

Modelling Hydroecomonitoring of Surface Water in Ukraine Using Machine Learning

Leonid Bytsyura¹, Anatoliy Sachenko^{1,2}, Taras Kapusta³, Khrystyna Lipianina-Honcharenko¹ and Ruslan Brukhanskyi¹

¹ West Ukrainian National University Ternopil, Ukraine

² Kazimierz Pulaski University of Radom, Radom, Poland

³ Center of Advanced Training of Water Management Personnel of Ministry of Ecology of Ukraine, Kyiv

Abstract

Providing the world's population with good-quality drinking water is one of the most important current global challenges. The authors suggest using artificial intelligence tools for factor analysis of water pollution of various origins, modelling probable parameter series, predicting the parameters of ongoing biological and chemical processes, and identifying data with a low level of reliability. The proposed method of analyzing environmental information allows to identify the factor influence of polluting compounds and to model the requested data series. The construction of the machine learning model described in the study involves selecting the most efficient information processing algorithm, adapting it to the training data set to build the required model design, further testing and calculating metrics. In order to improve the forecasting accuracy, a meta-classifier has been developed that combines several basic classifiers as part of an assembly model.

Integration of the described methods into a hydroecological monitoring data processing system can increase the overall performance and correctness of information analysis, as well as make it possible to model scenarios of development of physical and chemical parameters.

Keywords

modelling, machine learning, hydroecological monitoring.

1. Introduction and State of The Arts


One of the world's most important challenges today is to provide the world's population with drinking water of good quality. Systematic researches in recent decades have shown that the quality of water in surface water sources is deteriorating almost everywhere. The establishment of chemical and biological pollution is determined using agreed-upon laboratory methods. Practice proves that the time spent on transportation of the selected samples has a dramatic impact on the quality of research. For example, there are more than 20 regional water laboratories and 4 national laboratories in the system of State agency of water resources in Ukraine. Insufficient laboratory equipment at the regional level and overloading of national laboratories leads to a loss of correctness of indicators. The discrepancy between the data of the authors' research and these laboratories reaches 30-40%. In addition, water quality measurements are usually based on a single point without spatial coverage and insufficient sampling.

A promising approach to water quality monitoring is to actually assess the quality of water in lakes and rivers with the help of using remote sensing images, as each substance has a unique spectral character [27, 32]. The relationship between the percentage of the print and the wavelength when a substance is exposed to the electromagnetic spectrum is known as the spectral signature, which is unique to each substance [1, 26]. Thus, the amount of a pollutant in water can be estimated from the intensity of reflection at different wavelengths by creating an empirical statistical

ProfIT AI 2024: 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024), September 25–27, 2024, Cambridge, MA, USA

✉ l.bytsyura@wunu.edu.ua (L. Bytsyura); as@wunu.edu.ua (A. Sachenko); tiua@ukr.net (T. Kapusta); xrustya.com@gmail.com (Kh. Lipianina-Honcharenko); r.brukhanskyi@wunu.edu.ua (R. Brukhanskyi)

ORCID 0000-0002-9476-011X (L. Bytsyura); 0000-0002-0907-3682 (A. Sachenko); 0009-0002-1533-0494 (T. Kapusta); 0000-0002-2441-6292 (Kh. Lipianina-Honcharenko); 0000-0002-9360-1109 (R. Brukhanskyi)

 © 2024 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

regression between them [29]. Water quality data is expected to grow rapidly as monitoring methods improve [30, 31].

Promising tools for monitoring data processing include forecasting of a multidimensional time series using stacked networks LSTM. The structural design of the LSTM network can be applied based on its effectiveness in predicting time series and in learning long-term dependencies [2]. Water quality monitoring can be carried out manually, as well as with the help of autonomous vehicles using modern robotic systems [3]. At the same time, the construction of integrated water quality monitoring systems based on the Internet of Things requires the compatibility of various sensors and devices, which will allow monitoring water quality in real time [4, 35]. Such tools should reflect the basin principle of information display. At present, there are modern tools for accumulating and analyzing spatial sub-basin and river section data, which is used in HydroATLAS, obtained from the global HydroSHEDS database [5, 24]. Meanwhile the construction of an integrated system for processing monitoring information should take into accounts at least four data layers: national, transborder, regional and global [6]. Forecasting the hydroecological state of water resources is solved with machine learning tools [16, 21-23, 37]. Among the machine learning methods used to assess the physical, chemical and ecological state of water bodies are support vector regression (SVR), artificial neural networks (ANN), random forest (RF) and gradient boosting machine (GBM) [33].

