National College of Ireland

Enhancing Water Quality Prediction Through Explainable Machine Learning

MSc Research Project
MSc Data Analytics

Pramod Akuthota
Student ID: 23408979

School of Computing
National College of Ireland

Supervisor: Arghir Nicolae Moldovan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Pramod Akuthota
Student ID:	23408979
Programme:	MSc Data Analytics **Year:** 2025-2026
Module:	MSc Research Practicum Part 2
Supervisor:	Arghir Nicolae Moldovan
Submission Due Date:	11-12-2025
Project Title:	Enhancing Water Quality Prediction Through Explainable Machine Learning
Word Count:	9749 **Page Count:** 26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Pramod Akuthota

Date: 11-12-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Water Quality Prediction Through Explainable Machine Learning

Pramod Akuthota
23408979

Abstract

The increasing importance of predicting water quality is a result of environmental degradation, urbanization, and the need for quick access to freshwater resources. Traditional laboratory methods for measuring water quality are inefficient, leading to a rise in demand for automated and data-driven means of assessing water quality. Machine learning (ML) and deep learning (DL) models are able to provide good predictive performance. However, their black-box nature makes it difficult to trust them in real-world applications. This paper seeks to fill that gap by integrating Random Forest with LIME in order to provide both high predictive accuracy and interpretability. The physicochemical characteristics of water are normalized using the SelectKBest based on ANOVA statistical analyses, and class imbalance is handled with SMOTE. The Random Forest model can learn how features interact with each other non-linearly. However, the interpretation of feature importance is not as clear to the end user as the output of LIME would indicate. Together, the integrated framework would enable reliable classification of water quality along with transparent decision support. This integrated framework demonstrated its ability to provide high-quality results across individual datasets from Canada, China, England, Ireland and the United States, including a score of R2 = 0.99 and a very high classification accuracy.

1 Introduction

1.1 Background

Rapid industrialization, urban growth, agricultural effluents, and hydrological alterations caused by climate have become a major international issue that has entered water quality degradation. These stresses have enhanced the variability in main physicochemical parameters of pH, turbidity, dissolved oxygen and conductivity, which has a direct impact on the health of the ecosystem and human wellbeing. Conventional laboratory-based water quality measurement approaches are usually time-consuming, resource-demanding, and incapable of offering real-time monitoring, which leaves considerable loopholes in the timely identification of pollution incidences (Rahu et al., 2023). With the growing level of stress on water resource systems, there is an urgent necessity of intelligent, automated, and predictive monitoring methods, which may aid in sustaining the management of the environment and the population health.

The recent innovations in the field of IoT sensing networks and data-driven modelling have provided the opportunities to improve the precision and regularity of water quality evaluation. But the environmental data is normally nonlinear, noisy and very dynamic and demands models that can handle complex spatial-temporal relationships. Although machine learning and deep learning have great predictive power, most models do not have interpretable insights, which restricts their applicability in policymaking scenarios, resource planning, and risk management (Dasgupta et al., 2025). It is based on these challenges that this research aims at coming up with a system that does not only work on prediction of water quality effectively, but also on the factors behind the choices made by the model, thus narrowing the gap between prediction and the practical decision support.

Various types of machine learning algorithms such as Random Forests, Support Vector Machine, and regression-based hybrid classification and forecasting water quality parameters were studied and confirmed to have a significant increase in predictive accuracy and computational efficiency (Jibrin et al., 2024; Rahu et al., 2023). As of recently, more intricate temporal patterns and nonlinearity in environmental data have been modelled using deep learning architectures, including CNN-LSTM, BiLSTM, and Gated Liquid Neural Networks, which further increases the performance of the models (Chen et al., 2025; Gohari Nezhad & Emami Niri, 2025; Singh et al., 2025)Although such advances have been made, the current models remain more of a black box, which is not fully interpretable, and it is not easy to have the domain experts to validate or trust the predictions made by them particularly in real time and critical situations.

1.2 Objectives

The paper intends to explain a predictable water quality model with good predictive capabilities and clear logic. The key objectives are:

- To create machine learning and predict water quality based on physicochemical parameters.
- To enhance quality of data and model robustness via preprocessing, outlier management and feature scaling.
- To use ANOVA-based feature selection (SelectKBest) and SMOTE-based class balancing to improve the performance of the model.
- To enhance interpretability using LIME explanations to discover influential features of predicting models.
- To test the generalizability through combined and country-wise experimental research on different regions.

Research Question

How can a machine learning and deep learning framework, strengthened through ANOVA-based feature selection and SMOTE-based class balancing, and enhanced with LIME explainability, improve both the predictive accuracy and interpretability of water quality predictions?

This research question aligns with the identified need for predictive models that not only perform well but also reveal which variables influence water quality outcomes while also using ANOVA and SMOTE in preprocessing. The use of explainable AI techniques brings greater transparency to the water-quality model by allowing environmental authorities, decision-makers, and water-resource managers to understand and trust its outputs.

Contributions of the work:

- Explanation of model predictions using LIME and the Random Forest is explained based on the features that had the greatest impact on the classification of water quality, including the relative importance of each feature value when making the classification.
- Pre-processing of model input data to guarantee model performance and the data quality is improved, dimensionality is reduced, and the class imbalance that exists between each of the water quality classes is addressed using multiple types of preprocessing (outlier analysis, MinMax scaling, ANOVA based feature selection [SelectKBest] and SMOTE.
- General benchmarking of techniques: Makes a comparative analysis of the Machine Learning (Decision Tree, XGBoost, Random Forest) and Deep Learning (ANN, LSTM) models based on regression and classification measures.
- Multi-country validation and generalizability: Testing is performed on a big dataset of five countries (Canada, China, England, Ireland, USA), combined and country-wise experiments to determine the consistency between regions.

Structure of the Report

The rest of this report is structured in the following way: Section 2 will conduct a review of related work on predicting water quality with the help of machine learning, deep learning, and explainable AI, and identify gaps in the research. Section 3 describes the dataset, preprocessing (scaling, ANOVA feature selection, SMOTE), model development, and explainability by the use of LIME, and Section 4 presents and discusses the results in terms of performance comparison and feature-influence explainability. Lastly, Section 5 wraps up the research by summarizing the main results, limitations and future research directions.

2 Literature review

The issue of water quality prediction is becoming more significant as the environmental degradation, climatic variability, and high rate of urbanization become more dangerous to freshwater resources. As modern technologies of high sensing and data analytics are emerging, scientists have turned to the machine learning, deep learning, and hybrid intelligent systems to simulate the complex physicochemical interactions in water bodies. Recent research indicates that besides the promise of such computational techniques with respect to their ability to improve predictive accuracy, interpretability, scalability and adaptability are

also required in real-world monitoring settings. This literature review will be a synthesis of the major findings of literature available, the analysis of the new trends, the gaps in the research, which will inspire a more transparent and strong water quality prediction framework.

2.1 Machine Learning Approaches for Water Quality Prediction

Most environmental models have been developed through machine learning (ML) because it has the advantage of analysing nonlinear interactions among different water quality parameters. Traditional ML models like Support Vector machines (SVM), random forest (RF), decision tree (DT) and XGBoost have demonstrated good predictive performance in terms of classifying and assessing the water quality status. The models are useful with multidimensional, partially incomplete or noisy data, which is essential in the application of the models in the environment where the conditions of data collection are highly diverse. Research such as (Rahu et al., 2023) show that the integration of ML and into-based sensor networks can help to maintain constant monitoring, anomalies, and correct classification of water quality levels or indices.

