Machine learning approaches for water potability prediction: addressing class imbalance with SMOTE

Elina Stepanova[1], *Vasiliy* Orlov[2,4*], *Vladislav* Kukartsev[2,3,4], *Ivan* Pinchuk[3,5], and *Elena* Suprun[2,4]

[1]Krasnoyarsk State Agrarian University, Krasnoyarsk, Russia
[2]Reshetnev Siberian State of Science and Technology, Krasnoyarsk, Russia
[3]Siberian Federal University, Krasnoyarsk, Russia
[4]Artificial Intelligence Technology Scientific and Education Center, Bauman Moscow State Technical University, Moscow, Russia
[5]SPE «Radiosvyaz JSC, ERP department, 19 Dekabristov Str., Krasnoyarsk, 660021, Russia

Abstract. Ensuring access to safe drinking water is a fundamental public health priority. Traditional methods for assessing water quality are labor-intensive and require specialized equipment, which may not be feasible for continuous monitoring. This study explores the use of machine learning models to predict water potability based on various chemical properties. Specifically, we evaluate the performance of Logistic Regression and Random Forest models in the presence of class imbalance, a common issue in environmental datasets. To mitigate this, we apply the Synthetic Minority Over-sampling Technique (SMOTE). Our results indicate that before applying SMOTE, both models exhibited a strong bias towards the majority class (non-potable water), achieving an accuracy of 69.36% and a ROC-AUC of 0.63. However, the application of SMOTE significantly improved the model's ability to identify potable water samples, particularly for the Random Forest model, which achieved an accuracy of 67.07% and a ROC-AUC of 0.64 post-SMOTE. In contrast, the Logistic Regression model showed a decline in performance after SMOTE, suggesting the need for further optimization or alternative approaches. This study highlights the importance of addressing class imbalance in machine learning tasks, especially for critical applications like water quality assessment. Our findings suggest that the Random Forest model, combined with SMOTE, offers a robust solution for predicting water potability. These insights can aid environmental scientists and public health officials in implementing more efficient and accurate water quality monitoring systems. Future research should explore a broader range of models and advanced techniques to further enhance prediction accuracy.

1 Introduction

Access to clean and safe drinking water is a fundamental human right and a critical component of public health. Contaminated water sources can lead to severe health issues,

* Corresponding author: vasi4244@gmail.com

including gastrointestinal diseases, neurological disorders, and even death. The traditional methods of water quality assessment involve laboratory-based chemical analysis, which, although accurate, are time-consuming, costly, and require specialized equipment and personnel [1-3].

In recent years, advancements in machine learning (ML) have provided new avenues for automating and improving the accuracy of water quality assessments. By analysing complex patterns in large datasets, ML models can predict water potability efficiently and accurately. This approach not only reduces the time and cost associated with traditional methods but also provides a scalable solution for continuous water quality monitoring [4, 5].

This study investigates the application of machine learning techniques to classify water potability based on various chemical attributes. The dataset utilized in this research includes measurements of pH, hardness, solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, and turbidity, along with a binary target variable indicating potability [6, 7].

The primary objectives of this research are:

- Data Preprocessing: To preprocess the water quality dataset by addressing missing values and normalizing the data to ensure that all features contribute equally to the model's predictions.
- Model Evaluation: To evaluate the performance of Logistic Regression and Random Forest classifiers in predicting the potability of water.
- Class Imbalance Handling: To tackle the class imbalance in the dataset using the Synthetic Minority Over-sampling Technique (SMOTE), which helps improve the model's performance on the minority class.
- Comparison and Analysis: To compare the performance of the models before and after the application of SMOTE and determine the most effective model for water potability classification.

In the subsequent sections, we present a comprehensive analysis of the dataset, outline the pre-processing techniques employed, and describe the implementation and evaluation of the machine learning models. Through this study, we aim to highlight the potential of machine learning in enhancing water quality assessment processes, ultimately contributing to better public health and resource management [8-10].

By leveraging machine learning, we can develop robust systems for real-time water quality monitoring that can be deployed in various environments, from urban centres to remote rural areas, ensuring that safe drinking water is accessible to all.