The concept of an intelligent surface water quality monitoring system has been continuously expanded by the idea of combining online water quality monitoring with various online automatic monitoring devices for data collection, communication protocol and software for data interpretation [28]. Implementing an online surface water monitoring system involves calibration and verification methods. Challenges in design and implementation are highlighted as indicators for future improvement and study [7, 25].

It is worth noting that classical distributed measurement systems (DMS) under the new control paradigm are integrated into more complex CyberPhysical Systems (CPS) along with the physical infrastructure [18].

The review of existing environmental water quality monitoring systems demonstrates the features that affect the quality and reliability of the generalised results of data processing. Given the biological and chemical processes, the most vulnerable parameter of such studies is the time, i.e. the speed of processing the selected sample or test [34, 36]. The existing problem of incompatibility of data formats is quite critical, making it difficult or impossible to display their analysis on all four layers of a potential integrated monitoring information processing system. Other factors such as insufficient coverage of environmental data sources, inconsistent application of methodologies, and human factors, also contribute to the poor quality of existing studies.

The authors propose to use artificial intelligence tools for factor analysis of water pollution of various origins, modelling probable parametric series, forecasting the parameters of ongoing biological and chemical processes, identifying data with a low level of reliability, and, as a result, eliminating the negative impact of the identified factors and increasing the speed of functioning of the integrated environmental information processing system.

2. Methods and Materials

The authors propose an approach that includes seven main stages of data collection, processing and visualization of results.

To collect data at Stage 1, within the framework of the State Water Monitoring System, the State Hydrometeorological Service monitors the hydrochemical state of water at 151 water objects and carries out hydrobiological observations at 45 water objects. Data on 46 parameters are obtained, which make it possible to assess the chemical structure of water, biogenic parameters, the presence of suspended particles and organic matter, major pollutants, heavy metals and pesticides. Chronic water toxicity is monitored at 8 water objects. Indicators of radioactive contamination of surface waters are determined [8].

The State Ecological Inspectorate of the Ministry of the Ecology of Ukraine collects water samples and obtains data on 60 measured parameters.

The State Water Agency of Ukraine monitors rivers, reservoirs, canals, irrigation systems and reservoirs within integrated water management systems, water supply systems, transborder watercourses and reservoirs in the areas of influence of nuclear power plants. Water quality is monitored for physical and chemical parameters in 72 reservoirs, 164 rivers, 14 irrigation systems, 1 estuary and 5 mixed-use canals. In addition, water management organizations monitor the content of radio nuclides in surface waters as part of radiation monitoring.

The Sanitary and Epidemiological Service of the Ministry of Health of Ukraine monitors sources of centralised and decentralised drinking water supply, as well as recreational areas along rivers and reservoirs.

Enterprises of the State Geological Service of the Ministry of Ecology of Ukraine monitor groundwater conditions. At the monitoring sites, the level of groundwater occurrence or availability and its natural geochemical composition are assessed. The enterprises determine 22 parameters, including the concentration of heavy metals and pesticides [9; 10].

Pre-processing at the Stage 1 includes data cleaning, determination of missing data, coding of categorical variables, normalisation and scaling of numerical data according to principles [19].

At the Stage 2, the data are divided into training and test sets. This provides an objective measure of their effectiveness. The data are split in a 70:30 or 80:20 ratio, with the larger portion used for the training model and the smaller portion for testing one. Let's have the initial data set

$$D = \{(x_i, y_i)\}_{i=1}^n,$$

where x_i – quotient vectors, and y_i – corresponding marks. Then the training set

$$D_{train} = \{(x_i, y_i)\}_{i=1}^{k \cdot n},$$

where $k \in (0,1)$ contains k% of the total amount of data.

The testing set

$$D_{test} = \{(x_i, y_i)\}_{i=k \cdot n + 1}^n$$

contains correspondingly (1-k)% of the total amount of data. This approach allows the model to be trained on D_{train} and evaluate its performance on an independent set of D_{test} , which reduces the risk of overtraining and provides better generalisation.

At the Stage 3, the models are selected and their hyperparameters are set. Model selection involves comparing different algorithms and determining the most effective one. Classification models can be built on the basis of, for example, Decision Trees, Logistic Regression, Support Vector Machines and Neural Networks [14, 17, 19, 20].

