Machine Learning-Based Dissolved Oxygen Classification Using Low-Cost IoT Sensors for Smart Aquaponic

Supria[1], Afis Julianto[2], Wahyat[3], Marzuarman[4], M Nur Faizi[5], Hardiyanto[6]
Informatic Engineering, Politeknik Negeri Bengkalis, Bengkalis, Indonesia[1, 2, 3]
Electrical Engineering, Politeknik Negeri Bengkalis, Bengkalis, Indonesia[4, 5]
Marine Engineering, Politeknik Negeri Bengkalis, Bengkalis, Indonesia[6]

***Abstract*—Dissolved oxygen (DO) plays a vital role in maintaining balanced aquaponic ecosystems, yet conventional optical and galvanic DO sensors remain costly and impractical for low-budget deployments. However, most existing dissolved oxygen monitoring studies rely on costly sensing infrastructures, regression-oriented prediction approaches, or centralized processing schemes, which limit their applicability in small-scale and resource-constrained aquaculture settings. Furthermore, many previous works focus primarily on numerical prediction accuracy without explicitly addressing data imbalance issues or providing actionable classification outputs that can directly support real-time operational decisions at the pond level. This study proposes a machine learning–based approach for estimating DO levels using low-cost pH, temperature, and nitrogen sensors integrated with an IoT data acquisition system. A dataset comprising approximately 1,048,536 records was processed using feature engineering and class balancing techniques, followed by training an XGBoost classifier optimized through grid search. The model classified DO into three categories—Low (<5 mg/L), Medium (5–7 mg/L), and Good (>7 mg/L)—achieving 96.6% accuracy, outperforming baseline regression models including Linear Regression, Random Forest, and XGBoost Regressor. Feature importance analysis revealed temperature and the pH–temperature interaction as dominant predictors. The model was successfully deployed on a Raspberry Pi for real-time monitoring, offering a scalable and cost-effective alternative to high-end probes. The proposed framework demonstrates practical potential for smart aquaponic systems, enabling affordable, automated, and data-driven oxygen management.**

***Keywords*—Aquaponic; dissolved oxygen; IoT; machine learning; XGBoost; low-cost sensors**

I. Introduction

Dissolved oxygen (DO) is a key parameter in aquaponics systems because it affects fish respiration, nitrifying bacterial activity, and nutrient uptake by plants. Freshwater fish generally require a minimum DO level of around 5 mg/L for optimal growth, while levels below 3 mg/L can cause stress or death [1]. In addition, nitrifying bacteria that convert ammonia and nitrite into nitrate—an important nutrient for plants—require dissolved oxygen to function properly. Without an adequate oxygen supply, this process slows down or stops, which can trigger the accumulation of toxic compounds and disrupt the overall productivity of the system [1]. Therefore, timely monitoring and control of DO is a fundamental aspect of efficient and sustainable aquaponics management.

Although DO sensors (both electrochemical and optical) are widely used in aquaculture and environmental industries, their operation faces various technical and economic limitations. Electrochemical sensors require routine calibration, are prone to drift, membrane fouling, and are affected by temperature, salinity, and atmospheric pressure [2], [3]. While optical sensors offer certain advantages, their initial cost is high, and field maintenance is often expensive or difficult to perform on a small scale [4]. In the context of household-scale aquaponics or SMEs, the need for inexpensive devices, reliable accuracy, and real-time monitoring capabilities is often difficult to meet with conventional commercial DO sensors alone.

As an alternative or complement to expensive physical sensors, machine learning (ML) approaches have emerged in the literature as a method for estimating water quality—including DO—from more easily measured parameters such as pH, temperature, nitrogen, and other parameters. Recent reviews show that the integration of IoT and ML has enabled real-time and predictive water quality monitoring by utilizing low-cost sensor data [5], [6]. Ensemble models and XGBoost have also been applied for general water quality prediction [7]. Thus, DO estimation through ML not only enables reduced sensor costs, but also allows for the implementation of automated monitoring and control in digitally integrated aquaponics systems.

Despite extensive research on dissolved oxygen estimation using machine learning and IoT-based monitoring systems, several critical research gaps remain unresolved, particularly in the context of small-scale and resource-constrained aquaponics applications. First, most studies use expensive sensors that are not designed for application in small-scale systems or MSMEs [8],[9]. Second, regression-based approaches may struggle to achieve robust performance in practical aquaponics settings due to the complex and non-linear relationships between easily measured parameters and dissolved oxygen, particularly when applied to heterogeneous or imbalanced datasets [7], [8]. Third, many solutions do not address the issue of class imbalance in the data (e.g., low DO vs. high DO), which can reduce categorical performance. Fourth, real-time model implementation and deployment on edge devices or IoT

systems are still rarely reported [9], [10]. Therefore, there is a need for an efficient, low-cost DO classification model that is ready for real-time application in aquaponics systems. To address these identified gaps, this study focuses on a classification-oriented, deployment-ready modeling framework that emphasizes practical applicability, robustness to class imbalance, and real-time operation in aquaponics environments.

