International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 259-281
https://doi.org/10.70917/ijcisim-2026-4402
Article

A Hybrid Deep Learning–SVM Framework for Water Quality Prediction Using CNN Feature Extraction and Data Balancing

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

[1]Department of Computer Science and Engineering, LNCT University, Bhopal, MP 462022 India
[2] Department of Computer Science and Engineering, LNCT University, Bhopal, MP 462022 , India
Email: [1] shilpa.agnihotri28@gmail.com [2] prajeets@lnctu.ac.in

Abstract: Prediction of water quality (WQP) and the safety of drinking water is one of the most difficult tasks given the intricate interrelationship of physicochemical and biological factors, the imbalance in the data, and the other features within the dataset. The proposed approach fuses deep learning techniques and machine learning to establish a hybrid method. In particular, the input features are extracted using convolutional neural networks (CNNs) from the input data with 64 features first, followed by the classification using an RBF kernel as the input on support vector machine (SVM). Pre-processing steps include data cleaning through imputation of both moving mean and global mean. The data is addressed when the data is imbalanced using SMOTE method. This reduces the imbalanced distribution from 7:1 to nearly 1:1 after the algorithm's operation. Then the predictions made are checked for success with a pair of benchmarks, where the general prediction of water quality (WQP) has 20 predictors as well as the predicted potable water has a total of nine predictors, respectively. For comparison, SMOTE-based techniques greatly enhance baseline algorithms effectiveness. Our proposed hybrid technique applied as combination of CNN-RF and CNN-SVM techniques performs better, which is evident from (the WQP dataset), whose accuracy is 96.7% and 97.25%, while the accuracy of the water potability dataset using the same techniques is 82% and 82.87%.
Keywords: Water Quality Prediction, Hybrid CNN-SVM, Water Potability, Machine Learning, Deep Learning, Radial Basis Function

1. Introduction:

Water quality prediction (WQP), is essential for humanity it's honestly fundamental if we want to protect the environment and ourselves, and use our water resources responsibly. Every year, more pollutants seep into rivers and lakes water. It is since, more people, more factories, more stuff being produced in recent times. That makes it even more important to keep an eye on water quality, so it don't end up facing nasty consequences. For ages, people have relied on collecting samples and running lab tests to figure out what's in the water. Sure, these methods work. They're tried and true, but let's be real, they're slow and take a lot of effort. It needs skilled technicians, and everything has to be done carefully. That means water quality reports come with a pretty big delay, and by the time you get results, things might already have changed. Thus, it required smart AI solution to predict WQP. As shown in Figure 1, this is an attempt at visualizing the evolution of water quality prediction systems due to the rise in environmental threats. The first section of the figures shows the difficulties caused by human activities (industrialization, urbanization, intensive agriculture), the problems associated with lowering the quality of surface water types (rivers, lakes) and subsurface water. Consequently, the pollution levels and water condition deteriorate, and in order to control and predict the changes, it is necessary.

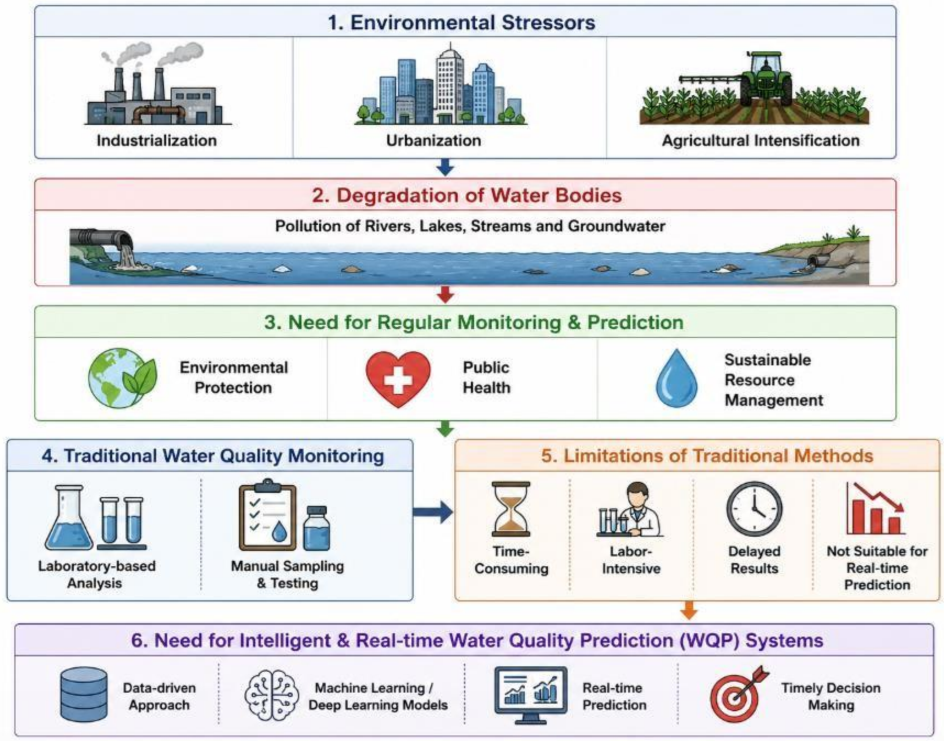


Figure 1: Major challenges and issues for continuous and accurate WQP requirement

This portion of the schematic reflects on the relevance of these findings to ecological protection and public health, as well as sustainable resource management. As in Figure 1, the traditional water quality control techniques used include laboratory analysis and sampling. While these strategies have enjoyed great fame for reliability and precision, they may fall short of what is needed in today's world. That gap can be fatal in some scenarios where there needs to be an answer to pollution incidents and environmental threats. Current research works such as a recent trend in the literature indicate that machine learning (ML) and deep learning (DL) methods are taken as possible solutions to solving this problem. ML and DL algorithms have been capable of addressing the intricate relationships among various water quality parameters including pH, dissolved oxygen, turbidity, temperature, and biochemical oxygen demand (BOD). Popular conventional ML algorithms such as Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbours are commonly used because of their robustness and capability to withstand changes in input data. Out of these algorithms, SVM has become one of the most popular approaches used to solve environmental problems. Especially in the case of complex high-dimensional data, applying SVM performs very well for classification and regression. Additionally, SVM also is not complicated over computations and ensures that maximum margin optimization will prevent overfitting.

But most traditional ML methods rely on manual feature selection methods, which might not be able to detect complex spatial or hidden dependencies and characteristics in water quality datasets. To tackle this problem, deep learning models, with CNNs in particular, have become increasingly popular in recent years. CNNs can capture hierarchical structures on-the-fly from the input by automatic feature extraction and could raise prediction accuracy with less need for manually generated features. Because CNNs are designed for image classification, they can be

used for structured environmental data as well, if they are properly adapted. A good deep learning architecture is trained on a large amount of data in an appropriately balanced way. But as mentioned above, that balance is rarely found in practical water quality databases.

Usually, one has far more data on common conditions than on unusual cases, which may include, for instance, specific classes of pollution. In this case, the machine learning technique just disregards less frequently occurring samples and trains on more frequently occurring instances. This reduces system efficiency owing to its inability to manage classes that appear less frequently. To address this issue, there are different ways of balancing the classes, including SMOTE, random over- or under-sampling.

This research presents a combined deep learning-SVM approach for WQP based on CNNs to extract the features and SVMs for the classification process, in accordance with these problems and possibilities. The goal is to label the multi-feature water quality data as safe or unsafe quality using a binary classification problem. This hybridization takes advantage of deep learning networks for representation learning and SVMs for their excellent generalization capabilities, which increases prediction performance and robustness.

Research presents a comprehensive hybrid predictive model that integrates deep learning and classical ML in depth to improve the efficiency of water quality assessment. Our approach mainly combined automatic feature extraction based on CNN and an optimized classifier based on SVM and exploit the complementary strengths of both models. The CNN architecture is responsible to learn complex, non-linear, and high-level feature representations on the raw input data. Thanks to its hierarchical structure and fully connected layers, the network can produce compact yet informative 64-dimensional feature vectors, which are capable of capturing inherent trends and underlying relationships between water quality parameters. The deep feature embeddings are then exploited as inputs for the SVM classifier, which is provided with an RBF kernel to process the nonlinear decision boundaries.

In addition, the SVM method also automatically selects kernel scales and standardizes features for high margin separation while improving generalization effectiveness on diverse data. Not only does this integrate the process, but this also helps improve classification accuracy and reduce the risk of overfitting, which is a typical issue in a standalone deep learning model, especially when trained with small datasets. Generalizability is challenged on two different water quality data repositories. Moreover, the study handles a significant difficulty found in practical environmental datasets, which is the class imbalance. To mitigate this challenge, the SMOTE approach is used as a data balancing methodology. Synthetic samples generated by SMOTE for classes having less information or a unique feature category in datasets are likely to give an even distribution of data to those classes, thus enabling lower bias in these models and more reliable and robust prediction in specific categories of water quality conditions. The novelty of the proposed framework is realized by combining the CNN derived deep feature representations with an optimized SVM classifier along with an efficient data balancing method. Additionally, the scheme includes a solid data preprocessing pipeline with a two-stage missing value imputation technique mixing moving mean and global mean approaches.

Framework provide a customized data augmentation to improve model data consistency, which is implemented using the SMOTE method and synthetic samples create in order to balance the class distributions of the dataset by augmenting the minority class. Its lightweight CNN architecture is computationally efficient, allowing the model to effectively address real-time and resource-constrained applications like IoT-based environmental monitoring systems. The hybrid

model in contrast to the traditional single model or the only system with the single handcrafted model or the end-to-end deep learning, has shown that the model makes both the feature learning and the classification efficiency trade-off in balance. This leads to relatively better prediction accuracy and capability and a better performance on generalization ability and resilience towards the noise and imbalance introduced by the given dataset. Our idea provides a novel technique in that it is scalable, accurate and robust to handle real-time water quality prediction, supporting real-time environment monitoring and decision-making scenarios.

For the rest of paper, it will first describe the associated works for hybrid DL based research in the section 2. In section 3 the two completely different water quality databases, which are used in the generalizability test used in this study, have been described. The proposed data processing and feature extraction DL framework is presented in section 4. The proposed method for safe and unsafe water quality classification with SVM-RBF-based optimization is presented in section 5. Section 6 will qualitatively and quantitatively present the results of the sequential ML based classifiers as well as the proposed hybrid Approach. The main conclusions, scope and discussions are presented later on.