2 Materials and methods

2.1 Data Sources

The dataset utilized in this study was obtained from a public repository, specifically designed to provide comprehensive water quality information. The dataset includes measurements of various chemical properties of water samples, which are crucial indicators of water potability. The key features and their descriptions are as follows:

- pH: Measures the acidity or alkalinity of water. The pH scale ranges from 0 to 14, with values below 7 indicating acidity and above 7 indicating alkalinity. Drinking water typically has a pH between 6.5 and 8.5 [11-13].
- Hardness: Represents the concentration of calcium and magnesium ions in water. It is expressed in milligrams per liter (mg/L) and affects the taste and scaling properties of water [11].

- Solids (Total Dissolved Solids - TDS): Measures the combined content of all inorganic and organic substances contained in water. High levels of TDS can affect the taste and may indicate the presence of contaminants [11, 14-16].
- Chloramines: Measures the concentration of chloramines used for water disinfection. While effective at killing bacteria, high levels of chloramines can cause health issues [11].
- Sulfate: Represents the concentration of sulfate ions in water. Sulfate levels above a certain threshold can cause a laxative effect and have an unpleasant taste.
- Conductivity: Measures the water's ability to conduct electrical current, which correlates with the concentration of ions in water.
- Organic Carbon: Indicates the amount of organic compounds in water, which can originate from natural or man-made sources. High levels of organic carbon can lead to the formation of harmful disinfection byproducts.
- Trihalomethanes (THMs): Measures the concentration of byproducts formed when chlorine reacts with organic matter in water. High levels of THMs are associated with health risks such as cancer and reproductive issues.
- Turbidity: Indicates the cloudiness or haziness of water caused by large numbers of individual particles. High turbidity can protect microorganisms from disinfectants and indicate the presence of pathogens.

The target variable, Potability, is a binary indicator where:

- 0: Water is not potable.
- 1: Water is potable.

The dataset contains 3276 samples with the aforementioned features. However, it has some missing values, particularly in the ph, Sulphate, and Trihalomethanes columns. Addressing these missing values is a crucial part of the data pre-processing step.

2.2 Data Description

2.2.1 Handling Missing Values

Missing values in the dataset were addressed by imputing the mean values of the respective columns. This approach ensures that the dataset remains complete without introducing bias that could arise from more sophisticated imputation methods. The columns with missing values and the number of missing entries is:

- ph: 491 missing values
- Sulfate: 781 missing values
- Trihalomethanes: 162 missing values.

2.2.2 Feature Scaling

To ensure that each feature contributes equally to the model's predictions, all features were scaled using the StandardScaler. This method standardizes the features by removing the mean and scaling to unit variance [17].

2.2.3 Handling Class Imbalance

The dataset exhibits a class imbalance, with 1586 samples labeled as non-potable (class 0) and 1034 samples labeled as potable (class 1). To address this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. SMOTE

generates synthetic samples for the minority class, thereby balancing the class distribution and improving the model's ability to generalize [18, 19, 20, 21].

Figure 1 shows the distribution of each feature and the target variable (Potability) in the dataset.

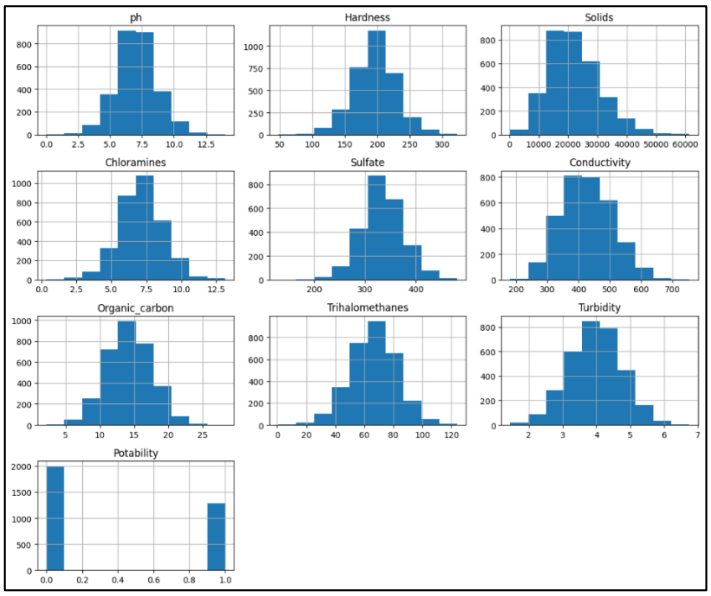

Fig. 1. Distribution of each feature and the target variable

Figure 2 displays the correlation matrix of the dataset, indicating the relationships between the various features and the target variable.

