Bozok J Eng Arch, 2025; 4(1) 56-65
DOI: 10.70700/bjea.1590677

Bozok
Journal of Engineering and Architecture

e-ISSN: 3023-4298
Research Article

Predicting dissolved oxygen levels in aquatic ecosystems using machine learning models

Çağrı ARISOY[1]*, Enes Eren SÜZGEN[1], Gülbahar YILDIZ[1]

[1] Yozgat Bozok Üniversitesi, Mühendislik-Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Yozgat, Türkiye

ARTICLE INFORMATION

Article History:

Received
25.11.2024
Accepted
07.04.2025
Published
31.06.2025

Keywords:

Machine Learning Algorithms
Aquatic Ecosystems
Brisbane River
Water Quality

ABSTRACT

This research investigates the performance of several machine learning algorithms in forecasting dissolved oxygen (DO) levels in the Brisbane River, utilizing physicochemical parameters alongside water flow data. We examined algorithms such as Linear Regression, Support Vector Regression, Random Forest, Gradient Boosting, XGBoost, and K-Nearest Neighbors, employing evaluation metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and Mean Absolute Percentage Error (MAPE). Among the models, ensemble techniques, particularly Random Forest and XGBoost, exhibited enhanced predictive accuracy and robustness in identifying intricate, non-linear relationships. Analysis revealed that key variables, including pH, salinity, and specific conductance, were significant predictors, a finding corroborated by the correlation matrix. This study underscores the promise of machine learning, particularly ensemble approaches, in improving water quality monitoring and management, providing valuable insights for ecological sustainability and informed policymaking.

Makine öğrenimi modelleri kullanarak su ekosistemlerindeki çözünmüş oksijen seviyelerinin tahmini

MAKALE BİLGİSİ

Makale Tarihleri:

Geliş tarihi
25.11.2024
Kabul tarihi
07.04.2025
Yayın tarihi
31.06.2025

Anahtar Kelimeler:

Makine Öğrenimi Algoritmaları
Su Ekosistemleri
Brisbane Nehri
Su Kalitesi

ÖZET

Bu araştırma, Brisbane Nehri'ndeki çözünmüş oksijen (DO) seviyelerini tahmin etmek için çeşitli makine öğrenimi algoritmalarının performansını, fizikokimyasal parametreler ve su akış verileri ile incelemektedir. Lineer Regresyon, Destek Vektör Regresyonu, Random Forest, Gradient Boosting, XGBoost ve K-En Yakın Komşu algoritmaları değerlendirilmiş; Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and Mean Absolute Percentage Error (MAPE) gibi değerlendirme metrikleri kullanılmıştır. Modeller arasında, özellikle Random Forest ve XGBoost gibi topluluk (ensemble) teknikleri, karmaşık ve doğrusal olmayan ilişkileri belirlemede üstün tahmin doğruluğu ve dayanıklılık sergilemiştir. Analizler, pH, tuzluluk ve spesifik iletkenlik gibi anahtar değişkenlerin önemli öngörücüler olduğunu ve bu bulguların korelasyon matrisi ile desteklendiğini ortaya koymuştur. Bu çalışma, makine öğreniminin, özellikle topluluk yaklaşımlarının, su kalitesinin izlenmesi ve yönetiminin iyileştirilmesindeki potansiyelini vurgulamakta ve ekolojik sürdürülebilirlik ile bilinçli politika oluşturma için değerli içgörüler sunmaktadır.

1. INTRODUCTION

Dissolved oxygen (DO) is a critical parameter for the health of aquatic ecosystems. Predicting and monitoring DO levels is particularly important for maintaining the ecological balance of water bodies such as rivers, lakes, and streams [1,2]. In urban and

ORCID ID: Çağrı Arısoy: 0009-0005-0296-537X; Enes Eren Süzgen: 0009-0001-5442-930X; Gülbahar Yıldız: 0009-0004-3951-599X

***Corresponding author(s):** Çağrı Arısoy
Tel:+90 539 4469509
Fax: +90 354 2421005
E-mail: cagri.arisoy@yobu.edu.tr

To cite this article: Ç. Arısoy, E.E. Süzgen, G. Yıldız, "Predicting dissolved oxygen levels in aquatic ecosystems using machine learning models", Bozok Journal of Engineering and Architecture, vol. 4, no, 1, pp. 56-65, June 2025.

industrially influenced water sources like the Brisbane River, accurately forecasting DO levels is essential for monitoring water quality and managing environmental risks [3]. A decline in DO levels in aquatic ecosystems can have fatal consequences for marine life and is considered a significant indicator of environmental degradation [4].

