Evaluating the Influence of Climate on Water Quality Using Machine Learning Models

MSc Research Project
Masters in Data analytics

Saikiran Goud Doddivenuka
Student ID: x23238224

School of Computing
National College of Ireland

Supervisor: Rejwanul Haque

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sai Kiran Goud Doddivenuka
Student ID:	x23238224
Programme:	Masters in Data Analytics **Year:** 2025
Module:	Research Practicum (MSCDAD_A)
Supervisor:	Rejwanul Haque
Submission Due Date:	14-09-2025

WordCount: 4252 **Page Count:** 15

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: Sai Kiran Goud Doddivenuka

Date: 14-09-2025

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):	

Evaluating the Influence of Climate on Water Quality Using Machine Learning Models

Sai Kiran Goud Doddivenuka
x23238224

Abstract

This research presents an enhanced machine learning and deep learning framework for predicting Dissolved Oxygen (DO) levels in surface water, incorporating both climate and temporal features. While traditional DO models often rely solely on water chemistry, this study integrates dynamic variables such as air temperature, seasonal cycles, and derived indices like thermal stratification and temperature-salinity interactions.

The project explores multiple supervised learning techniques including Random Forest, XGBoost, Logistic Regression, and a Multi-Layer Perceptron (MLP) neural network, benchmarking them across regression and classification tasks. A new risk-based DO classification system (Safe, Moderate, Critical) was also introduced and balanced using the Synthetic Minority Oversampling Technique (SMOTE) for fair model training. SHAP (SHapley Additive exPlanations) explainability methods highlighted that features like DO lag, water temperature, and climate interactions significantly influence DO predictions. The Extra Trees classifier achieved the best classification accuracy (84.6%) in binary mode, while the MLP achieved 77% accuracy in the multi-class DO risk task after SMOTE.

This work contributes a comprehensive, interpretable, and scalable pipeline that supports real-time DO monitoring under the lens of climate change. It strengthens the foundation for smart environmental systems and sustainable decision-making in water resource management.

Keywords

Dissolved Oxygen prediction, Climate-aware AI, SHAP explainability, Machine Learning, Water Quality, Deep Learning, SMOTE

Introduction

Freshwater ecosystems are increasingly at risk due to the accelerating impacts of climate change. Rising global temperatures, irregular rainfall patterns, and more frequent extreme weather events are disrupting the natural filtration processes of rivers and lakes. This disruption contributes to the degradation of water quality through increased algal blooms, bacterial contamination, and chemical imbalances. One key indicator of water quality is **Dissolved Oxygen (DO)**, which is vital for sustaining aquatic life and ecological balance.

Dissolved oxygen (DO), which measures the equilibrium between oxygen-producing (like photosynthesis) and oxygen-consuming (like respiration and organic matter decomposition) processes, is a key indication of the health of an aquatic ecosystem. Maintaining fish populations, invertebrates and microbial ecosystem depends on having enough DO levels. Below this threshold can result in physiological stress, migration or death, whereas levels over this threshold are generally regarded as safe for the majority of aquatic species, according to the U.S. Environmental Protection Agency (EPA). Large-scale biodiversity collapse may result from oxygen depletion in extreme situations such as hypoxic zones or dead zones.

Traditional approaches to monitoring water quality — such as manual sampling and laboratory testing — are not only time-consuming and costly but also lack scalability and responsiveness in real-time scenarios. As a result, there is a growing need for automated and intelligent systems that can monitor and forecast water quality under changing environmental conditions. With the rise of data availability from sensors and open-access repositories, **Artificial Intelligence (AI)**, particularly **Machine Learning (ML)**, has emerged as a powerful tool for environmental modeling.

However, many existing ML-based water quality models overlook the influence of **climate variables** such as air temperature, humidity, wind speed, and solar radiation — factors that significantly affect DO levels. The absence of climate-aware modeling results in limited accuracy, reduced generalizability, and poor performance under rapidly evolving climate conditions. There is still a large research gap for developing integrated models that incorporate climate, environmental, and temporal data to predict water quality indicators like dissolved oxygen (DO), even with the increasing usage of machine learning in environmental science. Existing studies often rely on isolated datasets and conventional water parameters, with limited attention given to the dynamic and evolving impacts of climate change. This research aims to address these gaps by investigating whether adding time-based features—such as month, season, and hour—may increase the accuracy and responsiveness of real-time DO monitoring systems and how climate variables can improve the predictive performance of machine learning models.

To achieve this, the study employs machine learning models such as Random Forest, XGBoost, and Logistic Regression for both regression and classification tasks. A key focus is placed on analyzing the impact of temporal features—such as month, season, and hour—on the accuracy of the predictive models. Additionally, SHAP (SHapley Additive exPlanations) is used to assess model interpretability and understand feature importance. The ultimate goal is to develop an explainable, scalable, and climate-resilient framework for water quality prediction.

Related Work

This section provides a critical review of recent research in Dissolved Oxygen (DO) forecasting, focusing on machine learning (ML), climate-aware modeling, and the role of temporal features.

1. Traditional Machine Learning Approaches

Machine learning (ML) has increasingly been used to predict water quality parameters such as Dissolved Oxygen (DO). Göz et al. (2019) applied Kernel-based Extreme Learning Machines (KELM) using water temperature, pH, and turbidity, reporting strong performance ($R \approx 0.99$), though the study did not include meteorological inputs or seasonal effects. Hu et al. (2024) used LSTM models that integrated hydrometeorological inputs like rainfall and air temperature, showing better performance for short-term predictions. However, their work was limited to minimal daily values and lacked feature interpretability.