Setting of hyperparameters is a critical step in creating machine learning models, as properly selected hyperparameters can significantly improve model productivity. In this case, it is advisable to use the cross-validation methods GridSearchCV and RandomizedSearchCV to optimise hyperparameters. GridSearchCV systematically searches all possible combinations of hyperparameters from a given set.

Let $\Theta = (\theta_1, \theta_2, \dots, \theta_n)$ – be the set of hyperparameters, in this case GridSearchCV will look for the most optimal meanings using modal estimate for every combination $\theta_i \in \Theta$.

Method RandomizedSearchC performs a random search for hyperparameters in a given space, which reduces computational costs. It randomly selects combinations θ_i from a given distribution Θ and estimates the model for each selected combination. Both methods use k-fold cross-validation, where the original data set D is divided into k subsets, and the model is trained on k-1 subsets and

tested on the remaining subset, repeating the process k times. The optimal hyperparameters are those that maximise the average performance of the model over all k partitions.

At the Stage 4 the model is trained by adapting the algorithm to the training data set. This process is usually based on minimising the loss function $L(\Theta)$, which quantifies the difference between the model predictions $f(x_i; \Theta)$ and the actual labels y_i . During training, the model parameters are optimised, for example, using gradient descent methods. Such training allows the model to adapt to the specifics of the data, reducing the error on the training set and increasing the ability to generalise to new data.

At the Stage 5, the effectiveness of the models is evaluated by testing them and calculating metrics to select the most productive model. Model testing involves adapting the trained model on a test data set to objectively evaluate its performance and generalisation ability [15]. Let us assume

$$D_{test} = \{(x_j, y_j)\}_{j=1}^m$$

is the testing set of data, where x_j – quotient vectors, a y_j – corresponding marks. After training, the model $f(\Theta)$ predicts the meaning of the function

$$\hat{y}_j = f(x_j; \Theta).$$

The productivity of the model is evaluated by computing metrics that compare the predictions of \hat{y}_j with the actual labels of y_j . The most common metric for classification is accuracy [15]

$$Accuracy = \frac{1}{m} \sum_{j=1}^m I(\hat{y}_j = y_j),$$

where $I(\cdot)$ – indicator function. Other important metrics are precision

$$Precision = \frac{TP}{TP+FP},$$

and completeness

$$Recall = \frac{TP}{TP+FN},$$

where TP , FP , FN are the number of true positive, false positive and false negative predictions, respectively. The latter two metrics characterise the quality of classification in more detail, especially when the data is unbalanced. This approach allows us to comprehensively evaluate the model's performance on independent data and identify its strengths and weaknesses.

The best models are selected based on accuracy metrics.

At the Stage, the meta-classifier is trained and evaluated for accuracy. This training is based on combining the predictions of several base models to build a generalised algorithm. First, the training data (x_{train}, y_{train}) is used to train k base models f_1, f_2, \dots, f_k , each of which generates its own predictions $\hat{Y}_{f_1}, \hat{Y}_{f_2}, \dots, \hat{Y}_{f_k}$. These predictions are combined into a new feature matrix Z , where $Z = [\hat{y}_{f_1}, \hat{y}_{f_2}, \dots, \hat{y}_{f_k}]$. The meta-classifier ggg is then trained on the new dataset (Z, y_{train}) to optimally combine the predictions of the base models. Analytically, this can be expressed as finding a function $g: \mathbb{R}^k \rightarrow \mathbb{R}$ which minimizes some loss $L(g(\mathbf{Z}), \mathbf{y})$, where L is the loss function that determines the difference between the meta-classifier's predictions and the actual class values.

Evaluation of the meta-classifier accuracy on the test data includes a classification report and the feature matrix mentioned above.

At the final Stage 7, a graph of the accuracy of various models, including the meta-classifier, is created.

3. Case Study

Monitoring of rivers is carried out according to various parameters, the main ones being biological chemical, physicochemical and hydromorphological [10; 11]. Determination of the ecological state of a surface water body is based on the use of a complex of biotic and abiotic components inherent in aquatic ecosystems. The following classes are used to classify the ecological status of a surface water body: I – ‘excellent’, indicated in blue; II – ‘good’, indicated in green; III – ‘satisfactory’, indicated in yellow; IV – the state corresponding to the ecological condition ‘poor’, indicated in orange [12]. A list of chemical pollutants is defined for surface water monitoring [13].

The chemical state of a surface water body is determined on the basis of environmental quality standards.

