INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING
ISSN: 2229-838X e-ISSN: 2600-7916

IJIE

Vol. 17 No. 9 (2025) 71-84
https://publisher.uthm.edu.my/ojs/index.php/ijie

Long-term Continuous Monitoring of Dissolved Oxygen Concentration Based on Multispectral Sensors and Machine Learning

Nhut-Thanh Tran[1]*, Chanh-Nghiem Nguyen[1], Quoc-Hung Pham[2,3], Chi-Ngon Nguyen[1]

[1] *Faculty of Automation Engineering, College of Engineering, Can Tho University, Can Tho city, 94000, VIETNAM*
[2] *Department of Computer Engineering, University of Information Technology, Ho Chi Minh city, 720400, VIETNAM*
[3] *Vietnam National University, Ho Chi Minh city, 720400, VIETNAM*

*Corresponding Author: nhutthanh@ctu.edu.vn
DOI: https://doi.org/10.30880/ijie.2025.17.09.006

Article Info

Received: 3 May 2025
Accepted: 5 December 2025
Available online: 31 December 2025

Keywords

Dissolved oxygen monitoring, multispectral sensor, machine learning, visible and near infrared spectroscopy

Abstract

Dissolved oxygen (DO) is a crucial indicator of water quality and requires continuous monitoring across various applications. Although optical sensors are widely used for DO measurement, their large-scale deployment for long-term monitoring remains challenging due to high costs and the need for periodic replacement of sensing probes. This study presents a novel non-contact DO monitoring system that integrates a low-cost multispectral sensor with an optimized machine learning framework, offering a practical solution for long-term, continuous monitoring. A compact spectroscopic sensing unit was developed to continuously acquire absorbance data from water samples across 18 wavebands ranging from 410 to 940 nm. Multiple machine learning models were trained under different configurations, and several waveband selection algorithms were applied to identify the optimal predictive model. The neural network regression model utilizing four wavebands (460, 585, 680, and 760 nm) achieved the best result with a coefficient of determination and a root mean square error of 0.99 and 0.22 mg/L, respectively. These findings demonstrate the high accuracy and practical potential of the proposed system for long-term DO monitoring in aquaculture and environmental applications.

1. Introduction

Dissolved oxygen (DO) is one of the important water quality indicators in many fields such as aquaculture, wastewater treatment, and environmental monitoring. For example, in aquaculture, changes in oxygen concentration in water can be achieved through major processes such as photosynthesis of aquatic plants, respiration of fish, shrimp, and other organisms, and diffusion of oxygen at the water surface [1]. The DO concentration must reach a suitable threshold for organisms to survive in an aquatic environment. Particularly, a DO concentration of 5 mg/L or higher is considered ideal. DO concentrations below 2 mg/L cause difficulties in development and even the risk of mass mortality [2]. Therefore, continuous monitoring and timely detection of DO deficiencies are essential for maintaining stable oxygen conditions and preventing adverse impacts [3].

Three main methods are commonly used to measure DO concentration: iodometric titration, electrochemical, and optical methods. Among these methods, iodometric titration is considered the most accurate and reliable [4].

However, this is a complex and time-consuming chemical analysis method that is typically performed in a laboratory. This makes the iodometric titration method difficult to apply for continuous DO measurements in the field [4, 5]. The electrochemical method is widely used for many practical applications because of its fast, simple operation, and low cost. The main disadvantage of the electrochemical method is that the measurement device must be maintained and calibrated regularly. In addition, the measurement process of this method consumes oxygen from the measured water [4]. Therefore, this method is only suitable for instantaneous testing of DO concentrations and is not suitable for long-term continuous measurements.

Optical methods mainly based on the principle of fluorescence quenching have received much attention from research groups and equipment manufacturers because they have high accuracy, do not consume oxygen during the measurement process, and require little calibration. Therefore, this measurement method is more suitable for continuous measurements of DO concentrations in the field [4, 5]. However, the cost of optical sensors is often quite high. Moreover, the probe of these optical sensors must be replaced periodically because the luminescent dye inside degrades over time, reducing its sensitivity to oxygen [6].

In addition to the specific advantages and limitations of electrochemical and optical methods, both approaches share a common feature: their sensing elements must be in direct contact with the water sample. As a result, the internal electronic components of these sensors are more susceptible to reduced durability when exposed to water, particularly under conditions of high salinity, humidity, or elevated temperature during long-term use. Furthermore, in certain aquaculture applications, such as intensive shrimp farming or fish hatcheries, dissolved oxygen (DO) levels must be continuously monitored throughout the culture period. Under such conditions, the frequent replacement of sensing probes in electrochemical or optical sensors not only increases equipment and labor costs but may also disrupt continuous monitoring.

To address these limitations, particularly the drawbacks of frequent sensor replacement and the challenges of immersing electrochemical and optical sensors directly in water for long-term monitoring, recent studies have explored non-contact measurement approaches for water quality. In particular, methods based on absorption spectroscopy combined with machine learning have received increasing attention. Maryam et al. [7] performed non-contact pH measurements using the absorbance at 670 nm combined with a support vector machine classification model, achieving a classification accuracy of up to 95%. Devansh et al. [8] tested non-contact total dissolved solids (TDS) measurements using a multispectral sensor combined with some machine learning models. The experimental results in their study showed that the neural network model with 18 wavebands in the region from 410 to 940 nm yielded the best TDS prediction results with an accuracy of over 90%. Suarin et al. [9] employed an artificial neural network model combined with absorption spectroscopy at wavelengths of 670, 770, 810, and 950 nm to predict pH and total ammonia nitrogen in water. Their study achieved coefficients of determination (R^2) of 0.71 for pH and 0.91 for total ammonia nitrogen, respectively.

For DO measurement, Aika et al. [10] used absorbance data in the ultraviolet region combined with a simple regression model to analyze DO concentrations at five different saturation levels (3, 25, 50, 75, and 100%). Their results indicated that the wavelength of 210 nm had a great correlation coefficient in the range of 0.90 to 0.99, depending on the water type. Tran et al. [11] also built a DO concentration prediction system using absorption spectroscopy in the visible and near-infrared region combined with a multiple linear model with R^2 of 0.85 and $RMSE$ (root mean square error)of 0.68 mg/L. Navid et al. [12] built an IoT system to monitor DO concentrations using two wavelengths at 660 and 880 nm. In their study, the Orthogonal Distance Regression (ODR) model was used to predict DO concentration with R^2 of 0.98 and $RMSE$ of 0.39 mg/L. In addition, Lehua et al. [13] used remote sensing images at different wavelengths to build a DO monitoring model in coastal waters. A multiple linear model using images at wavelengths of 483, 613, and 655 nm gave the best performance with R^2 of 0.70 and $RMSE$ of 0.55 mg/L.

These studies demonstrate that absorption spectroscopy combined with machine learning models holds great potential for the continuous monitoring of water quality parameters. However, absorption spectroscopy in the ultraviolet region is associated with health risks and high equipment costs. Moreover, these previous studies have focused on applying only one or a limited number of machine learning algorithms in developing prediction models. This indicates that the systematic development of an optimal DO prediction model has not yet been fully explored.

