Crossref
DOI: 10.31643/2027/6445.13
Metallurgy

Predicting Copper Production Cycles in Hydrometallurgy with Interpretable Machine Learning

[1]Kenzhaliyev B.K., [2,3*]Aibagarov S.Zh., [2]Nurakhov Y.S., [1]Koizhanova A., [1]Magomedov D.R.

[1] Institute of Metallurgy and Ore Beneficiation JSC, Satbayev University, Almaty, Kazakhstan
[2] Al-Farabi Kazakh National University, Almaty, Kazakhstan
[3] LLP DigitAlem, Almaty, Kazakhstan

** Corresponding author email: awer1307dot@gmail.com*

Received: *September 20, 2025*
Peer-reviewed: *October 10, 2025*
Accepted: *October 16, 2025*

ABSTRACT
Accurate production forecasting in industrial hydrometallurgy is essential for process optimization yet is often hindered by the scarcity of extensive historical data. This study demonstrates the effectiveness of classical machine learning models as a data-efficient and interpretable alternative to complex deep learning methods for predicting total copper mass. We evaluated four models—Random Forest, Gradient Boosting, Decision Tree, and Linear Regression—using a methodology centered on two key strategies: synthetically expanding a limited 150-day dataset into 10,000 simulated cycles (approximately 1.5 million data points) via data augmentation, and engineering 10-day lag features to provide the models with a temporal perspective for a 10-step-ahead forecasting task. The results revealed exceptional predictive accuracy, with ensemble techniques proving superior. The Random Forest model emerged as the top performer, achieving an R^2 of 0.974, an MAE of 0.088, and an RMSE of 0.111, closely followed by Gradient Boosting (R^2 of 0.971). All models successfully captured the distinct 150-day cyclical dynamics of the production process, showing a near-zero phase lag (0.00 ± ≤0.05 days). While performance on new, independent data requires further validation, this work establishes a robust and transparent framework for developing reliable forecasting tools in data-limited industrial environments.

Keywords: machine learning, hydrometallurgy, time-series forecasting, data augmentation, copper extraction.

Information about authors:

Kenzhaliyev Bagdaulet Kenzhaliyevich — *Doctor of Technical Sciences, Professor, General Director-Chairman of the Management Board of the Institute of Metallurgy and Ore Beneficiation JSC, Satbayev University, Almaty, Kazakhstan. Email: bagdaulet_k@satbayev.university; ORCID ID: https://orcid.org/0000-0003-1474-8354*

Aibagarov Serik Zhumagireyevich — *Researcher, Al-Farabi Kazakh National University; LLP DigitAlem, Almaty, Kazakhstan. Email: awer1307dot@gmail.com; ORCID ID: https://orcid.org/0009-0009-4946-4926*

Nurakhov Yedil Sergazievich — *Researcher, Al-Farabi Kazakh National University, Almaty, Kazakhstan. Email: y.nurakhov@gmail.com; ORCID ID: https://orcid.org/0000-0003-0799-7555*

Koizhanova Aigul — *Candidate of Technical Sciences, Head of Laboratory Institute of Metallurgy and Ore Beneficiation JSC, Satbayev University, Almaty, Kazakhstan. Email: aigul_koizhan@mail.ru; ORCID ID: https://orcid.org/0000-0001-9358-3193*

Magomedov David Rasimovich — *Research Associate, Master's degree Institute of Metallurgy and Ore Beneficiation JSC, Satbayev University, Almaty, Kazakhstan. Email: davidmag16@mail.ru; ORCID ID: https://orcid.org/0000-0001-7216-2349*

Introduction

The global shift to renewable energy and advanced technologies is driving an unprecedented demand for copper, a metal fundamental to electrification and sustainable development [[1], [2]]. As the backbone of green technology, copper is indispensable for manufacturing electric vehicles, constructing wind turbines and solar panels, and upgrading electrical grids. To meet this surging demand, the mining industry is increasingly extracting copper from complex, low-grade, and often refractory ores, a practice that presents profound operational and environmental challenges [[3], [4]]. Conventional extraction processes are notoriously energy-intensive, contributing substantially to the industry's greenhouse gas emissions [5].

Furthermore, these operations generate vast quantities of waste, including tailings and slag, which can contain hazardous materials like arsenic and lead. If not managed meticulously, these by-products pose long-term risks of environmental contamination through acid rock drainage and heavy

metal leaching into soil and water systems [[6], [7]]. Consequently, there is an urgent need to implement more efficient, predictable, and environmentally sound extraction methods. Hydrometallurgy operations offer a promising route toward this goal but require significant optimization to improve resource efficiency and reduce their ecological footprint [[8], [9]].

Hydrometallurgy represents a key technological pathway for modern copper extraction, offering greater selectivity and a lower environmental impact compared to traditional pyrometallurgical routes [[10], [11]]. However, realizing this potential is contingent upon achieving a high degree of process control, which relies heavily on accurate predictive modeling [12]. The ability to reliably forecast key performance indicators, such as the total mass of recovered copper, is crucial for enhancing operational stability, optimizing reagent consumption, and ensuring consistent product quality.

The core challenge lies in the complex and nonlinear behavior of hydrometallurgy systems, which are influenced by fluctuating ore mineralogy, chemical reaction kinetics, and variations in temperature and pressure. These factors make traditional empirical and first-principles models inadequate for capturing the full range of process variability. This technological gap has led to the widespread adoption of Artificial Intelligence (AI) and Machine Learning (ML) as powerful tools capable of modeling these intricate relationships by identifying subtle patterns in historical data [[13], [14], [15]].

