Application of Machine Learning for Water Pollution Monitoring*

Leonid Bytsyura[1,†], Lesia Dubchak[1,†], Baran Anzhelika[2,†], Maksym Palka[1,†] and Nazar Vivchar[1,†]

[1] *West Ukrainian National University, 11 Lvivska Str., Ternopil, 46008, Ukraine*
[2] *Osnabrueck University , Süsterstraße 28, 49074 Osnabrueck, Germany*

Abstract

Water pollution monitoring is crucial for environmental protection, public health, and sustainable resource management. This study investigates the effectiveness of machine learning models in predicting the Water Pollution Index (WPI) based on monitoring data from the Ikva River in Ukraine. Data from two control points were collected at 10-day intervals from January 2021 to September 2023, covering 20 key physicochemical parameters. Three machine learning models—Linear regression, Random Forest regressor, and XGBoost Regressor—were evaluated using raw, standardized, and polynomial-transformed data. The results indicate that ensemble methods, particularly Random Forest, outperform other models in accuracy, with the best prediction achieved after data standardization. The findings highlight the potential of machine learning for enhancing water quality assessment and environmental decision-making.

Keywords

water pollution monitoring, machine learning, water quality prediction, environmental management, data preprocessing

1. Introduction

Water pollution monitoring plays an important role in ensuring environmental safety, preserving aquatic ecosystems and protecting public health. Water quality directly affects various areas of life, including healthcare, agriculture, industry and ecosystems. Polluted water can contain hazardous chemicals, heavy metals and pathogenic microorganisms that cause serious diseases. In agriculture, the use of such water leads to the accumulation of toxic substances in soil and plants, which negatively affects food security. Industrial processes also depend on water quality, as its pollution can reduce production efficiency. In addition, the deterioration of water resources harms natural ecosystems, causing fish kills, water pollution and loss of biodiversity. That is why regular monitoring is a key element in timely detection of problems, predicting the consequences of pollution and developing effective measures to minimize it [1-3].

Modern approaches to monitoring and controlling water pollution have been significantly improved by the development of technology. Traditional methods involve laboratory analysis of water samples, where physicochemical and biological parameters are evaluated, as well as the use of bioindicators, such as algae and mollusks, to determine the level of pollution. A significant breakthrough in this area has been provided by automated monitoring systems that use sensors and gauges for continuous real-time monitoring of water quality. Internet of Things technologies allow for centralized collection and analysis of this data, which allows for faster response to potential threats [4].

The Second International Conference of Young Scientists on Artificial Intelligence for Sustainable Development (YAISD), May 8-9, 2025, Ternopil, Ukraine

* Corresponding author.

† These authors contributed equally.

l.bytsyura@wunu.edu.ua (L. Bytsyura); dlo@wunu.edu.ua (L.Dubchak); abaran@uni-osnabrueck.de(A.Baran)

0000-0002-9476-011X (L. Bytsyura); 0000-0003-3743-2432 (L.Dubchak); 0000-0003-2134-1325 (A.Baran)

Remote sensing methods, which include satellite monitoring and the use of drones, play a special role. These technologies allow assessing the state of water bodies over large areas by analyzing the spectral characteristics of water and detecting signs of pollution, such as oil spills or algal blooms. Another promising direction is the application of machine learning and artificial intelligence methods. Based on historical data and environmental factors, algorithms predict changes in water quality and help identify the main factors of its pollution. The neural networks and regression algorithms make it possible to identify hidden patterns in changes in the hydrochemical parameters of water bodies, which significantly increases the accuracy of forecasts [4-6].

Water management systems have also been transformed by intelligent platforms that integrate different data sources: field measurements, satellite imagery, meteorological indicators. This allows for a comprehensive approach to assessing the state of water resources and developing effective strategies for their conservation [7].

Thus, modern technologies open up new opportunities for water quality control. The combination of automated sensor systems, remote sensing and artificial intelligence methods provides rapid detection of pollution and effective prediction of their consequences [8-10]. The integration of these approaches allows not only to control the level of pollution of water bodies, but also to implement integrated management of water resources, which is an important step towards their preservation and restoration.

2. Water pollution monitoring

This study analyzed the water quality of the Ikva River, which flows in Western Ukraine. Monitoring data from two control points were used to conduct the analysis:

- Dubno (Dub) – located upstream (upper data collection point);
- village of Sapaniv (Sap) – located downstream (lower data collection point).

