JOURNAL OF INTERNATIONAL SCIENTIFIC RESEARCH
Volume 3, Issue 3, September, 2025 Online ISSN: 3030-3508
https://spaceknowladge.com

METHOD FOR MONITORING THE PHYSICAL AND CHEMICAL PROPERTIES OF LIQUIDS BASED ON MACHINE LEARNING

Kengesbayev Salauat Kuanishbayevich
PhD Student, Department of Broadcasting Systems, TIUT
Email: salawatkenesbaev@gmail.com

ARTICLE INFO

ARTICLE HISTORY:

Received:26.09.2025
Revised: 27.09.2025
Accepted:28.09.2025

KEYWORDS:

Machine learning, XGBoost, multi-output regression, spectroscopy, ATR method, drinking water, physicochemical parameters

ABSTRACT:

This article presents an approach developed for real-time monitoring of the physical and chemical parameters of drinking water using machine learning (ML) algorithms and Attenuated Total Reflection (ATR) spectroscopy. Although traditional chemical analysis methods provide high accuracy, they are time- and resource-intensive and do not allow automated monitoring. The proposed methodology employs spectral absorption values combined with multi-output regression (Multioutput Regressor) and the XGBoost model to simultaneously predict DOC, NH_4, PO_4, SO_4, NO_3, and NO_2 parameters. The results demonstrate that the XGBoost algorithm ensures the highest accuracy and stability, making it a reliable approach for assessing drinking water quality.

INTRODUCTION

Traditional chemical analysis methods (titrimetry, chromatography, gravimetry) can provide high accuracy; however, they are time-consuming, require the use of expensive reagents, and are not suitable for real-time monitoring [1]. Therefore, in recent years, spectroscopic methods and machine learning (ML) algorithms have been widely applied for liquid monitoring.

ML algorithms enable the processing of large volumes of multidimensional data and the identification of complex relationships within them [2]. This approach offers the following advantages:

- Multivariate analysis: the ability to determine multiple physicochemical parameters simultaneously.
- Flexibility: applicability to different types of liquids by retraining the model.
- Accuracy: improved precision by accounting for nonlinear and high-dimensional correlations.
- Automation: capability to perform real-time monitoring of liquid properties.

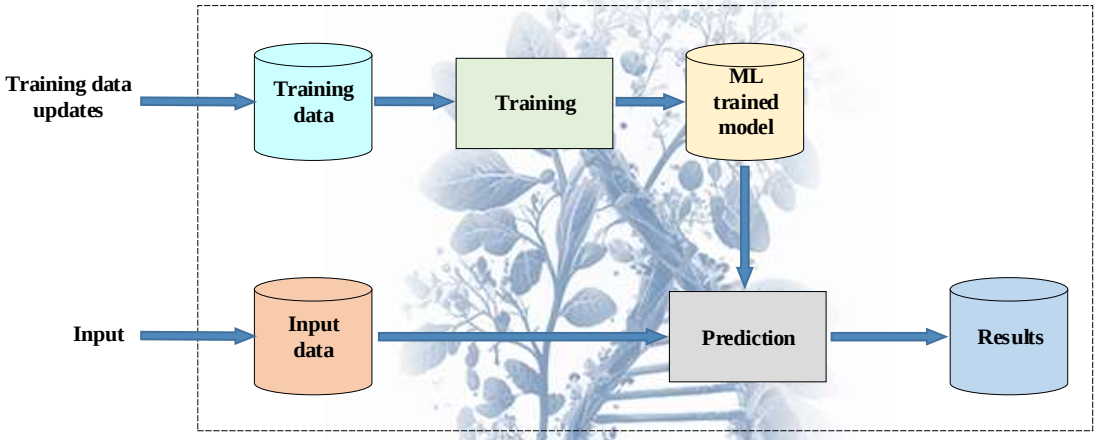

Fig. 1. General workflow of the ML algorithm

Machine learning algorithms typically operate on spectroscopic data. Spectroscopic techniques such as ATR, ultraviolet-visible (UV-VIS), near-infrared (NIR), and Raman spectroscopy allow the identification of the molecular and ionic composition of liquids across different ranges [3]. The ATR method records molecular vibrations in the infrared region, providing a unique "fingerprint" of the components present in the liquid [4]. This approach is particularly advantageous for monitoring multiple parameters simultaneously in multicomponent solutions.