This study proposes a novel non-contact DO monitoring system that integrates a low-cost multispectral sensor with an optimized machine learning framework to enable long-term, continuous, and cost-effective monitoring without requiring periodic replacement of sensing probes. The proposed system is particularly suitable for real-world aquaculture applications, such as intensive shrimp farming or fish hatcheries, where continuous DO monitoring is essential to maintain optimal water quality and ensure healthy stock development. In this study, a compact sensing unit was constructed to continuously acquire spectral data of water samples at multiple wavebands in the visible and near-infrared regions. To identify the optimal prediction model, several feature selection algorithms and a range of popular machine learning models were systematically explored. The main contributions of this study can be summarized as follows:

- Development of a compact, low-cost, non-contact DO monitoring system integrating a multispectral sensor with an optimized machine learning framework for long-term, continuous measurements.

Int. Journal of Integrated Engineering Vol. 17 No. 9 (2025) p. 71-84

- Systematic exploration of multiple feature selection algorithms and machine learning models to identify the optimal prediction approach, enhancing accuracy and robustness.
- Use of transmittance measurement with a flow-through cuvette to minimize maintenance and enable practical, real-world applications in aquaculture and environmental monitoring, such as intensive shrimp farming or fish hatcheries.

2. Materials and Methods

2.1 Experimental Setup

The proposed system is divided into four main parts: a sensing box, a computer, a water delivery system, and a water tank. These main components are connected, as shown in Fig. 1. Water samples with different DO concentrations were continuously transported from the water tank to the sensing box and returned to the water tank via a water delivery system. The light intensities in the sensing box after passing through the water samples were recorded by a multispectral sensor. The sensing box consisted of a multispectral sensor, a light source, and a flow-through cuvette. This box was covered with black aluminum composite panels. The captured intensity data were then sent to the computer to calibrate the model and to predict the DO concentration in the development and implementation phases, respectively.

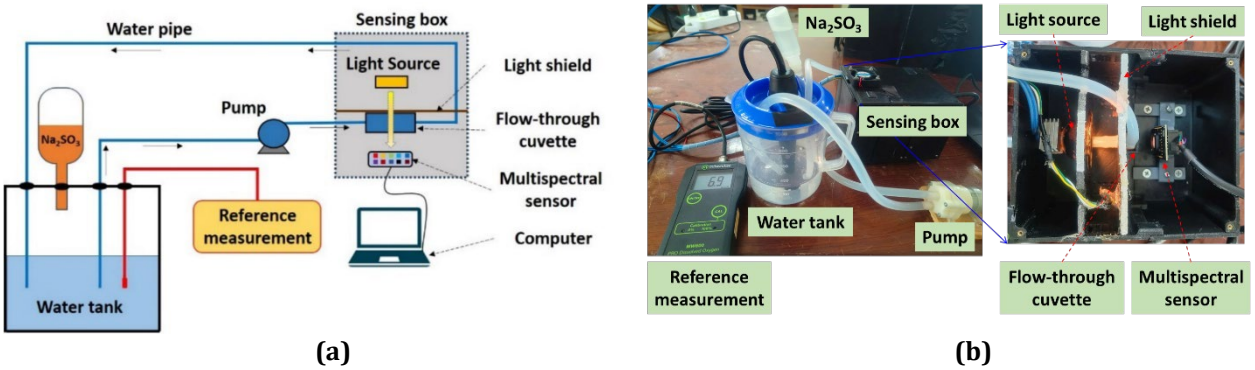


Fig. 1 *The proposed non-contact DO monitoring system: (a) schematic diagram; (b) actual experimental setup*

According to the literature review, the absorption wavelengths of molecular oxygen appear in the ultraviolet, visible, and near-infrared regions [10, 14–17]. Toward creating a continuous DO monitoring system with low investment cost, simplicity in system design, and high predictability, a low-cost multispectral sensor AS7265x (ams-OSRAM AG, Styria, Austria) was used. This sensor has 18 wavebands covering both the visible and near-infrared regions (410, 435, 460, 485, 510, 535, 560, 585, 610, 645, 680, 705, 730, 760, 810, 860, 900, and 940 nm) with a full wave at half maximum of 20 nm for each waveband. With the integration of many modern manufacturing technologies in the sensor (on-chip interference filters, on-chip micro-controller, in-factory calibration, built-in temperature sensor, and low power consumption), the use of this sensor can help reduce the complexity in the system construction process, and the manufacturability and reproducibility of the proposed system can also be significantly improved. This multispectral sensor has been successfully applied in various recent studies, such as for fruit internal quality assessment [18–20], detection of pathogenic bacteria [21], detection of pesticide residues [22], milk quality control [23], and liquid adulteration detection [24].

To obtain a strong and stable light source, a 20 W OSRAM halogen lamp (model 64258, OSRAM GmbH, Munich, Germany) was used in this study. The lamp was powered by a 12 V stabilized power supply to ensure stable light intensity. A flow-through glass cuvette with an optical path length of 10 mm was used. In this study, transmittance measurement was adopted. Accordingly, a light shield with a circular hole of 10 mm in diameter was placed on the cuvette to ensure that only the light that had passed through the water sample was captured by the sensor.

A 1-liter plastic water tank was used to store water samples with different DO concentrations. The water tank was covered by a lid with two holes for the water inlet and outlet, one hole for installing a reference meter probe, and one hole for adding the sodium sulphite (Na_2SO_3) solution. The water delivery system was responsible for continuously circulating water samples from the tank through the sensing box and back to the tank, while ensuring that no leakage occurred and no oxygen from the external environment entered. The water system consisted of flexible plastic pipes and a water pump. Additionally, a reference DO meter was used to provide accurate DO concentrations. Although the Na_2SO_3 bottle and reference DO meter were not part of the system, they were only used to assist in building the prediction model.

In this study, a Windows 10 laptop computer with a typical configuration (Intel Core i3-3110M CPU 2.4 GHz, 6 GB RAM) was used for data acquisition, processing, and training of the machine learning models. MATLAB 2024b software was installed on this laptop to perform model training and extract the characteristics of the machine learning models.

2.2 Water Preparation

To change the concentration of DO in water, mechanical and chemical methods are commonly used. Mechanical methods increase DO levels through aeration or agitation, while chemical methods rely on chemical reactions that either raise or lower DO concentrations. Among these, the addition of sodium sulfite (Na_2SO_3) to water is considered a simple and economical approach and has been widely used in recent studies to create controlled low-oxygen conditions [12, 25, 26]. Therefore, in this study, 400 mL of distilled water was poured into a water tank, and Na_2SO_3 solution was then added to the distilled water tank and the chemical reaction to reduce oxygen in water is as follows:

$$2Na_2SO_3 + O_2 = 2Na_2SO_4 \quad (1)$$

A total of 68 water samples with different oxygen concentrations ranging from 0.1 to 7.1 mg/L were prepared by adding a few drops of Na_2SO_3 solution to the water tank and shaking gently for approximately 2 minutes. The actual DO concentration of the water samples was determined using a reference device.

2.3 Data Acquisition

After adding the Na_2SO_3 solution, each water sample was simultaneously recorded with the proposed system and the reference DO meter to obtain the absorbance data and the corresponding measured DO values. The data acquisition process used in this study is illustrated in Fig. 2.