In addition to classification, ML models that emphasize regression have been of great popularity in predicting continuous water quality variables. According to research conducted by (Jibrin et al., 2024), hybrid regression methods like Gaussian Process Regression (GPR), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and regression trees have been found to be valuable in predicting groundwater and surface water indices. These models are ideal in the representation of nonlinear trends in physicochemical variables and they provide adaptable prediction mechanisms of rules. (Shyu et al., 2023) also show that soft sensor systems based on Support Vector Regression (SVR), Partial Least Squares (PLS), and Cubist Regression are capable of estimating such parameters as chemical oxygen demand, turbidity, and suspended solids using the real-time sensor inputs. This type of regression lowers the frequency of laboratory testing, and makes water quality monitoring cost effective, reliable and suitable in decentralized and dynamic settings.

Machine learning has been significant in changing the conventional environmental monitoring systems to smart and automated systems as well. The IoT-based ML systems as described by (Rahu et al., 2023) demonstrate how the continuous data streams can be analysed in real-time and provided with near real-time evaluation of the water quality conditions. Despite the strength of the ML techniques, they can be relatively weak in capturing the temporal dependencies over longer durations and reacting swiftly to sudden changes in the patterns of the environment. This is because the number of fixed feature sets and limited memory components limits them to modelling complex and evolving ecosystems.

2.2 Deep Learning Approaches for Water Quality Prediction

Deep learning (DL) has proven to be a more effective method of modelling water quality since it is capable of automatically extracting meaningful features of raw and high-

dimensional data. Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) are models that have been demonstrated to significantly enhance predictive accuracy when used to solve tasks related to water potability classification, pollution forecasting and prediction of hydrological variables. Studies by (Dasgupta et al., 2025; Rahu et al., 2023) explain the ability of DL systems to acquire complex temporal and spatial associations that traditional ML models cannot. These architectures are capable of processing various datasets that have information on chemical, physical and meteorological data and identifying latent patterns and interactions that affect water quality. Their capacity to learn many layers of representation enables DL to be especially useful in the modelling of large-scale environmental systems that are uncertain and nonlinear.

Hybrid deep learning models have been of interest to integrate the benefits of the various types of neural networks, promoting predictive accuracy and situational awareness. CNN-LSTM and CNN-BiLSTM architectures combine convolutional networks to extract spatial features with recurrent networks to learn the long-time temporal covariance in water quality data. As an illustration, (Gohari Nezhad & Emami Niri, 2025) demonstrate that hybrid CNN-LSTM is better than standalone models because the former can learn the spatial variations in physicochemical parameters and the temporal sequence of behaviour together. These models are most effective in situations whereby the environmental variables vary at a slow pace or suddenly because of rain, seasonal, or an occurrence of contaminations. Their stratified structures enable them to generalize across various water systems whilst ensuring that they perform very well in predicting under different conditions.

Recently, sophisticated DL models like Gated Liquid Neural Networks (GLNNs) have been used on water quality prediction due to their capability to deal with irregular and noisy data and constantly changing data. GLNNs control the flow of information by adaptively controlling gating mechanisms that allow the model to give more emphasis to important temporal information and deemphasize irrelevant or redundant data. It is noted that such structures are particularly useful in capturing dynamic interactions of parameters including dissolved oxygen, pH, conductivity and turbidity (Singh et al., 2025). Their natural flexibility suits them better to the environmental data that do not have solid patterns and which contain numerous anomalies. In general, the DL approaches provide an effective way to model environmental complexity, which would guarantee reliability and scalability in the long-term water quality prediction.

2.3 Emerging Trends: Hybrid, Explainable, and Attention-Driven Models

Hybrid modelling is a relatively new phenomenon that is propelled by the growing necessity to be able to embody multiscale spatial-temporal interactions in environmental data. The models are the result of integration of the advantages of convolutional, recurrent and statistical decomposition methods to generate robust and noise-resistant predictions. Hybrid CNN-LSTM models, e.g., are commonly known to be able to jointly learn localized correlations between chemical parameters and long-term behavioural patterns that are caused

by seasonal or climatic conditions. The study by (Gohari Nezhad & Emami Niri, 2025) and (Chen et al., 2025) has shown that these hybrid systems are more effective than standalone models in the analysis of complex conditions in the environment. Prediction accuracy can be further enhanced with decomposition-based hybrids, including Ensemble Empirical Mode Decomposition (EEMD) with deep networks that isolate noise factors and show patterns underlying the dynamics of water temperature and pollutants. These developments point to the direction of the field towards multi-layered learning systems that can interpret different environmental cues.

Neural networks that can assign importance to particular input features or time steps dynamically are known as attention-based neural networks and have been demonstrated to have very flexible mechanisms. Attention mechanisms are also important, unlike the traditional DL models where all inputs are treated equally, the most impactful variables are focused on when making a prediction. Research such as (Dasgupta et al., 2025) and (Singh et al., 2025) indicate that the attention layers could be useful in accentuating the prevailing physicochemical variables in the critical events, such as the rainfall-driven runoff, seasonal shifts, or immediate spikes of contamination. This enables the models to respond to changes in the environment and becomes more effective in making predictions in fast-changing situations. Attention systems are also useful in increasing robustness by deemphasizing irrelevant or noisy inputs a valuable quality when dealing with water quality data, which frequently has missing or varying sensor measurements.

Recent survey studies indicate how AI aids aquaculture activities like water-quality monitoring, biomass estimation, and disease behaviour analysis and state that no single generalized model exists across heterogeneous ponds and recommends federated learning with UAV-assisted data collection (Karim et al., 2025). AquaSurveil presents a multi-agent water-health surveillance system that combines robots, IoT nodes, and privacy-preserving learning and shows good performance in anomaly-detection and monitoring on a large-scale multi-country surface-water dataset (Alevizos et al., 2025).

2.4 Research Gaps

Although machine learning and deep learning methods have made a lot of progress in predicting water quality, there are still a number of gaps and limitations that have not been addressed. The existing research is predominantly based on the black-box predictive models, including Random Forests, XGBoost, LSTMs, or hybrid CNN-LSTM networks, and commonly lacks the transparency of the prediction generation process (Chen et al., 2025; Rahu et al., 2023). Although these models are highly accurate, they do not usually give interpretable information on how each of the physicochemical parameters affects the ultimate predictions, and therefore environmental experts cannot easily trust and prove their results. Moreover, the optimization of predictive performance is studied by a multitude of other researchers, who do not consider such problems as feature sensitivity, biasing on the noise or missing sensor data, and instability to a sudden environmental change (Jibrin et al., 2024; Singh et al., 2025). Little has also been done to the universality of models between water

bodies, seasonal cycles, and geographic regions, so more adaptive, informative and powerful prediction systems are required.

Also, explainable ML models with the assistance of LIME can mitigate the generalization and adaptability challenges that are present in previous research. LIME allows the model developer to optimise feature engineering, preprocess noisy sensor data in IoT sensors, and create more robust hybrid models that react to environmental anomalies by pointing out changes in model behaviour across varying temporal patterns, seasons or sensor conditions (Dasgupta et al., 2025; Gohari Nezhad & Emami Niri, 2025). The incorporation of LIME into the IoT-based monitoring pipelines would make the predictions regarding the water quality not only accurate but also credible and open-minded, giving the policymakers and hydrological researchers something specific to act upon. This forms the basis of creating next-generation smart water monitoring systems in which predictive functionality is enhanced by interpretability, strength, and explainability by science.

