MODELLING DISSOLVED OXYGEN IN INTENSIVE AQUACULTURE SYSTEMS: LINEAR REGRESSION vs RANDOM FOREST APPROACHES

Ahmad Yani[1,*], Asthervina W. Puspitasari [1], Hendra Poltak[1], Ahmad Fahrizal[2], Rusli[1], and Ramadhona Saville[3]

[1]Aquaculture Study Program, Marine and Fisheries Polytechnic of Sorong, Jalan Kapitan Patimura Tg. Kasuari, Maladummes District, Sorong 98401, Southwest Papua, Indonesia
[2]Aquatic Resources Management Study Program, Faculty of Fisheries, Universitas Muhammadiyah Sorong, Jalan Pendidikan No. 27, Malaimsimsa District, Sorong 98412, Southwest Papua, Indonesia
[3]Tokyo University of Agriculture, 1-1-1 Sakuragaoka, Setagaya City, Tokyo 1568502, Japan

(Submitted: 24 September 2025; Final revision: 11 December 2025; Accepted: 11 December 2025)

ABSTRACT

Dissolved oxygen (DO) is a critical water quality parameter in intensive aquaculture systems because its fluctuations directly affect farmed fish. Accurate prediction of DO is challenging due to complex, often nonlinear interactions among physicochemical and biological variables. Despite increasing interest in machine learning applications, comparative evaluations between traditional linear models and ensemble-based approaches in aquaculture contexts remain limited. This study aimed to analyse key variables associated with DO dynamics, compare the predictive performance of linear regression (LR) and random forest (RF) models, and identify dominant predictors relevant to aquaculture management. A publicly available aquaculture water quality dataset from Mendeley Data was analysed. Data were preprocessed by outlier removal and normalization, then split into training (70%) and test (30%) sets, and model robustness was assessed using 5-fold cross-validation. Dissolved oxygen concentrations ranged from 0.21 to 10.17 mg L^{-1} (mean = 5.19 mg L^{-1}). Pearson correlation analysis showed positive associations between DO and ammonia ($r = 0.60$), biochemical oxygen demand ($r = 0.55$), and nitrite ($r = 0.52$), and negative associations with hydrogen sulphide ($r = -0.55$) and turbidity ($r = -0.53$). These relationships reflected indirect, management-mediated effects rather than direct causation. The RF model slightly outperformed LR ($R^2 = 0.515$ vs. 0.470), demonstrating the advantage of non-linear modelling. The feature importance analysis identified ammonia, hydrogen sulphide, nitrite, and biochemical oxygen demand as the dominant predictors. Although predictive accuracy remained moderate, the results highlight key drivers of DO variability and support the use of machine learning as a decision-support tool for smart aquaculture management.

KEYWORDS: dissolved oxygen; intensive aquaculture; linear regression; machine learning; random forest; water quality modelling

ABSTRAK: Pemodelan Oksigen Terlarut dalam Sistem Budidaya Intensif: Pendekatan Regresi Linier vs Random Forest

*Correspondence: Aquaculture Study Program, Marine and Fisheries Polytechnic of Sorong, Jalan Kapitan Patimura Tg. Kasuari, Maladummes District, Sorong 98401, Southwest Papua, Indonesia
Email: ahmad.yani73@kkp.go.id

Oksigen terlarut (dissolved oxygen = DO) merupakan parameter kualitas air yang sangat penting dalam sistem akuakultur intensif karena fluktuasinya secara langsung memengaruhi komoditas yang dibudidayakan. Prediksi DO yang akurat menjadi tantangan karena adanya interaksi yang kompleks dan sering kali bersifat nonlinier antara variabel fisikokimia dan biologis. Meskipun minat terhadap penerapan machine learning terus meningkat, evaluasi komparatif antara model linier tradisional dan pendekatan berbasis ensemble dalam konteks akuakultur masih terbatas. Penelitian ini bertujuan untuk menganalisis variabel-variabel utama yang berkaitan dengan dinamika DO, membandingkan kinerja prediktif model regresi linier (LR) dan random forest (RF), serta mengidentifikasi prediktor dominan yang relevan untuk pengelolaan akuakultur. Dataset kualitas air akuakultur yang tersedia secara publik dari Mendeley Data dianalisis dalam penelitian ini. Data dipraproses melalui penghapusan pencilan dan normalisasi, kemudian dibagi menjadi data training (70%) dan pengujian (30%), dengan ketahanan model dievaluasi menggunakan validasi silang lima lipatan. Konsentrasi DO berkisar antara 0,21 hingga 10,17 mg L^{-1} (rata-rata = 5,19 mg L^{-1}). Analisis korelasi Pearson menunjukkan hubungan positif antara DO dan amonia ($r = 0,60$), kebutuhan oksigen biokimiawi (BOD; $r = 0,55$), serta nitrit ($r = 0,52$), dan hubungan negatif dengan hidrogen sulfida ($r = -0,55$) dan kekeruhan ($r = -0,53$). Hubungan tersebut mencerminkan efek tidak langsung yang dimediasi oleh praktik pengelolaan, bukan hubungan kausal langsung. Model RF menunjukkan kinerja yang sedikit lebih baik dibanding LR ($R^2 = 0,515$ vs. 0,470), yang menegaskan keunggulan pemodelan nonlinier. Analisis kepentingan fitur mengidentifikasi amonia, hidrogen sulfida, nitrit, dan kebutuhan oksigen biokimiawi sebagai prediktor dominan. Meskipun akurasi prediksi masih tergolong moderat, hasil penelitian ini menyoroti faktor-faktor utama yang memengaruhi variabilitas DO dan mendukung penerapan machine learning sebagai alat pendukung keputusan dalam pengelolaan akuakultur cerdas.

KATA KUNCI: akuakultur intensif; machine learning; oksigen terlarut; pemodelan kualitas air; random forest; regresi linear

INTRODUCTION

Dissolved oxygen (DO) is one of the most important indicators of water quality in aquaculture, with direct implications for fish health, growth, and productivity. Low DO can cause physiological stress, reduced growth, increased disease susceptibility, and, in severe cases, mass mortality (Abdel-Tawwab *et al.*, 2014; Schäfer *et al.*, 2021). Accurate prediction of DO dynamics is therefore essential for sustainable aquaculture management.

DO levels are influenced by multiple environmental factors, including temperature, pH, salinity, turbidity, respiration, and photosynthesis (Greig *et al.*, 2007; He *et al.*, 2011). These interactions are highly complex and often nonlinear, which limits the effectiveness of traditional linear regression models in capturing such variability (Jeong *et al.*, 2024; Jongjaraunsuk *et al.*, 2024; Liu *et al.*, 2023b; Xu *et al.*, 2023; Zhang *et al.*, 2019).