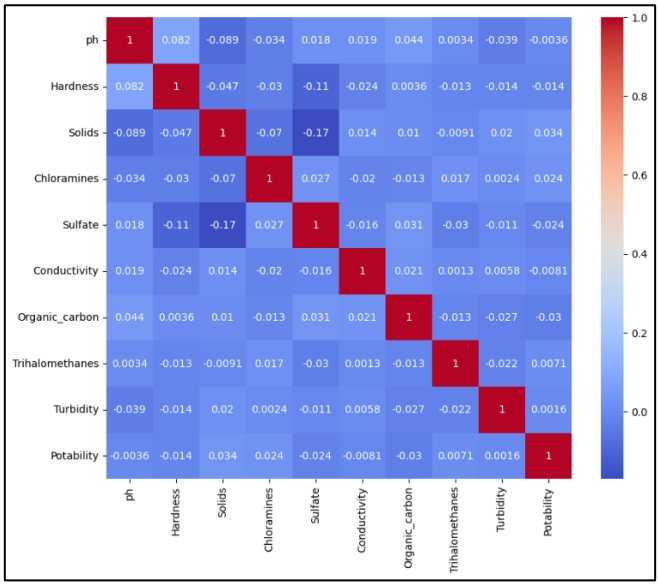

Fig. 2. Correlation matrix

In the following sections, we describe the implementation and evaluation of machine learning models used to classify water potability, both before and after applying SMOTE.

3 Implementation of Machine Learning Models

3.1 Model Selection

In this study, we evaluated two widely used machine learning models for the classification of water potability: Logistic Regression and Random Forest. These models were chosen due to their effectiveness in binary classification tasks and their interpretability.

3.2 Model Training and Evaluation

The dataset was divided into training and testing sets using an 80:20 split. The training set was used to train the models, while the testing set was used to evaluate their performance.

3.2.1 Logistic Regression

Logistic Regression is a linear model for binary classification that estimates the probability that a given input belongs to a certain class. The implementation steps are as follows [22, 23, 24] :

- Model Training: The logistic regression model was trained on the training data using the Scikit-learn library. The logistic regression model is defined as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

  where p is the probability of the target variable being 1 (potable), β are the coefficients, and x are the input features.
- Model Evaluation: The trained model was evaluated on the testing set. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were computed to assess the model's performance. These metrics provide a comprehensive view of the model's classification ability, considering both precision and recall [25, 26, 27].

3.2.2 Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and merges their results to improve accuracy and control over-fitting. The implementation steps are as follows:

- Model Training: The random forest classifier was trained on the training data. Random Forest builds several decision trees during training and outputs the mode of the classes (classification) of the individual trees.
- Model Evaluation: Similar to logistic regression, the performance of the random forest model was evaluated on the testing set using the same metrics. Random Forest often provides better generalization and handles the variance-bias trade-off effectively [28].

3.3 Addressing Class Imbalance

To improve the model's performance on the minority class (potable water), we applied the Synthetic Minority Over-sampling Technique (SMOTE) to the training data. SMOTE generates synthetic samples for the minority class to balance the class distribution [29].

3.3.1 Training and Evaluation with SMOTE

- SMOTE Application: SMOTE was applied to the training data to create a balanced dataset. SMOTE works by selecting a minority class instance and generating new synthetic instances along the line segments joining it to its nearest neighbors in the feature space.
- Model Training: Both logistic regression and random forest models were retrained on the balanced dataset. The application of SMOTE aims to address the imbalance by providing the models with more representative samples of the minority class.
- Model Evaluation: The retrained models were evaluated on the testing set. The same metrics were used to assess the performance improvements resulting from the application of SMOTE. This step ensures that the improvements are quantified and validated.

4 Results

This section presents the performance of the Logistic Regression and Random Forest models in predicting water potability. The models were evaluated both before and after applying the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. The results are detailed below, highlighting key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

4.1 Results Before Applying SMOTE

4.1.1 Logistic Regression

Before applying SMOTE, the Logistic Regression model was evaluated on the testing set. The performance metrics are as follows:

- Accuracy: 69.36%;
- Precision (Class 0: 0.70, Class 1: 0.66);
- Recall (Class 0: 0.89, Class 1: 0.36);
- F1-score (Class 0: 0.79, Class 1: 0.47);
- ROC-AUC: 0.63.