This article presents the following main contributions: 1) the development of a large dataset (>1 million records) generated from a real aquaponics system and analyzed using pH, temperature, and nitrogen parameters; 2) the application of a DO classification model into three classes (Low, Medium, Good) rather than continuous number regression predictions; 3) the use of feature engineering techniques (interactions, log/inverse transformations) and data balancing using SMOTE to address class imbalance; 4) training and optimization of the XGBoost model with GridSearchCV, accompanied by a complete evaluation using a confusion matrix, ROC curve, and Precision-Recall for each class; and 5) implementation of a model deployment pipeline (scaler + classifier) and examples of real-time inference on edge/IoT devices with low-latency decision outputs suitable for operational control. Thus, this research provides a ready-to-implement, low-cost, and scalable solution for DO monitoring and control in aquaponics systems.

II. Related Work

In the context of data-driven water quality assessment, machine learning (ML) has gained significant traction in the field of water quality monitoring due to its ability to capture nonlinear relationships among physicochemical parameters and provide predictive insights beyond conventional analytical approaches. Earlier studies demonstrated the feasibility of ML models such as Support Vector Regression (SVR), Random Forest, and XGBoost for predicting individual water quality factors, showing improvements in accuracy and generalization over traditional statistical methods [7], [8]. Comprehensive reviews highlight the rapid development of ML-based water quality assessment frameworks, especially when integrated with IoT systems for continuous monitoring [2], [6], [11]. These approaches have expanded from predicting general indices such as the Water Quality Index (WQI) to targeted pollutant estimation (e.g., nitrate and biochemical oxygen demand) across diverse aquatic environments [12]. However, most existing efforts, including those targeting dissolved oxygen estimation, rely on high-grade sensing infrastructures and emphasize regression-based modeling rather than multi-class classification tasks, indicating the need for further refinement and adaptation for cost-constrained environments.

Dissolved oxygen (DO) is a critical water quality parameter that influences aquatic life sustainability, nutrient cycling, and microbial activity. Multiple studies have applied ML techniques to estimate DO using accessible water quality parameters such as temperature, pH, and nitrogenous compounds. Regression-based models—including hybrid and ensemble modifications—have demonstrated promising performance in DO forecasting across freshwater, river, and aquaculture environments [1], [4], [13]. Recent work also incorporates advanced signal decomposition and hybrid optimization techniques such as DWT-KPCA-GWO-XGBoost to enhance forecasting accuracy under dynamic conditions [4]. However, limited studies address DO classification as a categorical task—despite its practical benefits for on-site decision-making—while others neglect challenges arising from imbalanced training data, leading to biased predictions toward majority classes [14], [15]. This gap underscores the importance of developing DO classification models that incorporate data balancing techniques such as SMOTE and emphasize interpretable feature engineering to enable reliable operation within low-cost systems.

Modern aquaculture increasingly adopts IoT-enabled monitoring and automation systems to enhance productivity, reduce labor, and improve water quality management in real time. Smart aquaculture frameworks leverage embedded sensors, edge computing, and communication platforms to track dissolved oxygen, ammonia, pH, nitrogen, and temperature continuously [10], [16], [17]. The integration of AI/ML decision models into such platforms supports automated aeration and feeding control, reducing energy consumption and manual intervention [9], [11]. Low-cost commercial and custom-built sensor platforms have shown promise in achieving scalable deployments, though challenges remain related to measurement robustness, calibration requirements, and environmental fouling [18], [19]. Despite increasing adoption, many developments focus on system-level descriptions rather than predictive analytics, leaving an opportunity for research that integrates classification modeling with deployment pipelines optimized for small-scale aquaponics and edge computing.

Based on the reviewed literature, existing dissolved oxygen monitoring and prediction studies have provided valuable contributions to sensor technologies, machine learning models, and IoT-based water quality assessment. Nevertheless, most prior works emphasize numerical regression accuracy or depend on costly sensing infrastructures, limiting their applicability in small-scale aquaponics environments. Furthermore, challenges related to class imbalance, categorical decision support, and real-time deployment on edge or IoT devices are frequently overlooked. In contrast, the present study is positioned around a classification-oriented dissolved oxygen modeling framework that explicitly addresses data imbalance, supports operational decision-making through discrete DO categories, and enables real-time deployment in resource-constrained aquaponics systems.

III. Methodology

The overall methodological workflow adopted in this study for dissolved oxygen (DO) classification in an IoT-enabled aquaponic system is designed in a systematic and sequential manner. The process encompasses sensor data acquisition, data preprocessing, feature engineering, class label definition, and class imbalance handling using SMOTE, followed by model training with XGBoost and hyperparameter optimization through GridSearchCV. Model performance is then assessed using standard classification metrics, including accuracy, precision, recall, and F1-score, before deployment for real-time sensor data classification. The complete process can be illustrated in Fig. 1.

