FMDB *Transactions on Sustainable Environmental Sciences*

A Systematic Comparative Analysis of Data Preprocessing Strategies for Water Quality Safety Classification

K. B. Susha[1,*], H. Jayamangala[2]

[1,2]Department of Computer Applications, Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamil Nadu, India.
sushakb@gmail.com[1], jayamangala.scs@vistas.ac.in[2]

Abstract: Monitoring water quality is a key element to maintaining the sustainability of our environment and public health. Water-quality safety classification systems based on machine learning have been developed as effective tools for predicting potable and non-potable conditions. Nevertheless, classification performance depends significantly on the effectiveness of data preprocessing techniques before model training. Although machine learning has been increasingly used to assess water quality, few studies have systematically tested the effects of various preprocessing methods on classification results. This study compares and contrasts the most significant preprocessing methods, including missing data imputation, outlier detection, normalisation, feature selection, dimensionality reduction, and class balancing. A systematic experimental framework is created from publicly available water-quality datasets. The following classification models are evaluated: Random Forest, Support Vector Machine, XG-Boost, and Artificial Neural Networks, with varying preprocessing conditions. The experimental results show that the accuracy, precision, recall, F1-score, and computational efficiency of classification can be significantly influenced by preprocessing. Comprehensive feature selection, along with powerful normalisation and advanced imputation strategies, delivers the highest overall performance. The results offer useful recommendations for selecting appropriate pre-processing pipelines for water-quality safety classification systems and for developing reliable intelligent water-monitoring systems.

Keywords: Water Quality Classification; Data Preprocessing; Machine Learning; Feature Selection; Data Imputation; Water Potability Prediction; Environmental Monitoring.

Received on: 20/03/2025, **Revised on:** 27/05/2025, **Accepted on:** 14/08/2025, **Published on:** 05/06/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSESS

DOI: https://doi.org/10.69888/FTSESS.2026.000714

Cite as: K. B. Susha and H. Jayamangala, "A Systematic Comparative Analysis of Data Preprocessing Strategies for Water Quality Safety Classification," *FMDB Transactions on Sustainable Environmental Sciences*, vol. 3, no. 2, pp. 54–70, 2026.

1. Introduction

Water is a necessity for human health, industry, and environmental sustainability, and must be safe and clean [1]. These factors, such as rapid urbanisation, industrialisation, and agriculture, pollute water, underscoring the importance of an intelligent water-quality assessment system and continuous monitoring. Process testing methods are typically costly and time-consuming and are not suitable for online applications [4]. Hence, machine learning techniques have received significant attention for automating water quality classification and prediction to enhance water safety [2]. Typical challenges in real-world water quality data include missing data, noise, faulty sensor readings, redundant features, high dimensionality, and skewed class

[*]Corresponding author.

distributions [3]. These are all factors that affect machine learning performance and can lead to incorrect classification results. The most important step is data pre-processing, which converts raw environmental data into structured, reliable, and usable data for predictive modeling [5]. Several approaches to preprocessing have been suggested in the literature, including statistical imputation [6], interpolation methods [7], normalization procedures [8], outlier removal [10], feature selection algorithms [11], and dimensionality reduction [12]. The effects of these methods on water quality safety classification have been investigated separately, but limited information is available [9]. This study comprehensively evaluates the effectiveness of various pre-processing techniques and the best pre-processing pipelines for water-quality classification problems [31]. The results yield several positive recommendations for researchers and practitioners in environmental monitoring engaged in the design of intelligent water-quality monitoring systems. The major objectives are as follows:

- To review systematically water quality classification studies in which preprocessing methods have been used.
- To study the impact of different preprocessing approaches on the classification accuracy, an experiment was conducted.
- To find the optimum pre-processing of the data that will lead to a successful water quality safety prediction process.

2. Literature Review

Recent work shows that the performance of machine learning techniques in environmental monitoring systems can be significantly affected by data preprocessing. Water quality datasets are commonly encountered in practice, and various techniques for handling missing values, such as mean imputation, KNN imputation, and multiple imputation, are widely used. Moving-average filtering, Kalman filtering, and wavelet denoising are popular noise-reduction techniques for enhancing the quality of sensor data. Water quality prediction has emerged as a significant research area due to growing concerns about the safety of drinking water and sustainable resource management. Several techniques, including machine learning, deep learning, preprocessing, dimensionality reduction, and statistical methods, have been used to enhance prediction accuracy. Some recent studies are briefly summarised below, which highlight new developments, problems, and prospects. Feature consistency and improved model convergence are achieved through normalization techniques such as Min-Max Scaling and Z-score Standardization. Various feature selection methods can be used to identify non-redundant yet important features, such as Recursive Feature Elimination (RFE), Mutual Information, and Random Forest Importance. High-dimensional water-quality datasets are often analyzed using Principal Component Analysis (PCA). To enhance the accuracy of water potability prediction in resource-limited environments, Phorah et al. [12] highlighted the need for data cleaning, feature engineering, and dimensionality reduction. The study showed that the model's performance can be greatly improved with strong preprocessing.

However, it was still an investigation based on reviews rather than the integration of an intelligent prediction framework. To improve the accuracy of the prediction model, Shim et al. [13] added wave information acquired during ultrafiltration to wavelet-enhanced CNN and LSTM models. They detected an increase in turbidity when membrane damage occurred, as measured by their procedure. The study, however, was conducted for a specific desalination application, and the results may not be extrapolated to other water-quality assessment applications. Panahi et al. [14] proposed a hybrid VMD ensemble learning model for predicting water quantity and quality. The method was applied to simulate non-linear streamflow data at multiple frequencies to enhance forecast accuracy. While it did work, it focused primarily on forecasting tasks, and the preprocessing steps were somewhat computationally intensive and may pose a problem for real-time applications. Pipelzadeh and Mastouri [15] proposed combining EMD, EEMD, and VMD pre-processing techniques with Model Tree algorithms to estimate TDS and EC concentrations. The conclusion was that performance improvements were observed compared to stand-alone models, although not all were significant. The study focused only on the selected water quality parameters, not on water quality classification scenarios. Lokman et al. [16] provided an overview of machine learning and statistical methods for water quality prediction and classification and emphasized the advantages of hybrid machine learning and statistical predictive models. Interpretability and data quality issues were identified in the study. This was descriptive, with no new structure to address these shortcomings.

Mounir et al. [17] investigated methods for augmenting data for imbalanced water quality datasets. They found that applying SMOTE-ENN improved recall for minority classes. The paper successfully addressed the contamination-detection problem in safety-critical areas. The study focused mainly on balancing strategies; however, it did not include advanced deep learning architectures capable of extracting complex feature representations. Given the shortcomings of current water quality classification models, Chen et al. [18] developed an innovative BS-FAMLP model that combines the benefits of GBDT, Bayesian optimization, attention mechanisms, and MLPs. The model's classification performance was excellent, and the computation was very efficient. However, it has limitations, including shallow learning components in the architecture and the lack of explicit modeling of temporal relationships in water-quality data. Chatterjee et al. [19] evaluated various machine learning methods and demonstrated that water potability prediction could be improved by applying PCA. They found that dimensionality reduction is a key issue in feature optimization. The study, however, mainly used traditional classifiers and did not employ cutting-edge deep learning methods to automatically extract features. Dritsas and Trigka [20] discussed efficient

machine learning models for predicting water quality based on data, with a particular focus on implementing automated water quality assessment systems. The study helped gain insight into the application of machine learning to water monitoring. However, there was less emphasis on explainable AI, hybrid deep learning models, and adaptive feature optimization strategies.

Mittal et al. [21] used the SVM, Decision Tree, and Random Forest classifiers for water quality monitoring and water potability assessment. The study found that machine learning can effectively assess water quality parameters. However, traditional classifiers were not very effective in capturing the complex interactions and highly nonlinear relationships between water quality attributes. Saab and Zéhil [22] focused on statistical techniques used in water quality monitoring and forecasting, highlighting their importance for identifying water quality patterns and supporting decision-making. The review reiterated the significance of statistical analysis. However, for high-dimensional, nonlinear environmental data, statistical tools are insufficient for modeling. Previous studies have shown the advantages of various individual preprocessing methods, but these have been conducted under inconsistent experimental conditions and without comparative studies. The importance of pre-processing, Dimensionality reduction, ML, and Statistical techniques for water quality prediction is established by previous work. Most approaches, however, tend to focus on individual components rather than end-to-end frameworks. Numerous studies regarding deep learning are application-oriented, and traditional machine learning models are less suitable for high-dimensional, imbalanced, and nonlinear data. There has, however, been less discussion about the optimization of adaptive features, explainable AI, attention-based learning, and temporal-spatial dependency modelling. Moreover, the deployment of real-time applications in IoT-enabled environments remains underexplored. The restrictions indicate the need for an explainable, unified, and efficient computational hybrid deep learning framework for robust water quality classification and prediction [23].