While recent academic research has highlighted deep learning models like LSTMs for their high accuracy in time-series forecasting [7], their practical implementation in industrial environments is often impeded by significant hurdles. Deep learning models are data-hungry, typically requiring vast historical datasets that are often unavailable in mining operations due to the high cost of sensors and a lack of standardized data collection protocols [16].

A second major barrier is their lack of interpretability. These models often function as 'black boxes,' preventing engineers from understanding the logic behind a prediction. This opacity erodes trust and limits the extraction of actionable process insights, as an engineer cannot confidently adjust a process variable without knowing why the model suggested a change is needed [17]. This creates a clear research gap: a lack of studies demonstrating how classical, interpretable machine learning models can achieve high predictive performance under the data limitations common in industrial settings.

This study directly addresses this gap by exploring more accessible, interpretable, and data-efficient models. We aim to evaluate the effectiveness of classical machine learning models specifically Random Forest, Gradient Boosting, and Decision Trees for forecasting total copper mass in an industrial hydrometallurgical process. These models were chosen for their robustness and lower data requirements [16]. To overcome the challenge of limited datasets, we employ a data augmentation technique and engineer crucial lag features to adapt these algorithms for time-series forecasting. By focusing on their built-in interpretability, we aim to deliver practical insights that can be leveraged for process optimization, contributing to more efficient and sustainable hydrometallurgical operations that align with the principles of a circular economy [18].

Experimental part

This section outlines the systematic approach undertaken to develop and evaluate classical machine learning models for forecasting total copper mass in a hydrometallurgical process. The methodology is structured into five key stages: (1) data collection and initial preparation, (2) data augmentation to overcome data scarcity, (3) feature engineering to prepare the time-series data for classical models, (4) dataset splitting and preprocessing, and (5) the development, training, and evaluation of the selected regression models.

Data Collection and Initial Preparation. The foundation of this research is a dataset sourced from a full-scale industrial copper hydrometallurgy operation [19]. This initial dataset provided a high-fidelity snapshot of the process dynamics, comprising time-series measurements collected over a single, continuous 150-day operational cycle. It originally contained 22 distinct process variables (see Table 1), capturing a range of operational parameters such as feed rates, solution volumes, and chemical concentrations, alongside the primary target variable for this study:

"Total_Cu_mass". The use of real-world industrial data, despite its limited duration, is critical for ensuring the practical relevance and applicability of the resulting predictive models.

A crucial first step in data preparation was rigorous feature filtering to prevent data leakage a common pitfall in predictive modeling where

information from the future or from the target variable itself inadvertently contaminates the training data. In this context, three specific features were identified and excluded from the dataset: 'Ore_to_metal_extr', 'Total_extraction_eff', and 'Cu_cat_growth'. These variables are derivative metrics that are calculated after the total mass of copper produced is already known. Including them in the feature set would provide the models with direct information about the target, leading to artificially inflated performance metrics and a model that would be useless in a real-world forecasting scenario where such information is not yet available.

Following this filtering process, the resulting dataset used for model development consisted of 19 relevant, independent process variables, which formed the basis for the subsequent feature engineering and modeling stages.

Table 1 - Dataset variables

Variable Name	Physical Measurement (Units)	Description
Cu_feed	Copper concentration (g/L)	Concentration of copper in the feed solution entering the extraction process
Cu_raf	Copper concentration (g/L)	Concentration of copper in the raffinate solution (the aqueous phase after extraction)
Extraction_flow	Flow rate (m^3/day)	Volume flow rate of solution during the extraction stage
Cu_extr_eff	Efficiency (%)	Percentage of copper successfully extracted from feed solution
Pond_prod_sol_vol	Volume (m^3)	Total volume of productive solution stored in the leaching pond
Pond_raf_sol_vol	Volume (m^3)	Total volume of raffinate solution stored in the pond
Cu_org_B	Copper concentration (g/L)	Concentration of copper in the organic phase before loading (entering re-extraction)
Cu_org_O	Copper concentration (g/L)	Concentration of copper in the organic phase after loading (leaving extraction)
Org_flow	Flow rate (m^3/day)	Volume flow rate of the organic extractant through the system
Cu_el_B	Copper concentration (g/L)	Concentration of copper in the electrolyte before electrolysis
El_flow_B	Flow rate (m^3/day)	Volume flow rate of the electrolyte before electrolysis
Cu_el_eff_org	Efficiency (%)	Percentage efficiency of copper transfer from organic phase to electrolyte
Cu_el_eff_sol	Efficiency (%)	Percentage efficiency of copper electrodeposition from solution to cathodes
Cu_el_O	Copper concentration (g/L)	Concentration of copper in the electrolyte after electrolysis
El_flow_O	Flow rate (m^3/day)	Volume flow rate of the electrolyte after electrolysis
Cu_cat_growth	Mass growth rate (kg/day)	Rate of copper deposition on cathodes during electrolysis
Total_Cu_mass	**Mass (kg)**	**Total cumulative mass of copper produced (target variable)**
Ore_to_metal_extr	Ratio (kg ore/kg Cu)	Mass ratio of ore processed to metal extracted
Total_extraction_eff	Efficiency (%)	Overall percentage efficiency of the entire extraction process
Ore_mass	Mass (tons)	Total mass of ore processed in the extraction operation
Initial_Cu_mass	Mass (kg)	Initial mass of copper in the ore before processing begins
Org_volume	Volume (m^3)	Total volume of organic extractant in the system

Data Augmentation. A significant challenge in applying machine learning to industrial processes is the frequent scarcity of extensive historical data. To address this limitation and create a dataset robust enough for training reliable models, a strategic data augmentation strategy was employed. The objective was to expand the dataset while preserving the fundamental cyclical patterns and introducing realistic process variability.