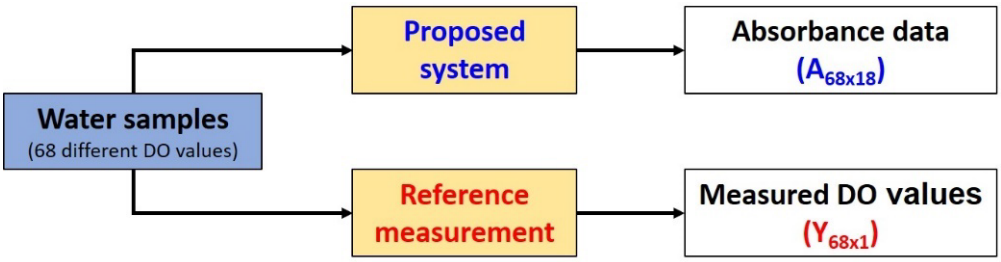


Fig. 2 *Procedure for obtaining absorbance data and measured DO values*

2.3.1 Absorbance Data

Water samples with different DO concentrations were continuously pumped from the water tank through the flow-through cuvette by the water delivery system. At the cuvette position, the light intensity of the water samples, $I(\lambda_i)$, after being illuminated by the light source, was collected by the AS7265x sensor. For each water sample, the sensor read the light intensity 20 times at an interval of 400 ms, and the average value of these 20 readings was used as the light intensity of the water sample. This light intensity was used to calculate the absorbance of the water sample, $A(\lambda_i)$, according to Equation (2), as follows:

$$A(\lambda_i) = log_{10}\left(\frac{I_0(\lambda_i)}{I(\lambda_i)}\right) \quad (2)$$

where $I_0(\lambda_i)$ is the light intensity measured when the cuvette was filled with the standard solution (zero oxygen solution) provided by Hanna Instruments (Model HI7040L, Hanna Instruments, Rhode Island, USA).

2.3.2 Reference Measurement

After dropping the Na_2SO_3 solution into the water, the water sample was gently shaken, and the water delivery system continuously pumped the water through the water pipe to ensure that the DO concentration was well mixed and homogeneous in both the water tank and the cuvette. A portable DO meter (MW 600 Pro, Milwaukee, Szeged, Hungary) was used to determine the DO concentration in the water sample. This is an electrochemical-based DO measurement device designed for reliable and accurate determination of dissolved oxygen in water. It has a measuring range of 0.0–19.9 mg/L with a resolution of 0.1 mg/L and an error of ±1.5% at 25°C. Calibration is performed manually using two points (zero and slope). The device can operate under a wide temperature range (0–50°C) and up to 95% humidity, with automatic temperature compensation from 0 to 30°C. With up to 70 hours of continuous operation, it is particularly suitable for laboratory measurements. In addition, the continuous

Int. Journal of Integrated Engineering Vol. 17 No. 9 (2025) p. 71-84

pumping of water in the proposed system helps evenly mix the Na_2SO_3 solution in the water sample and creates a water flow through the MW600 Pro probe to help this meter operate accurately.

2.4 Regression Analysis

The absorbance data obtained with the proposed system and the DO values measured with the MW600 Pro meter were combined to build machine learning models for predicting the DO concentration. The process of building machine learning models was carried out through the following steps: selecting effective wavebands, training different machine learning models, and evaluating the training results to select the best prediction model. To select an optimal prediction model, various feature selection algorithms and machine learning models have been combined comprehensively. Therefore, both types of feature selection (i.e., the filter and wrapper) have been used. In addition, some popular machine learning models were also applied in this study.

2.4.1 Waveband Selection

The AS7265x sensor simultaneously provided spectral data at 18 wavebands in the range of 410 to 940 nm. To reduce the complexity and improve the prediction performance of the regression model, several feature selection algorithms were implemented, including the minimum redundancy maximum relevance (MRMR), F-test, and sequential forward selection (SFS) algorithms. MRMR and F-test are filter-type feature selection algorithms in which the importance of features is evaluated through feature characteristics such as feature variance or feature relevance to the response. SFS is a wrapper-type feature selection algorithm in which the training process starts by using only one of the features and selecting the feature that best meets the pre-defined training criteria. This training process is repeated with the selected feature sets and the remaining unselected features until the pre-definition training criteria are satisfied. MRMR, F-test, and SFS algorithms have also been commonly used in previous studies to determine the optimal combination of features for machine learning models [27–29].

2.4.2 Training and Evaluation of Regression Models

In this study, various popular regression algorithms, including multiple linear, decision tree, support vector machine, random forest, and neural network, were proposed and tested to predict DO concentration. As listed in Table 1, the configuration values for the machine learning models were used to evaluate multiple variations of each algorithm, thereby expanding the search space for identifying the optimal model in this study.

Table 1 *Several configurations for each proposed algorithm*

Algorithm	Hyperparameter
Multiple linear regression (MLR)	Linear, interaction, robust, stepwise
Decision tree regression (DTR)	Minimum leaf size: 4, 12, 36
Support vector regression (SVR)	Linear, quadratic, cubic, Gaussian
Random forest regression (RFR)	Minimum leaf size: 8; number of learners: 30
Neural network regression (NNR)	Connected layer: 1, 2, 3; layer size: 10, 25, 100

The Regression Learner tool of MATLAB software (version 2024b) was used to train the machine learning models. During training, the k-fold cross-validation technique was applied to validate and compare the performance of the regression models. In this study, the coefficient of determination (R^2) and root mean square error of cross-validation ($RMSECV$) were used to evaluate and select the optimal model. A model is considered better when it has a larger R^2 and a smaller $RMSECV$. R^2 and $RMSECV$ were calculated using Equations (3) and (4), respectively.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{3}$$

$$RMSECV = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{4}$$

Where y_i are the DO values measured by the MW600 Pro meter, $\hat{y}_i$ are the DO values predicted by the proposed model, n is the number of water samples, and $\overline{y}$ is the average value of the measured DO values.

3. Results and Discussion

3.1 Spectral Responses

In this study, a total of 68 water samples spanning a wide range of DO concentrations were collected. Table 2 presents the statistical summary of the recorded light intensities and the corresponding DO concentrations for these samples. The wavebands at 560 and 585 nm exhibited the largest variations in intensity, with coefficients of variation of approximately 19%. This was followed by the wavebands at 645, 705, and 940 nm, which also showed high coefficients of variation of 14.4%, 9.6%, and 10.0%, respectively. Although these wavebands displayed substantial intensity variability, the correlation between their spectral intensities and DO concentration has not yet been established and requires further analysis. The DO concentration data were intentionally generated within the range of 0.1 to 7.1 mg/L, with an interval of 0.1 mg/L, to fully cover the levels typically observed in practical water samples. The dataset has a mean DO concentration of 3.51 mg/L, a standard deviation of 2.10 mg/L, and a coefficient of variation of 58.3%, indicating substantial variability. This variability is beneficial for modeling purposes, as it enables the proposed system to learn from a wide spectrum of DO conditions.