3. Materials and Methods

3.1. Dataset Description

The study aims to use the Water Potability Dataset, an open-source benchmark dataset widely adopted for water quality assessment and potable water prediction. There are 3,276 water samples, each consisting of 9 physicochemical parameters that affect water safety and suitability for drinking. The aim is to categorise water samples into two categories: Potable (Safe for Drinking) and Non-Potable (Unsafe for Drinking). The data set includes challenges such as missing values, class imbalance, and unequal scales in the data parameters, making it useful for assessing the effectiveness of data preprocessing strategies. The measurements taken are vital chemical and physical characteristics of water that impact humans and environmental health.

3.1.1. Key Features in the Dataset

Table 1 presents the main characteristics of the Water Potability Dataset and their relevance to evaluating water quality and safety for human consumption. The data set includes nine physicochemical parameters, which represent different chemical and physical characteristics of water, such as acidity (pH), mineral presence (Hardness, Solids), disinfectant level (Chloramines), ionic composition (Sulfate, Conductivity), organic matter (Organic Carbon), disinfection by-products (Trihalomethanes), and water clarity (Turbidity).

Table 1: Key features in the water potability dataset

Feature	Description	Significance in Water Quality Assessment
pH	Measures the acidity or alkalinity of water	Determines whether water falls within safe drinking standards
Hardness	Concentration of calcium and magnesium salts	Influences taste, scaling, and water usability
Solids	Total dissolved solids (TDS) present in water	Indicates the presence of dissolved minerals and contaminants
Chloramines	Concentration of chloramines used for disinfection	Helps eliminate harmful microorganisms
Sulfate	Sulfate ion concentration in water	Excess levels may cause health and taste issues
Conductivity	The ability of water to conduct electricity	Reflects the concentration of dissolved ions
Organic Carbon	Amount of organic matter dissolved in water	Indicates potential contamination sources
Trihalomethanes	By-products formed during water disinfection	High concentrations may pose health risks
Turbidity	Measure of water cloudiness caused by suspended particles	Indicates water clarity and contamination levels
Potability	Target class (0 = non-potable, 1 = Potable)	Represents overall drinking water safety

The features provide information specific to water quality that makes it suitable for human consumption and may identify potential contamination sources or concerns. The target variable, Potability, is a binary classification task in which water samples are classified as potable or non-potable, and machine learning models are developed for water safety classification. All these attributes provide a comprehensive picture of water quality conditions and serve as a basis for analysing the effectiveness of various data preprocessing methods and classification algorithms.

3.1.2. Challenges in the Dataset

The Water Potability Dataset contains several real-life issues common to environmental monitoring problems. Effective imputation techniques need to be applied to address information loss caused by missing values in important attributes, such as pH, Sulfate, and Trihalomethanes. The dataset is also class-imbalanced, with more non-potable than potable samples, which could lead to a classification model skewed toward the majority class. In addition, the different magnitudes of physicochemical parameters need to be normalized or standardized to represent the features better. Outliers and measurement noise in the measurement process can adversely affect the model's performance, and feature correlations can lead to redundant features. The difficulty has made the need for robust pre-processing techniques evident, and the data set is a suitable testing ground for implementing and systematically assessing different pre-processing techniques for water quality safety classification.

3.2. Preprocessing Strategies

The proposed architecture shown in Figure 1 represents a suitable pipeline for water quality safety classification, comprising a multistage process of data preprocessing, feature engineering, machine learning, and evaluation. Its initial input is raw water-quality data from environmental datasets, consisting of physicochemical parameters.

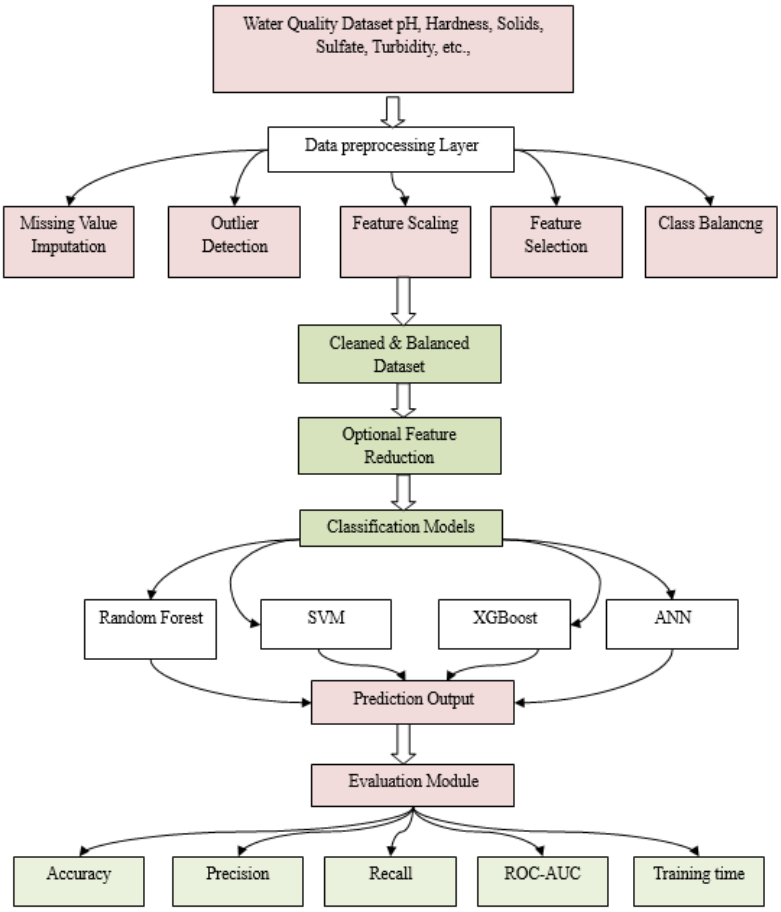

Figure 1: Proposed system architecture for water quality safety classification using a hybrid preprocessing and machine learning framework

The most crucial part is preprocessing, where missing values are imputed using statistical and distance-based methods, and outliers are removed using IQR and Z-score methods. Feature scaling ensures that the attributes are scaled similarly, and feature

selection techniques discard features that do not contribute much to the prediction process while retaining those that do. Class imbalance is addressed using SMOTE and ADASYN, which improve the model's accuracy in identifying minority-class samples. Additionally, dimensions are reduced using PCA on the cleaned data, which is then fed to several classification algorithms, including Random Forest, SVM, XG-Boost, and ANN. All of these are separate models and can independently classify a body of water as potable or non-potable. The evaluation module computes various metrics, including Accuracy, Precision, Recall, F1-score, ROC-AUC, and computational efficiency. According to the Architecture, the hybrid pre-processing pipeline using XG-Boost is the best overall solution. This modular concept will enable scalability, robustness, and versatility for real-world water quality monitoring applications.

3.2.1. Strategy S1: Baseline

The baseline strategy is the reference strategy used to assess the effectiveness of the various preprocessing techniques. This method uses the original water quality data set without preprocessing, transformation, or enhancement. All available features are preserved in their original format, even missing values, outliers, and different scales. The idea behind this strategy is to provide a standard for evaluating how the learning process's performance improves when preprocessing is used. However, because environmental data are noisy and incomplete, and may contain inconsistencies, the baseline model may be less accurate in prediction and exhibit weaker generalisation. The original dataset is used directly and employed without any modification, and it is represented by the dataset Eq (1):

$$X_{baseline} = X_{raw} \tag{1}$$

Where X_{raw} represents the original water quantity dataset. However, it provides some insight into the dataset's intrinsic quality and underscores the need for preprocessing. The study can quantitatively evaluate the contribution of each proposed preprocessing method to improving water-quality safety classification performance and model robustness relative to the baseline configuration.

3.2.2. Strategy S2: Missing Value Imputation

Handling missing data is one of the most significant challenges in environmental monitoring datasets, and if not properly addressed, it can affect the performance of machine learning algorithms. The strategy uses three common imputation techniques: Mean Imputation, Median Imputation, and K-Nearest Neighbours (KNN) Imputation. The mean imputation is to replace missing features with the mean of the feature, and the median imputation is to replace missing features with the median. The KNN imputation method first computes the similarity between the two neighbouring samples, then imputes the missing value based on that similarity.