In spectral data, the relationship between absorbance and concentration is generally expressed by the Beer–Lambert law. However, in multicomponent solutions, the interaction of signals often leads to nonlinear behavior of this relationship. Therefore, the use of machine learning algorithms is appropriate for identifying such complex dependencies.

II. MAIN PART

Multi-output regression (Multioutput Regressor) is a wrapper model that extends single-output regressors to simultaneously process multiple target variables. As illustrated in Fig. 2, a separate regression model is trained for each output variable. This approach is particularly convenient for predicting the quality indicators of liquids based on spectral data.

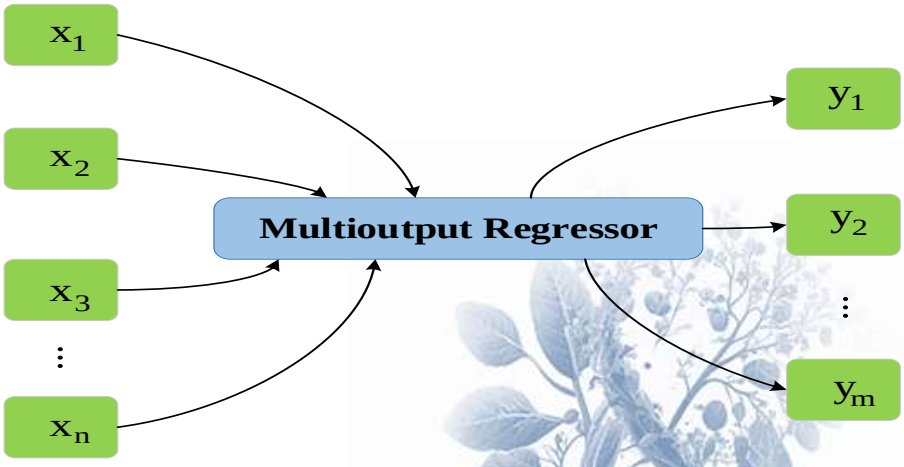

Fig. 2. Input and output data flow in the multi-output regression model

By training a separate regression for each output and combining the results, the Multioutput Regressor enables efficient and independent prediction of each target variable. Although the models are independent for different outputs, this method supports parallel computation and allows multi-output tasks to be performed more conveniently without the need to develop specialized algorithms [5].

XGBoost Regressor is a powerful ensemble machine learning model based on the principle of gradient boosting. The model sequentially constructs multiple decision trees, where each new tree is directed at correcting the errors of the previous ones. As a result, the final model is formed as a weighted sum of the outputs of all trees.

As illustrated in Fig. 3, the XGBoost algorithm builds a new tree at each step, minimizing the errors made in previous predictions. This approach enables the model to effectively learn complex and nonlinear relationships while employing specific mechanisms to reduce overfitting [6].

