines have shown consistently strong performance on standard potability and WQI datasets, as evidenced by Dritsas and Trigka (2023), who achieved near-perfect accuracy through stacked ensemble learning supported by SMOTE balancing and cross-validation. Work has proved the importance of data balancing for WQP task.

Parallel efforts, such as the hybrid ML models proposed by Biju et al. (2024) and the stacked-feature-selection approach by Elshewey et al. (2025), confirm that hybridization and feature optimization significantly enhance classification robustness compared with single-model baselines. Complementing these advances, Shams et al. (2024) demonstrated that systematic hyperparameter tuning using grid search substantially improves prediction accuracy in both regression- and classification-based WQI models, highlighting the sensitivity of model performance to search

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

space design. Beyond standard benchmark datasets, several studies focus on real-world temporal and spatial water quality forecasting. Farzana et al. (2023) and Li et al. (2023) found that deep learning approaches such as LSTM and time-graph convolutional networks provide more reliable long-term predictions for reservoir and marine-ranch environments, while hybrid frameworks combining wavelet transforms, ANN, and LSTM (Wu & Wang, 2022) capture multiscale temporal variations more effectively than single models. In parallel, the integration of soft-computing techniques—such as fuzzy logic combined with ANN in Trach et al. (2022)—has improved interpretability and uncertainty handling in WQI estimation.

The importance of uncertainty quantification is further emphasized by Uddin et al. (2023), who propose a novel probabilistic modelling framework to quantify multi-model and input uncertainty, addressing a critical gap in conventional deterministic ML studies. Comparative assessments of ML algorithms for potability classification (Musleh, 2024; Sinap, 2024; Kamath et al., 2025) consistently show that tree-based and ensemble methods outperform linear and distance-based algorithms, though their reliance on small, imbalanced public datasets limits generalizability. Additional works investigating hybrid optimization approaches, such as bio-inspired ML models (Bui et al., 2020) and deep neural synthesis of heavy-metal parameters (Moeinzadeh et al., 2023), further reveal that complex hybrid and deep models can extract richer feature representations and achieve higher predictive accuracy, albeit at the expense of interpretability and computational cost. Overall, the literature converges on the conclusion that hybrid ML/AI frameworks, uncertainty-aware modelling, and temporally- and spatially-aware deep learning architectures offer the most promising direction for robust and operationally useful water quality prediction. However, common limitations—including dataset imbalance, limited external validation, overfitting risks in deep or stacked models, and a lack of transferable evaluation across geographical and seasonal contexts, and highlight the need for more comprehensive, large-scale, and generalizable modelling frameworks in future research.

3. Water Quality Prediction Parameters and Methods

Water quality prediction (WQP) system or methods accuracy relies on analyzing multiple physicochemical and biological parameters that collectively determine the safety and suitability of water for human consumption and environmental health. Figure 3 has presented the visual representation of various parameters used for WQP system or data. Key parameters for WQP are pH, turbidity, hardness, and total dissolved solids (TDS) reflect the fundamental physical and chemical properties of water. Chemical parameters may include percentage content of nitrate, fluoride, sulfate, chloramines, and their conductivities help identify contamination from agricultural, industrial, or municipal sources. Other chemical and biological parameters may be Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), and Chemical Oxygen Demand (COD) assessing the biological and chemical oxygen balance crucial for aquatic life. Additionally, disinfection byproducts such as trihalomethanes (THMs) indicate treatment-related chemical reactions. Collectively, these parameters form the foundation for accurate water quality modeling and predictive analysis, enabling early detection of pollution and effective water resource management. This research is aimed at designing the hybrid ML based predictive moles using dataset of various existing parameters. The paper considered the various WQP parameters to solve the binary classification problem for WQP.