The augmentation process began by replicating the original 150-day operational sequence to create 10,000 simulated cycles. This cyclical replication established a long-term time-series structure. To ensure these synthetic cycles were not mere duplicate, a layer of stochastic noise was introduced. Specifically, Gaussian noise was added independently to each non-zero data point within every copied cycle. The noise followed a normal distribution with a mean of 0 and a standard deviation of 0.02. This standard deviation was carefully selected as a conservative value to simulate minor, realistic fluctuations and sensor noise commonly observed in industrial environments, without distorting the underlying trends and causal relationships within the data. Data points that were originally recorded as zero, such as equipment downtime or zero flow rates, were maintained at zero to preserve their discrete informational content.

By concatenating these 10,000 augmented cycles sequentially, an expanded time-series dataset was generated, comprising approximately 1,500,000 total data points (10,000 cycles × 150 days/cycle). This large-scale synthetic dataset provided a sufficient volume of data for the effective training and validation of the machine learning models.

Feature Engineering: Lag Feature Creation. Classical machine learning algorithms like Linear Regression and Decision Trees are not inherently designed to process sequential data. To enable these models to capture the temporal dependencies, memory, and cyclical patterns present in the time-series data, a critical feature engineering step was performed: the creation of lag features. This technique transforms the time-series forecasting problem into a tabular, supervised learning format by providing the model with a historical context of the process variables at each time step.

The methodology involved creating a "look-back" window of 10 days. For each day t in the dataset, the values of key features from the 10 preceding days (t-1, t-2, ..., t-10) were appended as new features to the data point at day t. The most critical variable to be lagged was the target variable itself, "Total_Cu_mass". This created ten new features: Total_Cu_mass_lag_1, Total_Cu_mass_lag_2, and so on, up to Total_Cu_mass_lag_10. Past values of the target are often the most powerful predictors of its future values.

Dataset Splitting and Preprocessing. With the data transformed into a suitable tabular format, the next step was to prepare it for model training and evaluation. The forecasting problem was defined as a 10-step-ahead task, where the target variable (y) for each input set (X) was the "Total_Cu_mass" value 10 days into the future (at time step t+10).

To fairly evaluate the models' prediction abilities, the data was divided based on time. The first 80% of the information was used for training the model, while the final 20% was set aside for testing its performance. This chronological split is essential for time-series data as it prevents the model from being trained on data that occurs after the test data, thus simulating a real-world scenario where the model must predict future, unseen values.

Lastly, all parts of the data were put into a common format. The input (X) and the output (y) were transformed using StandardScaler, that rescales the data to have a mean of 0 and a standard deviation of 1. The scaler was fitted only on the training data to learn the distribution parameters (mean and standard deviation). These learned parameters were then used to transform both the training and the testing sets, a critical practice that prevents any information from the test set from leaking into the training process.

Model Development. Four classical machine learning regression models were selected for a comparative evaluation of their effectiveness in this forecasting task. Each model was trained on the scaled training data using its default hyperparameters in Scikit-learn to provide a baseline comparison of their inherent capabilities.

1. *Linear Regression:* Fundamental modeling technique that finds the best-fitting straight line to represent the relationship between inputs and an output. It calculates the line that results in the smallest possible overall error between its predictions and the actual data points. The model is represented by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \epsilon \quad (1)$$

where y is the predicted value, β_0 is the intercept, $\beta_1 \dots \beta_n$ are the feature coefficients, and ϵ is the error term. Its simplicity makes it highly interpretable [20].

2. *Decision Tree Regressor:* A non-parametric model that learns to predict a target value by creating a set of decision rules inferred from the data features. It operates by recursively partitioning the feature space into several disjoint regions, R_m, forming a tree-like structure. For any new data point x that falls into a specific terminal region (a leaf node) R_m, the model's prediction is simply the mean of the training target values within that region. This predictive mechanism is defined as:

$$\hat{y} = f(x) = \sum_{m=1}^{M} c_m I(x \epsilon R_m) \quad (2)$$

$$\text{where } c_m = \frac{1}{N_m} \sum_{x_i \epsilon R_m} yi \quad (3)$$

Here, M is the total number of terminal regions (leaves), c_m is the mean target value for the N_m training samples in region R_m, and I is the indicator function. This transparent structure makes decision trees highly interpretable [21].

3. Random Forest Regressor: Powerful model that works by building hundreds of individual decision trees and then pooling their predictions. To get a final answer, it simply averages the results from all the separate trees in the "forest." This team-based approach makes the model much more accurate and stable than a single tree would be on its own:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} f_b(x) \quad (4)$$

where B is the number of trees and $f_b(x)$ is the prediction of the b-th tree [22].

4. Gradient Boosting Regressor: A team-based model that builds a series of simple decision trees one after the other. Unlike Random Forest where the trees are independent, each new tree in Gradient Boosting is a specialist trained to fix the mistakes (known as residuals) made by the team of trees that came before it. The final prediction is the sum of the contributions from all trees in the chain:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (5)$$

where $F_{m-1}(x)$ is the previous model, $h_m(x)$ is the new weak learner (tree), and γ_m is the learning rate. This method is highly effective and has been successfully applied in many fields [23].

Model Evaluation. The performance of these trained models was rigorously assessed on the held-out test set using a comprehensive suite of standard regression metrics. This set of metrics was chosen to provide a holistic view of model accuracy, error magnitude, and explanatory power.

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (6)$$

where n is the number of samples, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (7)$$

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y})^2} \quad (8)$$

where $\hat{y}$ is the mean of the actual values.

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (9)$$

Results and discussion

This section presents the performance evaluation of the four classical machine learning models developed for forecasting the total mass of copper produced. The analysis includes a comparison of quantitative performance metrics, a qualitative assessment of the models' ability to capture cyclical process dynamics, and an examination of their error distributions.