Today, traditional methods for monitoring and predicting water quality are increasingly being replaced by sophisticated, data-driven approaches. Environmental factors such as rainfall, sunlight, and temperature significantly affect the oxygen-holding capacity of water bodies, complicating the prediction of DO levels [5]. The dynamic nature of these environmental impacts has necessitated using machine learning (ML) techniques in situations where traditional predictive models fall short [3]. These limitations have driven interest in data-driven approaches, particularly ML and deep learning (DL) techniques, for improved accuracy in DO prediction. Accurate DO forecasting not only aids in safeguarding water resources but also contributes to the early identification of environmental risks. Research in literature highlights the successful application of ML and DL techniques for predicting DO levels in various river systems. From the Mississippi River to rivers in China, studies demonstrate the robust accuracy and efficiency of these techniques.

In this study, ML models including Linear Regression (LR), Support Vector Regression (SVR), Gradient Boosting, XGBoost, K-Nearest Neighbors (KNN), and Random Forest Regressor were employed to predict DO levels in the Brisbane River. The models were evaluated using performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R^2. Dimensionality reduction was performed using Principal Component Analysis (PCA), and predicted DO values were compared with actual values through visualizations. A multivariate series analysis was also conducted using the Long Short-Term Memory (LSTM) model, incorporating water quality variables such as temperature and pH.

Many studies in the literature use only short-term data or a limited number of physico-chemical variables. This leads to loss of accuracy, especially in ecosystems where seasonal variations are intense and parameters such as current direction and salinity are critical. This study aims to contribute to overcome this deficiency by addressing the multidimensional data collected from the Brisbane River between [04.08.2023 - 27.06.2024].

This study aims to contribute to more accurate predictions of DO levels in the Brisbane River, offering new perspectives for water quality management and ecological sustainability. Compared to existing studies, this research provides a comprehensive analysis of the performance of various models and identifies the most effective one. Moreover, the study addresses gaps such as the challenges current models face in long-term predictions and the inability to account for the complexity of environmental variables fully. By improving the reliability of DO predictions in the Brisbane River, this research seeks to bridge these gaps. The findings of this study are expected to enhance river management and conservation efforts, enabling stakeholders to make more informed decisions regarding the protection of aquatic ecosystems. Reliable DO prediction models can proactively identify potential oxygen deficits or surpluses, thereby preventing environmental degradation and supporting sustainable water management practices.

1.1 Literature Review

Dissolved oxygen (DO) prediction has been the subject of extensive research, with significant advancements in machine learning (ML) and deep learning (DL) techniques enhancing the accuracy and robustness of forecasting models. In this section, we review key studies to highlight the evolution of methodologies and their application to diverse aquatic ecosystems.

First, Granata et al. [1] developed two models, AR-RBF (Additive Regression of Radial Basis Function) and MLP-RF (Multilayer Perceptron with Random Forest), using data from the Mississippi River [6] to forecast short- and medium-term DO levels. The models were evaluated under two scenarios: (A) incorporating water temperature and previous DO levels, and (B) using only previous DO values. The AR-RBF model outperformed MLP-RF, achieving high short-term accuracy (RMSE: 0.28 mg/L, MAPE: 2.5%) and acceptable medium-term results (RMSE: 0.93 mg/L, MAPE: 8.2%), highlighting the importance of environmental variables in improving predictions. Wang et al. [2] employed hybrid DL models such as IVMD-MAGRU and CEEMDAN-AGRU to predict DO using data from Chinese rivers**.** The IVMD-MAGRU model achieved exceptional accuracy (RMSE: 0.097, NSE: 0.925), demonstrating the effectiveness of hybrid approaches in handling time series data. Wei Liu et al [7] developed a Support Vector Regression (SVR)-based model (MIC-SVR) with Maximum Information Coefficient (MIC) variable selection technique to predict dissolved oxygen (DO) levels and identify influencing factors using the Tanjiang River dataset. The data set includes water quality and meteorological parameters and consists of 1740 data samples. The MIC method showed that total phosphorus (TP), pH, electrical conductivity (EC), water temperature (WT) and chemical oxygen demand (CODMn) were the most influential factors on DO changes. The MIC-SVR model provided higher prediction accuracy by reducing the RMSE by 4.46% and increasing the NSE

by 45.85% compared to the conventional SVR model. This model is recommended as a powerful decision support tool for water quality monitoring and management. Similarly, Shadkani et al. [8] developed hybrid models integrating TPAFFNN and LSTM, achieving notable improvements in DO prediction accuracy for the Illinois and Des Plaines Rivers (RMSE: 0.241 mg/L and 0.450 mg/L, respectively). Qiulin Li et al [9]. developed an explainable machine learning model (XAI) with Bayesian optimization (BO) for the prediction of dissolved oxygen (DO) levels using a Mississippi River dataset. The data set includes parameters such as temperature (T), river flow (Q), water level (G), pH, turbidity (Tu) and conductivity (Sc) collected at the Baton Rouge station. Three different models were compared in the study: Support Vector Regression (SVR), Regression Tree (RT) and Boosting Ensemble. The BO-SVR model performed the best (with $R^2 = 0.97$, RMSE = 0.395 mg/L and MAE = 0.303 mg/L). The SHAP analysis revealed that temperature, flow rate and water level have the greatest influence on DO levels. The model also proved to be reliable for long-term forecasts, with an accuracy level of $R^2 > 0.75$ for forecasts up to 30 days. Lim et al. [3] proposed an LSTM-based model for hourly and daily DO forecasting in the Oncheon Stream watershed [10], achieving an average 25.6% increase in prediction accuracy using hourly data and obtaining R^2 values exceeding 0.9 for short-term predictions.