Such standards are set on two levels: the maximum permissible concentration and the average annual concentration. If water is found to be contaminated with chemicals, it is necessary to measure their content in bottom sediments and confirm their bioaccumulation. Two classes are used to classify the chemical state of a surface water body. For graphical representation, each class is denoted by a different colour: Class I chemical status, which corresponds to the chemical status of ‘good’, is indicated in blue; Class II chemical status, which corresponds to the chemical status of ‘less than good’, is indicated in red.

As a result of monitoring of the surface water massif conducted by the laboratory of the Regional Department of the State Agency of Water Resources of Ukraine for the Dniester River Basin in 2021-2023, physicochemical indicators were obtained, including heavy metals, for a total of 27 items. The data were processed using the tools of this study including an evaluation of various classification models based on accuracy and other important metrics. According to the results of the analysis (Table 1), the selected models Random Forest, Gradient Boosting and XGBoost showed the highest accuracy above 90%.

Table 1
Results of the classification models

Model	Accu- racy	0_Preci- sion	0_Recall	0_F1- Score	1_Preci- sion	1_Recall	1_F1- Score	Macro Avg Preci- sion	Macro Avg Recall	Macro Avg F1- Score	Weigh- ted Avg Preci- sion	Weigh- ted Avg Recall	Weighted Avg F1- Score
Logistic Regression	0,88	0,95	0,91	0,93	0,64	0,78	0,70	0,79	0,84	0,81	0,90	0,88	0,89
K-Nearest Neighbors	0,85	0,89	0,93	0,91	0,57	0,44	0,50	0,73	0,69	0,70	0,83	0,85	0,84
Support Vector Machine	0,83	0,83	1,00	0,91	0,00	0,00	0,00	0,41	0,50	0,45	0,68	0,83	0,75
Decision Tree	0,88	0,95	0,91	0,93	0,64	0,78	0,70	0,79	0,84	0,81	0,90	0,88	0,89
Random Forest	0,92	0,91	1,00	0,96	1,00	0,56	0,71	0,96	0,78	0,83	0,93	0,92	0,91
Gradient Boosting	0,90	0,91	0,98	0,94	0,83	0,56	0,67	0,87	0,77	0,81	0,90	0,90	0,90
AdaBoost	0,87	0,91	0,93	0,92	0,63	0,56	0,59	0,77	0,74	0,75	0,86	0,87	0,86
XGBoost	0,90	0,93	0,95	0,94	0,75	0,67	0,71	0,84	0,81	0,82	0,90	0,90	0,90
LightGBM	0,88	0,91	0,95	0,93	0,71	0,56	0,63	0,81	0,75	0,78	0,88	0,88	0,88
Naive Bayes	0,63	1,00	0,56	0,72	0,32	1,00	0,49	0,66	0,78	0,60	0,88	0,63	0,68

The main ions that determine the chemical type of water are: HCO₃, SO₄²⁻, Cl⁻, Ca²⁺, Mg²⁺, Na⁺, K⁺. In freshwater, their content reaches 95% of all salts. The mineralisation of the river waters studied in this research mainly depends on natural factors.

The pH of river waters depends on the concentration of calcium Ca(HCO₃)₂ and magnesium Mg(HCO₃)₂ in the water, carbon dioxide (CO₂), humic acids and rock diversity. Thus, the pH of water reflects the geochemical situation of the territory.

The biogenic substances of the water resources under study include primarily nitrogen and phosphorus compounds, which are part of the tissues of living organisms and are vital for the development of aquatic plants and animals. The concentration of nutrients is an indicator of the biological and biochemical processes taking place in water bodies.

In general, the hydrochemical composition of the water resources that were the subject of this study is characterised by the presence of nutrients, trace elements and specific pollutants, as well as chemical compounds of agronomic origin. A significant proportion of the substances identified as the main factors are caused by the inflow of untreated wastewater from enterprises, surface-slope runoff from agricultural land and domestic wastewater.

To identify common factors a network diagram of the influence of factors for different classifiers is used, including Decision Tree, Random Forest, Gradient Boosting, and XGBoost (Fig. 1). Red nodes represent factors used by three or more models, orange nodes represent factors shared by two models, and blue nodes represent factors used by only one model. Green nodes represent the models themselves. The importance of each factor is represented by a number on the edges connecting the models to the factors.