The goal of this preprocessing technique is to reconstruct information lost without distorting the distribution of the original data. Correctly handling missing values enables the use of all samples for training and testing without loss of information, thereby yielding a more reliable model. For water quality datasets, such as those with attributes like pH, Sulfate, or Trihalomethanes, which often have missing data, this strategy is especially important. The mean imputation replaces missing values with the mean of the corresponding features, and it is denoted by Eq. (2):

$$X_{i,j} = \begin{cases} x_{i,j} \;\; if \; x_{i,j} \neq \emptyset \\ \bar{x}_j \;\; if \; x_{i,j} = \emptyset \end{cases} \tag{2}$$

Where $\bar{x}_j$ is the mean value of the features j.

3.2.3. Strategy S3: Outlier Treatment

Values that are very different from the other values are called outliers and may be due to sensor problems or errors, or to unusual environmental conditions. These extreme values may impair the operation of machine learning algorithms, including their ability to learn and make predictions. They can negatively affect machine learning models, particularly by altering their statistical distributions and shifting their decision boundaries. To overcome this problem, two popular outlier detection and treatment techniques, namely the Interquartile Range (IQR) Method and Z-Score Filtering, are used in this proposed strategy. The IQR method will be used to detect points outside the quartile range, and the Z-score method will be used to detect samples that are significantly different from the feature mean by a given number of standard deviations. Z-Score for outlier detection is represented by Eq. (3):

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{3}$$

Where value satisfying $|Z_i| > 3$ is considered an outlier. After identifying outliers, they can be either corrected or eliminated to minimize their impact on model learning. This preprocessing method helps make the data more consistent and enables classifiers to understand meaningful patterns within the dataset. By reducing the influence of outliers, outlier treatment improves classification performance and more accurately predicts water quality safety.

3.2.4. Strategy S4: Data Normalization

Numerical ranges and units of measurement for water quality parameters can vary greatly. The pH, for example, will generally be 0–14, while the dissolved solids or conductivity could be in the thousands. These differences can negatively affect distance-based and gradient-based machine learning algorithms. Three normalization techniques are used: Min-Max Scaling, Z-Score Standardization, and Robust Scaling. Min-Max Scaling (also known as "scaling between a range") is a technique for scaling feature values to a fixed range, typically 0 to 1. Z-Score Standardization rescales features to a zero mean and unit variance based on the mean and standard deviation. Unlike other statistics, such as the mean, which is easily affected by outliers, Robust Scaling uses median and interquartile range statistics. Normalization balances the importance of all features during model training and helps prevent features with higher magnitudes from dominating the learning process. In general, normalized datasets lead to better convergence, better models, and better classification results.

3.2.5 Strategy S5: Feature Selection

The goal of feature selection is to find the most informative features and remove redundant or irrelevant features from the data. If you can trim unnecessary features, you can boost computational efficiency, reduce the risk of overfitting, and improve model interpretation. The strategy used is Correlation Analysis, Recursive Feature Elimination (RFE), and Random Forest Feature Importance. While correlation analysis can detect highly correlated features that can make the model redundant, RFE can be used to exclude less significant features based on the model's performance. Random Forest Importance is based on the principle of measuring each feature's relative importance by how much its removal reduces the model's predictions. In water quality assessment, certain parameters may provide only limited additional information if other parameters in the system are highly correlated with them. Thus, selecting the most influential features will yield a smaller, more efficient data set. Feature Selection: Selects only the most discriminative features, boosting classifier efficiency and ensuring the retention of critical information for water safety prediction. This approach also helps to create more straightforward and comprehensible machine learning models. The importance of the features using random forest is given by Eq. (4):

$$FI_j = \frac{1}{T}\sum_{t=1}^{T} \Delta I_{j,t} \quad (4)$$

Where $\Delta I_{j,t}$ is the impurity reduction contributed by features j in tree t.

3.2.6. Strategy S6: Dimensionality Reduction

Dimensionality reduction reduces the dimensionality of data while retaining most of the information. In this study, the main method of dimensionality reduction used is Principal Component Analysis (PCA). PCA generates a new set of variables, called principal components, which capture as much variance as possible from the original set. PCA: projection of data on a smaller number of components, minimizing the redundancy of the features and the influence of the correlation between the features. The Principal Component Analysis (PCA) is denoted by Eq. (5):

$$z = XW \quad (5)$$

Where X is the normalization features matrix, W is the matrix of principal component eigenvectors, and Z is the reduced - dimensional features space. This helps reduce data volume while preserving the key aspects of the data. When using water quality models, dimensionality reduction can improve computational efficiency, shorten training time, and make complex relationships among water quality variables easier to visualize. Also note that dimensionality reduction can help combat overfitting, particularly when the training sample is small. The effectiveness of PCA is evaluated by comparing classification (before transformation) with classification (after transformation) and determining whether there is any relationship between the two, which can be compactly represented in the classification space after transformation that retains the capability of prediction.

3.2.7. Strategy S7: Class Balancing

A problem in classification problems is the uneven distribution of classes, where one class is much larger than the others. The Water Potability Dataset has more non-potable than potable samples, which could bias the machine learning model toward the majority class and lead it to ignore the minority class. In this regard, the study uses Synthetic Minority Over-sampling Technique

(SMOTE) and Adaptive Synthetic Sampling (ADASYN). SMOTE is a technique for synthesizing minority-class samples by interpolating existing minority-class samples to create additional samples that do not simply duplicate the originals. SMOTE generates synthetic minority samples as given by Eq. (6):

$$x_{new} = x_i(\lambda(x_{nn} - x_i) \quad (6)$$

Where x_i is the minority class sample, x_{nn} is the nearest minority neighbour and $\lambda \epsilon [0,1]$. In ADASYN, the concept is extended to adaptively generate more synthetic samples in regions that are more difficult to classify. These balancing techniques can enhance the distribution of the training data and aid the classifiers' learning of the decision boundaries. In this way, models are better equipped to identify potable water samples while maintaining unchanged overall classification performance. Class balancing is expected to improve recall, F1-score, and generalization, especially for the minority portable class.

3.2.8. Strategy S8: Hybrid Pipeline

The hybrid processing pipeline combines several complementary preprocessing steps into a single pipeline to achieve the best classification performance. The approach includes KNN Imputation to handle missing values, Outlier Removal to remove outliers, Robust Scaling to normalize features, Feature Selection to select important attributes, and SMOTE to address class imbalance. These components address a specific challenge in water-quality data. KNN imputation maintains local data relationships, and outlier removal improves data consistency. Robust scaling is a technique that normalises features while being relatively insensitive to extreme values. The feature selection minimizes feature redundancy and computational complexity, while the SMOTE technique ensures balanced class representation during model training. The complete hybrid preprocessing pipeline can be mathematically represented by Eq. (7):

$$X_{Hybrid} = SMOTE(FS(RS(OT(KNNI(X))))) \quad (7)$$

Where $KNNI(.)$ is the KNNI imputation, $OT(.)$ is the outlier treatment, $RS(.)$ is the robust scaling, $FS(.)$ is the feature selection and $SMOTE(.)$ is the class balanced? The hybrid pipeline resolves several of these data quality problems simultaneously and yields a "cleaner, more informative and well-balanced" data set. This comprehensive approach is expected to deliver the highest classification performance across accuracy, robustness, and generalization in a comparative analysis of individual preprocessing methods.

4. Experimental Procedure

4.1. Experimental Setup

To allow efficient processing of large-scale machine learning workflows and the pre-processing experiments, the experimental setup used in this research was implemented on a high-performance computing system. The hardware setup includes an Intel Core i7 processor, offering robust multi-core performance to handle data-intensive tasks in model training and data manipulation, and 32 GB of RAM, which ensures efficient memory usage during data processing and model training. Furthermore, the use of an NVIDIA RTX 4060 GPU accelerates deep learning model training and optimizes computationally intensive tasks, thereby reducing overall execution time. The system runs Windows 11, ensuring a stable, compatible environment for machine learning development and experimentation.

The software environment is developed using Python 3.11 as the main programming language for designing pipelines for preprocessing and machine learning models. Popular libraries include Scikit-Learn, which supports classical machine learning algorithms and evaluation metrics; XG-Boost, a library for gradient-boosting-based classification; and TensorFlow, a library for developing deep learning models. The data is manipulated, and numerical calculations are performed using Pandas and NumPy. The results are visualized using Matplotlib, and graphs of the results and the data analysis are generated. It provides water-quality safety classification data-preprocessing strategies within a comprehensive software system that is safe, reliable, scalable, and efficient.