2. Literature Review

WQP has become an important domain of research owing to growing environmental contamination from pollution, industrialization, and the demand for sustainable management of natural water resources. Several studies based on ML and DL have been done recently to try improving the classification performance. This section has sequentially reviewed the most topical works on each of these topics. Huang et al. [1] (2025) introduced a hybrid deep learning and machine learning architecture for urban water quality prediction. Their approach mixes feature extraction via deep neural networks with the final prediction with traditional ML models enhancing both accuracy and robustness. Nevertheless, the model is computationally intensive and relies on large-scale datasets for better performance. Muhammad et al. [2] (2022), used a Support Vector Machine (SVM) for drinking water classification. The procedure of pre-processing water characteristics and SVM classification is used. Although the model performed well, sensitivity towards kernel choice and tuning of parameters resulted in a generalization effect. Zaky et al. [3] (2023) used a Decision Tree (DT) classifier to predict water potability. The model is interpretable and simple to apply, but it is prone to overfitting and instability when dealing with noisy or imbalanced data. Sangwan and Bhardwaj [4] (2024) established a full ML scheme, using several classifiers and feature engineering. Their methods allow for comparison and selection towards better predictions. It is, however, computationally expensive to implement, and there is a lack of real-time applicability. Kumar and Jeevanantham [5] (2024) described a review on hybrid deep learning frameworks employing optimized feature selection. This work emphasizes the significance of feature optimization approaches. The limitation is that it is theory-based with very little experimental validation. Afzal et al. [6] (2024) introduced a machine learning (ML)-based predictive modeling strategy for water quality assessment. This involves preprocessing, feature selection, and comparative analyses of models. Performance is impacted by dataset imbalance and scalability limitation. Zhu et al. [7] (2025) proposed a hybrid deep learning model that leverages CNN-based feature extraction and sequence modeling. This model correctly incorporates spatial-temporal dependencies. Despite improved accuracy, it entails higher computational cost and requires extensive training data.

Karthick et al. [8] (2024) proposed a data-driven ML strategy using data augmentation techniques. It increases the diversity and robustness of the dataset. Nevertheless, the synthetic

data generating approach could be biased and lead to unrealistic patterns. Thota et al. [9] (2024) proposed the CNN-SVM and CNN-RF hybrid models for water-based microplastic detection. The pairing adds accuracy to classification. However, the model suffers from complex architecture design and longer training time. Abuzir and Abuzir [10] (2022) used machine learning classifiers for water quality classification. The methodology compares several models and indicates the success of ensemble methods. It has disadvantages such as lack of real-time implementation and dependency on dataset issues. Alex and Brindha [11] (2025) have also introduced a hybrid preprocessing pipeline with wavelet denoising and ensemble ML models. This enhances the data quality prior to the classification process. The preprocessing pipeline, however, imparts computational overhead and adds more complexity. Tverdokhlib et al. [12] (2022) presented an analytical and ML-based approach to the water suitability for consumption. Although it is effective, the method suffers from reliance on defined thresholds and domain-specific tuning. Ahmed et al. [13] (2019) proposed supervised ML models to efficient WQP that would be primarily based on regression and classification methods. Although the study shows well predictive ability, it does not consider temporal dynamics. Haghiabi et al. [14] (2018) utilized ML systems including ANN and regression model for water quality prediction. While proven to be effective, the models depend on manual feature engineering and are also noise sensitive. Ahmed et al. [15] (2019) have evaluated several ML approaches such as RF, ANN, and SVM and demonstrated improved prediction accuracy. Nevertheless, issues of data preprocessing and parameters tuning are raised in the research.

Azrour et al. [16] (2022) presents a comparative study of ML algorithms for WQP and finds ensemble models that outperform individual classifiers. However, they are still computationally expensive and lack interpretability. Zrouri and El Farissi [17] (2024) investigated the utilization of IoT and ML for predictive water quality monitoring. The approach combines real-time sensor observations and cloud prediction. Despite this, concerns about sensor noise, communication delay, and security threats persist. Kaddoura [18] (2022) reviewed ML algorithms for predicting sustainable drinking water quality. The study focuses on sustainability but suffers from limited dataset diversity and generalization issues. Bhalerao and Mohan [19] (2023) developed ML based water quality monitoring models with different classifiers. Good though the research was, it does not have the benefit of being deployed into real-time systems. Suwansukho and Wutthithanyawat [20] (2024) applied deep neural network (DNN) in edge computing devices to perform real-time water quality testing. This minimizes latency, however, limitations on hardware and processing capabilities apply. Anand et al. [21] (2023) applied CNN for predicting water quality. The model automatically features are extracted which is enhanced performance. Yet CNN based models requires big data and high computational capabilities.

Yan et al. [22] (2024) reviewed the ML methods for WQP in detail within the past five years. The study reveals trends and some issues but not experimental verification. Luthra et al. [23] (2024) applied ML techniques on the water quality prediction using the prediction of the class classification accuracy enhancement. The method works well, but is constrained by the performance from specific datasets. Helaly et al. [24] (2025) introduced ML and DL developments in WQP, and highlighted hybrid methods as the most promising ones. But difficulties such as system complexity and interpretability still go unresolved. Islam et al. [25] (2025) proposed an IoT-based intelligent water quality monitoring system with ML model integration. Though it allows for real-time monitoring, sensor reliability and scaling of a system's operations becomes an issue. Cardia et al. [26] (2026) proposed a multivariate ML-based estimation method for water quality estimation. Using statistical and ML, we can use them. With the exception of one specific case, this is a great but poorly done modeling that relies on quality multivariate data and lack the contextual environment adaptability. Review overview is summarized in Table 1.

Author and Reference	**Methodology + Dataset**	**Performance**	**Limitations**
Huang et al. [1] (2025)	Hybrid DL + ML using urban water datasets	95% accuracy	High computational cost
Muhammad et al. [2] (2022)	SVM classifier using drinking water dataset	91–99% accuracy	Sensitive to kernel tuning
Zaky et al. [3] (2023)	Decision Tree using water potability dataset	54.3% accuracy	Overfitting, instability
Sangwan and Bhardwaj [4] (2024)	Multi-ML framework using water quality dataset	RF 92– SVM 96%	High complexity, no real-time
Abuzir and Abuzir [10] (2022)	ML classifiers using water dataset	68 % accuracy	No real-time system
Alex and Brindha [11] (2025)	Wavelet + Ensemble ML using water potability dataset	99% accuracy	High preprocessing cost
Ahmed et al. [13] (2019)	Supervised ML using water dataset	85.7 % accuracy	No temporal modelling
Haghiabi et al. [14] (2018)	ANN + Regression using water dataset	~88–93% accuracy	Manual feature engineering
Azrour et al. [16] (2022)	ML comparison using water dataset	~90–95% accuracy	Low interpretability
Kaddoura [18] (2022)	ML models using drinking water dataset	73.1 % accuracy	Limited generalization
Bhalerao and Mohan [19] (2023)	ML models using water dataset accuracy is maximum for ANN	82–97% accuracy	No deployment
Anand et al. [21] (2023)	CNN using water dataset accuracy of 80^	70-80% accuracy	Needs large data
Luthra et al. [23] (2024)	ML models using water dataset	70.01 % Score	Dataset dependency
Cardia et al. [26] (2026)	Multivariate ML using multivariate dataset	94–97% accuracy	Needs high-quality data

Table 1 Summary of Review and Limitations Problem statements and gaps

According to the reviewed papers (Table 1), a number of the common limitations and research questions hamper the development of a highly consistent water quality prediction (WQP) system and encourage the need for advanced mix between methods. Figure 2 shows a diagram with circular clusters summarizing the main restrictions and research needs for Water Quality Prediction (WQP).

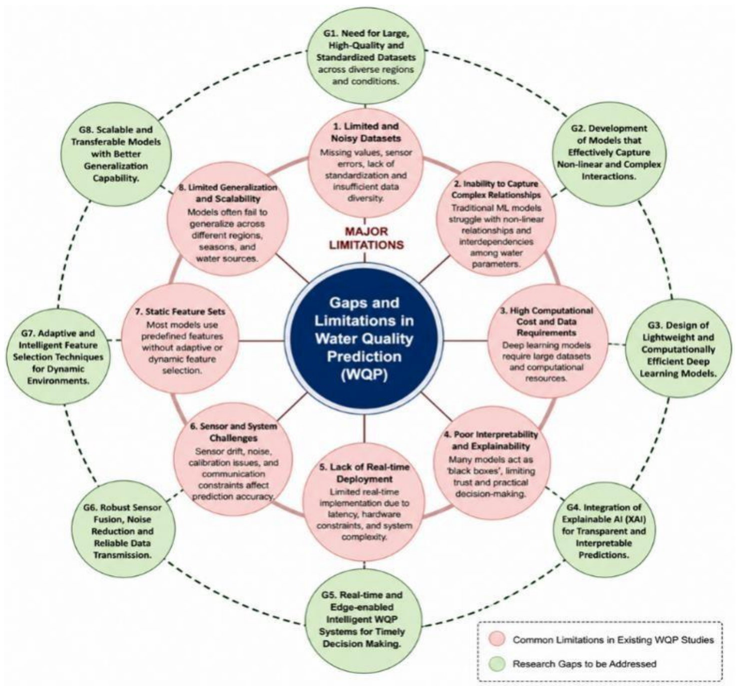


Figure 2 Research Gaps and challenges of WQP problem

The inner circle points out some of the key limitations (noisy and limited datasets, inability to capture complex relationships, high computational cost, poor interpretability, lack of real-time deployment, sensor-related issues, static feature selection, and weak generalization). By contrast the outer ring highlights comparable research gaps around large and standardized datasets, advanced nonlinear modeling, lightweight deep learning architectures, explainable AI integration, real-time edge-enabled systems, robust sensor fusion, adaptive feature selection, and scalable models. The diagram can indeed be seen that the above-mentioned interconnected gaps of the WQP systems can be solved to develop efficient, accurate and intelligent hybrid WQP systems.

Another prevalent challenge is the absence of real-time deployment capability, which is particularly concerning in IoT-based systems, where the effect of latency, sensor drift, and communication constraints all interfere with prediction accuracy. What's more, most of the existing models are static, without either adaptive or dynamic feature selection, making it hardly possible for them to respond to the dynamic changes in the environment. On a research side, there is a glaring need for lightweight, scalable, and explainable hybrid framework designs combining ML, DL, and optimization techniques. The future work can include robust preprocessing of the data (e.g., denoising and imputation), intelligent feature selection, ensemble learning, and edge-enabled real-time processing, and the application of explainable AI (XAI) for trust and decision-making. Filling these gaps can give rise to efficient, adaptive, and high-accuracy hybrid models for the next generation water quality predictions.

3. Proposed WQP Methodology

Figure 3 shows the holistic structure of the proposed hybrid CNN–SVM method for water quality prediction which can be categorized into three basic main stages: data preparation, CNN feature learning and feature classification with evaluation. For the first phase, termed Data Preparation, raw water quality data is first loaded and pre-processed. This involves dealing with missing values, normalization and class balancing for quality and consistency of data. The

resulting data set is segmented into training and test datasets via the train–test split algorithm, for unbiased model assessment.