The most common factors are located in the centre of the network, such as FLR_AVG, which is used by all models, and CAD_AVG and FLR_MAX, which are important for three models. These factors have the greatest importance and impact on the classification result, as evidenced by their high importance values (0.200 for Decision Tree, 0.091 for Random Forest, 0.202 for Gradient Boosting, and 0.162 for XGBoost).

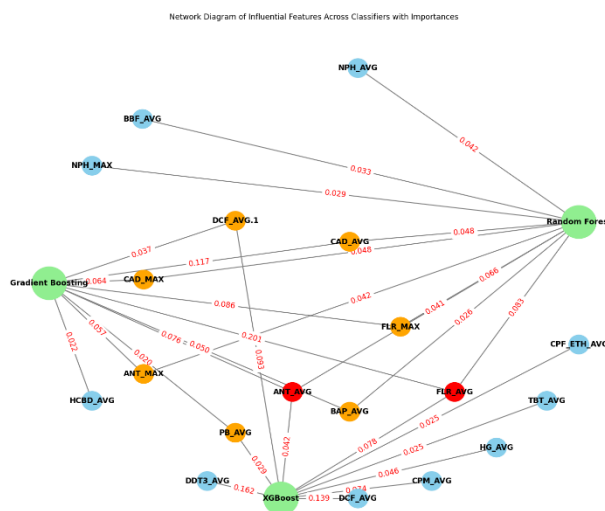


Figure 1: Network diagram of influential features across classifiers with importances

To improve the prediction accuracy, we developed a meta-classifier that combines several basic classifiers as part of an ensemble model. Picture 2 shows the StackingClassifier, which includes four basic classifiers (which were identified in the previous step): Decision Tree, Random Forest, Gradient Boosting, and XGBoost. To set up the meta-classifier (RandomForestClassifier), we used RandomizedSearchCV with the parameters: n_estimators from 50 to 300, max_depth from 4 to 7, min_samples_split from 2 to 10, min_samples_leaf from 1 to 4, and bootstrap (True/False). The best parameters were searched for using five-fold cross-validation (cv=5), 50 iterations, and a random state (random_state=42). The best meta-classifier was used in StackingClassifier as a final estimator.

After training the ensemble model on the training data (Fig. 2), the accuracy of the classifier was 0.9423. Testing the model on the test data showed an accuracy of 94.23%, with macro averages of precision (0.91), recall (0.88), and f1-score (0.89). For class 0, the model achieved an accuracy of 0.95, recall of 0.98 and f1-score of 0.97, while for class 1, the accuracy was 0.88, recall 0.78 and f1-score 0.82. The confusion matrix showed that the model correctly classified 42 out of 43 cases in class 0 and made only 1 mistake, while the model correctly classified 7 out of 9 cases in class 1 and made 2

mistakes. These results confirm the high efficiency of using ensemble methods to improve classification accuracy.

```

Stacking Classifier - Accuracy: 0.9423
      precision    recall  f1-score   support

     0       0.95     0.98     0.97         43
     1       0.88     0.78     0.82          9

   accuracy          0.94         52
  macro avg       0.91     0.88     0.89         52
 weighted avg     0.94     0.94     0.94         52

[[42  1]
 [ 2  7]]

```

Figure 2: Training results of the ensemble model on the training data

The graph (Fig. 3) shows the accuracy of the five classifiers: Decision Tree, Random Forest, Gradient Boosting, XGBoost and Stacking Ensemble. According to the results, Random Forest showed the highest accuracy among the individual models, reaching 92.31%. The accuracy of Decision Tree, Gradient Boosting, and XGBoost was 90.38%. The highest accuracy was demonstrated by Stacking Ensemble, which combines all of these models, reaching 94.23% accuracy. This shows the effectiveness of using ensemble methods that combine several models to improve the overall classification accuracy.

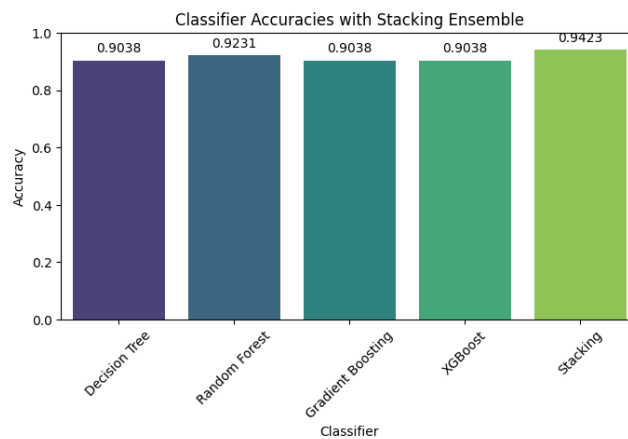


Figure 3: Comparison of classifiers on the test set

Based on the cross-validation evaluation on the new dataset (Fig. 4), the ensemble classifier Stacking showed the highest accuracy with an average value of 0.8135, which indicates its superiority compared to individual models.

