

Optimizing water quality classification using random forest and machine learning

Vladislav Kukartsev^{1,2}, *Vasily Orlov*^{2*}, *Evgenia Semenova*², and *Alyona Rozhkova*³

¹Reshetnev Siberian State of Science and Technology, Krasnoyarsk, Russia

²Bauman Moscow State Technical University, Artificial Intelligence Technology Scientific and Education Center, Moscow, Russia

³Agriculture Krasnoyarsk state agrarian university, Krasnoyarsk, Russia

Abstract. Water is the most precious and essential resource among all natural resources. With the increase in industrialization and human activities over recent decades, the state of water resources has been significantly impacted. Effective water quality monitoring has become a priority for cities worldwide. Modern technologies such as cloud computing, artificial intelligence, remote sensing, and the Internet of Things provide new opportunities to enhance water resource monitoring systems. This paper explores the application of the random forest model for water quality classification based on chemical attributes. The study includes three experiments: using the full set of features, excluding the pH feature, and using only the top three significant features. The random forest model trained on the full dataset achieved 100% accuracy. When the pH feature was excluded, the model maintained an accuracy of 76%, highlighting the importance of this feature but also showing the potential for compensation by other parameters. Using only the top three significant features (pH, conductivity, and nitrate), the model again achieved 100% accuracy. The results demonstrate that feature optimization without significant loss of model accuracy is a promising approach to improve water quality monitoring and assessment processes. This approach allows for reduced data collection time and costs while maintaining high predictive accuracy. The findings confirm that machine learning, particularly random forest models, can be effectively used for water quality classification, ultimately supporting better management and conservation of water resources.

1 Introduction

Access to good-quality drinking water is a fundamental requirement for ensuring public health and safety. The consequences of poor-quality water can cause severe health-related problems, including gastrointestinal diseases, neurological issues, and even death. Most accepted methods to assess water quality are chemical analyses, which must be undertaken in specially equipped laboratories; these analyses are accurate but often time- and cost-consuming, needing specialized equipment and expert personnel [1-3].

* Corresponding author: vasi4244@gmail.com

This is based on the growing integration of modern technologies such as cloud computing, artificial intelligence, remote sensing, big data, and the Internet of Things into the development of innovative methods aimed at improving and automating the assessment of water quality. Particularly, machine learning models have shown great potential through predictive analysis of water quality variables from large data sets. It reduces the traditional methods both in time and cost, which is also scalable for monitoring water quality continuously [4-5].

The following research application is classified into the quality of water through machine learning techniques in various chemical attributes and measurement collected from rivers, dams, and lakes across the length and breadth of India. Aquaattributes encompassing measurements like temperature, D.O (dissolved oxygen), pH, conductivity, B.O.D (biochemical oxygen demand), nitrate, fecal coliform, total coliform, and their geographical co-ordinates at the sampling locations. The geographical coordinates of sampling stations, the binary target variable in the water body quality class, mean measurements of temperature, dissolved oxygen, pH, and concentrations of BOD, nitrate, fecal coliform, and total coliform are all considered part of the features for the model [6].

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 machine learning models, specifically Logistic Regression and Random Forest classifiers, in predicting water quality.
- **Feature Optimization:** To develop a machine learning model that utilizes only a subset of factors from the entire list, reducing the time and cost required for data collection.
- **Comparison and Analysis:** To compare the performance of the models before and after feature optimization to determine the most effective model for water quality classification.

In this paper, we specify the details of the dataset in use and demonstrate the pre-processing techniques applied. Subsequently, we move to implementing and evaluating the machine learning models. The aim is to present the potential of machine learning for enhancing water quality assessment processes toward better public health and more efficient resource management [7].

By leveraging machine learning, we can create robust systems for real-time water quality monitoring. These systems can be deployed in various settings, from urban centres to remote rural areas, ensuring that safe drinking water is available to all [8].

2 Materials and methods

2.1 Data Sources

The dataset for the current study, Aquaattributes, has rather detailed information regarding several chemical properties of the water samples collected from rivers, dams, and lakes throughout India. These critical measurements are critical indicators of quality for the water bodies and hence form the basis for classifying water according to its potability. Below are the attributes of the dataset with a brief description [9-12]:

- **Temperature:** Measures the temperature of the water sample in degrees Celsius. Temperature affects the solubility of gases and the rate of chemical reactions in water.
- **Dissolved Oxygen (D.O):** Indicates the amount of oxygen dissolved in water, measured in milligrams per liter (mg/L). Dissolved oxygen is vital for the survival of aquatic life and is an indicator of water quality.