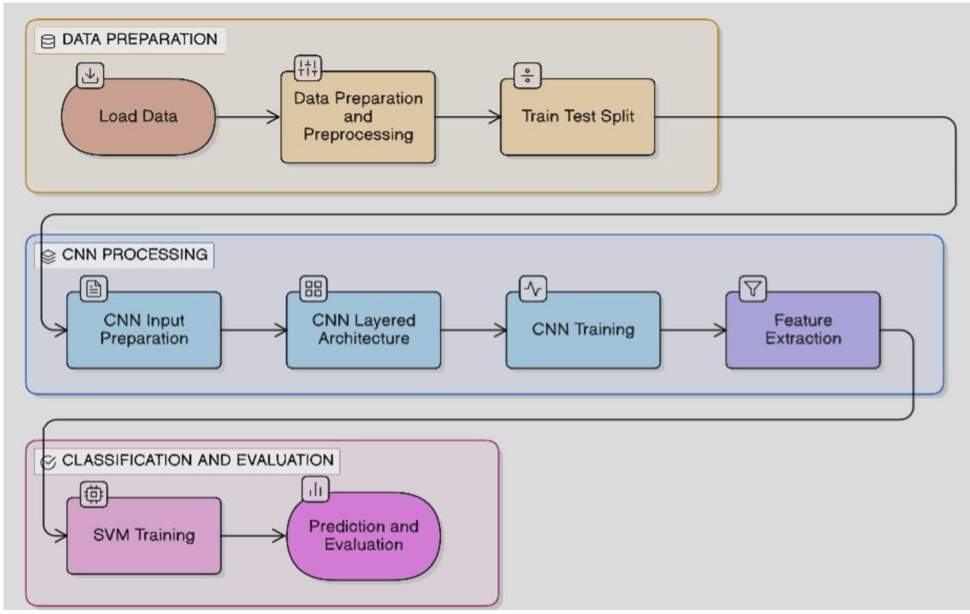


Figure 3 Proposed system block diagram for hybrid approach of WQP

In the next stage of the model pipeline, CNN Processing, deep features are extracted. The pre-processed input datasets are initially modified in a CNN fed form. It is subsequently fed through the described CNN layered architecture with convolutional layers, batch normalization, and activation functions, where intricate nonlinear relationships between water quality parameters were revealed. The network is implemented with labelled data, and the high-level discriminative features are created from the intermediary layers in a compact feature representation. The selected CNN features can be transferred into a SVM classifier in the last stage, Classification and Evaluation. The SVM with a suitable kernel (e.g., RBF) can be used for classification, by maximizing the margin among the water quality classes. The trained model is then applied to make predictions on test data and is evaluated based on its accuracy, precision, recall and F1-score. By using the deep feature learning capability of CNN + classification power of SVM, this hybrid model ensures better accuracy and generalization for water quality prediction.

4. Problem Formulation:

The dataset in the suggested water quality forecasting framework is defined as $X = \{x_i\}_{i=1}^{m} \in R^{m \times n}$ where m refers to the total number of water samples collected from various locations and n means the 20 physicochemical properties in WQP database, whereas the pH, turbidity, dissolved oxygen (DO), total dissolved solids (TDS), conductivity, and temperature for the water_potability database are given. A label $y_i \in \{1,2, \dots, k\}$ indicates a sample x_i; the name of it represents the water quality class (e.g., safe, moderate, polluted) or pre-defined Water Quality Index (WQI) categories.

$$\mathrm{X} = \{x_i\}_{i=1}^{m} \in \mathrm{R}^{m \times n}, \mathrm{y} = \{y_i\}_{i=1}^{m}, y_i \in \{1,2, \dots, k\} \quad (1)$$

Where m is the number of samples, n is the number of features, and k is the number of classes. Its goal is to understand a mapping:

$$f\colon \mathrm{R}^n \rightarrow \{1,2, \dots, k\} \quad (2)$$

So, it results in the lowest possible classification error. The key is to learn a predictive function

$f\colon R^n \rightarrow \{1,2, \dots, k\}$,, that maps the input feature space to the correct water quality class with low classification error. This formulation allows the construction of an intelligent model that can accurately evaluate water quality conditions based on observed environmental parameters, supporting real-time monitoring and decision-making in water resource management.

4.1 Data Preprocessing

For enhancing the prediction performance, the water quality dataset is pre-processed to eliminate missing values and to minimize data imbalance. This pre-processing is achieved in three steps as follows;

4.2 Missing Value Imputation

For our proposed water quality prediction model, missing data in the input features are handled using a two-stage imputation strategy to ensure data completeness and reliability. Initially, a local interpolation approach is applied using a moving mean window of size five.

For a missing feature value x_i,, it is replaced by the average of its neighbouring values, specifically from $x_{i-2}$to x_{i+2}. For each feature value x_i,:

$$x_i = \begin{cases} \frac{1}{5}\sum_{j=i-2}^{i+2} x_j, & \text{if } x_i \text{ is missing} \\ x_i, & \text{otherwise} \end{cases} \tag{3}$$

Remaining missing values are replaced by average:

$$x_i = \mu = \frac{1}{m}\sum_{j=1}^{m} x_j \tag{4}$$

This approach maintains local trends and continuity in the data, particularly important in the case

of environmental parameters of pH, turbidity, and dissolved oxygen that develop gradually. Where missing data still exist after this step (for instance, at boundary points or in very sparse areas), a global mean imputation is made. Here the missing value is replaced with the overall mean μ computed across all available samples. This two-level imputation process increases the data quality, mitigates bias, and guarantees that the dataset is suitable for successful training of machine learning models.

4.3 Data Normalization

Feature normalization is carried out in this framework for water quality prediction to guarantee equal contribution of all input variables to the learning of the model. The use of raw data directly can bias the model towards features with larger magnitudes as the dimensions of water quality parameters like pH, turbidity, dissolved oxygen, and conductivity may vary across units and range. To mitigate this, min–max normalization is done for each feature, which converts each one of the values to a normalized scale from 0 to 1 as.

$$x_i^{\text{norm}} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{5}$$

Concisely, we normalize every feature value using the difference between the minimum and maximum values of the dataset such that the smallest value is mapped to 0 and the largest value is mapped to 1. This scaling increases the numerical stability, speed of convergence and the performance of machine learning algorithms, particularly hybrid ones like CNN and SVM that can be sensitive to the scaling of the features.

4.4 Data Balancing (SMOTE)

In the proposed water quality prediction model, the class imbalance is solved by implementing the Synthetic Minority Over-sampling Technique (SMOTE) to enhance the learning capability of the classifier. For many real-world water datasets, certain classes (for example, polluted or unsafe water) have considerably fewer samples than others, resulting in biased model outputs. To mitigate this, SMOTE generates synthetic samples for the minority class through interpolation of existing samples.

$$\tilde{x}_i = x_i + \delta(x_{nn} - x_i),\ \delta \sim U(0,1) \tag{6}$$

A new sample as $\tilde{x}_i$ is established by instances from the minority class as $\tilde{x}_i$ and one of its nearest neighbors represented by x_{nn}. then generating a point along the line segment joining them with a random factor given as $\delta \in (0,1)$. This contributes a significant expansion towards the feature space of the minority class. For extreme imbalanced cases where there are very few samples available, synthetic data is generated by overlaying small Gaussian noise with the existing sample: $\tilde{x}_i = x_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$..

$$\tilde{x}_i = x_i + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \quad (7)$$

This prevents overfitting while maintaining variability in the dataset. SMOTE-based balancing improves classification performance, generalization, and ensures that all water quality classes are well represented in the dataset in training the model.

4.5 CNN based Features Extraction

In the proposed water quality prediction framework, the Convolutional Neural Network (CNN) is employed as a nonlinear feature extractor to learn complex patterns from the input data. Unlike traditional methods that rely on manually engineered features, the CNN automatically captures meaningful and discriminative representations from the input feature space.

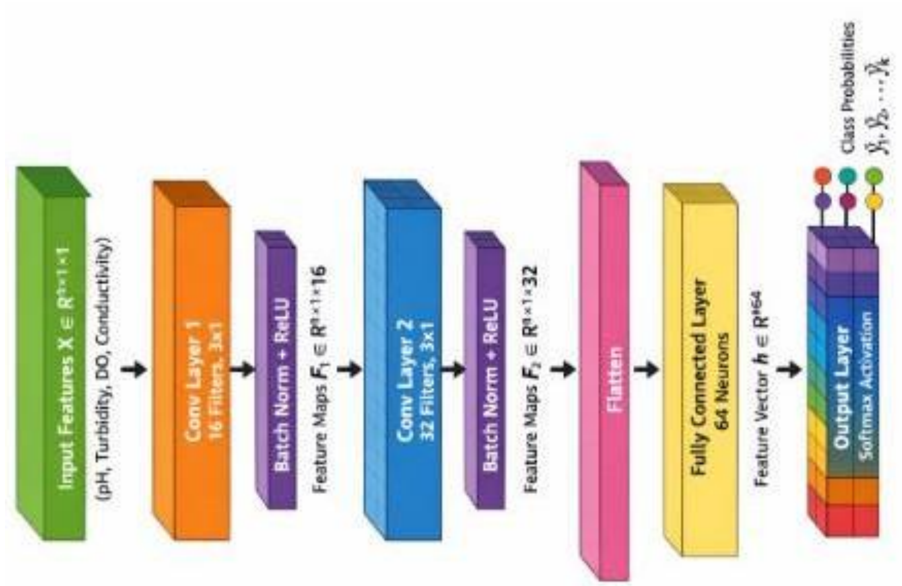


Figure 4 CNN layered architecture diagram used for DL feature extractions

Figure 4 illustrates the proposed Convolutional Neural Network (CNN) architecture designed for water quality prediction based on physicochemical input features such as pH, turbidity, dissolved oxygen (DO), and conductivity. The model begins with an input layer where the feature vector is represented in the form $R^{n\times1\times1}$. This is followed by the first convolutional layer consisting of 16 filters of size 3 × 1, which extracts initial feature patterns and generates feature maps of dimension $R^{n\times1\times16}$. A batch normalization and ReLU activation layer is then applied to enhance training stability and introduce nonlinearity. Subsequently, a second convolutional layer with 32 filters of the same kernel size further refines the learned representations, producing deeper feature maps of size $R^{n\times1\times32}$. Again, batch normalization and ReLU activation are employed to improve convergence and generalization. The output is then passed through a flattening layer, which converts the multidimensional feature maps into a one-dimensional vector.