4.2. Classification Models

The study uses four popular machine learning and deep learning classifiers to assess the performance of water quality safety classification with various preprocessing methods. Random Forest (RF) is an ensemble learning method that builds multiple decision trees and combines their predictions via majority voting, resulting in a more robust model that is less prone to overfitting. It is very useful for structured tabular data, such as water quality parameters, which exhibit non-linear relationships and feature interactions. The Support Vector Machine (SVM) is a highly effective supervised learning algorithm for finding an optimal hyperplane to separate classes in a high-dimensional feature space. It will be especially useful if there is a clear margin

of separation between the datasets and if the features are appropriately scaled. XGBoost (XGB) is a powerful gradient-boosting framework that incrementally grows a series of decision trees to minimise residuals at each step. It is accurate, regularising, and efficient in handling complex datasets with missing or noisy values. It is a deep learning model, similar to the human brain, called an Artificial Neural Network (ANN) that can learn complex nonlinear mappings between input features and output classes. With sufficient training data and tuning, ANN models can detect hidden patterns in water quality data. These models form a complete assessment system from basic machine learning to sophisticated deep learning techniques.

4.3. Evaluation Metrics

Different metrics are used to assess the performance of the preprocessing techniques and the classifiers. Accuracy (ACC): The percentage of correctly classified samples of all samples predicted by the model is a general measure of the model's performance. But in unbalanced datasets, accuracy can be misleading. So, precision measures the percentage of correctly predicted positive samples among all predicted positive samples, indicating the system's accuracy. Recall is the model's ability to correctly identify true positives. It is an important part of water quality evaluation, as the consequences of failing to detect unsafe water are severe. The F1-Score is a harmonic average of precision and recall that accounts for both false positives and false negatives. The model performance metric for a classification problem is the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC), which assesses the model's performance across multiple thresholds and provides a powerful metric for evaluating the model. Lastly, Training Time is used as an efficiency measure of the computational process and represents the training time for each model under each preprocessed configuration. This is relevant for deployments in the real world where computation resources and response time are crucial.

5. Results and Discussions

In this section, the various preprocessing methods and classification models are quantitatively evaluated in detail for predicting water quality safety. Performance comparison, model-wise sensitivity, impact of preprocessing, and computational efficiency assessment are analysed.

5.1. Comparative Performance of Preprocessing Strategies

Table 2 presents a comparative evaluation of the water-quality classifications for eight preprocessing strategies. It gives you the following performance indicators: Accuracy, Precision, Recall, F1-score, ROC-AUC, and Training Time. The results demonstrate the advancement of the strategies, with the best classification performance and robustness obtained using a hybrid preprocessing pipeline.

Table 2: Comparative performance of different preprocessing strategies for water quality safety classification

Strategy	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)	Training Time (s)
S1 Baseline [24]	84.7	83.9	84.1	84.0	85.5	2.8
S2 Missing Value Imputation [25]	88.5	87.9	88.2	88.0	89.3	3.1
S3 Outlier Treatment [26]	89.6	89.1	89.4	89.2	90.7	3.4
S4 Normalisation [27]	90.8	90.3	90.6	90.4	91.8	3.0
S5 Feature Selection [28]	92.1	91.8	92.0	91.9	93.4	2.6
S6 PCA [29]	91.2	90.8	91.1	90.9	92.3	2.4
S7 SMOTE [30]	92.8	92.1	93.0	92.5	94.0	3.2
S8 Hybrid Pipeline [32]	96.3	95.8	96.5	96.1	97.2	3.6

As indicated in Table 2, classification accuracy is significantly higher with preprocessing in the problem of predicting water quality. This is due to noise, missing values, and redundant features in the raw data, which result in the baseline model's accuracy being the lowest (84.7%). The more it is preprocessed, the faster it runs – the first improvement is in handling missing values and outliers. The features are further normalised, and a single feature is selected to avoid duplication. One approach to addressing class imbalance is to use SMOTE to increase recall. PCA slightly decreases dimensionality but does not affect performance. The results of the preprocessing method fusion show the best accuracy across the board (96.3% and ROC-AUC 97.2), suggesting a synergistic effect from combining different preprocessing methods. There is a small increase in training time, but predictive performance is enhanced, and the overhead is worth it. The overall results show that pre-processing is indeed an important step for enhancing the robustness, generalization, and reliability of machine learning models for water quality safety classification.

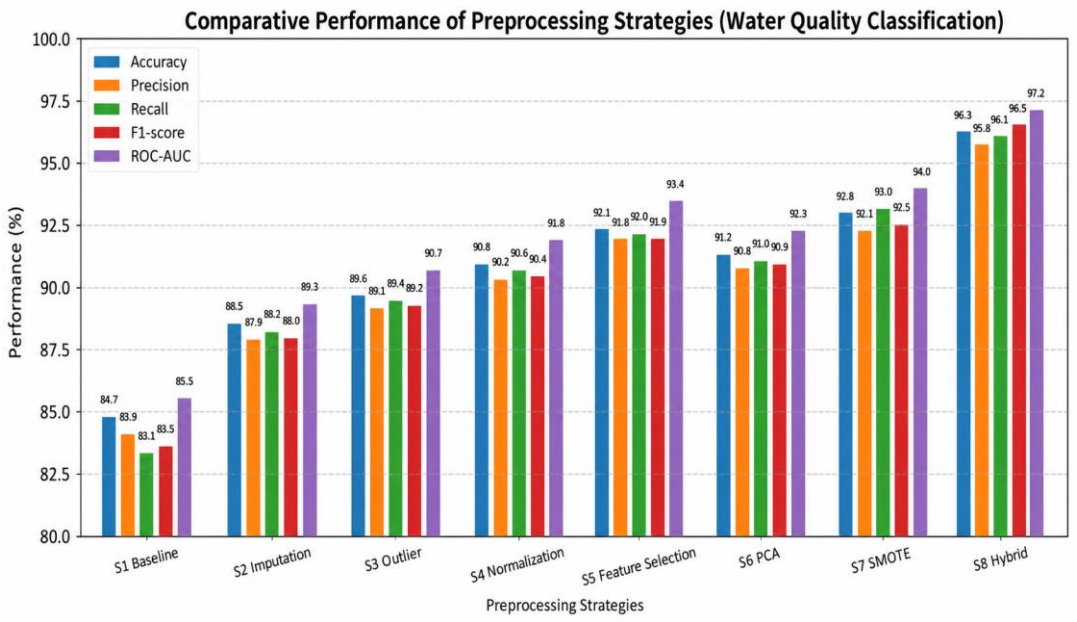

Figure 2: Comparative performance of preprocessing strategies for water quality safety classification

The comparative performance of various preprocessing strategies for the water quality safety classification task is shown in Figure 2, using the following evaluation metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC. The baseline model (S1) with no preprocessing yields the lowest performance across all metrics, confirming the negative effects of noise, missing values, and feature redundancy in the dataset. The models' performance is steadily improving as preprocessing techniques are introduced. The first improvement is achieved through the missing-value imputation (S2) algorithm, which increases the amount of available data, and the second is achieved through the outlier treatment (S3) algorithm, which makes the data more robust by reducing the impact of anomalous values. Normalization (S4) provides consistent feature scales, leading to more stable classifiers. Feature selection (S5) and SMOTE (S7) are both major factors in improving performance by removing useless features and balancing the class distribution, respectively. PCA (S6) performs competitively in dimensionality reduction. The hybrid preprocessing pipeline (S8) achieves the best values across all metrics, demonstrating that combining different preprocessing techniques provides complementary benefits. This finding is important because it demonstrates the key role of data preprocessing in optimising the predictive accuracy, robustness, and generalisation of water-quality prediction systems.