Received: November 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

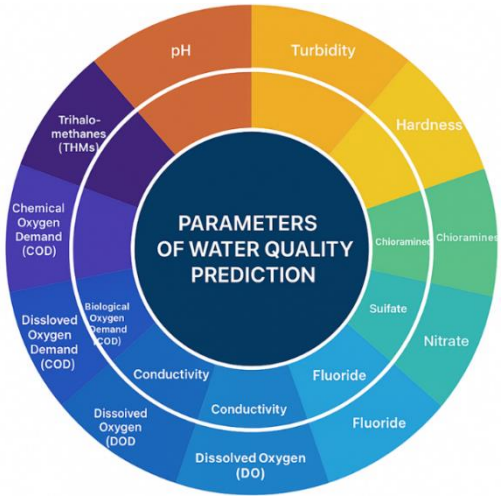


Figure 3. Parameters of the WQP system or data

Figure 4 illustrates the various possible machine learning (ML) applications for predicting and managing water quality across diverse environmental and operational contexts. Figure 4 highlights how ML techniques such as neural networks (NN), SVM, random forest (RF), and DL models are utilized to analyze complex datasets derived from sensors, laboratories, and remote sensing systems. These models enable accurate prediction of key water quality indicators as mentioned above in Figure 3 which are essential for assessing potability and ecosystem health. Beyond prediction, ML supports real-time anomaly detection, identifying contamination events or sudden parameter deviations to ensure timely interventions. It also aids in source contamination tracing, distinguishing between industrial discharge, agricultural runoff, and natural pollution sources. Furthermore, time-series forecasting models such as LSTM and ARIMA are applied to monitor long-term water quality trends, supporting preventive management strategies. It is observed that taking advantage of hybrid competitions of WQP methods using ML approaches may improve the prediction accuracy, thus this paper has proposed investigating various hybrid ML solutions for achieving this task.

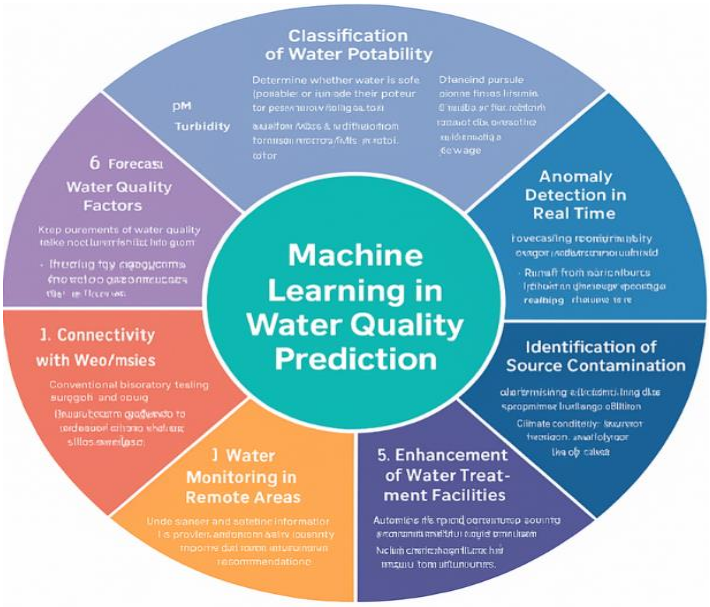


Figure 4. Various possible ML applications for predicting water quality

Received: November 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4. Proposed Methodology

The process begins with the dataset, which undergoes initial data exploration to understand its structure and contents. Following this, the data is normalized using Z-score normalization to standardize the variables. The normalized data is then used in two parallel streams: one involves applying a hybrid deep learning model for prediction, where the performance is evaluated using metrics like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The other stream involves classification tasks, evaluated based on accuracy, sensitivity, specificity, and precision. Additionally, machine learning models are employed for correlation analysis, determining the relationships between variables within the dataset. The entire workflow is designed to ensure a comprehensive analysis of the dataset, leveraging both deep learning and traditional machine learning approaches.

4.1 Data Loading and Cleaning

Let the dataset be:

$$\mathrm{D} = \{(\mathrm{x}_i, y_i)\}_{i=1}^{N}, \mathrm{x}_i \in R^d, y_i \in \{0,1\} \quad (1)$$

where $d$is the number of features and $y_i$represents potability.

Handling Missing Values:

- Original missing entries in D are removed:

$$\mathrm{D}_{clean} = \mathrm{D} \setminus \{\mathrm{x}_i \mid \mathrm{x}_i \text{ has NaN values}\} \quad (2)$$

- Randomly mask 10% of the data for imputation:

$$\mathbf{x_{i,j}} = \text{NaN for randomly selected indices} \quad (3)$$

- Impute missing values with column mean:

$$\mathbf{x}_{i,j} = \frac{1}{N_j}\sum_{i=1}^{N_j}\mathbf{x}_{i,j} \text{ if } \mathbf{x}_{i,j} \text{ is NaN} \quad (4)$$

where $N_j$is the number of non-missing values in feature j.

4.2. Feature Scaling

Min-max normalization of features:

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \min(x_{:,j})}{\max(x_{:,j}) - \min(x_{:,j})} \forall i,j \quad (5)$$

4.3. Correlation Analysis

Compute the correlation matrix:

$$\mathrm{C} = \mathrm{corr}(\tilde{\mathrm{X}}), \mathrm{C}_{jk} = \frac{\mathrm{cov}(\tilde{x}_j, \tilde{x}_k)}{\sigma_j \sigma_k} \quad (6)$$

where $\tilde{x}_j$is the j-th normalized feature.

4.4. Class Balancing with SMOTE

For class $c \in \{0,1\}$with fewer samples, synthetic data is generated

for $i = 1$ to n_{synth}:

$$\mathbf{x}_{synth} = \mathbf{x}_i + \mathrm{rand}(0,1) \cdot (\mathbf{x}_{neighbor} - \mathbf{x}_i) \quad (7)$$

end

where $\mathrm{x}_{neighbor}$is a randomly chosen neighbor of the same class.