Figure 3 shows the confusion matrix for logistic regression before SMOTE.

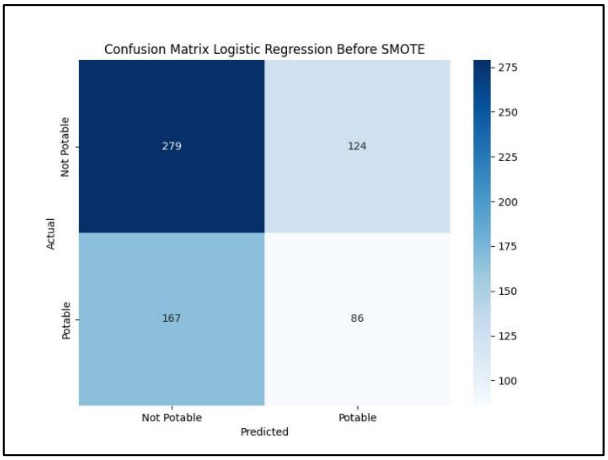

Fig. 3. Confusion Matrix for Logistic Regression before SMOTE

4.1.2 Random Forest

The performance of the Random Forest model before applying SMOTE is as follows:

- Accuracy: 69.36%;
- Precision (Class 0: 0.70, Class 1: 0.66);
- Recall (Class 0: 0.89, Class 1: 0.36);
- F1-score (Class 0: 0.79, Class 1: 0.47);
- ROC-AUC: 0.63.

Figure 4 shows the confusion Matrix for Random Forest before SMOTE.

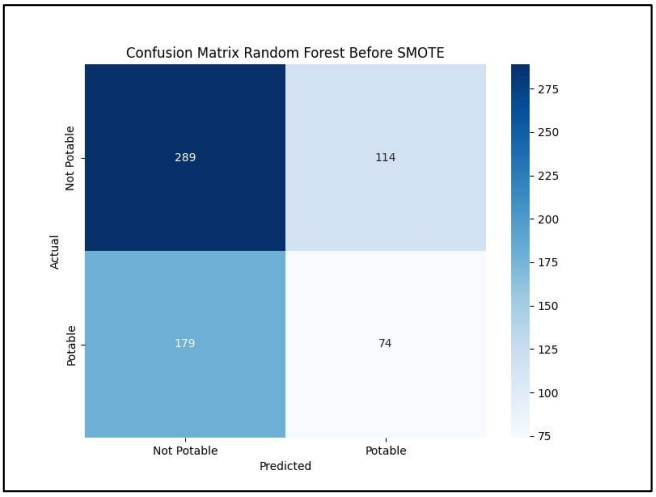

Fig. 4. Confusion Matrix for Random Forest before SMOTE

4.2 Results After Applying SMOTE

4.2.1 Logistic Regression with SMOTE

After applying SMOTE to balance the classes, the Logistic Regression model was retrained and evaluated. The performance metrics are as follows:

- Accuracy: 50.46%;
- Precision (Class 0: 0.63, Class 1: 0.37);
- Recall (Class 0: 0.51, Class 1: 0.49);
- F1-score (Class 0: 0.56, Class 1: 0.42);
- ROC-AUC: 0.50.

Figure 5 shows the confusion Matrix for Logistic Regression after SMOTE.

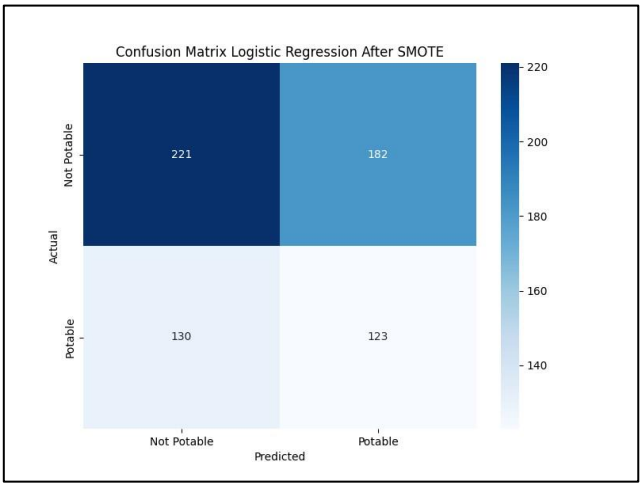

Fig. 5. Confusion Matrix for Logistic Regression after SMOTE

4.2.2 Random Forest with SMOTE

The performance of the Random Forest model after applying SMOTE is as follows:

- Accuracy: 67.07%;
- Precision (Class 0: 0.73, Class 1: 0.56);
- Recall (Class 0: 0.76, Class 1: 0.52);
- F1-score (Class 0: 0.74, Class 1: 0.54);
- ROC-AUC: 0.64.