Table 1. Related Work Summary: Methods, Datasets, Results, and Limitations

Author(s) & Year	Methodology / Approach	Main Model(s) Used	Dataset & Availability	Results / Performance	Limitations / Gaps
Shams et al. (2024)	Grid-search optimization for Water Quality Index (WQI) and Classification (WQC).	Gradient Boosting (GB), MLP Regressor	7 features, 1991 samples. Publicly available	GB Accuracy = 99.5%, MLP R^2 = 0.998.	Small feature set, no temporal analysis, static data.
Gohari Nezhad & Emami Niri (2025)	Hybrid deep model for predicting water saturation profiles.	CNN–LSTM Hybrid	10,674 points from 4 wells (South Pars Gas Field).	R^2 = 0.968 (10-fold validation).	Domain-specific (petroleum), complex architecture.
Jibrin et al. (2024)	Predictive hybrid approach for WQI & GWQI in arid regions.	GPR, ANFIS	60 groundwater samples (Eastern Province, KSA).	GPR RMSE = 0.0169, ANFIS RMSE = 0.0401.	Small regional dataset, no real-time or deep learning.
Shyu et al. (2023)	Soft-sensor model for real-time onsite wastewater monitoring.	SVR, Cubist Regression (CUB)	56 records (OWTS dataset, South Africa).	COD R^2 = 0.96 (SVR), TSS R^2 = 0.99 (CUB).	Very small dataset, E. coli prediction weak.
Rahu et al. (2023)	IoT + ML framework for WQI/WQC prediction.	MLP, Random Forest	Rohri Canal (Pakistan). 600–6000 records.	MLP R^2 = 0.93, RF Accuracy = 91%.	Needs real-time validation, limited sensor scope.
Arumugam et al. (2023)	Multi-layer deep neural network for potable vs non-potable classification.	Deep Neural Network (DNN)	UCI Water Potability Dataset. Publicly available	Accuracy=98%.	Limited explainability, no multi-feature fusion.
Lavanya & Sivakumar (2024)	Supervised learning approach for predicting water potability using regression-based ML	Decision Tree (DT), Support Vector Machine (SVM)	Water Quality Dataset – 2,293 samples, 9 physicochemical parameters. Publicly	DT Accuracy = 62.21%, SVM = 59.84%.	Small dataset, low accuracy, static data, no temporal or spatial variation.

	algorithms with correlation and preprocessing analysis.		available.		
Patel (2023)	Dual-stage ML framework for water potability prediction and estimation of atmospheric water generation using AQUOMIST dataset.	XGBoost, Support Vector Machine (SVM)	Water Quality Dataset (3,300 samples) and AQUOMIST Dataset (23,780 records from NASA Power Data Access Viewer, 29 Indian cities). Publicly sourced.	SVM Accuracy = 68.85%, XGBoost Accuracy = 98.6%.	High computational cost, climatic dependency, lacks physical or attenuation-based modeling.

3 Methodology

3.1 KDD (Knowledge Discovery in Databases) Methodology

In this work, the KDD approach is used to develop an explainable water quality prediction workflow, which is aimed at extracting insights instead of creating a real-time system. The selection of data entails the determination of key physicochemical parameters (e.g., ammonia, BOD, dissolved oxygen, orthophosphate, pH, temperature, nitrogen, nitrate) and the desired outputs (CCME_Values to be regressed and CCME_WQI classes to be classified). Preprocessing/cleaning is followed by an enhancement in data reliability to deal with missing values, correct inconsistency, noise reduction and treatment of outliers. Then, the data are transformed into a model input by standardizing/scaling the data, classifying the data as labels, and eliminating features based on ANOVA (SelectKBest); in case of imbalance, the training data are smoothed with SMOTE to facilitate balanced training.

During the data mining phase, the framework of the Random Forest is trained to acquire nonlinear relations between the variables of water quality and provide the predictions of the values of the CCME as well as the categories of WQI. To determine generalization, the evaluation phase measures performance based on regression (R2, MSE/MAE) and classification (accuracy, precision, recall, F1-score) on test data. To be interpretable, LIME gives local explanations giving which features most affected individual predictions. Lastly, the knowledge that has been extracted is provided in the form of brief descriptions of key drivers and output of explanations, emphasizing on transparency and analysis worth, but not deployment.

3.2 Dataset

Table 2. Dataset Overview and Key Attributes

Attribute	Summary
Records (rows)	**2,827,977**
Countries	**5** (Canada, China, England, Ireland, USA)
Waterbody Types	**12** unique types (e.g., River, Lake, Canal, Sea Water, Coastal, Estuarine, Bay, etc.)

Date Range	**1940-01-02 → 2023-12-10**
Measured Parameters (features)	Ammonia, BOD, Dissolved Oxygen, Orthophosphate, pH, Temperature, Nitrogen, Nitrate
Water Quality Index (numeric)	CCME_Values (range: **0.0 → 100.0**)
Water Quality Class (label)	CCME_WQI (Excellent / Good / Fair / Marginal / Poor)

In this dataset, there are 2,827,977 records of water-quality measurements of 5 different countries (Canada, China, England, Ireland, USA) in 12 different waterbody types (including rivers, lakes, canals, coastal/estuarine waters) during 1940-01-02 to 2023-12-10. Every record has the measured water parameters Ammonia, BOD, Dissolved Oxygen, Orthophosphate, pH, Temperature, Nitrogen and Nitrate which collectively illustrate the chemical and physical state of the water. The dataset also includes the CCME Water Quality Index as a numerical value (CCME_Values, 0-100) as well as a corresponding category (CCME_WQI) which will categorize the overall water quality as Excellent, Good, Fair, Marginal or Poor.

3.3 Contributions

The given work adds to the list of already existing end-to-end, explainable water-quality prediction workflows based on the KDD methodology, tailored to large-scale multi-country physicochemical data. It builds a full preprocessing and transformation chain, which encompasses outlier treatment, label encoding, Minmax normalization, an ANOVA-based feature selection (SelectKBest) and SMOTE-based class balancing, to guarantee that the training data is accurate and equitable across the water-quality groups. In addition to this, the work undertakes extensive benchmarking by assessing various machine learning and deep learning models in both regression and classification metrics to offer a robust comparative insight into the behaviour and performance of various models.

Besides that, the study enhances practical utility by integrating high predictive performance with transparency with LIME-based explainability, that gives local, human-readable reasons behind predictions of the Random Forests. The method measures the impact of features and interprets feature-value intervals so that domain experts can understand which physicochemical situations drive predictions to certain CCME_WQI classes. Lastly, the framework is tested on combined and country-specific data of five countries (Canada, China, England, Ireland and the USA) and shows consistent trends in performance and allows generalization to a wide range of regional water conditions.

3.4 Data Preprocessing

The preprocessing is initiated by the process of reading the dataset with the help of a standard data-loading process. The initial few records are loaded and then the data are checked to see that it has been loaded correctly and to get an initial feeling of the structure. This check can be used to ensure that all the variables expected by the dataset are captured and the format is appropriate to analyse. Many of the column names within the data already have long descriptive words or units in brackets. Such names are reduced to smaller and easy to

understand names like changing the name ammonia into the ammonia, biochemical oxygen demand, dissolved oxygen, orthophosphate, pH, temperature, nitrogen, and nitrate columns. This renaming is done to generate clean and uniform names of variables that are easier to read and are easier to analyse later without losing the semantic meaning of each parameter.