Moreover, Hu et al. [11] advanced hybrid CNN-LSTM models for daily minimum DO concentration predictions in the Oyster River. While hybrid models demonstrated superior stability, the standalone LSTM achieved similar accuracy (R^2: 0.865). These studies underscored the utility of hybrid and deep learning models in capturing the temporal dynamics of DO fluctuations. Shaghaghi et al. [12] introduced DOxy, an IoT-based DO monitoring system integrating ML models such as SVM and ODR for real-time predictions. The system achieved high accuracy (RMSE: 0.186) and demonstrated potential for aquaculture sustainability. Lin et al. [13] combined GBR and LSTM in a two-stage hybrid model for DO prediction in lakes, effectively simulating seasonal changes with R^2 values ranging from 0.6 to 0.7 and NMAE below 0.15. Rajagopal et al. [14] proposed the LSTM-CSOA-COOA model, achieving remarkable accuracy (RMSE: 0.01, MAPE: 1.2%) for weekly DO forecasting in the Vaigai River. Nong et al. [15] enhanced SVR models using feature selection methods like LASSO and RF, coupled with optimization algorithms such as Genetic Algorithm (GA) and Random Search (RS). This wavelet-based hybrid model achieved superior performance (RMSE: 0.251, R^2: 0.911). Roushangar et al. [16] developed the SBO-LSTM hybrid model for the Savannah River, surpassing standalone LSTM, SVM, and GPR models (R^2: 0.981, RMSE: 0.034) and highlighting water temperature's critical role in DO dynamics. Bolick et al. [17] used RF models for urban stream DO predictions in South Carolina, achieving NSE values above 0.9 across most locations and identifying location-specific water quality interactions. Rajesh and Rehana [5] assessed climate change impacts on Indian rivers using KNN-LSTM models, revealing a potential 2–12% reduction in DO saturation by 2100 due to increasing river water temperatures. Their findings emphasized the strong correlation between air and river temperatures and the implications for future water quality management. Garabaghi et al. [4] applied RF models with dimensionality reduction to predict DO in the Büyük Menderes River, achieving high performance (PCC: 0.8052, RMSE: 1.2805). Wang et al. [18] demonstrated that SVM and LSTM models could reliably predict DO in the Pearl River Basin, with SVM excelling in generalization (R^2: 0.775) and LSTM effectively capturing temporal patterns.

2. MATERIALS AND METHODS

2.1. Dataset

The dataset used in the study includes measurements of water physicochemical parameters and in situ readings of water flow direction and velocity at thirty-minute intervals. The data were recorded as they may be useful for building time series models and investigating correlations between measurements, as well as investigating how the river changes throughout the year. The dataset is publicly available through the Queensland Government's open data portal. The dataset used includes data taken at specific times each day between 04.08.2023 and 27.06.2024.

Table 1 provides a description of the dataset features, detailing each measurement included in the analysis. In addition to this, Figure 1 presents the correlation matrix of the dataset's features, highlighting the relationships between different physicochemical parameters. The correlation matrix reveals how certain parameters are interrelated, which can provide insights into which features may have stronger influences on dissolved oxygen levels.