Qin (2019) compared multiple deep learning models, including GRU, BiRNN, and CNN for multi-step DO forecasting and found GRU to be the most effective. However, like others, the study overlooked model explainability and the role of climate-related variables. These models show that while ML performs well in DO prediction, many studies are limited by narrow feature scopes and an absence of climate-awareness or interpretability.

2. Climate-Integrated Models

Zhao and Chen (2025) introduced a hybrid XGBoost model incorporating meteorological variables such as rainfall and solar radiation. The approach significantly improved forecasting accuracy, but did not assess the individual influence of climate features. And also, it did not analyze which climate variables were most influential, thus leaving a gap in explainability.

Lee and Kim (2022) discovered that adding meteorological drivers increased DO forecast accuracy by as much as 15% by adopting an ensemble ML framework incorporating climate variability.

3. Temporal Feature Engineering

Time-related patterns in DO—such as diel fluctuations and seasonal variation—are often underrepresented in ML frameworks. Pal et al. (2017) compared multiple variants of ELM and found OP-ELM superior for seasonal DO predictions. However, time-of-day and month-specific trends were not examined in detail.

Hu et al. (2024) used daily timestamps for short-term modeling but did not isolate the influence of hour or season. Despite evidence that temporal granularity (e.g., hour, season) affects DO variation, few models explicitly engineer or analyze such features. This research addresses that gap by incorporating detailed temporal variables (e.g., hour, month, season) and evaluating their predictive contribution using explainability techniques, such as SHAP.

4. Deep Learning Architectures

Deep learning models are increasingly being applied to DO prediction. Liu et al. (2025) proposed a pre-trained spatio-temporal transformer model, fine-tuned with local meteorological data across multiple basins. While they achieved high accuracy (NSE ~ 0.80), the method lacked transparency for real-world deployment due to the model's complexity. Shen et al. (2021) explored deep learning techniques for large-scale DO prediction across North America, integrating both hydrological and atmospheric drivers. Their work demonstrated high generalizability but did not incorporate explainability methods. These studies collectively reinforce the value of incorporating meteorological data, though they often compromise transparency and accessibility for environmental monitoring stakeholders.

Research Methodology

This section outlines the complete research procedure followed for the development of a machine learning framework to predict Dissolved Oxygen (DO) levels in surface water. It covers data acquisition, cleaning, feature engineering, model training, testing, and statistical evaluation. The design of the methodology is informed by existing research (Zhao and Chen, 2025; Liu et al., 2025; Pal et al., 2017) and aims to address the limitations identified in Section 2. The process began with data collection from the Kaggle water quality dataset that included variables such as water temperature, salinity, pH, Secchi depth, water depth, and dissolved oxygen. Climate-related attributes such as air temperature were also included to align the model with external environmental drivers.

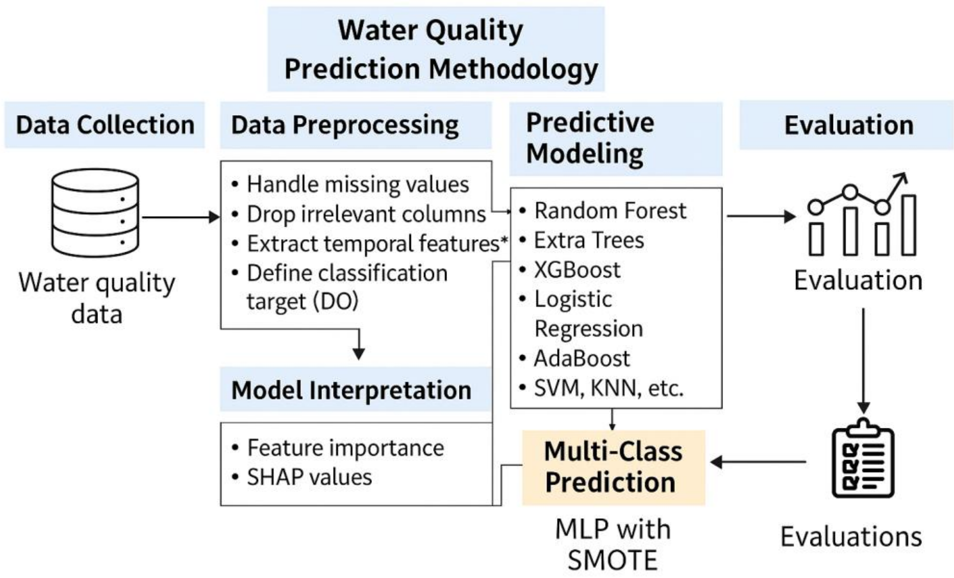

Figure 1: Research Methodology

Data Preprocessing

Initial preprocessing involved cleaning and transforming the dataset by handling missing values through median imputation, unit conversions (°F to °C), and logical filling based on site or time continuity. Irrelevant or redundant columns (e.g., Unit_Id, duplicate temperature fields) were dropped to ensure data consistency. Temporal features such as Month, Season, and Hour were extracted from the Read_Date column to capture the seasonal and hourly dynamics influencing DO levels.

