IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Hybrid machine learning forecasting of aquatic ecosystem dynamics using sensor-based monitoring systems

Gabit Shuitenov[1], Alua Turginbayeva[2*], Madi Muratbekov[3], Akbota Abylayeva[4], Serik Altynbek[1]

[1]Department of Information Systems and Technologies, Institution "Esil University", Astana, Republic of Kazakhstan.
[2]Department of Computer and Software Engineering, Faculty of Information Technologies, Non-profit Joint Stock Company "L.N. Gumilyov Eurasian National University", Astana, Republic of Kazakhstan.
[3]Department of Information Security, Faculty of Information Technologies, Non-profit Joint Stock Company "L.N. Gumilyov Eurasian National University", Astana, Republic of Kazakhstan.
[4]Department of Fundamental Mathematics, The Mechanics and Mathematics faculty, Non-profit Joint Stock Company "L.N. Gumilyov Eurasian National University", Astana, Republic of Kazakhstan.

Corresponding author: Alua Turginbayeva (*Email:* tasheart@mail.ru)

Abstract

This study aims to develop an integrated hybrid machine learning model that combines sensor-based environmental monitoring for the accurate prediction of aquatic ecosystem dynamics, focusing on key water quality indicators. A novel hybrid framework integrating CatBoost and XGBoost regressors was constructed, optimized through LASSO feature selection, and enhanced by SHAP analysis for interpretability. Real-time data on dissolved oxygen, hardness, transparency, and nutrient content were collected using IoT-enabled multi-parameter water sensors in lakes in Northern Kazakhstan. The hybrid model outperformed individual algorithms, achieving an RMSE of 0.362 for dissolved oxygen predictions. SHAP analysis revealed that nitrate nitrogen, total phosphorus, pH, and suspended solids were the most significant influencing factors. Additionally, the system effectively forecasted impacts on biota, indicating a potential reduction in phytoplankton and an increase in zooplankton populations in 2024. The integration of hybrid machine learning with real-time monitoring significantly improves prediction accuracy and interpretability, providing a practical decision-support tool for environmental agencies and water resource managers to proactively monitor and manage water bodies under the pressures of climate change and anthropogenic influences.

Keywords: Aquatic ecosystems, CatBoost, XGBoost, IoT, Machine learning, Sensor-based monitoring, SHAP analysis, Water quality prediction.

DOI: 10.53894/ijirss.v8i5.9168
Funding: This work is supported by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Kazakhstan (Grant number: AP23489791).
History: Received: 19 June 2025 / **Revised:** 23 July 2025 / **Accepted:** 25 July 2025 / **Published:** 7 August 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

With increased ecological stress and climatic variability, timely and accurate assessment of aquatic ecosystems is essential [1, 2]. Conventional sampling methods are limited by temporal lags and labor-intensive processing, making them unsuitable for continuous high-resolution environmental analysis [3, 4]. In contrast, sensor-based water quality monitoring systems provide high-frequency, real-time measurements of key parameters, offering a dynamic view of aquatic system health [5-7]. However, despite advances in sensor technologies, transforming high-frequency water quality data into actionable ecological insights remains challenging due to the complexity of the data and the lack of robust predictive frameworks [8].

The primary objective of this study is to develop a hybrid machine learning model that accurately predicts key water quality parameters and provides interpretable insights into the drivers affecting aquatic ecosystems. Previous studies have primarily focused on either traditional statistical approaches or single-model machine learning techniques [9], lacking comprehensive hybrid solutions that integrate interpretability and real-time sensor data for aquatic environments. This study aims to address the following research question: (1) Can a hybrid CatBoost–XGBoost model enhance the prediction accuracy of water quality parameters compared to individual models? (2) What environmental factors most significantly influence these predictions? (3) How can SHAP analysis enhance the interpretability of model outputs for ecological decision-making?

To address these questions, the research involved developing a sensor-based monitoring system, preprocessing and feature selection using LASSO regression, constructing and evaluating a hybrid machine learning model, and applying SHAP analysis to interpret the results. Thus, the presented study fits organically into the modern scientific context, where the integration of machine learning methods and ecological modeling is becoming an essential tool for the sustainable management of natural resources. Future development in this area is associated with the improvement of interpretable machine learning methods, an increase in the volume of available data for analysis, and the introduction of artificial intelligence technologies for modeling complex environmental processes [10, 11] and the creation of complex monitoring systems using satellite data [12, 13] and IoT devices [14, 15].

In this study, we propose a hybrid machine learning framework that integrates CatBoost and XGBoost regressors, optimized through LASSO feature selection and enhanced by SHAP analysis for interpretability. Unlike previous studies that relied mainly on single-model approaches or did not incorporate explainability, our method combines the strengths of two gradient boosting techniques, offering a transparent understanding of feature importance. This integrative approach enables more accurate and interpretable predictions of key water quality indicators and biotic responses, setting it apart from traditional black-box models used in earlier research.

2. Literature Review

Monitoring and forecasting aquatic ecosystems have become critical topics in environmental research due to the accelerating effects of climate change, urbanization, and industrial impacts. Traditional manual sampling methods often fail to capture high-frequency fluctuations, limiting the effectiveness of water quality assessments [16]. Recent advancements in remote sensing and Internet of Things (IoT) technologies have enabled continuous, real-time monitoring of key water parameters, providing more dynamic and accurate insights [17, 18]. Machine learning has emerged as a powerful tool for analyzing complex environmental datasets. Chen, et al. [19] demonstrated that tree-based ensemble models, such as random forests, outperform linear models in predicting nutrient concentrations and eutrophication events in lake systems [20]. Similarly, Li et al. used deep learning models to predict chlorophyll a dynamics, demonstrating significant improvements over classical statistical approaches [21]. However, these models often function as black boxes, limiting their interpretability for practical decision-making in ecosystem management [22]. Hybrid machine learning approaches have recently gained traction as solutions to improve predictive performance while maintaining interpretability. For example, Wang et al. combined support vector machines with genetic algorithms to optimize forecasts of water quality parameters, achieving higher accuracy and reduced overfitting Wang, et al. [23]. Xu, et al. [18] proposed a hybrid long short-term memory (LSTM) and convolutional neural network (CNN) model to predict dissolved oxygen levels, demonstrating improved temporal resolution and robustness to noise [24]. Nevertheless, these studies often neglect the integration of real-time sensor data and fail to provide transparent explanations for model predictions.

To address these gaps, interpretability techniques such as SHAP (Shapley Additive Explanations) have been introduced. According to Zhang et al., integrating SHAP values into water quality prediction models allows stakeholders to understand the influence of each environmental factor, thereby supporting more informed management strategies [25]. However, despite these advancements, comprehensive frameworks that combine hybrid models, real-time sensor data, and interpretability remain limited.

In summary, existing studies highlight the potential of machine learning and hybrid approaches for aquatic ecosystem monitoring; however, they often lack the integration of real-time data and explainability. This underscores the need for comprehensive, interpretable hybrid models, motivating the development of our proposed CatBoost–XGBoost framework, which is integrated with SHAP analysis.

3. Materials and Methods

The analytical framework presented in this study is based on a hybrid approach that combines CatBoost and XGBoost regressors. This design enhances prediction accuracy and robustness, while integrating SHAP analysis enables a detailed interpretation of feature contributions. Compared to earlier works that typically used standalone machine learning models without interpretability mechanisms, our method provides both high-performance predictions and actionable ecological insights. To analyze and predict changes in dynamic aquatic ecosystems, this paper employs machine learning methods that process multivariate data, identify key influencing factors, and predict changes in water quality parameters. The developed methodology comprises several stages: data collection and preprocessing, feature selection, building and training predictive models, evaluating their effectiveness, and conducting an interpretable analysis of feature significance.

3.1. Sensor-Based Monitoring System

Water quality data were collected using multi-parameter sensors (e.g., YSI EXO2 or equivalent), which are capable of measuring dissolved oxygen (DO), pH, conductivity, turbidity, temperature, and nutrient concentrations. The sensors were deployed in a lake in Northern Kazakhstan, connected to LoRaWAN modules for low-power, long-range wireless communication. Measurements were transmitted every 15 minutes to a cloud-based platform for storage and analysis. The collected data were preprocessed and served as input for the machine learning pipeline, described in Section 2.2.

3.2. Data Preprocessing and Hybrid Machine Learning Model Development

The hybrid model development algorithm, as illustrated in Figure 1comprises several key stages, each of which plays a crucial role in ensuring the accuracy and effectiveness of the forecasting. The developed hybrid model provides a detailed discussion of the execution steps and interactions of various machine learning methods and tools used in the hybrid model, which combines XGBoost [26, 27] and CatBoost [28-30]. A detailed mathematical description of how data moves along the chain and interacts between stages is presented below.