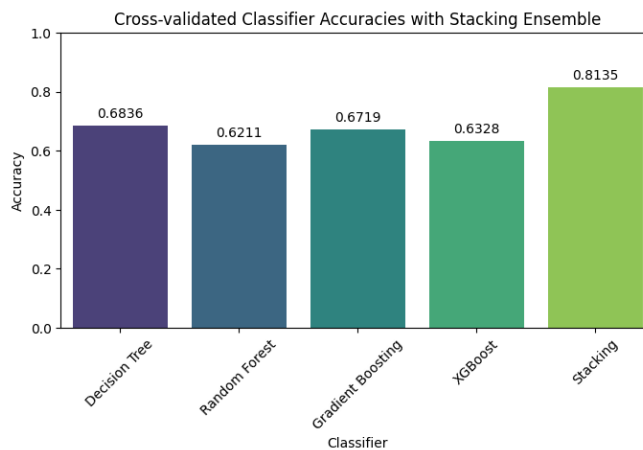


Figure 4: Comparing classifiers on a new data set

4. Conclusion

Systematic research in recent decades has shown that the water quality of surface water sources is getting worse and worse almost everywhere from year to year.

The authors have analyzed various classification models, focusing on their accuracy and other key metrics. According to the results, the Random Forest, Gradient Boosting, and XGBoost models demonstrated the highest accuracy rates, exceeding 90%. In particular, the Random Forest model achieved an accuracy of 92% with a macro average accuracy of 96%.

Based on the results of our investigation, models with an accuracy of more than 90%, namely Random Forest, Gradient Boosting and XGBoost, were selected for further analysis of the influence of factors. The network diagram of the influence of factors for different classifiers represents the importance of each factor by the number on the edges connecting the models to the factors. The most common factors are located in the centre of the network, such as FLR_AVG as well as CAD_AVG and FLR_MAX. It has been proved experimentally that these factors have the greatest importance and impact on the classification result. Other factors, such as DDT3_AVG and CPM_AVG, are used by only one model and have a lower importance. This distribution demonstrates the dependence of different models on different factors, providing a deeper understanding of the impact of each factor on the classification process.

References

- [1] Y. Chen, D. Han, Water quality monitoring in smart city: A pilot project, *Automation in Construction* 89 (2018) 307-316.
- [2] P. Boccadoro, V. Daniele, P. Di Gennaro, D. Lofù, P. Tedeschi, Water quality prediction on a Sigfox-compliant IoT device: The road ahead of WaterS, *Ad Hoc Networks* 126 (2022). <https://doi.org/10.1016/j.adhoc.2021.10274>
- [3] V. L. Dsouza, S. F. Dsouza, M. Sarosh, S. Kukkilaya, V. Chilimbi, S. R. Fernandes, Remotely controlled boat for water quality monitoring and sampling, *Materials Today: Proceedings* 47 (2021) 2391-2400.
- [4] S. A. H. AlMetwally, M. K. Hassan, M. H. Mourad, Real Time Internet of Things (IoT) Based Water Quality Management System, *Procedia CIRP* 91 (2020) 478-485. DOI:10.1016/j.procir.2020.03.107.
- [5] S. Linke, B. Lehner, C. Ouellet Dallaire, Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Sci Data* 6 (2019) 283. DOI:10.1038/s41597-019-0300-6.
- [6] S. Pecora, & H. F. Lins, E-monitoring the nature of water. *Hydrological Sciences Journal* 65 (2020) 683–698. DOI:10.1080/02626667.2020.1724296
- [7] M. I. H. Zaidi Farouk, Z. Jamil, M. F. A. Latip, Towards online surface water quality monitoring technology: A review, *Environmental Research* 238 (2023) 117147, DOI:10.1016/j.envres.2023.117147.
- [8] Dung Dong Thi, Ana Miranda, Charlene Trestrail, Helena De Souza, Khuong V. Dinh, Dayanthi Nugegoda. Antagonistic effects of copper and microplastics in single and binary mixtures on development and reproduction in the freshwater cladoceran *Daphnia carinata*. *Environmental Technology & Innovation*, Volume 24, 2021..
- [9] O. V. Lototska, L. O. Bytsyura, Monitoring of surface water resources in Ukraine and its legislative basis. *Herald of social hygiene and health care organization of Ukraine* 88 (2021) 79-84
- [10] T. Ya. Kapusta, M. Ya. Siviyy, L. O. Bytsyura, Analysis of the state of study of the rivers of the Dniester basin within Ternopil Oblast, *Hydrology, hydrochemistry and hydroecology* 66 (2022) 68-80.
- [11] V. K. Khilchevskiy, T. Ya. Kapusta, L. O. Bytsyura, Characterization of the chemical composition of water and the hydrochemical regime of the left-bank tributaries of the Dniester within Ternopil Oblast, *Hydrology, hydrochemistry and hydroecology* 69 (2023) 31-50.