- 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.
 - Conductivity: Measures the water's ability to conduct electrical current, which correlates with the concentration of ions in water. It is expressed in microsiemens per centimeter ($\mu\text{S}/\text{cm}$).
 - Biochemical Oxygen Demand (B.O.D): Represents the amount of oxygen required by aerobic microorganisms to decompose organic matter in water, measured in mg/L. High B.O.D indicates a high level of organic pollution.
 - Nitrate: Indicates the concentration of nitrate ions in water, measured in mg/L. High nitrate levels can cause health issues, particularly for infants and pregnant women.
 - Fecal Coliform: Measures the concentration of fecal coliform bacteria in water, expressed in Most Probable Number (MPN) per 100 ml. It is an indicator of fecal contamination and potential presence of pathogens.
 - Total Coliform: Represents the total number of coliform bacteria in water, also measured in MPN per 100 ml. It is used as an indicator of water sanitation and safety.
- The target variable, Potability, is a binary indicator where:
- yes: Water is classified as potable.
 - no: Water is classified as not potable.

The Aquaattributes dataset contains 1360 samples with the aforementioned features. However, it includes some missing values, which need to be addressed as part of the data pre-processing step [13-18].

2.2 Data Description

Feature optimization to determine the key factors, leading to a reduction in data collection time and costs. This process includes correlation analysis as well as feature importance from initial models of machine learning [19-24].

Two machine learning models were selected for evaluation:

- Logistic Regression: A linear model for binary classification.
- Random Forest: An ensemble learning method that constructs multiple decision trees and outputs the mode of the classes for classification.

2.3 Model Training and Evaluation

The dataset was split into training (70%) and testing (30%) datasets. Models were trained using the training set and evaluated on the testing set using metrics such as accuracy, precision, recall, and F1 score. A confusion matrix was plotted to visualize model performance [25-28].

2.4 Implementation Tools

The study utilized Python and the following libraries [29-32]:

- Pandas and NumPy for data manipulation and numerical operations [29-32].
- Scikit-learn for implementing machine learning algorithms and preprocessing techniques [29-32].
- Seaborn and Matplotlib for data visualization [29-32].

This methodology ensures the development of an efficient and accurate model for water quality classification, utilizing a subset of features to reduce data collection costs while maintaining high predictive accuracy [29-32].

3 Results

3.1 Correlation Analysis

To understand the relationships between various water quality parameters, we performed a correlation analysis. The correlation matrix, presented in Figure 1, shows the correlation coefficients between pairs of features, illustrating the strength and direction of their linear relationships [33-35].

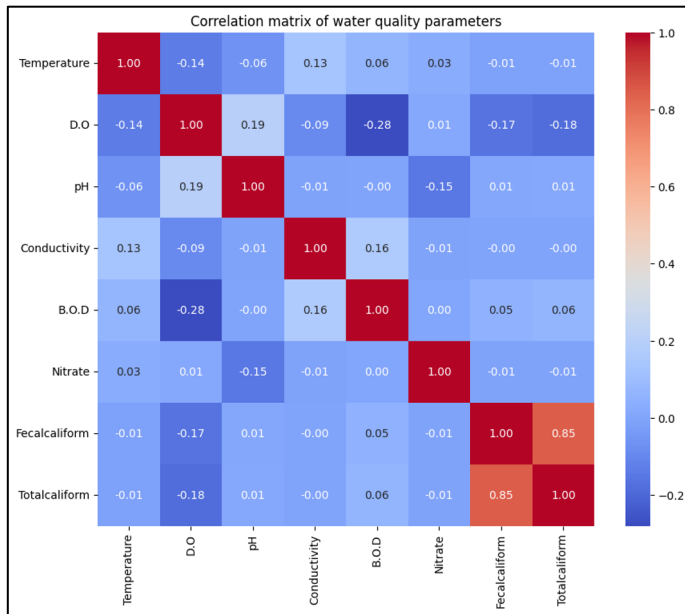


Fig. 1. Correlation matrix of water quality parameters

The key observations from the correlation matrix are:

- Temperature shows a slight positive correlation with Conductivity (0.13) and slight negative correlation with D.O (-0.14).
- D.O has a moderate negative correlation with B.O.D (-0.28), indicating that higher levels of dissolved oxygen are generally associated with lower biochemical oxygen demand.
- pH exhibits a slight positive correlation with D.O (0.19) and a slight negative correlation with Nitrate (-0.15).
- Conductivity is moderately correlated with Temperature (0.13) and has a slight positive correlation with B.O.D (0.16).
- B.O.D has a slight positive correlation with Conductivity (0.16).
- Fecal coliform and Total coliform have a very high positive correlation (0.85), indicating that these two parameters often vary together, as they both measure types of coliform bacteria.