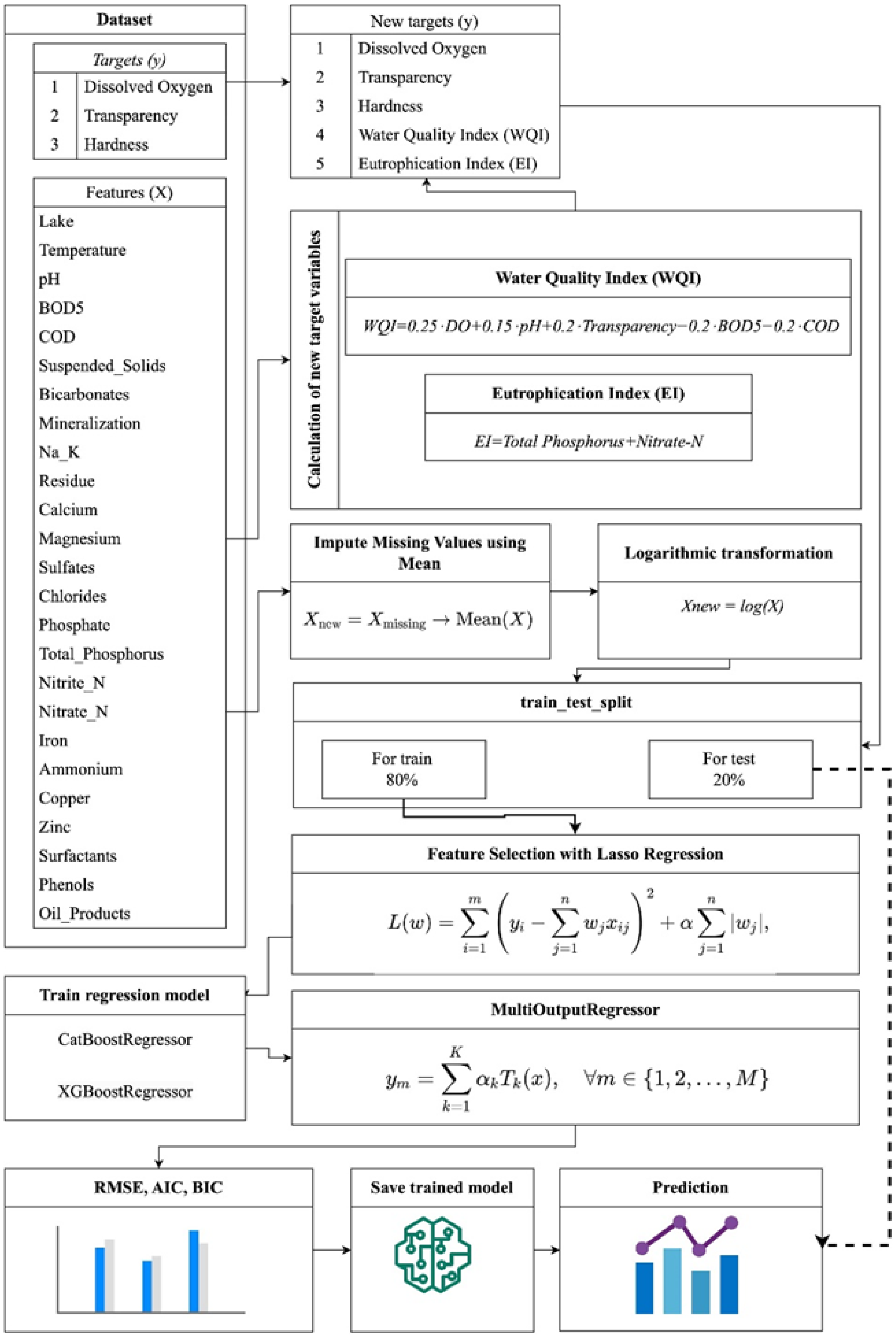
Dataset
Targets (y)
1 Dissolved Oxygen
2 Transparency
3 Hardness
Features (X)
Lake
Temperature
pH
BOD5
COD
Suspended_Solids
Bicarbonates
Mineralization
Na_K
Residue
Calcium
Magnesium
Sulfates
Chlorides
Phosphate
Total_Phosphorus
Nitrite_N
Nitrate_N
Iron
Ammonium
Copper
Zinc
Surfactants
Phenols
Oil_Products
New targets (y)
1 Dissolved Oxygen
2 Transparency
3 Hardness
4 Water Quality Index (WQI)
5 Eutrophication Index (EI)
Calculation of new target variables
Water Quality Index (WQI)
$WQI=0.25\cdot DO+0.15\cdot pH+0.2\cdot Transparency-0.2\cdot BOD5-0.2\cdot COD$
Eutrophication Index (EI)
$EI=Total\ Phosphorus+Nitrate\text{-}N$
Impute Missing Values using Mean
$X_{\text{new}} = X_{\text{missing}} \rightarrow \text{Mean}(X)$
Logarithmic transformation
$Xnew = log(X)$
train_test_split
For train 80%
For test 20%
Feature Selection with Lasso Regression
$L(w) = \sum_{i=1}^{m}\left(y_i - \sum_{j=1}^{n} w_j x_{ij}\right)^2 + \alpha \sum_{j=1}^{n} |w_j|,$
Train regression model
CatBoostRegressor
XGBoostRegressor
MultiOutputRegressor
$y_m = \sum_{k=1}^{K} \alpha_k T_k(x), \quad \forall m \in \{1, 2, \ldots, M\}$
RMSE, AIC, BIC
Save trained model
Prediction

Figure 1.
Algorithm for developing a hybrid model.

In this study, a comprehensive methodology for constructing a hybrid machine-learning model to analyze and predict water quality parameters was developed. The method comprises several key stages, each designed to enhance the accuracy of the forecast and the interpretability of the results. The sequence of stages is presented below:

1. Data loading and preprocessing. Step one is to read data from a file. The data consists of several water quality indicators, including Dissolved Oxygen, Hardness, Transparency, Biological Oxygen Demand (BOD5), and other chemical and physical properties. Categorical data is encoded to numerical data using LabelEncoder. String values (e.g., lake names, periods) are encoded as numbers in this step to enable machine learning algorithms to process these features. Let us consider a data set X of size m × n, where m is the sample size and n is the number of features (variables), which can be numeric or categorical. The transformed categorical variables are passed to the model for further calculations.

2. Filling gaps. Gaps in the data are filled using the SimpleImputer method with the parameter strategy='mean'. For each feature (variable), the mean is calculated, and the gaps are replaced by this mean (1):

$$x_{ij} = \frac{1}{m}\sum_{k=1}^{m} x_{ij} \qquad (1)$$

where x_{ij} is the missing value for variable j in row i. This step is essential to ensure the model can work correctly with complete data.

3. Calculation of new target variables. One of the most critical steps is the development of two new indices to characterize water quality: the Water Quality Index (WQI) and the Eutrophication Index (EI). The Water Quality Index (WQI), which is computed through formula (2):

$$WQI = 0.25 \cdot DO + 0.15 \cdot pH + 0.2 \cdot Transparency - 0.2 \cdot BOD5 - 0.2 \cdot COD \qquad (2)$$

where DO is dissolved oxygen, pH is acidity, BOD5 is biological oxygen demand, and COD is chemical oxygen demand. The Eutrophication Index, which is calculated as (3):

$$EI = Total_{phosphorus} + Nitrate\_N \qquad (3)$$

where Total_Phosphorus is the total phosphorus content, Nitrate_N is the concentration of nitrates in water. Now, the matrix of target variables y includes (4):

$$y = [DO, Hardness, Transparency, WQI, EI] \qquad (4)$$

4. Logarithmic transformation of features. During data preparation, all variables with exclusively positive values are subjected to a logarithmic transformation to stabilize variance and enhance the model's learning capacity (5).

$$X_{new} = \log(X) \qquad (5)$$

where X is the original variable, and X_{new}\ is its logarithmic transformation. Logarithmic transformation is useful when the data has a skewed distribution or strong outliers. This transformation is applied to all X matrix features with values greater than zero.

5. Splitting the data into training and test sets. The data is divided into a training set (train) and a test set (test). The training set comprises 80% of the total data, while the test set accounts for 20%. You can train the model using one-half of the data and evaluate its accuracy on the other half, which is independent, by this split.

6. Feature selection using Lasso regression. The next step uses Lasso regression (L1 regularization) to select the most significant features. The loss function that is minimized for feature selection is presented as (6):

$$L(w) = \sum_{i=1}^{m}\left(y_i - \sum_{j=1}^{n} w_j x_{ij}\right)^2 + \alpha \sum_{j=1}^{n}\left|w_j\right| \qquad (6)$$

where L(w) is the loss function, w_j are the weights of the features, and α is the regularization coefficient. The selected features are used to train the models, thereby increasing their efficiency and reducing data redundancy. This method allows you to zero out the coefficients of insignificant features, leaving only the most important ones for forecasting. Features with non-zero coefficients are selected for further use in the model.

7. Training a hybrid multi-task model. A multi-tasking model is used to solve the multi-task forecasting problem. The multi-task model is built on either CatBoostRegressor or XGBRegressor and enables the simultaneous prediction of multiple target variables. This makes the process more efficient since the same features can influence multiple target variables. The main differences between CatBoost and XGBoost are:

- CatBoost builds gradient-boosting models based on categorical features and utilizes specialized algorithms to handle missing data. It minimizes the following loss function (7):

$$L_{CatBoost} = \sum_{i=1}^{N}(y_i - \sum_{k=1}^{K} \alpha_k T_k(x_i))^2 \qquad (7)$$

where $T_k(x_i)$ is the decision tree at step k, α_k is the weight of the tree, and k is the number of trees.

- XGBoost uses a tree-based boosting method with a broader range of regularization functions (8):

$$L_{XGBoost} = \sum_{i=1}^{N}(y_i - \sum_{k=1}^{K} \alpha_k T_k(x_i))^2 + \lambda \sum_{k=1}^{K} T_k(x_i)^2 \qquad (8)$$

Here, λ is a regularization parameter for decision trees, which helps to prevent overfitting. XGBoost also optimizes computations through parallelism and second-order regularization, enhancing the model's robustness. Both models can be employed for multi-task learning using the MultiOutputRegressor method, which enables the construction of independent models for each target variable (9):

$$y_m = \sum_{k=1}^{K} \alpha_k T_k(x), \forall m \in (1,2,\dots,M) \qquad (9)$$

where M is the number of target variables (in this case, 5).

8. Prediction on test data. After training, the model performs a prediction on test data (10):

$$\widehat{y_{test}} = f(X_{test}) \qquad (10)$$

where y_{test} are the predicted values on the test data.

9. Evaluation of the quality of models (RMSE, AIC, BIC). To evaluate the quality of trained models, several metrics are used, such as:

- RMSE (root mean square error) (11) measures the average deviation of predicted values from the true ones.

$$MSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \qquad (11)$$

where y_i is the actual value, $\hat{y}_i$ is the value predicted by the model.

- AIC (Akaike Information Criterion) (12) and BIC (Bayesian Information Criterion) (13) are measures that compare the quality of model fit relative to its complexity. Lower values of these indicators suggest a higher quality of forecasting by the model.

$$IC = 2p - 2\log L \tag{12}$$

$$BIC = \log(n) \cdot p - 2\log L \tag{13}$$

where p is the number of model parameters, n is the number of observations, and L is the maximum likelihood value.

9. Saving the trained model. The trained model is saved for future use on new data. This allows you to avoid retraining the model and to use it immediately for forecasting.

10. Forecasting. After training, the model forecasts values on the test set. Parameters such as Dissolved Oxygen, Hardness, Transparency, Water Quality Index, and Eutrophication Index are forecasted.

10. SHAP analysis to explain factors. SHAP analysis explains the significance of features on the model results. It illustrates the impact of each feature on the prediction of each target variable. SHAP values provide insights into which factors significantly impact parameters such as dissolved oxygen, hardness, or eutrophication.

The hybrid model combines XGBoost and CatBoost, enabling more accurate prediction of target variables compared to using only one model. Lower RMSE values confirm this. Using SHAP analysis helps you better understand which features have the most significant impact on each index, which is especially important for interpreting the results. Feature selection using Lasso regression reduces the model's complexity, improving its learning ability and prediction speed. Different regressors, such as CatBoost and XGBoost, offer greater flexibility in selecting the optimal method for a specific task. Thus, the hybrid model, which utilizes XGBoost and CatBoost, demonstrated superiority in both learning and prediction due to a more accurate data fit and better consideration of various factors.