Monitoring data was collected at 10-day intervals from January 2021 to September 2023, providing sufficient data for a statistically significant analysis and predictive modeling. This data collection approach allowed for tracking seasonal changes in water quality and analyzing long-term trends.

Monitoring included the collection and analysis of 20 key physicochemical indicators that are critical for determining the ecological status of a water [4]: water temperature (°C); hydrogen pH; dissolved oxygen (mg O_2/dm^3); magnesium (mg/dm^3); chlorides (mg/dm^3) – High chloride levels indicate pollution from industrial waste, road salt, or sewage discharge, which can be toxic to freshwater organisms; sulfates (mg/dm^3); sum of ions (mg/dm^3); hardness (mg-eq/dm^3); hydrocarbonates (mg/dm^3); calcium (mg/dm^3); nitrates (mg/dm^3); dichromate oxidation capacity (mg O/dm^3); BOD_5 (biological oxygen consumption for 5 days, mg O_2/dm^3); ammonium nitrogen (mg N/dm^3); nitrite nitrogen (mg N/dm^3); phosphates (mg P/dm^3); electrical conductivity (μS/cm); total phosphorus (mg P/dm^3); data collection region (categorical variable).

For a comprehensive assessment of water quality, the Water Pollution Index (WPI) was used, which is calculated as the arithmetic mean of the ratios of pollutant concentrations to their regulatory values [11]:

$$WPI = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{S_i}$$

where X_i– actual concentration of the i-th pollutant;

S_i– normative value (maximum permissible concentration) for the i-th pollutant;

n– the number of indicators used for the calculation.

WPI values are interpreted according to the following scale:

WPI < 1 – clean water;

$1 \leq WPI < 2$ – moderately polluted water;

2.1 ≤ WPI < 4 – polluted water;
4.1 ≤ WPI < 6 – very polluted water;
WPI > 6 – extremely polluted water.

This approach to assessing water quality allows us to obtain an integral indicator that takes into account the impact of various pollutants, normalized relative to their potential harmfulness (through the use of MPC).

Data preprocessing occurs by removing duplicate records that may have occurred due to errors during data collection, checking data types and converting them to appropriate formats (numeric, categorical), analyzing the structure of gaps and filling them with median values to preserve the statistical properties of the sample, detecting and processing outliers using the interquartile range (IQR) method.

The creation of additional features is performed by isolating time components (month, season) from the observation date to take into account seasonality and calculate relationships between certain parameters that have an ecological justification.

3. Comparison of different approaches to data processing

The study of the Water Pollution Index (WPI) prediction models involves the analysis of different approaches to machine learning and comparison of their effectiveness. The goal was to determine the most reliable method for predicting the state of water resources, taking into account different types of input data and the peculiarities of their processing [12-17].

In this study, machine learning is an approach in which a computer model automatically finds patterns in data and then uses them to predict new outcomes.

Three different machine learning models were used to predict WPI at Sap based on data from Dub:

1. Linear regression— a basic model that establishes a linear relationship between the input features and the target variable. The optimal coefficients are determined by minimizing the sum of the squares of the deviations between the predicted and actual values.
2. Random forest regressor is an ensemble method based on the construction of a set of decision trees. The following hyperparameters were used to build the random forest model:

- number of trees (n_estimators): 100;
- maximum tree depth (max_depth): optimized through cross-validation;
- minimum number of samples to split (min_samples_split): 2;
- minimum number of samples in a leaf (min_samples_leaf): 1.

3. XGBoost regressor— an efficient gradient boosting algorithm that sequentially builds decision trees, each of which corrects the errors of the previous ones. Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. Shortly after its development and initial release, XGBoost became the go-to method and often the key component in winning solutions for a range of problems in machine learning competitions. XGBoost can be used directly for regression predictive modeling. The following hyperparameters were used for XGBoost:

- learning rate (learning_rate): 0.1;
- maximum tree depth (max_depth): 3;
- number of trees (n_estimators): 100;
- L1 regularization (alpha): 0;
- L2 regularization (lambda): 1.

To compare the effectiveness of different data processing approaches, each model was trained and evaluated on three different datasets:

raw data — original values without additional transformations;

standardized data — after applying StandardScaler;

data with polynomial features — with the inclusion of second-order interactions between features.

This approach allowed us to determine which combination of data processing method and machine learning algorithm is optimal for predicting the WPI.