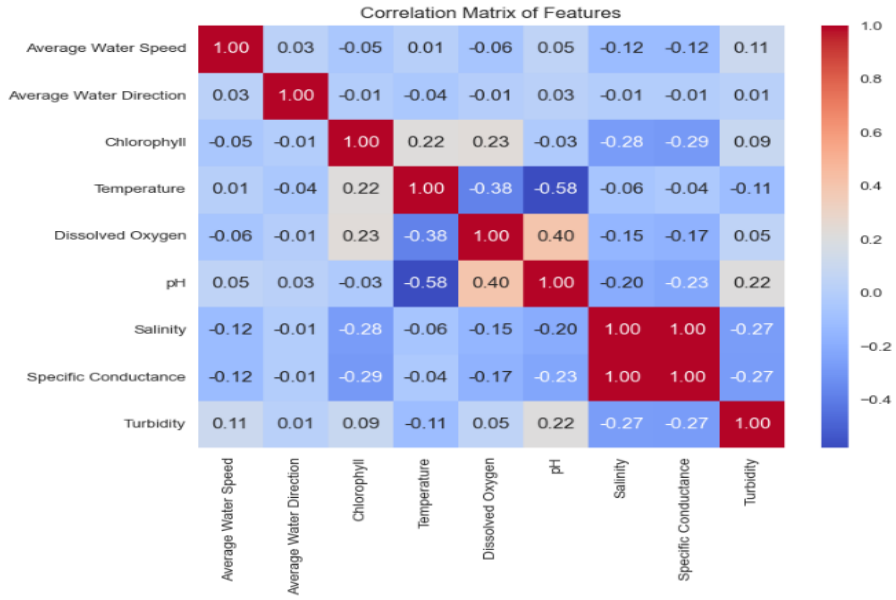

Figure 1. Correlation matrix of features

The dataset is also publicly available on the Kaggle platform [19]. When the data in the data set is analyzed, it is seen that some values are not recorded. These values are removed from the dataset and the operations performed on the recorded data given in Table 1.

Table 1. Description of dataset features

Name	Description
Timestamp	Date and time of measurement
Record Number	The number that uniquely identifies each record.
Average Water Speed	The average speed of the water (m/s).
Average Water Direction	The average direction in which the water moves (in degrees).
Chlorophyll	The concentration of chlorophyll in the water
Temperature	Temperature of the water (°C).
Dissolved Oxygen	Amount of dissolved oxygen in the water (mg/L)
pH	The degree to which the water is acidic or basic
Salinity	Salinity of water (ppt)
Specific Conductance	Electrical conductivity of water
Turbidity	Degree of turbidity of the water

2.2. Machine Learning

Machine learning (ML) is a rapidly evolving field that forms the backbone of many modern technological advancements, offering systems the ability to automatically learn from data and improve performance over time without explicit programming. ML algorithms are capable of identifying complex patterns in large datasets, which makes them invaluable in fields ranging from healthcare to finance, climate modeling, and autonomous systems. Their ability to handle both structured and unstructured data enables a wide range of applications, such as predictive modeling, classification, and clustering. The importance of ML continues to grow as the volume of data increases exponentially, and its integration into decision-making processes becomes essential for developing intelligent systems that can adapt to new challenges [20].

2.2.1. Linear Regression

Linear regression is a classical statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between variables and minimizes the sum of squared errors (Ordinary Least Squares - OLS) to find the best-fit line. The method's foundation can be traced back to the work of Gauss, who introduced the least squares method in 1809 [21].

2.2.2. Support Vector Regression (SVR)

Support Vector Regression (SVR) is an extension of the Support Vector Machine (SVM) algorithm, designed for regression tasks. It aims to find a function that approximates the target variable within a certain margin of tolerance. SVR handles both linear and non-linear regression by using kernel functions. The principles of SVR were first introduced by Cortes and Vapnik in 1995 [22], and it has been widely adopted due to its robustness in handling complex datasets.

2.2.3. XGBoost Regressor

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable implementation of the gradient boosting algorithm. It is particularly suited for structured/tabular data and offers several enhancements over traditional boosting methods, including regularization techniques and parallel processing. XGBoost was developed by Chen and Guestrin, and it quickly gained popularity due to its performance in machine learning competitions [23].

2.2.4. Gradient Boosting Regressor

Gradient Boosting Regressor is a powerful ensemble learning technique that sequentially builds models by minimizing the error of previous models. At each iteration, a new weak learner, typically a decision tree, is added to correct the errors made by the previous learners. This approach was formalized by Friedman in 2001 and has since become a widely used technique for both classification and regression tasks [24].

2.2.5. K-Nearest Neighbors Regressor (KNN)

K-Nearest Neighbors (KNN) Regressor is a simple and intuitive algorithm that predicts the target value based on the kkk closest neighbors in the feature space. The prediction is made by averaging the output values of the nearest data points. The KNN method was introduced by Cover and Hart in 1967 and remains a fundamental algorithm in machine learning for both classification and regression tasks [25].

2.2.6. Random Forest Regressor

Random Forest Regressor is an ensemble learning method that constructs multiple decision trees and merges their results to improve accuracy and prevent overfitting. Each tree is built using a random subset of the data, ensuring diversity among the trees and reducing variance. The Random Forest algorithm was developed by Breiman in 2001 and has become a staple in machine learning for its effectiveness in both classification and regression problems [26].

2.3. Evaluation Metrics

2.3.1. Mean Squared Error (MSE)

Mean Squared Error (MSE) is one of the most commonly used metrics for regression tasks. It measures the average of the squared differences between the predicted values and the actual values. The formula for MSE is given as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \quad (1)$$

where y_i represents the actual values, $\hat{y}_i$are the predicted values, and n is the number of observations. MSE was first formalized in the least squares method by Carl Friedrich Gauss [27].