Figure 6 shows the confusion Matrix for Random Forest before SMOTE.

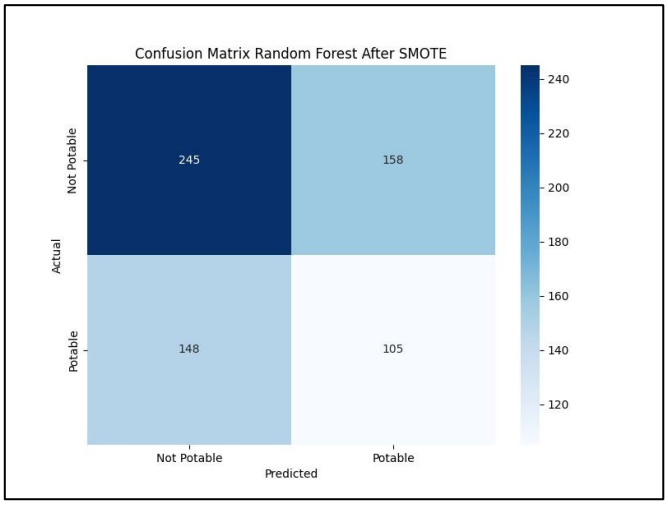

Fig. 6. Confusion Matrix for Random Forest before SMOTE

5 Discussion

The study aimed to evaluate the effectiveness of two machine learning models, Logistic Regression and Random Forest, in predicting the potability of water samples based on various

chemical properties. The results from the analysis highlighted several key insights regarding the performance of these models, both before and after applying the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance [30-33].

5.1 Model Performance Before SMOTE

Both Logistic Regression and Random Forest models demonstrated a clear bias towards the majority class (non-potable water) before applying SMOTE. This was evident from the high recall and F1-scores for the non-potable class, contrasted with the significantly lower recall and F1-scores for the potable class.

- Logistic Regression: Achieved an accuracy of 69.36% with a ROC-AUC score of 0.63. The model had a high recall of 0.89 for the non-potable class but only 0.36 for the potable class, indicating that the model was less effective at identifying potable water [34, 35].
- Random Forest: Similarly achieved an accuracy of 69.36% with a ROC-AUC score of 0.63. The performance metrics mirrored those of Logistic Regression, with a high recall for the non-potable class and a low recall for the potable class [34, 35].

These results underscore the challenge posed by class imbalance, where the model tends to favour the majority class, leading to poor detection of the minority class (potable water) [34, 35].

5.2 Model Performance After SMOTE

Applying SMOTE to balance the classes resulted in notable changes in model performance. The introduction of synthetic samples for the minority class (potable water) aimed to provide the models with a more balanced representation of both classes.

- Logistic Regression with SMOTE: The performance metrics revealed a trade-off. While the recall for the potable class improved to 0.49, the overall accuracy dropped to 50.46%, and the ROC-AUC score decreased to 0.50. This suggests that Logistic Regression struggled to generalize with the newly introduced synthetic samples.
- Random Forest with SMOTE: Showed improved performance with a balanced class representation. The accuracy was 67.07%, and the ROC-AUC score increased to 0.64. The recall for the potable class improved to 0.52, and the F1-score for both classes became more balanced.

5.3 Comparison and Insights

The results indicate that the Random Forest model is more robust in handling class imbalance compared to Logistic Regression. The use of SMOTE enhanced the model's ability to identify potable water samples without significantly compromising the overall accuracy and other performance metrics.

- Effectiveness of SMOTE: The application of SMOTE proved beneficial for the Random Forest model, as it helped mitigate the bias towards the majority class and improved the model's sensitivity to the minority class. However, for Logistic Regression, SMOTE did not yield a similar improvement, highlighting the need for model-specific strategies when dealing with class imbalance [36, 37].
- Model Selection: Random Forest emerged as the preferred model for this classification task due to its superior performance and ability to handle imbalanced datasets more effectively. Its ensemble nature and capacity to capture complex patterns in the data contributed to its robustness.