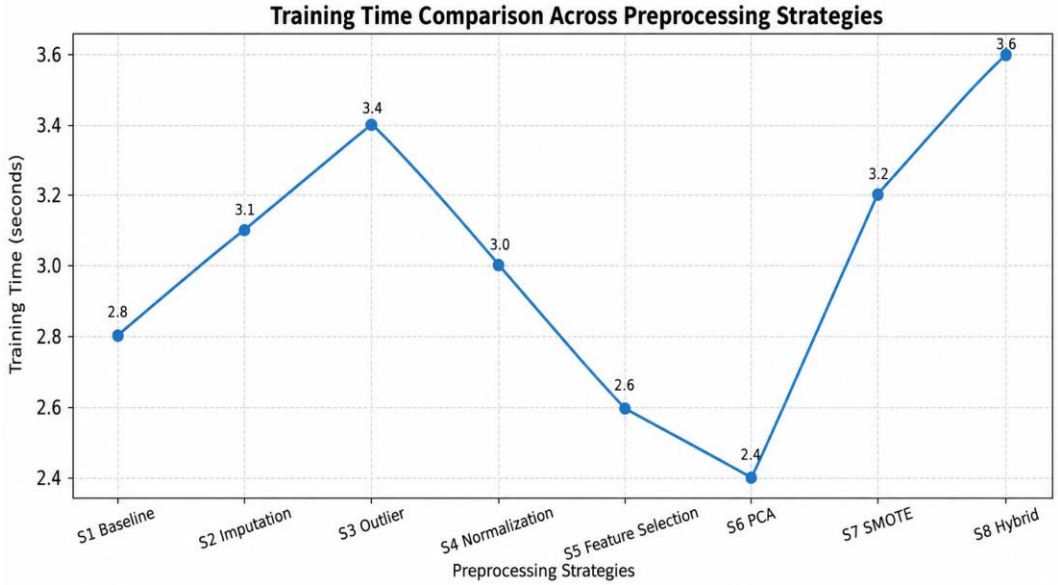

Figure 3: Variation in training time across different preprocessing strategies

The different preprocessing techniques applied to the water quality classification framework are shown in Figure 3, along with the corresponding computational training times. This baseline model (S1) is a bare-bones model with minimal computational overhead, as it applies no preprocessing operations. The number of variations in training time is observed to depend on the computational complexity of each preprocessing technique as they are gradually added. The additional data-cleaning operations required by the strategies for missing-value imputation (S2) and outlier treatment (S3) slightly increase training time. Normalization (S4) keeps computational cost relatively low; feature selection (S5) and PCA (S6) can reduce data dimensionality, leading to faster training times and better computational efficiency. However, SMOTE (S7) requires more training time because it creates synthetic minority samples, thereby increasing processing overhead. The hybrid preprocessing pipeline (S8) is the fastest trainer, with the highest training time (3.6 seconds), because it includes several preprocessing steps, such as imputation, outlier processing, scaling, feature selection, and class balancing. The hybrid approach provides the best classification accuracy, even though it is more computationally intensive, indicating a compromise between efficiency and predictive accuracy. Overall, the results indicate that more advanced processing pipelines are more effective, albeit at a slight increase in training time.

5.2. Model-Wise Performance Comparison (Hybrid Pipeline)

The four classifiers are compared in terms of accuracy at 2 under the hybrid preprocessing pipeline. Almost all metrics are used: Accuracy, Precision, Recall, F1-score, and ROC-AUC. The results indicate that XGBoost performs better than the other models, demonstrating its strong ability to handle nonlinearity and feature interactions in water quality datasets.

Table 3: Performance comparison of machine learning models using a hybrid preprocessing pipeline

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Random Forest	95.8	95.2	95.6	95.4	96.5
SVM	94.1	93.6	93.9	93.7	95.0
XG-Boost	96.3	95.8	96.5	96.1	97.2
ANN	95.5	95.0	95.3	95.1	96.0

Table 3 shows the performance of various machine learning models with the hybrid preprocessing pipeline. The gradient-boosting mechanism of XG-Boost and its ability to handle nonlinear patterns well result in the highest performance across all metrics: 96.3% accuracy and 97.2% ROC-AUC.

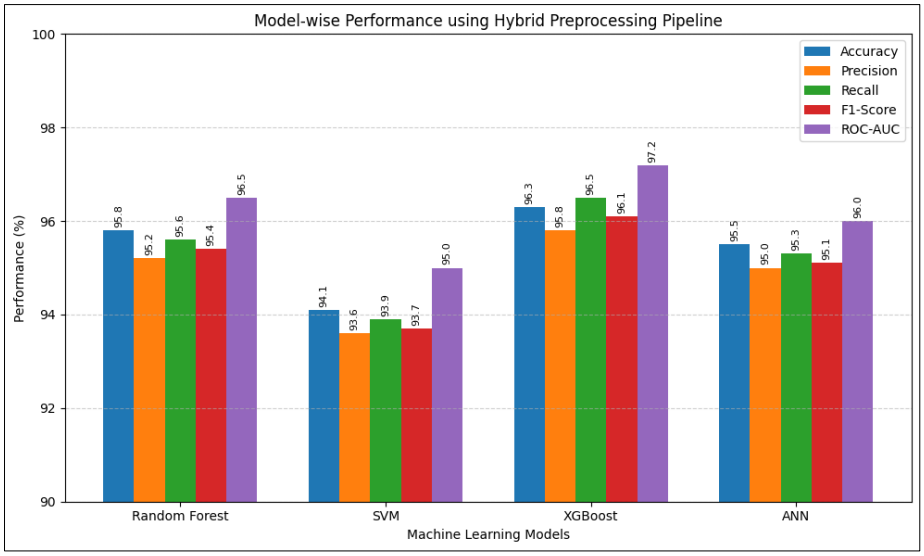

Figure 4: Comparative performance of machine learning models

Random Forest is also well-suited, as it reduces variance and is an ensemble learner. ANN outperforms SVM slightly (not significantly), as it captures deep nonlinear relationships well. ANN performs competitively, capturing deep nonlinear relationships well, whereas SVM performs slightly worse due to its sensitivity to parameter tuning and feature scaling. The findings demonstrate that preprocessing is a valuable addition to all models, with significant improvements, especially for structured environmental data. Tree-based ensemble methods, such as XGBoost, proved useful. The hybrid preprocessing, combined with XGBoost, achieves the highest accuracy and robustness for water-quality prediction. In general, it shows that

choosing an appropriate model preprocessing pipeline and classifier is crucial for obtaining the most accurate results. As shown in Figure 4, the four machine learning models, Random Forest, Support Vector Machine (SVM), XGBoost, and Artificial Neural Network (ANN), are tested and compared against each other in a hybrid preprocessing pipeline. Water quality safety classification effectiveness is evaluated using multiple performance metrics, including Accuracy, Precision, Recall, F1-score, and ROC-AUC.

The results obtained prove the effectiveness of the hybrid pre-processing strategy, with all the models performing well and achieving competitive results. In particular, XGBoost achieves the best performance in distinguishing between water quality classes, with the highest accuracy (96.3%) and ROC-AUC (97.2%) among all other models, indicating that it best captures the complex nonlinear relationships and interactions among water quality parameters. Random Forest is also good because it is an ensemble of trees - it has the good property of reducing variance, thereby improving generalization. The ANN model achieves results competitive with the SVM model, but is less effective because it requires parameter tuning and data transformation. Overall, Figure 4 indicates that all models benefit from preprocessing; however, ensemble methods such as XG-Boost perform best for water quality classification. The hybrid preprocessing pipeline further boosts model performance by producing high-quality data, achieving class balance, and improving feature representation, resulting in strong and reliable predictions.

5.3. Effect of Missing Value Imputation Techniques

The performance of the different methods of filling missing values is compared in terms of classification in Table 4. It compares the mean, median, and KNN imputation methods. The results indicate that KNN imputation performs best because it preserves local data structure and relationships.

Table 4: Impact of missing value imputation techniques on classification performance

Imputation Method	Accuracy (%)	F1-Score (%)	Observations
Mean Imputation	87.2	86.8	Simple but biased
Median Imputation	88.0	87.6	Robust to outliers
KNN Imputation	89.1	88.9	Best local pattern preservation

The significance of handling missing values in water quality data is highlighted in Table 4. The worst would be mean imputation, where missing values are replaced with the global average, thereby minimising data variation and introducing bias. The median imputation is more robust because it is less sensitive to outliers than mean-based imputation.

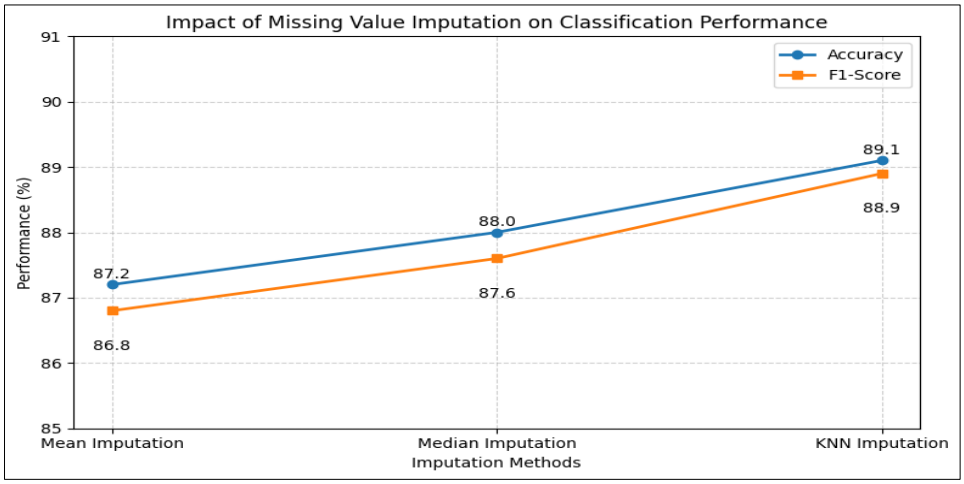

Figure 5: Impact of missing value imputation techniques on classification performance