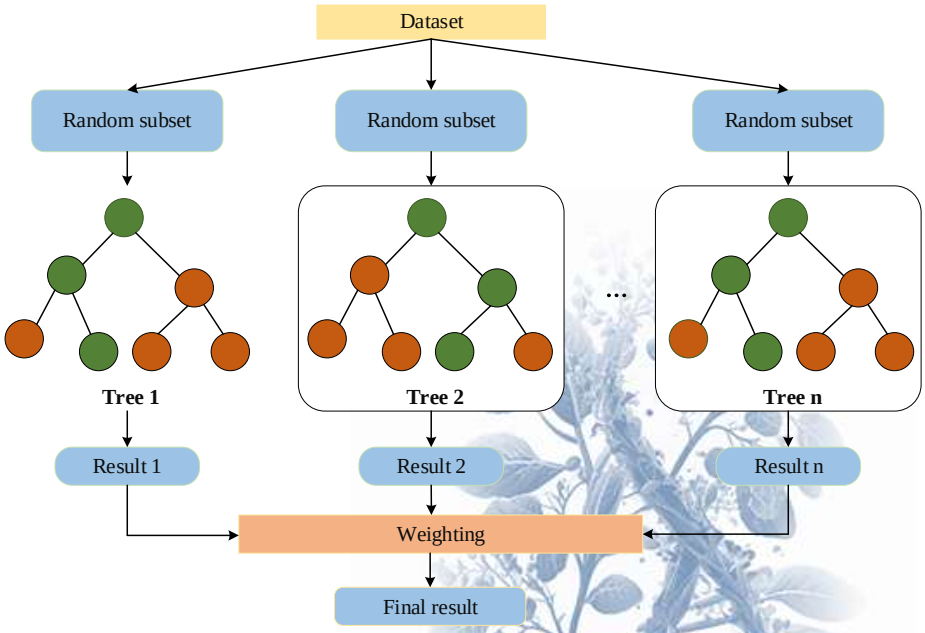

Fig. 3. Decision tree structures of the XGBoost model

In our article, XGBoost is employed as the core model within the Multioutput Regressor framework. This approach allows the simultaneous prediction of multiple physicochemical parameters of liquids using spectral data (e.g., absorbance values). For each output, a separate XGBoost regressor is constructed, and the results are combined to produce the final output.

Another important advantage of XGBoost is its ability to analyze feature importance. This makes it possible to determine which wavelengths (spectral ranges) are most significant in identifying the target parameters. Thus, the model not only performs prediction but also provides insights into the key spectral features influencing the physicochemical properties of liquids.

Various machine learning methods have been applied in spectroscopic prediction: Deep Learning can capture complex correlations but requires large datasets, while SVM is highly sensitive to kernel selection. We chose XGBoost because it strikes a balance between accuracy, efficiency, and interpretability, effectively learns nonlinear relationships, and enables simultaneous prediction of multiple parameters.

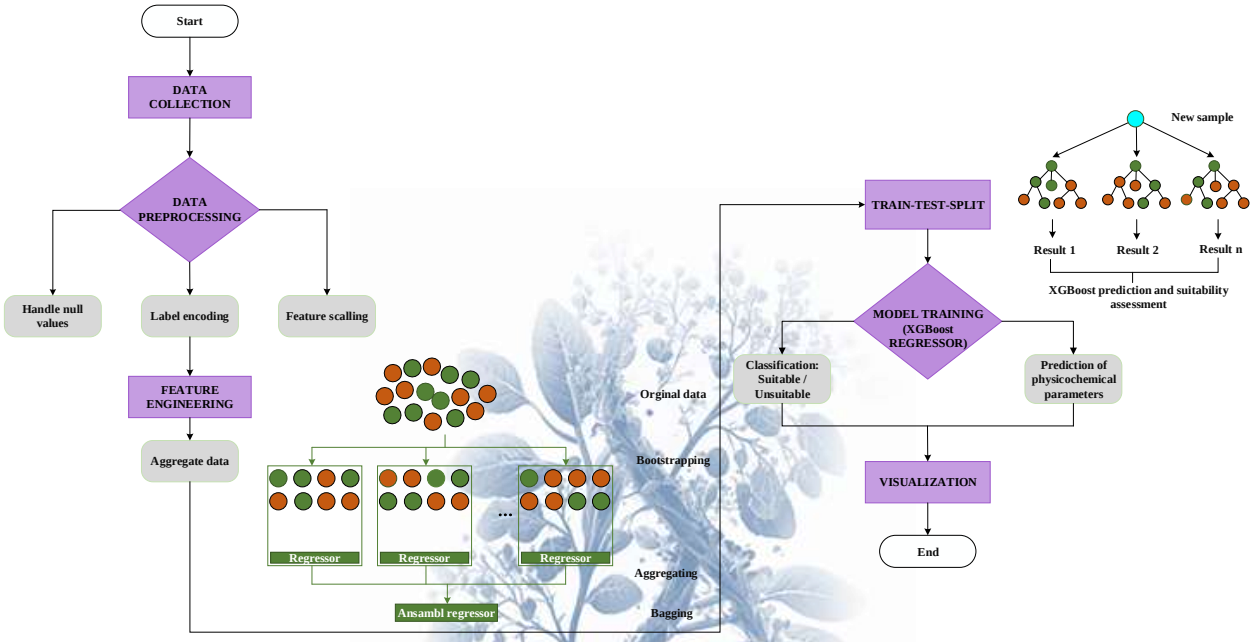

Fig. 4. Block diagram of the machine learning process: data collection, preprocessing, and step-by-step analysis flow