Table 2 *Descriptive statistics of the light intensities at the 18 wavebands and DO concentration for 68 water samples*

Variable	Minimum (a.u.)	Maximum (a.u)	Mean (a.u.)	Standard deviation (a.u.)	Coefficient of variation (%)
410 nm	81	116	84.8	6.5	7.7
435 nm	62	82	67.8	4.4	6.4
460 nm	88	121	104.5	6.3	6.0
485 nm	235	284	256.7	10.0	3.9
510 nm	349	385	364.8	7.0	1.9
535 nm	232	260	248.6	4.3	1.7
560 nm	142	256	183.3	35.1	19.1
585 nm	340	609	440.2	83.3	18.9
610 nm	148	181	172.6	5.5	3.2
645 nm	265	414	323.6	46.7	14.4
680 nm	155	170	164.6	4.0	2.4
705 nm	1194	1622	1372.1	131.7	9.6
730 nm	122	139	135.7	2.3	1.7
760 nm	259	289	279.5	5.8	2.1
810 nm	1182	1375	1306.1	44.1	3.4
860 nm	1062	1188	1139.5	41.6	3.7
900 nm	1581	2111	1845.6	142.3	7.7
940 nm	805	1079	903.6	90.8	10.0
DO concentration	0.1	7.1	3.5	2.1	58.3

As shown in Fig. 3, the average absorbance of the water sample groups at different DO levels exhibits two clear peaks at 610 and 680 nm. In particular, at the 680 nm waveband, the absorbance of the water samples increased gradually as the DO values increased. This suggests that the 680 nm waveband is related to the DO concentration in water. In fact, the absorption wavelength in the 680 nm region corresponds to an electronic transition in the oxygen molecule. In addition, some wavebands in the near-infrared region (760, 810, and 860 nm) exhibited high absorption. These absorption wavebands are also related to electronic transitions in molecular oxygen.

Int. Journal of Integrated Engineering Vol. 17 No. 9 (2025) p. 71-84

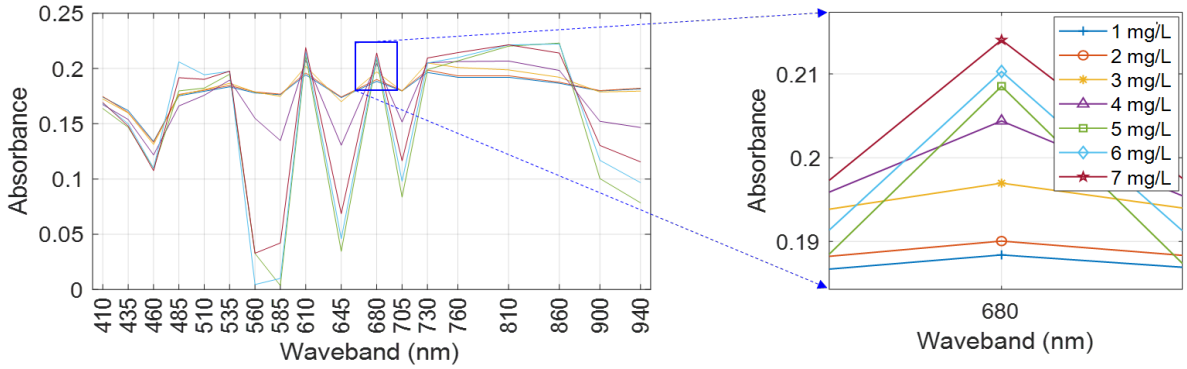


Fig. 3 *Typical spectra of water samples with different oxygen concentrations*

3.2 Wavebands Selection

After performing an analysis of the MRMR and F-test on the absorbance dataset, the importance scores of the 18 wavebands are shown in Fig. 4. The MRMR algorithm ranked the four wavebands at 410, 610, 760, and 860 nm with the highest importance scores (Fig. 4a). Meanwhile, the wavebands at 610, 680, 760, and 810 nm had the highest importance scores in the F-test analysis (Fig. 4b). In addition, the two wavebands at 610 and 760 nm were ranked high in both analysis methods, and these two wavebands also had high absorbance, as shown in Fig. 3.

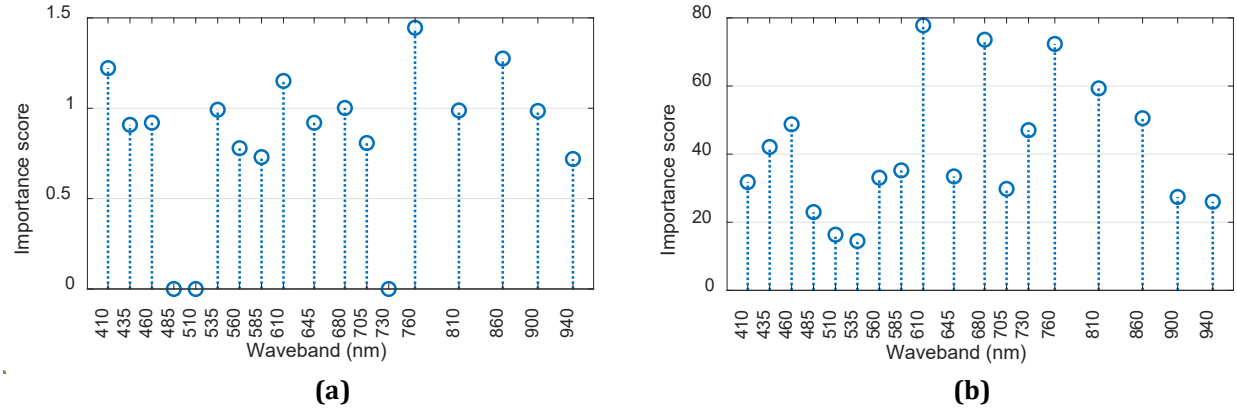


Fig. 4 *The important scores of the 18 wavebands using (a) the MRMR algorithms; (b) F-test algorithm*

For the SFS algorithm, the correlation coefficient between the absorbance of each waveband and the DO value was calculated, as shown in Fig. 5. The 680 nm waveband showed the highest correlation coefficient among the 18 wavebands. Therefore, this waveband was first selected. Subsequently, the 680 nm waveband was combined with one of the remaining wavebands, and the correlation coefficient was calculated. This process was repeated, and a set of four wavebands (680, 460, 585, and 760 nm) yielded the highest correlation coefficients.

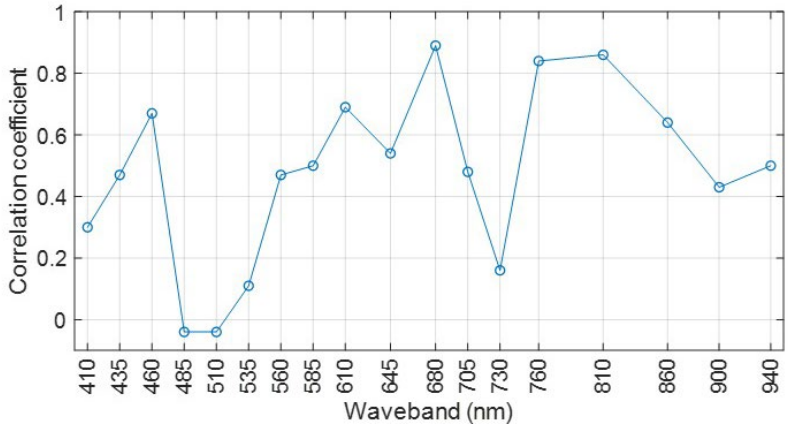


Fig. 5 *The correlation coefficient between absorbance and DO value of each waveband*

Through the waveband selection process using three different algorithms, the selected wavebands of each algorithm were also slightly different: the set of the MRMR selected wavebands included 410, 610, 760, and 860

nm; the set of the F-test selected wavebands included 610, 680, 760, and 810 nm; and the set of SFS selected wavebands included 460, 585, 680, and 760 nm. Due to the difference in the method of measuring the feature importance, the selected waveband sets of the algorithms were slightly different. However, the waveband of 760 nm was selected by all three algorithms. This waveband has also been noted as one of the absorption wavelengths involved in transitions between different electronic states of the oxygen molecule. In addition, 680 nm was the waveband commonly selected by the F-test and SFS algorithms, and the waveband of 610 nm was selected by the MRMR and F-test algorithms. Therefore, the three selected waveband sets were sequentially combined with the five machine learning models to compare and select the model with the best prediction performance.