In recent years, machine learning (ML) approaches have been increasingly applied to water quality prediction, offering improved performance by modelling nonlinear interactions among variables (Yang *et al.*, 2023). In line with this, hybrid frameworks combining machine learning algorithms using decomposition methods have demonstrated strong potential in improving DO prediction. By analyzing time-series patterns, these

methods not only enhance predictive accuracy but also provide interpretability of the underlying dynamics (Tong *et al.*, 2024). Such innovations are particularly valuable for real-time monitoring and management of DO in recirculating aquaculture systems (RAS), where precise control is essential to maintain resilient and healthy aquatic systems (Lee *et al.*, 2024; Yao *et al.*, 2023).

Although DO levels in aquaculture have traditionally been managed through operational practices such as aeration, feeding strategies, and control of biological load (Boyd & McNevin, 2020; Chakravarty *et al.*, 2022), these approaches can be limited in responsiveness and precision. Recent advances in smart aquaculture systems have transformed DO management by integrating automated monitoring with advanced control technologies. Through continuous analysis of key water quality parameters, including temperature and ammonia, these systems enable more accurate and timely adjustments to maintain stable DO conditions (Nagamora *et al.*, 2022; Vo *et al.*, 2021). Many platforms now incorporate machine learning algorithms that enhance predictive capabilities and support proactive interventions. Compared to simpler models that rely on historical averages, machine-learning-based approaches offer a more reliable and dynamic means of sustaining optimal DO levels (Kuang *et al.*, 2020).

Although the ensemble structure of random forests (RFs) is recognized for its effectiveness in managing multicollinearity in water quality data, comparative studies on this application remain limited. The present investigation aims to address this specific gap by: (1) Analyzing key environmental variables affecting DO dynamics, (2) Developing predictive models using linear regression (LR) and random forest (RF), (3) Comparing their predictive performance, and (4) Identifying the most influential predictors of DO to support aquaculture management. By addressing these objectives, this research contributes to the growing literature on ML applications in aquaculture and provides insights into how predictive modelling can enhance water quality monitoring and management.

MATERIALS AND METHODS

Data Source

This study used Water Quality Dataset (Veeramsetty *et al.*, 2024) published online at Mendeley data that can be accessed online at https://data.mendeley.com/datasets/y78ty2g293/1, which consists of 4,300 observations with 15 water quality variables. These include temperature, pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), ammonia, nitrite, hydrogen sulphide (H_2S), phosphorus, calcium, alkalinity, hardness, turbidity, CO_2, and plankton counts. The dataset represents intensive aquaculture systems in fish ponds, but the specific sampling locations and collection periods were not provided in the original source.

Data Analysis

Pre-Processing Data

Several pre-processing stages were employed prior to modelling, which are:

1. Handling missing data. Observations with incomplete values were excluded rather than imputed to ensure consistency in the input features
2. Outlier removal. Extreme values were removed utilising the interquartile range (IQR) method, defined as values outside $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$.
3. Normalisation. Predictor variables were scaled into the 0-1 range using min-max scaling to balance their influence.
4. Data splitting. In this study, the dataset was randomly divided into a training set (70%) and a testing set (30%). The 70–30 ratio is one of the most widely used practices in predictive modelling, as it provides a balanced trade-off between training and evaluation. Allocating 70% to training ensures the model captures underlying

patterns effectively, while 30% to testing provides a sufficiently large, independent sample to evaluate generalization performance (Kuang *et al.*, 2020; Kuhn & Johnson, 2013). Alternative strategies, such as 80–20 or 90–10, are often applied when datasets are small and maximizing the training size is critical. However, given that the present dataset contains more than 4,000 observations, the 70–30 split was considered appropriate to avoid overfitting while maintaining robust evaluation. This approach is consistent with prior work in environmental and aquaculture-related water-quality modelling, where a 70–30 train–test split is commonly used (Heddam & Kisi, 2017; Kuang *et al.*, 2020; Yang *et al.*, 2023).

To minimize bias from random partitioning, the evaluation was further strengthened using 5-fold cross-validation, which provides a more stable estimate of model accuracy across multiple data splits

Modelling

In the modelling stage, this study used Python's Integrated Development and Learning Environment (Python 3.13.7 with Tcl/Tk 8.6.16) with two different algorithmic approaches: linear regression and random forest regression. The linear regression model was used as a baseline to examine the linear relationships between water-quality predictor variables (such as temperature, pH, turbidity, and ammonia concentration) and DO. This model was trained using the training set and subsequently generated predictions for DO levels on the testing set.

For comparison, the random forest regression algorithm was also used, an ensemble method based on decision trees that is capable of capturing non-linear relationships and interactions between variables (Biau & Scornet, 2016; Schonlau & Zou, 2020). In this study, the random forest parameters were set with the number of trees at 200, the maximum depth limited to 10 levels, and a random state of 42 to ensure the consistency of the results. After the model was trained with the training data, DO predictions were generated on the testing data for further evaluation.

Model Evaluation

The performance of both models was evaluated using three statistical metrics: root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). RMSE measures the standard deviation of the residuals (prediction errors) between actual and predicted values; MAE assesses the average absolute error (Hodson, 2022); and R^2 indicates the proportion of variance in dissolved oxygen explained by the model.

In addition to the train-test split evaluation, this study also implemented K-Fold Cross Validation with $k = 5$ to ensure the reliability of the results. In this procedure, the dataset was divided into five balanced folds, and the model was subsequently trained and tested five times, with each fold alternately serving as the test data. The RMSE, MAE, and R^2 values from each fold were then averaged to obtain a more stable performance estimate.

Dominant Factor Analysis

In addition to the predictive performance evaluation, this study also conducted a feature importance analysis on the random forest model to identify the water quality parameters that have the most significant influence on DO dynamics. A brief summary of the research method in this study is shown in Figure 1.

RESULTS AND DISCUSSION

Descriptive Statistics

In this section, the results of the descriptive statistics were presented. Before data pre-processing, the dataset consisted of 4,300 observations with 15 water quality variables.

The dataset description has been clarified to specify that the data represent intensive fish pond aquaculture systems, based on secondary data from Mendeley Data. After performing data pre-processing, there was no missing data. In the outlier removal step, 3,455 samples were obtained with 15 variables. Descriptive statistics of the data after preprocessing were presented in Table 1.