In this study, the data collection process encompasses spectral absorption values obtained through ATR spectroscopy along with laboratory results. The acquired data undergoes preprocessing, including handling of missing values, normalization, and formatting into the required structure. In the subsequent stage, feature engineering is performed to extract the most relevant wavelengths from the spectral data, and a consolidated dataset is constructed. This dataset is then processed using an ensemble approach, where random sampling is applied, multiple regression models are trained, and their results are aggregated. The dataset is divided into training and testing subsets, and a multi-output regression model is built based on the XGBoost algorithm. When new samples are introduced into the model, outputs from each regressor are combined to produce the final prediction. In the final stage, the predicted physicochemical parameters are visualized and compared with international standards, leading to a conclusion on whether the liquid is classified as “suitable” or “unsuitable.”

The block diagram of the machine learning process encompasses the stages of data collection, preprocessing, feature selection, and model training. At the end of this process, the results generated by the model must be scientifically evaluated. For multi-output regression, several evaluation metrics are applied[7]:

Root Mean Square Error (RMSE) – this metric represents the square root of the differences between the predicted results and the actual observations. RMSE expresses the model's errors in the original units and is highly sensitive to large deviations. It is calculated using the following formula:

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

where y_i is the actual (observed) value, $\hat{y}_i$ is the value predicted by the algorithm, and N is the total number of observations.

R^2 – this is a metric that measures how well the independent variables in the model explain the variance of the dependent (target) variable. It represents the proportion of variance explained by the model:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (2)$$

where $SS_{res} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ – is the residual sum of squares of the model, i.e., the sum of squared differences between the actual values (y_i) and the model predictions $(\hat{y}_i)$; $SS_{tot} = \sum_{i=1}^{N}(y_i - \bar{y})^2$ – is the total sum of squares, which measures how far the actual values (y_i) deviate from their mean $(\bar{y})$; N is the number of observations.

We use the above evaluation metrics because RMSE accounts for large errors with high sensitivity, while R^2 reflects the overall goodness of fit of the model.

RESULTS AND DISCUSSION

During the testing process, primary attention was focused on the indicators that are widespread in drinking water and are considered critical according to ecological and sanitary standards. In particular, special emphasis was placed on determining dissolved organic carbon (DOC), ammonium ions (NH_4^+), sulfates (SO_4^{2-}), orthophosphates (PO_4^{3-}), nitrates (NO_3^-), and nitrites (NO_2^-). These components are regarded as key indicators in assessing organic pollution, nutrient levels, and the overall ecological quality of water.

In the course of the study, a dataset of size (505, 269) was constructed and used to test various ML algorithms, including Linear Regression (LR), Random Forest (RF), Support Vector Regression (SVR), and XGBoost. These algorithms were applied to predict the physicochemical parameters of liquids (using drinking water as an example) based on ATR

spectroscopic absorption measurements, and their performance was evaluated using RMSE and R^2 metrics.

The obtained results demonstrated the following (Fig. 5):

- The XGBoost algorithm provided the highest accuracy and the lowest error overall, establishing itself as the most optimal approach.
- The SVR algorithm achieved relatively high R^2 values for certain parameters but showed less stability compared to XGBoost.
- The RF algorithm produced comparatively stable results; however, the errors for some parameters were higher.
- LR, used as a simple baseline model, showed the lowest accuracy when compared with the other algorithms.