3.3 Performance of Regression Models

The proposed regression models were trained and evaluated using a 10-fold cross-validation method. Table 3 shows the best performance of each regression model with different options using the three sets of selected wavebands and a set of all 18 wavebands. The best-tuned hyperparameters for the five models varied with the waveband selection method. For MLR, the linear option was selected in most cases, while a stepwise option was applied when using MRMR. The DTR model consistently performed best with a tree size of 4 across four sets of waveband combinations. For SVR, linear kernels were preferred for the All 18 wavebands and the SFS selected wavebands, whereas Gaussian kernels yielded better performance with MRMR and F-test selected wavebands. The RFR model achieved optimal results with a leaf size of 8 and 30 learners regardless of the waveband selection method. Finally, for the NNR, the number of connected layers and the size of the first layer varied depending on the feature set: using all 18 wavebands required a single layer with 100 neurons, MRMR used one layer with 25 neurons, and both F-test and SFS selected waveband sets performed best with two layers of 10 neurons each.

Table 3 shows that all prediction models demonstrate excellent performance with an average R^2 greater than 0.9 and $RMSE$ less than 0.6 mg/L for both the case using all 18 wavebands and the cases using the wavebands selected by the MRMR, F-test, and SFS algorithms. In the case using all 18 wavebands, the NNR model showed the best prediction performance with an R^2 of 0.98 and an $RMSE$ of 0.26. This result reflects the characteristics of the NNR model for automatically learning from complex data. The models with the following prediction performance were RFR, SVR, MLR, and DTR with R^2 of 0.96, 0.96, 0.95, and 0.91, respectively, and $RMSE$ of 0.41, 0.42, 0.47, and 0.62 mg/L, respectively.

Table 3 *Performance comparison of the regression models across various waveband selections*

Model	All 18 wavebands		MRMR wavebands		F-test wavebands		SFS wavebands	
	R^2	$RMSE$ (mg/L)	R^2	$RMSE$ (mg/L)	R^2	$RMSE$ (mg/L)	R^2	$RMSE$ (mg/L)
MLR	0.95	0.47	0.86	0.78	0.88	0.72	0.97	0.35
DTR	0.91	0.62	0.97	0.36	0.97	0.36	0.96	0.40
SVR	0.96	0.42	0.97	0.38	0.97	0.37	0.97	0.34
RFR	0.96	0.41	**0.97**	**0.35**	**0.97**	**0.35**	0.95	0.45
NNR	**0.98**	**0.26**	0.79	0.95	0.93	0.55	**0.99**	**0.22**
Average	0.95	0.44	0.91	0.56	0.94	0.47	0.97	0.35

Compared with the models with all 18 wavebands, the prediction performance of the DTR, SVR, and RFR models with the MRMR selected wavebands was improved. In particular, there was a significant improvement in the prediction results of the DTR model with the R^2 value increasing from 0.91 to 0.97. However, the MLR and NNR models showed worse performance. The models with the F-test selected wavebands also showed a similar prediction performance to the models with the MRMR selected wavebands. Although the DTR, SVR, and RFR models improved their prediction performance, the performance of the MLR and NNR models was worse. Unlike the models with the waveband combinations selected by the MRMR or F-test, the models with the SFS selected wavebands obtained good results (R^2 from 0.95 or higher, $RMSE$ from 0.45 mg/L or lower), and the performance of these models was improved compared to using all 18 wavebands. In particular, the NNR model gave the highest prediction results with R^2 and $RMSE$ of 0.99 and 0.22 mg/L, respectively.

Compared to the average prediction performance of all five models with all 18 wavebands, the models with the SFS selected wavebands proved to be more effective with average R^2 and $RMSE$ values of 0.97 and 0.35 mg/L, respectively. Meanwhile, the average R^2 and $RMSE$ values of the models with the MRMR selected wavebands and with the F-test selected wavebands were not as good. In the waveband selection process, the filter-based MRMR and F-test algorithms measured the importance of features based only on the properties of the features themselves and did not consider the performance of the model. Therefore, the prediction performance of the

Int. Journal of Integrated Engineering Vol. 17 No. 9 (2025) p. 71-84

models with these selected wavebands improved only in some cases, specifically for DTR, SVR, and RFR. In contrast, SFS is a wrapper-based feature selection algorithm, the waveband selection process directly measured the change in the prediction performance of the model, so the prediction results were improved in almost all cases. The prediction results in Table 3 show that the waveband selection process was effective, reducing the complexity of the model, from 18 wavebands to 4 wavebands, and improving the prediction properties.

To further assess the practical applicability of the proposed models, prediction speed and model size were evaluated for the models demonstrating high prediction performance within each group of selected wavebands. Specifically, the models with the highest predictive performance for each waveband group included the NNR model with all 18 wavebands (NNR-Full), the RFR model with MRMR-selected wavebands (RFR-MRMR), the RFR model with F-Test-selected wavebands (RFR-FTest), and the NNR model with SFS-selected wavebands (NNR-SFS). MATLAB's Regression Learner tool was used to calculate the prediction speed and model size of these models. Table 4 presents the evaluation criteria for prediction performance alongside the metrics of prediction speed and model size for the models of interest.

Table 4 *Key evaluation metrics of the top-performing models for each selected waveband set*

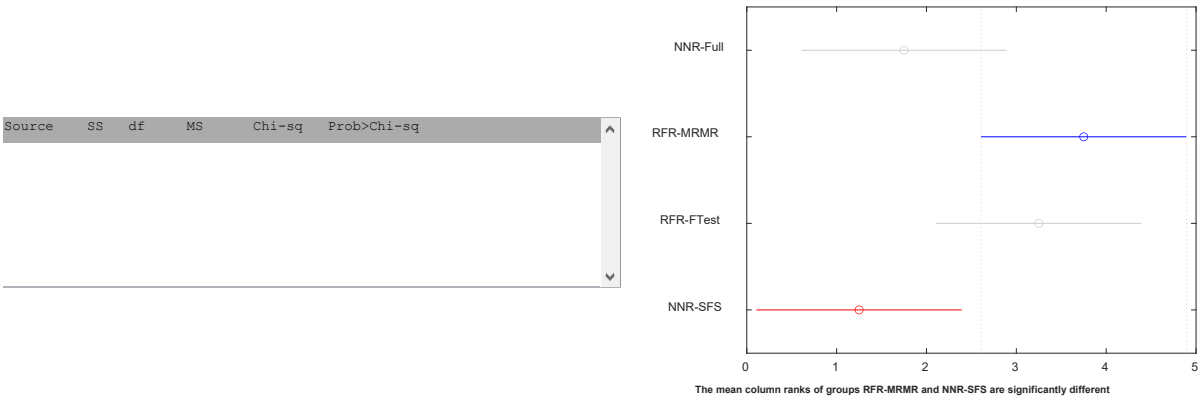
Metrics	NNR-Full	RFR-MRMR	RFR-FTest	NNR-SFS
R^2	0.98	0.97	0.97	0.99
RMSECV (mg/L)	0.26	0.35	0.35	0.22
Prediction speed (objects/second)	489	135	150	415
Model size (KB)	22.85	117.23	116.85	7.21