6 Conclusion

This study explored the application of machine learning models to classify water potability based on various chemical properties. The main focus was on evaluating the performance of Logistic Regression and Random Forest models, particularly in addressing class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE) [38, 39].

Before applying SMOTE, both Logistic Regression and Random Forest models exhibited a clear bias towards the majority class, which was non-potable water. This bias resulted in higher recall and F1-scores for the non-potable class, while the potable class was less accurately identified. Both models achieved an accuracy of 69.36% and a ROC-AUC of 0.63, indicating reasonable performance but underscoring the need for better identification of potable water samples [38, 40].

The application of SMOTE effectively balanced the class distribution in the training data, which led to improved recall for the potable class in both models. The Random Forest model showed significant improvements after applying SMOTE, achieving an accuracy of 67.07% and a ROC-AUC of 0.64. This suggests that Random Forest is robust in handling class imbalance and is well-suited for the task of classifying water potability. In contrast, the Logistic Regression model experienced a decline in overall accuracy to 50.46% and a ROC-AUC of 0.50 after applying SMOTE, indicating that it may require further tuning or alternative approaches to manage imbalanced data effectively.

This study highlights the importance of addressing class imbalance in machine learning tasks, particularly in critical applications such as water quality assessment. The effective use of SMOTE with the Random Forest model demonstrates how balancing techniques can enhance model performance. These findings emphasize the potential of the Random Forest model as a reliable tool for classifying water potability, providing valuable support for environmental scientists and public health officials in monitoring and ensuring water quality [41-43].

Despite the promising results, this study has limitations that warrant further research. The focus on only two machine learning models suggests that exploring a broader range of models and techniques could yield further improvements. Additionally, the dataset used had inherent limitations, such as missing values and potential measurement errors. Future research could benefit from employing more sophisticated imputation techniques and thorough data cleaning methods to enhance data quality and model performance. Hyperparameter tuning for both models and the exploration of more advanced machine learning models, such as Gradient Boosting Machines or Neural Networks, may also provide further enhancements.

In conclusion, this study provides valuable insights into the application of machine learning for water quality assessment, highlighting the critical role of handling class imbalance to achieve reliable and accurate predictions.

References

1. Martyushev N. V. et al. Energies **16**. 2. 729 (2023)
2. Shutaleva A. et al. Sustainability **15**. 4. 3011 (2023)
3. Rezanov V. A. et al. Metals **12**. 12. 2135 (2022)
4. Martyushev N. V. et al. Materials **16**. 9. 3490 (2023)
5. Kukartsev V. A. et al. Metals **13**. 2. 337 (2023)
6. V. Singh, N. K. Wallia, A. Kudake and A. Raj, *Water Potability Prediction Model Based on Machine Learning Techniques,* 2023 World Conference on Communication & Computing (WCONF), RAIPUR, India, 2023, pp. 1-7, doi: 10.1109/WCONF58270.2023.10235096.