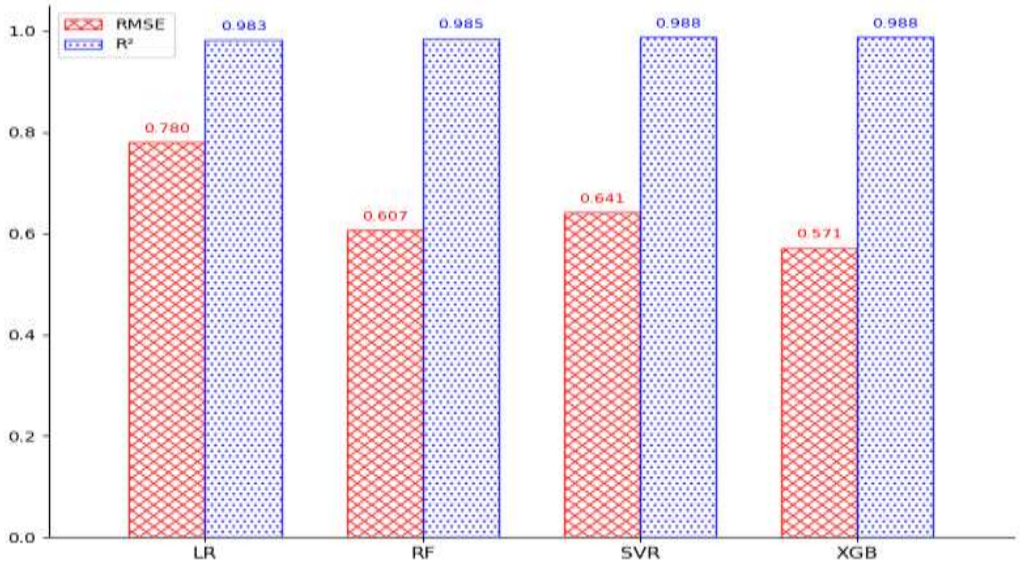

Fig. 5. Comparison of machine learning algorithms based on average RMSE and R^2 values

Thus, in this study, the XGBoost algorithm is recommended as the primary choice for multi-parameter physicochemical prediction.

Analysis of the dataset identified 12 critical wavelengths for drinking water, which were designated as sensitive ranges for evaluating physicochemical parameters. These selected wavelengths are presented in Fig. 6.