This flattened vector is fed into a fully connected layer with 64 neurons, serving as the feature extraction layer that captures the most discriminative information from the input data. Finally, the extracted features are fed into an output layer with $k$neurons corresponding to the number of water quality classes. A softmax activation function is applied to generate class probabilities, where each output

$\hat{y}_j$represents the likelihood of the input belonging to class j. The network is trained using the cross-entropy loss function, which measures the difference between the predicted probabilities and the true class labels. This architecture effectively combines feature learning and classification, enabling accurate and robust water quality prediction. The overall architecture effectively combines convolutional feature learning and dense classification to achieve accurate and robust prediction of water quality.

Mathematically, the CNN performs a transformation;

$$\varphi_{\text{CNN}}: \mathrm{R}^n \rightarrow \mathrm{R}^d,$$

where the original n-dimensional input features are mapped into a lower-dimensional latent feature space of size $d = 64$. This transformation enables the model to encode nonlinear relationships among water quality parameters such as pH, turbidity, dissolved oxygen, and conductivity. The extracted feature vector provides a compact and informative representation of the input data, which is subsequently used by the classifier (e.g., SVM) for accurate prediction. This approach improves model robustness, enhances classification performance, and reduces the dependency on manual feature selection. The details of CNN feature layer is shown in Table 2.

Feature Source Layer	Layer Name	Number of Features	Description
Fully Connected Layer	Feature-Layer	64	Learned feature representation used as input to SVM classifier

Table 2 CNN Feature Extraction Layer Details

The features extracted from the CNN model are the activations obtained from the fully connected layer. The fully connected layer named feature-Layer generates a 64-dimensional feature vector, which serves as a compact and discriminative representation of the input data. These extracted features are subsequently used as input to the SVM classifier for improved WQP and classification performance.

Convolution Operation

In the proposed CNN-based feature extraction framework, the convolution operation is the fundamental building block used to learn spatial and feature-level representations from the input data. At each layer l, the feature map $F^{(l)}$ is obtained by applying a set of learnable filters $W^{(l)}$ to the output of the previous layer $F^{(l-1)}$, followed by the addition of a bias term $b^{(l)}$. This operation is mathematically expressed as:

$$F^{(l)} = \sigma\left(W^{(l)} * F^{(l-1)} + b^{(l)}\right)$$

where $*$denotes the convolution operation. The convolution process involves sliding the filters across the input feature map to capture local patterns and dependencies among the input features. After convolution, a nonlinear activation function is applied to introduce nonlinearity into the model. In this work, the Rectified Linear Unit (ReLU) activation function is used, defined as;

$$\sigma(z) = \max(0, z),$$

Which effectively retains positive values while suppressing negative ones. This not only helps in learning complex nonlinear relationships among water quality parameters but also improves computational efficiency and mitigates issues such as vanishing gradients

In the proposed water quality prediction model, the Convolutional Neural Network (CNN) is trained using the cross-entropy loss function, which is widely used for multi-class classification problems. This loss function measures the difference between the true class labels and the predicted probability distribution generated by the network. Mathematically, the loss is defined as

$$\mathcal{L}_{\text{CNN}} = \sum_{i=1}^{m} \sum_{c=1}^{k} y_{i,c} \log(\hat{y}_{i,c})$$

where m is total number of training samples and $k$is the number of classes. Term $y(i,c)$ represents the true label (one-hot encoded) for sample $i$belonging to class c, while $\hat{y}(i,c)$ denotes the predicted probability for that class obtained through the softmax function. The loss function is evaluated and is minimized for improving the accuracy of detection.

4.6 Hybrid CNN-SVM Classifier

In the proposed water quality prediction framework, the SVM classifier is applied on the high-level features extracted by the CNN. Instead of directly classifying raw input data, the CNN first transforms the original feature space into a more discriminative latent representation $h \in Rd$, capturing complex nonlinear relationships among water quality parameters. These learned features are then used as input to the SVM for final classification. The SVM aims to learn an optimal decision boundary in this transformed feature space, defined as:

$$f(\mathrm{h}) = \mathrm{w}^T\phi(h) + b$$

where w is defined as the weight vector, variable b is the bias term, and $\phi(\cdot)$ represents a nonlinear mapping induced by the kernel function (typically the Radial Basis Function, RBF). This mapping allows the SVM to separate data points that are not linearly separable in the original space. In this hybrid approach, the CNN acts as an automatic feature extractor, while the SVM serves as a powerful classifier that maximizes the margin between different classes. The SVM optimizes the decision boundary by solving a convex optimization problem, ensuring better generalization and robustness, especially for small or imbalanced datasets. By combining CNN's deep feature learning capability with SVM's strong classification performance, the proposed model achieves higher accuracy and improved reliability in water quality prediction tasks

4.7 Optimization Objective

In the proposed hybrid CNN–SVM framework, the SVM) is trained on the features extracted by the CNN to determine an optimal decision boundary. The objective of the SVM is to maximize the margin between different classes while minimizing classification errors. This is formulated as a convex optimization problem:

$$\min_{\mathrm{w},b,\xi} \frac{1}{2} \|\mathrm{w}\|^2 + C \sum_{i=1}^{m} \xi_i$$

subject to:

$$y_i(\mathrm{w}^T\phi(\mathrm{h}_i) + b) \geq 1 - \xi_i$$

Here, w represents the weight vector, $b$is the bias term, and ξ_i are slack variables that allow for misclassification of some samples. The parameter Ccontrols the trade-off between maximizing the margin and minimizing classification errors. A larger value of Cemphasizes correct classification, while a smaller value allows a wider margin with some misclassification. This formulation ensures robust generalization, especially when dealing with complex and noisy water quality data.

4.8 Radial Basis Function (RBF) Kernel

To handle nonlinear separability in the feature space, the SVM employs the Radial Basis Function (RBF) kernel, defined as:

$$K(\mathrm{h}_i, \mathrm{h}_j) = \exp(-\gamma \|\mathrm{h}_i - \mathrm{h}_j\|^2)$$

where $\gamma$is a kernel parameter that controls the spread of the Gaussian function. The RBF kernel implicitly maps the input features into a higher-dimensional space, enabling the SVM to construct nonlinear decision boundaries. This is particularly useful in water quality prediction, where relationships among parameters are often complex and nonlinear.

4.9 Hybrid Model Prediction

The overall hybrid model integrates deep learning and machine learning as:

$$f(x) = \mathrm{SVM}(\phi_{\mathrm{CNN}}(x))$$

In this formulation, the CNN first processes the input data $x$and extracts high-level feature representations ϕCNN(x). These features are then fed into the SVM classifier, which performs classification by identifying the optimal separating hyperplane. Thus, the CNN is responsible for automatic feature learning, while the SVM ensures optimal margin-based classification, leading to improved prediction accuracy and generalization. Once the model is trained, prediction is performed by assigning each input sample to the class with the highest posterior probability:

$$\hat{y}_i = \arg\max_c P(y = c \mid \mathrm{h}_i)$$

where h_i is representing the CNN-extracted feature vector for the i-th sample. The classifier outputs probabilities for each class, and the class with the maximum probability is selected as the final prediction. This approach ensures reliable and accurate classification of water quality categories.

4.10 Proposed Hybrid Models (CNN+SVM)

as clear from Figure 5 illustrating the complete workflow of the proposed hybrid CNN–SVM model for water quality prediction using deep learning techniques. The process begins with the decision to analyse water quality data, followed by loading the dataset. An initial check is performed to ensure that a valid dataset is available. The system then evaluates the presence of missing values; if detected, appropriate preprocessing techniques such as moving mean and global mean imputation are applied to handle incomplete data. The proposed sequential flow chart is illustrated in the Figure 5. The flowchart introduces an integrated pipeline that combines preprocessing of missing physicochemical data with intelligent feature learning to enable accurate and reliable water potability prediction

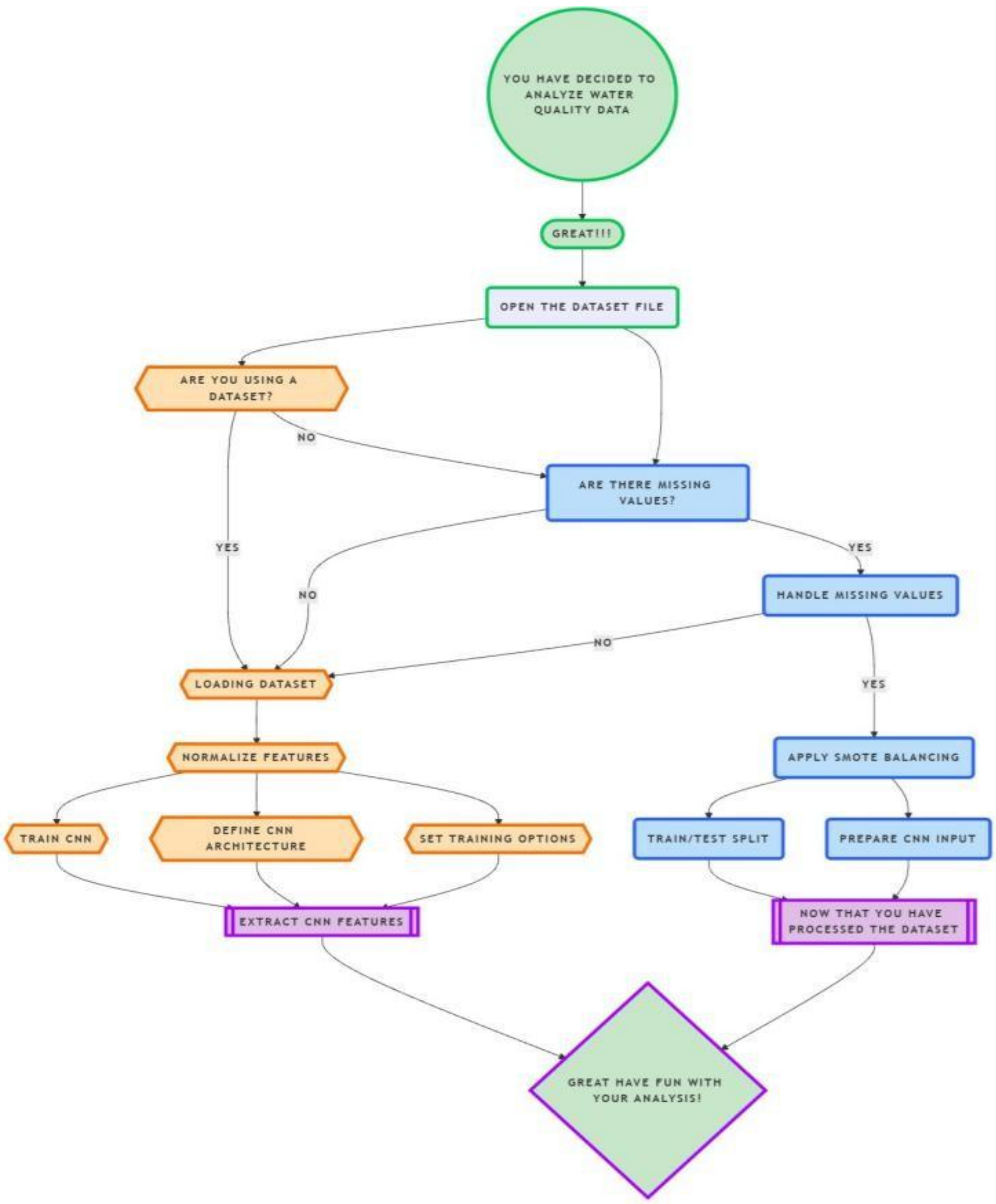


Figure 5 Flow Chart of the Proposed Hybrid CNN-SVM methodology for WQP

5. Results and Discussions

This research has presented the results of the water quality prediction or classifications. The two-distinct database are used one for water quality prediction and for water potability database. both are unique in their features.