In order to comprehend the behaviour of the data at various locations, a separation of the dataset into separate country-level subsets is carried out. Each subset is then calculated to provide summary statistics of the water-quality parameters, analysis of distributions, central tendency, and spread of the parameters within each country. This analysis gives an understanding of the difference in measurements between geographical locations and allows determining possible anomalies or excessive values in some areas. Outlier analysis of each parameter of water-quality is done by the quartile and interquartile range. The first, second and third quartile are calculated and then interquartile range is calculated. Through this range, inner and outer fences are computed to give boundaries exceeding which data points can be regarded as potentially or possibly anomalous. The percentage of values that fall outside these calculated limits is also printed and this will aid in knowing the extent of outliers in each parameter and also aid in decision making regarding cleaning the data. Once the ranges of acceptable values are determined, each numerical column is cleaned of observations which are out of the range of acceptable interquartile based values. Individual upper and lower thresholds are calculated on ammonia, biochemical oxygen demand, dissolved oxygen, orthophosphate, pH, temperature, nitrogen and nitrate. Any dataset that fails to fall within such thresholds is dropped. This is done to make sure that the statistical associations are not skewed by extreme or erroneous values and to make sure that the machine-learning models are not trained in a way that Favors these values. Some columns like the name of the country, geographic region, type of water body, and date are eliminated in the dataset. Such variables can either cause redundancy, possess too many levels of categorical variables, or have no direct relationship with the target of the prediction. Dropping them will reduce the dataset to be only meaningful numeric predictors and avoids unnecessary noise influencing the performance of a model.

The data is cleaned, divided into two parts, predictor features, and target variable which is the values of the CCME index. The preprocessing will prepare the data to be scaled, feature selected and ultimately trained to create a regression model by isolating the target variable. In the process of classification, the target variable should be coded in terms of numerical class codes and thus, such categorical variables like CCME WQI will be coded to machine readable numerical values. The categorical columns will automatically be identified and each of them that incorporates the label of CCME WQI is then coded into integers through label encoding. This is to guarantee that the classification algorithms are able to properly describe the target classes as discrete categories and not as text-based classes.

All numeric predictors are normalized under a scaling method that brings all the predictors to a normalized range between zero and one. Such normalization is necessary to make the variables having different scales contribute equally to the machine-learning model. It also assists the distance-based algorithms and gradient-based optimizers to work better by

balancing the feature magnitude. The normalized data is separated into training and testing partitions. Part of the data is also set aside in final evaluation, and the rest is model training. This divide makes sure that the trained models are applied to unseen data and this gives a fair representation of the predictive power of the models.

ANOVA is a statistical method that can be used to identify the most strongly related predictors of changes in the target variable. In this technique, the contribution of each of the features to differences in the class labels is estimated and the features are ranked by their statistical significance. The top five features determined by ANOVA are chosen as the reduced feature set which enhances the efficiency of the model, reduces noise and presents the most relevant variables that can be used to make a precise prediction.

4 Design Specification

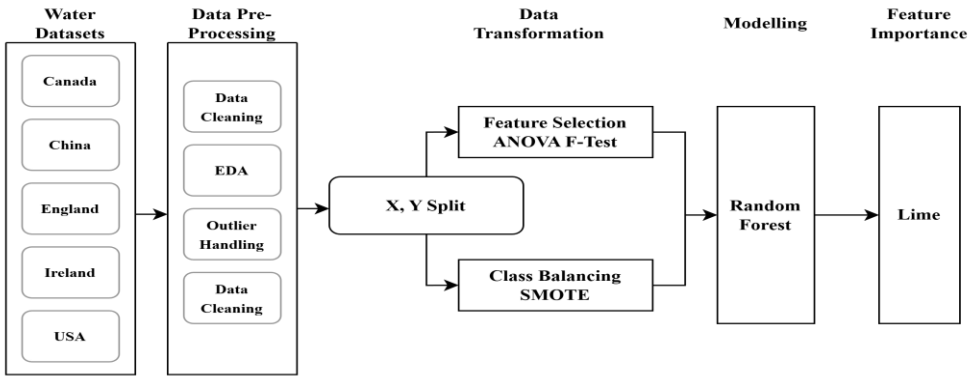

Figure 1. Overall Workflow of the Water Quality Prediction Framework

System Design Overview

In figure 3, illustrates the end-to-end pipeline adhered to in the process of water quality classification and explanation on datasets of USA, Ireland, England, Canada and China. Raw physicochemical records are initially acquired and run through data pre-processing, during which data cleaning is applied to eliminate missing/noisy sensor values, EDA is applied to conceive distributions and relationship among parameters, and outlier handling is carried out to limit the impact of abnormal readings that may bias learning. Once the data is prepared, it is arranged into X (features of input) and Y (target water-quality class/label) and further divided into training and testing sets to assure that it is not biased in any way. In the transformation stage, ANOVA-based feature selection (SelectKBest) is applied to retain the most statistically significant predictors, reducing dimensionality and keeping only the features that best discriminate between water-quality categories. Since water-quality classes are often imbalanced, SMOTE is applied to the training data to synthetically increase minority-class samples and improve fairness in learning. The selected and balanced dataset is then used to train a Random Forest model, chosen for its robustness and ability to capture nonlinear relationships in environmental variables. Finally, LIME is applied as the interpretability module: it explains individual Random Forest predictions by showing which

features (and value ranges/intervals) push the decision toward a particular class, thereby providing transparent, human-understandable reasoning behind the model outputs.

5 Implementation

The models in this study are implemented using a publicly available surface-water dataset rather than data collected through a deployed system. The data sample is borrowed by (Karim et al., 2025). To preprocess and enhance learning, ANOVA (SelectKBest) is used to choose the most significant physicochemical features and SMOTE is used on the training set to deal with the issue of class imbalance. Random Forest model is then interpreted with the help of LIME that explains predictions by giving the most influential features and feature-interval contributions.

5.1 Machine Learning Models Implementation

Application of machine learning models in the given study will start with the Decision Tree Regressor which is trained using the training dataset that will be prepared to learn the nonlinear relationship between the physicochemical features and the values of CCME water quality. The model is trained and then predicts the test set and evaluation metrics are calculated to measure the predictive performance of the model. A scatter diagram that compares the values of the actual and predicted values is drawn to visually compare the model accuracy with the ideal prediction line. This assists in knowing the extent of the Decision Tree in capturing underlying water quality patterns and in pointing out any deviations or inconsistencies in prediction behaviour.

The next implemented XGBoost Regressor is an upgraded and optimized boosting-based model with a high generalization power and a high capacity to work with tabular data. Once the model has been fitted with the training dataset, the model will then make predictions on the test set, and then performance measures are determined to test accuracy. The visual comparison of the actual and predicted values is plotted, as in the case of the Decision Tree output, which allows assessing the predictive strength of XGBoost visually. The gradient-boosting model can be more reliable and accurate and, therefore, a useful part of the machine learning phase of the entire water quality prediction system.

5.2 Deep Learning Models Implementation

The deep learning models used in the current research start with the Artificial Neural Network (ANN), which is built on a Sequential architecture comprising of two hidden layers each with 64 neurons and is activated by the ReLU function. This model takes the complete list of pre-processed physicochemical parameters as an input, and these are ammonia, biochemical oxygen demand, dissolved oxygen, orthophosphate, pH, temperature, nitrogen, and nitrate. The input layer transmits these variables to the network where the ANN can learn nonlinear relationships between these features. The model is trained with Adam optimizer and mean squared error as the loss function and is trained on 30 epochs. The ANN predicts

the values of CCME targets of the test set after training and the performance of the ANN is determined. This form of deep learning is useful in capturing complicated interactions in the dataset, which enhances predictive performance over traditional machine learning models.

The Long Short-Term Memory (LSTM) model is applied to use the temporal dependencies that exist in the physicochemical parameters and CCME water quality patterns. The dataset is reshaped to satisfy the needs of LSTM before training, and each of the parameters, including ammonia, BOD, DO, orthophosphate, pH, temperature, nitrogen, and nitrate, is considered a sequential input feature to the network. The LSTM model has 1 LSTM layer that consists of 50 units then the dense output neuron to predict continuous values of the CCME. The model is then trained in 30 epochs after compilation with Adam optimizer and mean squared error loss. The predictions made on the test set are assessed in terms of standard measures. The LSTM model uses physicochemical parameters sequentially to capture the underlying temporal trends and water quality variability, which make the model useful in understanding environmental variability.