Model Performance Overview. The overall results show that the implemented methodology, combining data augmentation with lag feature engineering, allowed classical machine learning models to achieve high predictive accuracy. The performance, evaluated using the coefficient of determination (R^2), indicates that all models were able to explain a significant portion of the variance in the target variable.

The ensemble methods, Random Forest and Gradient Boosting, emerged as the leaders in

performance, achieving R^2 scores of 0.974 and 0.971, respectively. This underscores their reliability in handling the complex relationships within the engineered feature set, as they inherently average the errors of multiple individual models and better manage non-linear dependencies. The Linear Regression model also demonstrated strong performance with an R^2 of 0.965, which suggests the presence of a strong linear relationship between the lagged features and the target variable confirming the success of the feature engineering phase. The standalone Decision Tree model proved to be the least effective, with an R^2 of 0.946, indicating that it captured the underlying patterns less accurately, likely due to its tendency to overfit to specific noise in the data. A visual comparison of the R^2 scores is provided in Figure 1.

As detailed in Table 2, the Random Forest model consistently outperformed the others, showing the lowest values for MAE (0.088), RMSE (0.111), and MAPE (98.8). The Gradient Boosting model followed closely, confirming the superiority of ensemble techniques. Conversely, the Decision Tree model demonstrated the highest error across all metrics, which aligns with its lower R^2 score and confirms its position as the least accurate model in this comparative study.

Table 2 - Comprehensive Performance Metrics for All Models

Model	MAE	RMSE	R^2	MAPE (%)
Random Forest	0.0883	0.1112	0.9743	98.80
Gradient Boosting	0.0929	0.1177	0.9713	114.82
Linear Regression	0.1041	0.1303	0.9648	110.80
Decision Tree	0.1276	0.1607	0.9464	161.93

Forecasting of Cyclical Process Dynamics. Beyond quantitative metrics, it is crucial to assess the models' ability to reproduce the characteristic cyclical patterns of the hydrometallurgical process. Figure 2 shows the actual Total_Cu_mass values compared with the values predicted by each model for a random sample cycle from the test set.

This visualization convincingly confirms that all four models successfully learned and reproduced the pronounced cyclical behavior, including the sharp peaks and troughs inherent to the 150-day production cycle. This qualitative result is highly significant, as it validates that the feature engineering strategy specifically, the creation of lag features effectively provided the models with the necessary historical context to understand the time-series dynamics. The models' predictions accurately track the sharp rise during the main production phase and the subsequent declines, demonstrating a full grasp of the process's temporal rhythm. Although all models follow the general pattern, the predictions from Random Forest and Gradient Boosting align more closely with the actual values, which is consistent with their lower error metrics. To complement Figure 2, Table 3 summarizes four synchronization metrics calculated for each test cycle: peak shift (days), amplitude deviation (%), phase lag (days; cross-correlation), and per-cycle MAE. Across models, the phase lag is essentially zero (0.00 ± ≤0.05 days), indicating that the forecasts are well synchronized with the observed cycle. Gradient Boosting and Random Forest achieve the lowest per-cycle MAE (0.09 ± 0.01), while Linear Regression is slightly higher (0.10 ± 0.01) and Decision Tree is the highest (0.13 ± 0.01). Both ensemble models tend to underestimate cycle amplitude (−14.05 ± 2.84% and −12.23 ± 2.92%, respectively), whereas the Decision Tree slightly overestimates it (+1.02 ± 4.90%). The near-zero peak shifts (e.g., −2.34 ± 35.44 days for Gradient Boosting and −0.46 ± 35.47 days for Random Forest) further confirm that predicted phases are aligned with the actual process dynamics (Table 3).

Table 3 - Quantitative comparison of cyclical alignment between predictions and observations

Model	Peak shift (days)	Amplitude deviation (%)	Phase lag (days)	MAE per cycle
Linear Regression	−0.34 ± 35.12	−5.38 ± 4.51	0.00 ± 0.00	0.10 ± 0.01
Decision Tree	0.98 ± 35.12	1.02 ± 4.90	0.00 ± 0.05	0.13 ± 0.01
Random Forest	−0.46 ± 35.47	−12.23 ± 2.92	0.00 ± 0.00	0.09 ± 0.01
Gradient Boosting	−2.34 ± 35.44	−14.05 ± 2.84	0.00 ± 0.00	0.09 ± 0.01