Dataset Head:

	aluminium	ammonia	arsenic	barium	cadmium	chloramine	chromium	copper
0	1.65	9.08	0.04	2.85	0.007	0.35	0.83	0.17
1	2.32	21.16	0.01	3.31	0.002	5.28	0.68	0.66
2	1.01	14.02	0.04	0.58	0.008	4.24	0.53	0.02
3	1.36	11.33	0.04	2.96	0.001	7.23	0.03	1.66
4	0.92	24.33	0.03	0.20	0.006	2.67	0.69	0.57

	flouride	bacteria	...	lead	nitrates	nitrites	mercury	perchlorate
0	0.05	0.20	...	0.054	16.08	1.13	0.007	37.75
1	0.90	0.65	...	0.100	2.01	1.93	0.003	32.26
2	0.99	0.05	...	0.078	14.16	1.11	0.006	50.28
3	1.08	0.71	...	0.016	1.41	1.29	0.004	9.12
4	0.61	0.13	...	0.117	6.74	1.11	0.003	16.90

	radium	selenium	silver	uranium	is_safe
0	6.78	0.08	0.34	0.02	1
1	3.21	0.08	0.27	0.05	1
2	7.07	0.07	0.44	0.01	0
3	1.72	0.02	0.45	0.05	1
4	2.41	0.02	0.06	0.02	1

Table 3 screen shot of the WQP data features as head

The head of the WQP dataset is shown in Table 3, with 20 input features and 1 target variable. The 20 input features contain parameters from chemical indicators (aluminium, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, mercury, perchlorate, radium, selenium, silver, and uranium) and biological indicators (bacteria, lead, nitrates, and nitrites). All input features are extremely relevant as they cover various physical and chemical properties and microbiology of the water body, which ultimately affects the drinkability of water. The presence of heavy metals (arsenic, cadmium, mercury, lead), nitrates, and nitrites implies pollution by chemicals, whereas the concentration of bacteria indicates pollution due to biological contaminants. Chloramine and fluoride are associated with the water treatment process. The target variable is_safe (0/1), where 0 means water is unsafe, and 1 means water is safe for human consumption.

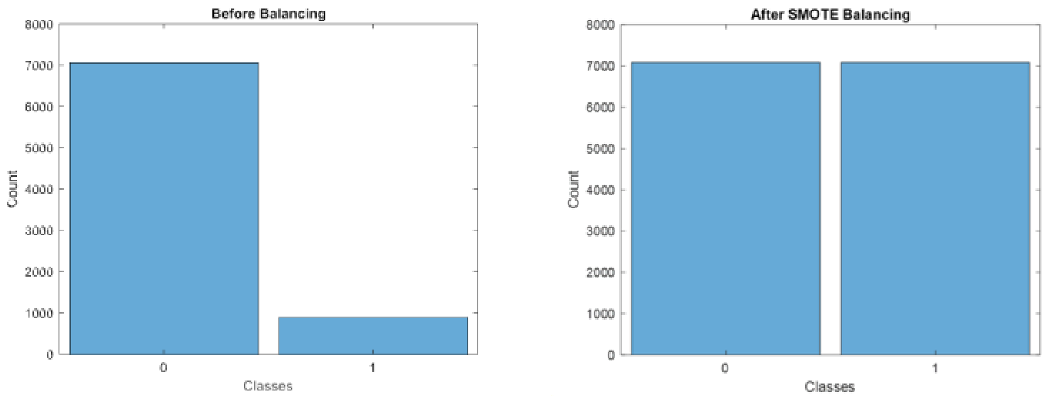
	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

Table 4 water potability dataset description as head

Table 4 provides a description of the water-potability dataset consists of key physicochemical parameters used to assess whether water is safe for human consumption. It includes nine input features such as pH, hardness, total dissolved solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, and turbidity, which collectively describe the chemical composition and quality of water. The target variable, Potability, classifies water as safe (1) or unsafe (0). The presence of some missing values indicates the need for preprocessing before model development. Overall, the dataset provides a balanced and practical representation of real-world water conditions, making it suitable for training machine learning models for accurate water quality prediction.

5.1 Results of the Oversampling (SMOTE)

Figure 6 presents a clear comparison of the class distribution in the water potability dataset before and after applying SMOTE oversampling. In Figure 6(a), the dataset exhibits a strong class imbalance, where the non-potable class (0) significantly outnumbers the potable class (1). Such imbalance can lead to biased model learning, where the classifier tends to favour the majority class, resulting in poor detection of potable water instances.



a) before Oversampling b) with SMPOTE Oversampling

Figure 6 the representation of the data class distribution before and after SMOTE

In contrast, Figure 6(b) shows the distribution after applying the Synthetic Minority Over-sampling Technique (SMOTE), where the minority class is augmented with synthetically generated samples to achieve a nearly balanced distribution between both classes. This balancing improves the model's ability to learn from both classes effectively, enhances generalization, and leads to more reliable and unbiased prediction performance in water quality assessment. It is clear from Figure 6 that before SMOTE, the class distribution is highly imbalanced (approximately 7:1 in favor of the non-potable class), whereas after SMOTE, it becomes nearly balanced at a 1:1 ratio between potable and non-potable classes.

5.2 Results of ML classifier without Using CNN

In this section the results of the WQP or classification performance of various conventional ML methods are presented. The comparative evaluation of the four confusion matrices as given in the Figure 7 a) to d), reveals notable differences in classification performance across the applied models for water quality prediction. The baseline Decision Tree (DT) model without SMOTE achieves an accuracy of approximately 93.5%, which marginally improves to 93.7% after applying SMOTE, indicating that oversampling has a limited but positive effect on DT performance. In contrast, the Random Forest (RF) with SMOTE demonstrates the best overall performance, achieving the highest accuracy of approximately 96.6%, reflecting its superior ability to generalize and effectively handle class imbalance through ensemble learning. However, the SVM model with SMOTE exhibits a significantly lower accuracy of approximately 64.7%, despite achieving very high true negative classification, suggesting poor sensitivity toward the minority (potable) class. This indicates that individual SVM classifier might struggles to balance precision and recall in the presence of synthetic data. Thus, it creates the straight forward demand of designing the hybrid classification approach capable of increasing the accuracy of the classification. Although using data balancing (SMOTE) method is capable of improving the accuracy. Similarly, the highly imbalance water-potability dataset results for different ML methods are presented in the Figure 8.