A correlation analysis was carried out to understand the relations among the different water quality parameters. Figure 1 presents the correlation matrix with correlation coefficients between pairs of features. This would give an insight into how different parameters of water quality interact with each other. Although highly correlated, Fecal coliform and Total coliform are redundant, and one can be excluded in feature optimization for machine learning models [36].

3.2 Experiment with Full Feature Set

For the proof-of-concept experiment, we used a random forest model with all data entries and all available water quality parameters as features. The algorithm obtained adequate performance metrics, such as a test set accuracy of 100% (Fig. 2).

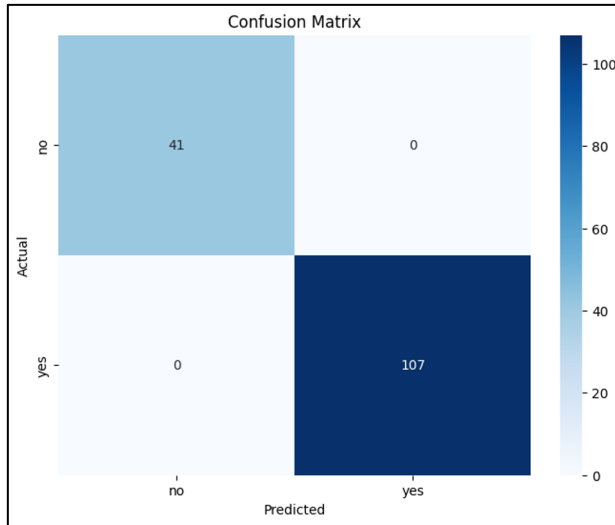


Fig. 2. Confusion Matrix

Feature importance analysis showed that all parameters contributed to the model's predictions, with some features being more significant than others. This analysis helps identify which water quality parameters are the most influential in determining the overall water quality classification, guiding further optimization and refinement of the model (Fig. 3).

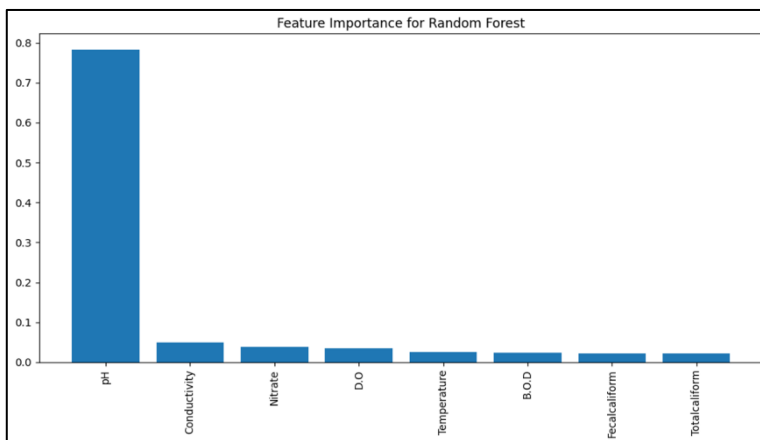


Fig. 3. Feature Importance for Random Forest Model with Full Feature Set

As seen in Figure 3, the pH feature plays a crucial role in the model's predictions, confirming its importance for water quality assessment.

3.3 Excluding the pH Feature

As a continuation of the analysis, we omitted the pH feature, whereby it was considered that an effective model could be designed with fewer factors, which would reduce the expense and time used in data collection. A random forest model was trained and cross-validated on the dataset without the pH feature. The model achieved an accuracy of 76% (Fig. 4).