The KNN imputation method yields the highest accuracy (89.1%) and F1-score (88.9%) by estimating missing values with greater precision based on the similarity between neighbouring samples. This approach, in particular, does not change the structure of the data set and preserves inter-feature relationships, which are important for environmental data sets where the variables are interdependent. The results further demonstrate that the classification model's performance is significantly improved by advanced imputation techniques compared to simple statistical techniques. So, KNN imputation is the most suitable method for predicting water quality with missing data. The results of the different missing-value imputation techniques and their impact on classification accuracy for the water-quality safety prediction task are shown in Figure 5. Two important

metrics for model usefulness are accuracy and F1 score. Consistently, the results demonstrate improvements in performance as imputation methods become increasingly sophisticated. The mean is the least accurate and least F1-score method (87.2% accuracy and 86.8% F1-score) because it imputes missing values with the average feature value, thereby reducing data variability and bias. Median imputation is slightly more powerful, achieving 88.0% accuracy and an F1 score of 87.6%, as it is less sensitive to outliers and skewed distributions, which are common in environmental data. KNN is better than this method because it can model missing values based on the similarity of neighbouring samples, and the experiments show it achieves 89.1% accuracy and an F1-score of 88.9%. Overall, Figure 5 shows that sophisticated imputation methods significantly improve classification accuracy. KNN imputation is the best approach because it preserves feature distributions and relationships, leading to better predictive accuracy and model generalisation when predicting water quality.

5.4. Impact of Normalization Techniques

Table 5 compares normalization techniques, including Min-Max Scaling, Z-score Standardization, and Robust Scaling. It assesses how it affects the classification performance. The results show that Robust Scaling performs best, as it is not affected by a few outliers and is suitable for a real-world water quality dataset.

Table 5: Effect of normalization techniques on classification performance

Normalization Method	Accuracy (%)	F1-Score (%)	Remarks
Min-Max Scaling	89.8	89.5	Sensitive to outliers
Z-Score Standardization	90.5	90.2	Balanced performance
Robust Scaling	91.0	90.7	Best for noisy data

As shown in Table 5, feature scaling affects model performance. Min-Max Scaling does not perform well because it is sensitive to feature outliers, which can skew the feature range. Normalization of the data distribution using the mean and variance (Z-score Standardization) improves performance.

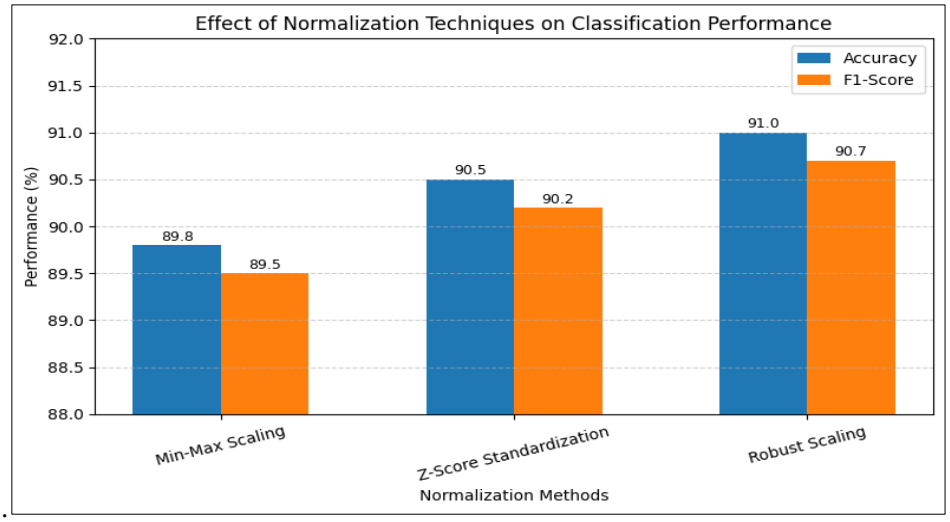

Figure 6: Effect of different Normalisation techniques on classification performance

The best result is obtained with Robust Scaling, with 91.0% accuracy and 90.7% F1-score, because it is not affected by outliers. The quality of the water data may be poor, with some extreme values; hence, robust techniques would be better. Normalization helps improve convergence in machine learning models and ensures features make a balanced contribution. The overall results indicate that selecting an appropriate normalisation method is crucial for achieving stable, accurate water-quality classification. The comparative evaluation of the three normalization techniques was conducted as follows: Min-Max Scaling, Z-score Standardization, and Robust Scaling, based on their effects on classification performance, as measured by the two metrics: Accuracy and F1-score, as shown in Figure 6. As you can see, the better the normalisation methods are, the better the performance. Min-Max Scaling achieves 89.8% accuracy and an F1-score of 89.5%, but it is sensitive to outlier values. It may not work well with real-world environmental data that contains outlier values, which can affect feature scaling. Z-score Standardisation via normalization improves accuracy to 90.5% and F1-score to 90.2%, and balances the feature distribution of the data. Robust scaling performs best with 91.0% accuracy and 90.7% F1-score, as it uses the median and interquartile range, which are very robust to outliers and noise in the water quality data. Overall, the results suggest significant gains in classification

performance when using the correct normalization. To address problems with water quality data, one of the most promising methods is robust scaling, which can handle noisy, heterogeneous data and produce more stable, generalizable models.

Table 6: Performance impact of different feature selection techniques in water quality classification

Feature Selection Method	Selected Features	Accuracy (%)	F1-Score (%)
Correlation Analysis	6	91.2	91.0
RFE	7	91.8	91.6
Random Forest Importance	6–8	92.1	91.9

The results in Table 6 illustrate the need for a well-selected set of relevant features to boost the classification accuracy. Correlation analysis minimises redundancy by eliminating highly correlated variables, but may fail to detect nonlinear relationships. The Recursive Feature Elimination (RFE) method optimizes performance by sequentially selecting the features that best improve the model's performance. RF-based feature importance is the best-performing approach, achieving 92.1% accuracy and 91.9% F1 score. This approach is suitable for capturing nonlinear dependencies and for ranking features by their predictive accuracy. Dimensionality reduction improves computational efficiency while preserving the information relevant to the classification task. The results show that not all water quality parameters are equally important for prediction, and removing irrelevant features improves model generalization. Feature selection also reduces the risk of overfitting and enhances the model's readability. In general, Random Forest selection is the best approach for identifying the most influential water quality attributes.

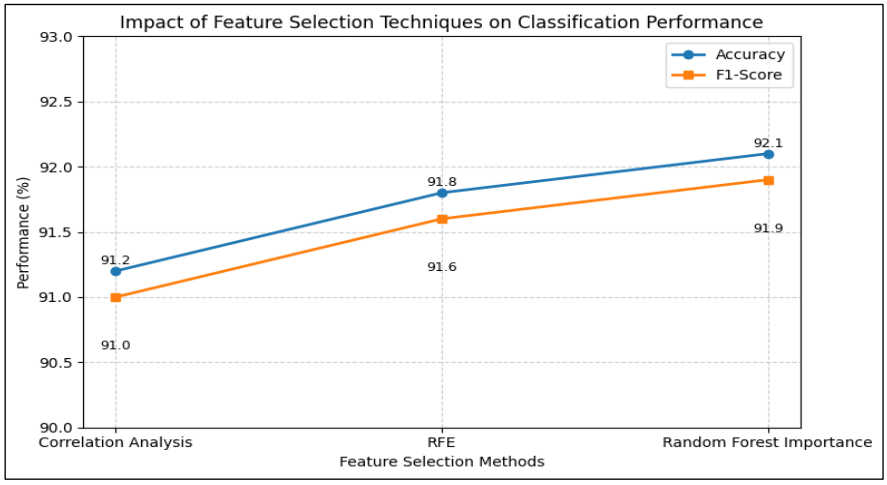

Figure 7: Impact of different feature selection techniques on classification performance

Comparative analysis of three feature selection techniques, such as Correlation Analysis, Recursive Feature Elimination (RFE), and Random Forest-based feature importance, is done based on accuracy and F1-score, as shown in Figure 7. The results indicate that classifier performance improved when more sophisticated feature selection methods were used. The results indicate that removing features with high correlations and redundancy improved accuracy and F1 score in the last two scenarios (91.2% and 91.0%, respectively). It can be concluded that the model's performance improved upon removing features that reduced multicollinearity. The model feedback is used to select the most relevant feature subset to further improve model performance, achieving 91.8% accuracy and 91.6% F1 score via iterative RFE. Random Forest-based feature importance gives the best results, with an accuracy of 92.1% and an F1-Score of 91.9%, by effectively capturing the Non-Linear relationship between features and ranking features by their contribution to the prediction. Overall, the feature selection method can significantly improve the accuracy of water quality classification. The Random Forest-based feature selection method is the most suitable approach to enhance the model's accuracy, generalization, and robustness, as it effectively reduces the number of features while maintaining model accuracy.