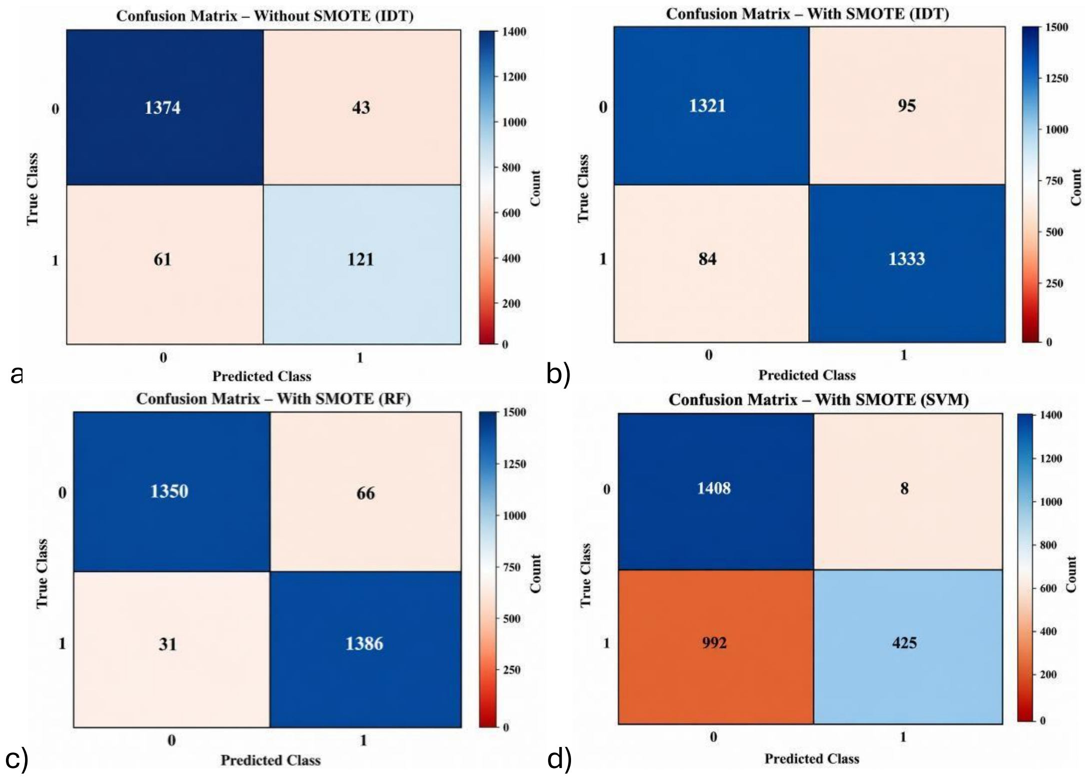


Figure 7 ML based classifier validation result on waterQuality database

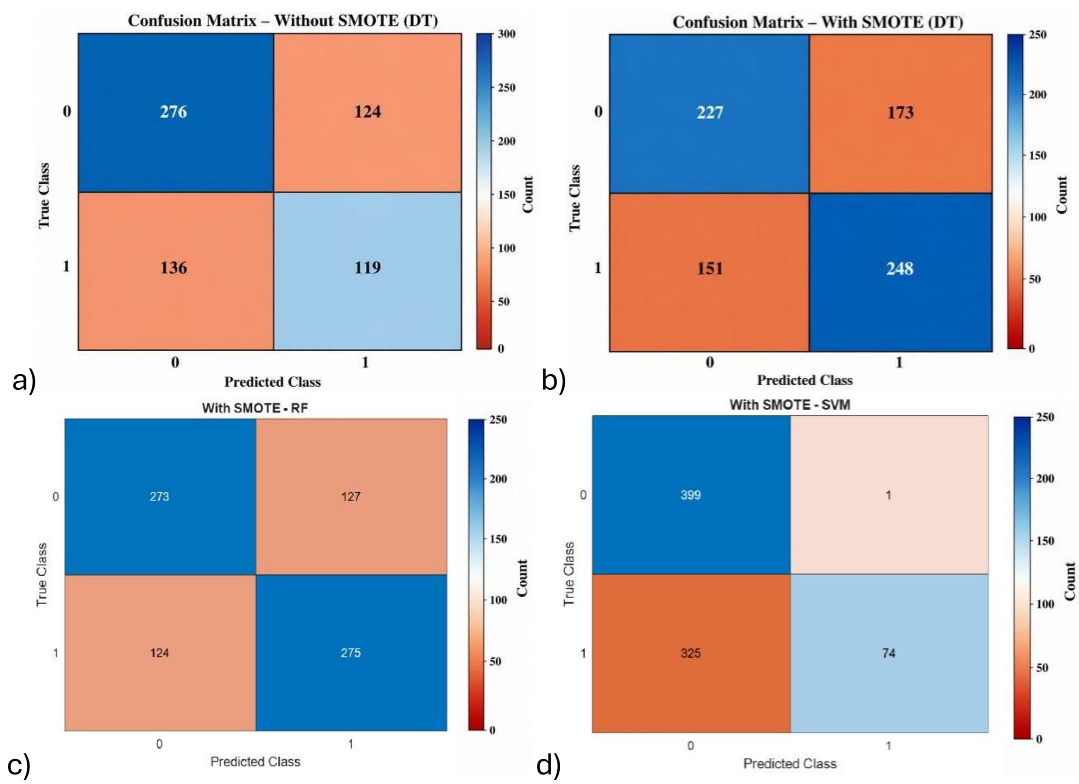


Figure 8 ML methods Validation for water-potability dataset

The performance comparison of four ML models, including DT, RF, SVM as the baseline model, in their validation phase based on SMOTE or non-SMOTE is illustrated in Figure 7. Without SMOTE, the DT model (Fig. 8a) displays decent accuracy, although there is significant

misclassification of the data point between potable and non-potable water samples. With SMOTE, the DT model (Fig. 8b) can obtain better class balance in terms of increased recall value of minority samples. With SMOTE, the RF model (Fig. 8c) achieves the highest level of consistency in classification, which indicates that this model provides the best accuracy and least false prediction. On the other hand, with SMOTE, the SVM model (Fig. 8d) is precise for detecting non-potable water samples but less sensitive to detect potable samples because of class overlap despite the oversampling.

It is apparent that there are significant disparities in accuracy among the four models. Without using SMOTE, the DT model produced an accuracy of 60.3% with notable class imbalance. When SMOTE was used accuracy level dropped slightly to 59.4%. The RF with SMOTE model generated the best accuracy at 68.6%, indicating the model's superior generalization and good classification accuracy for both classes. On the other hand, the SVM with SMOTE produced the worst accuracy at 59.2%. This model showed high accuracy in predicting non-potable water but low sensitivity for potable water, resulting in lower accuracy scores.

Model	Precision	Recall	F1-score
Decision Tree (No SMOTE)	0.738	0.665	0.700
Decision Tree (SMOTE)	0.933	0.941	0.937
Random Forest (SMOTE)	0.955	0.978	0.966
SVM (SMOTE)	0.981	0.300	0.459

Table 5 Comparative quantitative performance of Models for WQP database

Model	Precision	Recall	F1-score
Decision Tree (No SMOTE)	0.490	0.467	0.478
Decision Tree (SMOTE)	0.589	0.622	0.605
Random Forest (SMOTE)	0.684	0.689	0.686
SVM (SMOTE)	0.987	0.185	0.311

Table 6 Comparative quantitative performance of Models for water_potability database

The tables below show how various models perform in comparison on the WQP dataset (Table 5) and the water potability dataset (Table 6). In regards to the WQP dataset, use of SMOTE greatly improves model performance, especially in case of DT where the F1-score improves from 0.700 to 0.937. The model that performs best using SMOTE is the RF since it has the highest accuracy in terms of precision (0.955), recall (0.978) and F1-score (0.966). Although the model that has the highest precision is SVM (0.981), its recall is quite low (0.300), giving a poor F1-score (0.459), which indicates poor model performance since the model cannot effectively identify the minority class. On the other hand, the model that has the best performance on the water potability dataset when SMOTE is used is the RF. The F1-score increases from 0.478 to 0.605 in case of DT. These initial validations offer significant challenge of handling the data imbalance in case of water-portability database. Therefore, in this research a hybrid approach of DL based on CNN and SVM combination is proposed.

5.3 CNN Training Results

The training process of the Convolutional Neural Network (CNN) for water quality prediction is carried out using an optimized deep learning configuration to ensure efficient learning and high accuracy. The model is trained for a maximum of 10 epochs, during which the network iteratively updates its weights using the Adam optimization algorithm. A mini-batch size is employed to process subsets of data, enabling faster convergence and improved generalization.

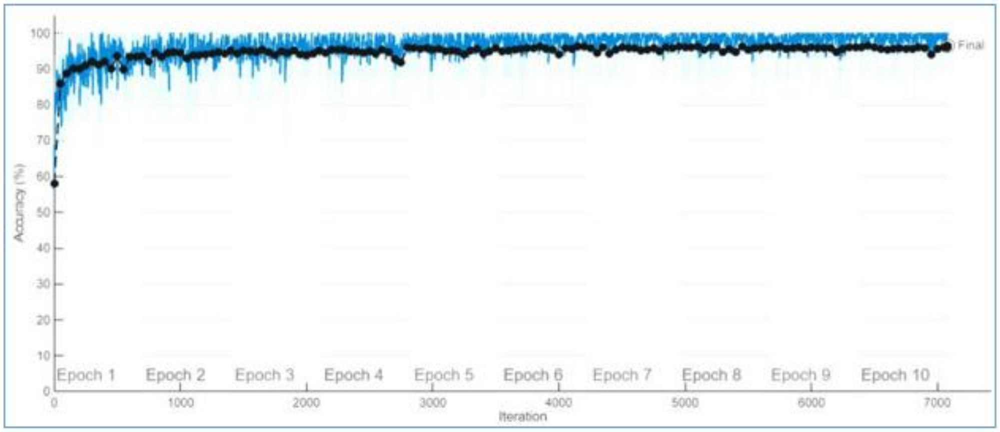


Figure 9 data Heatmap Confusion matrix

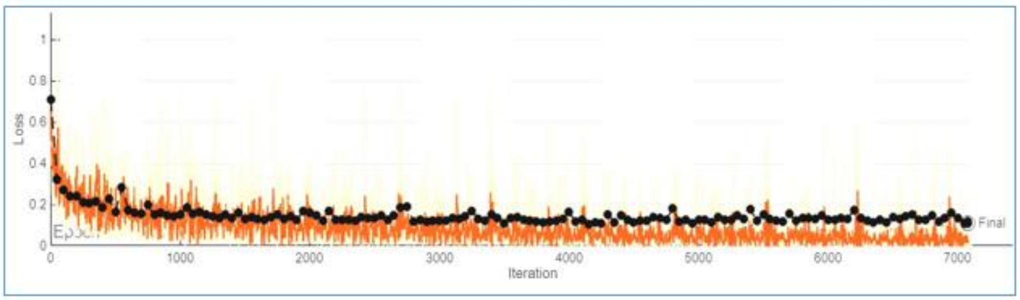


Figure 10 Loss function validation for hybrid WQP system

During training, the model performs forward propagation to extract hierarchical features from the input data trading accuracy is validated as in Figure 9, followed by backpropagation to minimize the cross-entropy loss function as shown in Figure 10. The learning rate is maintained constant at 0.001, ensuring stable and consistent weight updates throughout the training process. Validation is performed periodically after every 50 iterations to monitor model performance and prevent overfitting.

The training process consists of 7080 total iterations, with 708 iterations per epoch and above results are shown for the WQP database, and is executed using a single GPU to accelerate computation. The total training time recorded is approximately 6 minutes and 37 seconds. Upon completion, the model as shown in Figure 9 achieves a validation accuracy of 96.58%, indicating strong predictive capability and effective learning of underlying patterns in the water quality dataset. The lower loss function as in Figure 10 signifies the good quality of training and validation.

5.4 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 11 have presented the results of the validate confusion matrix for the hybrid approached of various ML models for WQP. Similarly, the Figure 12 presented the Confusion matrix for the hybrid CCN-RF and CNN-SVM approaches for the water_potability database (DB). It is clear from the Figure that significant accuracy improvement is achieved by the proposed hybrid model over previous validation results. The Figure 13 shows the percentage accuracy comparison between two hybrid models, CNN–RF and CNN–SVM, on two datasets: Water Quality DB and water_potability DB. As far as Water Quality DB is concerned, CNN–RF has an accuracy of 96.70% whereas CNN–SVM has a higher accuracy than CNN–RF with 97.25%, which signifies that both models have almost perfect accuracy and CNN–SVM has a slight edge over CNN–RF.

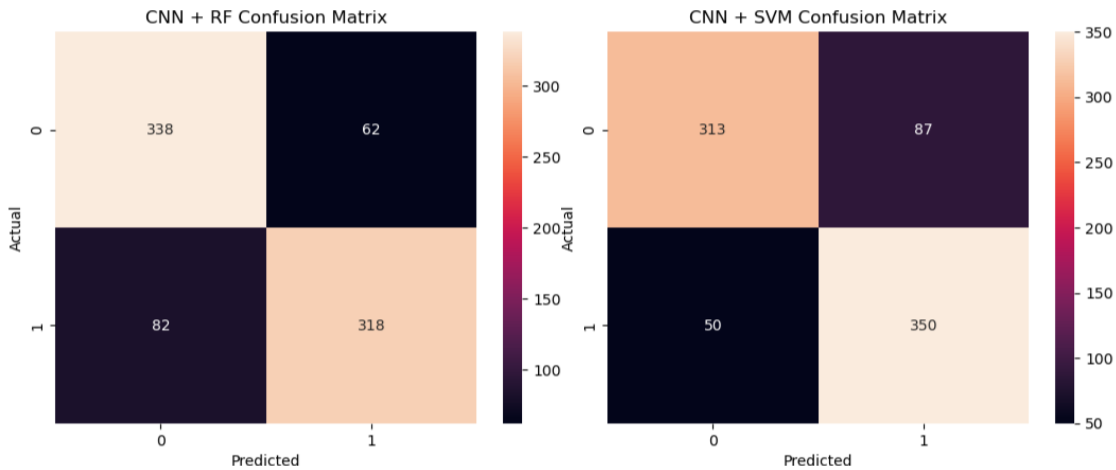
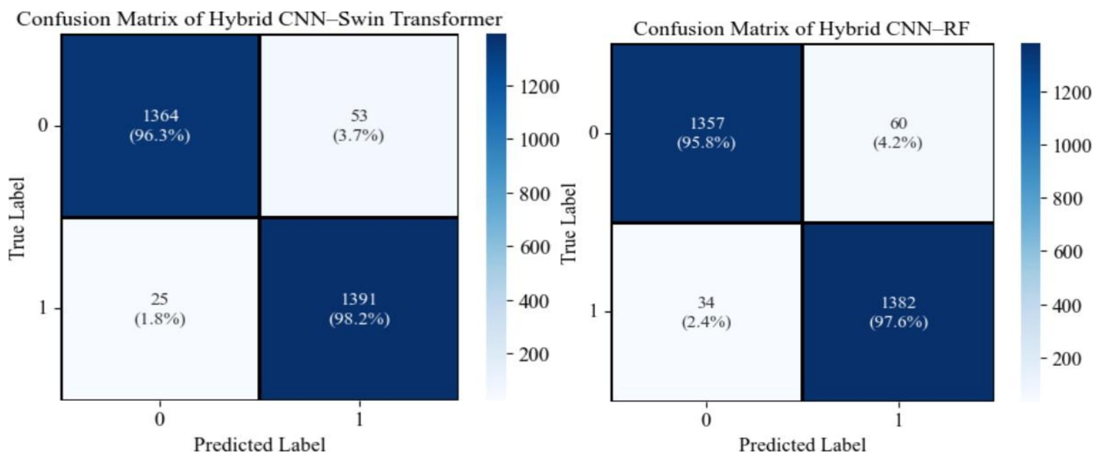