After SMOTE, the new dataset becomes:

$$\mathrm{D}_{balanced} = \mathrm{D}_{clean} \cup \mathrm{D}_{synth} \quad (8)$$

4.5 Train and Split Data

The data is trained and splinted at the 80:20 basis as used by the most of the existing cases.

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5. ML Methodologies and Modeling

The three different ML DL models are being tested in this work for highly imbalanced WQP dataset. Then after validation the hybrid combination of these modes are investigated for the accuracy improvement.

5.1 Support Vector Machine (SVM)

Train SVM with kernel $K$and box constraint C:

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i \quad (9)$$

subject to:

$$y_i(w^\top \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \quad (10)$$

where $\phi(\cdot)$is the kernel function (linear, RBF, polynomial)

5.2 Decision Tree and XGBoost

- **Decision Tree:** recursively partition feature space to minimize impurity (e.g., Gini index or entropy):

$$\text{Impurity}(S) = 1 - \sum_{c} p_c^2 \quad (11)$$

- **XGBoost (LogitBoost):** ensemble of weak learners using additive model:

$F_m(x) = F_{m-1}(x) + \nu h_m(x), h_m(x)$ is weak learner (12)

5.3 Deep Neural Network (CNN)

Feed-forward network:

$$\begin{aligned} h_1 &= \text{ReLU}(W_1 x + b_1) \\ h_2 &= \text{ReLU}(W_2 h_1 + b_2) \\ y_{pred} &= \text{softmax}(W_3 h_2 + b_3) \end{aligned} \quad (13)$$

with dropout applied:

$h_i = h_i \cdot \text{mask}, \text{mask} \sim \text{Bernoulli}(1-p)$ (14)

5.4 Proposed Hybrid Models (ML + CNN)

Finally, in this work the hybrid combination of the CNN model with ML methods like SVM and DT are designed and investigated for the performance for WQP task. the block diagram of the prosed hybrid-based testing of WQP is illustrated in the Figure 6 and are mathematically described further.

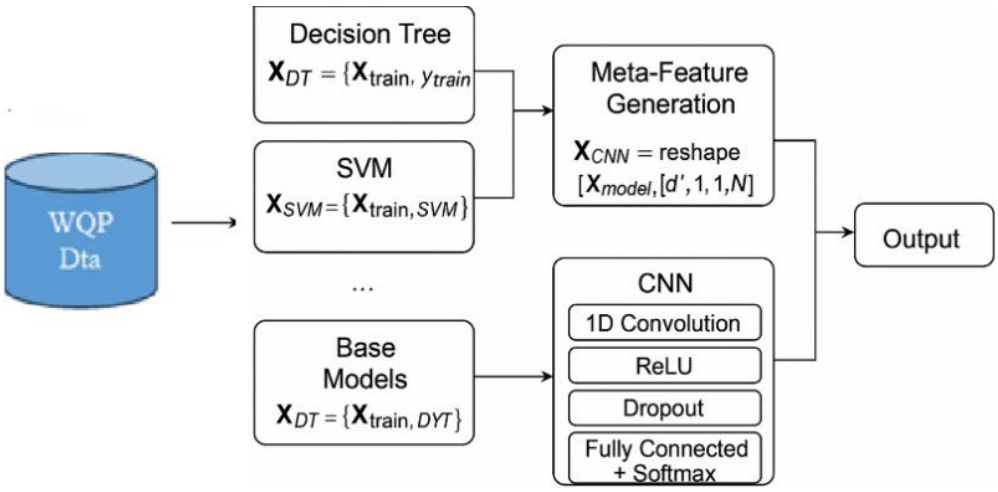


Figure 5. Proposed Hybrid Model testing methodology

Received: November 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1. Generate meta-features from base models:

$$\mathrm{X}^{DT} = [\mathrm{X}_{train}, \hat{\mathrm{y}}_{train}^{DT}], \quad (15)$$

$$\mathrm{X}^{SVM} = [\mathrm{X}_{train}, \hat{\mathrm{y}}_{train}^{SVM}], \quad (16)$$

$$\mathrm{X}^{zGBoost} = [\mathrm{X}_{train}, \hat{\mathrm{y}}_{train}^{SzGBoost}] \quad (17)$$

2. Reshape for CNN input:

$$\mathrm{X}_{CNN} = \mathrm{reshape}(\mathrm{X}^{model}, [d', 1, 1, N]) \quad (18)$$