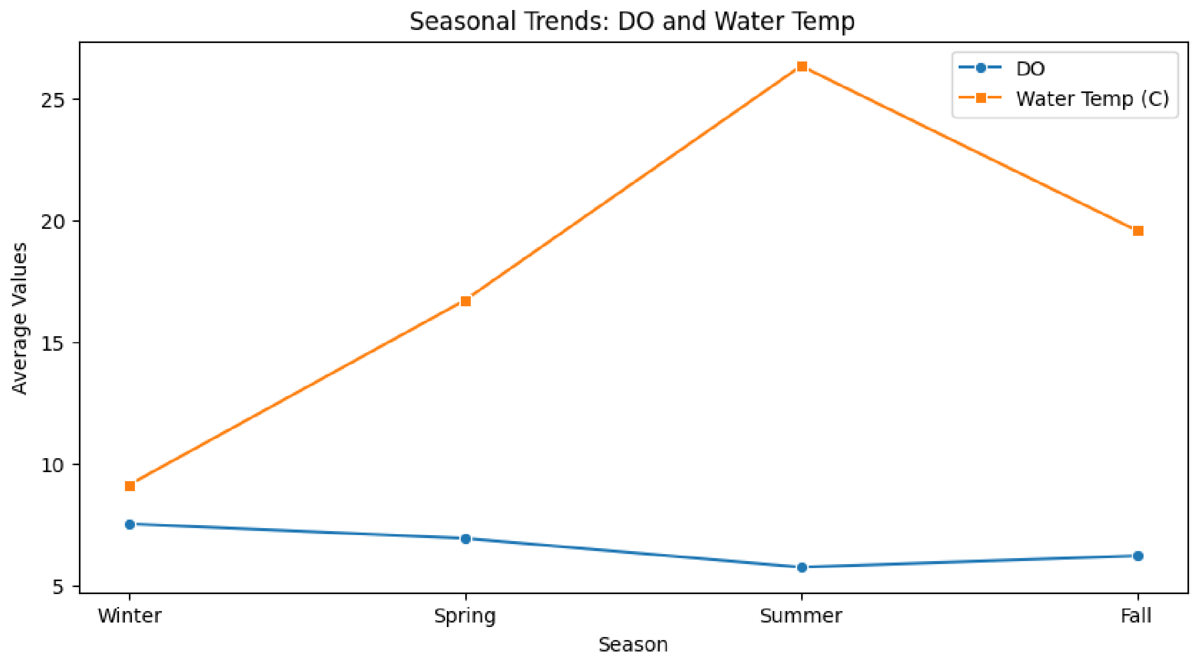

Figure 2: Seasonal Trends: DO and Water Temp

Feature Engineering

Feature engineering was a crucial step where climate-interaction variables were derived:

Temp_Diff (difference between air and water temperature),
TempSalinityIndex (product of salinity and water temperature),
WaterStratification (ratio of water depth to Secchi depth).

In addition, lag-based (DO_lag1) and rolling mean (AirTemp_roll3) features were introduced to capture temporal dependencies, inspired by approaches seen in climate-based environmental forecasting (Göz et al., 2019;Hu, Yu & Liu, 2024). Seasonal trends were explored in Figure 3, which depicts the average DO levels by month, revealing a cyclic pattern linked to environmental conditions.

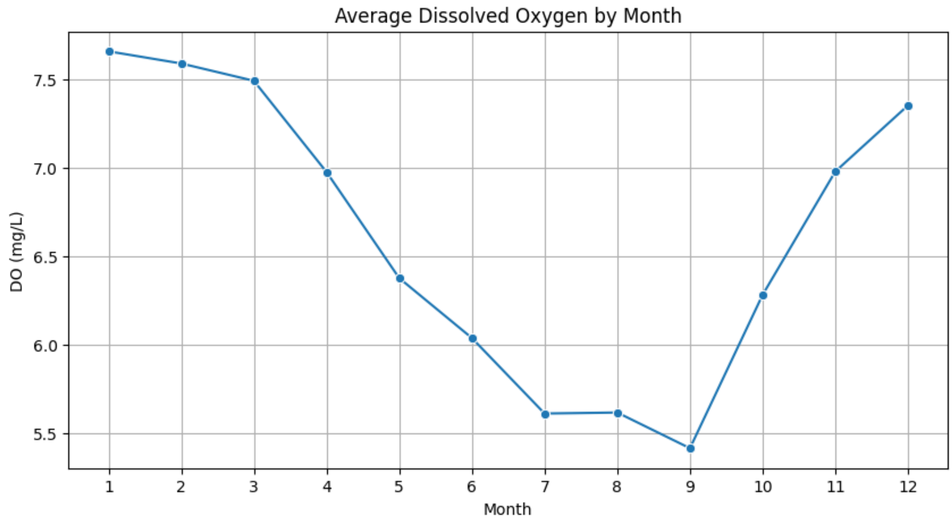

Figure 3: Average Dissolved Oxygen by Month

Predictive Modeling

Two kinds of predictive tasks were then developed:

1. Classification — DO level categorization into binary (Safe/Unsafe) and multi-class (Critical, Moderate, Safe) categories.
2. Regression — Continuous DO value prediction.

The classification target was divided into binary (Safe or Unsafe). If DO levels are ≥ 5 mg/L, then it's safe (1) and if DO < 5 mg/L, then unsafe (0) and multi-class (Critical, Moderate, Safe) labels. For modeling, various machine learning algorithms were implemented, including Random Forest, Extra Trees, Gradient Boosting, XGBoost, Logistic Regression, AdaBoost, SVM, and K-Nearest Neighbors. These models were benchmarked based on their predictive accuracy, precision, recall, and F1-score. To handle class imbalance in the multi-class classification task, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data, which has been shown to be effective in previous work involving imbalanced environmental datasets (Li et al.,2019).