4. Results

The study conducted a comprehensive evaluation of water quality in North Kazakhstan's Zerendi and Kopa lakes from June 2015 to September 2023. The total number of observations was 4,351, including physicochemical water parameters such as dissolved oxygen (DO), hardness, transparency, pH, temperature, BOD5, COD, and the content of nitrates, phosphates, sulfates, minerals, and suspended matter. Both the eutrophication index (EI) and the water quality index (WQI) were calculated from the data and served as target variables for developing a hybrid predictive model based on CatBoostRegressor and XGBRegressor algorithms. The data revealed high seasonal variability in physicochemical parameters within the aquatic environment, as revealed by the 2021 data collected in June–September within the framework of the general monitoring from 2015 to 2023, was observed. As a result water temperature in Lake Zerendi ranged from 7.2°C in May to 22.6°C in July, while in Lake Kopa, it ranged from 8.0°C to 23.2°C. Dissolved oxygen levels in Zerendi were 13.19 mg/dm³, significantly above the sanitary norm of at least 5 mg/dm³, whereas in Kopa, levels fluctuated from 6.02 to 9.12 mg/dm³. A notable difference was observed in chemical oxygen demand (COD): in September 2021, it was 64.5 mg/dm³ in Zerendi, exceeding the maximum permissible level of 30 mg/dm³ for recreational waters, while in Kopa, it reached 37.7 mg/dm³. Suspended matter content was 16.2 mg/dm³ in Kopa and decreased to 5.2 mg/dm³ in Zerendi during summer. Mineralization levels exceeded 1000 mg/dm³, with Zerendi ranging from 1204 mg/dm³, and Kopa approaching the critical threshold of 998 mg/dm³. The 2021 data, part of long-term monitoring, clearly indicate the influence of climatic and seasonal conditions on water quality, underscoring the importance of regular environmental monitoring to prevent eutrophication and ecosystem degradation.

The study pays particular attention to analyzing the influence of individual factors on the state of aquatic systems using interpretable machine learning methods, such as SHAP analysis. SHAP (SHapley Additive exPlanations) is a method for explaining a machine learning model that evaluates the contribution of each feature to the final model predictions. The SHAP summary presents the influence of various factors (or features) on the model predictions, illustrating how much each feature contributes to or detracts from the predicted value. The SHAP summary analysis enables us to discuss which features (e.g., the content of dissolved substances in water, temperature, etc.) have the most significant impact on the model predictions. The higher the SHAP value for a particular feature, the greater its influence on the final model result. Figure 2 illustrates the SHAP summary, displaying the results of the analysis of the impact of different factors on the prediction of dissolved oxygen concentration in a reservoir. The most significant influences are from parameters such as nitrate nitrogen (Nitrate_N), pH, and total phosphorus (Total_Phosphorus). These parameters have a substantial impact on the photosynthesis processes of aquatic plants, leading to an increase in the concentration of oxygen in the water. Nitrate-N promotes the growth of marine plants and phytoplankton, thereby increasing oxygen production through photosynthesis. There is also pH, an alkaline condition that initiates photosynthesis, and phosphorus supports plant growth at a high rate. Temperature also harms it, as the former lowers water's capacity to retain oxygen as its value rises. Suspended solids, such as those in water, lower water transparency, inhibit photosynthesis, and harm the amount of oxygen. Therefore, the diagram illustrates that chemical parameters indicating nutrient and photosynthesis conditions are the primary determinants of the dissolved oxygen level in a water body.

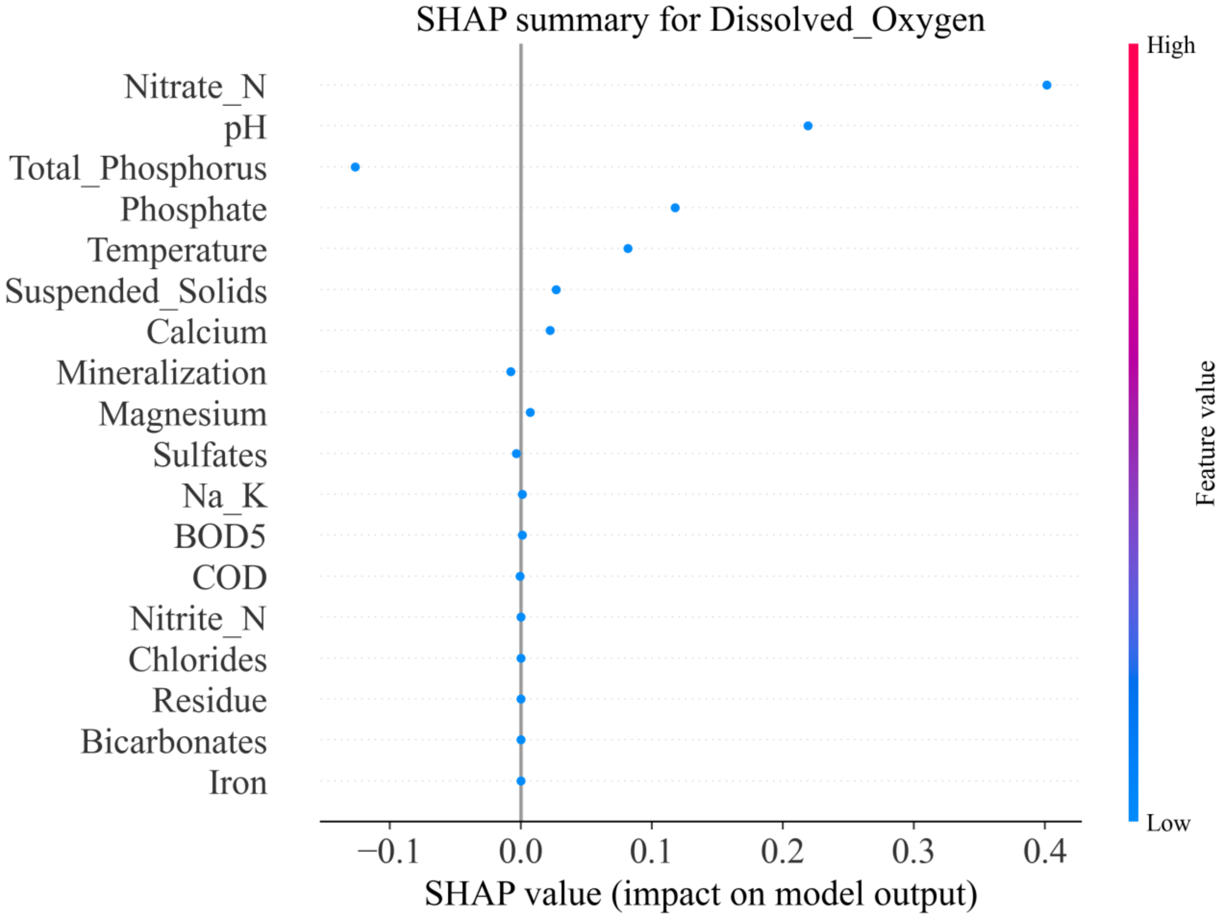

Figure 2.
Results of analysis of the effect of different factors on the prediction of the level of dissolved oxygen in a reservoir.

Figure 3 shows the SHAP summary, which presents the factors influencing water hardness prediction. Suspended solids, chlorides, and salinity are the most significant contributors to hardness. This also confirms that water hardness is primarily determined by mineral salts present in dissolved form, specifically calcium and magnesium. Suspended solids may carry such minerals and contribute to increased hardness. Salinity is directly proportional to dissolved salts and, therefore, naturally contributes to hardness. Chloride is a prominent ion in water and also adds to hardness. pH, temperature, and nitrite are less fundamental parameters affecting chemical equilibrium and salt solubility in water. The hardness of water is thus almost entirely dependent on suspended solids and dissolved salts, emphasizing the importance of water chemistry in forecasting.

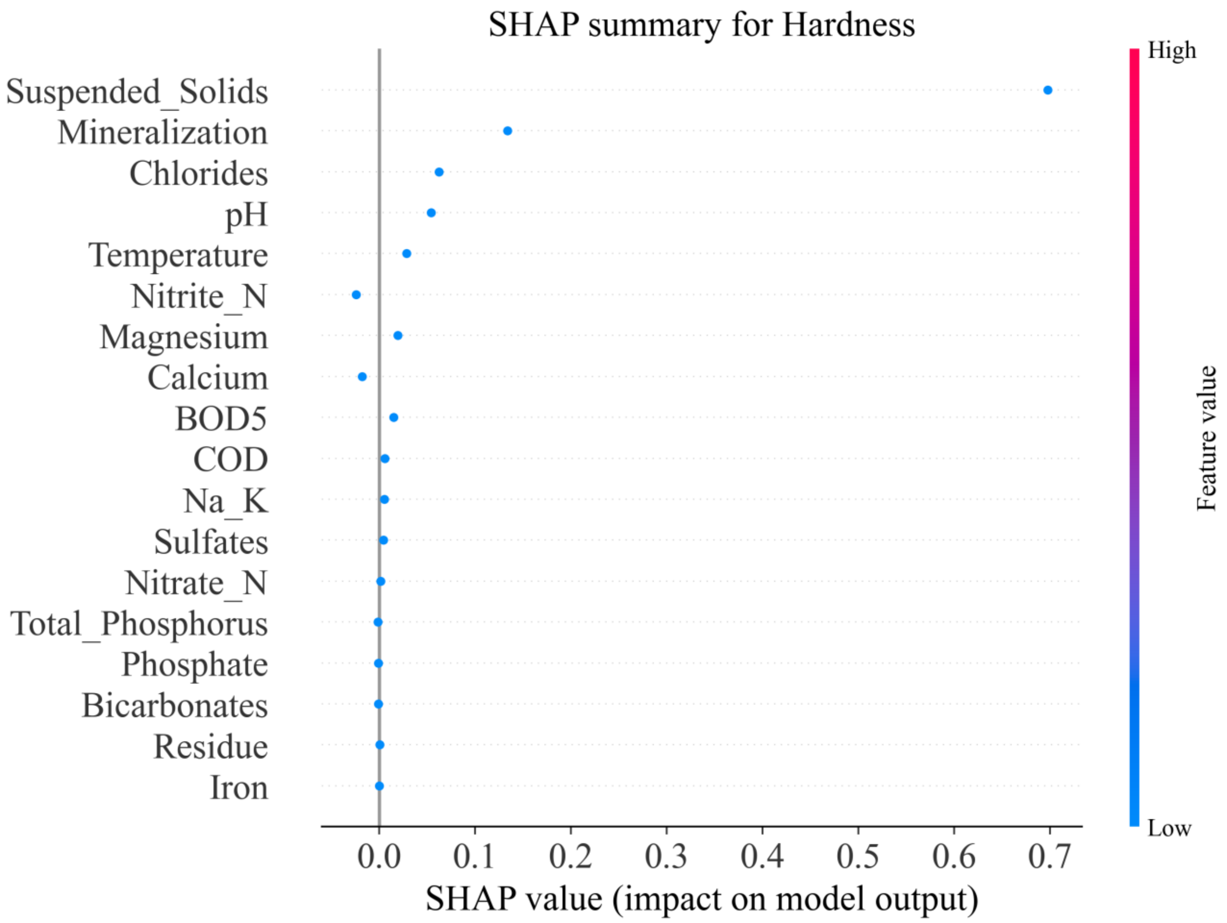

Figure 3.
Factors affecting the prediction of water hardness.