CNN performs 1D convolution, ReLU, dropout, fully connected + softmax for classification:

$$\mathrm{y}_{hybrid} = \mathrm{CNN}(\mathrm{X}_{CNN}) \quad (19)$$

The proposed hybrid framework integrates machine learning (ML) models with a Convolutional Neural Network (CNN) to enhance classification performance by leveraging meta-features. First, predictions from multiple base learners—such as Decision Trees, SVM, and Gradient Boosting—are combined with the original training data to form enriched meta-feature sets (e.g., $X^{DT} = [X_{train}, \hat{y}_{train}^{DT}]$, $X^{SVM} = [X_{train}, \hat{y}_{train}^{SVM}]$, $X^{zGBoost} = [X_{train}, \hat{y}_{train}^{zGBoost}]$). These meta-features capture both input characteristics and model-specific learning patterns. Next, each meta-feature set is reshaped into a 4-dimensional tensor format suitable for CNN processing, represented as

$$X_{CNN} = \mathrm{reshape}(X^{model}, [d', 1, 1, N]). \quad (20)$$

This transformation enables the CNN to treat the meta-features as structured sequences. Finally, the CNN applies 1D convolution layers followed by ReLU activation, dropout regularization, and fully connected layers with softmax output to generate the final predictions. The hybrid model output is expressed as

$$y_{hybrid} = \mathrm{CNN}(X_{CNN}), \quad (21)$$

This allowing the network to learn high-level representations from the combined strengths of multiple ML models.

.

5. Result and Discussion

This paper has compared and investigated the performance of the hybrid ML and Dl approaches for water quality prediction (WQP). the data ued in the water_potability data available publicly on Koggle. The water quality dataset consists of 3,276 samples, each containing key physicochemical indicators used to assess the safety of drinking water. These features include pH, hardness, total dissolved solids (TDS), chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity, all of which influence the chemical balance, mineral content, and overall purity of water. The dataset also includes a potability label indicating whether the water is safe for human consumption. the screen shot of the data used for the investigation is given in Table 3 as example. The selected parameters align with international guidelines such as WHO and EPA, ensuring that the dataset captures realistic variations found across natural and treated water sources. This comprehensive feature set makes the dataset suitable for predictive modeling and analysis of drinking water quality.

Table 1 screenshot of the WQP data features

ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
NaN	204.89	20791	7.3002	368.52	564.31	10.38	86.991	2.9631	0
3.7161	129.42	18630	6.6352	NaN	592.89	15.18	56.329	4.5007	0
8.0991	224.24	19910	9.2759	NaN	418.61	16.869	66.42	3.0559	0
8.3168	214.37	22018	8.0593	356.89	363.27	18.437	100.34	4.6288	0
9.0922	181.1	17979	6.5466	310.14	398.41	11.558	31.998	4.0751	0
5.5841	188.31	28749	7.5449	326.68	280.47	8.3997	54.918	2.5597	0
10.224	248.07	28750	7.5134	393.66	283.65	13.79	84.604	2.673	0
8.6358	203.36	13672	4.563	303.31	474.61	12.364	62.798	4.4014	0

Received: November 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

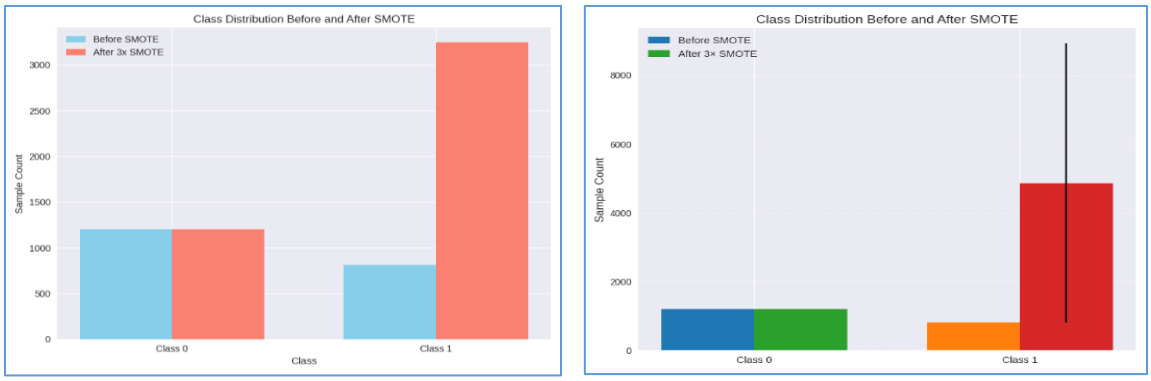
Table 1 provides a concise statistical overview of the water quality dataset by presenting the minimum, median, and maximum values for each measured parameter, along with the count of missing records. The data shows wide variability across features such as pH, solids, and conductivity, indicating diverse water conditions across the sampled locations. Some parameters, notably pH, sulfate, and trihalomethanes, contain missing values that may require preprocessing before model development. The potability feature ranges from 0 to 1, reflecting the classification label for safe or unsafe drinking water. Overall, this summary helps in understanding the distribution, scale differences, and data quality issues essential for further analysis and predictive modeling.