Results of the descriptive analysis

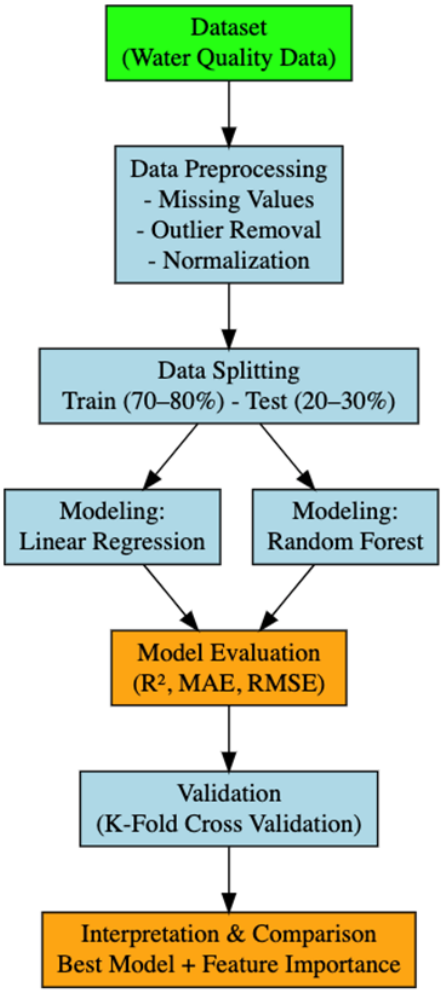

Figure 1. Schematic overview of the research workflow for modelling dissolved oxygen in intensive aquaculture systems using linear regression and random forest approaches

Table 1. Descriptive statistics of dissolved oxygen and key water quality parameters (mean, standard deviation, minimum, quartiles, and maximum) in intensive aquaculture systems after data preprocessing

Variable	Mean	STD	Min	25%	50%	75%	Max
DO (mg L^{-1})	5.19	1.50	0.21	3.98	4.95	6.41	10.17
Ammonia (mg L^{-1})	0.02	0.01	0.00	0.01	0.02	0.04	0.07
H_2S (mg L^{-1})	0.01	0.01	0.00	0.01	0.02	0.02	0.03
Nitrite (mg L^{-1})	0.52	0.67	0.00	0.01	0.02	1.03	2.87
BOD (mg L^{-1})	2.77	1.56	1.00	1.50	2.00	4.04	8.51

Note: STD = standard deviation, Min = minimum value, Max = maximum value, BOD = biochemical oxygen demand.

revealed that the mean DO level was 5.19 mg $L^{-1,}$ with an interval of 0.21–10.17 mg L^{-1}. This indicated that DO conditions varied across the data. Ammonia and H_2S were relatively low and remained stable, while nitrite and BOD showed greater variation, with maximum values reaching 2.87 mg L^{-1} and 8.51 mg L^{-1}, respectively. This suggests heterogeneity in water quality, particularly in parameters related to biological processes.

Figure 2 presents the distribution of DO within the data. The DO distribution shows that the majority of samples fall within the 4-6 mg L^{-1} interval. Some samples with low DO (about 0-2 mg L^{-1}) and others with high DO (nearly 10 mg L^{-1}). This pattern indicates that variation within the dataset is quite varied. The majority of the samples are in relatively good condition, with moderate DO levels. As depicted in Figure 2, the right-skewed histogram of dissolved oxygen (DO) indicates a higher frequency of low-DO concentrations, underscoring the need for predictive alert systems.

Figure 3 compares the boxplots of DO

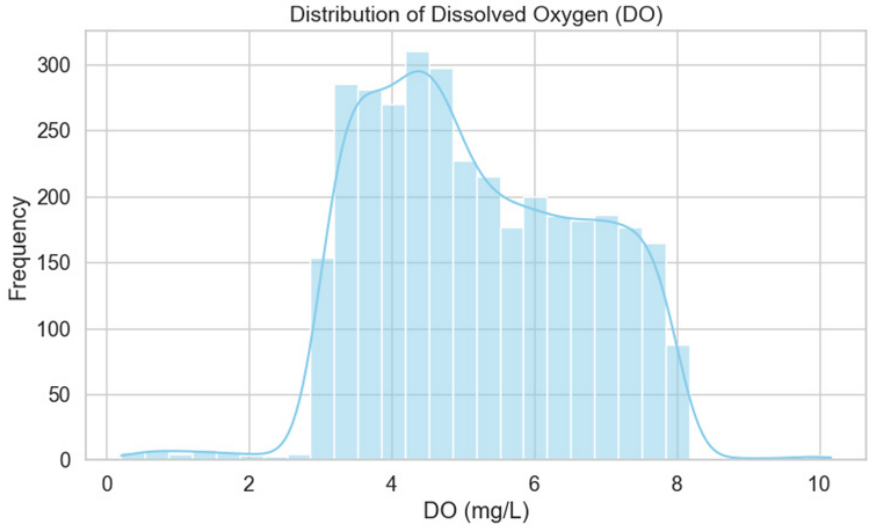

Figure 2. Frequency distribution of dissolved oxygen concentrations in intensive aquaculture systems after data preprocessing

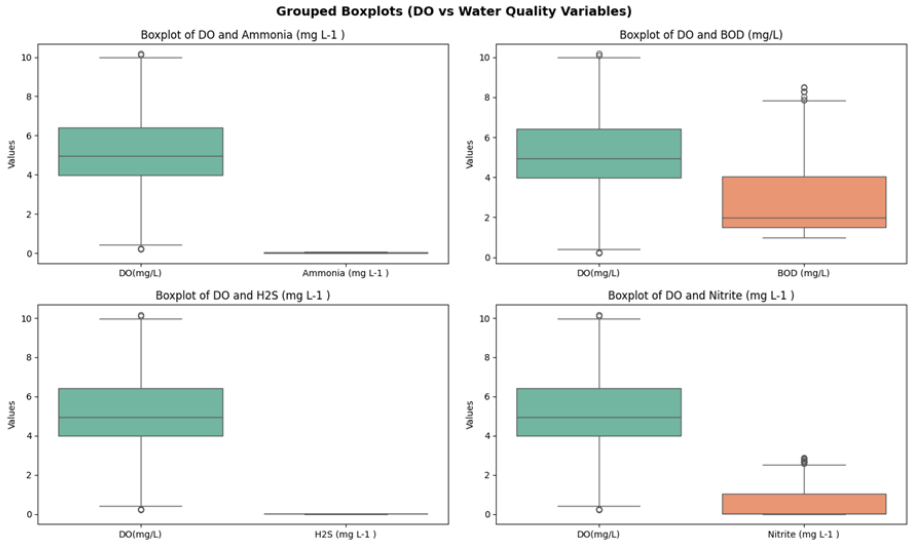

Figure 3. Boxplot comparison of dissolved oxygen concentrations and selected key water quality parameters in intensive aquaculture systems

concentrations with those of the main water quality variables: ammonia, BOD, H_2S, and nitrite. DO distribution is relatively wide, with several outliers, suggesting that its levels fluctuated during the observation period. In contrast, ammonia and H_2S levels tended to be low and stable with little variation. Meanwhile, BOD and nitrite showed a much wider spread in their data, along with a few outliers. This pattern indicates that, even though DO levels were volatile, most other key water quality variables remained relatively stable and still had a significant impact on dissolved oxygen dynamics.

Moving into Figure 4, it shows a violin plot comparing DO with the main water quality parameters, consisting of ammonia, BOD, H_2S, and nitrite. DO distribution shows a wide range, with a high density of data clustered in the 4–6 mg L^{-1} zone, indicating that the majority of samples are concentrated around these values. In contrast, ammonia and H_2S have narrow distributions with relatively low values. BOD and nitrite, on the other hand, show greater variation, although most of their data also remains in the lower range. The shape of these violin plots confirms that DO is more volatile than most other variables, suggesting it is potentially more sensitive to changes in environmental conditions.