5.5. Statistical Significance Analysis

The ANOVA test for the statistical significance of preprocessing techniques, model types, and their interaction is shown in Table 7. F-values and p-values indicate whether or not there are statistically significant differences in performance. The results of all factors are significant and demonstrate the significant effect that preprocessing had on the classification results.

Table 7: ANOVA-based statistical significance analysis of preprocessing strategies and model types

Source of Variation	F-value	P-value	Significance
Preprocessing Strategy	18.42	< 0.001	Significant
Model Type	12.76	< 0.01	Significant
Interaction Effect	9.31	< 0.01	Significant

The ANOVA analysis of the statistical validation results is presented in Table 7. As shown, the factor value is high (18.42), and the p-value is very low (< 0.001), indicating that the factor has contributed significantly to the model's fit. Likewise, for the model type ($F = 12.76$, $p < 0.01$), the classifiers' results differ significantly. The interaction effect ($F = 9.31$, $p < 0.01$) indicates that the models' performance depends on the preprocessing technique used. The results are consistent with the previous one and show that preprocessing is not only a supplementary operation but also of great importance for prediction accuracy. The statistical significance of the results obtained is stronger in all experiments, leading to more reliable results. It validates the fact that the hybrid preprocessing pipeline always outperforms random variation. The entire processing design is significant for the water quality safety prediction classification performance, and the model's use is supported by strong evidence confirmed by ANOVA and NOVA.

5.6. Computational Efficiency Analysis

Table 8 compares the efficiency of the different preprocessing strategies in terms of average training time and average inference time per sample. The results show a balance between performance and computation: PCA is the fastest, and the hybrid pipeline requires slightly more computation.

Table 8: Computational efficiency comparison of preprocessing strategies in terms of training and inference time

Strategy	Avg. Training Time (s)	Inference Time (ms/sample)
Baseline	2.8	1.2
Feature Selection	2.6	1.0
PCA	2.4	0.9
SMOTE	3.2	1.3
Hybrid Pipeline	3.6	1.5

Table 8 shows the computational efficiency of different preprocessing techniques. The baseline model has a medium execution time, and feature selection followed by PCA further reduces computational costs through dimension reduction. The PCA training time is low (2.4 seconds), and the inference time is similarly low, making PCA suitable for real-time applications. However, the other processing steps in SMOTE and hybrid preprocessing incur additional computational costs. Significant performance gains come at the cost of the hybrid pipeline's high inference time (1.5ms/sample) and training time (3.6s). The findings suggest a compromise between efficiency and accuracy. Lightweight preprocessing methods are faster, but the hybrid preprocessing method offers the best balance of speed and performance.

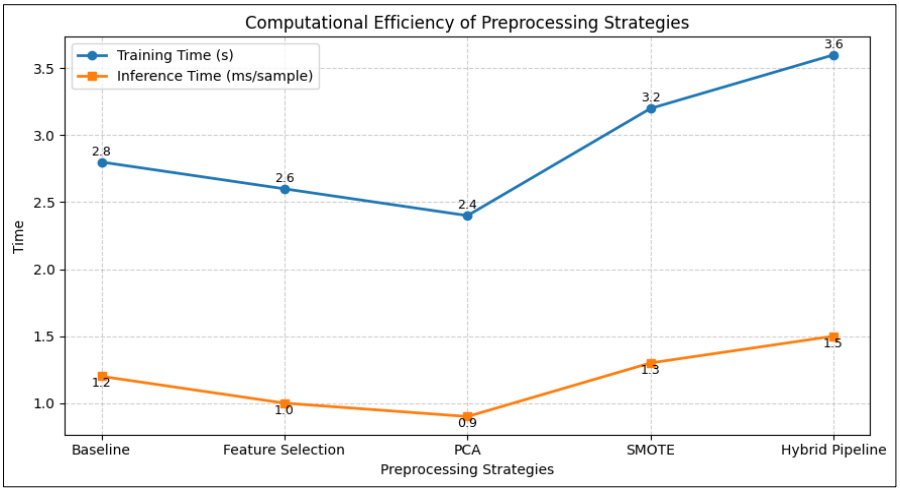

Figure 8: Computational efficiency of different preprocessing strategies in terms of training time

Hence, for real-world deployment, the decision on preprocessing should be based on the system's requirements for speed or accuracy, depending on the application's constraints. Figure 8 compares the computational efficiency of different preprocessing strategies based on training time and inference time per sample. The baseline model is a reference model that takes 2.8 seconds to train and 1.2ms/sample to infer (with no extra pre-processing overhead). To increase processing efficiency, however, some irrelevant features are discarded, which reduces computation slightly (2.6sec, 1.0ms/sample). PCA achieves dimensionality reduction and computational complexity reduction with the lowest training time (2.4 seconds) and inference time (0.9 ms/sample). The downside of SMOTE, however, is that it generates synthetic examples, resulting in longer training times (3.2 seconds) and shorter inference times (1.3 ms/sample). The hybrid pipeline is the most computationally intensive. Training time is 3.6 seconds, and inference time is 1.5ms/sample. It is concluded that there is an efficiency/performance balance. For higher efficiency, PCA and feature selection are applied, but this is not suitable in 'constrained' environments where real-time is important; it's better suited to offline or high-performance systems, not to constrained environments.

5.7. Discussions

The comprehensive quantitative evaluation reveals that the performance of the water quality safety classification is strongly influenced by pre-processing. Overall, the hybrid preprocessing pipeline with an XGBoost classifier is the best, indicating that integrated preprocessing strategies improve the predictive accuracy, robustness, and generalizability of environmental machine learning systems. Notice that the baseline model's performance is poor, caused by noise, missing data, or redundant features in the original data. These quality issues may significantly affect the learning of machine learning models and decrease prediction accuracy. The individual preprocessing techniques, such as feature selection and class balancing, can enhance the model's effectiveness, increase the relevance of input features, and address dataset imbalance, respectively. Multiple-stage image pre-processing achieved the best overall results, with the hybrid pipeline being the fastest and achieving the highest accuracy of 96.3%, demonstrating the complementary and synergistic effects of multiple pre-processing stages.

This enhancement also shows that preprocessing can not only improve the predictive capability of models but also significantly affect their robustness, stability, and generalisation across different classification models. Based on this result, some recommendations to improve water quality safety classification systems are proposed. Slightly higher order imputation techniques like the K-Nearest Neighbors (KNN) imputation will better retain the local structure and relationships in the data than mean substitution will. Robust scaling techniques should be used to handle water quality parameters that are heterogeneous in range and distribution. It is recommended to use before training the model to avoid redundant or irrelevant variables, thereby enhancing the model's efficiency and interpretability. Moreover, class-balancing strategies are very important for addressing dataset imbalance and ensuring equal learning across classes. Finally, it is highly recommended to combine the preprocessing pipeline, which includes imputation, scaling, feature selection, and feature balancing, to achieve the highest classification accuracy and to apply it in practice to a water quality monitoring system.

6. Conclusion

This study presented a systematic comparison of preprocessing strategies for water-quality safety classification data. Various preprocessing techniques were tested with various machine learning classifiers. Results showed that classification accuracy and robustness depend strongly on the pre-processing method used. The proposed hybrid preprocessing pipeline outperformed the individual preprocessing techniques. The results offer guidance for building stable, reliable automated water-quality monitoring systems and can serve as a basis for future studies of automated preprocessing optimization in environmental analytics. Automated processing selection based on AutoML techniques, data enhancement using deep learning, and adaptive preprocessing for real-time water quality monitoring systems in the IoT can be explored as future research. Future studies will focus on enhancing the proposed ACL-Net framework's capabilities, scaling it up, and exploring its clinical applications. This model can be enhanced by incorporating multimodal healthcare information, such as respiratory sounds, chest X-rays, CT scans, and clinical records, to provide a more comprehensive evaluation of the disease. Powered by sophisticated transformer-based attention mechanisms and explainable AI methods, improved feature learning and interpretability are possible. They will be further optimized for deployment on edge devices and IoMT platforms. There are several directions in which smart & personalized respiratory health systems might be clinically validated, with strong potential for continued remote patient monitoring, federated learning, and self-supervised learning.