Figure 4 of the SHAP summary illustrates the parameters that have the most significant impact on water clarity. The chart indicates that the most important parameters influencing this parameter are chemical oxygen demand (COD), biochemical oxygen demand (BOD), sulfates, and salinity. Water with a high organic content, as indicated by a high BOD, is not clear because the organic matter settles out as suspended solids, which stimulates biological activity and thus reduces clarity. High levels of sulfate would cause the water to be cloudy due to chemical reactions and the formation of insoluble compounds, which also reduce clarity. COD is linked to organic contaminants in the water, and the greater the COD, the greater the quantity of organic matter that detracts from water clarity. Elevated salinity is typically associated with reduced clarity, as salts bind to inorganic and organic matter particles, thereby increasing turbidity. Calcium and other elements, such as temperature and magnesium, have a lesser effect on clarity but can alter the chemical makeup of water and induce suspended particles that impact vision. For this reason, organic matter, mineral salts, and general pollution will largely account for water transparency, as indicated by evidence in the diagram.

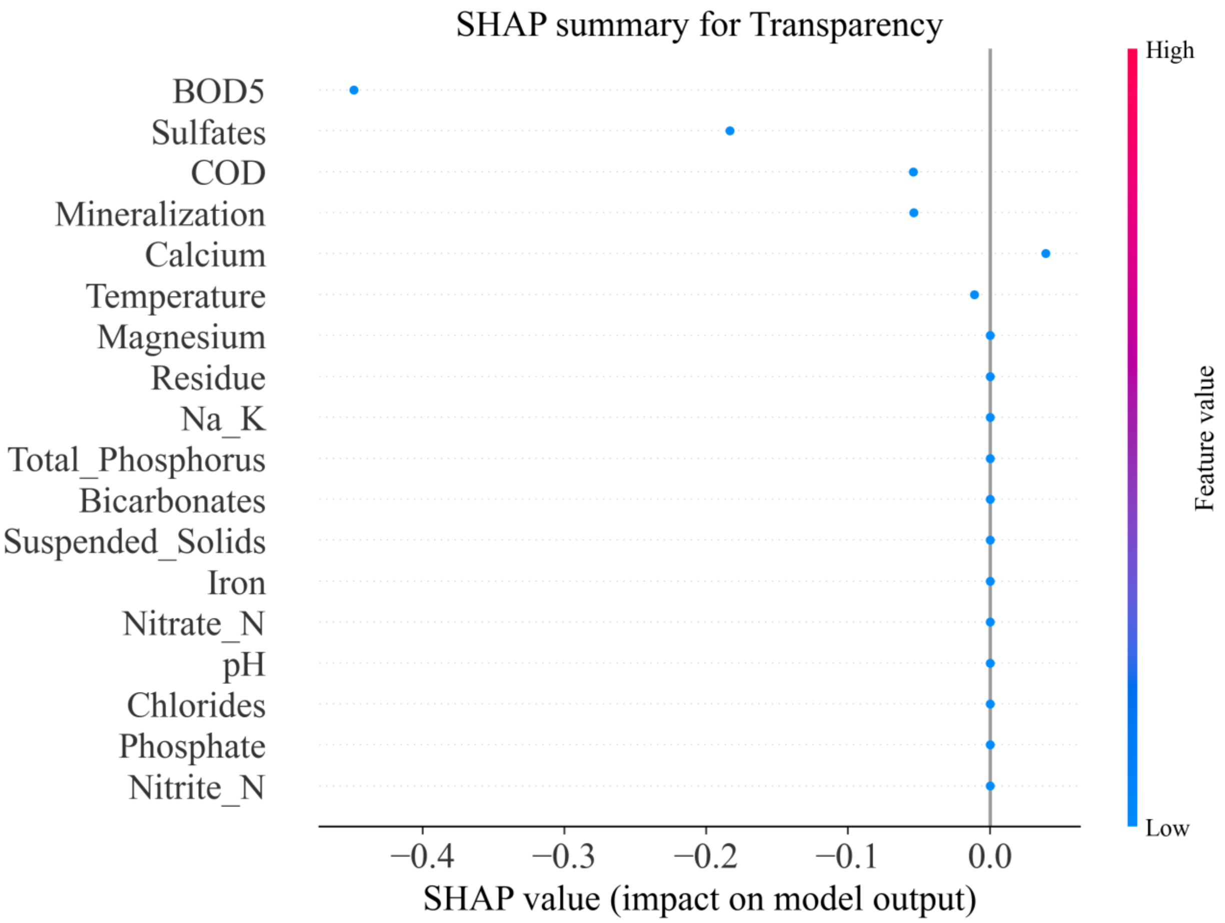

Figure 4.
Factors that have the most significant impact on water transparency.

Figure 5 the SHAP summary explains the factors that have the most significant influence on the Water Quality Index. From the summary, the major influencing parameters are BOD, pH, magnesium, phosphates, temperature, and COD. Among these, BOD is the most vital parameter, as it has the most significant effect on the water quality index, indicating the amount of organic matter in the water that microbes have decomposed. The presence of organic water pollution is evidenced by high BOD levels, which lower water quality. The pH level is second in importance, as it must not be too low or too high; otherwise, it will adversely affect living organisms and alter the chemical reactions within the water body. Both phosphates and magnesium significantly contribute to the chemical composition of water. Excessive magnesium indicates mineral contamination, and the presence of phosphates typically signifies eutrophication, which reduces water quality. Temperature also influences chemical and biological processes, such as oxygen solubility and microbial activity, since an increase in temperature leads to increased metabolic processes, reduced organic decomposition rates, and poorer water quality. COD, or chemical oxygen demand, measures the amount of organic and inorganic impurities in water; high COD levels indicate contamination that degrades water quality. Therefore, the most significant characteristics affecting water's chemical properties, such as pH and mineralization, and indicating chemical and organic contamination, determine its overall quality.

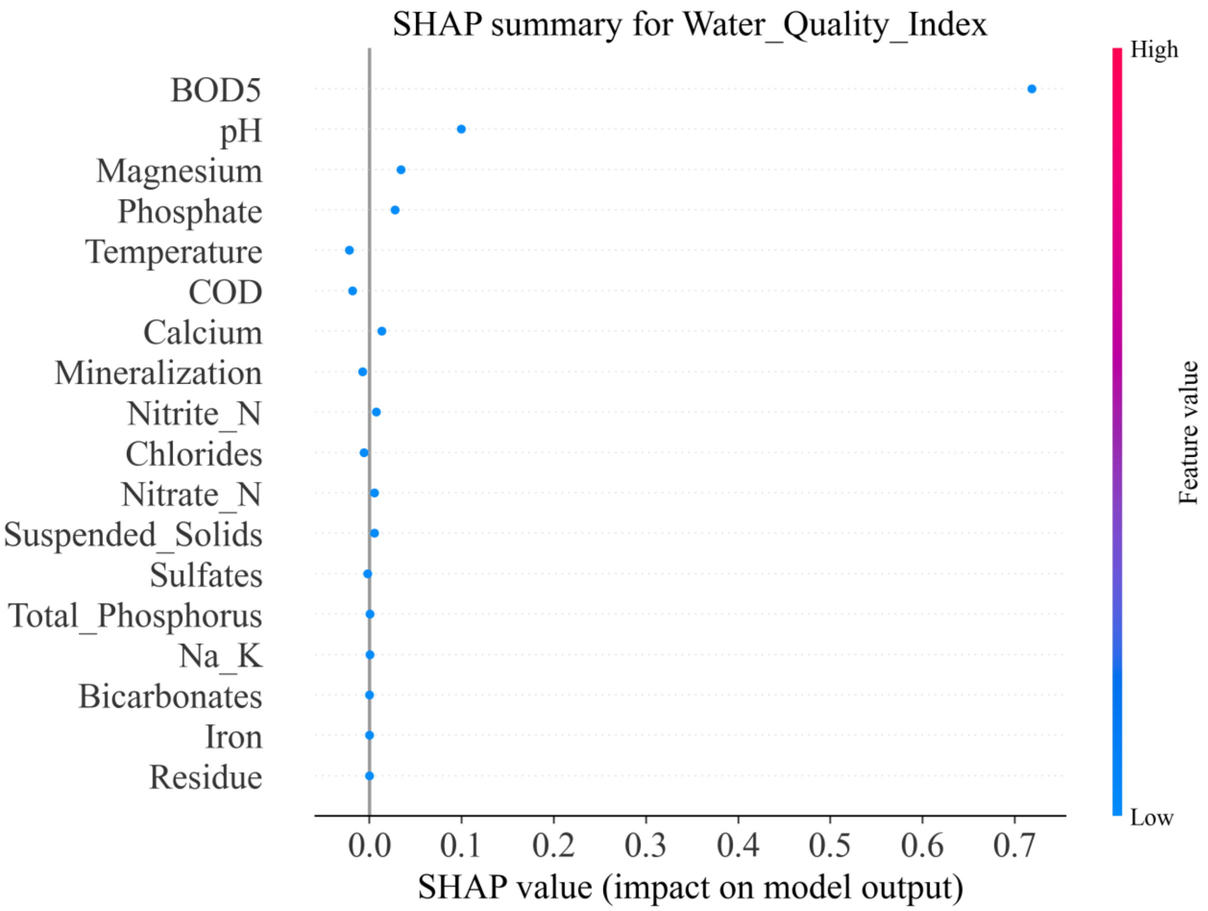

Figure 5.
Factors that have the most significant impact on the water quality index.

Figure 6 of the SHAP summary for the Eutrophication Index illustrates the factors that most significantly impact this indicator. pH is one of the key parameters influencing the eutrophication index, as water acidity can dramatically impact the growth of algae and microorganisms, thereby contributing to eutrophication processes. High or low pH would accelerate or retard such processes. Temperature also affects eutrophication, for it is involved in biological activity and the rate of chemical reactions in water bodies. Increased temperatures, however, allow for faster growth of phytoplankton and promote eutrophication. Sulfates also affect biodegradation and microbial growth, as they can function with other water constituents. Nitrates, as one of the nutrients that stimulate algae growth, enhance phytoplankton growth in large quantities, which is a factor contributing to water body eutrophication. Suspended solids cause water turbidity and decrease light penetration, thereby affecting eutrophication processes. Thus, the chemical determining factors of water composition, i.e., pH, sulfates, nitrates, and temperature, are the principal reasons why the eutrophication index varies and reflects the effect of these factors on the reservoir and biocenosis condition.