2.3.2. Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) is the square root of the MSE and provides an interpretable error in the same units as the dependent variable, making it easier to understand the scale of the error. The formula for RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \quad (2)$$

RMSE is commonly used in fields like environmental science and engineering, where interpretability is important. The use of RMSE is well-documented in numerous applications, particularly in model evaluation [28].

2.3.3. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) calculates the average absolute difference between the actual and predicted values, making it less sensitive to outliers compared to MSE. The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

MAE has been widely used since its introduction in early regression analysis for its simplicity and robustness to large errors [29].

2.3.4. R-Squared (R^2)

R-squared (R^2) is a statistical measure that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is defined as:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (4)$$

where $\bar{y}_i$ is the mean of the actual values. R^2 was introduced by Galton and Pearson in the late 19th century and is widely used to evaluate the goodness of fit of regression models [30].

2.3.5. Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) is a percentage-based error metric that provides a relative measure of prediction accuracy. It is defined as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}}{y_i}\right| \times 100 \quad (5)$$

MAPE was popularized by Spyros Makridakis, who utilized it extensively in forecasting applications due to its interpretability and effectiveness in fields like economic forecasting and demand planning [31].

3. RESULTS AND DISCUSSION

In this study, multiple machine learning algorithms were evaluated based on their performance in predicting dissolved oxygen levels. The models were assessed using five different evaluation metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and Mean Absolute Percentage Error (MAPE). The results for each model are presented in Table 2 below:

Table 2. Results of algorithms

Algorithm/Metrics	MSE)	RMSE	MAE	R^2	MAPE
Linear Regression	0.2094	0.4576	0.3591	0.5074	5.4608
Support Vector Regression	0.2770	0.5262	0.4064	0.3488	6.2104
XGBoost Regressor	0.0673	0.2594	0.1895	0.8416	2.8845
Gradient Boosting Regressor	0.1230	0.3507	0.2673	0.7107	4.0706
K-Nearest Neighbors Regressor	0.3141	0.5604	0.4367	0.2614	6.6528
Random Forest Regressor	0.0560	0.2367	0.1643	0.8681	2.5097

The results of this study indicate that ensemble-based methods, particularly the Random Forest Regressor and XGBoost Regressor, consistently outperformed the other machine learning algorithms across various evaluation metrics. Both models achieved the lowest values for Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), which are critical indicators of a model's predictive accuracy. The Random Forest Regressor demonstrated the best overall performance, with the lowest MSE (0.0560), RMSE (0.2367), and MAE (0.1643), suggesting that it can effectively model the relationship between the input features and the target variable (dissolved oxygen levels) with minimal error. Similarly, XGBoost exhibited strong results with an MSE of 0.0673 and RMSE of 0.2594, showing its robustness in handling complex, non-linear patterns in the dataset.

Figure 2 shows the feature importance scores across the models used to predict dissolved oxygen levels. Key features, such as pH, specific conductance, and salinity, consistently emerged as the most influential predictors, especially in ensemble models like Random Forest and XGBoost. Secondary features, including temperature and chlorophyll, also contributed to the model's accuracy, while average water speed, average water direction, and turbidity had lower importance scores. These findings suggest that the models effectively prioritize critical water quality parameters, enhancing prediction reliability.

The R-squared (R^2) values, which reflect how well each model explains the variance in the target variable, were also highest for the Random Forest and XGBoost models. Although these R^2 values (0.8681 for Random Forest and 0.8416 for XGBoost) are relatively high, it is important to note that R^2 values closer to 1 would indicate a better fit, and these models still leave some variance unexplained. However, compared to the other models, these results show that Random Forest and XGBoost are more capable of capturing the underlying trends in the data.