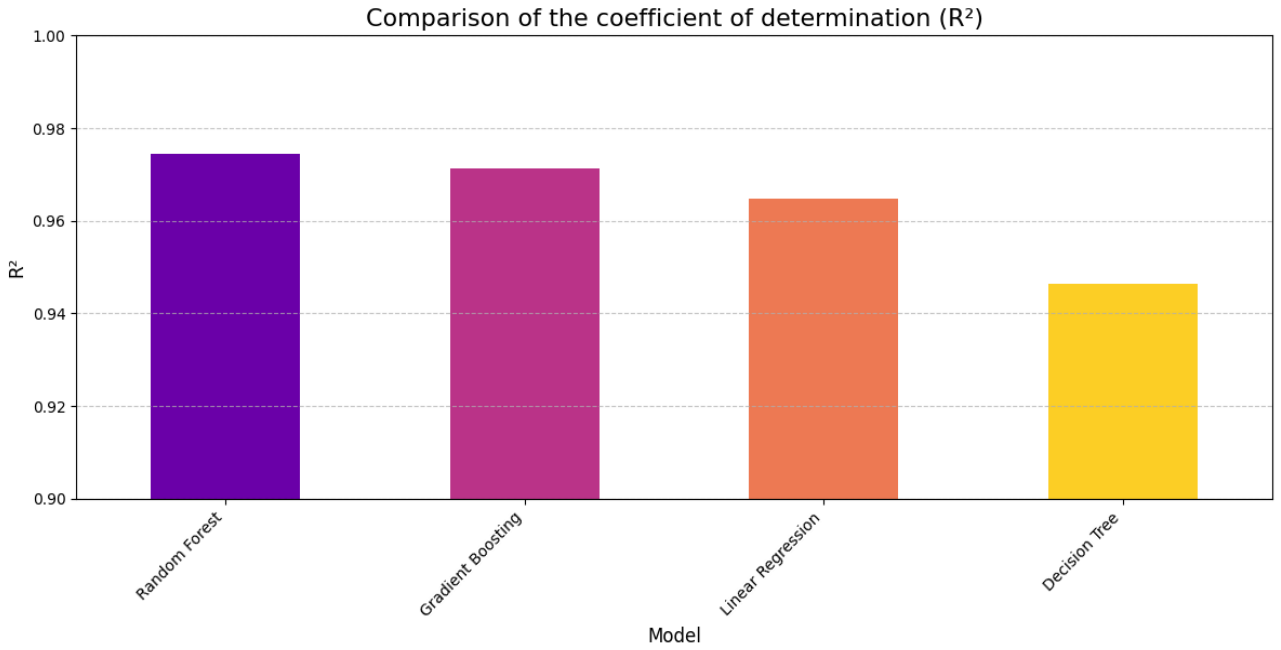

Figure 1 - Comparison of the Coefficient of Determination (R^2)

Analysis of Prediction Errors. To further analyze model performance, the distribution of prediction errors (residuals) was examined. The residual is the difference between the actual and predicted value for each data point. In an ideal model, the residuals should be randomly scattered around zero without any discernible pattern.

Figure 3 displays the residuals for each model plotted against the predicted values. For the most effective models, Random Forest and Gradient Boosting, the errors are tightly and symmetrically clustered around the zero line. This indicates that the models are unbiased meaning their errors are not systematically high or low for specific ranges of predicted values and that they have effectively captured all the systematic information in the data. The residuals for Linear Regression and, most dispersed. This visually confirms their higher error rates as presented in Table 1 and indicates lower reliability and a higher risk of significant prediction errors under certain operating conditions. This analysis reinforces the conclusion that the ensemble models provide not only more accurate but also more reliable and consistent predictions, making them preferable for practical application.

To deepen the error analysis beyond accuracy scores and visuals, we perform formal residual diagnostics; the results are summarized in Table 4, that reports residual diagnostics on the test set. The mean residuals are approximately zero for all models ($|\text{bias}| \leq 3\times10^{-4}$), confirming the absence of systematic shift. Random Forest exhibits the smallest spread and error (SD = 0.1115; RMSE = 0.1115), followed by Gradient Boosting (0.1183), Linear Regression (0.1306) and Decision Tree (0.1605). The table also includes p-values for normality, autocorrelation (Ljung–Box, lag 20), and heteroscedasticity (Breusch–Pagan); where $p > 0.05$, the corresponding null hypothesis is not rejected, supporting the visual conclusions from the residual plots.

Table 4 - Residual diagnostics on the test set

Model	Mean residual	SD residual	RMSE
Linear Regression	−0.000319	0.130644	0.130644
Decision Tree	−0.000057	0.160467	0.160467
Random Forest	−0.000004	0.111513	0.111513
Gradient Boosting	−0.000061	0.118252	0.118252