Correlation Analysis

The Pearson correlation analysis in Table 2 revealed significant relationships between DO and other water quality parameters, both positive and negative.

Dissolved oxygen (DO) in the aquaculture system suggested strong positively correlated with ammonia, BOD, and nitrite, likely due to management practices such as aeration rather than direct causation. Alkalinity, calcium, hardness, phosphorus, and plankton were

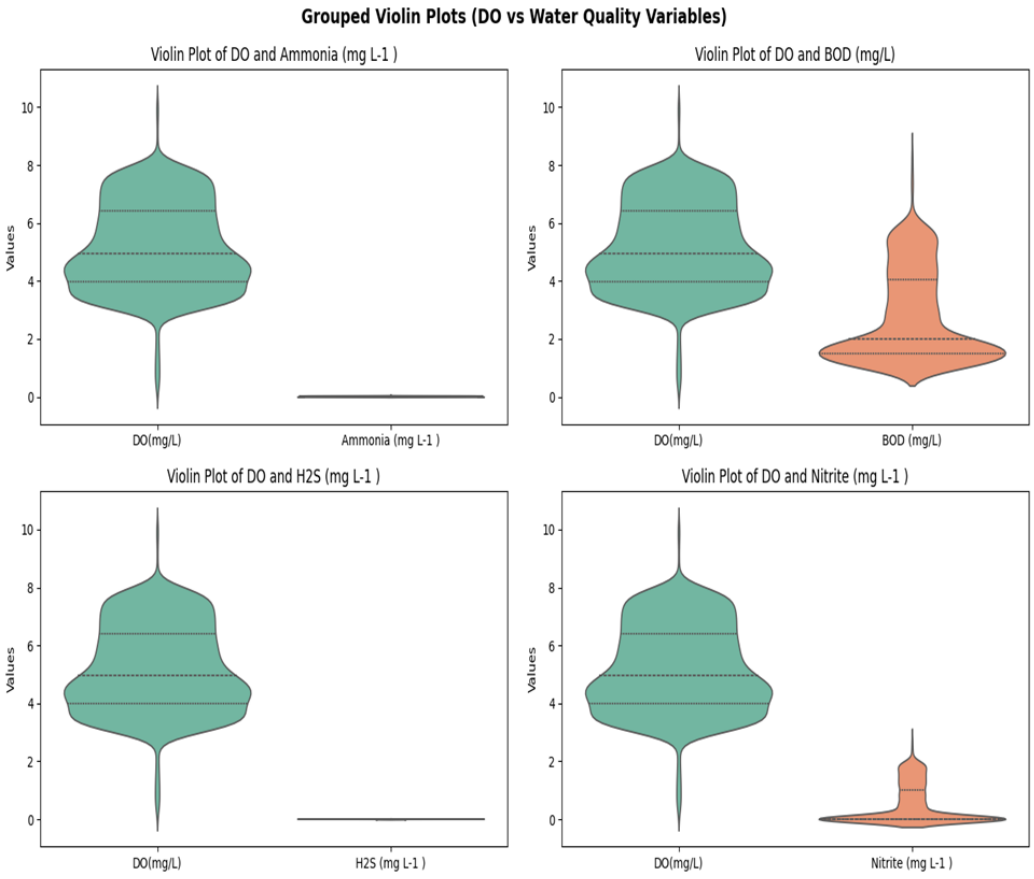

Figure 4. Violin plots illustrating the distribution and density of dissolved oxygen and selected key water quality parameters in intensive aquaculture systems

weakly positively correlated with DO, suggesting only minor roles in DO dynamics. Negative correlations with turbidity, H_2S, and CO_2 in line with ecological principles, as these factors reduce oxygen availability (Boyd & McNevin, 2020; Das & Behera, 2008). Temperature and pH showed no significant effect, indicating stability. Overall, DO is mainly influenced by biological activity and decomposition processes, making it a key predictor for regression and machine learning models.

The relationship between DO and water quality variables was visualised using various approaches. Figure 5 showed DO distribution

Table 2. Pearson correlation coefficients (r) between dissolved oxygen concentrations and other water quality parameters in intensive aquaculture systems

Variable	Pearson r	*p*-value
Ammonia (mg L^{-1})	0.60	0.00
BOD (mg L^{-1})	0.55	0.00
Nitrite (mg L^{-1})	0.52	0.00
Alkalinity (mg L^{-1})	0.21	0.00
Calcium (mg L^{-1})	0.17	0.00
Hardness (mg L^{-1})	0.10	0.00
Phosphorus (mg L^{-1})	0.10	0.00
Plankton (No. L^{-1})	0.05	0.00
Temperature (^{0}C)	0.00	0.87
pH	-0.03	0.06
CO_2 (mg L^{-1})	-0.08	0.00
Turbidity (cm)	-0.53	0.00
H_2S (mg L^{-1})	-0.55	0.00

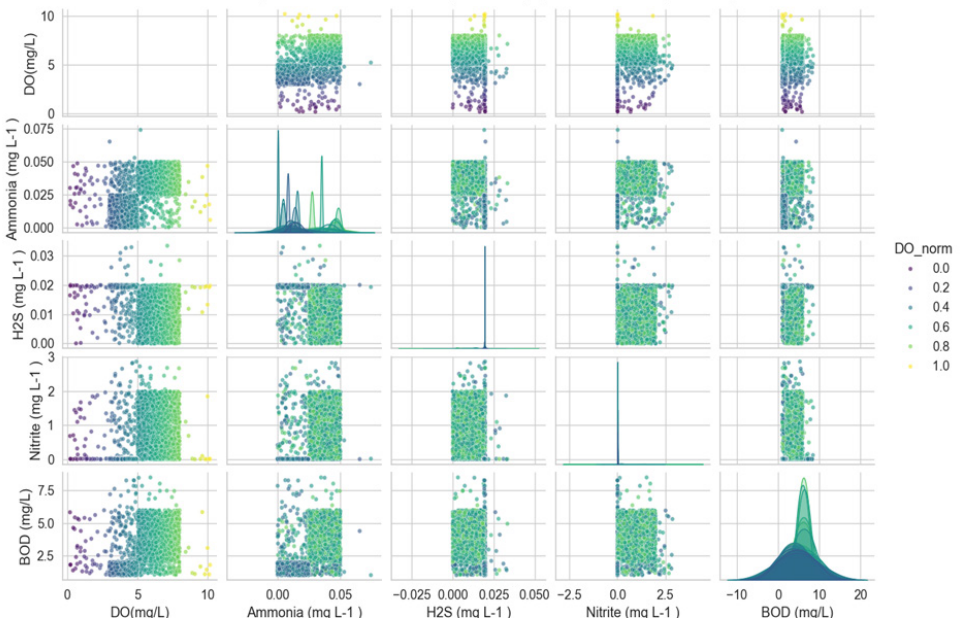

Figure 5. Scatter matrix illustrating the relationships between dissolved oxygen and ammonia, biochemical oxygen demand, hydrogen sulphide, and nitrite in intensive aquaculture systems

patterns in the presence of ammonia, H_2S, nitrite, and BOD. DO tended to be positively correlated with ammonia, nitrite, and BOD, and negatively correlated with H_2S. However, distribution patterns were widely dispersed, indicating non-linear relationship between variables.