Table 2 statistical summary of the water quality data features

Parameter	Min	Median	Max	Missing Values
pH	0	7.0368	14	491
Hardness	47.432	196.97	323.12	0
Solids	320.94	20,928	61,227	0
Chloramines	0.352	7.1303	13.127	0
Sulfate	129	333.07	481.03	781
Conductivity	181.48	421.88	753.34	0
Organic Carbon	2.2	14.218	28.3	0
Trihalomethanes	0.738	66.622	124	162
Turbidity	1.45	3.955	6.739	0
Potability	0	0	1	0

5.1 Results of the Oversampling (SMOTE)

Figure 6 illustrates the impact of applying SMOTE (Synthetic Minority Over-sampling Technique) on the class distribution within the dataset. The Figure 4 a) is showing data balancing for 4 times oversampling. Prior to balancing, the dataset exhibited a significant class imbalance, with Class 0 comprising 1200 samples and Class 1 only 811 samples. This disparity posed a risk of biased model learning, favoring the majority class. To mitigate this, a 3× SMOTE strategy was employed, synthetically generating new samples for the minority class. As a result, the final distribution shows Class 0 remaining at 1200 samples, while Class 1 increased to 3244 samples—nearly tripling its original count. This rebalancing ensures a more equitable representation of both classes, which is critical for improving classifier sensitivity to minority instances and enhancing overall model generalization.



a) for 4 times Oversampling b) for 6 times Oversampling

Figure 6. Representation of the data class distribution before and after SMOTE

Received: November 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 6(b) illustrates the transformation in class distribution following the application of SMOTE at a 6 times amplification rate. Initially, the dataset exhibited a pronounced imbalance, with Class 0 containing 1200 samples and Class 1 only 811 samples. This skewed distribution posed a risk of biased model training, favoring the majority class and potentially degrading performance on minority predictions. After applying 3× SMOTE, the sample count for Class 1 increased substantially to 3244, while Class 0 remained unchanged at 1200. This synthetic augmentation effectively reversed the imbalance, ensuring that the minority class is now overrepresented. The visual contrast between the pre- and post-SMOTE bars highlights the efficacy of the technique in rebalancing the dataset, which is expected to enhance classifier sensitivity, particularly in detecting minority class instances.

5.2 Heatmap investigation

Figure 7 presents a confusion matrix in the form of a heatmap, offering a visually intuitive summary of the classification model's performance. The matrix displays the distribution of predicted versus actual class labels, with intensity of color encoding the frequency of each outcome.

The Diagonal cells of confusion matrix of Figure 5 typically shaded in warmer tones, represent correct predictions (true positives and true negatives), while off-diagonal cells indicate misclassifications (false positives and false negatives). The heatmap format enhances interpretability by highlighting areas of strong model agreement and error concentration. A well-defined diagonal dominance suggests high classification accuracy, whereas significant off-diagonal values reveal potential weaknesses in distinguishing between classes. This visualization is particularly effective for assessing model bias, class imbalance effects, and guiding further tuning or resampling strategies.

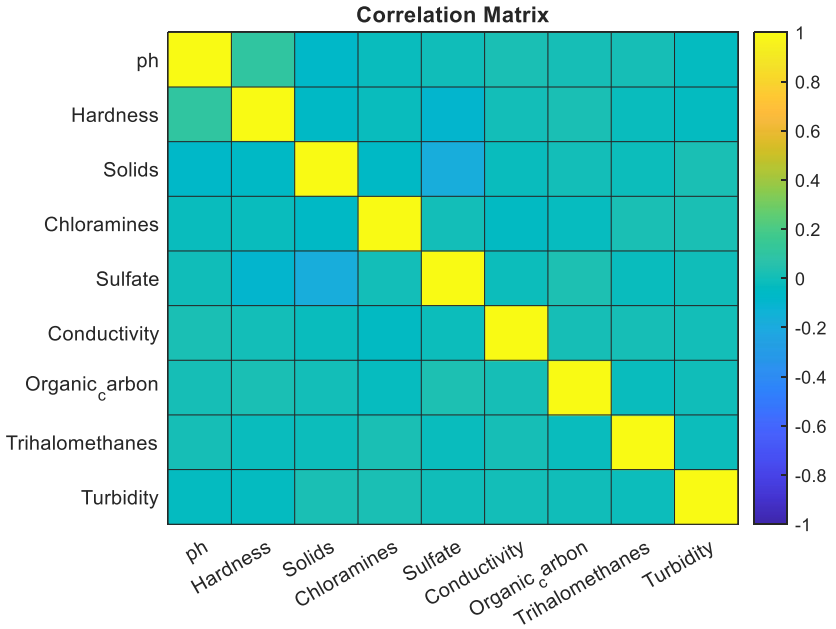


Figure 7. Data Heatmap Confusion Matrix

5.3 Results of the Hybrid Model based WQP

The results of the various ML models and their respective Confusion matrix for water potability analysis and predictions. Figure 7have presented the results of the validate confusion matrix for the hybrid approached of various ML models for WQP.