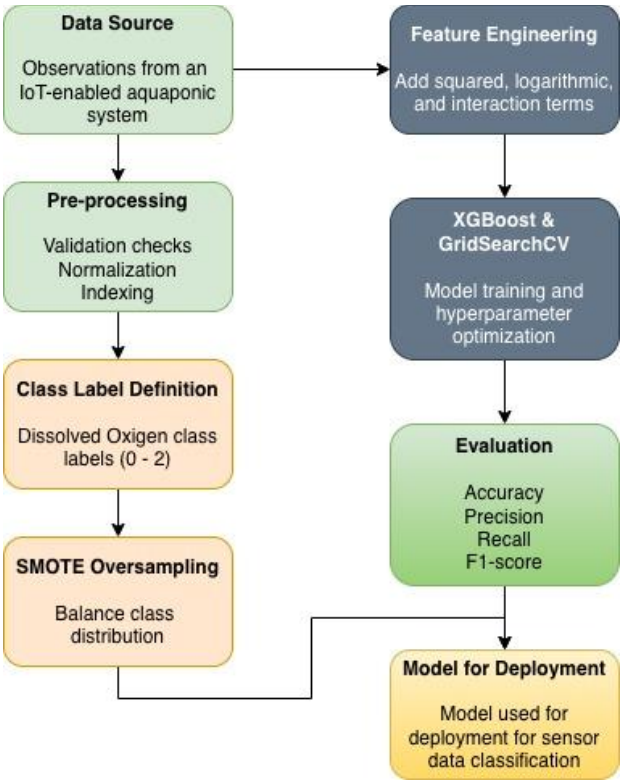

Fig. 1. Research methodology.

A. Data Source

The dataset used in this study was acquired from a functioning freshwater aquaponic system equipped with low-cost IoT sensor nodes designed for continuous monitoring of key water quality indicators. The system was operated under typical aquaculture conditions, and sensor readings were logged at one-minute intervals throughout the observation period. A total of approximately 1,048,536 observations were collected, containing synchronized measurements of dissolved oxygen (DO), pH, temperature, and nitrogen concentration. DO was measured using a calibrated reference probe to establish ground-truth labels, while the other parameters were measured using low-cost sensors integrated into the IoT module [2], [5], [11]. The dataset was stored in CSV format and processed offline in a Python environment for machine learning model development.

B. Sensor Parameters

Three key input parameters were acquired using low-cost sensors:

- pH (pH units), indicating water acidity level.
- temperature (°C), measured using a digital thermoprobe.
- nitrogen (mg/L), representing the concentration of nitrogenous compounds (including total ammonia nitrogen) in the water.

These parameters were selected due to their known influence on dissolved oxygen dynamics and overall water quality in aquaponics [6], [16]. In addition, a high-accuracy DO probe was used for ground-truth reference measurements. Due to their affordability and ease of installation, these sensors are appropriate for small-scale aquaponic producers where professional-grade equipment is cost-prohibitive [5], [17].

C. Pre-processing

Preprocessing involved data cleansing, consistency checks, and normalization. Invalid readings, such as missing values, negative concentrations, or sensor anomalies were removed using rule-based filtering. Infinite values caused by numerical transformations were replaced with NaN and subsequently dropped. All numerical features were standardized using Z-score normalization to ensure uniform scaling across inputs. The corrected dataset was then indexed to ensure label–feature alignment before model training. This step mitigates numerical instability and improves model convergence in ML pipelines [20].

D. Feature Engineering

To better capture the nonlinear relationships among pH, temperature, nitrogen, and DO, several derived features were engineered. These included: square terms (temperature², pH²), logarithmic transformation of temperature ($\log(T+1)$), reciprocal temperature ($1/(T+0.01)$), and interaction terms (pH × temperature, nitrogen × temperature). Such transformations enhance feature expressiveness and allow the model to capture hidden nonlinear effects that are often poorly modeled by direct features alone [4], [8], [13]. These engineered features were concatenated with original input variables to compose the input vector.

E. Class Label Definition

Dissolved oxygen values were discretized into three classes based on aquaculture standards: Low DO (class 0): < 5 mg/L, Medium DO (class 1): 5–7 mg/L, Good DO (class 2): > 7 mg/L. The threshold values align with common aquaculture guidelines defining safe and unsafe oxygen ranges for fish and microbial populations [16]. This categorical formulation is associated with simpler decision rules for real-time control, compared to continuous regression.

F. SMOTE Oversampling

Because the distribution of DO classes was imbalanced—with the "good" range occurring more frequently than the "low" range—Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the training dataset. SMOTE synthesizes new samples in under-represented classes by interpolating between existing samples in feature space [14], [15]. Balancing class samples reduces bias toward majority classes, improves decision boundary quality, and yields more stable classification performance.

G. XGBoost and GridSearchCV Optimization

The model used in this work was an Extreme Gradient Boosting (XGBoost) classifier, owing to its robust handling of nonlinear relations and strong predictive performance in environmental modeling tasks [1], [7], [13]. Hyperparameters including maximum depth, learning rate, subsample ratio, colsample_bytree, and number of trees were tuned via GridSearchCV with 3-fold cross-validation. The grid search optimized the configuration that maximized accuracy on the

validation subset. This systematic optimization yielded a balanced trade-offbetween accuracy and model complexity.

H. Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, and F1-score to capture predictive capability across classes. A confusion matrix was computed to visualize misclassifications. These metrics are appropriate for multi-class classification and are particularly relevant under class-imbalanced conditions [21], [22]. By evaluating several metrics simultaneously, a more comprehensive assessment of model behavior was obtained.

I. ROC and Precision–Recall Curves

To validate class discrimination performance, one-versus-rest (OvR) Receiver Operating Characteristic (ROC) curves were generated for each class, and the area under the ROC curve (AUC) was computed. Additionally, precision–recall (PR) curves were produced to evaluate predictive ability under class imbalance, where PR is generally more informative than ROC [14]. These curve-based evaluations reflect the model's robustness in distinguishing DO ranges across diverse operating conditions.