To identify the best model among the top-performing models within each selected waveband set, a nonparametric statistical testing approach was employed. First, the Friedman test was conducted to assess significant differences among the models, with the null hypothesis stating that no differences exist between them. If the null hypothesis was rejected, the Nemenyi test was subsequently applied to determine which pairs of models differed significantly. Prior to applying the Friedman test, the evaluation metrics of the models were sorted and ranked from best to worst. Table 5 presents the ranking of the top-performing models for each selected waveband set based on their evaluation metrics from Table 4. Models with the same metric values were assigned an equal division of ordinal ranks. For instance, the RFR-MRMR and RFR-FTest models both had an R^2 value of 0.97 and were therefore assigned a tied rank of 3.5.

Table 5 *Ranking of top-performing models for each selected waveband set*

Metrics	NNR_Full	RFR_MRMR	RFR_FTest	NNR_SFS
R^2	2	3.5	3.5	1
RMSECV (mg/L)	2	3.5	3.5	1
Prediction speed (sample/second)	1	4	3	2
Model size (KB)	2	4	3	1
Average	1.75	3.75	3.25	1.25

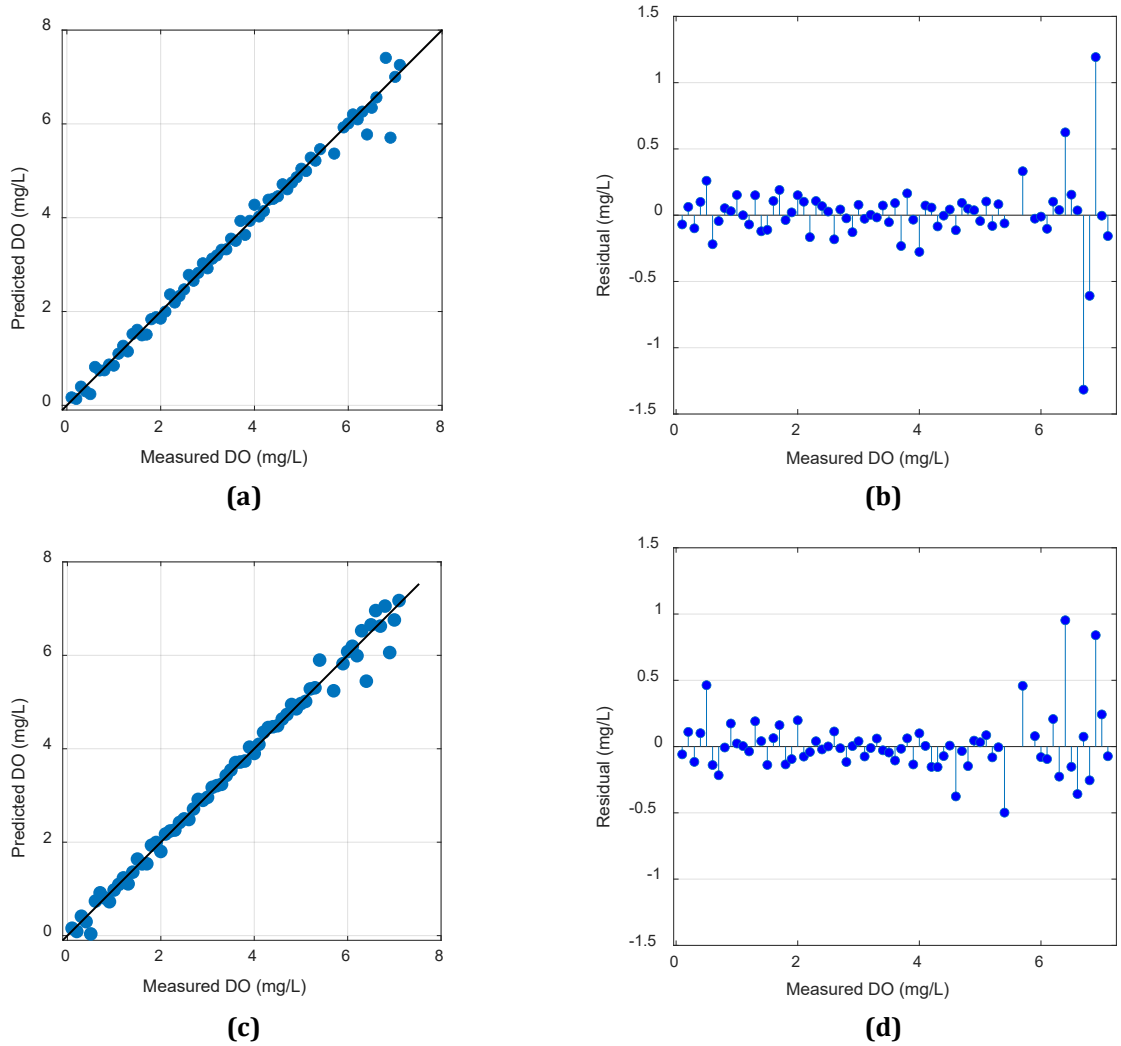
Subsequently, the data in Table 5 were imported into MATLAB, where the *friedman()* function was used to perform the Friedman test, followed by the *multcompare()* function to conduct the Nemenyi test. The results of the statistical analysis are presented in Fig. 6. The Friedman test (Fig. 6a) indicates a significant difference in performance among the models, with a p-value of 0.0132. Subsequent pairwise comparisons using the Nemenyi test (Fig. 6b) reveal that the RFR-MRMR and NNR-SFS models differ significantly, with the NNR-SFS model exhibiting superior performance.



(a) **(b)**

Fig. 6 *Analysis of model rankings using (a) the Friedman test; (b) pairwise comparisons with the Nemenyi test*

In an overall comparison, the NNR model using the SFS selected wavebands gave the best performance and was selected as an optimal model in this study. The scatter plots comparing the actual DO values and the NNR-predicted DO values using all 18 wavebands and the SFS-selected waveband are presented in Fig. 7. The average absolute residual of the NNR model with all 18 wavebands and with the selected SFS wavebands was almost equivalent with values of 0.14 mg/L and 0.13 mg/L, respectively. Furthermore, more than 94% of the error values were within ±0.5 mg/L.