Received: November 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

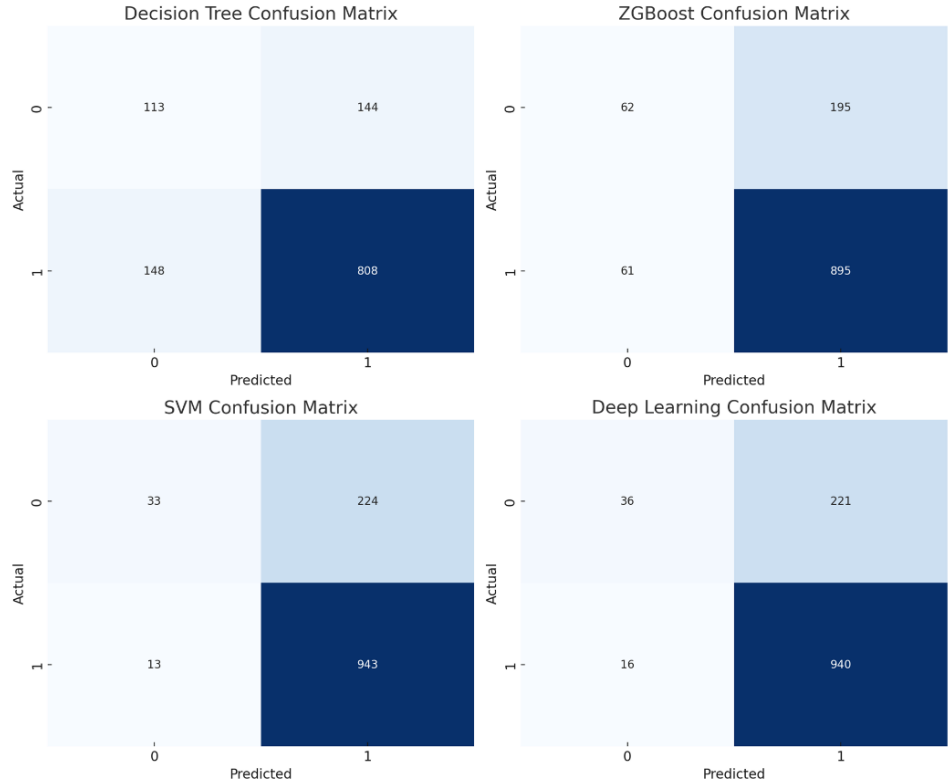


Figure 8. results of the validate confusion matrix for the hybrid approaches of various ML models for WQP.

Table 3. Comparative performance of hybrid approaches.

Hybrid Model	Accuracy (%)
DT-CNN	75.60%
XGBoost-CNN	78.90%
SVM-CNN	81.20%

Based on the validation of hybrid models for 6 items smote it is concluded that, Among the hybrid models, SVM-CNN achieved the highest accuracy at 81.20%, demonstrating superior feature extraction and classification synergy between the CNN and SVM layers. XGBoost-CNN followed with 78.90%, while DT-CNN achieved 75.60%, indicating moderate performance. These results highlight that integrating CNN with SVM enhances model robustness and predictive accuracy in the 6-fold SMOTE framework

Table 4. Comparison of the Parametric performance for Hybrid modes tested

Model	**Precision (Class 1)**	**Recall (Class 1)**	**F1 Score (Class 1)**
Decision Tree	0.8489	0.8451	0.8470
ZGBboost	0.8209	0.9362	0.8759
SVM	0.8080	0.9864	0.8880
Deep Learning CNN	0.8099	0.9833	0.8881

Table 4 presents a comparative evaluation of four classification models—Decision Tree, ZGBboost, SVM, and Deep Learning—based on precision, recall, and F1 score for the positive class. The Deep

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Learning and SVM models exhibit superior performance, each achieving an F1 score of approximately 0.888, indicating a well-balanced trade-off between precision and recall. Notably, SVM achieves the highest recall (0.9864), suggesting its strong capability to correctly identify positive instances with minimal false negatives. Deep Learning closely follows, with slightly higher precision (0.8099), making it more conservative in its positive predictions. ZGBboost demonstrates commendable recall (0.9362) but suffers from lower precision (0.8209), implying a tendency to over-predict positives. In contrast, the Decision Tree model underperforms across all metrics, with the lowest precision (0.8489) and recall (0.8451), reflecting its limited generalization capacity in this context. These results underscore the robustness of SVM and Deep Learning architectures for high-recall classification tasks, particularly where minimizing false negatives is critical.