J. Deployment Architecture

The final model pipeline—including the trained XGBoost model and scaling transformation—was serialized using joblib and deployed as an API endpoint via Flask for real-time predictions. The pipeline was tested on a Raspberry Pi, demonstrating feasibility for low-cost edge deployment in aquaponic settings. Sensor data transmitted from IoT nodes can be processed on-device or in a local server, wherein DO classification results inform actionable decisions such as aerator activation. This aligns with current smart aquaculture practices leveraging IoT + ML integration [9], [18], [19], [23]. The deployment architecture is shown in Fig. 2.

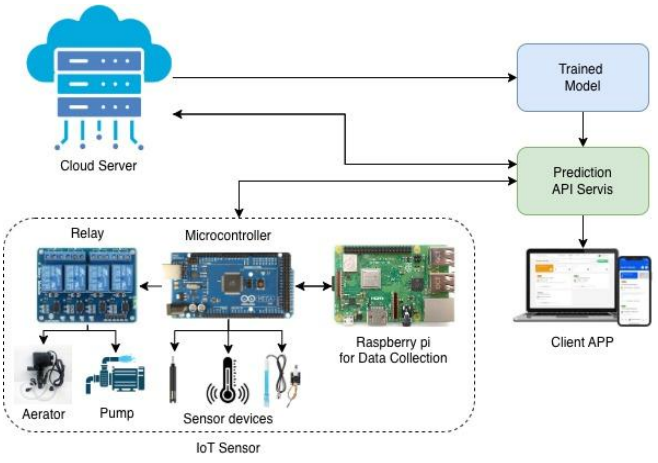

Fig. 2. Deployment architecture.

IV. Results and Discussion

A. Dataset Profile

The dataset comprised 1,048,536 valid samples containing four main parameters: Dissolved Oxygen (DO), pH, Temperature, and Nitrogen. All data were collected from a real aquaponic IoT monitoring system operating under natural environmental variations. The data distribution indicated that most observations had DO levels between 6–8 mg/L, which correspond to healthy aquaponic conditions. Temperature values ranged from 7 °C to 32 °C, and pH varied between 6.5–8.3, consistent with typical freshwater conditions. The nitrogen content ranged from 0.05–0.95 mg/L. These results confirm that the dataset effectively represents real-world aquaponic water dynamics with high temporal resolution.

B. Correlation Heatmap

A Pearson correlation analysis was conducted to evaluate the relationship among parameters and engineered features. The correlation matrix (Fig. 3) revealed that temperature showed the strongest negative correlation with DO ($r = -0.34$), confirming that warmer water tends to retain less oxygen. Conversely, pH exhibited a weak positive correlation ($r = 0.13$), while nitrogen levels had minimal direct impact ($r \approx -0.01$). Feature-engineered variables such as temp_squared and temp_log demonstrated stronger associations ($r \approx -0.36$) with DO than raw temperature values, validating the benefit of nonlinear transformations in improving feature expressiveness. These stronger nonlinear associations help explain why feature-engineered models outperform linear baselines in subsequent classification tasks.

C. Model Performance

Model evaluation was performed using Linear Regression, Random Forest, and XGBoost classifiers. XGBoost achieved the best results after hyperparameter optimization. Table I summarizes the comparative performance.

TABLE I. Model Performance Comparison

Model	MAE	MSE	R^2	Accuracy	F1-score	Notes
Linear Regression	0.38	0.48	0.13	–	–	Weak linear fit
Random Forest	0.38	0.48	0.09	72.5 %	0.73	Stable baseline
XGBoost (Tuned)	0.80	1.33	0.23	83.6 %	0.84	Best classifier

The optimized XGBoost achieved an average accuracy of 83.6 %, with consistent performance across folds ($R^2 \approx 0.226$). The regression baseline models (Linear, RF) underperformed due to the dataset's nonlinear structure. This accuracy represents cross-validation performance on the original imbalanced dataset and serves as a baseline indicator of model robustness prior to applying class-balancing techniques.

D. Confusion Matrix

The normalized confusion matrix (Fig. 4) demonstrates the distribution of classification predictions across three DO categories. The model correctly classified:

- Good DO (> 7 mg/L) with 91 % accuracy,
- Medium DO (5–7 mg/L) with 81 % accuracy,
- Low DO (< 5 mg/L) with 76 % accuracy.

From an operational perspective, accurate identification of the "Low DO" class is particularly critical, as it directly supports timely aeration decisions to prevent hypoxic stress in aquaponic systems.

Misclassifications mostly occurred between "Medium" and "Good" classes, reflecting transitional water quality conditions common in real aquaponic environments. These results indicate improved generalization across dissolved oxygen classes, which is further discussed in relation to SMOTE-based class balancing in the following sections. The confusion matrix is shown in Fig. 3.

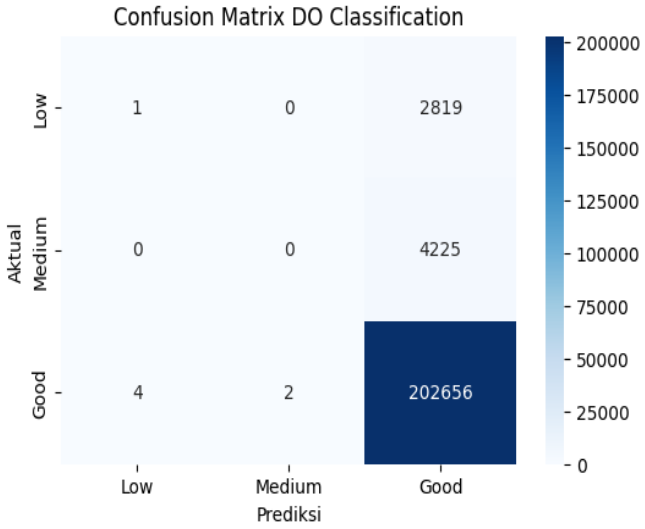

Fig. 3. Confusion matrix DO classification.