(a) **(b)**

(c) **(d)**

Fig. 7 *Scatter plot and residual of NNR models (a, b) with all 18 wavebands; (c, d) with the SFS selected wavebands*

Int. Journal of Integrated Engineering Vol. 17 No. 9 (2025) p. 71-84

3.4 Discussion

In this study, using a systematic framework that applied multiple feature selection algorithms and various machine learning models with different configurations, the NNR model with four wavebands selected by the SFS algorithm (460, 585, 680, and 760 nm) showed the best prediction performance with $R^2$of 0.99 and $RMSE$ of 0.22 mg/L.

The four SFS selected wavebands in this study are also consistent with those in previous studies. Although the 760 nm wavelength is a weak absorption wavelength of molecular oxygen, this wavelength can still be used to directly measure oxygen by spectrophotometric techniques [5]. The 760 nm wavelength was also recorded as one of the absorption wavelengths of oxygen in the study of [15–17]. The 680 nm waveband was the absorption peak in this study, shown in Fig. 3, and was also the wavelength related to the oxygen absorption recorded in the study of [17]. Two wavebands in the visible region of 460 nm and 585 nm have also been used to monitor the oxygen concentration in water [11]. Notably, these four selected wavebands correspond to or are close to the wavelengths reported in the study of oxygen absorption in the visible and near-infrared regions [14]. This shows that the wavebands selected in this study were well-established and reliable.

The prediction results of this study were compared to those of the previous studies, as shown in Table 6. The prediction performance of the NNR model with the SFS selected wavebands was superior to that of previous studies in terms of both R^2 and $RMSECV$. Study [10] employed shortwave UV bands (190, 210, 240, 250 nm) with a quantile regression (QR) model, achieving a coefficient of determination of 0.96, whereas our NNR model using visible to near-infrared bands (460, 585, 680, 760 nm) reached 0.99 with an RMSECV of 0.22 mg/L. Study [11] applied visible to near-infrared bands (460, 485, 585, 900 nm) with MLR, resulting in lower accuracy (R^2 = 0.85, RMSECV = 0.68 mg/L) compared to our approach. Similarly, Study [12], which used two near-infrared bands (660, 880 nm) with ODR (R^2 = 0.98, RMSECV = 0.39 mg/L), and Study [13], which employed visible bands (483, 613, 655 nm) with MLR (R^2 = 0.73, RMSECV = 0.55 mg/L), were both outperformed by our NNR model. Overall, these results demonstrate that the SFS-selected wavebands with NNR model in this study were highly effective for predicting DO concentrations and outperform those used in previous studies.

Table 6 *Prediction performance of the regression models in this study and previous studies*

Reference	Waveband set (nm)	Model	R^2	$RMSECV$ (mg/L)
[10]	190, 210, 240, 250	QR	0.96	-
[11]	460, 485, 585, 900	MLR	0.85	0.68
[12]	660, 880	ODR	0.98	0.39
[13]	483, 613, 655	MLR	0.73	0.55
This study	460, 585, 680, 760	NNR	0.99	0.22

In addition, by using transmittance measurement with a flow-through cuvette, the proposed system can continuously monitor the oxygen concentration without using any additional color indicator or measuring membrane. This helps the proposed system increase durability and minimize the need for maintenance or component replacement. With excellent prediction performance and almost no need for maintenance or periodic replacement, the proposed system has great potential for practical applications involving long-term continuous DO measurements.

Although the proposed system demonstrated a cost-effective, non-contact spectrometric design with strong predictive capability for DO concentration in water, it remains relatively bulky due to the use of a computer for data acquisition and processing. In addition, the experimental water samples were not diverse, and the influence of pH or temperature on the water samples was not considered. The following issues need to be considered:

- Further optimization of hardware design is needed to create a portable and energy-efficient device, such as considering the use of suitable photo-diodes combined with powerful, low-power micro-controllers.
- A large number of data samples must be collected under various environmental conditions to build a comprehensive and sustainable model.
- Further investigation is required to evaluate the influence of environmental factors (such as temperature, salinity, and pH) on the performance of machine learning models.

4. Conclusions

This study demonstrated a systematic framework for accurate, non-contact monitoring of dissolved oxygen (DO) in water using a low-cost multispectral sensor integrated with an optimized machine learning model. Of the models tested, the NNR model with four wavebands (460, 585, 680, and 760 nm) selected by the SFS algorithm achieved the highest predictive performance (R^2 = 0.99, RMSECV = 0.22 mg/L). These wavebands are consistent

with oxygen absorption wavelengths reported in previous studies, which supports the reliability of the waveband selection. Compared with previous research, the proposed NNR model outperformed existing methods in both accuracy and robustness. These results highlight the potential of the proposed framework for long-term, continuous DO monitoring in aquaculture, environmental management, and other water quality applications, and indicate that similar frameworks could be applied to additional water quality parameters in future research.

Acknowledgement

The authors would like to thank Mr. Nguyen Nhat Linh, Mr. Nguyen Quoc Bao, and Mr. Pham Nguyen Anh Duy for their assistance in preparing data samples.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Nhut-Thanh Tran;* ***data collection:*** *Nhut-Thanh Tran, Chanh-Nghiem Nguyen;* ***analysis and interpretation of results:*** *Chanh-Nghiem Nguyen, Quoc-Hung Pham;* ***draft manuscript preparation:*** *Nhut-Thanh Tran, Chi-Ngon Nguyen. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Boyd, C.E., Torrans, E.L., Tucker, C.S. (2018) Dissolved Oxygen and Aeration in Ictalurid Catfish Aquaculture, *Journal of the World Aquaculture Society 49(1),* 7–70, https://doi.org/10.1111/jwas.12469

[2] Rahman, A., Dabrowski, J., McCulloch, J. (2020) Dissolved oxygen prediction in prawn ponds from a group of one step predictors, *Information Processing in Agriculture 7(2)*, 307–317, https://doi.org/10.1016/j.inpa.2019.08.002

[3] Anuar, F. F., Md Zain, B. A., & Al-Shaibani, N. (2018) Comparative Study on Flexible Link Aerator Using Arduino Programming and Dissolved Oxygen Meter, *International Journal of Integrated Engineering 10(4)*, pp. 1-5, https://doi.org/10.30880/ijie.2018.10.04.001

[4] Wei, Y., Jiao, Y., An, D., Li, D., Li, W., Wei, Q. (2019) Review of Dissolved Oxygen Detection Technology: From Laboratory Analysis to Online Intelligent Detection, *Sensors 19(18),* 3995, https://doi.org/10.3390/s19183995

[5] Wang, X.-d., Wolfbeis, O.S. (2014) Optical methods for sensing and imaging oxygen: materials, spectroscopies and applications, *Chem. Soc. Rev. 43(10),* 3666–3761 (2014) https://doi.org/10.1039/C4CS00039K

[6] Zhang, Y., Chen, L., Lin, Z., Ding, L., Zhang, X., Dai, R., Yan, Q., Wang, X.D. (2019) Highly Sensitive Dissolved Oxygen Sensor with a Sustainable Antifouling, Antiabrasion, and Self-Cleaning Superhydrophobic Surface, *ACS Omega 4(1)*, 1715–1721, https://doi.org/10.1021/acsomega.8b02464

[7] Saberi, M., Gardner, S.D., Haider, M.R. (2021) A machine learning based smart contactless ph sensing and classification, *In: 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS),* pp. 1049–1052, https://doi.org/10.1109/MWSCAS47672.2021.9531918