The scatter plot details in Figure 6, along with the linear regression line and the R2 value, strengthen the previous findings. DO revealed a positive correlation with ammonia (R^2=0.361), nitrite (R^2=0.272), and BOD (R^2=0.307); DO was negatively correlated with H_2S (R^2=0.305). The moderate R^2 values indicate that while the analysed variables have a tangible influence, other factors also contribute to DO dynamics.

Moving to heatmap correlation (Figure 7), it gives a summary of the strength of

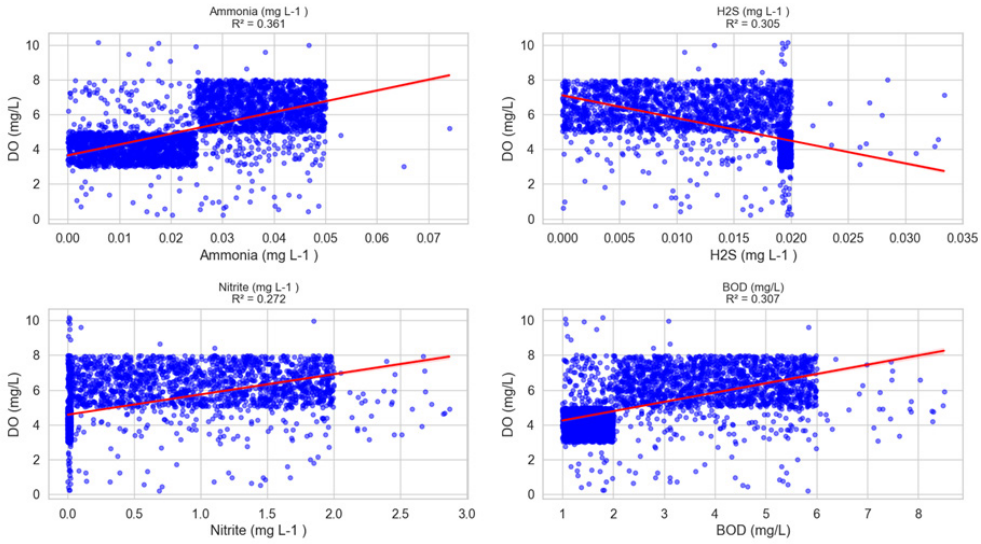

Figure 6. Scatter plots with linear regression lines showing the relationships between dissolved oxygen and ammonia, biochemical oxygen demand, hydrogen sulphide, and nitrite in intensive aquaculture systems

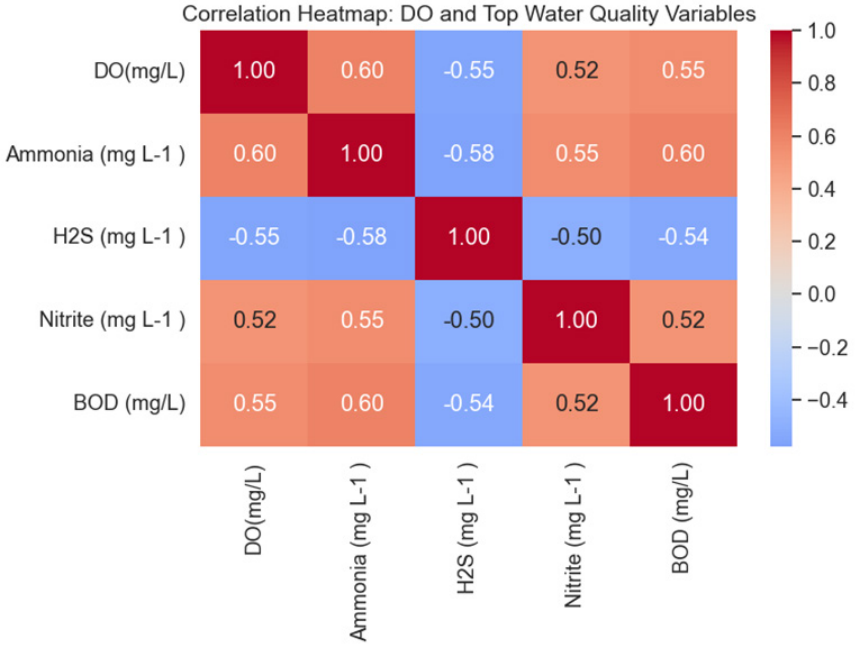

Figure 7. Correlation heatmap illustrating the strength and direction of relationships between dissolved oxygen, ammonia, biochemical oxygen demand, hydrogen sulphide, and nitrite in intensive aquaculture systems

variable interactions. DO is positively strongly correlated with ammonia (r = 0.60), BOD (r = 0.55), and nitrite (r = 0.52), and negatively correlated with H_2S (r = -0.55). This pattern is consistent with the principles of aquatic ecology, in which the accumulation of organic matter and decomposition processes can reduce dissolved oxygen levels (Barnes & Mann, 2009).

Overall, the combination of a scatter matrix, a scatter plot, and a heatmap correlation revealed that chemical-biological parameters (ammonia, nitrite, BOD, and H_2S) were the dominant factors influencing DO, while physical variables such as temperature and pH were less significant in this study. These findings also emphasise the need for a non-linear modelling approach, like random forest, to capture complex interactions between variables.

Model Performance Comparison

The correlation analysis shows that several water quality parameters, e.g., ammonia, BOD, nitrite, turbidity, and H_2S, have a reasonably strong relationship with DO. On the other hand, variables such as temperature and pH did not show any significant correlation within the observations.

This finding suggests that DO dynamics in an intensive aquaculture system are more complex than can be explained solely by physical factors. The varied relationships, both positive and negative, between DO and the environmental variables also point to non-linear interactions that would be difficult for a simple linear model to capture. For this reason, we evaluated the predictive models using two approaches: linear regression as a baseline model and random forests, a machine-learning-based model capable of capturing non-linear relationships and interactions between variables.

Evaluation of model performance (see Table 3) revealed that both linear regression (LR) and random forest (RF) have limited predictive ability in explaining DO variance in the intensive aquaculture system. The linear regression model yielded an R^2 of 0.470, indicating it could only explain about 47% variance in the DO data, with a prediction error RMSE of 1.099 mg L^{-1} and an MAE of 0.828 mg L^{-1}.