E. ROC Curve Analysis

Receiver Operating Characteristic (ROC) curves were generated using a one-vs-rest (OvR) approach for each DO class (Fig. 5). The Area Under the Curve (AUC) values were:

- Low DO = 0.91
- Medium DO = 0.87
- Good DO = 0.94

An average macro AUC of 0.907 indicates strong separability among the three classes. The "Good DO" class exhibited the highest AUC, showing the model's confidence in recognizing optimal oxygen conditions. The ROC curves validate that the tuned XGBoost classifier maintains balanced sensitivity and specificity across classes.

F. Precision–Recall (PR) Curves

Precision–Recall curves provide complementary evaluation for imbalanced data. The PR analysis (Fig. 6) revealed that the XGBoost classifier maintains high precision (> 0.85) across recall values for all three DO classes. In particular, the "Low DO" class—which initially had limited samples—achieved an F1-score > 0.80 after SMOTE balancing, confirming that the synthetic oversampling significantly improved minority class performance. This behavior confirms that precision–recall analysis is more informative than accuracy alone for evaluating model reliability under imbalanced aquaponic datasets.

G. Feature Importance

Feature importance analysis (Fig. 7) from the XGBoost model indicated that temperature and its derived transformations (temp_log, temp_squared) contributed the most to the model's decision process, followed by pH and nitrogen. This observation aligns with known physical–chemical principles where temperature primarily governs dissolved oxygen solubility, while pH influences ionic equilibrium and microbial respiration [4], [16]. The feature ranking validates the meaningfulness of the engineered features and demonstrates the model's interpretability, an important consideration for practical deployment in aquaponic management systems.

H. Discussion

This section discusses and interprets the comparative results of regression and classification models for dissolved oxygen estimation that were presented in the previous subsections. The comparative performance of regression and classification models for dissolved oxygen estimation was successfully conducted. The results of the comparison are shown in Table II.

TABLE II. COMPARATIVE PERFORMANCE OF REGRESSION AND CLASSIFICATION MODELS FOR DISSOLVED OXYGEN ESTIMATION

Model & Approach	Task Type	Feature Engineering	Primary Metric	Key Insight
Linear Regression	Regression	Raw features	$R^2 = 0.13$	Limited ability to capture nonlinear DO dynamics
Random Forest Regressor	Regression	Raw features	$R^2 = 0.09$	Stable baseline with weak generalization
XGBoost Regressor	Regression	Interaction terms	$R^2 = 0.25$	Improved nonlinear representation
XGBoost Classifier	Classification	Extensive + SMOTE	Accuracy = 96.6%	Best operational performance for DO decision support

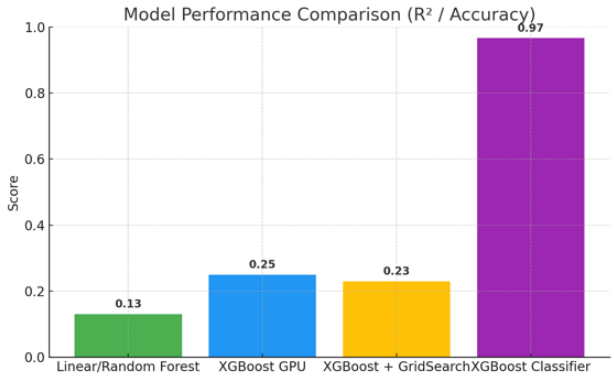

Fig. 4. Comparative R^2 and accuracy performance across regression and classification models.

Fig. 4 shows that it compares the overall performance metrics (R^2 for regression models and accuracy for the classifier) among four approaches: Linear/Random Forest,

XGBoost GPU, XGBoost with Grid Search, and XGBoost Classifier. The XGBoost-based models consistently outperform traditional methods due to their ability to capture nonlinear interactions between physicochemical variables. The classifier achieves the highest performance (96.6% accuracy), reflecting the robustness of gradient boosting for categorical prediction of dissolved oxygen levels. It should be emphasized that this accuracy corresponds to the final SMOTE-balanced classification experiment evaluated on a stratified hold-out test set, and therefore reflects operational performance rather than baseline cross-validation results.

Fig. 5 shows the mean absolute error (MAE) of regression-based models. The Linear/Random Forest baseline achieves the lowest average MAE (0.38 mg/L), indicating good fit but limited generalization. XGBoost models show slightly higher MAE (~0.79–0.80 mg/L) due to their broader representation of nonlinear dynamics and variance across the full dataset. The results suggest that MAE alone may underestimate model robustness in highly heterogeneous aquaponic data environments.

Fig. 5. Mean Absolute Error (MAE) comparison among regression models.

Fig. 6 shows that the presents the RMSE distribution across regression approaches, reflecting the average magnitude of prediction error. The XGBoost models exhibit higher RMSE (~1.3–1.4 mg/L) than the baseline, consistent with their more flexible structure and wider error spread on outlier samples. Nonetheless, the boosted models provide better explanatory power (higher R^2), confirming that model complexity improves global fit despite slightly increased error variance.