5.3 Random Forest Model Implementation

Feature Selection Using ANOVA and SMOTE

Prior to the ANOVA application, eight physicochemical variables Ammonia, BOD, DO, Orthophosphate, pH, Temperature, Nitrogen and Nitrate were incorporated into the dataset as the possible predictors of the CCME WQI classification. To determine which of these features had the greatest significance, ANOVA F-test was used by applying the SelectKBest method which analyses the degree to which each feature is variable between the various water quality classes. In this approach, each parameter is assigned an F-value with the higher the value the stronger is the discriminatory power. The parameter k of SelectKBest is the top-ranked features to retain, by using five as the value of k, the top five features with the highest F-values were retained to use in the model training. The top five features identified after the calculation of the scores were Ammonia, BOD, Orthophosphate, Temperature, and Nitrate, which indicated that the variables mentioned above influence the differentiation between categories of CCME_WQI the most. After the feature selection was done, SMOTE (Synthetic Minority Oversampling Technique) was used on the training data only in order to correct the imbalance in classes by synthesizing samples that represented the underrepresented classes. Such a combination of ANOVA to determine the statistically significant predictors and SMOTE to balance the data set makes sure that the machine learning models are trained on meaningful and balanced data, which enhances predictive power and minimizes bias.

Random Forest Model Training and Performance Evaluation

Random Forest model was used as the primary classification tool in this research due to the high capacity in learning sophisticated patterns of environmental data and the accuracy in dealing with varied water quality characteristics. The model was set up with a n_estimator of 100 i.e. the model constructs 100 separate decision trees and uses the outputs to come up with a more consistent and reliable final forecast. The training was done using the feature-selected

and SMOTE-balanced dataset in such a way that all the CCMEWQI classes were well represented during the training. After the model had been trained, the model was applied to the test set to predict water quality classes, and the performance of the model was measured in terms of accuracy, macro precision, macro recall, and macro F1-score. Such measures give an objective assessment of all classes and show the level to which the Random Forest classifier can detect trends in physicochemical parameters and classify water quality levels in the developed prediction structure.

LIME-Based Feature Contribution Analysis

Once the Random Forest classifier was trained, LIME was used to produce local explanations and the features that were the most influential to the predictions made by the model. LIME tested the impact of variations in every feature on the probability of the predicted classes by sampling 50 random test instances, and taking notes of the strength of each feature. These personal contributions were then averaged to give an overall measure named Avg_Contribution and this measure represents the average absolute contribution a feature makes to the decision of the model over all the sampled explanations. Larger values of Avg_Contribution demonstrate that the feature has always been a more important influence on the classification output of the Random Forest. The findings indicated that the Orthophosphate possessed the largest Avg_Contribution value (0.064773), implying that it had the greatest and the most stable effect on the decision to make predictions. The temperature was also followed with a moderate contribution (0.016548), implying that it has a significant, but less dominant, contribution. The contribution values of BOD, Ammonia, and Nitrate were lower (0.005568, 0.004968, and 0.003965 respectively), which implies that they have a lesser impact on the model than the most important features. The above Avg_Contribution values have been used to measure the importance of each physicochemical parameter in the determination of water quality classes and this has been found to be clearly interpretable and has given more credibility to the decisions made by the model as indicated in figure 4.

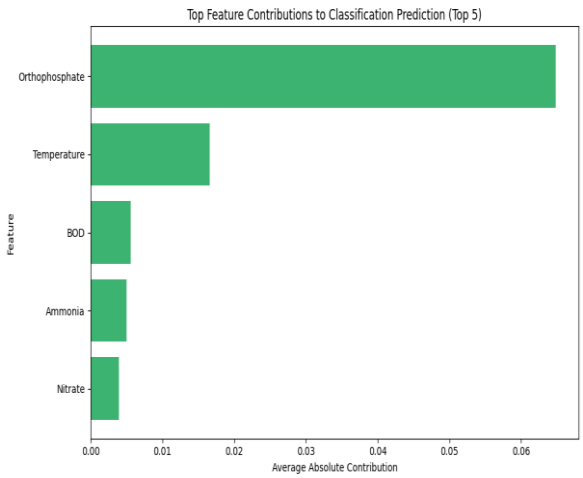

Figure 2. Top 5 Feature Contributions

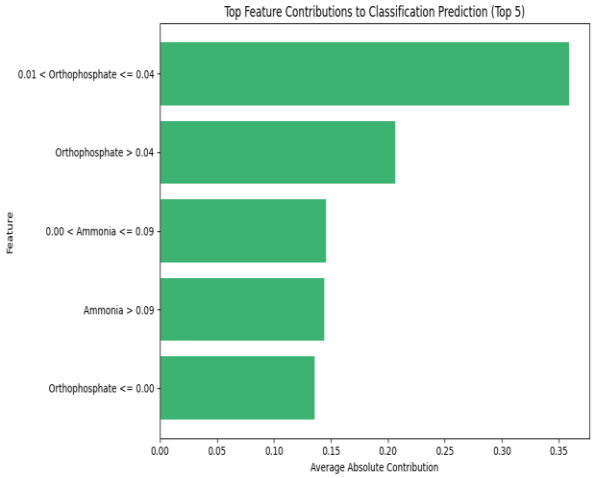

Figure 3. Top 5 Influential Feature Intervals

Interval-Level Interpretability from LIME

When LIME is used on the Random Forest classifier as in figure 5, it operates by repeatedly perturbing each of the test instances and measuring the change in the prediction response to the input features in this case, LIME automatically subdivides each continuous physicochemical parameter into intervals which make sense, such as 0.01 < Orthophosphate ≤ 0.04, Orthophosphate > 0.04, 0.00 < Ammonia ≤ 0.09, Ammonia > 0.09, Orthophosphate ≤ 0.00. and assigns a contribution value showing how strongly each interval pushes the prediction toward or away from a given class. By sampling 50 random test samples and averaging the absolute LIME contributions, the model reveals how consistently each feature interval influences classification across many predictions. The high Avg_Contribution of 0.01 < Orthophosphate ≤ 0.04 (0.359063) shows it is the most dominant influence on the classifier, meaning that measurements in this range strongly steer the Random Forest's decision. The next influential interval, Orthophosphate > 0.04 (0.206708), also shows that higher orthophosphate levels significantly shape predicted water-quality categories. Ammonia intervals including 0.00 < Ammonia ≤ 0.09 (0.145888) and Ammonia > 0.09 (0.144091) exert moderate influence, indicating that different ammonia concentrations play an important but secondary role in classification. Finally, Orthophosphate ≤ 0.00 0.135724 still contributes meaningfully, though with slightly lower strength. Together, these averaged contribution values summarize how strongly each feature interval repeatedly impacts prediction decisions, providing transparent, model-agnostic evidence of which physicochemical conditions drive the water-quality classifications in the Random Forest model. This is for Canada Dataset. For the remaining four countries, the LIME analysis showed similar patterns, confirming that the explanatory behaviour of the model remained consistent. Overall, the findings demonstrated that all datasets followed the same general interpretability trend as observed in the Canada dataset.