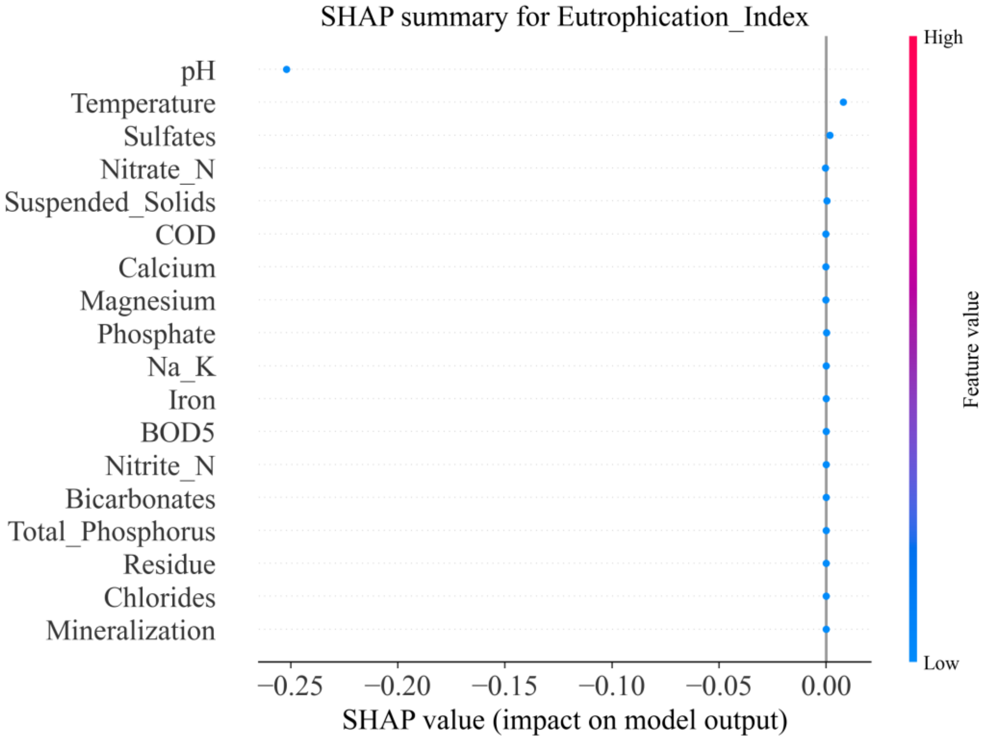

Figure 6.
Factors that have the most significant impact on the eutrophication index.

Figure 7 SHAP summary shows the overall impact of the different factors on the lake water condition. SHAP values represent the extent to which each variable affects the model's prediction; the higher the SHAP value, the greater the variable's impact on the outcome. Suspended solids are a significant variable, as they can result in transparency loss and influence the chemistry, thereby decreasing the quality of the aquatic environment for aquatic organisms. The second high-priority factor is nitrates (Nitrate_N), as they directly contribute to eutrophication through the acceleration of algae growth and the removal of general water quality. Biochemical Oxygen Demand (BOD5) is also a gauge of the organic content of water and a good indicator; a high reading indicates organic water pollution. Phosphates and sulfates both have vital functions in the state of the water. Additionally, sulfates influence chemical reactions in the water, while phosphates stimulate the growth of phytoplankton, which in turn deoxygenates the water. Since the health and processes of aquatic life in an ecosystem are affected by changing values from non-affecting ones, pH is considered. Temperature affects the chemical and biological processes within the water body, such as oxygen dissolution and the rate of development in organisms. Thus, chemical water parameters such as nitrates, phosphates, and sulfates, as well as biological conditions in BOD5, and physical parameters like suspended solids and temperature, play a leading role in determining the health of the water body. These observations highlight the importance of monitoring and controlling key chemical and biological parameters to ensure the health of the water body.

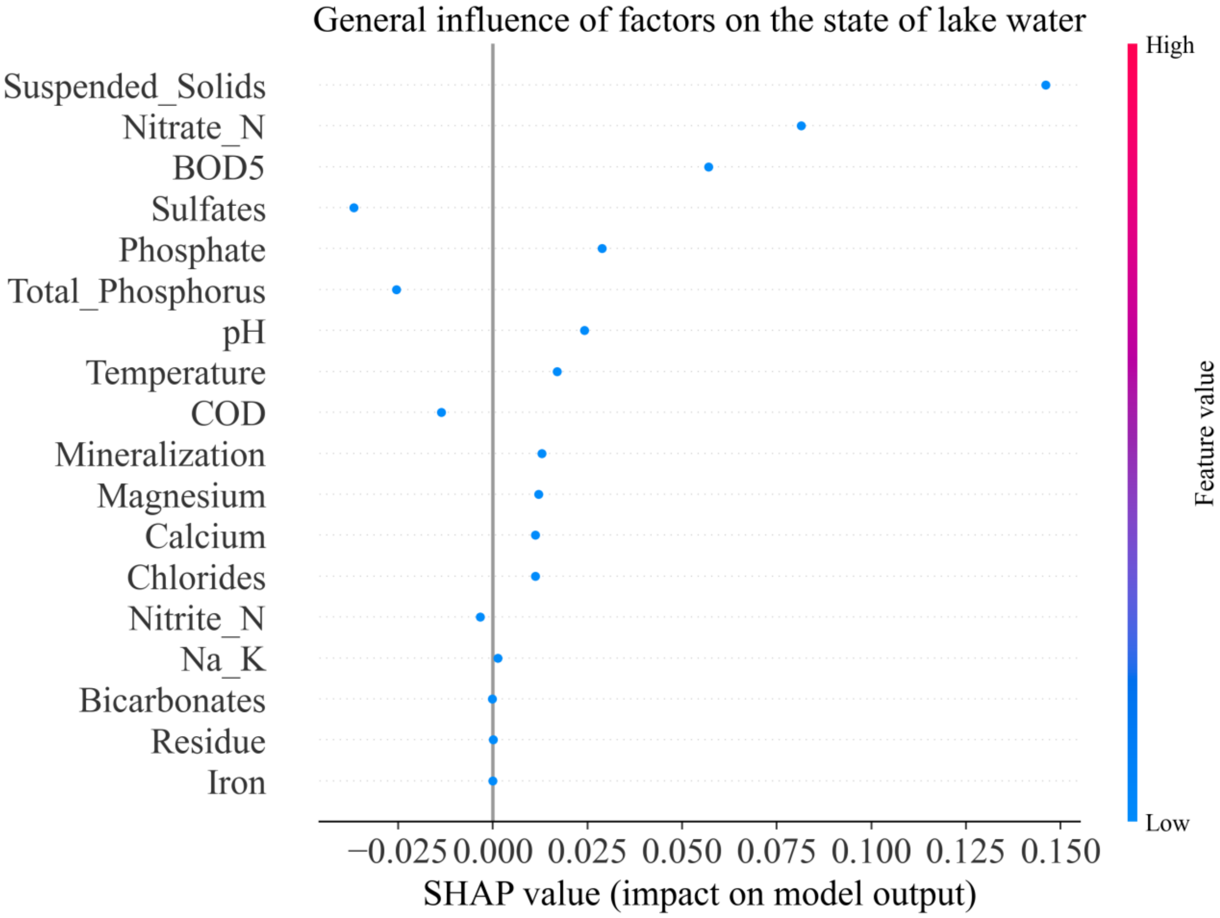

Figure 7.
The general influence of various factors on the state of lake water.

Figure 8 SHAP for phytoplankton illustrates the overall influence of various factors on phytoplankton development in aquatic environments. SHAP values allow us to observe the contribution each feature makes to the model's predictions. Suspended solids have the most significant impact on phytoplankton growth, possibly because phytoplankton utilize suspended particles for nourishment or because they serve as indicators of changing water conditions that promote eutrophication. Nitrates and BOD5 are also influential in a positive manner, as nitrates supply the primary nitrogen source necessary for phytoplankton proliferation, and high BOD5 levels may indicate the presence of organic matter utilized by phytoplankton as a nutritional source. Sulfates and phosphates are significant because they are major nutrients for phytoplankton; sulfates influence water chemistry by altering nutrient availability. Total phosphorus acts as a limiting nutrient for phytoplankton growth. Since pH and temperature can vary, deviations from neutral pH might harm phytoplankton's growth environment. Conversely, temperature directly affects the rate of biological processes. The most important factors for phytoplankton include suspended solids, nitrates, BOD5, sulfates, and phosphates. These chemical and physical indicators serve as early signs of eutrophication and can significantly alter the ecosystem of the water body.

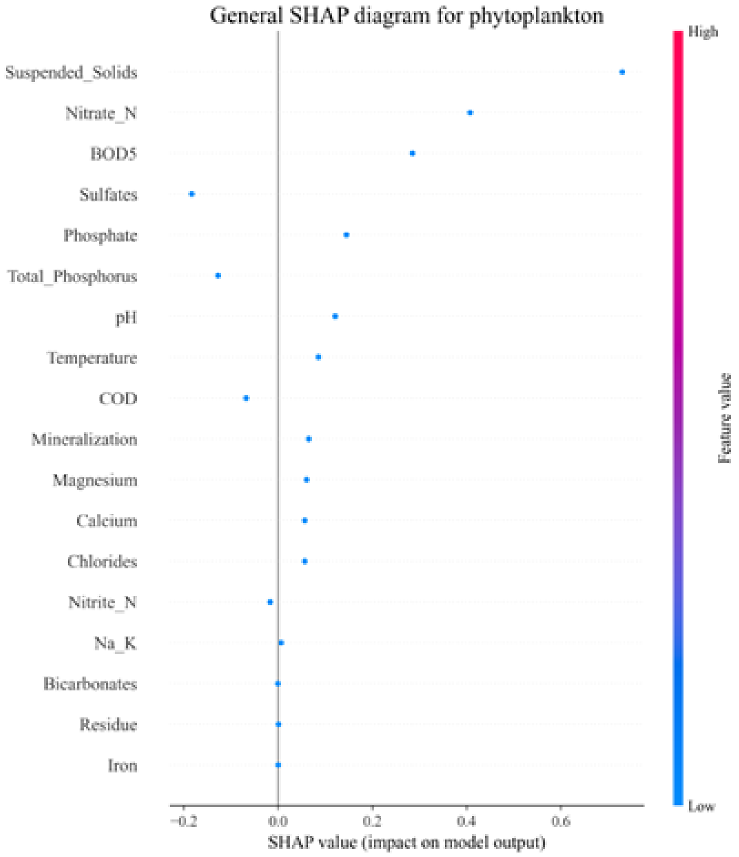

Figure 8.
The general influence of various factors on the development of phytoplankton in a reservoir.