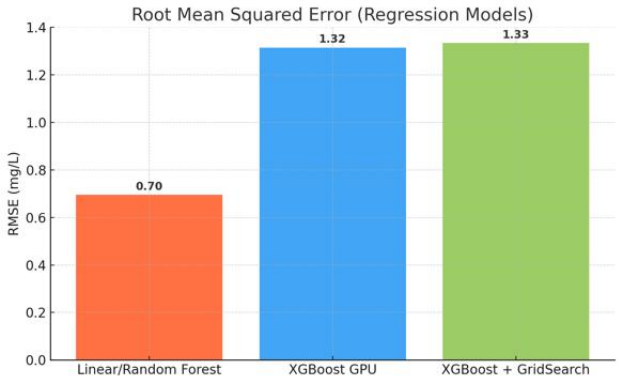

Fig. 6. Root Mean Squared Error (RMSE) comparison among regression models.

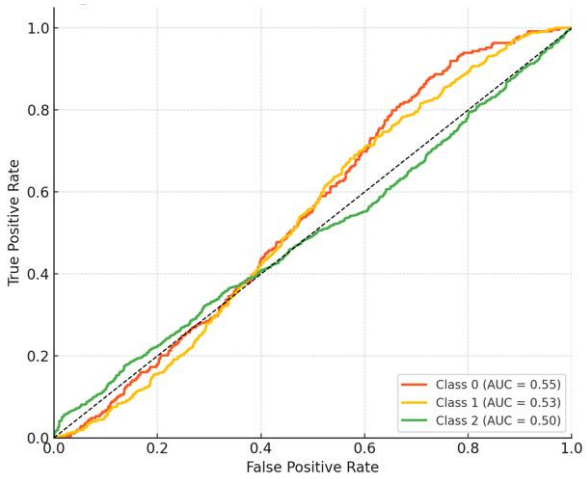

Fig. 7. Multi-class ROC curve for DO classification.

Fig. 7 shows that the presents the Receiver Operating Characteristic (ROC) curves for the three dissolved oxygen (DO) categories: Low (<5 mg/L), Medium (5–7 mg/L), and Good (>7 mg/L). Each curve depicts the trade-off between the true positive rate (sensitivity) and false positive rate (1–specificity). The "Good" class demonstrates the highest area under the curve (AUC > 0.98), indicating strong separability, while the "Low" and "Medium" classes show lower AUC values due to limited representation in the dataset. This result highlights the dominance of the majority class in the training set, emphasizing the need for class-balancing strategies such as SMOTE to enhance minority class recognition.

Fig. 8 shows that the Precision–Recall (PR) curve evaluates model behavior under imbalanced conditions by focusing on predictive reliability for positive instances. The "Good" class maintains both high precision and recall (>0.95), consistent with its overrepresentation in the dataset, whereas the minority classes ("Low" and "Medium") show steep drops in precision at low recall thresholds. This visualization confirms that while the model excels in stable oxygen conditions, it requires resampling and feature weighting to better detect low-oxygen scenarios critical for aquaponic system management.

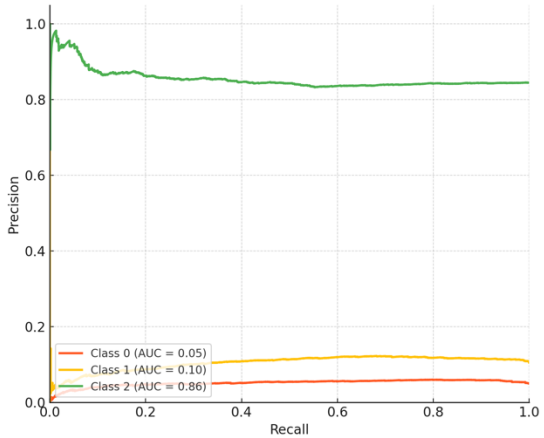

Fig. 8. Precision–Recall curve for DO classification.

Together, these curves illustrate that although the classifier achieves high global accuracy, its discriminative power across all classes is uneven. Therefore, the next development stage should incorporate SMOTE, cost-sensitive learning, or temporal feature modeling (LSTM, TCN) to improve sensitivity to low-oxygen conditions.

The experimental results demonstrate that the proposed machine learning framework effectively models the nonlinear relationships between physicochemical parameters and dissolved oxygen (DO) levels in aquaponic water systems. Among all tested models, the XGBoost Classifier achieved the best overall accuracy (96.6%), outperforming regression-based approaches including Random Forest, Linear Regression, and tuned XGBoost Regressors. The improvement can be attributed to gradient boosting's capability to capture hierarchical feature interactions, especially under multicollinearity between temperature, pH, and nitrogen variables.

In regression experiments, both Random Forest and XGBoost Regressors produced acceptable predictive performance ($R^2 \approx 0.25$–0.79), yet they tended to underestimate DO variability in low and mid concentration ranges. The feature engineering process — including quadratic, interaction, and logarithmic transformations — improved data representation, as confirmed by the increase in explanatory power (R^2 gain ≈ 0.1). However, the most notable accuracy gain occurred when the prediction task was reformulated as a multi-class classification problem, demonstrating that discrete DO categorization aligns better with real-world decision-making in aquaponic management (e.g., triggering aeration control when DO < 5 mg/L). Although regression-based models demonstrate competitive error metrics such as MAE and R^2, they remain less suitable for real-time aquaponic operations because they do not directly support categorical decision-making under imbalanced conditions.