6 Evaluation Results

6.1 EDA (Exploratory Data Analysis)

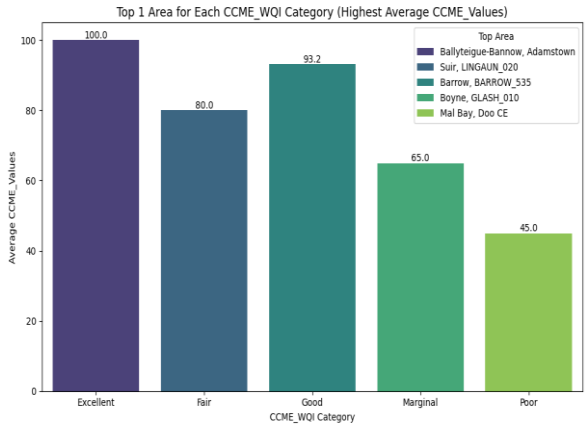

Figure 4. Top 1 Area for Each CCME_WQI

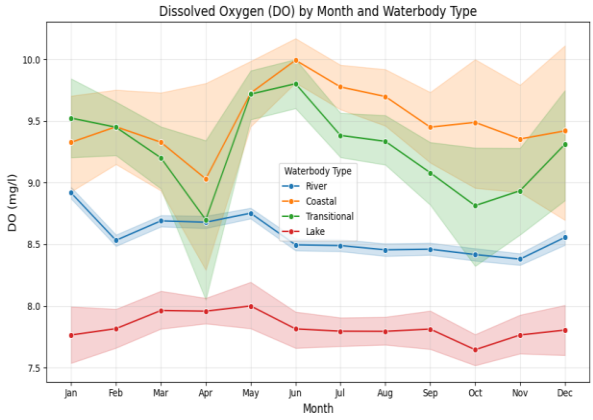

Figure 5. DO by Month and Waterbody Type

In figure 1, This bar chart displays the highest-scoring monitoring area within each CCME_WQI water quality category, based on average CCME_Values. The plot shows that the Excellent and Good categories have the highest scores overall, while Marginal and Poor categories correspond to areas with noticeably lower water quality. In figure 2, This line plot illustrates how dissolved oxygen levels vary across months for different waterbody types, highlighting that coastal and transitional waters generally maintain higher DO concentrations throughout the year compared to rivers and lakes. Seasonal trends are visible across all categories, with slight increases around mid-year and modest declines during late autumn.

6.2 Experiment 1

Table 3. Performance Comparison Before Removing Outliers 5 Countries Combine dataset

Models	Regression				Classification			
	MSE	RMSE	MAE	R2	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.031	0.177	0.0184	0.99	0.9507	0.9574	0.9507	0.9520
Random Forest	0.020	0.142	0.0128	0.99	0.9980	0.9980	0.9980	0.9980
XGBoost	0.218	0.467	0.151	0.99	0.9907	0.9907	0.9907	0.9907
ANN	0.322	0.567	0.248	0.99	0.9739	0.9744	0.9739	0.9738
LSTM	0.299	0.547	0.162	0.99	0.9748	0.9750	0.9748	0.9748

Regression Results

For the combined 5-country dataset before removing outliers, all models achieved a very high regression fit with $R^2 = 0.99$, indicating that each method explains nearly the same proportion of variance in the CCME continuous values. However, the models differ clearly in prediction error (MAE). Random Forest gives the best regression accuracy with the lowest MAE = 0.0128 ($R^2 = 0.99$), showing very close agreement between predicted and actual values. Decision Tree performs next with MAE = 0.0184 ($R^2 = 0.99$), but its higher MAE suggests fewer stable predictions compared to the ensemble approach. XGBoost shows a much larger error with MAE = 0.151 ($R^2 = 0.99$), meaning its predictions deviate more on average even though the overall fit remains high. In the case of deep learning, the MAE of the LSTM = 0.162 (R 2 = 0.99) and ANN = 0.248 (R 2 = 0.99) are the lowest error rates, which means that these neural networks fail to decrease the error in regression as effectively as the Random Forest and Decision Tree do on the combined data.

Classification Results

To compare the performance on the same combined dataset, the differences in performance are better observed with the help of Accuracy and F1-score which show correct prediction of classes and balanced performance in the classes. Random Forest provides the best scores Accuracy = 0.9980 and F1-score = 0.9980, which demonstrates almost perfect classification

and highly predictable results on all classes. XGBoost has also a high level of results Accuracy = 0.9907 and F1-score = 0.9907, which proves that boosting can be used to classify water-quality, but a little slower than Random Forest. Decision Tree results in relatively smaller figures with Accuracy = 0.9507 and F1-score = 0.9520, indicating that it is more sensitive to noise and imbalance and cannot work with complex boundaries. ANN has Accuracy = 0.9739 and F1-score = 0.9738 and LSTM has almost the same result with Accuracy = 0.9748 and F1-score = 0.9748. On the whole, the ensemble techniques, in particular, the Random Forest, provide the most trustworthy classification results with the combined 5-country dataset, prior to the removal of outliers.

6.3 Experiment 2

Table 4. Performance Comparison After Removing Outliers 5 Countries Combine dataset

Models	Regression				Classification			
	MSE	RMSE	MAE	R2	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.0001	0.012	0.0002	0.99	0.9997	0.9997	0.9997	0.9997
Random Forest	0.0001	0.012	0.0002	0.99	0.9999	0.9999	0.9999	0.9999
XGBoost	0.065	0.255	0.037	0.99	0.9958	0.9958	0.9958	0.9958
ANN	0.352	0.593	0.347	0.99	0.9959	0.9965	0.9959	0.9961
LSTM	0.040	0.200	0.027	0.99	0.9955	0.9961	0.9955	0.9957

Regression Results

Once outliers are eliminated in the combined 5-country data, the regression performance of the tree-based models improves significantly, and R 2 is consistently high at 0.99 with all the techniques. Decision Tree indicates that the error decreased significantly with the MAE = 0.0002 (R 2 = 0.99), which means that the outliers were eliminated, and the model created very close predictions of the actual values. Random Forest compares this regression error rate with MAE = 0.0002 (R 2 = 0.99), and it is highly stable and accurate even when cleaned. XGBoost does moderately well with MAE = 0.037 (R 2 = 0.99) which is better than previously but still higher than Decision Tree and Random Forest. Under deep learning, MAE is relatively low LSTM = 0.027 (R 2 = 0.99) and its improvement and precision are very good compared to XGBoost regression. Nevertheless, ANN has the greatest regression error with MAE = 0.347 (R 2 = 0.99) indicating that it is unable to reduce the error in prediction after the elimination of outliers as compared to the other models.

Classification Results

In order to classify after the removal of outliers, the overall performance is very high in all models, and it proves that data cleaning enhance the classification and consistency of models. The Decision Tree has an almost perfect accuracy of 0.9997 and an F1-score of 0.9997, which means that there is a high level of reliability in the classification of the classes.

Random Forest has the highest Accuracy = 0.9999 and F1-score = 0.9999, which proves that it is robust and has good generalization when the noisy extremes are removed. XGBoost also has a good entry score of Accuracy = 0.9958 and F1-score = 0.9958, which also exhibits consistent predictions but a little lower compared to the best performing tree ensembles. ANN achieves the highest Accuracy = 0.9959 and F1-score = 0.9961, which implies that the model has a noticeable increase in the consistency of classifications despite having a high regression MAE. LSTM achieves Accuracy = 0.9955 with F1-score = 0.9957, with equally high levels of classification performance, and shows that the outlier handling is beneficial with both traditional and neural methods of multi-country water quality classification.

6.4 Experiment 3

Table 5. Performance Comparison random forest model across individual country wise dataset

Countries	Random Forest				Random Forest			
	Regression				Classification			
	MSE	RMSE	MAE	R2	Accuracy	Precision	Recall	F1 Score
Canada	1.743	1.320	0.329	0.98	0.9666	0.9322	0.9274	0.9298
China	0.029	0.172	0.0107	0.99	0.9984	0.9984	0.9984	0.9984
England	0.025	0.1607	0.016	0.99	0.9974	0.9974	0.9974	0.9974
Ireland	0.023	0.152	0.008	0.99	0.9983	0.9984	0.9983	0.9983
USA	0.036	0.191	0.006	0.99	0.9994	0.9994	0.9994	0.9994

In Experiment 3, for the five-country datasets (Canada, China, England, Ireland, and USA).