The random forest model performed slightly better, with an R^2 of 0.515 and lower prediction errors (RMSE = 1.051 mg L^{-1}; MAE = 0.780 mg L^{-1}). This suggests that RF is better at capturing non-linear relationships between variables than linear regression, though the improvement in accuracy is relatively small. RF's edge ($\Delta R^2 = 0.045$) likely stems from bagging reducing variance in nonlinear interactions, but persistent errors (RMSE$\sim$1 mg L^{-1}) suggest unmodeled temporal dynamics or interactions (e.g., temperature*BOD), warranting lagged variables in future work.

Visually, the scatter plot of observed versus predicted values (see Figure 3) shows that both LR and RF still exhibit significant scatter from the ideal line. However, the predictions from the RF model tend to cluster more closely to this line than those from the LR model.

Therefore, RF can be considered more suitable for modelling DO dynamics, though further exploration with other machine learning approaches or the addition of more relevant predictor variables would be necessary to significantly improve accuracy.

Table 3. Comparison of predictive performance metrics (RMSE, MAE, and R^2) for linear regression and random forest models in dissolved oxygen prediction

Model	RMSE	MAE	R^2
Linear regression	1.099	0.828	0.470
Random forest	1.051	0.780	0.515

Variable Importance

Analysis of feature importance random forest model (see Table 4, Figure 9) revealed that the most influential parameter for projecting DO is ammonia (0.34), followed by H_2S (0.26) and nitrite (0.14). These three parameters have the greatest impact, given their biological and chemical roles in organic decomposition and their potential to degrade water quality (Akhtar *et al.*, 2021).

Other variables, such as BOD (0.05), alkalinity (0.03), hardness (0.03), pH (0.03), and calcium (0.03), had a moderate impact. Factors such as phosphorus, plankton, temperature, CO_2, and turbidity had lower importance scores (all around 0.02). While these factors do not contribute much on their own, they still help to make the model more accurate, especially in certain environmental situations.

These findings are consistent with the previous correlation analysis, confirming that chemical-biological factors, including organic matters and decomposition process (ammonia, H_2S, nitrite, and BOD), are the main determinants of DO dynamics in intensive aquaculture systems. Therefore, monitoring and controlling these parameters should be a priority to maintain water quality stability.

Implications for Aquaculture Management

Findings of this study revealed that DO dynamics in an intensive aquaculture system were influenced by a combination of chemical and biological factors, with ammonia, H_2S, nitrite, and BOD as the dominant variables. This situation aligns with aquatic ecology theory, which states that the decomposition of organic matter and the accumulation of toxic compounds play important roles in reducing DO concentrations (Boyd & Tucker, 2012). These findings are also consistent with a study by Liu *et al.* (2023) and Zhang *et al.* (2020), who reported that organic waste parameters are closely related to DO fluctuation in intensive aquaculture systems.

Pearson correlation analysis in this study

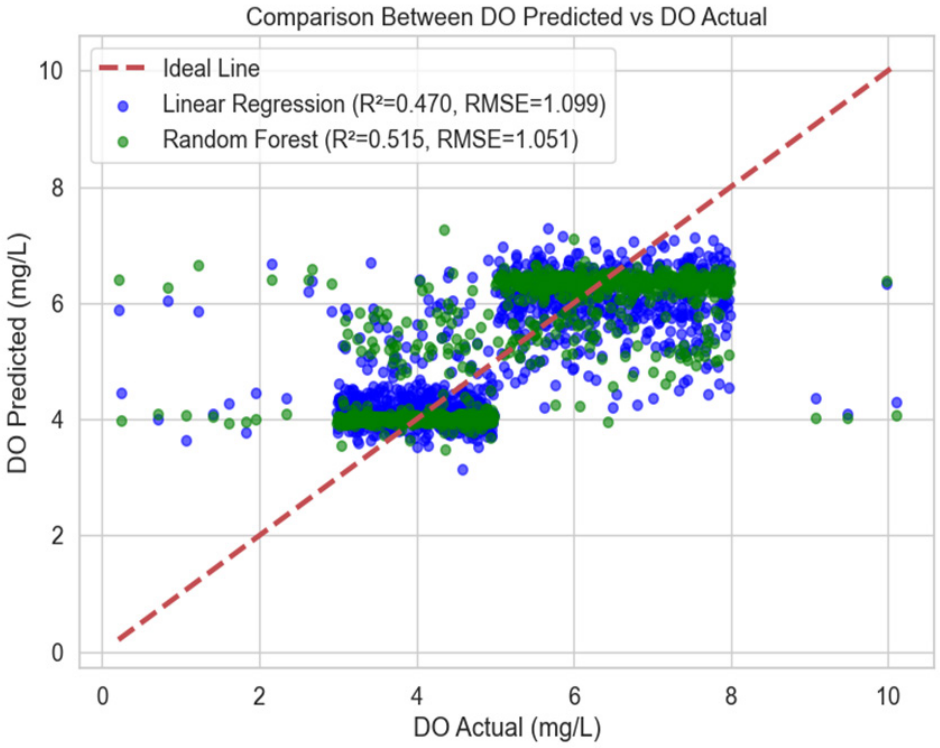

Figure 8. Scatter plots of observed versus predicted dissolved oxygen concentrations generated by linear regression and random forest models in intensive aquaculture systems

Table 4. Feature importance scores of individual water quality variables derived from the random forest model for predicting dissolved oxygen in intensive aquaculture systems

Variable	Importance
Ammonia (mg L^{-1})	0.34
H_2S (mg L^{-1})	0.26
Nitrite (mg L^{-1})	0.14
BOD (mg L^{-1})	0.05
Alkalinity (mg L^{-1})	0.03
Hardness (mg L^{-1})	0.03
pH	0.03
Calcium (mg L^{-1})	0.03
Phosphorus (mg L^{-1})	0.02
Plankton (No. L^{-1})	0.02
Temperature (^{0}C)	0.02
CO_2 (mg L^{-1})	0.02
Turbidity (cm)	0.02