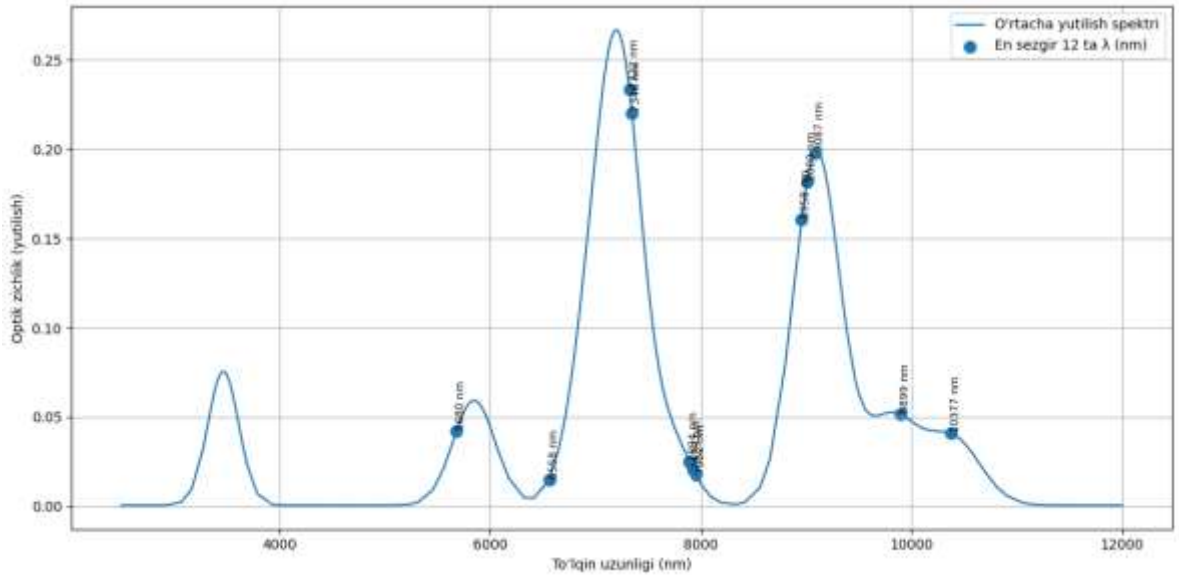

Fig. 6. Selected key wavelengths for drinking water parameters

Based on these key wavelengths, paired LED parameter sets were developed for integration into the device. For each parameter (DOC, NH_4, PO_4, SO_4, NO_3, NO_2), two LED pairs were selected. This approach enhances measurement accuracy and improves spectral sensitivity. The choice of LED pairs was guided by statistical analysis of the dataset, identifying wavelengths with significant influence on the measurement parameters. Furthermore, a pre-established dataset for these parameters was available, enabling the ML module to train regression models specifically on this basis. The selected LED parameter pairs are presented in Table 1.

Table 1. Selected LED pairs for determining the parameters

Parametrs	LED-1 (nm)	LED-2 (nm)
DOC (lab_doc_mg_l)	5680	7322
NH_4 (lab_nh4_mg_l)	6558	7929
PO_4 (lab_po4_mg_l)	9899	10377
SO_4 (lab_so4_mg_l)	8958	9087

NO_3 (lab_no3_mg_l)	9009	9087
NO_2 (lab_no2_mg_l)	9087	7894

During the dataset analysis, the wavelength of 11,618 nm was selected as the optimal point for the compensation channel. At this range, the average absorption value was very low (≈0.00054), and the signal dispersion (std) was also observed to be minimal. This ensures high stability of the compensation signal.

The ML module was trained on the dataset using regression models and validated on the test set. The results demonstrated that the model can accurately predict the key chemical parameters of drinking water (DOC, NH_4, PO_4, SO_4, NO_3, NO_2). Fig. 7 shows a comparison between the actual values and the model's predicted values. The fact that the data points for each parameter are located very close to the ideal line (red dashed line) confirms the effectiveness of the machine learning approach.