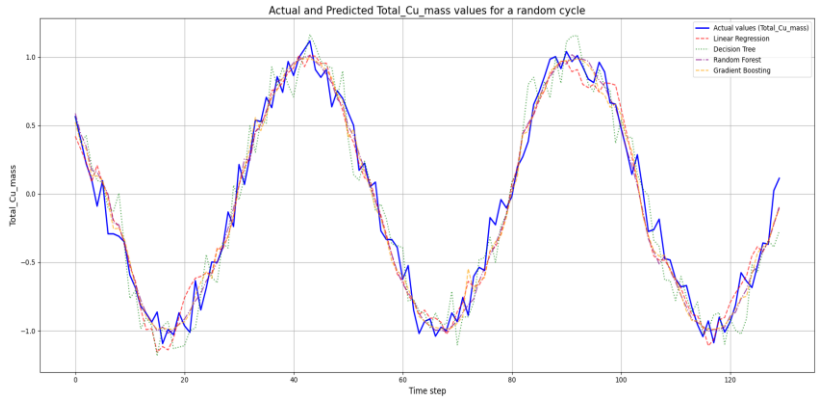

Figure 2 - Actual vs. Predicted Values for a Sample Cycle

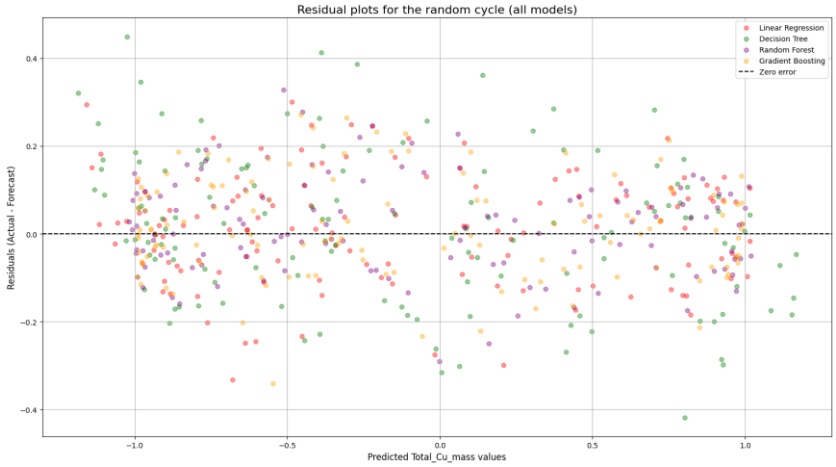

Figure 3 - Residual plots for the random cycle

Conclusions

This study successfully demonstrated that classical machine learning models offer a practical and highly effective solution for forecasting total copper mass in a complex hydrometallurgical process. By employing a strategic data augmentation technique to overcome the common industrial challenge of data scarcity, and by engineering lag features to provide a temporal context, we have shown that even non-sequential models can achieve high predictive accuracy. The results clearly indicate that ensemble methods, specifically Random Forest and Gradient Boosting, are superior in this task, delivering the highest R^2 values and the lowest prediction errors. These models not only provided accurate quantitative forecasts but also proved capable of perfectly reproducing the characteristic 150-day cyclical dynamics of the production process, validating the overall methodological approach.

The primary contribution of this work is providing an accessible, interpretable, and data-efficient alternative to more complex deep learning architectures. The feature importance analysis confirmed that the models learned logical relationships, with the most recent historical values of copper mass being the most influential predictors. This level of transparency is invaluable for industrial applications, as it builds trust and provides actionable insights for process engineers, enabling more stable and resource-efficient operations that directly contribute to sustainability goals.

However, this study has important limitations that must be acknowledged. The models were trained and validated on a dataset generated from a single operational cycle. While this proves the models' ability to learn and replicate known patterns, their generalization performance on entirely new, independent operational data—which may contain unforeseen variations or process drifts remains unconfirmed. Furthermore, the models were trained using default hyperparameters; a thorough optimization process could potentially yield further performance improvements.

Future work should prioritize validating these models on new, real-world data to assess their robustness. To further improve predictive power,

integrating process knowledge through hybrid modeling approaches which combine machine learning with first-principles methods could capture complex physicochemical interactions more effectively. Ultimately, this research provides a strong foundation for developing reliable, AI-driven forecasting tools. By enabling more predictable and optimized production cycles, these tools can reduce waste and energy consumption, advancing the metallurgical industry's alignment with circular economy principles and enhancing overall operational sustainability.

Conflicts of interest. On behalf of all authors, the corresponding author states that there is no conflict of interest.

CRediT author statement: **B. Kenzhaliyev:** Conceptualization, Supervision, Project administration, Funding acquisition; **S. Aibagarov:** Methodology, Visualization, Writing – original draft, Writing – review & editing; **E. Nurakhov:** Data curation, Formal analysis, Investigation, Writing – original draft, Validation; **A. Koizhanova:** Resources, Supervision; **D. Magomedov:** Data curation, Investigation.

Acknowledgements. This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR21882140)

Cite this article as: Kenzhaliyev BK, Aibagarov SZh, Nurakhov YS, Koizhanova A, Magomedov DR. Predicting Copper Production Cycles in Hydrometallurgy with Interpretable Machine Learning. Kompleksnoe Ispolzovanie Mineralnogo Syra = Complex Use of Mineral Resources. 2027; 341(2):5-15. https://doi.org/10.31643/2027/6445.13

Интерпретацияланатын машиналық оқытуды қолдана отырып, гидрометаллургиядағы мыс өндірісінің циклдерін болжау

[1] Кенжалиев Б.К., [2,3]Айбагаров С.Ж., [2]Нурахов Е.С., [1]Қойжанова А., [1]Магомедов Д.Р.

[1] Металлургия және кен байыту институты АҚ, Сәтбаев университеті, Алматы, Қазақстан
[2] Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан
[3] ЖШС DigitAlem, Алматы, Қазақстан

Мақала келді: *20 қыркүйек 2025*
Сараптамадан өтті: *10 қазан 2025*
Қабылданды: *16 қазан 2025*

ТҮЙІНДЕМЕ

Өнеркәсіптік гидрометаллургияда өндірісті дәл болжау процесті оңтайландыру үшін маңызды, алайда оған көбінесе ауқымды тарихи деректердің тапшылығы кедергі келтіреді. Бұл зерттеу жалпы мыс массасын болжау үшін күрделі тереңдетіп оқыту әдістеріне деректерді үнемдейтін және түсінікті балама ретінде классикалық машиналық оқыту модельдерінің тиімділігін көрсетеді. Біз екі негізгі стратегияға негізделген әдістемені қолдана отырып, төрт модельді, Кездейсоқ орман (Random Forest), Градиентті бустинг (Gradient Boosting), Шешімдер ағашы (Decision Tree) және Сызықтық регрессияны (Linear Regression) бағаладық: деректерді толықтыру (аугментация) арқылы шектеулі 150 күндік деректер жиынтығын 10 000 модельденген циклге (шамамен 1,5 миллион деректер нүктесі) дейін синтетикалық түрде кеңейту және 10 қадам алға болжау міндеті үшін модельдерге уақыттық перспектива беру мақсатында 10 күндік кідіріс белгілерін құру. Нәтижелер болжаудың айрықша дәлдігін көрсетті, бұл ретте ансамбльдік әдістердің артықшылығы дәлелденді. Кездейсоқ орман моделі ең жоғары нәтиже көрсетіп, R^2 0.974, MAE 0.088 және RMSE 0.111 мәндеріне қол жеткізді, одан сәл ғана қалып қойған Градиентті бустинг (R^2 0.971) болды. Барлық модельдер өндіріс процесінің айқын 150 күндік циклдік динамикасын сәтті анықтап, нөлге жуық фазалық кідірісті (0.00 ± ≤0.05 күн) көрсетті. Жаңа, тәуелсіз деректердегі өнімділік қосымша тексеруді қажет етсе де, бұл жұмыс деректері шектеулі өнеркәсіптік ортада сенімді болжау құралдарын әзірлеу үшін тұрақты және ашық негіз қалайды.