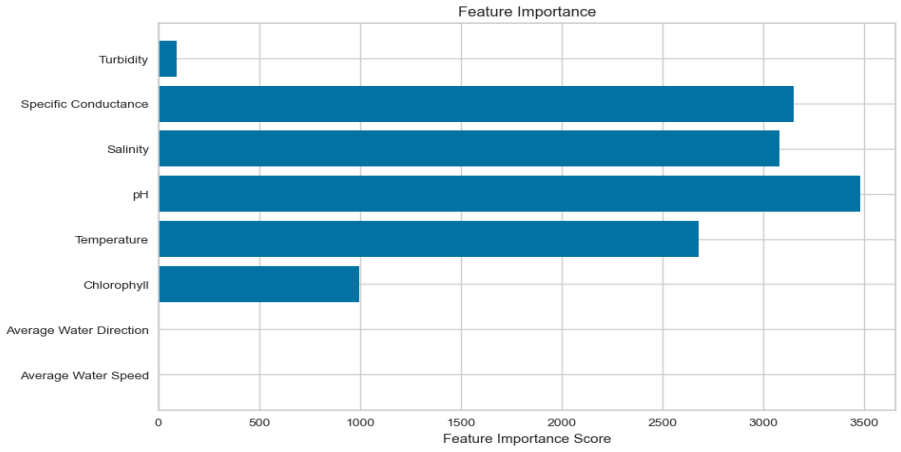

Figure 2. Feature importance graphic

On the other hand, simpler models such as K-Nearest Neighbors and Support Vector Regression showed lower performance across all metrics. K-Nearest Neighbors, for instance, had the highest MSE (0.3141) and RMSE (0.5604), indicating that it struggled to generalize well to the data. Support Vector Regression also had relatively high error rates and lower accuracy, suggesting it may not be well suited for this particular regression task. The Mean Absolute Percentage Error (MAPE) results further support these findings, with Random Forest and XGBoost achieving lower MAPE values (0.8681 and 0.8416, respectively), which indicates better relative prediction accuracy. Conversely, K-Nearest Neighbors (MAPE = 0.2614) and Support Vector Regression (MAPE = 0.3488) showed less accuracy, particularly in cases where the actual values were lower or more variable. To further illustrate the effectiveness of the Random Forest Regressor, Figure 3 shows the comparison between the predicted and actual dissolved oxygen levels over time for the third fold of the 5-fold cross-validation. This visualization demonstrates that the Random Forest model closely tracks the real dissolved oxygen values, highlighting its predictive accuracy. Additionally, Figure 4 presents a residual plot for the Random Forest Regressor model in the third fold, displaying residuals for both training and testing datasets. The residuals are generally centered around zero, with the training R^2 reaching 0.981 and the test R^2 at 0.871, indicating strong model performance and low error across both sets.

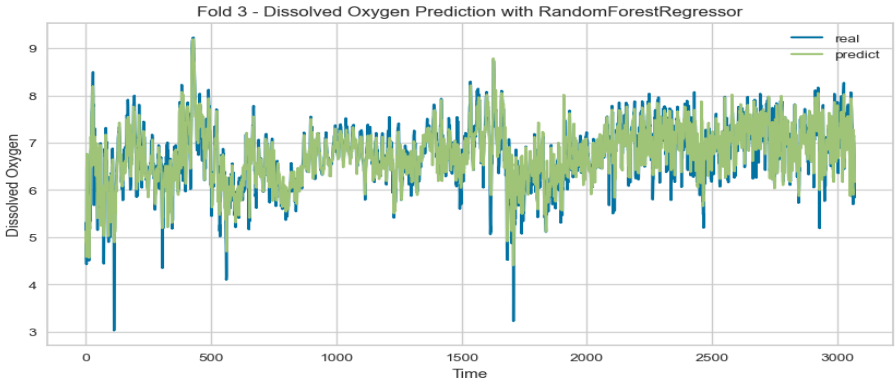

Figure 3. Dissolved Oxygen Prediction with RandomForestRegressor

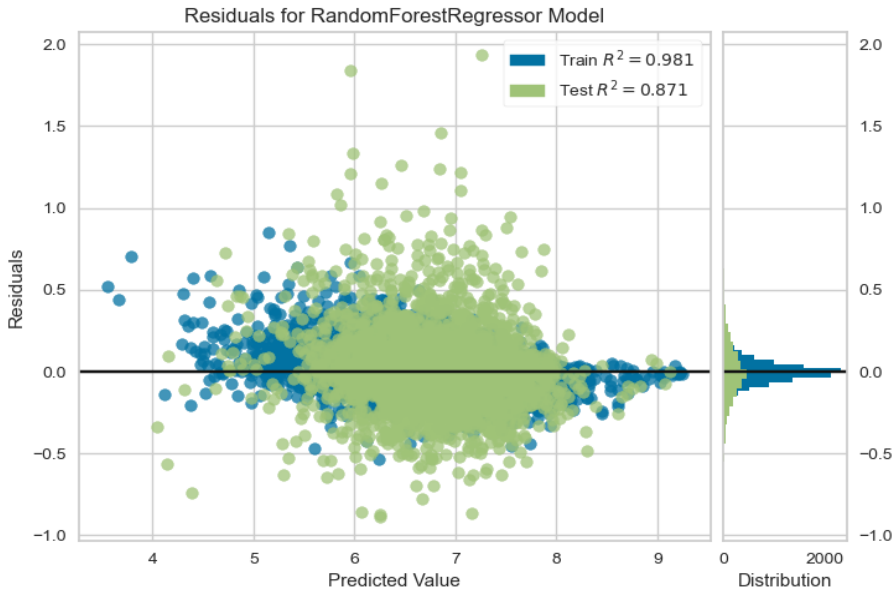

Figure 4. Residuals for RandomForestRegressor Model

In summary, the ensemble methods (Random Forest and XGBoost) emerged as the most effective models for predicting dissolved oxygen levels, demonstrating superior performance across both absolute and relative error metrics. Their ability to handle complex data patterns and variability in the dataset likely contributed to their strong results. Simpler models, including Linear Regression, K-Nearest Neighbors, and Support Vector Regression, while providing some level of accuracy, were less effective in modeling the intricacies of the dataset and thus showed higher error rates. These findings suggest that ensemble approaches are better suited for this type of predictive task, especially when working with environmental datasets that may exhibit non-linear patterns and variability.