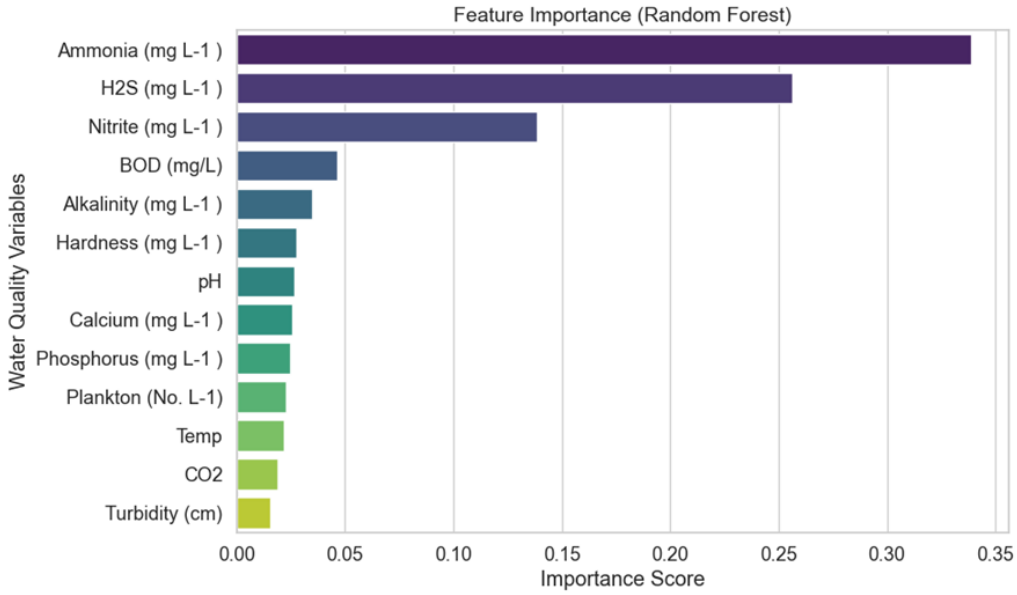

Figure 9. Feature importance ranking of water quality parameters influencing dissolved oxygen prediction based on the random forest model in intensive aquaculture systems

showed positive relationships between DO and ammonia, nitrite, and BOD, and a negative relationship with H_2S. Even though the R^2 value in the scatter regression plot showed a moderate relationship, indicating that DO dynamics are not influenced by a single factor but are a result of complex interactions among variables. This pattern explains why the simple linear approach has a limitation in its ability to achieve high predictive accuracy, as evidenced by the relatively low performance of linear regression ($R^2 = 0.470$).

Conversely, random forests showed better predictive performance (R^2 = 0.515), though the improvement in accuracy over linear regression remained limited. This indicates that machine learning models are indeed superior at capturing non-linear relationships and interactions among variables, but it should be noted that the predictive quality is highly dependent on the completeness and variability of the input data (Wang *et al.*, 2018). Therefore, further exploration using other algorithms such as Gradient Boosting or Neural Networks presents an opportunity to enhance DO prediction accuracy in future research.

Analysis of feature importance in random forest emphasised that ammonia, H_2S, nitrite, and BOD were the main determinants of DO, whilst physical variables, e.g., temperature and pH, have a relatively small contribution. This finding is highly relevant to aquaculture practices, as it provides a basis for prioritizing water-quality monitoring. Furthermore, this finding also supports the application of the Internet of Things (IoT), in which sensors can be focused on key parameters to support a real-time monitoring system and integration with a predictive model. In this respect, an early warning system can be developed to detect potential decline in DO, allowing for faster and more efficient interventions.

Overall, this study strengthens the understanding that managing DO in intensive aquaculture systems requires an integrated approach combining water-quality monitoring, predictive modelling, and IoT-based automation. This concept not only improves prediction accuracy and management effectiveness but also supports the transformation toward a more sustainable and smart aquaculture system.

CONCLUSIONS

This study demonstrates that dissolved oxygen (DO) dynamics in intensive aquaculture systems are primarily associated with chemical and biological water quality parameters, particularly ammonia (NH_3), hydrogen sulphide (H_2S), nitrite, and biochemical oxygen demand (BOD). Correlation analysis and feature importance results consistently indicate that these variables play a dominant role in explaining DO variability, whereas physical parameters such as temperature and pH exhibit relatively minor or statistically insignificant effects within the observed dataset. Importantly, the observed positive relationships between DO and ammonia, nitrite, and BOD should be interpreted with caution. These associations do not imply a direct causal increase in DO by these parameters. Instead, they likely reflect management-driven interactions, such as increased aeration and water exchange in response to elevated organic loading or nitrogenous waste. In intensive aquaculture systems, higher ammonia or BOD concentrations often trigger corrective management actions (e.g., aeration), which in turn elevate DO levels. Conversely, the negative relationship between DO and H_2S is ecologically intuitive, as H_2S accumulation is associated with anaerobic conditions and oxygen depletion.

The absence of a strong relationship between DO and pH further suggests that pH remained relatively stable across observations or was effectively buffered by system management practices. As a result, pH did not emerge as a key driver of DO variability in this study, despite its theoretical relevance in aquatic chemistry.

Model comparison results confirm that the random forest approach outperformed linear regression (R^2 = 0.515 vs. 0.470), highlighting the importance of non-linear modelling techniques for capturing the complex and indirect interactions among water quality variables. Nevertheless, the moderate predictive performance of both models indicates that DO dynamics cannot be fully explained by individual parameters alone, but rather by their combined and interacting effects.

From a practical perspective, these findings emphasise that effective DO management in intensive aquaculture should prioritise

monitoring and controlling biologically driven parameters, particularly ammonia, H_2S, nitrite, and BOD, while recognising that their relationships with DO are often indirect and mediated by management interventions. This clarification strengthens the interpretation of the results and supports the application of predictive models and IoT-based monitoring systems as decision-support tools for smart and sustainable aquaculture management. This study has some limitations that need to be acknowledged. The predictive accuracy of both linear regression and random forest models remained moderate, with R^2 values under 0.60, suggesting that the models could only partially capture dissolved oxygen (DO) dynamics. This limitation may be attributed to the restricted size (length x width) and depth of ponds, the point of sensor measurements, the aeration method, the variability of the dataset, and the absence of other potentially relevant predictors, such as chlorophyll-a, photosynthetic activity, or fish biomass. In addition, the study focused only on two modelling approaches, which may not fully represent the potential of advanced machine learning algorithms.

ACKNOWLEDGMENTS

The authors sincerely thank Politeknik Kelautan dan Perikanan Sorong (POLTEK KP Sorong) for the institutional support, facilities, and technical assistance that made this research possible

AUTHOR CONTRIBUTION

AY: conceptualisation, formal analysis, funding acquisition, investigation, resources, validation, writing, writing review, supervision, and editing; AWP: conceptualisation, formal analysis, writing, writing review, supervision, validation, and editing; HP: conceptualisation, resources, formal analysis, writing review, writing, funding acquisition, and editing; AF: conceptualisation, resources, formal analysis, writing, writing review, validation, and editing; R: conceptualisation, data curation, methodology, data analysis, visualisation, writing, writing review, software, validation, and editing. RS: methodology, writing review, and validation.

DECLARATION OF COMPETING INTEREST AND USE OF GENERATIVE AI

The authors declare no competing interests. Generative AI tools were used in this article solely to improve readability and grammar.