The data set used for training and testing the models contains 194 data vectors, where each vector has 19 features. To assess the performance of the models and prevent overfitting, the data was divided into training and testing sets in a ratio of 80:20. The division was carried out using stratification by time periods to ensure the representativeness of both sets.

Two main metrics were used to assess the quality of forecasts:

- mean absolute error (MAE)— average absolute deviation between predicted and actual values;
- coefficient of determination (R^2)— a measure of the proportion of variance in the dependent variable that is explained by the model. R^2 takes values from 0 to 1, where 1 means perfect prediction and 0 means the model is no better than the mean.

The first stage of the study was devoted to the analysis of the forecasting results using conventional raw data. The application of linear regression demonstrated moderate results with a mean absolute error (MAE) of 5.2283 and a coefficient of determination (R^2) of 0.6520. This indicates the presence of significant limitations in the linear approach to forecasting complex hydrochemical relationships.

More promising results were shown by the Random Forest model, which achieved MAE=3.5758 and R^2=0.7022. Such indicators indicate the ability of ensemble methods to better capture nonlinear dependencies in data on the state of water bodies. The XGBoost model demonstrated slightly lower efficiency with MAE=5.1899 and R^2=0.5797, which emphasized the importance of the correct choice of machine learning algorithm.

The next step was to process the standardized data, which allowed to significantly improve the prediction results. Linear regression showed MAE=0.3032 and R^2=0.6520, indicating limited scaling efficiency for this type of model. In contrast, Random Forest demonstrated a significant improvement in prediction quality with MAE=0.2236 and R^2=0.8041, which is the best result among all previous tests. The XGBoost model also showed an improvement with MAE=0.2428 and R^2=0.6790, confirming the positive effect of data preprocessing. Comparison of the results showed that scaling of input data can be a critical factor in improving prediction accuracy for some machine learning algorithms.

The third variant of the study involved the use of data with added polynomial features, which allowed modeling more complex nonlinear dependencies. Linear regression unexpectedly showed the best results in this configuration with MAE=0.2390 and R^2=0.8201, indicating the effectiveness of introducing additional nonlinear characteristics. Random Forest showed MAE=0.2641 and R^2=0.7491, slightly inferior to linear regression, but still demonstrating high quality of prediction. XGBoost with MAE=0.2844 and R^2=0.6271 confirmed its suitability for solving prediction problems, although it did not achieve the highest performance.

A graphical representation of the results of testing machine learning methods is shown in Figure 1.

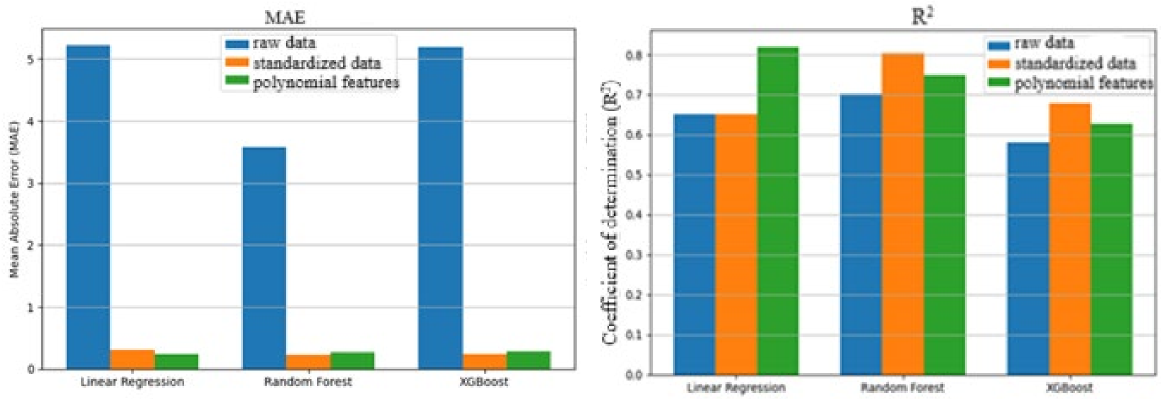

Figure 1: Results of testing machine learning methods

4. Conclusion

The results of the study confirm the effectiveness of machine learning methods for predicting Water Pollution Index. It was found that ensemble methods, in particular Random Forest, demonstrate the best accuracy compared to linear regression and XGBoost. Additional data processing, including standardization and creation of polynomial features, allows for an increase in the accuracy of predictions.