4. CONCLUSIONS

In conclusion, this study demonstrates the effectiveness of machine learning, particularly ensemble models like Random Forest and XGBoost, in predicting dissolved oxygen levels and analyzing water quality data. These models achieved high predictive accuracy and identified key physicochemical parameters, such as pH, specific conductance, and salinity, as critical predictors. The findings underscore the potential of machine learning for environmental monitoring, providing valuable insights for resource management and policymaking, while encouraging further research to enhance model accuracy with additional temporal and spatial data. In addition, the study's outcomes can guide water management policies on critical issues such as immediate and future prediction of dissolved oxygen levels, early detection of pollution increases, and prevention of fish kills. The model outputs enable relevant institutions to monitor sudden changes in water quality and take protective measures in a timely manner. Furthermore, the generalizability of the Brisbane River results to other aquatic ecosystems depends on the physicochemical parameter distribution of the water body and the ecological characteristics of the region. For example, the importance of variables such as salinity and specific conductance may increase in estuarine or coastal ecosystems with higher salinity. Therefore, integration of additional parameters (e.g. nutrient salts, heavy metals) and recalibration with local data are recommended to adapt the model to different ecosystems.

Looking ahead, future studies could improve long-term prediction accuracy by including more environmental and climatic factors (e.g. rainfall, seasonal temperature variations, nutrient salts, etc.) using LSTM or Transformer-based deep learning models. Moreover, the performance of the models may be further enhanced through hyperparameter optimization techniques such as Grid Search, Random Search, or Bayesian Optimization.

AUTHOR CONTRIBUTIONS

The first author contributed to conceptualization, supervision, validation, investigation, resource management, and the preparation of the original draft. The second author was responsible for methodology, formal analysis, data curation, visualization, writing—review, and editing. The third author contributed to visualization, software development, investigation, writing—review, and editing. All authors have read and approved the final version of the manuscript.

CONFLICT OF INTEREST

The authors confirm that they have no conflicts of interest to disclose.

ETHICS

This article does not present any ethical issues for publication.

REFERENCES

[1] F. Granata, S. Zhu, and F. di Nunno, "Dissolved oxygen forecasting in the Mississippi River: advanced ensemble machine learning models," Environmental Science: Advances, 2024, doi: 10.1039/d4va00119b.

[2] Z. Wang, Q. Wang, Z. Liu, and T. Wu, "A deep learning interpretable model for river dissolved oxygen multi-step and interval prediction based on multi-source data fusion," Journal of Hydrology, vol. 629, p. 130637, Feb. 2024, doi: 10.1016/J.JHYDROL.2024.130637.

[3] H. Lim, H. Shin, J. Lee, J. Do, I. Song, and Y. Jin, "Prediction of Dissolved Oxygen Factor at Oncheon Stream Watershed Using Long Short-Term Memory Algorithm," Water (Switzerland), vol. 16, no. 17, Sep. 2024, doi: 10.3390/w16172363.

[4] F. H. Garabaghi, S. Benzer, and R. Benzer, "Modeling dissolved oxygen concentration using machine learning techniques with dimensionality reduction approach," Environmental Monitoring and Assessment, vol. 195, no. 7, pp. 1–23, Jul. 2023, doi: 10.1007/S10661-023-11492-3/FIGURES/21.

[5] M. Rajesh and S. Rehana, "Impact of climate change on river water temperature and dissolved oxygen: Indian riverine thermal regimes," 123AD, doi: 10.1038/s41598-022-12996-7.

[6] https://waterdata.usgs.gov/monitoring-location/07374000. [Accessed: 13-September-2024].

[7]Liu, W., Lin, S., Li, X., Li, W., Deng, H., Fang, H., & Li, W. (2024). Analysis of dissolved oxygen influencing factors and concentration prediction using input variable selection technique: A hybrid machine learning approach. Journal of Environmental Management, 357, 120777, doi: 10.1016/J.JENVMAN.2024.120777.

[8] S. Shadkani, Y. Hemmatzadeh, A. Saber, and M. Mohammadi Sergini, "Enhanced predictive modeling of dissolved oxygen concentrations in riverine systems using novel hybrid temporal pattern attention deep neural networks," Environmental Research, vol. 263, Dec. 2024, doi: 10.1016/j.envres.2024.120015.

[9] Li, Q., He, J., Mu, D., Liu, H., & Li, S. (2025). Dissolved Oxygen Modeling by a Bayesian-Optimized Explainable Artificial Intelligence Approach. Applied Sciences, 15(3), 1471, doi: 10.3390/app15031471.