Figure 11 Results of for water_potability dataset a) for RF and b) for SVM



a) for CNN-SVM and b) for CNN-RF

Figure 12 Results of for WQP dataset

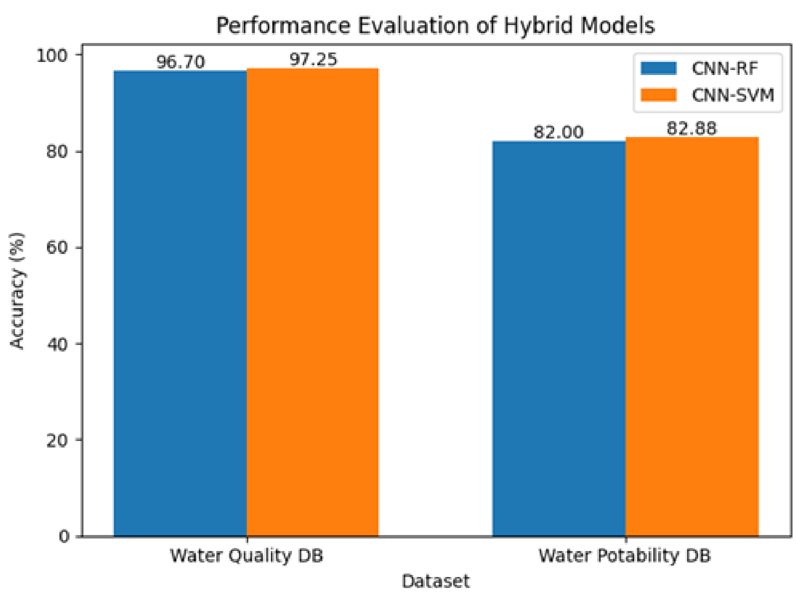


Figure 13 Performance Evaluation of hybrid model

Figure 13 Accuracy comparisons result for hybrid CCN-RF and CNN-SVM models

On the other hand, when it comes to Water Potability DB, as in Figure 13 the CNN–RF has 82.00% accuracy whereas CNN–SVM has a slight edge over CNN–RF by obtaining an accuracy of 82.88%.

The comparison of class-wise performance measures presented in Table 7 and Table 8 clearly shows the inherent strengths of the hybrid CNN–RF and CNN–SVM approaches for the Water Potability and Water Quality datasets. In the case of the Water Potability dataset (Table 7), the hybrid CNN–RF approach showed equal performance for Class 0 with precision of 80.48%, recall of 84.50%, and F1-score of 82.44%, whereas the hybrid CNN–SVM approach had higher precision (86.23%) and lower recall (78.25%) for Class 0, which gave it a similar F1-score of 82.04%. As far as Class 1 is concerned, the hybrid CNN–RF approach achieved 83.68% precision and 79.50% recall (F1 = 81.54%), while the hybrid CNN–SVM approach increased recall dramatically (87.50%) at the cost of reduced precision (80.09%), giving it a superior F1-score of 83.63%.

Class	**For Hybrid CNN-RF**			**For Hybrid CNN-SVM**		
	Precision (%)	**Recall (%)**	**F1 Score (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Class 0	80.48	84.50	82.44	86.23	78.25	82.04
Class 1	83.68	79.50	81.54	80.09	87.50	83.63

Table 7 Comparative Class-wise Performance for water potability dataset

Class	**For Hybrid CNN-RF**			**For Hybrid CNN-SVM**		
	Precision (%)	**Recall (%)**	**F1 Score (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Class 0	97.56	95.76	96.66	98.20	96.26	97.22
Class 1	95.84	97.60	96.71	96.33	98.23	97.27

Table 8 Comparative Class-wise Performance Metrics for CNN -SVM WQP database

Unlike the other datasets, the Water Quality dataset (as in Table 8) shows nearly perfect classification performance for both models, where the CNN–SVM model outperforms the CNN–RF model slightly. Specifically, for class 0, the CNN–RF model gives an accuracy of 97.56%, recall of 95.76%, and F1 of 96.66%. However, the CNN–SVM model performs slightly better,

giving an accuracy of 98.20%, recall of 96.26%, and F1 of 97.22%. Likewise, for class 1, the CNN–RF model obtains an accuracy of 95.84%, recall of 97.60%, and F1 of 96.71%. In contrast, the CNN–SVM model obtains an accuracy of 96.33%, recall.

Figure 14 shows the performance of the Hybrid CNN-SVM classifier on a per-class basis, emphasizing how consistent it is with respect to both the classes. In terms of Class 0, the precision, recall, and F1-score of the classifier were 98.20%, 96.26%, and 97.22%, respectively, showing that the classifier has performed exceptionally well in identifying cases where water samples are not potable and has been very sensitive. On the other hand, when the same measures were considered for Class 1, the model produced scores of 98.23% (precision), 96.33% (recall), and 97.27% (F1-score). It is evident from the above discussion that the performance of the Hybrid CNN-SVM model has been fairly similar in terms of precision, recall, and F1-score across both classes, thereby signifying that it does not exhibit any bias towards either of the two classes.

These results are further justify from the Figure 15 depicts the Receiver Operating Characteristic (ROC) curve for the Hybrid CNN–SVM model applied to the *Water Potability dataset*. The curve demonstrates the trade-off between the True Positive Rate (sensitivity) and the False Positive Rate across varying classification thresholds. The ROC curve rises steeply toward the top-left corner, indicating strong discriminative ability of the model. The Area Under the Curve (AUC) is reported as 0.9725, which reflects excellent classification performance, signifying that the CNN–SVM model can reliably distinguish between potable and non-potable water samples. The curve is contrasted against the diagonal baseline of a random classifier

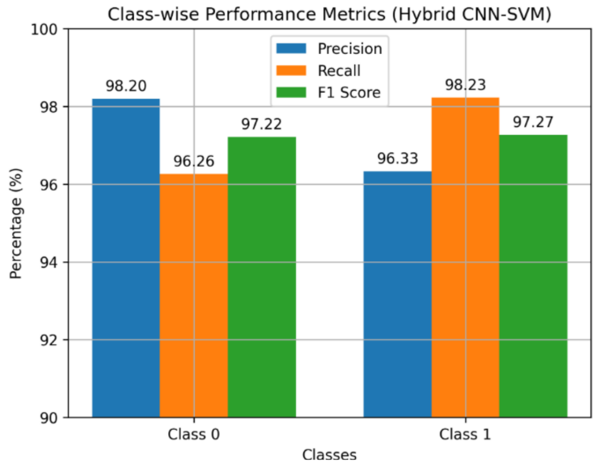


Figure 14 Results of the class vise performance comparisons for proposed CNN-SVM

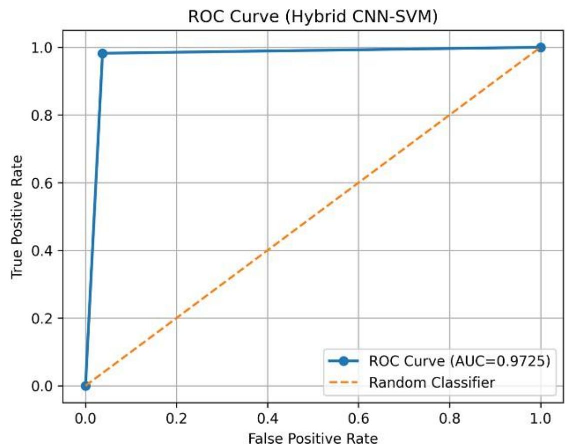


Figure 15 ROC curve and AUC result for the CNN-SVM WQP dataset

Model	Precision (Class 1)	Recall (Class 1)	F1 Score (Class 1)
P. Kamath RF Ensemble (2025)	73.0	75.0	74.02
Karthick et al (2024) LightGBM	94.90	74.00	83.10
Karthick et al (2024) RF	95.50	65.00	77.40
Karthick et al (2024) CatBoost	93.80	68.00	78.80
Proposed Deep CNN+RF Model	95.84	97.60	96.71
Proposed Deep CNN+SVM Model	96.33	98.23	97.27

Table 9 Comparison of the State of art performance for proposed hybrid modes tested

The comparative findings presented in Table 9 prove the dominance of the proposed hybrid CNN-SVM method compared to other previous state-of-the-art methods. In relation to the precision measure of Class 1, the hybrid CNN-SVM model attained a value of 96.33%, which is an increase of +23.33% from the RF Ensemble method of P. Kamath (73.0%), as well as +1.43% from the highest value among all traditional ensembles (RF with a value of 95.50%). On the other hand, for recall, CNN-SVM scored 98.23%, a high increase of +23.23% from Kamath's RF Ensemble (75.0%), as well as +33.23% from RF (65.0%). Thus overall, it is concluded that proposed hybrid CNN-SVM offers significant edge over the existing ML based approaches and offers the relatively significant edge on bot datasets. for water-potability dataset prosed hybrid approach offers nearly 15 % hick in accuracy signifies the good generalizability of model.

6. Conclusions and Future Scope

This study shows that our proposed hybrid CNN-SVM model significantly outperforms both classic machine learning methods and other hybrid models towards prediction of water characteristics. Since the feature extraction of CNN is deep and the classification of SVM is powerful, the hybrid approach achieves better parametric performance between different datasets.