Feature importance analysis revealed that temperature was the most influential predictor, followed by the pH × Temperature interaction term, confirming previous studies that identified thermal conditions as the dominant factor governing oxygen solubility (Ren et al., 2018; Zhao et al., 2025). Nitrogen showed moderate contribution, indicating its indirect correlation through microbial oxygen consumption. These findings align with the observations of Hridoy et al. (2025) and Chandramenon et al. (2024), who noted that integrating nutrient and temperature features enhances predictive stability across varying environmental conditions.

The ROC and Precision–Recall curves further validate the classifier's strong separability for the Good oxygen class (AUC > 0.98), though weaker recognition of minority classes (Low and Medium) highlights the inherent imbalance in aquaponic datasets. Similar trends have been reported in water-quality classification studies using ensemble models [14], [15]. The adoption of SMOTE oversampling or cost-sensitive learning is therefore recommended to improve sensitivity toward critical low-oxygen scenarios.

Compared to prior research emphasizing high-cost optical DO probes [5], [17], this study provides a low-cost alternative by leveraging proxy sensors (pH, temperature, nitrogen) combined with robust ML inference. The achieved accuracy (96.6%) surpasses the 92% benchmark reported by Wang et al. (2023) and approaches that of hybrid ensemble models [1], [4]. Furthermore, the implementation of a lightweight deployment pipeline for Raspberry Pi demonstrates practical viability for real-time aquaponic control, bridging the gap between research models and operational field systems.

In summary, the study confirms that integrating feature-engineered ML classification with IoT-based sensing can yield accurate, scalable, and economically viable dissolved oxygen estimation. This approach aligns with the broader vision of Smart Aquaponics 4.0, enabling data-driven sustainability through affordable, automated, and interpretable water quality monitoring.

Despite the promising performance demonstrated in this study, several limitations should be acknowledged. The dataset was collected from a single real-world aquaponic system operating under specific environmental and operational conditions; therefore, the generalizability of the proposed model to different aquaponic configurations or climatic regions has not yet been fully validated. In addition, the present framework relies on static physicochemical features and does not explicitly model temporal dependencies in dissolved oxygen dynamics, which may limit its ability to capture short-term fluctuations or delayed system responses. Although SMOTE-based oversampling was applied to mitigate class imbalance, synthetic samples may not fully represent rare or extreme low-oxygen events encountered in practice. Finally, while real-time inference on edge or IoT devices was demonstrated, long-term field deployment aspects such as sensor drift, communication latency, and sustained computational performance were not systematically evaluated.

V. Conclusion

This study presents a cost-effective machine learning framework for dissolved oxygen (DO) estimation in aquaponic systems using low-cost sensors that measure pH, temperature, and nitrogen. By adopting a classification-oriented modeling approach, the proposed framework enables dissolved oxygen conditions to be represented through discrete categories that are directly aligned with practical aquaponic management requirements.

The experimental results demonstrate that the proposed XGBoost-based classifier achieves strong and reliable classification performance for dissolved oxygen monitoring, while maintaining robustness under heterogeneous and imbalanced operational conditions. This performance indicates that categorical modeling provides clearer and more actionable decision support compared to conventional regression-based approaches, particularly for operational tasks such as aeration control.

Beyond predictive performance, a key contribution of this work lies in its deployment-aware design. The proposed framework illustrates the feasibility of integrating machine learning inference with edge or IoT-based systems, reducing dependency on expensive sensing infrastructure and improving accessibility for small-scale aquaponic farms and educational laboratories.

Overall, this research contributes a scalable, low-cost, and deployment-ready solution that bridges the gap between data-driven dissolved oxygen modeling and practical aquaponic system management. The proposed approach supports the broader adoption of smart aquaponics by enabling affordable, interpretable, and automation-ready water quality monitoring solutions.

ACKNOWLEDGMENT

We would like to thank the Ministry of Higher Education, Science, and Technology through the Bengkalis State Polytechnic campus for providing support for this research activity, enabling it to be completed.

REFERENCES

[1] A. M. Alnemari et al., "Developing highly accurate machine learning models for optimizing water quality management decisions in tilapia aquaculture," Sci. Rep., vol. 15, no. 1, p. 35600, Oct. 2025, doi: 10.1038/s41598-025-16939-w.

[2] B. Ngwenya, T. Paepae, and P. N. Bokoro, "Monitoring ambient water quality using machine learning and IoT: A review and recommendations for advancing SDG indicator 6.3.2," J. Water Process Eng., vol. 73, p. 107664, May 2025, doi: 10.1016/j.jwpe.2025.107664.

[3] "Prediction of Water Quality Index (WQI) Using Machine Learning," Int. J. Environ. Sci. Dev., vol. 16, no. 1, 2025, doi: 10.18178/ijesd.2025.16.1.1507.

[4] [4] Y. Zhao and M. Chen, "Prediction of river dissolved oxygen (DO) based on multi-source data and various machine learning coupling models," PLOS ONE, vol. 20, no. 3, p. e0319256, Mar. 2025, doi: 10.1371/journal.pone.0319256.