The study provided a deeper understanding of how different machine learning methods respond to the type of hydrochemical data processing when predicting the Water Pollution Index (WPI). In particular, the high efficiency of machine learning as a tool for modeling the state of water resources was confirmed. It has been found that the choice of input data processing method (standardization, feature expansion) can have no less impact on forecast accuracy than the choice of model itself. It has also been demonstrated that even basic models, such as linear regression, can achieve high accuracy when properly tuned.

The proposed approach can be useful for environmental monitoring and management of water resources, contributing to more effective decision-making on their protection and restoration [18-20].

Further research may concern automatic selection of hyperparameters, in particular the use of AutoML or Bayesian optimization methods to improve the performance of models without manual tuning. Also relevant is the development of an interactive WPI forecasting system, which makes it possible to obtain forecasts based on new data quickly.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using these tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] L.Bytsyura, A. Sachenko, T. Kapusta, Kh. Lipianina-Honcharenko, R. Brukhanskyi. Modelling Hydroecomonitoring of Surface Water in Ukraine Using Machine Learning. ProfIT AI 2024: 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024), September 25-27, 2024, Cambridge, MA, USA. P. 245-254 ISSN 1613-0073 https://ceur-ws.org/Vol-3777/paper15.pdf

[2] Statistical Framework for Recreational Water Quality Criteria and Monitoring: monograph / by L. J. Wymer. – Hoboken: John Wiley & Sons, 2007. – 216 p.

[3] Water Quality Monitoring and Management: basis, technology and case studies / edited by Daoliang Li, Shuangyin Liu. – London: Academic Press, 2018. – 368 p.

[4] Alalam, S.; Ben-Souilah, F.; Lessard, M.-H.; Chamberland, J.; Perreault, V.; Pouliot, Y.; Labrie, S.; Doyen, A. Characterization of Chemical and Bacterial Compositions of Dairy Wastewaters. *Dairy* 2021, *2*, 179-190. https://doi.org/10.3390/dairy2020016

[5] MODERN TECHNOLOGIES AND PROCESSES OF IMPROVING THE QUALITY OF LIFE IN GLOBAL CONDITIONS: monograph / edited by M. Bezpartochnyi. – Riga: Baltija Publishing, 2022. – 410 p.

[6] Smart Water Technology for Sustainable Water Management: emerging research and opportunities / edited by Fadi Al-Turjman. – Hershey: IGI Global, 2020. – 230 p.

[7] Introduction to Environmental Data Analysis and Modeling / by Moses Eterigho Emetere. – Cham: Springer, 2020. – 124 p.

[8] Broadening the Use of Machine Learning in Hydrology: challenges and opportunities / edited by Chaopeng Shen, Xiaowei Jia, L. Ruby Leung. – Washington: American Geophysical Union, 2021. – 304 p.

[9] Reshaping Environmental Science Through Artificial Intelligence: emerging research and opportunities / edited by A. J. Tallón-Ballesteros. – Hershey: IGI Global, 2020. – 300 p.

[10] Scaling and Uncertainty Analysis in Ecology: methods and applications / edited by Jianguo Wu, K. Bruce Jones, Harbin Li. – Dordrecht: Springer, 2006. – 338 p.

[11] Mobarok Hossain, Pulak Kumar Patra, Water pollution index – A new integrated approach to rank water quality, Ecological Indicators, Volume 117, 2020, 106668, https://doi.org/10.1016/j.ecolind.2020.106668.

[12] Hrystyna Lipyanina, Anatoliy Sachenko, Taras Lendyuk, Serhiy Nadvynychny, Sergii Grodskyi. Decision Tree Based Targeting Model of Customer Interaction with Business Page. CMIS-2020 Computer Modeling and Intelligent Systems. CEUR Workshop Proceedings (CEUR-WS.org) Vol-2608 urn:nbn:de:0074-2608-1. ISSN 1613-0073. Computer Science - Information Systems - Information Technology. Pp. 1001-1012

[13] Shakhovska, N., Kaminskyy, R., Zasoba, E., & Tsiutsiura, M. (2018). ASSOCIATION RULES MINING IN BIG DATA. International Journal of Computing, 17(1), 25-32. https://doi.org/10.47839/ijc.17.1.946

[14] O. Duda et al., "Data Processing in IoT for Smart City Systems," 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Metz, France, 2019, pp. 96-99, doi: 10.1109/IDAACS.2019.8924262.