Figure 9 illustrates a SHAP summary for zooplankton, highlighting the impact of various factors on their condition and abundance. SHAP values describe the significance of every feature to model prediction. Suspended solids impact zooplankton the most since suspended solids concentrations alter water clarity and zooplankton habitat conditions. Nitrate provides aquatic life, including zooplankton, with a source of nitrogen they require, as algae and other zooplankton they consume depend on nitrate concentration. pH has the highest impact on the health conditions of zooplankton, as changes in water acidity directly affect their health. Total phosphorus also significantly affects the growth of phytoplankton and the food of zooplankton. Salinity and phosphate concentration have a moderate impact on zooplankton, regulating water chemistry and exerting an indirect influence on ecosystem processes. Other parameters such as temperature, chlorides, magnesium, BOD5, and COD have minimal effects but can nonetheless fluctuate zooplankton's living conditions. Suspended matter, nitrates, and water acidity have the most significant impact on zooplankton and present their overarching influence on the water body's ecosystem balance.

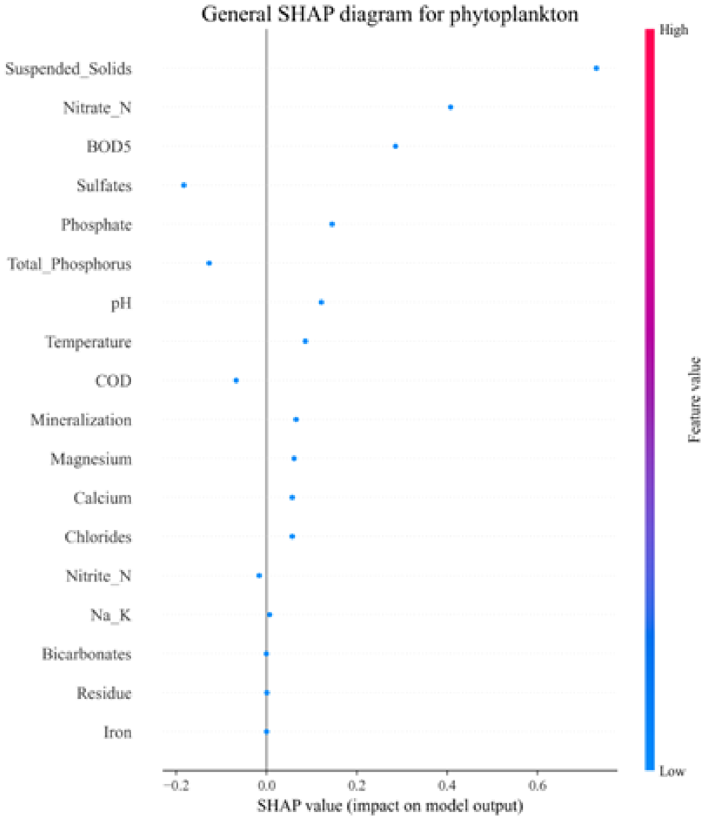

Figure 9.
The Influence of Various Factors on the State and Abundance of Zooplankton.

The SHAP analysis showed that the parameters Nitrate_N, pH, and Total_Phosphorus had the most significant impact on dissolved oxygen, while temperature and Suspended Solids had an adverse effect. Water hardness was determined by mineralization, suspended solids, and chlorides, while clarity decreased with increasing BOD5, COD, and sulfate levels. The most important factors affecting the WQI and EI indices were BOD5, pH, phosphate, temperature, and nitrates. Despite CatBoost's higher training quality scores, the XGBRegressor model demonstrated better generalization ability on the test data, making it preferable for predictive environmental analysis tasks. According to these predictions, water in 2024 will be sufficient for the growth of zooplankton but insufficient for phytoplankton. This may indicate a nutrient imbalance or water clarity issue that restricts photosynthesis. This may indicate some nutrient imbalance or water clarity, which limits photosynthesis. The objective of our study was to develop and analyze a hybrid machine learning model that combines two gradient boosting methods, namely XGBoost and CatBoost. Both approaches have proven to be powerful tools for regression problems, mainly when the data contains categorical features and missing values. However, the study's goal was not only to evaluate them separately but also to investigate how they work in a hybrid system, enabling improved forecast quality through the combined strengths of both models. We compared standard models (XGBoost and CatBoost) and their hybrid versions for forecasting five key lake water quality indices: Dissolved Oxygen (DO), Hardness, Transparency, Water Quality Index (WQI), and Eutrophication Index (EI). An essential aspect of our study was to assess the reservoir biocenosis by forecasting the impact of these factors on indices such as phytoplankton and zooplankton, which allowed us to determine the relationship between the chemical and physical characteristics of water and the state of the biota in aquatic ecosystems [26]. A hybrid model that combines the strengths of both methods was built to improve the accuracy of forecasts. Each method utilizes its advantages in the hybrid system: CatBoost for efficient training on complex data with categorical features and XGBoost for more accurate estimates of test data. The hybrid model demonstrated the best results for all target variables during both training and forecasting stages. The following facts support this: the overall RMSE of the hybrid model was lower than that of the standard models (for example, for dissolved oxygen, the hybrid model showed RMSE=0.362, which is better than both standard models). The AIC and BIC values of the hybrid model were also lower, indicating its accuracy and efficiency in managing complexity. The hybrid method generalized the data more effectively, reducing test data errors while maintaining high training data accuracy. This effectiveness is due to the two methods complementing each other's weaknesses when combined: the flexibility of XGBoost improved performance on unseen data, while the accuracy of CatBoost enhanced the model's learning ability.

One of the primary objectives of the study was to predict the impact of various factors on the biocenosis of a reservoir, particularly on phytoplankton and zooplankton. Of the five indicators analyzed, a SHAP analysis was employed, which enabled the study of the impact of various factors on the ultimate results. Phytoplankton was most strongly affected by variables that include the eutrophication index, which is directly related to levels of nitrogen and phosphorus in water. Nitrates and phosphorus are also firmly integrated into phytoplankton predictions, again confirming their contribution to eutrophication. Zooplankton is influenced by water chemistry variables, such as dissolved oxygen and water hardness, highlighting the importance of water chemistry in sustaining zooplankton populations. Based on the research, the following conclusions can be drawn. The hybrid model performed more successfully than the baseline models at both the training and forecasting phases because it was capable of combining the advantages of both approaches (XGBoost and CatBoost). CatBoost performed better in training, but XGBoost performed better in the prediction stage, which can be attributed to the flexibility and adaptability of the XGBoost method on test data. SHAP analysis revealed that the most significant parameters influencing the prediction of water quality are dissolved oxygen, nitrates, transparency, and total mineralization, thereby verifying the theory of the dependence of chemical and physical water parameters on the state of the biocenosis. It is important to note that the hybrid model has the potential for further application in environmental monitoring tasks, especially when high accuracy of forecasts is required in complex and heterogeneous data conditions. Thus, the results of our study confirm that the application of hybrid machine learning methods can significantly improve the quality of forecasts in tasks related to ecosystem analysis and monitoring and offer new approaches to water resource quality management.

To improve the accuracy of forecasts, a hybrid model was built that combines the strengths of the CatBoost and XGBoost methods. CatBoost provided efficient training on complex data with categorical features, while XGBoost demonstrated high accuracy on test data due to its flexibility and adaptability. The hybrid model outperformed the standard models for all target variables at both the training and forecasting stages. This is confirmed by a decrease in the overall RMSE value (for example, for dissolved oxygen, the RMSE was 0.362, which is better than both standard models), as well as lower AIC and BIC values, indicating an optimal balance between accuracy and model complexity. The union of these two approaches made it possible to compensate for their weaknesses: the incredible learning capacity of CatBoost and the flexibility of XGBoost helped to enhance the model's generalization capacity. One of the research tasks was to predict the influence of environmental factors on the biocenosis of a reservoir, specifically on the populations of phytoplankton and zooplankton. Using SHAP analysis for five target indicators, key influencing factors were identified: the content of phosphorus and nitrates proved decisive in predicting the state of phytoplankton through the eutrophication index. For zooplankton, the most significant factors were dissolved oxygen and water hardness, emphasizing the importance of the chemical composition of the aquatic environment for their vital activity. Accordingly, the research outcomes demonstrate that the hybrid model is emerging as a viable option for environmental monitoring operations that require high accuracy in processing complex and heterogeneous data. The use of hybrid machine learning methods opens up new avenues for effective water quality management and forecasting of ecosystem states. With the acceleration of climate change and the emergence of anthropogenic drivers, predictive models are becoming an inevitable tool for the surveillance and regulation of water bodies. Hybrid models based on CatBoostRegressor and XGBRegressor algorithms were constructed and evaluated in the present research. CatBoostRegressor proved to be a better performer, as indicated by lower RMSE, AIC, and BIC values during model training. The root mean square error (RMSE) analysis, shown in Figure 10 suggests that the CatBoostRegressor model yields more accurate predictions of water quality parameters, which is particularly crucial for evaluating environmentally sensitive indicators such as oxygen levels, transparency, and eutrophication.

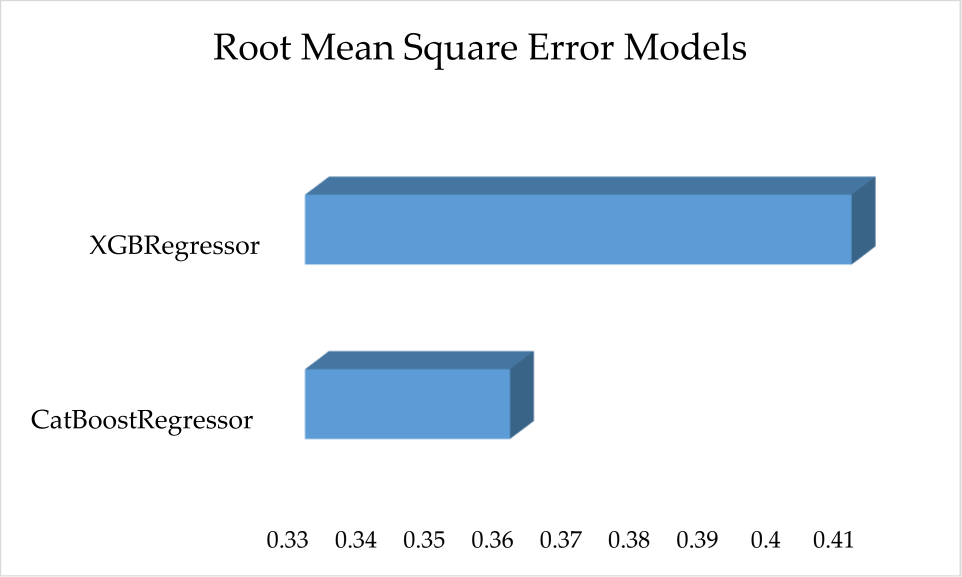

Figure 10.
Root Mean Square Error of CatBoostRegressor and XGBRegressor Models.