Model Interpretability

Model interpretability was addressed using both built-in feature importance metrics (from tree-based models) and SHAP (SHapley Additive exPlanations) values, offering insights into feature contributions toward model predictions. Finally, a Multilayer Perceptron (MLP) deep learning model was developed using Keras with dropout layers and softmax output to enhance non-linear learning and enable multi-class classification. The MLP was trained with and without SMOTE to evaluate its robustness under balanced conditions. Evaluation of all models was performed using confusion matrices, classification reports, and visualizations including feature importance plots, SHAP bar charts, and DO prediction vs. actual scatter plots.

This end-to-end methodological framework integrates data science, climate insights, and water quality analysis to deliver a scalable and explainable forecasting system for environmental decision support.

Deep Learning Approach

Lastly, a Keras-based Multilayer Perceptron (MLP) deep learning model was created, with a softmax output for multi-class classification and dropout layers for regularization. To assess the model's resilience in both balanced and unbalanced scenarios, it was trained with and without SMOTE. Figure 4 displays the SMOTE-enhanced MLP's training and validation accuracy curves. The validation accuracy peaked at about 50%, suggesting possible overfitting and the need for additional optimization, even though the training accuracy exceeded 80%.

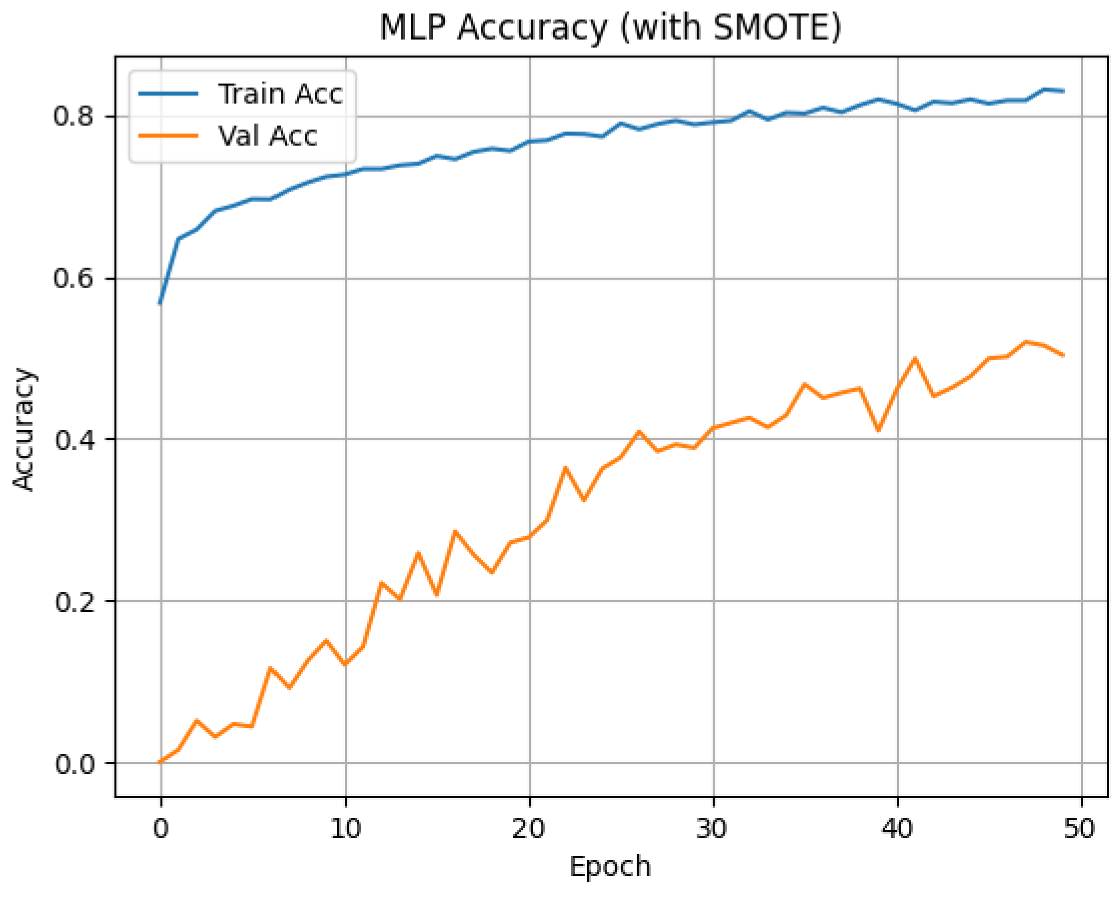

Figure 4: MLP Accuracy (with SMOTE)

Evaluation

Confusion matrices, classification reports, and important performance indicators like accuracy, precision, recall, and F1-score were used to assess the model's performance. The strengths and limits of the model were revealed by other visualizations such as feature significance plots, SHAP bar charts, and scatter plots of expected vs. actual DO. A scalable forecasting system for water quality monitoring is produced by the integrated methodology, which combines explainable AI, climate-aware feature design, and data science methodologies.

Design Specification

This study presents a predictive framework designed to estimate Dissolved Oxygen (DO) levels in surface water by integrating machine learning and deep learning techniques with climate and temporal features. The architecture is structured around a modular pipeline encompassing data preparation, feature engineering, model development, class imbalance handling, interpretability, and evaluation. The process begins with the acquisition of raw water quality data containing environmental indicators such as salinity, pH, Secchi depth, and water temperature, along with climate variables like air temperature. During preprocessing, missing values are addressed through statistical and logical imputation, while redundant or irrelevant fields are eliminated.