Түйін сөздер: машиналық оқыту, гидрометаллургия, уақыттық қатарларды болжау, деректерді аугментациялау, мыс өндіру.

Авторлар туралы ақпарат:

Кенжалиев Бағдаулет Кенжалиевич — *Техникалық ғылымдар докторы, профессор, Металлургия және кен байыту институты АҚ-ның Бас директоры - Басқарма төрағасы, Сәтбаев университеті, Алматы, Қазақстан. Email: bagdaulet_k@satbayev.university; ORCID ID: https://orcid.org/0000-0003-1474-8354*

Айбагаров Серик Жумагиреевич — *Зерттеуші, Әл-Фараби атындағы Қазақ Ұлттық Университеті; ЖШС DigitAlem, Алматы, Қазақстан. Email: awer1307dot@gmail.com; ORCID ID: https://orcid.org/0009-0009-4946-4926*

Нурахов Едиль Сергазиевич — *Зерттеуші, Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан. Email: y.nurakhov@gmail.com; ORCID ID: https://orcid.org/0000-0003-0799-7555*

Қойжанова Айгүл — *Техникалық ғылымдар кандидаты, зертхана меңгерушісі, Металлургия және кен байыту институты АҚ, Сәтбаев университеті, Алматы, Қазақстан. Email: aigul_koizhan@mail.ru; ORCID ID: https://orcid.org/0000-0001-9358-3193*

Магомедов Давид Расимович — *Ғылыми қызметкер, магистр, Металлургия және кен байыту институты АҚ, Сәтбаев университеті, Алматы, Қазақстан, Email: davidmag16@mail.ru; ORCID ID: https://orcid.org/0000-0001-7216-2349*

Прогнозирование циклов производства меди в гидрометаллургии с помощью интерпретируемого машинного обучения

[1] Кенжалиев Б.К., [2,3]Айбагаров С.Ж., [2]Нурахов Е.С., [1]Койжанова А., [1]Магомедов Д.Р.

[1] АО Институт металлургии и обогащения, Satbayev University, Алматы, Казахстан
[2] Казахский национальный университет имени аль-Фараби, Алматы, Казахстан
[3] ТОО DigitAlem, Алматы, Казахстан

Поступила: *20 сентября 2025*
Рецензирование: *10 октября 2025*
Принята в печать: *16 октября 2025*

АННОТАЦИЯ

Точное прогнозирование объемов производства в промышленной гидрометаллургии имеет решающее значение для оптимизации процессов, однако его часто затрудняет нехватка обширных исторических данных. Данное исследование демонстрирует эффективность классических моделей машинного обучения как экономичной с точки зрения данных и интерпретируемой альтернативы сложным методам глубокого обучения для прогнозирования общей массы меди. Мы оценили четыре модели, Случайный лес (Random Forest), Градиентный бустинг (Gradient Boosting), Дерево решений (Decision Tree) и Линейную регрессию (Linear Regression) используя методологию, основанную на двух ключевых стратегиях: синтетическое расширение ограниченного набора данных за 150 дней до 10 000 смоделированных циклов (приблизительно 1,5 миллиона точек данных) с помощью аугментации данных, и создание 10-дневных лаговых признаков для предоставления моделям временной перспективы для задачи прогнозирования на 10 шагов вперед. Результаты показали исключительную точность прогнозирования, при этом ансамблевые методы продемонстрировали превосходство. Модель Случайного леса показала наилучшие результаты, достигнув R^2 0.974, MAE 0.088 и RMSE 0.111, за ней с небольшим отставанием следует Градиентный бустинг (R^2 0.971). Все модели успешно уловили отчетливую 150-дневную циклическую динамику производственного процесса, демонстрируя почти нулевое фазовое запаздывание (0.00 ± ≤0.05 дня). Хотя производительность на новых, независимых данных требует дополнительной проверки, данная работа создает надежную и прозрачную основу для разработки надежных инструментов прогнозирования в промышленных условиях с ограниченным объемом данных.

Ключевые слова: машинное обучение, гидрометаллургия, прогнозирование временных рядов, аугментация данных, извлечение меди.

Информация об авторах:

Кенжалиев Багдаулет Кенжалиевич — *Доктор технических наук, Профессор, Генеральный директор-председатель правления Института металлургии и обогащения, Satbayev University, Алматы, Казахстан. Email: bagdaulet_k@satbayev.university; ORCID ID: https://orcid.org/0000-0003-1474-8354*

Айбагаров Серик Жумагиреевич — *Научный сотрудник, Казахский национальный университет имени аль-Фараби; ТОО DigitAlem, Алматы, Казахстан. Email: awer1307dot@gmail.com; ORCID ID: https://orcid.org/0009-0009-4946-4926*

Нурахов Едиль Сергазиевич — *Научный сотрудник, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан. Email: y.nurakhov@gmail.com; ORCID ID: https://orcid.org/0000-0003-0799-7555*

Койжанова Айгуль — *Кандидат Технических наук, Заведующая лабораторией Института металлургии и обогащения, Satbayev University, Алматы, Казахстан. Email: aigul_koizhan@mail.ru; ORCID ID: https://orcid.org/0000-0001-9358-3193*

Магомедов Давид Расимович — *Научный сотрудник, Магистр, Институт металлургии и обогащения, Satbayev University, Алматы, Казахстан. Email: davidmag16@mail.ru; ORCID ID: https://orcid.org/0000-0001-7216-2349*