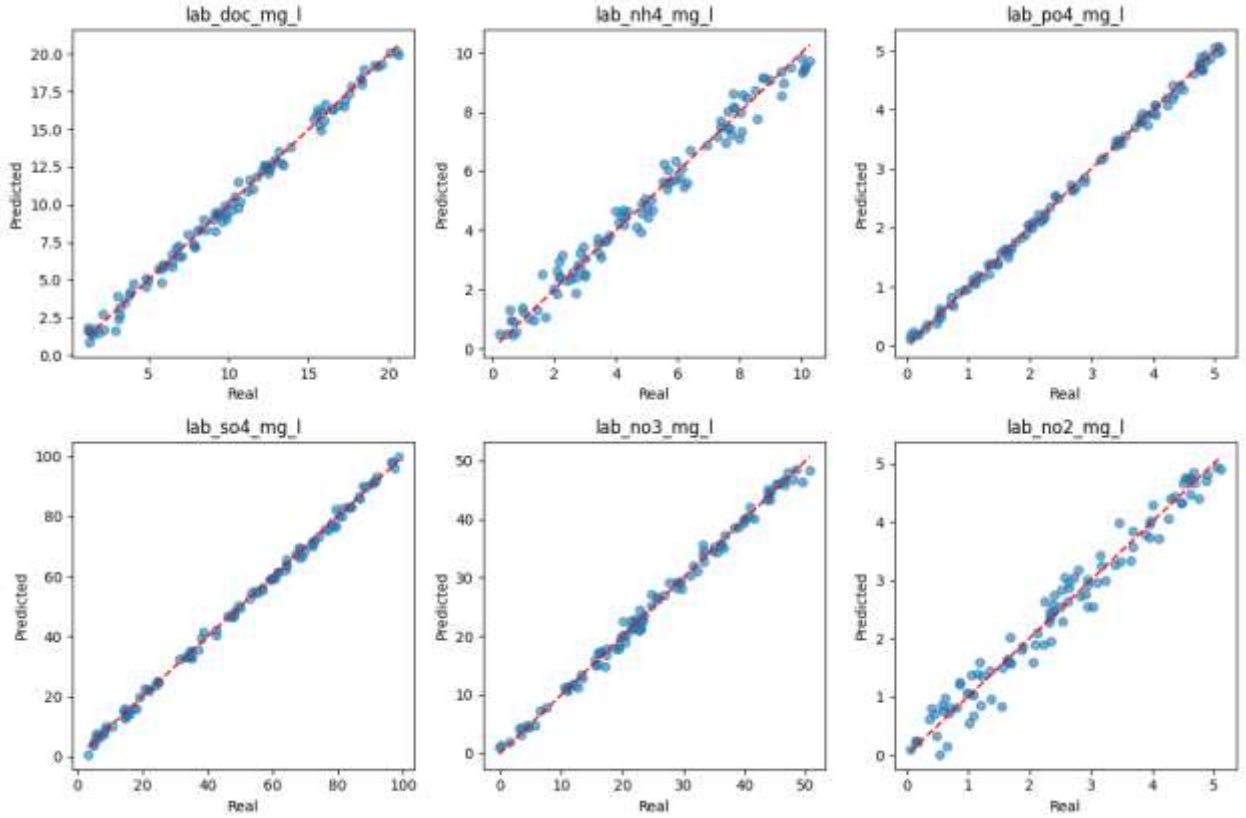

Fig. 7. Prediction accuracy of the chemical parameters of drinking water

The conducted experiments confirmed the high accuracy and stability of the device and demonstrated its capability to effectively monitor the physicochemical parameters of liquids in real time.

CONCLUSION

In this article, the potential of using ATR spectroscopy and machine learning algorithms to assess the key physicochemical parameters of drinking water was investigated. The obtained results demonstrated that the Multioutput regression model based on XGBoost is capable of accurately predicting parameters such as DOC, NH_4, PO_4, SO_4, NO_3, and NO_2. Evaluation metrics, including RMSE and R^2, confirmed the superiority of the XGBoost algorithm compared to other methods.

Furthermore, 12 critical wavelengths were identified and designated as the most sensitive ranges for measurement. Optimal LED pairs were selected for each parameter, while the wavelength of 11,618 nm was determined as the most stable point for the compensation channel. This approach enhanced the reliability and stability of the device's results.

Thus, the developed methodology enables fast, automated, and highly accurate monitoring of drinking water quality. These findings may serve as a scientific foundation for the future development of intelligent optoelectronic devices.

References

1. C. Pasquini, "Near Infrared Spectroscopy: Fundamentals, practical aspects and analytical applications," Journal of the Brazilian Chemical Society, vol. 14, no. 2, pp. 198–219, 2003.

2. Кенгесбаев С.К. СОВЕРШЕНСТВОВАНИЕ ОПТИКОЭЛЕКТРОННЫХ СИСТЕМ УПРАВЛЕНИЯ НА БАЗЕ НПВО С ИСПОЛЬЗОВАНИЕМ МАШИННОГО ОБУЧЕНИЯ. Innovations in Science and Technologies. (2024). Innovations in Science and Technologies, 6, 166. Volume 1, ISSN: 3030-3451. С 166-180.

3. P. R. Griffiths and J. A. de Haseth, Fourier Transform Infrared Spectrometry. 2nd ed. Hoboken, NJ: John Wiley & Sons, 2007.

4. Н. Р. Рахимов, В. А. Жмудь, В. А. Трушин, И. Л. Рева, И. А. Сатволдиев, "Оптоэлектронные методы измерения и контроля технологических параметров нефти и нефтепродуктов," Автоматика и программная инженерия, №2(12), С 85–98, 2015.

5. Paliwal, A.; Subramanian, G.; Ramsundar, B.; Pande, V. MolPROP: Predicting Multiple Molecular Properties Simultaneously using Language and Graph Representations. J. Cheminf. 2024, 16 (1),46.

6. M. Wiens, A. Verone-Boyle, N. Henscheid, J. T. Podichetty, and J. Burton, "A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications," Clinical and Translational Science, vol. 18, no. 3, pp. e70172, 2025. doi: 10.1111/cts.70172

7. D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Computer Science, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.