7. Degtyareva, K. et al. *Use of Computer Simulation Tools to Simulate Processes at the Foundry*. 2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE.
8. R. G. de Luna et al., *A Comparative Study of Machine Learning Techniques for Water Potability Classification,* TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON), Chiang Mai, Thailand, 2023, pp. 1345-1350, doi: 10.1109/TENCON58879.2023.10322335.
9. Degtyareva, K. et al. *Automated System for Accounting of Customers and Orders.* In 2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-4). IEEE.
10. Golik V. I. et al. MIAB. Mining Inf. Anal. Bull. **(11-1):**175-189 (2023)
11. Panfilova T. A. et al. MIAB. Mining Inf. Anal. Bull. **(11-1):**239-251 (2023)
12. Suprun E. et al. BIO Web of Conferences **84**. 01008 (2024)
13. Orlov V. et al. E3S Web of Conferences **460**. 07002 (2023)
14. Kravtsov K. et al. E3S Web of Conferences **458**. 09022 (2023)
15. Tynchenko V. S. et al. E3S Web of Conferences **458**. 01011 (2023)
16. Semenova E. et al.. *Using UML to Describe the Development of Software Products Using an Object Approach.* In 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-4). IEEE.
17. Tynchenko V. S. et al. AIP Conference Proceedings **2700**, 1 (2023)
18. Chernykh N. et al. *Comparative Analysis of Existing Measures to Reduce Road Accidents in Western Europe.* In 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE.
19. Volneikina E. et al. *Simulation-Dynamic Modeling Of Supply Chains Based On Big Data.* In 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE.
20. Filina O. A. et al. Energies **17**. 1. 17 (2023)
21. Boychuk, I. P., Grinek, A. V., Martyushev, N. V., Klyuev, R. V., Malozyomov, B. V., Tynchenko, V. S., Kondratiev, S. I. (2023). Energies, **16(24),** 8101.
22. P. Bupe et al., *Embedded Optical Waveguide Sensors for Dynamic Behavior Monitoring in Twisted-Beam Structures,* 2024 IEEE 7th International Conference on Soft Robotics (RoboSoft), San Diego, CA, USA, 2024, pp. 139-144, doi: 10.1109/RoboSoft60065.2024.10521938.
23. Golik V. I. et al. Materials **16**. 21. 7004 (2023)
24. Malozyomov B. V. et al. Energies **16**. 13. 5046 (2023)
25. R. G. de Luna et al., *A Comparative Study of Machine Learning Techniques for Water Potability Classification,* TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON), Chiang Mai, Thailand, 2023, pp. 1345-1350, doi: 10.1109/TENCON58879.2023.10322335.
26. Malashin I. P. et al. Polymers **16**. 1. 115 (2023)
27. Malozyomov B. V. et al. Energies **16**. 13. 4907 (2023)
28. Tynchenko V. S. et al. Journal of Physics: Conference Series **2373**, 6, 062015 (2022)
29. Nelyub V. A. et al. (2023). *Correlation Analysis and Predictive Factors for Building a Mathematical Model.* In Proceedings of the Computational Methods in Systems and Software (pp. 14-25). Cham: Springer International Publishing.

30. Degtyareva K. V. et al. E3S Web of Conferences **458**. 02002 (2023)
31. Gantimurov A. et al. E3S Web of Conferences **431**. 03005 (2023)
32. Tynchenko V. S. et al. E3S Web of Conferences **458**. 07003 (2023)
33. Tynchenko V. et al. *Application of U-Net Architecture Neural Network for Segmentation of Brain Cell Images Stained with Trypan Blue.* In International Conference on High-Performance Computing Systems and Technologies in Scientific Research, Automation of Control and Production (pp. 170-181). Cham: Springer Nature Switzerland.
34. N.Ratnakar, K.Tejeshkumar, K.Venkata Sravani, S.Jeevitha. *Detection of Arc Fault by using ML Algorithm,* 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2024, pp. 1-6, doi: 10.1109/ICAECT60202.2024.10469113.
35. Martyushev, N. V. et al. Materials, **16(13),** 4529 (2023)
36. Nidhi, D. K., Seppä, I., Farahnakian, F., Zelioli, L., Heikkonen, J., Kanth, R. (2023). *Enhancing Minerals Prospects Mapping with Machine Learning: Addressing Imbalanced Geophysical Datasets and Data Visualization Approaches.* In 2023 34th Conference of Open Innovations Association (FRUCT) (pp. 125-135). IEEE.
37. Tynchenko V. et al. Mathematics, **12(2),** 276 (2024).
38. K. E. Chirchi and B. Kavya, *Unraveling Patterns in Healthcare Fraud through Comprehensive Analysis,*2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2024, pp. 585-591, doi: 10.23919/INDIACom61295.2024.10498727.
39. Rogova D. et al. *Software System for Modeling Temperature Distribution During the Electron Beam Welding.* In 2022 IEEE International Conference on Design & Test of Integrated Micro & Nano-Systems (DTS) (pp. 1-6). IEEE.
40. Kurashkin S. et al. Procedia Computer Science, **200**, 83-90 (2022)
41. Tynchenko V. et al. Procedia Computer Science, **200**, 843-851 (2022).
42. Tynchenko V. et al. *Software for modeling brazing process of spacecraft elements from widely used alloys.* In 2022 21st International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE.
43. Panfilova T. A., Kukartsev V. A., Tynchenko V. S.,Mikhalev A. S., Xiaogang Wu, MIAB. Mining Inf. Anal. Bull. (**7-1):**72-82 (2024). DOI: 10.25018/0236_1493_2024_71_0_72