Figure 11 presents a comparative assessment of the quality of two machine learning models, CatBoostRegressor and XGBRegressor, by the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The two metrics enable us to assess the trade-off between the model's accuracy and complexity: the lower the values, the better the model quality for the number of parameters.

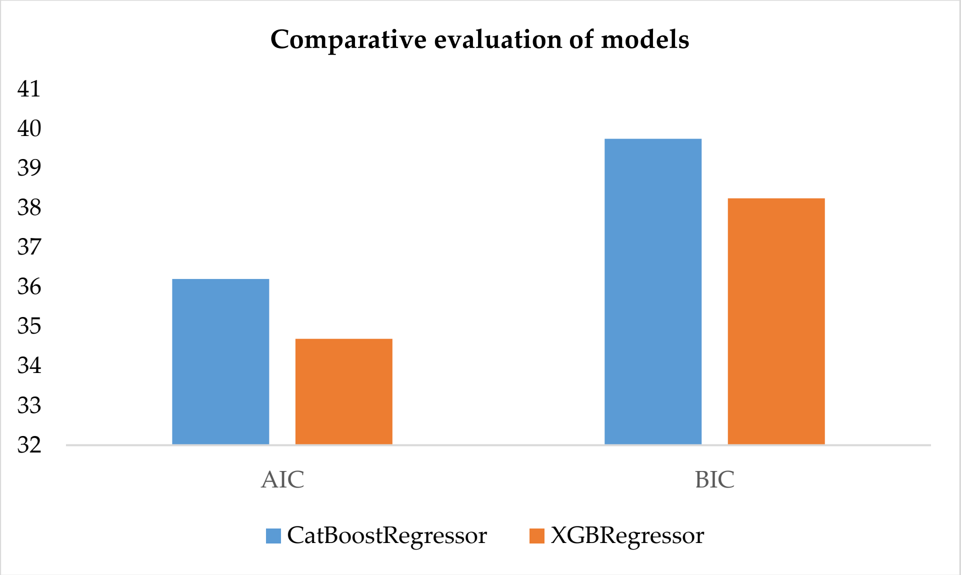

Figure 11.
Comparative evaluation of CatBoostRegressor and XGBRegressor models.

According to Figure 12:

- According to the AIC criterion, the XGBRegressor model showed a lower value compared to the CatBoostRegressor (approximately 34.7 vs. 36.2), indicating better information efficiency with minimal overfitting.
- According to the BIC criterion, the advantage of XGBRegressor is also observed (approximately 38.2 vs. 39.8), which further confirms its preference as a model with fewer parameters and better generalization ability.

Thus, the visualized data show that the XGBRegressor has higher statistical validity than the CatBoostRegressor and can be recommended as the primary model for predicting water quality parameters in this study.

After training the hybrid model, we obtained the following forecasts for 2024 for key water quality indicators and the eutrophication index. The choice favored the XGBoost model, which demonstrated the best forecast accuracy on the test data. Figure 12 shows the predicted values of key water quality indicators for the reservoir in 2024. A scale from -5 to 5 is used to normalize the data, allowing for a clearer display of each factor's influence within a specified range.

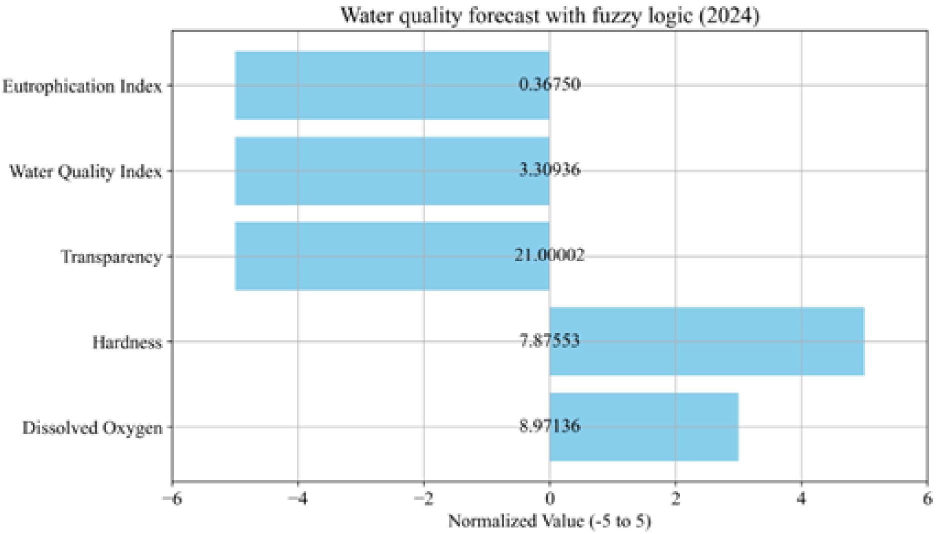

Figure 12.
Projected values of key water quality indicators for the reservoir for 2024.

The dissolved oxygen concentration is computed as 8.97136 mg/L, which approximates the value of 9.0 mg/L and indicates stability in water body conditions related to oxygen. Water hardness is calculated as 7.87553 mmol/L, a key indicator of the health of the ecosystem, more precisely, the mineralization and nutritional content of the water for organisms. The water transparency is computed as 21.00022 cm, indicating high openness to minimal suspended solids and organic matter. The water quality, as estimated by a water quality index, is 3.30936, indicating good water quality that meets cleanliness criteria and is sensitive to the ecosystem. An eutrophication index of 0.3675 indicates little eutrophication, suggesting a healthy environment with minimal threat from algal blooms and excessive algal biomass. The prediction is for consistent conditions in the reservoir, with minimal variations in water quality parameters, particularly in the most critical parameters, including dissolved oxygen levels and clarity. The XGBoost model could understand the interaction between temperature, dissolved solids, and water chemistry variables. Therefore, it was bound to provide a better prediction with less error compared to other models. The XGBoost model performed well in predicting water quality based on whether the predicted values were close to the actual data. Another plot displays the expected values for the effect of water quality on biota, specifically phytoplankton and zooplankton, in 2024. These findings show how water parameters influence the key components of the aquatic ecosystem, which is crucial in assessing the ecological status of the reservoir. As can be observed in Figure 13 the impact of -5.0 in the case of phytoplankton represents a considerable reduction in its population. This is because there is a shift in the quantity of nutrients or water clarity, thus impacting the resources of aquatic plants and phytoplankton, and ultimately affecting photosynthesis. On the other hand, the impact of +4.0 on zooplankton represents a boost in its population. A rise in the density of zooplankton can occur due to an enhancement in water quality, primarily through increased hardness and dissolved oxygen content, thereby making the environment more hospitable to life.

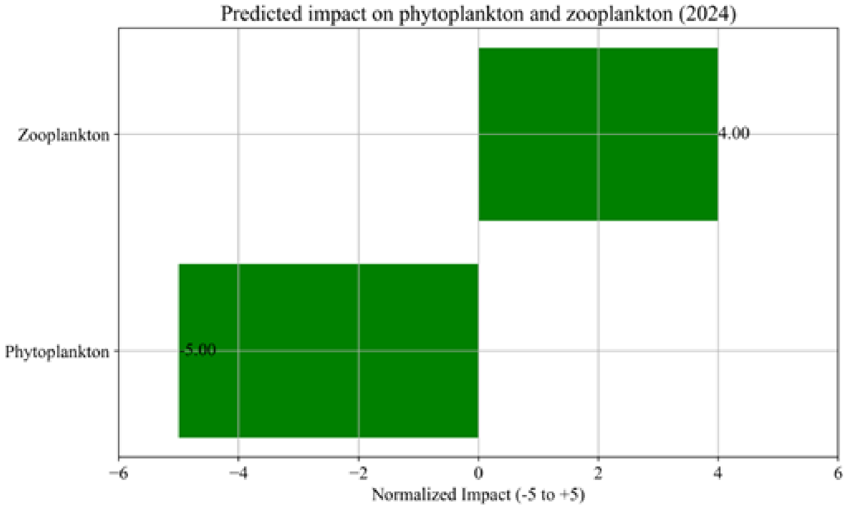

Figure 13.
Predicted impact on phytoplankton.

It is based on the complex interactions between the physical and chemical characteristics of the water body, which impact the abiotic factors that control the growth of phytoplankton and zooplankton. The fluctuating availability of nutrients, such as phosphates and nitrates, controls the primary producers, namely phytoplankton. They can be induced to grow in numbers by high levels of eutrophication but then promptly decline as soon as the conditions become unfavorable, such as reduced sunlight or transparency. In contrast, zooplankton, as consumers of phytoplankton, can reflect a growth in population towards more favorable environmental conditions, i.e., the highest oxygen levels and other favorable chemical conditions. This is a prerequisite for a sustainable ecosystem where zooplankton can regulate phytoplankton biomass. The projections indicate reduced phytoplankton populations, which could be an indicator of low eutrophication and favorable zooplankton conditions. Various models, including the statistical calibration model, have been calibrated and projected to run with high performance. These results highlight the importance of controlling and regulating the water body biosphere to prevent undesirable changes in aquatic ecosystems. The suggested hybrid model was found to be highly effective in predicting environmental parameters of marine ecosystems. A comparative analysis, conducted by applying RMSE, AIC, and BIC values, confirmed that the XGBRegressor model provides the best balance between prediction quality and resistance to overfitting, despite CatBoost showing higher accuracy on the training data. SHAP analysis not only explained model predictions but also identified the most significant factors affecting water parameters and the state of the biocenosis — nitrates, phosphates, BOD5, suspended matter, and temperature were the most important factors. The 2024 projection indicated a healthy state of the water bodies, characterized by sufficient dissolved oxygen, transparency, and low eutrophication values, as well as an optimal phytoplankton-zooplankton relationship. This supports the applicability of the model designed for decision-making and sustainable monitoring systems in managing water resource quality, particularly under climate change and human loading regimes.