- [12] The Order of the Ministry of Ecology and Natural Resources of Ukraine of 14.01.2019 No. 5 'On Approval of the Methodology for Assigning a Surface Water Body to One of the Classes of Ecological and Chemical Status of a Surface Water Body, as well as Assigning an Artificial or Significantly Modified Surface Water Body to One of the Classes of Ecological Potential of an Artificial or Significantly Modified Surface Water Body'. URL: <https://zakon.rada.gov.ua/laws/show/z0127-19#Text>. (in Ukrainian).
- [13] The Order of the Ministry of Ecology and Natural Resources of Ukraine of 06.02.2017 No. 45 'On Approval of the List of Pollutants for Determining the Chemical State of Surface and Groundwater Massifs and the Ecological Potential of an Artificial or Significantly Modified Surface Water Massif'. URL: <https://zakon.rada.gov.ua/laws/show/z0127-19#Text>. (in Ukrainian).
- [14] H. Lipyanina, V. Maksymovych, A. Sachenko, T. Lendyuk, A. Fomenko, & I. Kit, Assessing the investment risk of virtual IT company based on machine learning, in: Proceedings of the International Conference on Data Stream Mining and Processing, Cham: Springer International Publishing, August 2020, pp. 167-187.
- [15] K. Lipianina-Honcharenko, C. Wolff, A. Sachenko, I. Kit, & D. Zahorodnia, Intelligent method for classifying the level of anthropogenic disasters, *Big Data and Cognitive Computing* 7 (2023) 157.
- [16] Shafi, U., Mumtaz, R., Anwar, H., Qamar, A. M., & Khurshid, H. (2018, October). Surface water pollution detection using internet of things. In 2018 15th international conference on smart cities: improving quality of life using ICT & IoT (HONET-ICT) (pp. 92-96). IEEE.
- [17] V. Golovko, Y. Savitsky, T. Laopoulos, A. Sachenko and L. Grandinetti, Technique of learning rate estimation for efficient training of MLP, in: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks IJCNN 2000, Neural Computing: New Challenges and Perspectives for the New Millennium, Como, Italy, 2000, pp. 323-328, vol. 1. doi: 10.1109/IJCNN.2000.857856.
- [18] D. L. Carnì, D. Grimaldi, F. Lamonaca, L. Nigro, & P. F. Sciammarella, From distributed measurement systems to cyber-physical systems: A design approach, *International Journal of Computing* 16 (2017) 66-73. <https://doi.org/10.47839/ijc.16.2.882>.
- [19] S. Bhatia, M. Sharma, K. K. Bhatia, & P. Das, Opinion target extraction with sentiment analysis, *International Journal of Computing* 17 (2018) 136-142. <https://doi.org/10.47839/ijc.17.3.1033>
- [20] S. Anfilets, S. Bezobrazov, V. Golovko, A. Sachenko, M. Komar, R. Dolny, V. Kasyanik, P. Bykovyy, E. Mikhno, & O. Osolinskyi, Deep multilayer neural network for predicting the winner of football matches, *International Journal of Computing* 19 (2020) 70-77. <https://doi.org/10.47839/ijc.19.1.1695>
- [21] C. Arrighi, F. Castelli, Prediction of ecological status of surface water bodies with supervised machine learning classifiers, *Science of the Total Environment* 857 (2023) 159655. <https://doi.org/10.1016/j.scitotenv.2022.159655>
- [22] T. D. Acharya, A. Subedi, D. H. Lee, Evaluation of machine learning algorithms for surface water extraction in a landsat 8 scene of Nepal, *Sensors* 19 (2019) 2769. <https://doi.org/10.3390/s19122769>
- [23] A. Tariq, S. Qin, Spatio-temporal variation in surface water in Punjab, Pakistan from 1985 to 2020 using machine-learning methods with time-series remote sensing data and driving factors, *Agricultural Water Management* 280 (2023) 108228, <https://doi.org/10.1016/j.agwat.2023.108228>
- [24] K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, F. Liu, M. Zuo, X. Zou, J. Wang, Y. Zhang, D. Chen, X. Chen, Y. Deng, H. Ren, Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data, *Water Research* 171 (2020) 115454. <https://doi.org/10.1016/j.watres.2019.115454>
- [25] T. N. Do, D. M. T. Nguyen, J. Ghimire, et al., Assessing surface water pollution in Hanoi, Vietnam, using remote sensing and machine learning algorithms, *Environ Sci Pollut Res* 30 (2023) 82230–82247. <https://doi.org/10.1007/s11356-023-28127-2>