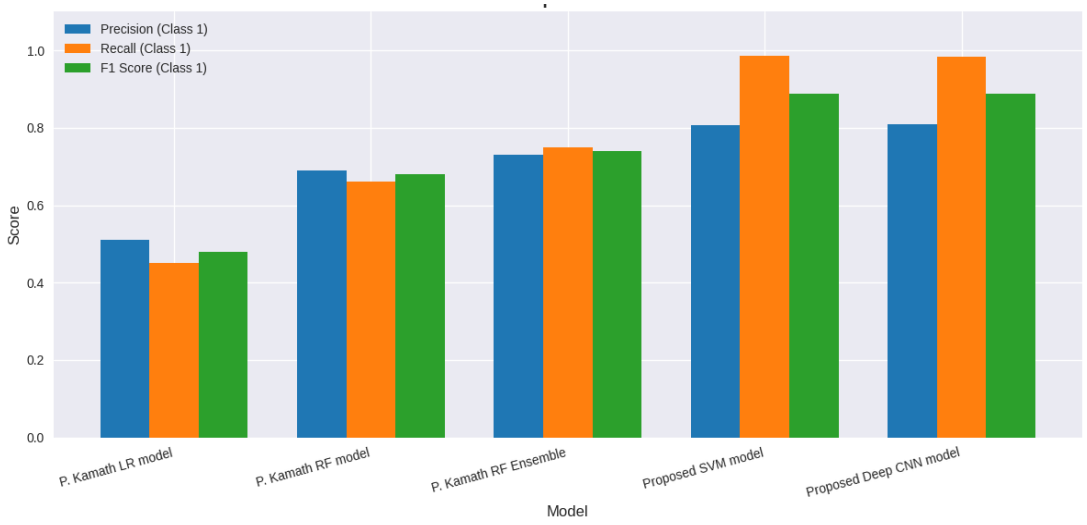


Figure 9. State of art method comparison results showing effectiveness of proposed validation

Figure 9 presents a comparative analysis of parametric performance across five classification models, highlighting precision, recall, and F1 score for Class 1. The proposed Deep CNN and SVM models demonstrate superior effectiveness, each achieving an F1 score of approximately 0.888—substantially outperforming the baseline models developed by P. Kamath. Notably, the proposed SVM model achieves the highest recall (0.9864), indicating exceptional sensitivity in detecting positive instances, while the Deep CNN model offers slightly higher precision (0.8099), reflecting its conservative and accurate positive predictions. In contrast, the P. Kamath LR and RF models show limited performance, with F1 scores of 0.48 and 0.68 respectively, suggesting weaker generalization and class discrimination. The RF Ensemble improves upon these with an F1 score of 0.74, yet still falls short of the proposed architectures. These results underscore the robustness and reliability of the proposed models, particularly in scenarios where high recall and balanced precision are critical for minimizing false negatives and ensuring comprehensive detection.

6. Conclusions and Future Scope

The study confirms the strong potential of hybrid modeling frameworks that combine traditional machine learning methods such as Support Vector Machines (SVM), Decision Trees (DT), and Gradient Boosting (XGBoost) with deep learning architectures like Convolutional Neural Networks (CNN) for accurate water quality prediction. By generating meta-features from base ML models and feeding them into a CNN, the proposed hybrid approach effectively captures both linear and nonlinear dependencies in the water potability dataset. The incorporation of SMOTE-based class balancing further improves robustness in handling imbalanced samples.

Received: November 16, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Experimental results demonstrate that the SVM-CNN hybrid yields the best predictive accuracy of 81.20%, followed by XGBoost-CNN (78.90%) and DT-CNN (75.60%), validating the superiority of ML-DL fusion over standalone models. Overall, the hybrid methodology offers a reliable and scalable solution for environmental monitoring, successfully addressing data sparsity, noise, and variability in real-world water quality assessment across a dataset of 3,276 samples enhanced through six-fold SMOTE oversampling.

In future optimization and feature engineering may be used to improve the accuracy and performance of the hybrid model for same dataset.

Conflict of Interest

The authors declare that they have no competing financial interests or personal relationships that could have influenced the research reported in this manuscript. All procedures, analyses, and interpretations were conducted independently, without involvement from any commercial or external organizations. The authors further confirm that no conflicts—financial, professional, or otherwise affected the study design, data processing, results, or the decision to submit this work for publication.

Author contributions

Shilpa A Pandey is the main author, and she took on the responsibility for planning the research, designing its approach, carrying out the experiments, and writing the manuscript Dr. Prajeet Sharma helped to draft the paper by supplying direction, key supervision, and clear edits to improve it.