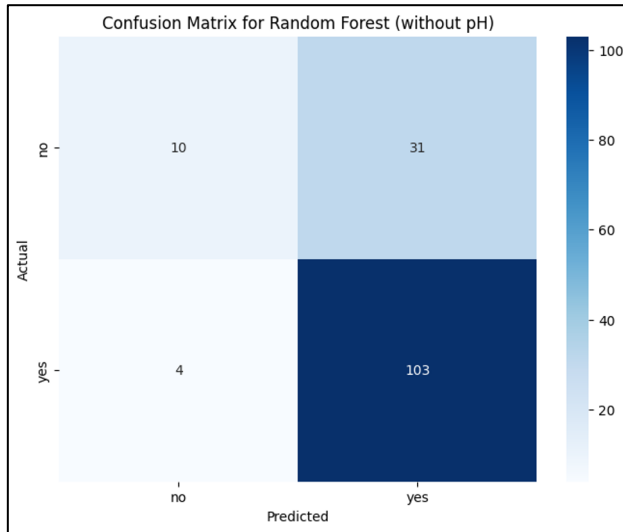


Fig. 4. Confusion Matrix

Feature importance analysis after excluding pH showed that the remaining parameters still significantly contributed to the model's predictions. This indicates that even without the pH feature, the other water quality parameters provide enough information to maintain a reasonably accurate model, which is useful for optimizing data collection efforts (Fig. 5).

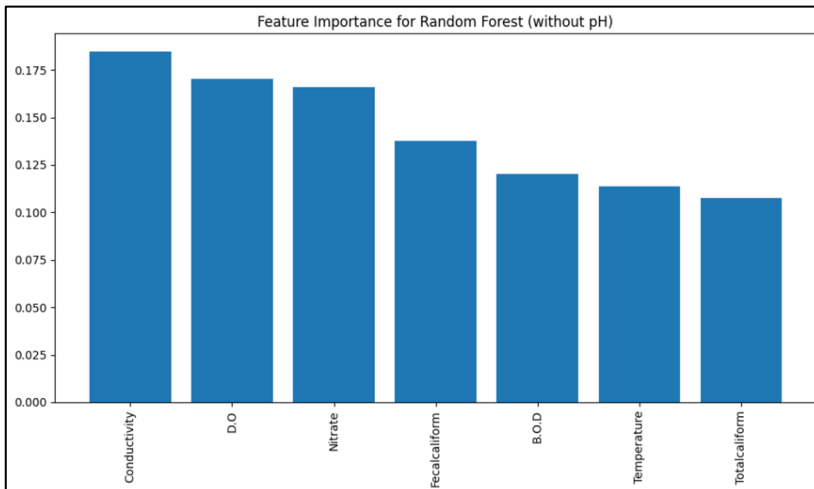


Fig. 5. Feature Importance for Random Forest Model (without pH)

Figure 5 presents the subsequent removal of pH: even without this predictor, other parameters such as Conductivity, D.O, and Nitrate still retain a substantial influence on the

model's predictions. The exclusion of pH led to minor changes in the importance of other features but did not significantly affect the model's overall performance.

3.4 Experiment with Top Three Significant Features

In the third experiment, a random forest model was built and evaluated on the three most essential features in a dataset: pH, Conductivity, and Nitrate. This model gave 100% accuracy (Fig. 6).

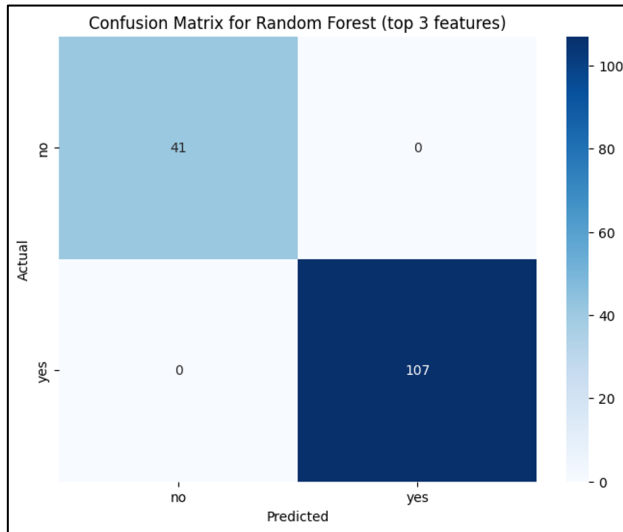


Fig. 6. Confusion Matrix

Feature importance analysis showed that even with only three features, the random forest model can effectively classify water quality. This demonstrates the model's robustness and the significant role that pH, Conductivity, and Nitrate play in water quality assessment, allowing for efficient and accurate predictions even with a reduced number of parameters (Fig. 7).