Regression Results Random Forest Country-wise

The Random Forest regression score of each country separately and it can be seen that the model is overall strong but differs across the characteristics of the national dataset. The regression fit in Canada is the lowest out of the five with R 2 = 0.98 and a significantly larger error of MAE = 0.329, which represents the larger deviations between predicted and actual values, presumably because of greater variability or more difficult patterns in that data. Conversely, China has a very high regression performance with R 2 = 0.99 and low MAE = 0.0107 which depicts very accurate predictions. England also has R 2 = 0.99 and MAE = 0.016, which proves that the quality of prediction is stable. On the same note, Ireland has a very good R2 = 0.99 and MAE = 0.008 which means that it has low average error. The USA dataset yields the lowest (smallest) regression error of MAE = 0.006 and R 2 = 0.99 indicating that Random Forest generalizes quite well on the water-quality trends in that country.

Classification Results Random Forest Country-wise

To classify, the performance of the Random Forest is found to be high in all countries with a close accuracy and F1 scores. Canada demonstrates a little less strength of classification than

other datasets, with Accuracy = 0.9666 and F1-score = 0.9298 which is good but lower balance between classes, which may be explained by class overlap or imbalance. China has almost perfect classification with Accuracy = 0.9984 and F1-score = 0.9984, which means that the prediction of the classes is very reliable. England comes next with Accuracy = 0.9974 and F1-score = 0.9974 with consistent and unchanging classification scores. The country of Ireland also has a good Accuracy = 0.9983 and F1-score = 0.9983 indicating that the classes are well separated. The USA dataset provides the best classification results with Accuracy = 0.9994 and F1-score = 0.9994, and only Canada has rather low but still good results.

6.5 Experiment 4

Feature Selection

Table 6. Performance Comparison random forest model across individual country wise dataset

Countries	Random Forest				Random Forest			
	After Feature Selection				Before Feature Selection			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Canada	0.4380	0.5713	0.6528	0.52	0.9666	0.9322	0.9274	0.92
China	0.9803	0.6820	0.6377	0.65	0.9984	0.9984	0.9984	0.99
England	0.7550	0.4464	0.5609	0.49	0.9974	0.9974	0.9974	0.99
Ireland	0.9059	0.3588	0.2901	0.30	0.9983	0.9984	0.9983	0.99
USA	0.9557	0.4906	0.5045	0.48	0.9994	0.9994	0.9994	0.99

In this Experiment 4, After the train and test split, we tested three various feature-selection methods in order to determine which physicochemical parameters would give the most information about predicting the classes of CCME_WQI. We have initially tested ANOVA (F-test) feature selection, that is, the features are ranked in order of their ability to show differences in their values between the water-quality classes (Excellent, Good, Fair, Marginal, Poor). That is, it puts a strong emphasis on parameters that display apparent statistical distinction among classes. Then we used Mutual Information (MI) that quantifies the amount of information each feature has about the class label, even nonlinear interactions that are not necessarily well measured by purely statistical tests of means difference. Finally, we experimented with Recursive Feature Elimination (RFE), a model-driven approach that repeatedly trains a classifier and progressively removes the least useful features until only the best subset remains. Using these three methods helped us cross-check which parameters consistently appear important and ensured that our feature choice was not dependent on a single technique.

Although we tested all three, we ultimately selected ANOVA as the final feature-selection method for this dataset because it provided the most reliable and stable top-feature subset for our multi-class water quality classification task, while also keeping the preprocessing pipeline

efficient and easy to justify scientifically. ANOVA directly supports our objective of improving class separability before training the Random Forest classifier, meaning the retained features are those most statistically significant for distinguishing between water-quality categories. This also strengthens interpretability: when the Random Forest predictions are later explained using LIME, the explanations focus on a smaller set of highly relevant parameters, making the decision reasoning clearer for environmental experts and decision-makers.

6.6 Experiment 5

Class Balancing

Table 7. Performance Comparison random forest model across individual country wise dataset

Countries	Random Forest				Random Forest			
	Before Class Balancing				After Class Balancing			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Canada	0.4380	0.5713	0.6528	0.52	0.9666	0.9322	0.9274	0.92
China	0.9803	0.6820	0.6377	0.65	0.9984	0.9984	0.9984	0.99
England	0.7550	0.4464	0.5609	0.49	0.9974	0.9974	0.9974	0.99
Ireland	0.9059	0.3588	0.2901	0.30	0.9983	0.9984	0.9983	0.99
USA	0.9557	0.4906	0.5045	0.48	0.9994	0.9994	0.9994	0.99

In Experiment 5, the focus is on class balancing because the CCME_WQI categories are not equally represented in the dataset. When class imbalance exists, a model can become biased toward the majority classes and may fail to recognize minority classes reliably, even if overall accuracy looks high. To handle this issue, we considered three balancing strategies as part of the experimental setup: SMOTE oversampling, Random Under sampling (RUS), and a Hybrid method (SMOTE followed by RUS). These variants are three approaches to addressing imbalance, which are creating more minority samples (SMOTE), fewer majority samples (RUS), or a combination of both methods (Hybrid) to achieve a more balanced distribution.

Even though the three strategies were incorporated as balancing options, in our final experiments we only used SMOTE. The reason behind the choice of SMOTE is that it enhances the fairness of learning by adding minority-class representation without giving up the majority-class information to allow the model to learn information about decision boundaries on all the categories of CCME_WQI more efficiently. Once we have used SMOTE, we made sure that the effects of the SMOTE are reflected by verifying the new class distribution and verifying that the formerly underrepresented classes have been upgraded and the total class distribution is more balanced. This helps to achieve more

accurate classification performance in all water-quality classes, with particular advantage to critical lower-quality classes like Marginal and Poor.

7 Discussion

To improve the classification of CCME_WQI water quality categories, this investigation combined the techniques of SMOTE, ANOVA F-test feature selection and LIME. The evidence of this dataset's class imbalance led to a bias in the Random Forest model, since it identified strong performance for the larger classes, but missed many of the smaller classes. However, following the application of SMOTE, the model showed increased recall and F1-score performance, which demonstrated that the model had improved performance in classifying the water quality categories. The ANOVA F-test feature selection effectively narrowed down the number of physicochemical parameters used for the classification to only the most important parameters statistically (DO, pH and Nitrate), which assisted in reducing the dimensionality of the feature space and increasing the overall efficiency of the training process with maintained or slight improvement in predictive accuracy, thus illustrating that not all variables are equally valuable for the classification.

To better understand model interpretability, individual predictions of LIME were generated so users had a local explanation for what contributed to that particular outcome. Random Forest is widely seen as a very accurate prediction method. However, since Random Forest works as a black box, it is nearly impossible to know exactly how specific environmental conditions influenced classification results when using Random Forest. To address this issue, LIME highlights which features are key contributors to each classification prediction, such as the effect of low dissolved oxygen (DO) and high nutrient concentrations on the overall water quality classification level. As such, LIME's explanations provide increased transparency and trustworthiness of machine learning models to environmental professionals and practitioners. Overall, the combination of SMOTE, feature selection via ANOVA, and LIME representation created a water quality classification system that has balanced, efficient, and interpretable results, and will allow true environmental decision-making in the real world.