[15] Morozov, V. V., Kalnichenko, O. V., & Mezentseva, O. O. O. M. (2020). THE METHOD OF INTERACTION MODELING ON BASIS OF DEEP LEARNING THE NEURAL NETWORKS IN COMPLEX IT-PROJECTS. International Journal of Computing, 19(1), 88-96. https://doi.org/10.47839/ijc.19.1.1697

[16] Sachenko, V. Kochan and V. Turchenko, "Intelligent distributed sensor network," IMTC/98 Conference Proceedings. IEEE Instrumentation and Measurement Technology Conference. Where Instrumentation is Going (Cat. No.98CH36222), St. Paul, MN, USA, 1998, pp. 60-66 vol.1, doi: 10.1109/IMTC.1998.679663.

[17] Lipianina-Honcharenko, K., Savchyshyn, R., Sachenko, A., Chaban, A., Kit, I., & Lendiuk, T. (2022). Concept of the Intelligent Guide with AR Support. International Journal of Computing, 21(2), 271-277. https://doi.org/10.47839/ijc.21.2.2596

[18] Vladov, Serhii, Lukasz Scislo, Valerii Sokurenko, Oleksandr Muzychuk, Victoria Vysotska, Serhii Osadchy, and Anatoliy Sachenko. 2024. "Neural Network Signal Integration from Thermogas-Dynamic Parameter Sensors for Helicopters Turboshaft Engines at Flight Operation Conditions" Sensors 24, no. 13: 4246. https://doi.org/10.3390/s24134246

[19] Bhatia, S., Sharma, M., Bhatia, K. K., & Das, P. (2018). OPINION TARGET EXTRACTION WITH SENTIMENT ANALYSIS. International Journal of Computing, 17(3), 136-142. https://doi.org/10.47839/ijc.17.3.1033

[20] M. Dyvak, "Parameters Identification Method of Interval Discrete Dynamic Models of Air Pollution Based on Artificial Bee Colony Algorithm," 2020 10th International Conference on Advanced Computer Information Technologies (ACIT), Deggendorf, Germany, 2020, pp. 130-135, doi: 10.1109/ACIT49673.2020.9208972.

[21] H. Lipyanina, A. Sachenko, T. Lendyuk, S. Nadvynychny and S. Grodskyi, "Decision Tree Based Targeting Model of Customer Interaction with Business Page", CMIS-2020 Computer Modeling and Intelligent Systems. CEUR Workshop Proceedings (CEUR-WS.org) Vol-2608 urn:nbn:de:0074-2608-1. Computer Science - Information Systems - Information Technology, pp. 1001-1012, [online] Available: https://doi.org/10.32782/cmis%2F2608-75.

[22] Komari, I.E., Fedorenko, M., Kharchenko, V., Yehorova, Y., Bardis, N.G., & Lutai, L. (2020). The Neural Modules Network with Collective Relearning for the Recognition of Diseases: Fault- Tolerant Structures and Reliability Assessment. International Journal of Circuits, Systems and Signal Processing. DOI:10.46300/9106.2020.14.102

[23] Dubchak, L.; Sachenko, A.; Bodyanskiy, Y.; Wolff, C.; Vasylkiv, N.; Brukhanskyi, R.; Kochan, V. Adaptive Neuro-Fuzzy System for Detection of Wind Turbine Blade Defects. Energies 2024, 17, 6456. https://doi.org/10.3390/en17246456

[24] Darmorost, M. Dyvak, N. Porplytsya, T. Shynkaryk, Y. Martsenyuk and V. Brych, "Convergence Estimation of a Structure Identification Method for Discrete Interval Models of Atmospheric Pollution by Nitrogen Dioxide," 2019 9th International Conference on Advanced Computer Information Technologies (ACIT), Ceske Budejovice, Czech Republic, 2019, pp. 117-120, https://ieeexplore.ieee.org/document/8779981

[25] V. Brych, V. Manzhula, B. Brych, N. Halysh, Y. Ursakii and V. Homotiuk, "Estimating the Efficiency of the Energy Service Market Functioning in Ukraine," 2020 10th International Conference on Advanced Computer Information Technologies (ACIT), Deggendorf, Germany, 2020, pp. 670-673, doi: 10.1109/ACIT49673.2020.9208858