Thus, the hybrid CNN-SVM model obtained a validation accuracy of 97.25% on the Water Quality Prediction (WQP) dataset and 82.88% on the water potability dataset, which significantly surpassed hybrid CNN-RF methods that obtained 96.70% and 82.00% respectively. The Area Under the Curve (AUC) for the hybrid CNN-SVM on the water potability dataset was 0.9725 as shown.

Thanks to incorporating SMOTE, this model manages some of the problems such as class imbalance well, while showing excellent generalizability and reliability, with high AUC values and balanced class-wise performance metrics. Compared to traditional ML models with

combined methods for instance SMOTE, the results showed much improved precision, recall and F1-scores from a hybrid CNN-SVM model. For instance, on the WQP sample, both the Random Forest with SMOTE obtained F1-score of 0.966 and the CNN-SVM achieved 0.9727.

Compared with the RF Ensemble method proposed by P. Kamath (2023), there was a

+23.33% precision increase and +23.23% recall improvement for Class 1 for proposed hybrid CNN-SVM model and outperformed state-of-the-art ensemble methods. The results verify that hybrid CNN-SVM architecture provides an effective and practical approach for the accurate and reliable water quality sampling, overcoming current state-of-the-art approaches and serving as a significant instrument for environmental monitoring and human health safeguarding.

As a future scope the optimize method can integrate with the hybrid CNN-SVM by introducing a more sophisticated feature selection to minimize dimensionality and increase computational efficiency. Considering other kernel functions or adaptive kernel learning in the SVM component may provide more effective performance among the nonlinear data from water quality. Studying ensemble hybrid models to combine few deep learning architectures with different classifiers to increase predictive ability and generalization to different water quality data could be promising

Acknowledgement

We gratefully acknowledge the support and facilities provided by Computer Science Department LNCT University,Bhopal in carrying out this research work. We also extend our appreciation to the guide who assisted in data preparation, experimentation, and constructive discussions that enriched the quality of this study.

Author contributions

All author contributes equally to the research, design the study, analyzed data and approved the final manuscript

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethic Approval

This study did not involve human participants or animal experiments and therefore did not require formal ethics approval. Publicly available, datasets were used in accordance with ethical research standards.

Data Availability

The dataset drinking water potability used in this study is openly available for legitimate research purposes but requires access approval to prevent misuse. The dataset can be requested via the project site: https://www.kaggle.com/datasets/adityakadiwal/water-potability

Declaration of Artificial Intelligence (AI) Assistance

The authors declare this manuscript was prepared without the assistance of generative AI. All conceptualization, analysis, interpretation and writing were developed independently by the authors.

Funding

The authors received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors for the conduct of this research.

References

1. Huang, T., Chau, K.Y., Zhan, S. et al. Hybrid deep learning and machine learning framework for high-precision water quality prediction in urban systems. Environ Dev Sustain (2025). https://doi.org/10.1007/s10668-025-06447-2
2. Z. Muhammad, N. A. J. Jailani, N. A. M. Leh and S. A. Hamid, "Classification of Drinking Water Quality using Support Vector Machine (SVM) Algorithm," *2022 IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE)*, Penang, Malaysia, 2022, pp. 75-80, doi: 10.1109/ICCSCE54767.2022.9935657.
3. Umar Zaky, Ahmad Naswin, Sumiyatun Sumiyatun "Performance Analysis of the Decision Tree Classification Algorithm on the Water Quality and Potability Dataset". *Indonesian Journal of Data and Science*, vol. 4, no. 3, Dec. 2023, pp. 145-
4. 50, https://doi.org/10.56705/ijodas.v4i3.113.
5. Vinita Sangwan, Rashmi Bhardwaj; Machine learning framework for predicting water quality classification. Water Practice and Technology 1 November 2024; 19 (11): 4499–
6. 4521. doi: https://doi.org/10.2166/wpt.2024.259
7. N. A. Kumar and V. Jeevanantham, "Optimized and Effective Feature Selection-Based Hybrid Deep Learning Framework for Accurate Water Quality Prediction: A Review," *2024 2nd International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR)*, Hyderabad, India, 2024, pp. 1-5, doi: 10.1109/ICAITPR63242.2024.10960056.
8. Mehak Afzal, Shujaat Ali, Hafiz Burhan Ul Haq, Rabia Younis, Hamid Ali, Amna Kosar, and Hafiz Muneeb Akhtar. (2024). Water Quality Assessment Through Predictive Modeling Employing Machine Learning Methods. *Journal of Computing and Biomedical Informatics*, *7*(02). Retrieved from https://jcbi.org/index.php/Main/article/view/607
9. Q. Zhu, X. Yu, L. Zhao and L. Xu, "A Hybrid Deep Learning Model for Water Quality Prediction," in *IEEE Access*, vol. 13, pp. 106406-106415, 2025, doi: 10.1109/ACCESS.2025.3580741.
10. Karthick K, S. Krishnan, R. Manikandan; Water quality prediction: a data-driven approach exploiting advanced machine learning algorithms with data augmentation. Journal of Water and Climate Change 1 February 2024; 15 (2): 431–452. doi: https://doi.org/10.2166/wcc.2023.403 s
11. P. Thota, K. Challapalli, H. Garikapati, V. M. S. Adusumilli, S. Anamalamudi and M. K. Enduri, "Microplastic Detection in Drinking Water: A Comparative Analysis of CNN-SVM and CNN-RF Hybrid Models," *2024 OITS International Conference on Information Technology (OCIT)*, Vijayawada, India, 2024, pp. 24-29, doi: 10.1109/OCIT65031.2024.00014.
12. Saleh Y. Abuzir, Yousef S. Abuzir; Machine learning for water quality classification. *Water Quality Research Journal* 1 August 2022; 57 (3): 152–164. doi: https://doi.org/10.2166/wqrj.2022.004
13. N. Alex and R.Brindha, "A Hybrid Preprocessing Pipeline Employing Ensemble Machine Learning Models and Wavelet Denoising for Intelligent Water Potability Prediction," *2025 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, Coimbatore, India, 2025, pp. 1411-1416, doi: 10.1109/ICoICI65217.2025.11254239.
14. Tverdokhlib, O., Shavaev, D., Matseliukh, Y., Gozhyj, A., Trzaskowska, A. M., Korobchynskyi, M., ... and Kalinina, I.. Analysis Method for Determining the Suitability of Water for Human Consumption. In *MoMLeT+ DS*, pp. 388-411 2022.
15. Ahmed, U.; Mumtaz, R.; Anwar, H.; Shah, A.A.; Irfan, R.; García-Nieto, J. Efficient Water Quality Prediction Using Supervised Machine Learning. Water 2019, 11, 2210. https://doi.org/10.3390/w11112210
16. Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi, Abbas Parsaie; Water quality prediction using machine learning methods. *Water Quality Research Journal* 1 February 2018; 53 (1):
17. 3–13. doi: https://doi.org/10.2166/wqrj.2018.025
18. Ali Najah Ahmed, Faridah Binti Othman, Haitham Abdulmohsin Afan, Rusul Khaleel Ibrahim, Chow Ming Fai, Md Shabbir Hossain, Mohammad Ehteram, Ahmed Elshafie, Machine learning methods for better water quality prediction, Journal of Hydrology, Volume 578, 2019, 124084, https://doi.org/10.1016/j.jhydrol.2019.124084.
19. Azrour, M., Mabrouki, J., Fattah, G. *et al.* Machine learning algorithms for efficient water quality prediction. *Model. Earth Syst. Environ.* **8**, 2793–2801 (2022). https://doi.org/10.1007/s40808-021-01266-6
20. Zrouri, A., El Farissi, I. (2024). Towards Predictive Water Quality: Synergies Between Machine Learning and Internet of Things. In: Serrhini, M., Ghoumid, K. (eds) Advances in Smart Medical, IoT and Artificial Intelligence. ICSMAI 2024. Information Systems Engineering and Management, vol 11. Springer, Cham. https://doi.org/10.1007/978-3-031-66850-0_17
21. Kaddoura, S. Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability. Sustainability 2022, 14, 11478. https://doi.org/10.3390/su141811478
22. Bhalerao, Omkar and Mohan, Allam, Water Quality Monitoring and Prediction Using Machine Learning Models (May 5, 2023). Proceedings of the KILBY 100 7th International Conference on Computing Sciences 2023 (ICCS 2023), Available at SSRN: https://ssrn.com/abstract=4487070 or http://dx.doi.org/10.2139/ssrn.4487070

23. N. Suwansukho and C. Wutthithanyawat, "Experimental Study of Real-Time Water Quality Analysis with Deep Neural Network (DNN) on Edge Computing," *2024 10th International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, Luang Prabang, Laos, 2024, pp. 138-143, doi: 10.1109/ICEAST61342.2024.10553703.
24. Anand, M and Sohitha, Chennareddy and Saraswathi, Galla and Lavanya, GV. (2023). Water quality prediction using CNN. Journal of Physics: Conference Series. 2484. 012051. 10.1088/1742-6596/2484/1/012051.
25. Yan, X.; Zhang, T.; Du, W.; Meng, Q.; Xu, X.; Zhao, X. A Comprehensive Review of Machine Learning for Water Quality Prediction over the Past Five Years. J. Mar. Sci. Eng. 2024, 12, 159. https://doi.org/10.3390/jmse12010159
26. Luthra, G., Kukkar, S., Harnal, S., Tiwari, R., Upadhyay, S., Chhabra, G. (2024). Water Quality Prediction Using Machine Learning. In: Kumar, R., Verma, A.K., Verma, O.P., Wadehra, T. (eds) Soft Computing: Theories and Applications. SoCTA 2023. Lecture Notes in Networks and Systems, vol 971. Springer, Singapore. https://doi.org/10.1007/978-981-97-2089-7_10
27. A. Helaly, M., Rady, S., Mabrouk, M. et al. Advancements in water quality prediction: a practical review of machine learning and deep learning approaches. Cluster Comput 28, 598 (2025). https://doi.org/10.1007/s10586-025-05221-3
28. Islam, Md and Roy, Sujan and Utsho, Fahad and Naznin, Laila and Sarkar, Ripa and R
29. Sarkar, Ratna. (2025). An IoT-enabled intelligent water quality monitoring system for tourist safety using machine learning. Discover Electronics. 2. 10.1007/s44291-025-00070-1.
30. Cardia, M., Chessa, S., Micheli, A., Giuliana Luminare, A., Gambineri, F. (2026). Water Quality Estimation Through Machine Learning Multivariate Analysis. In: Esposito, A., Faundez-Zanuy, M., Morabito, F.C., Pasero, E., Cordasco, G. (eds) Neural Networks: Overview of Current Theories and Applications. Smart Innovation, Systems and Technologies, vol 459. Springer, Singapore. https://doi.org/10.1007/978-981-95-4072-3_22
31. 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