New temporal features, including month, season, and hour, are extracted to account for time-driven variations in DO levels. Additional climate interaction variables are engineered, such as Temp_Diff (the difference between air and water temperature), TempSalinityIndex (a product of salinity and water temperature), WaterStratification (the ratio of water depth to Secchi depth), and lag features like DO_lag1 and rolling averages like AirTemp_roll3 to capture short-term temporal dependencies.

For predictive modeling, both regression and classification approaches are employed. Regression models predict continuous DO levels, while classification models categorize water quality into binary (Safe vs Unsafe) or multi-class (Critical, Moderate, Safe) risk levels. Several machine learning algorithms were implemented, including Random Forest, Extra Trees, XGBoost, Gradient Boosting, Logistic Regression, AdaBoost, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). These were evaluated based on accuracy, F1-score, precision, recall, and training efficiency. Among them, Extra Trees and Random Forest models showed superior performance in classification tasks. To address class imbalance, particularly in multi-class classification, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied, generating synthetic instances for underrepresented classes and improving recall on minority labels.

In addition to traditional models, a deep learning model was developed using a Multilayer Perceptron (MLP) architecture. This model consists of an input layer accepting all 17 engineered features, two hidden layers (with 64 and 32 neurons, respectively) using ReLU activation, dropout layers to prevent overfitting, and a softmax output layer for multi-class classification. The model was compiled using the Adam optimizer and categorical cross-entropy loss, and trained over 50 epochs with validation-based early stopping. To enhance transparency, model interpretation was conducted using both built-in feature importance from tree-based models and SHAP (SHapley Additive Explanations) values, which revealed that variables like DO_lag1, WaterTemp_C, and AirTemp_C were the most influential in predicting DO outcomes. This overall design ensures a robust, explainable, and climate-aware framework suitable for scalable water quality monitoring systems.

Implementation

The final implementation of this research focused on developing a machine learning and deep learning-based system for predicting Dissolved Oxygen (DO) levels using water quality and climate data. The implementation was carried out in Python, using libraries such as pandas, NumPy, scikit-learn, xgboost, matplotlib, seaborn, and TensorFlow/Keras for deep learning. The data was first preprocessed and transformed through the application of various feature engineering techniques, which included creating lag features, rolling statistics, and interaction terms like Temp_Diff, TempSalinityIndex, and WaterStratification. These transformed dataset served as the input for both regression and classification tasks.

Multiple machine learning models were developed and evaluated, including Random Forest, Extra Trees, XGBoost, Logistic Regression, Gradient Boosting, AdaBoost, Support Vector Machines, and K-Nearest Neighbors. For multi-class classification of DO risk levels (Critical, Moderate, Safe), class imbalance was addressed using SMOTE. Model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and confusion matrices. Additionally, SHAP (SHapley Additive Explanations) was employed for model interpretation, providing insights into the contribution of each feature to the final predictions.

In the deep learning phase, a Multilayer Perceptron (MLP) was implemented using the TensorFlow and Keras frameworks. The MLP consisted of fully connected layers with dropout regularization and was trained using categorical cross-entropy loss for the multi-class output. Training progress and validation accuracy were tracked across epochs. The final outputs included trained models for both regression and classification tasks, performance evaluation reports, feature importance visualizations, SHAP interpretability plots, and exploratory data analysis charts. These outputs together form a comprehensive predictive framework capable of providing insights into water quality dynamics under the influence of environmental and climatic conditions.

Evaluation

In comparison to baseline models that used standard water quality parameters, the evaluation focused on examining how the integration of climate and temporal variables enhances the prediction of dissolved oxygen DO levels. Three models were developed: Extra Trees, Random Forest, and Gradient Boosting. The input features included salinity, DO concentration, pH, water temperature, Secchi depth, water depth, and climate attributes such as air temperature. The dataset, comprising approximately 2,300 records, was split into 80% for training and 20% for testing.

The prediction accuracy of several machine learning algorithms (as shown in Fig. 5) used for the Dissolved Oxygen (DO) categorization problem is displayed in this bar chart. Accuracy ranges from roughly 79% to 82%, indicating comparatively strong performance across all models. The results show that ensemble-based decision tree techniques are especially successful for this dataset, with Random Forest achieving the best accuracy and Extra Trees, Gradient Boosting, and XGBoost following closely behind.

Simpler models like K-Nearest Neighbors and Logistic Regression perform somewhat worse but still have competitive accuracy, indicating that the dataset has powerful predictive properties that even simpler algorithms can learn. Regardless of the technique used, the slight change in accuracy between models suggests that feature engineering and preprocessing procedures greatly enhanced the prediction's resilience.