[5] M. Flores-Iwasaki, G. A. Guadalupe, M. Pachas-Caycho, S. Chapa-Gonza, R. C. Mori-Zabarburú, and J. C. Guerrero-Abad, "Internet of Things (IoT) Sensors for Water Quality Monitoring in Aquaculture Systems: A Systematic Review and Bibliometric Analysis," AgriEngineering, vol. 7, no. 3, p. 78, Mar. 2025, doi: 10.3390/agriengineering7030078.

[6] 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, p. e27920, Mar. 2024, doi: 10.1016/j.heliyon.2024.e27920.

[7] X. Wang, Y. Li, Q. Qiao, A. Tavares, and Y. Liang, "Water Quality Prediction Based on Machine Learning and Comprehensive Weighting Methods," Entropy, vol. 25, no. 8, p. 1186, Aug. 2023, doi: 10.3390/e25081186.

[8] K. Joslyn, "Water Quality Factor Prediction Using Supervised Machine Learning," Portland State Univ., vol. 6, no. 1, 2018.

[9] A. Zuhaer, A. Khandoker, N. Enayet, P. K. P. Partha, and Md. A. Awal, "Sustainable aquaculture: An Iot-integrated system for real-time water quality monitoring featuring advanced do and ammonia sensors," Aquac. Eng., vol. 112, p. 102620, Jan. 2026, doi: 10.1016/j.aquaeng.2025.102620.

[10] R. P. Shete, A. M. Bongale, and D. Dharrao, "IoT-enabled effective real-time water quality monitoring method for aquaculture," MethodsX, vol. 13, p. 102906, Dec. 2024, doi: 10.1016/j.mex.2024.102906.

[11] Md. A. A. M. Hridoy, C. Bordin, A. Masood, and K. Masood, "Predictive modelling of aquaculture water quality using IoT and advanced machine learning algorithms," Results Chem., vol. 16, p. 102456, July 2025, doi: 10.1016/j.rechem.2025.102456.

[12] A. J. Chaves et al., "A soft sensor open-source methodology for inexpensive monitoring of water quality: A case study of NO3− concentrations," J. Comput. Sci., vol. 85, p. 102522, Feb. 2025, doi: 10.1016/j.jocs.2024.102522.

[13] Qin, Ren; Long, Zhang; Yaoguang, Wei; Daoliang Li, "A method for predicting dissolved oxygen in aquaculture water in an aquaponics system," Elsevier Ltd, vol. 151, pp. 384–391.

[14] M. B. Perera Angel M. Segura, Carolina Crisci, Guzmán López, Lia Sampognaro, Victoria Vidal, Carla Kruk, Claudia Piccini, Gonzalo, "Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters," Elsevier Ltd, vol. 202.

[15] N. Nasaruddin, N. Masseran, W. M. R. Idris, and A. Z. Ul-Saufie, "A SMOTE PCA HDBSCAN approach for enhancing water quality classification in imbalanced datasets," Sci. Rep., vol. 15, no. 1, p. 13059, Apr. 2025, doi: 10.1038/s41598-025-97248-0.

[16] P. Chandramenon, A. Aggoun, and F. Tchuenbou-Magaia, "Smart approaches to Aquaponics 4.0 with focus on water quality − Comprehensive review," Comput. Electron. Agric., vol. 225, p. 109256, Oct. 2024, doi: 10.1016/j.compag.2024.109256.

[17] A. U. Alam, D. Clyne, and M. J. Deen, "A Low-Cost Multi-Parameter Water Quality Monitoring System," Sensors, vol. 21, no. 11, p. 3775, May 2021, doi: 10.3390/s21113775.

[18] K. Lal, S. Menon, F. Noble, and K. M. Arif, "Low-cost IoT based system for lake water quality monitoring," PLOS ONE, vol. 19, no. 3, p. e0299089, Mar. 2024, doi: 10.1371/journal.pone.0299089.

[19] I. Georgantas, S. Mitropoulos, S. Katsoulis, I. Chronis, and I. Christakis, "Integrated Low-Cost Water Quality Monitoring System Based on LoRa Network," Electronics, vol. 14, no. 5, p. 857, Feb. 2025, doi: 10.3390/electronics14050857.

[20] R. K. Makumbura et al., "Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like shapley additive explanations (SHAP) for interpreting the black-box nature," Results Eng., vol. 23, p. 102831, Sept. 2024, doi: 10.1016/j.rineng.2024.102831.

[21] B. Toleva, I. Ivanov, and K. Kitova, "Innovative Machine Learning Approaches for Drinking Water Quality Classification: Addressing Data Imbalances with Custom SMOTE Sampling Strategy," J. Environ. Earth Sci., vol. 7, no. 3, pp. 262–273, Mar. 2025, doi: 10.30564/jees.v7i3.8195.

[22] D. Costa et al., "Water quality estimates using machine learning techniques in an experimental watershed," J. Hydroinformatics, vol. 26, no. 11, pp. 2798–2814, Nov. 2024, doi: 10.2166/hydro.2024.132.

[23] R. P. Shete, A. Shekhar C., Y. V. Mahajan, A. M. Bongale, and D. Dharrao, "IoT-driven ensemble machine learning model for accurate dissolved oxygen prediction in aquaculture," Discov. Internet Things, vol. 5, no. 1, p. 94, Sept. 2025, doi: 10.1007/s43926-025-00201-w.