[8] Rana, D., Lulla, S., Kapadia, H., Dwivedi, A. (2023) Non-contact tds measurement by uvvis-nir spectrophotometric analysis, *In: 2023 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT),* pp. 1–6, https://doi.org/10.1109/CONECCT57959.2023.10234795

[9] Suarin, N. A. S., Lee, J. S., Seng Chia, K., Mohammad Fuzi, S. F. Z., & Gamal Al-Kaf, H. A. (2022) Artificial Neural Network and Near Infrared Light in Water pH and Total Ammonia Nitrogen Prediction, *International Journal of Integrated Engineering 14(4)*, pp. 228-238, https://doi.org/10.30880/ijie.2022.14.04.017

[10] Miura, A., Parra, L., Lloret, J., Catal´aIcardo, M. (2023) UV Absorption Spectrum for Dissolved Oxygen Monitoring: A low-Cost Proposal for Water Quality Monitoring, *Photonics 10(12),* 1336 https://doi.org/10.3390/photonics10121336

[11] Nhut Thanh, T., Nguyen Anh Duy, P., Nhat Linh, N., Quoc Bao, N., Chi Thanh, H., Chanh Nghiem, N. (2024) Continuous Monitoring of Dissolved Oxygen Concentration Using Lowcost Multispectral Sensors, *Journal of Science and Technique 02(02)*, 64–76, https://doi.org/10.56651/lqdtu.jst.v2.n02.857.pce

[12] Shaghaghi, N., Fazlollahi, F., Shrivastav, T., Graham, A., Mayer, J., Liu, B., Jiang, G., Govindaraju, N., Garg, S., Dunigan, K., Ferguson, P. (2024) DOxy: A Dissolved Oxygen Monitoring System, *Sensors 24(10),* 3253, https://doi.org/10.3390/s24103253

[13] Dong, L., Wang, D., Song, L., Gong, F., Chen, S., Huang, J., He, X. (2024) Monitoring Dissolved Oxygen Concentrations in the Coastal Waters of Zhejiang Using Landsat-8/9 Imagery, *Remote Sensing 16(11)*, 1951, https://doi.org/10.3390/rs16111951

[14] Bl´azquez-Castro, A. (2017) Direct 1O2 optical excitation: A tool for redox biology, *Redox Biology 13*, 39–59, https://doi.org/10.1016/j.redox.2017.05.011

[15] Newnham, D.A., Ballard, J. (1998) Visible absorption cross sections and integrated absorption intensities of molecular oxygen (O 2 and O 4), *Journal of Geophysical Research: Atmospheres 103(D22)*, 28801–28815, https://doi.org/10.1029/98JD02799

[16] Krasnovsky, A.A., Kozlov, A.S. (2016) Photonics of dissolved oxygen molecules. Comparison of the rates of direct and photosensitized excitation of oxygen and reevaluation of the oxygen absorption coefficients, *Journal of Photochemistry and Photobiology A: Chemistry 329*, 167–174, https://doi.org/10.1016/j.jphotochem.2016.06.026

[17] Krasnovsky, A.A., Kozlov, A.S., Benditkis, A.S., Goncharov, S.E. (2020) Oxygen activation in aerated solvents by red and infrared laser radiation: measurement of the absorption spectra of dissolved oxygen molecules, *In: 2020 International Conference Laser Optics (ICLO)*, pp. 1–1, https://doi.org/10.1109/ICLO48556.2020.9285531

[18] Tran, N.-T., Fukuzawa, M. (2020) A Portable Spectrometric System for Quantitative Prediction of the Soluble Solids Content of Apples with a Pre-calibrated Multispectral Sensor Chipset, *Sensors 20(20)*, 5883, https://doi.org/10.3390/s20205883

[19] Nguyen, C.-N., Phan, Q.-T., Tran, N.-T., Fukuzawa, M., Nguyen, P.-L., Nguyen, C.-N. (2020) Precise Sweetness Grading of Mangoes (Mangifera indica L.) Based on Random Forest Technique With Low-Cost Multispectral Sensors, *IEEE Access 8*, 212371–212382, https://doi.org/10.1109/ACCESS.2020.3040062

[20] Tran, N.-T., Phan, Q.-T., Nguyen, C.-N., Fukuzawa, M. (2021) Machine learning-based classification of apple sweetness with multispectral sensor, *In: 2021 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, pp. 23–27, https://doi.org/10.1109/SNPDWinter52325.2021.00014

[21] Bai, Z., Wang, B., Gao, T., Xu, X., Du, Z., Han, J., Hu, Y., Bai, Y., Wang, L., Wang, C., Li, D. (2025) A novel microfluidic colorimetric biosensor for rapid and automatic detection Escherichia coli O157:H7 in aquaponics water, *Computers and Electronics in Agriculture 229*, 109941, https://doi.org/10.1016/j.compag.2025.109941

[22] Aira, J., Olivares, T., Delicado, F.M. (2022) SpectroGLY: A Low-Cost IoT-Based Ecosystem for the Detection of Glyphosate Residues in Waters, *IEEE Transactions on Instrumentation and Measurement 71*, 1–10, https://doi.org/10.1109/TIM.2022.3196947

[23] Durgun, M. (2024) Real-Time Milk Quality Control Using Multi-Spectral Sensing and Edge Computing: Advancing On-Site Detection of Milk Components with XGBoost, *Applied Sciences 14(23)*, 10916, https://doi.org/10.3390/app142310916

[24] Shaikh, K., Waqas, A., Korai Baloch, U.A., Muneer, B., Memon, A. (2022) Cost-effective portable photonic sensor for liquid adulteration detection, *In: 2022 IEEE 19th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET)*, pp. 063–068, https://doi.org/10.1109/HONET56683.2022.10019183

[25] Durdevic, P., Yang, Z. (2019) Potential use of realtime dissolved oxygen sensors for oxygen scavenging feedback control, *IOP Conference Series: Materials Science and Engineering 504*, 012098, https://doi.org/10.1088/1757-899X/504/1/012098

[26] Miura, A.S., Parra, M., Lloret, J., Rodilla, M. (2021) Led optical sensor prototype to determine dissolved oxygen saturation in water, *In: 2021 Global Congress on Electrical Engineering (GC-ElecEng)*, pp. 115–120, https://doi.org/10.1109/GC-ElecEng52322.2021.9788384

[27] Wang, G., Lauri, F., Hassani, A.H.E. (2022) Feature Selection by mRMR Method for Heart Disease Diagnosis, *IEEE Access 10*, 100786–100796, https://doi.org/10.1109/ACCESS.2022.3207492

[28] Dhanya, R., Paul, I.R., Akula, S.S., Sivakumar, M., Nair, J.J. (2020) F-test feature selection in Stacking ensemble model for breast cancer prediction, *Procedia Computer Science 171*, 1561–1570, https://doi.org/10.1016/j.procs.2020.04.167

[29] Shafiee, S., Lied, L.M., Burud, I., Dieseth, J.A., Alsheikh, M., Lillemo, M. (2021) Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery, *Computers and Electronics in Agriculture 183*, 106036, https://doi.org/10.1016/j.compag.2021.106036