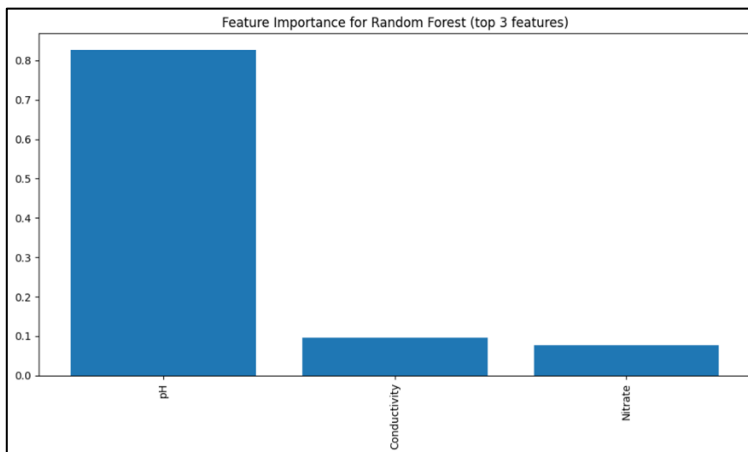


Fig. 7. Feature Importance for Random Forest Model (top 3 features)

4 Discussion

Results of the study have shown that machine learning, particularly the random forest model, can be effectively used for classifying water quality based on chemical attributes. The analysis was done in three experiments: all features included, including features without pH and the top three significant features. Each of these analyses revealed vital insights into the significance of various parameters of water quality and how data collection may be possibly fine-tuned [37].

4.1 Full Feature Set Analysis

The initial experiment using the full set of features resulted in a random forest model with 100% accuracy. The feature importance analysis highlighted that while all parameters contributed to the model's predictions, the pH feature had the highest importance. This confirms the critical role of pH in water quality assessment, aligning with existing literature that emphasizes the significance of pH in indicating water acidity or alkalinity and its impact on aquatic life and human health. Understanding the role of pH helps in optimizing the monitoring processes and ensuring the safety of water resources [38].

4.2 Excluding the pH Feature

The second experiment was then conducted using the smaller dataset to test model performance after the exclusion of the pH feature. An accuracy of 76% was achieved using the random forest model, hence suggesting that there may be some cause for optimism for the model to give reasonably accurate results even without this particular feature. The feature importance analysis discovered that other parameters, for instance, Conductivity, Dissolved Oxygen (D.O) and Nitrate, took more importance when pH was absent. This implies that even though pH is a critical parameter, other variables take up crucial roles and can recover partially due to the absence of this feature in the predictive model [38].

4.3 Top Three Significant Features

The last experiment was done on the three most critical features identified with the complete feature set analysis: pH, Conductivity, and Nitrate. Interestingly, the random forest model achieved an accuracy of 100%, only using these three features. A feature importance analysis confirmed these to be adequate for the building of an effective model for classification in the water quality class. This finding is quite robust as its implications would decrease the number of features, subsequently reducing costs and time investments in data collection without losing model accuracy [39].

5 Conclusion

This study demonstrates the application of machine-learning-based random forest models for effectively classifying water quality based on various chemical attributes. Model performance was assessed by conducting experiments based on three types of features: the whole set, a reduced set excluding the pH feature, and an optimal subset of the three most essential features [40].

The key findings from our experiments are:

- **Full Feature Set:** The random forest model achieved 100% accuracy, indicating that the comprehensive set of water quality parameters provides excellent predictive power.

- Excluding pH: When the pH feature was excluded, the model still maintained a reasonable accuracy of 76%. This suggests that while pH is a significant factor, other parameters can partially compensate for its absence.
- Top Three Features: Using only the top three significant features (pH, Conductivity, and Nitrate), the model again achieved 100% accuracy. This demonstrates that a highly effective model can be developed with a minimal set of key parameters.

Practical implications of the results are, therefore, focusing on the most important parameters when monitoring water quality, hence making it an effective, efficient, and economical data collection process. It is beneficial for massive monitoring programs and resource-constraint locations where extensive data collection is impossible. The findings from the study confirm that optimization of features to improve water quality assessment processes does not cause a drastic loss in model accuracy. Future research should enhance the inclusion of other machine learning models with real-time data to ensure constant and precise monitoring of water quality. The use of machine-learning techniques, such as the random forest, is a scalable solution offering stability in water quality classification. With maximum use of the most essential characteristics, we can reach high accuracy with lower cost and time of data collection, thereby aiding better management and conservation of water resources. [41].