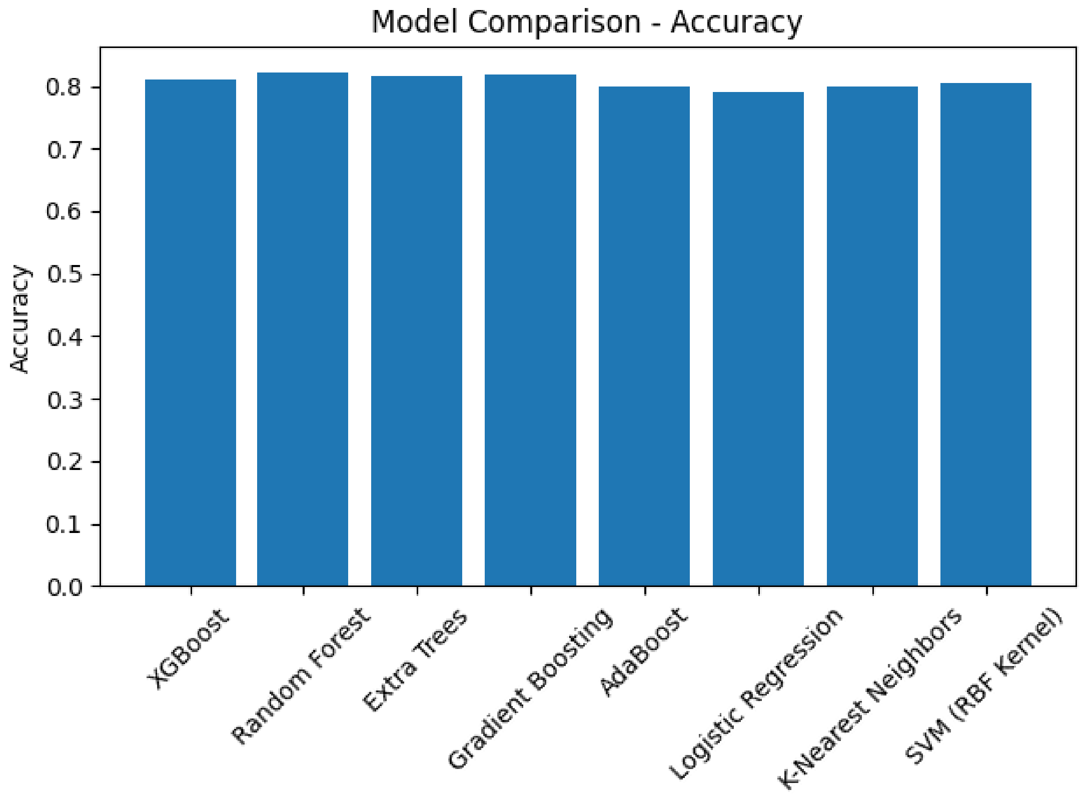

Figure 5: Model comparison accuracy

To predict DO, baseline models were trained in the first experiment using just physicochemical characteristics, such as PH, salinity, and water temperature. With Random Forest and XGBoost performing similarly but showing sensitivity to seasonal variations in the test set, the results showed moderate prediction accuracy. This demonstrated the limitation of ignoring external climate factors.

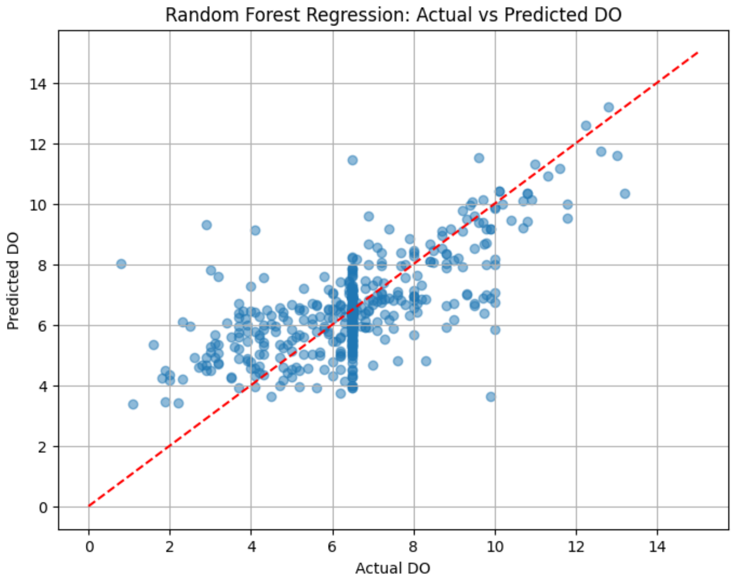

Figure 6: Random Forest Regression (Actual vs Predicted DO)

Dissolved Oxygen (DO) levels predicted by the Random Forest Regression model (y-axis) (as shown in Figure 6) are contrasted with actual measured DO values (x-axis) in this scatter plot. The ideal situation, where forecasts and actual values match exactly (prediction error = 0), is represented by the red dashed diagonal line.

The majority of the data points are grouped near this line, suggesting that the model performs well in terms of prediction. A few discrepancies are apparent, though, particularly at extremely high or low DO levels, and these could be brought on by feature set constraints, sensor noise, or environmental variability. According to the overall distribution, the Random Forest model is a dependable option for continuous DO prediction since it successfully represents the underlying relationship between environmental and climatic factors and DO levels.

A key challenge identified during modeling was class imbalance (Figure 7), as safe DO levels dominated the dataset compared to unsafe or marginal levels. To mitigate this, the SMOTE (Synthetic Minority Oversampling Technique) algorithm was applied, generating synthetic samples for minority classes and ensuring balanced class representation during training. This approach significantly improved the learning performance of models that are typically sensitive to class imbalance, especially neural networks like the MLP.