- [26] Y. Wang, G. Foody, X. Li et al., Regression-based surface water fraction mapping using a synthetic spectral library for monitoring small water bodies, *GIScience & Remote Sensing* 60 (2023) 2217573. doi:10.1080/15481603.2023.2217573
- [27] Z. Qiao, S. Sun, Q. Jiang, L. Xiao, Y. Wang, H. Yan, Retrieval of Total Phosphorus Concentration in the Surface Water of Miyun Reservoir Based on Remote Sensing Data and Machine Learning Algorithms, *Remote Sens* 13 (2021) 4662. <https://doi.org/10.3390/rs13224662>
- [28] M. I. Shah, W. S. Alaloul, A. Alqahtani, A. Aldrees, M. A. Musarat, M. F. Javed, Predictive Modeling Approach for Surface Water Quality: Development and Comparison of Machine Learning Models, *Sustainability* 13 (2021) 7515. <https://doi.org/10.3390/su13147515>
- [29] M. I. Shah, M. F. Javed, & T. Abunama, Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques, *Environ Sci Pollut Res* 28 (2021) 13202–13220. <https://doi.org/10.1007/s11356-020-11490-9>
- [30] M. I. Shah, M. F. Javed, A. Alqahtani, A. Aldrees, Environmental assessment based surface water quality prediction using hyper-parameter optimized machine learning models based on consistent big data, *Process Safety and Environmental Protection* 151 (2021) 324-340.
- [31] A. Yousefi, M. Toffolon, Critical factors for the use of machine learning to predict lake surface water temperature, *Journal of Hydrology* 606 (2022).
- [32] R. Borovskaya, D. Krivoguz, S. Chernyi, E. Kozhurin, V. Khorosheltseva, E. Zinchenko, Surface Water Salinity Evaluation and Identification for Using Remote Sensing Data and Machine Learning Approach, *J. Mar. Sci. Eng* 10 (2022) 257. <https://doi.org/10.3390/jmse10020257>
- [33] S. Rajabi-Kiasari, & M. Hasanlou, An efficient model for the prediction of SMAP sea surface salinity using machine learning approaches in the Persian Gulf, *International Journal of Remote Sensing* 41 (2019) 3221–3242. <https://doi.org/10.1080/01431161.2019.1701212>
- [34] Z. Di, M. Chang, P. Guo, Y. Li, & Y. Chang, Using real-time data and unsupervised machine learning techniques to study large-scale spatio-temporal characteristics of wastewater discharges and their influence on surface water quality in the Yangtze River Basin, *Water* 11 (2019) 1268.
- [35] U. Shafi, R. Mumtaz, H. Anwar, A. M. Qamar and H. Khurshid, Surface water pollution detection using internet of things, in: *Proceedings of the 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*, Islamabad, Pakistan, 2018, pp. 92-96, doi: 10.1109/HONET.2018.8551341
- [36] M. Naloufi, F. S. Lucas, S. Souihi, P. Servais, A. Janne, T. Wanderley Matos De Abreu, Evaluating the performance of machine learning approaches to predict the microbial quality of surface waters and to optimize the sampling effort, *Water* 13 (2021) 2457. <https://doi.org/10.3390/w13182457>
- [37] M. Zhu, J. Wang, X. Yang, Y. Zhang, L. Zhang, H. Ren, B. Wu, L. Ye. A review of the application of machine learning in water quality evaluation, *Eco-Environment & Health* 1, (2022) 107-116.