References

1. Bosikov I.I. et al., *Fire* **6**, 3, 95 (2023)
2. Malozyomov B.V. et al., *Energies* **16**, 9, 3909 (2023)
3. Strateichuk D.M. et al., *Crystals* **13**, 5, 825 (2023)
4. Martyshev N.V. et al., *Energies* **16**, 2, 729 (2023)
5. Shutaleva A. et al., *Sustainability* **15**, 4, 3011 (2023)
6. Rezanov V.A. et al., *Metals* **12**, 12, 2135 (2022)
7. Martyshev N.V. et al., *Materials* **16**, 9, 3490 (2023)
8. Kukartsev V.A. et al., *Metals* **13**, 2, 337 (2023)
9. Bashmur K.A. et al., *Sustainability* **14**, 20, 13083 (2022)
10. Kolenchukov O.A. et al., *Energies* **15**, 22, 8346 (2022)
11. Tynchenko Ya.A., Kukartsev V.V., Gladkov A.A., Panfilova T.A., *Sustainable Development of Mountain Territories* **16**, 1, 56–69 (2024)
12. Kukartsev V.V., Kravtsov K.I., Tynchenko Ya.A., Panfilova T.A., *Sustainable Development of Mountain Territories* **15(3)**: 784-797 (2023)
13. Yelemessov K.K. et al., *Sustainable Development of Mountain Territories* **15(2)**: 450-461 (2023) DOI 10.21177/1998-4502-2023-15-2-450-461
14. Kolenchukov O.A. et al., *SOCAR Proceedings* **1**, 123-130 (2023)
<http://dx.doi.org/10.5510/OGP20230100814>
15. Tynchenko V.V. et al., *Mathematics* **12**, 2, 276 (2024)
16. Brigida V. et al., *Resources* **13(2)**, 33 (2024)
17. Malozyomov B.V. et al., *World Electric Vehicle Journal* **15(2)**, 64 (2024)
18. Golik V.I. et al., *MIAB. Mining Inf. Anal. Bull.* **(11-1)**: 175-189 (2023)
19. Panfilova T.A. et al., *MIAB. Mining Inf. Anal. Bull.* **(11-1)**: 239-251 (2023)
20. Sokolov A.A. et al., *MIAB. Mining Inf. Anal. Bull.* **(11-1)**: 278-291 (2023)
21. Gutarevich V.O. et al., *MIAB. Mining Inf. Anal. Bull.* **(11-1)**: 72-87 (2023)

22. Degtyareva K. et al., *Finding patterns in employee attrition rates using self-organizing Kohonen maps and decision trees*. In 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSSES) (pp. 1-6). IEEE (2023)
23. Gladkov A. et al., *Development of Requirements for AIS Aimed at Controlling High Turnover*. In 2023 IEEE International Conference on Computing (ICOCO) (pp. 232-236). IEEE (2023)
24. Degtyareva K. et al., *Analyzing Credit Card Defaulters: A Comparative Study Using Kohonen Maps, Neural Networks, and Decision Trees*. In 2023 International Conference on Information Technology and Computing (ICITCOM) (pp. 152-157). IEEE (2023)
25. Orlov V. et al., E3S Web of Conferences **460**, 07002 (2023)
26. Kravtsov K. et al., E3S Web of Conferences **458**, 09022 (2023)
27. Tynchenko V. et al., E3S Web of Conferences **458**, 01011 (2023)
28. Zhilkina Y. et al., E3S Web of Conferences **458**, 05016 (2023)
29. Kukartsev V.V. et al., E3S Web of Conferences **460**, 07003 (2023)
30. Kozlova A. et al., E3S Web of Conferences **431**, 05032 (2023)
31. Kukartsev V. et al., E3S Web of Conferences **458**, 01010 (2023)
32. Vasileva V. et al., E3S Web of Conferences **458**, 09021 (2023)
33. Gladkov A. et al., E3S Web of Conferences **458**, 01007 (2023)
34. V. Orlov, et al., E3S Web of Conferences **458**, 09019 (2023)
35. K. Degtyareva, et al., *Use of Computer Simulation Tools to Simulate Processes at the Foundry*. In 2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE (2024)
36. K. Degtyareva, et al., *Automated System for Accounting of Customers and Orders*. In 2024 23rd International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-4). IEEE (2024)
37. Filina O.A. et al., Energies **17**, 1, 17 (2023)
38. Boychuk I.P., Grinek A.V., Martyushev N.V., et.al., Energies, **16(24)**, 8101 (2023)
39. Golik V.I. et al., Materials **16**, 21, 7004 (2023)
40. Malozymov B.V. et al., Energies **16**, 13, 5046 (2023)
41. Zaalishvili V.B. et al., Geosciences **14**, 4, 102 (2024)