[10] https://www.busan.go.kr/ihe/index. [Accessed: 13-September-2024].

[11] Y. Hu, C. Liu, and W. M. Wollheim, "Prediction of riverine daily minimum dissolved oxygen concentrations using hybrid deep learning and routine hydrometeorological data," Science of The Total Environment, vol. 918, p. 170383, Mar. 2024, doi: 10.1016/J.SCITOTENV.2024.170383.

[12] N. Shaghaghi et al., "DOxy: A Dissolved Oxygen Monitoring System," Sensors, vol. 24, no. 10, May 2024, doi: 10.3390/s24103253.

[13] S. Lin, D. C. Pierson, R. Ladwig, B. M. Kraemer, and F. R. S. Hu, "Multi-Model Machine Learning Approach Accurately Predicts Lake Dissolved Oxygen With Multiple Environmental Inputs," Earth and Space Science, vol. 11, no. 7, Jul. 2024, doi: 10.1029/2023EA003473.

[14] S. Rajagopal, S. S. Ganesh, A. Karthick, and T. Sampradeepraj, "Environmental water quality prediction based on COOT-CSO-LSTM deep learning," Environmental Science and Pollution Research, Sep. 2024, doi: 10.1007/s11356-024-34750-4.

[15] X. Nong, C. Lai, L. Chen, D. Shao, C. Zhang, and J. Liang, "Prediction modelling framework comparative analysis of dissolved oxygen concentration variations using support vector regression coupled with multiple feature engineering and optimization methods: A case study in China," Ecological Indicators, vol. 146, p. 109845, Feb. 2023, doi: 10.1016/J.ECOLIND.2022.109845.

[16] K. Roushangar, S. Davoudi, and S. Shahnazi, "The potential of novel hybrid SBO-based long short-term memory network for prediction of dissolved oxygen concentration in successive points of the Savannah River, USA," Environmental Science and Pollution Research, vol. 30, no. 16, pp. 46960–46978, Apr. 2023, doi: 10.1007/S11356-023-25539-Y/FIGURES/13.

[17] M. M. Bolick, C. J. Post, M. Z. Naser, and E. A. Mikhailova, "Comparison of machine learning algorithms to predict dissolved oxygen in an urban stream," Environmental Science and Pollution Research, vol. 30, no. 32, pp. 78075–78096, Jul. 2023, doi: 10.1007/S11356-023-27481-5/TABLES/7.

[18] X. Wang, Y. Li, Q. Qiao, A. Tavares, and Y. Liang, "Water Quality Prediction Based on Machine Learning and Comprehensive Weighting Methods," Entropy 2023, Vol. 25, Page 1186, vol. 25, no. 8, p. 1186, Aug. 2023, doi: 10.3390/E25081186.

[19] https://www.kaggle.com/datasets/downshift/water-quality-monitoring-dataset/data [Accessed: 3-September-2024].

[20] Jordan, M. I., & Mitchell, T. M. "Machine learning: Trends, perspectives, and prospects." Science, vol. 349, no. 6245, pp. 255-260, 2015. doi: 10.1126/science.aaa8415.

[21] Gauss, C. F. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Hamburg: Friedrich Perthes und I.H. Besser, 1809.

[22] Cortes, C., & Vapnik, V. "Support-vector networks." Machine Learning, vol. 20, no. 3, pp. 273-297, 1995. doi: 10.1007/BF00994018.

[23] Chen, T., & Guestrin, C. "XGBoost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, Aug. 2016, pp. 785-794. doi: 10.1145/2939672.2939785.

[24] Friedman, J. H. "Greedy function approximation: A gradient boosting machine." Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001. doi: 10.1214/aos/1013203451.

[25] Cover, T., & Hart, P. "Nearest neighbor pattern classification." IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967. doi: 10.1109/TIT.1967.1053964.

[26] Breiman, L. "Random forests." Machine Learning, vol. 45, no. 1, pp. 5-32, 2001. doi: 10.1023/A:1010933404324.

[27] Gauss, C. F. Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Hamburg: Friedrich Perthes und I.H. Besser, 1809.

[28] Willmott, C. J., & Matsuura, K. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." Climate Research, vol. 30, no. 1, pp. 79-82, 2005. doi: 10.3354/cr030079.

[29] Willmott, C. J., & Matsuura, K. "On the use of dimensioned measures of error to evaluate the performance of spatial interpolators." International Journal of Geographical Information Science, vol. 20, no. 7, pp. 801-820, 2006. doi: 10.1080/13658810600661574.

[30] Pearson, K. "Mathematical contributions to the theory of evolution—III." Philosophical Transactions of the Royal Society of London, Series A, vol. 187, pp. 253-318, 1896.

[31] Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. Forecasting: Methods and Applications. John Wiley & Sons, 1998.