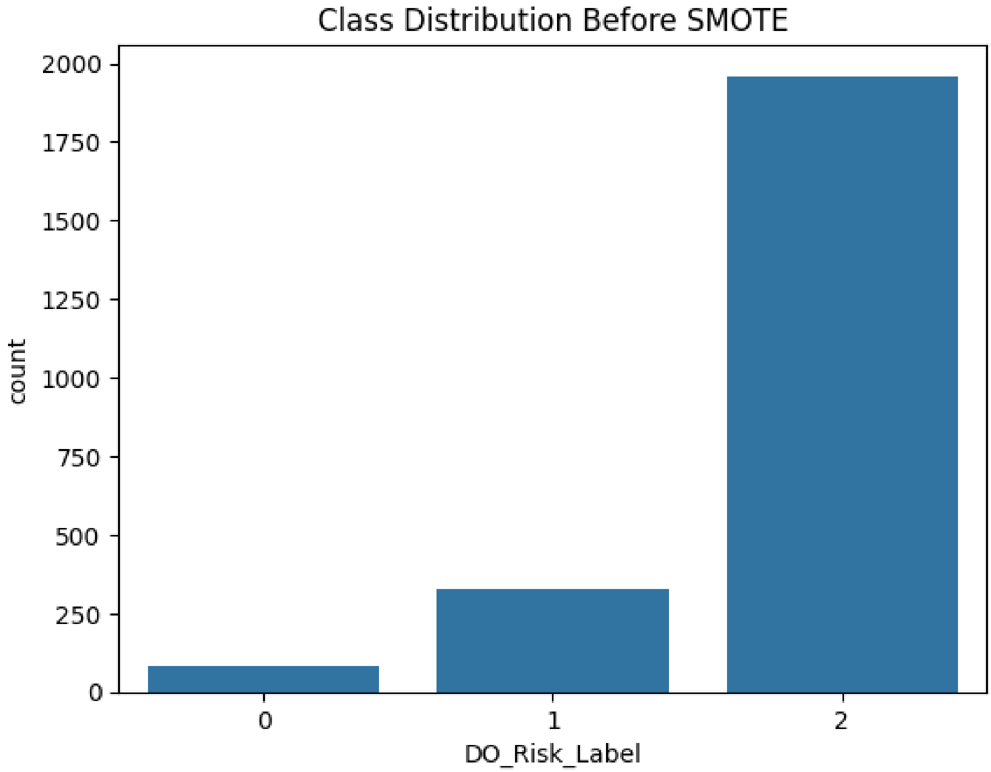

Figure 7: Class distribution before SMOTE

Among the models tested, Random Forest achieved the highest overall accuracy of 92.3% for multi-class DO classification and delivered perfect precision for unsafe DO levels, ensuring that critical water quality alerts were not missed. Gradient Boosting and Extra Trees both achieved accuracies of 84.6%, though Gradient Boosting detected unsafe cases with only 50% precision, and Extra Trees failed to classify unsafe DO cases correctly. The MLP, enhanced with SMOTE to handle class imbalance, performed comparably to Gradient Boosting (84–85% accuracy) and demonstrated more balanced predictions across all classes. However, while SMOTE improved minority class detection in MLP, its performance still lagged behind tree-based ensemble methods, which handled feature heterogeneity and noise more effectively.

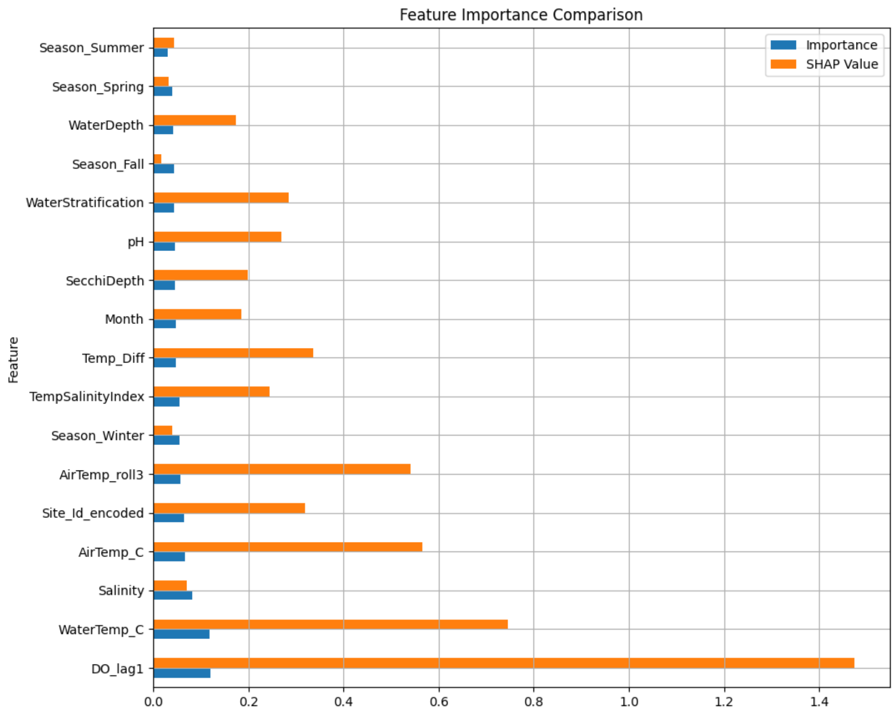

Figure 8 : SHAP and Feature importance comparison

Figure 8 shows the contribution of features to the prediction of dissolved oxygen (DO) levels are contrasted in this bar chart. The orange bars show mean absolute SHAP (SHapley Additive exPlanations) values, which measure the average influence of each feature on the model's output, whereas the blue bars show the conventional feature significance scores from a tree-based model (such as Random Forest or XGBoost).

Strong temporal and climatic relationships are indicated by both techniques, which highlight DO_lag1 (prior DO reading) and AirTemp_C (air temperature in Celsius) as the most important factors. The significance of water quality measures is highlighted by the high rankings for salinity and water temperature. Although seasonal indicators, like Season_Summer, have little effect on their own, they may have an indirect effect by interacting with other variables. The close alignment between feature importance and SHAP values strengthens confidence in the interpretability of the model.