5. Limitations and Future Work

Despite the promising results achieved in this study, several limitations should be acknowledged. First, the model's predictive accuracy is inherently dependent on the quality and quantity of available sensor data. In regions with limited monitoring infrastructure, the model's performance may decrease. Second, although SHAP analysis improves interpretability, it cannot fully capture complex causal relationships in aquatic ecosystems, which might require complementary ecological modeling. Additionally, this study focused on data from lakes in Northern Kazakhstan, which may limit the generalizability of the findings to other types of water bodies or regions with different climatic and hydrological conditions.

For future work, the hybrid framework could be enhanced by integrating satellite remote sensing data to improve spatial coverage and enable broader ecosystem assessments. Moreover, developing real-time early warning systems based on the proposed model can support proactive decision-making for water resource managers. Incorporating advanced uncertainty quantification methods and testing the model across different geographic locations and water body types will further validate its robustness. Finally, coupling machine learning predictions with mechanistic ecological models could provide a more comprehensive understanding of biotic and abiotic interactions in aquatic environments.

6. Conclusion

This study presents an intelligent forecasting approach for assessing aquatic ecosystem dynamics by integrating sensor-based environmental monitoring with machine learning techniques. The proposed hybrid model, which combines CatBoost and XGBoost regressors, enables the accurate prediction of key water quality indicators, including dissolved oxygen, hardness, transparency, the water quality index (WQI), and the eutrophication index (EI). To enhance model interpretability, SHAP analysis was applied to identify the most influential environmental features, including nitrate nitrogen, total phosphorus, pH, and suspended solids. The forecasting pipeline includes systematic data preprocessing, gap filling, logarithmic transformations, feature selection using LASSO regression, and multi-task model training. Evaluation metrics such as Root Mean Square Error (RMSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) confirmed the robustness of the hybrid approach compared to individual models. In particular, the XGBoost-based model showed superior generalization on test data, while CatBoost demonstrated strong performance during training. SHAP-based interpretation provided insights into the ecological drivers of water quality, revealing how physical and chemical parameters influence biotic components such as phytoplankton and zooplankton. This interpretable modeling approach supports ecological decision-making and provides a foundation for advanced environmental diagnostics. This study presents an integrated forecasting framework that combines sensor-based ecological data collection with interpretable machine-learning techniques to monitor and predict the dynamics of aquatic ecosystems. The hybrid CatBoost–XGBoost model demonstrates high accuracy and robustness, while SHAP analysis reveals critical environmental drivers, such as nitrate levels, total phosphorus, and suspended solids. The system's modular design allows integration into existing water monitoring infrastructure, providing a practical tool for environmental agencies and water resource managers. Future work will focus on expanding the system to include real-time alert mechanisms and satellite sensor fusion, thereby enhancing spatial resolution and accuracy.

References

[1] F. Yang and X. Xiong, "Carbon emissions, wastewater treatment and aquatic ecosystems," *Science of The Total Environment,* vol. 921, p. 171138, 2024.

[2] E. M. Albou, M. Abdellaoui, A. Abdaoui, and A. Ait Boughrous, "Agricultural practices and their impact on aquatic ecosystems–a mini-review," *Ecological Engineering & Environmental Technology,* vol. 25, 2024.

[3] M. Anas, A. Hayat, A. Falak, Q. Aslam, J. Fatima, and M. H. Saleem, "Micro/nanoplastics in aquatic ecosystems: An integrated review of occurrence, toxicological implications, case studies, methodologies, and future recommendations," *BioNanoScience,* vol. 14, no. 3, pp. 3440-3454, 2024.

[4] M. Banaee, A. Zeidi, N. Mikušková, and C. Faggio, "Assessing metal toxicity on crustaceans in aquatic ecosystems: A comprehensive review," *Biological Trace Element Research,* vol. 202, no. 12, pp. 5743-5761, 2024.

[5] G. Liu *et al.*, "Forecast urban ecosystem services to track climate change: Combining machine learning and emergy spatial analysis," *Urban Climate,* vol. 55, p. 101910, 2024. https://doi.org/10.1016/j.uclim.2024.101910

[6] Y. Liu, X. Huang, and Y. Liu, "Detection of long-term land use and ecosystem services dynamics in the Loess Hilly-Gully region based on artificial intelligence and multiple models," *Journal of Cleaner Production,* vol. 447, p. 141560, 2024. https://doi.org/10.1016/j.jclepro.2024.141560

[7] B. Almeida, J. David, F. S. Campos, and P. Cabral, "Satellite-based Machine Learning modelling of Ecosystem Services indicators: A review and meta-analysis," *Applied Geography,* vol. 165, p. 103249, 2024. https://doi.org/10.1016/j.apgeog.2024.103249

[8] S. R. Morshed, M. Esraz-Ul-Zannat, M. A. Fattah, and M. Saroar, "Assessment of the future environmental carrying capacity using machine learning algorithms," *Ecological Indicators,* vol. 158, p. 111444, 2024.

[9] A. E. Alprol, A. T. Mansour, M. E. Ibrahim, and M. Ashour, "Artificial intelligence technologies revolutionizing wastewater treatment: current trends and future prospective," *Water,* vol. 16, no. 2, p. 314, 2024. https://doi.org/10.3390/w16020314

[10] S. E. Bibri, J. Huang, and J. Krogstie, "Artificial intelligence of things for synergizing smarter eco-city brain, metabolism, and platform: Pioneering data-driven environmental governance," *Sustainable Cities and Society,* vol. 108, p. 105516, 2024. https://doi.org/10.1016/j.scs.2024.105516

[11] M. R. Anwar and L. D. Sakti, "Integrating artificial intelligence and environmental science for sustainable urban planning," *IAIC Transactions on Sustainable Digital Innovation,* vol. 5, no. 2, pp. 179-191, 2024.

[12] R. Naimaee, A. Kiani, S. Jarahizadeh, S. B. Haji Seyed Asadollah, P. Melgarejo, and A. Jodar-Abellan, "Long-term water quality monitoring: Using satellite images for temporal and spatial monitoring of thermal pollution in water resources," *Sustainability,* vol. 16, no. 2, p. 646. https://doi.org/10.3390/su16020646

[13] E. T. Wasehun, L. Hashemi Beni, and C. A. Di Vittorio, "UAV and satellite remote sensing for inland water quality assessments: A literature review," *Environmental Monitoring and Assessment,* vol. 196, no. 3, p. 277, 2024.

[14] P. Jayaraman, K. K. Nagarajan, P. Partheeban, and V. Krishnamurthy, "Critical review on water quality analysis using IoT and machine learning models," *International Journal of Information Management Data Insights,* vol. 4, no. 1, p. 100210, 2024.

[15] I. Essamlali, H. Nhaila, and M. El Khaili, "Advances in machine learning and IoT for water quality monitoring: A comprehensive review," *Heliyon,* vol. 10, no. 6, 2024.

[16] R. Moncelon *et al.*, "Drivers for primary producers' dynamics: New insights on annual benthos pelagos monitoring in anthropised freshwater marshes (Charente-Maritime, France)," *Water Research,* vol. 221, p. 118718, 2022. https://doi.org/10.1016/j.watres.2022.118718

[17] S. Kumar and L. K. Sharma, "Assessment of water and carbon use efficiency in the SAARC region for ecological resilience under changing climate," *Journal of Environmental Management,* vol. 326, p. 116812, 2023. https://doi.org/10.1016/j.jenvman.2022.116812

[18] J. Xu *et al.*, "Microcystin-leucine-arginine affects brain gene expression programs and behaviors of offspring through paternal epigenetic information," *Science of The Total Environment,* vol. 857, p. 159032, 2023. https://doi.org/10.1016/j.scitotenv.2022.159032

[19] X. Chen, D. Zou, H. Xie, G. Cheng, and C. Liu, "Two decades of artificial intelligence in education: Contributors, collaborations, research topics, challenges, and future directions," *Educational Technology & Society,* vol. 25, no. 1, pp. 28–47, 2022.

[20] H. Molaee Jafrodi, M. Gholami Parashkoohi, H. Afshari, and D. Mohammad Zamani, "Comparative life cycle cost-energy and cumulative exergy demand of paddy production under different cultivation scenarios: A case study," *Ecological Indicators,* vol. 144, p. 109507, 2022. https://doi.org/10.1016/j.ecolind.2022.109507

[21] J. Bastos-Arrieta and C. Palet, "Sustainable processes for the removal of heavy metals from aquatic systems," *Water,* vol. 15, no. 4, p. 761, 2023. https://doi.org/10.3390/w15040761

[22] J. Raimundo, J. Silva, and M. Oliveira, "Improving interpretability in machine learning-based water quality prediction," *Environmental Modelling & Software,* vol. 165, p. 105585, 2023.

[23] H. Wang, J. Xin, X. Zheng, Y. Fang, M. Zhao, and T. Zheng, "Effect of biofilms on the clogging mechanisms of suspended particles in porous media during artificial recharge," *Journal of Hydrology,* vol. 619, p. 129342, 2023. https://doi.org/10.1016/j.jhydrol.2023.129342

[24] E. Sinakou, V. Donche, and P. Van Petegem, "Action-orientation in education for sustainable development: Teachers’ interests and instructional practices," *Journal of Cleaner Production,* vol. 370, p. 133469, 2022. https://doi.org/10.1016/j.jclepro.2022.133469

[25] M. Badpa *et al.*, "Outdoor light at night and children's body mass: A cross-sectional analysis in the fr1da study," *Environmental Research,* vol. 232, p. 116325, 2023. https://doi.org/10.1016/j.envres.2023.116325

[26] A. Nageswari, U. Jyothi, G. Divya, T. Ammannamma, and V. Usha, "Water quality classification using XGBoost method," presented at the IEEE 6th International Conference on Cybernetics, Cognitive and Machine Learning Applications (ICCCMLA), pp. 302–306, 2024.

[27] G. Nagarajan, N. K. Reddy, Y. V. Kumar, A. R. Reddy, and T. Chandu, "Water quality classification using XG boost," presented at the International Conference on Trends in Quantum Computing and Emerging Business Technologies, pp. 1–3, 2024.

[28] J. Liu, Q. Chu, W. Yuan, D. Zhang, and W. Yue, "WQI improvement based on XG-BOOST algorithm and exploration of optimal indicator set," *Sustainability,* vol. 16, no. 24, p. 10991, 2024.

[29] F. U. Shah, A. U. Khan, A. W. Khan, B. Ullah, M. R. Khan, and I. Javed, "Comparative analysis of ensemble learning algorithms in water quality prediction," *Journal of Hydroinformatics,* vol. 26, no. 12, pp. 3041-3059, 2024.

[30] S. Liu *et al.*, "Probabilistic quantile multiple fourier feature network for lake temperature forecasting: Incorporating pinball loss for uncertainty estimation," *Earth Science Informatics,* vol. 17, no. 6, pp. 5135-5148, 2024.