8 Conclusion and Future Work

8.1 Conclusion

This study proved that accurate and interpretable water-quality forecasting could be done with the help of data-driven modelling by means of large-scale physicochemical records of five countries. It was executed in its entirety, including outlier processing, Minmax normalization, feature selection using ANOVA, and class balancing using SMOTE to enhance the quality of the data and the fairness of learning across water-quality categories. The various machine learning and deep learning models were tested in terms of regression and classification measures, in which ensemble methods in particular Random Forest performed best on the merged dataset and per-country. Besides predictive accuracy the study enhanced transparency through the application of LIME to the Random Forest classifier, which gave transparent understandings as to which features and feature-value ranges made the most significant contribution to the classification decision. In general, the findings

indicate that the use of predictive models with high predictive power in combination with statistically guided preprocessing and explainability methods can be used to provide reliable decision information to water-quality assessment using various geographic samples.

8.2 Future Work

This study can be furthered in the future by enhancing the temporal and spatial models and enhancing model robustness under varying environmental conditions. Because the dataset is multi-decade and a variety of waterbodies, future experiments can make use of explicit time-based structure (e.g. seasonal or yearly windows), sequence-aware models that utilize true time-series ordering instead of treating features as independent inputs. Other explainability techniques like SHAP can be added to the interpretation of LIME to compare the global and local interpretations as well as to ensure consistency of feature importance across nations. Systematic hyperparameter tuning, cross-country transfer testing, and more expressive feature engineering (e.g. interaction features or derived water-quality indicators) can also be used to improve the modeling stage. Lastly, the method can be scaled to an actual real-time monitoring environment by incorporating the streaming sensor data, automated drift detection and regular model updating to ensure the performance is retained in the ever-changing environmental patterns.

References

Alevizos, V., Yue, Z., Edralin, S., Xu, C., Gerolimos, N., & Papakostas, G. A. (2025). Multi-Agentic Water Health Surveillance. *Water*, *17*(17), 2653. https://doi.org/10.3390/w17172653

Balo Utku, E. D., Utku, A., & Kutlu, B. (2025). Enhancing water quality prediction with artificial intelligence: A hybrid convlstm model for spatio-temporal analysis. *International Advanced Researches and Engineering Journal*, *9*(2), 107–117. https://doi.org/10.35860/iarej.1679575

Biazar, S. M., Golmohammadi, G., Nedhunuri, R. R., Shaghaghi, S., & Mohammadi, K. (2025). Artificial Intelligence in Hydrology: Advancements in Soil, Water Resource Management, and Sustainable Development. *Sustainability*, *17*(5), 2250. https://doi.org/10.3390/su17052250

Chadalavada, S., Yaman, S., Sengur, A., Hafeez-Baig, A., Tan, R.-S., Datta Barua, P., Deo, R. C., Kobayashi, M., & Rajendra Acharya, U. (2025). Gated-LNN: Gated Liquid Neural Networks for Accurate Water Quality Index Prediction and Classification. *IEEE Access*, *13*, 69500–69512. https://doi.org/10.1109/ACCESS.2025.3561593

Chen, R., Wang, Q., & Javanmardi, A. (2025). A Review of the Application of Machine Learning for Pipeline Integrity Predictive Analysis in Water Distribution Networks. *Archives of Computational Methods in Engineering*, *32*(6), 3821–3849. https://doi.org/10.1007/s11831-025-10251-6

Dasgupta, R., Das, S., Banerjee, G., & Mazumdar, A. (2025). Perspectives on water quality analysis emphasizing indexing, modeling, and application of artificial intelligence for comparison and trend forecasting. *River*, *4*(2), 265–286. https://doi.org/10.1002/rvr2.70012

Gheisari, M., Shafi, J., Kosari, S., Amanabadi, S., Mehdizadeh, S., Fernandez Campusano, C., & Barzan Abdalla, H. (2025). Development of improved deep learning models for multi-step ahead forecasting of daily river water temperature. *Engineering Applications of Computational Fluid Mechanics*, *19*(1), 2450477. https://doi.org/10.1080/19942060.2025.2450477

Gohari Nezhad, A., & Emami Niri, M. (2025). Enhancing water saturation predictions from conventional well logs in a carbonate gas reservoir with a hybrid CNN-LSTM model. *Journal of Petroleum Exploration and Production Technology*, *15*(5), 89. https://doi.org/10.1007/s13202-025-01995-9

Jibrin, A. M., Al-Suwaiyan, M., Aldrees, A., Dan'azumi, S., Usman, J., Abba, S. I., Yassin, M. A., Scholz, M., & Sammen, S. Sh. (2024). Machine learning predictive insight of water pollution and groundwater quality in the Eastern Province of Saudi Arabia. *Scientific Reports*, *14*(1), 20031. https://doi.org/10.1038/s41598-024-70610-4

Karim, Md. R., Syeed, M. M. M., Rahman, A., Ayaz Rabbani, K., Fatema, K., Khan, R. H., Hossain, M. S., & Uddin, M. F. (2025). A Comprehensive Dataset of Surface Water Quality Spanning 1940-2023 for Empirical and ML Adopted Research. *Scientific Data*, *12*(1), 391. https://doi.org/10.1038/s41597-025-04715-4

Lahari, S. A., Kumawat, N., Amreen, K., Ponnalagu, R. N., & Goel, S. (2025). IoT integrated and deep learning assisted electrochemical sensor for multiplexed heavy metal sensing in water samples. *Npj Clean Water*, *8*(1), 10. https://doi.org/10.1038/s41545-025-00441-x

Lavanya, J., & Sivakumar, N. (2024). Predicting Water Quality using Machine Learning Techniques. *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, 554–560. https://doi.org/10.1109/ICIPCN63822.2024.00097

Paneru, B., Paneru, B., & Sapkota, S. C. (n.d.). *Enhancing Water Sustainability with AI methods: Analysis and Prediction of Seasonal Water Quality of Nepal Using Machine Learning Approach with SHAP Analysis*.

Pariyapurath, N. K., Tarunprasad, R., Maghimaa, M., Jagannathan, S., & Muthuvel, R. (2025). Integrating Artificial Intelligence into Modern Water Management Frameworks: Review. *International Journal of*

Advanced Science and Engineering, *11*(4), 4763–4774. https://doi.org/10.29294/IJASE.11.4.2025.4763-4774

Patel, S., Shah, K., Vaghela, S., Aglodiya, M., & Bhattad, R. (2023). *Water Potability Prediction Using Machine Learning*. In Review. https://doi.org/10.21203/rs.3.rs-2965961/v1

Rahu, M. A., Chandio, A. F., Aurangzeb, K., Karim, S., Alhussein, M., & Anwar, M. S. (2023). Toward Design of Internet of Things and Machine Learning-Enabled Frameworks for Analysis and Prediction of Water Quality. *IEEE Access*, *11*, 101055–101086. https://doi.org/10.1109/ACCESS.2023.3315649

Shams, M. Y., Elshewey, A. M., El-kenawy, E.-S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2023). Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*, *83*(12), 35307–35334. https://doi.org/10.1007/s11042-023-16737-4

Shyu, H.-Y., Castro, C. J., Bair, R. A., Lu, Q., & Yeh, D. H. (2023). Development of a Soft Sensor Using Machine Learning Algorithms for Predicting the Water Quality of an Onsite Wastewater Treatment System. *ACS Environmental Au*, *3*(5), 308–318. https://doi.org/10.1021/acsenvironau.2c00072

Singh, A., Bista, P., Deb, S. K., & Ghimire, R. (2025). Simulating cover crops impacts on soil water and nitrogen dynamics and silage yield in the semi-arid Southwestern United States. *Agricultural Water Management*, *307*, 109246. https://doi.org/10.1016/j.agwat.2024.109246

Song, X., Yan, H., Shen, Q., Sun, W., Li, P., & Wei, W. (2025). Treatment of water discharged from a Yellow River water purification plant: Optimization and application of enhanced coagulation technology. *RSC Advances*, *15*(22), 17476–17490. https://doi.org/10.1039/D5RA01783A