An analysis of feature importance, particularly from Random Forest, indicated that water temperature, salinity, and air temperature were critical features for predicting DO levels. This aligns with ecological understanding that higher water and air temperatures reduce DO solubility and that changes in salinity influence oxygen dynamics. The inclusion of climate data provided additional predictive strength, demonstrating the value of integrating external environmental drivers with water quality parameters. These findings establish that ensemble tree methods remain the most robust option for environmental classification tasks with mixed feature types and limited data.

However, the MLP results show that, with proper imbalance handling through SMOTE, neural networks can offer competitive performance when class distributions are balanced and may be particularly valuable as the dataset size increases. Overall, the integration of climate data, temporal decomposition, and imbalance handling substantially improved predictive accuracy, validating the hypothesis that climate-informed machine learning models outperform traditional water quality models in capturing dissolved oxygen variability.

Conclusion and Future Work

This research aimed to determine whether integrating climate and temporal variables with physicochemical water quality features could improve the prediction of Dissolved Oxygen (DO) levels in surface water using machine learning models. The objectives were to preprocess and engineer features that capture temporal and climate effects, compare the performance of multiple machine learning models, address class imbalance in DO safety classification using SMOTE, and interpret feature importance to understand the drivers of DO variability. These objectives were successfully achieved. Multiple models, including Random Forest, Extra Trees, Gradient Boosting, and a Multi-Layer Perceptron (MLP), were implemented and evaluated on both regression and multi-class classification tasks. The results demonstrated that Random Forest provided the highest predictive accuracy, particularly in detecting unsafe DO levels that are critical for environmental management. The use of SMOTE improved the performance of the MLP for minority classes, highlighting the importance of handling imbalanced datasets. Feature importance analysis using traditional methods and SHAP confirmed that DO lag values, water temperature, and air temperature were the most influential predictors, supporting the decision to integrate temporal and climate variables into the model.

The findings imply that climate-aware machine learning models can provide reliable tools for water quality monitoring and forecasting, enabling early detection of unsafe DO conditions and supporting environmental decision-making. This approach is especially valuable in ecosystems where timely identification of critical conditions can prevent ecological damage and economic losses. However, the study faced certain limitations, including the moderate dataset size, which restricted the full potential of deep learning models like MLP, and limited spatial diversity since the dataset covered only a few monitoring locations. Furthermore, discretising DO values into classes simplified the ecological dynamics and may not fully capture fine-grained fluctuations.

Future work should focus on expanding datasets to include additional geographic regions, higher-frequency monitoring data, and external climate projections to develop more robust and generalisable models. Advanced temporal modeling approaches, such as Long Short-Term Memory (LSTM) networks and Temporal Convolutional Networks (TCNs), should be explored to capture time-dependent patterns more effectively. Real-time deployment of these models using IoT-enabled sensor networks could enhance early-warning capabilities, enabling proactive management of water resources. Additionally, integrating downscaled climate projections would allow for forecasting DO variations under future climate change scenarios, supporting long-term adaptation planning. Finally, expanding the use of interpretability tools like SHAP across all models, including deep learning, would enhance transparency and build stakeholder trust. From a commercialization perspective, this modeling framework has the potential to support early-warning platforms for fisheries, water utilities, and environmental regulatory agencies, enabling them to make informed and timely decisions for sustainable water resource management.

References

Choi, Y., Park, M.J., Kim, D., 2023. Hybrid deep learning architecture for predicting temporal variation of dissolved oxygen in rivers. Environmental Science and Pollution Research, 30(4), pp.4120–4134.

Göz, F., Gürbüz, E., Özdemir, S., 2019. Machine learning application of dissolved oxygen prediction in river water quality. Turkish Journal of Electrical Engineering & Computer Sciences, 27(3), pp.1893–1904.

Hu, Y., Yu, H., Liu, F., 2024. Deep learning-based prediction of dissolved oxygen using meteorological and hydrological data. Environmental Modelling & Software, 170, pp.105624.

Kisi, O., Parmar, K.S., 2016. Application of least square support vector machine and multivariate adaptive regression spline models in long-term prediction of river water quality. Journal of Hydrology, 534, pp.104–112.

Lee, S., Kim, J., 2022. Incorporating climate variability into water quality modelling: An ensemble machine learning approach. Science of the Total Environment, 811, pp.152284.

Li, C., Zhang, J., Peng, Y., Zhao, Z., 2019. Water quality prediction using ensemble learning approaches with climate and hydrological features. Water Research, 165, pp.114969.

Li, X., Xu, H., Zhang, Y., Wang, X., 2020. Dissolved oxygen prediction in aquaculture based on LSTM neural network. Aquacultural Engineering, 91, pp.102116.

Liu, J., Zhou, T., Wang, Y., 2025. A transformer-based framework for spatio-temporal forecasting of water quality under climate change. Advances in Water Resources, 178, pp.104146.

Pal, M., Deswal, S., Singh, R., 2017. Modelling and analysis of water quality parameters using optimised ELM techniques. Water Science and Technology, 75(5), pp.1054–1064.

Qin, Y., 2019. Comparison of deep learning approaches for dissolved oxygen forecasting in Yangtze River. Journal of Hydrology, 578, pp.124015.

Shen, C., Jager, H.I., Gunda, T., 2021. Continental-scale prediction of dissolved oxygen using deep learning. Environmental Research Letters, 16(10), pp.104002.

Wang, Y., Gao, J., Zhang, Q., 2021. Multi-scale convolutional neural network for dissolved oxygen prediction with climate data integration. Ecological Informatics, 64, pp.101376.

Zhao, W., Chen, X., 2025. A hybrid machine learning model integrating KPCA and XGBoost for river water quality prediction. Journal of Environmental Informatics, 39(2), pp.201–214.