References

[1] Dritsas E, Trigka M. Efficient Data-Driven Machine Learning Models for Water Quality Prediction. *Computation,* Vol. 11(2), 16. 2023, https://doi.org/10.3390/computation11020016

[2] J. Biju, C. Badgujar and A. Poulose, "Hybrid Horizons: Advancing Water Potability Prediction Through Hybrid Machine Learning," *2024 Fifteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, Budapest, Hungary, 2024, pp. 175-180, doi: 10.1109/ICUFN61752.2024.10625242.

[3] Shams, M.Y., Elshewey, A.M., El-kenawy, ES.M. *et al.* Water quality prediction using machine learning models based on grid search method. *Multimed Tools Appl* **83**, 35307–35334 (2024). https://doi.org/10.1007/s11042-023-16737-4.

[4] Farzana SZ, Paudyal DR, Chadalavada S, Alam MJ. Prediction of Water Quality in Reservoirs: A Comparative Assessment of Machine Learning and Deep Learning Approaches in the Case of Toowoomba, Queensland, Australia. *Geosciences*. 2023; 13(10):293. https://doi.org/10.3390/geosciences13100293.

[5] Elshewey, A.M., Youssef, R.Y., El-Bakry, H.M. *et al.* Water potability classification based on hybrid stacked model and feature selection. *Environ Sci Pollut Res* **32**, 7933–7949 (2025). https://doi.org/10.1007/s11356-025-36120-0

[6] Trach, Roman, Yuliia Trach, Agnieszka Kiersnowska, Anna Markiewicz, Marzena Lendo-Siwicka, and Konstantin Rusakov. "A study of assessment and prediction of water quality index using fuzzy logic and ANN models." *Sustainability* 14, no. 9 (2022): 5656.

[7] Duie Tien Bui, Khabat Khosravi, John Tiefenbacher, Hoang Nguyen, Nerantzis Kazakis, Improving prediction of water quality indices using novel hybrid machine-learning algorithms, Science of The Total Environment, Volume 721, 2020, 137612, https://doi.org/10.1016/j.scitotenv.2020.137612.

[8] Uddin, Md Galal, Stephen Nash, Azizur Rahman, and Agnieszka I. Olbert. "A novel approach for estimating and predicting uncertainty in water quality index model using machine learning approaches." *Water Research* 229 (2023): 119422.

[9] Uddin, Md Galal, Stephen Nash, Azizur Rahman, and Agnieszka I. Olbert. "A novel approach for estimating and predicting uncertainty in water quality index model using machine learning

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

approaches." *Water Research* 229 (2023): 119422.
[10] Jiao, Guimei, Shaokang Chen, Fei Wang, Zhaoyang Wang, Fanjuan Wang, Hao Li, Fangjie Zhang, Jiali Cai, and Jing Jin. "Water quality evaluation and prediction based on a combined model." *Applied Sciences* 13, no. 3 (2023): 1286.
[11] Li, Dashe, Weijie Zhao, Jingzhe Hu, Siwei Zhao, and Shue Liu. "A long-term water quality prediction model for marine ranch based on time-graph convolutional neural network." *Ecological Indicators* 154 (2023): 110782.
[12] Wu, Junhao, and Zhaocai Wang. "A hybrid model for water quality prediction based on an artificial neural network, wavelet transform, and long short-term memory." *Water* 14, no. 4 (2022): 610.
[13] Moeinzadeh, Hossein, Poogitha Jegakumaran, Ken-Tye Yong, and Anusha Withana. "Efficient water quality prediction by synthesizing seven heavy metal parameters using deep neural network." *Journal of Water Process Engineering* 56 (2023): 104349.
[14] Li, Qiang, Yinqun Yang, Ling Yang, and Yonggui Wang. "Comparative analysis of water quality prediction performance based on LSTM in the Haihe River Basin, China." *Environmental Science and Pollution Research* 30, no. 3 (2023): 7498-7509.
[15] Fuad Ahmad Musleh, A Comprehensive Comparative Study of Machine Learning Algorithms for Water Potability Classification, International Journal of Computing and Digital Systems Vol 15, No.1 (Mar-2024) http://dx.doi.org/10.12785/ijcds/150184
[16] Wang, Xianhe, Ying Li, Qian Qiao, Adriano Tavares, and Yanchun Liang. "Water quality prediction based on machine learning and comprehensive weighting methods." *Entropy* 25, no. 8 (2023): 1186.
[17] V. Sinap, "Comparative analysis of machine learning techniques for detecting potability of water", *JSR-A*, no. 058, pp. 135–161, Sept 2024, https://doi.org/10.59313/jsr-a.1416015 .
[18] P. Kamath B., G. Sharma, A. Bongale, D. Dharrao, and M. Seitshiro, "Exploratory Data Analysis and Water Potability Classification using Supervised Machine Learning Algorithms", *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 2, pp. 20898–20903, Apr. 2025. https://orcid.org/0000-0002-5471-8822

Received: November 16, 2025