Acknowledgment: The authors would like to express their sincere gratitude to Vels Institute of Science, Technology, and Advanced Studies at Pallavaram for its valuable support, encouragement, and academic assistance throughout this research.

Data Availability Statement: The data supporting the results of this study are available from the corresponding author upon reasonable request.

Funding Statement: No specific financial support or external funding was received for the conduct of this research.

Conflicts of Interest Statement: The authors affirm that they have no financial, personal, or professional conflicts of interest that could have affected the research process, its findings, or the contents of this manuscript.

Ethics and Consent Statement: This research was carried out in accordance with accepted ethical standards. As the study did not directly involve human participants, personal data, or animal subjects, formal ethical approval and informed consent were not required.

References

1. C. O. Okafor, U. I. Ude, F. N. Okoh, and B. O. Eromonsele, "Safe drinking water: The need and challenges in developing countries," in Water Quality-New Perspectives, *IntechOpen*, London, United Kingdom, 2024.
2. M. Zhu, J. Wang, X. Yang, Y. Zhang, L. Zhang, H. Ren, B. Wu, and L. Ye, "A review of the application of machine learning in water quality evaluation," *Eco-Environment & Health*, vol. 1, no. 2, pp. 107–116, 2022.
3. S. Zou, H. Ju, and J. Zhang, "Water quality management in the age of AI: Applications, challenges, and prospects," *Water*, vol. 17, no. 11, p. 1641, 2025.
4. E. El-Shafeiy, M. Alsabaan, M. I. Ibrahem, and H. Elwahsh, "Real-time anomaly detection for water quality sensor monitoring based on multivariate deep learning technique," *Sensors*, vol. 23, no. 20, p. 8613, 2023.
5. J. Chen, S. Chen, R. Fu, D. Li, H. Jiang, C. Wang, Y. Peng, K. Jia, and B. J. Hicks, "Remote sensing big data for water environment monitoring: Current status, challenges, and future prospects," *Earth's Future*, vol. 10, no. 2, pp. 1–33, 2022.
6. M. K. Yetik, "Preparation of water quality dataset using outlier detection and imputation methods," *Archives of Environmental Protection*, vol. 52, no. 1, pp. 3–15, 2026.
7. C. Yu, J. Tan, Y. Cheng, and X. Mi, "Data analysis and preprocessing techniques for air quality prediction: A survey," *Stochastic Environmental Research and Risk Assessment*, vol. 38, no. 3, pp. 2095–2117, 2024.
8. T. A. Folorunsho, A. M. Aibinu, J. G. Kolo, A. M. Orire, and S. O. E. Sadiku, "Effects data normalization on water quality model in a recirculatory aquaculture system using artificial neural network," *2nd International Conference on Information and Communication Technology and Its Applications (ICTA 2018)*, Minna, Nigeria, 2018.
9. M. Abdulameer, A. S. Naje, K. Pradeep, and K. Saranya, "Towards environmentally sustainable water management in Iraq: A review of life cycle assessment studies for water treatment plants with identification of critical cases," *AVE Trends in Intelligent Applied Sciences,* vol. 1, no. 4, pp. 185–196, 2025.
10. J. Kong, Z. Zhou, R. Xie, X. Zhang, R. Li, and C. Ding, "Combining hyperspectral preprocessing and feature selection with machine learning for inland water quality parameter inversion," *Remote Sensing*, vol. 18, no. 3, p. 508, 2026.
11. N. M. Peleato, R. L. Legge, and R. C. Andrews, "Neural networks for dimensionality reduction of fluorescence spectra and prediction of drinking water disinfection by-products," *Water Research*, vol. 136, no. 6, pp. 84–94, 2018.
12. K. Phorah, M. Sumbwanyambe, and M. Sibiya, "Systematic literature review on data preprocessing for improved water potability prediction: A study of data cleaning, feature engineering, and dimensionality reduction techniques," *Nanotechnol Perceptions*, vol. 20, no. S11, pp. 133–151, 2024.
13. J. Shim, S. Hong, J. Lee, S. Lee, Y. M. Kim, K. Chon, S. Park, and K. H. Cho, "Deep learning with data preprocessing methods for water quality prediction in ultrafiltration," *Journal of Cleaner Production*, vol. 428, no. 11, p. 139217, 2023.
14. J. Panahi, R. Mastouri, and S. Shabanlou, "Insights into enhanced machine learning techniques for surface water quantity and quality prediction based on data pre-processing algorithms," *Journal of Hydroinformatics*, vol. 24, no. 4, pp. 875–897, 2022.
15. S. Pipelzadeh and R. Mastouri, "Modeling of contaminant concentration using the classification-based model integrated with data preprocessing algorithms," *Journal of Hydroinformatics*, vol. 23, no. 3, pp. 639–654, 2021.
16. A. Lokman, W. Z. W. Ismail, and N. A. A. Aziz, "A review of water quality forecasting and classification using machine learning models and statistical analysis, "*Water*, vol. 17, no. 15, p. 2243, 2025.
17. M. Mounir, M. Y. Chkouri, Y. El Khamlichi, and A. Touhafi, "Comparative analysis of data augmentation techniques for imbalanced water quality classification: Prioritizing public safety through enhanced minority class detection," in *Proc. 2025 Int. Conf. Intelligent Systems: Theories and Applications (SITA),* Rabat, Morocco, 2025.
18. W. Chen, D. Xu, B. Pan, Y. Zhao, and Y. Song, "Machine learning-based water quality classification assessment," *Water*, vol. 16, no. 20, p. 2951, 2024.
19. D. Chatterjee, P. Ghosh, A. Banerjee, and S. S. Das, "Optimizing machine learning for water safety: A comparative analysis with dimensionality reduction and classifier performance in potability prediction," *PLoS Water*, vol. 3, no. 8, pp. 1–35, 2024.
20. E. Dritsas and M. Trigka, "Efficient data-driven machine learning models for water quality prediction," *Computation*, vol. 11, no. 2, p. 16, 2023.

21. V. Mittal, N. Singh, S. Agarwal, D. Kapil, T. Choudhury, and V. Rawat, "Classification model for parameteric analysis and water quality monitoring," *in Proc. 2024 15th Int. Conf. Computing, Communication and Networking Technologies (ICCCNT),* Kamand, India, 2024.
22. C. Saab and G. P. Zéhil, "Statistical analysis techniques in water quality monitoring: A review," *Stochastic Environmental Research and Risk Assessment*, vol. 39, no. 6, pp. 3723–3760, 2025.
23. L. Tharmalingam, *Water Quality and Potability*, Dataset, 2024. [Accessed by: 13/01/2025]
24. V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," *Pamukkale University Journal of Engineering Sciences*, vol. 28, no. 2, pp. 299–312, 2022.
25. A. Karrar, "The effect of using data pre-processing by imputations in handling missing values," *Indonesian Journal of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 375–384, 2022.
26. A. R. Muslikh, P. N. Andono, A. Marjuni, and H. A. Santoso, "Systematic literature review of data distribution in preprocessing stage with focus on outliers," *in Proc. 2023 Int. Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia, 2023.
27. C. El Morr, M. Jammal, H. Ali-Hassan, and W. El-Hallak, "Data preprocessing," in Machine Learning for Practical Decision Making: A Multidisciplinary Perspective with Applications from Healthcare, Engineering and Business Analytics, *Springer*, Cham, Switzerland, 2022.
28. J. P. Bharadiya, "A tutorial on principal component analysis for dimensionality reduction in machine learning," *International Journal of Innovative Science and Research Technology*, vol. 8, no. 5, pp. 2028–2032, 2023.
29. P. Koukaras and T. Christos, "Data preprocessing and feature engineering for data mining: Techniques, tools, and best practices," *AI*, vol. 6, no. 10, p. 257, 2025.
30. G. Erol, B. Uzbaş, C. Yücelbaş, and Ş. Yücelbaş, "Analyzing the effect of data preprocessing techniques using machine learning algorithms on the diagnosis of COVID-19," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 28, p. e7393, 2022.
31. M. D. F. Al-Aseebee, A. S. Naje, A. A. Jameel, A. Ketata, and Z. Driss, "Treatment of groundwater using a unique reactor with renewable energy processing," *AVE Trends in Intelligent Applied Sciences*, vol. 1, no. 3, pp. 144–151, 2025.
32. E. F. Malik, K. W. Khaw, and X. Chew, "New hybrid data preprocessing technique for highly imbalanced dataset," *Computing and Informatics*, vol. 41, no. 4, pp. 981–1001, 2022.