REFERENCES

Abdel-Tawwab, M., Monier, M. N., Hoseinifar, S. H., & Faggio, C. (2019). Fish response to hypoxia stress: Growth, physiological, and immunological biomarkers. *Fish Physiology and Biochemistry, 45*(3), 997–1013. https://doi.org/10.1007/s10695-019-00614-9

Akhtar, N., Ishak, M. I. S., Bhawani, S. A., & Umar, K. (2021). Various natural and anthropogenic factors responsible for water quality degradation: A review. *Water, 13*(19), 2660. https://doi.org/10.3390/w13192660

Barnes, R. S. K., & Mann, K. H. (2009). *Fundamentals of aquatic ecology*. John Wiley & Sons.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test, 25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Boyd, C. E., & McNevin, A. A. (2020). Aerator energy use in shrimp farming and means for improvement. *Journal of the Word Aquaculture Society*, *52*(1), 6–29. https://doi.org/10.1111/jwas.12753

Boyd, C. E., & Tucker, C. S. (2012). *Pond aquaculture water quality management*. Springer. https://doi.org/10.1111/jwas.12753

Chakravarty, S. P., Sinha, A., Baishya, S., & Roy, P. (2022). Robust control of water quality in intensive aquaculture using multi-variable quantitative feedback theory: From an Indian context. *Asian Journal of Control*, *25*(4), 2790-2807. https://doi.org/10.1002/asjc.2979

Das, R., & Behera, D. (2008). *Environmental science: Principles and practice*. PHI Learning.

Greig, S., Sear, D., & Carling, P. (2007). A review of factors influencing the availability of dissolved oxygen to incubating salmonid embryos. *Hydrological Processes, 21*(3), 323–334. https://doi.org/10.1002/hyp.6188

He, J., Chu, A., Ryan, M. C., Valeo, C., & Zaitlin, B. (2011). Abiotic influences on dissolved oxygen in a riverine environment. *Ecological Engineering, 37*(11), 1804–1814. https://doi.org/10.1016/j.ecoleng.2011.06.022

Heddam, S., & Kisi, O. (2017). Extreme learning machines: A new approach for modeling dissolved oxygen concentration with and without water quality variables as predictors. *Environmental Science and Pollution Research, 24*(20), 16702–16724. https://doi.org/10.1007/s11356-017-9283-z

Hodson, T. O. (2022). Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Jeong, J., Awosile, B., Thakur, K. K., Stryhn, H., Boyce, B., & Vanderstichel, R. (2024). Longitudinal dissolved oxygen patterns in Atlantic salmon aquaculture sites in British Columbia, Canada. *Frontiers in Marine Science, 10*, 1289375. https://doi.org/10.3389/fmars.2023.1289375

Jongjaraunsuk, R., Taparhudee, W., & Suwannasing, P. (2024). Comparison of water quality prediction for red tilapia aquaculture in an outdoor recirculation system using deep learning and a hybrid model. *Water, 16*(6), 907. https://doi.org/10.3390/w16060907

Kuang, L., Shi, P., Hua, C., Chen, B., & Zhu, H. (2020). An enhanced extreme learning machine for dissolved oxygen prediction in wireless sensor networks. *IEEE Access, 8*, 198730–198739. https://doi.org/10.1109/ACCESS.2020.3033455

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Lee, S. Y., Jeong, D. Y., Choi, J., Jo, S. K., Park, D. H., & Kim, J. G. (2024). LSTM model to predict missing data of dissolved oxygen in land-based aquaculture farm. *ETRI Journal, 46*(6), 1047–1060. https://doi.org/10.4218/etrij.2023-0337

Liu, M., Lian, Q., Zhao, Y., Ni, M., Lou, J., & Yuan, J. (2023a). Treatment effects of pond aquaculture wastewater using a field-scale combined ecological treatment system and associated microbial characteristics. *Aquaculture, 563*, 739018. https://doi.org/10.1016/j.aquaculture.2022.739018

Liu, W., Liu, S., Hassan, S. G., Cao, Y., Xu, L., Feng, D., Cao, L., Chen, W., Chen, Y., Guo, J., Liu, T., & Zhang, H. (2023b). A novel hybrid model to predict dissolved oxygen for efficient water quality in intensive aquaculture. *IEEE Access, 11*, 29162–29174. https://doi.org/10.1109/ACCESS.2023.3260089

Nagamora, J., Courage, S., Angeles, H., Vertudes, R., Ken, J., Balangao, B., Halil, A., & Abdullah, S. (2022). An assessment of the control and monitoring functionalities of a developed small-scale aquaculture system. *International Journal of Biosciences*, *21*(4), 89–100. https://doi.org/10.12692/ijb/21.4.89-100

Schäfer, N., Matoušek, J., Rebl, A., Stejskal, V., Brunner, R. M., Goldammer, T., Verleih, M., & Korytář, T. (2021). Effects of chronic hypoxia on the immune status of pikeperch (*Sander lucioperca*). *Biology, 10*(7), 649. https://doi.org/10.3390/biology10070649

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal, 20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Tong, C., He, K., & Hu, H. (2024). Design and application of a new aeration device based on a recirculating aquaculture system. *Applied Sciences*, *14*(8), 3401. https://doi.org/10.3390/app14083401

Veeramsetty, V. A., Rajeshwarrao, & Bernatin, T. (2024). *Aquaculture – Water quality dataset* (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/y78ty2g293.1

Vo, T. T. E., Ko, H., Huh, J. H., & Kim, Y. (2021). Overview of smart aquaculture system: Focusing on applications of machine learning and computer vision. *Electronics, 10*(22), 2882. https://doi.org/10.3390/electronics10222882

Wang, K., Gopaluni, R. B., Chen, J., & Song, Z. (2018). Deep learning of complex batch process data and its application on quality prediction. *IEEE Transactions on Industrial Informatics, 16*(12), 7233–7242. https://doi.org/10.1109/TII.2018.2880968

Xu, W., Yang, C.-E., Luo, Y., Zhang, K., Chen, M., Jiang, S., Grossart, H. P., & Luo, Z.-H. (2023). Distinct response of total and active fungal communities and functions to seasonal changes in a semi-enclosed bay with mariculture (Dongshan Bay, southern China). *Limnology and Oceanography, 68*(5), 1048–1063. https://doi.org/10.1002/lno.12328

Yang, H., Sun, M., & Liu, S. (2023). A hybrid intelligence model for predicting dissolved oxygen in aquaculture water. *Frontiers in Marine Science, 10*, 1126556. https://doi.org/10.3389/fmars.2023.1126556

Yao, X., Zhang, G., Song, Y., & Chen, Y. (2023). Adaptive anti-disturbance control of dissolved oxygen in circulating water culture systems. *Symmetry, 15*(11), 2015. https://doi.org/10.3390/sym15112015

Zhang, X., Zhang, Y., Zhang, Q., Liu, P., Guo, R., Jin, S., Liu, J., Chen, L., Ma, Z., & Liu, Y. (2020). Evaluation and analysis of water quality of marine aquaculture area. *International Journal of Environmental Research and Public Health, 17*(4), 1446. https://doi.org/10.3390/ijerph17041446

Zhang, Y., Fitch, P., Vilas, M. P., & Thorburn, P. J. (2019). Applying multi-layer artificial neural network and mutual information to the prediction of trends in dissolved oxygen. *Frontiers in Environmental Science, 7*, 46. https://doi.org/10.3389/fenvs.2019.00046