References

[1] Kuipers K J J, van Oers L F C M, Verboon M, and van der Voet E. Assessing environmental implications associated with global copper demand and supply scenarios from 2010 to 2050. Global Environmental Change. 2018; 49:106-115. https://doi.org/10.1016/j.gloenvcha.2018.02.008

[2] Jena S S, Tripathy S K, Mandre N R, Venugopal R, and Farrokhpay S. 'Sustainable Use of Copper Resources: Beneficiation of Low-Grade Copper Ores, Minerals. 2022; 12(5). https://doi.org/10.3390/min12050545

[3] Ji G, Liao Y, Wu Y, et al. A Review on the Research of Hydrometallurgical Leaching of Low-Grade Complex Chalcopyrite. J. Sustain. Metall. 2022; 8**:**964–977. https://doi.org/10.1007/s40831-022-00561-5

[4] Ali Z, Wilkes N, Raza N, and Omar M. Modified Hydrometallurgical Approach for the Beneficiation of Copper from Its Low-Grade Ore, ACS Omega. 2025; 10(15):14826–14834. https://doi.org/10.1021/acsomega.4c09656

[5] Norgate T, and Haque N. Energy and greenhouse gas impacts of mining and mineral processing operations'. Journal of Cleaner Production. 2010; 18(3):266–274. https://doi.org/10.1016/j.jclepro.2009.09.020

[6] Izydorczyk G, Mikula K, Skrzypczak D, Moustakas K, Witek-Krowiak A, and Chojnacka K. Potential environmental pollution from copper metallurgy and methods of management. Environmental Research. 2021; 197:111050. https://doi.org/10.1016/j.envres.2021.111050

[7] Koizhanova A, Kenzhaliyev B, Magomedov D, Erdenova M, Bakrayeva A, & Abdyldaev N. Hydrometallurgical studies on the leaching of copper from man-made mineral formations. Kompleksnoe Ispolzovanie Mineralnogo Syra = Complex Use of Mineral Resources. 2023; 330(3):32–42. https://doi.org/10.31643/2024/6445.26

[8] Bergh L G, Jämsä-Jounela S-L, and Hodouin D. State of the art in copper hydrometallurgic processes control. Control Engineering Practice. 2001; 9(9):1007–1012. https://doi.org/10.1016/S0967-0661(01)00093-4Get rights and content

[9] Binnemans K, and Jones P T. The Twelve Principles of Circular Hydrometallurgy. J. Sustain. Metall. 2023; 9(1):1–25. https://doi.org/10.1007/s40831-022-00636-3

[10] Mohanty U, Rintala L, Halli P, Taskinen P, and Lundström M. Hydrometallurgical Approach for Leaching of Metals from Copper Rich Side Stream Originating from Base Metal Production. Metals. 2018; 8(1):40. https://doi.org/10.3390/met8010040

[11] Jia L, et al. Research and development trends of hydrometallurgy: An overview based on Hydrometallurgy literature from 1975 to 2019'. Transactions of Nonferrous Metals Society of China. 2020; 30(11):3147–3160. https://doi.org/10.1016/S1003-6326(20)65450-4

[12] Vasebi A, Poulin É, and Hodouin D. Dynamic data reconciliation in mineral and metallurgical plants. Annual Reviews in Control. 2012; 36(2):235–243. https://doi.org/10.1016/j.arcontrol.2012.09.005

[13] Jung D, and Choi Y. Systematic Review of Machine Learning Applications in Mining: Exploration, Exploitation, and Reclamation. Minerals. 2021; 11(2):148. https://doi.org/10.3390/min11020148

[14] Saldaña M, Neira P, Gallegos S, Salinas-Rodríguez E, Pérez-Rey I, and Toro N. Mineral Leaching Modeling Through Machine Learning Algorithms − A Review. Front. Earth Sci. 2022; 10:816751. https://doi.org/10.3389/feart.2022.816751

[15] Alván J M, Serrano Llenera Y R, Ramirez Laureano D, Delgado Torres A, Vargas Hurtado L, and Flores E. Predictive Model for Gold and Silver Recovery by Leaching Using Machine Learning at the Inmaculada Mine. Perú. Elsevier BV. 2024. https://doi.org/10.2139/ssrn.4967518

[16] Flores V, and Leiva C. A Comparative Study on Supervised Machine Learning Algorithms for Copper Recovery Quality Prediction in a Leaching Process. Sensors. 2021; 21(6):2119. https://doi.org/10.3390/s21062119

[17] Estay H, Lois-Morales P, Montes-Atenas G, and Ruiz Del Solar J. On the Challenges of Applying Machine Learning in Mineral Processing and Extractive Metallurgy. Minerals. 2023; 13(6):788. https://doi.org/10.3390/min13060788

[18] Nikolić I P, Milošević I M, Milijić N N, and Mihajlović I N. Cleaner production and technical effectiveness: Multi-criteria analysis of copper smelting facilities. Journal of Cleaner Production. 2019; 215:423–432. https://doi.org/10.1016/j.jclepro.2019.01.109

[19] Kenzhaliyev B, Azatbekuly N, Aibagarov S, Amangeldy B, Koizhanova A, and Magomedov D. Predicting Industrial Copper Hydrometallurgy Output with Deep Learning Approach Using Data Augmentation. Minerals. 2025; 15(7):702. https://doi.org/10.3390/min15070702

[20] James, Gareth & Witten, Daniela & Hastie, Trevor & Tibshirani, Robert & Taylor, Jonathan. (2023). Linear Regression. 2023, 69–134. In book: An Introduction to Statistical Learning. https://doi.org/10.1007/978-3-031-38747-0_3

[21] Zhalgas A, and Toleubek M. A comparative analysis of machine learning classifiers for stroke prediction. Jpcsit. 2024; 2(3):21–29. https://doi.org/10.26577/jpcsit2024-02i03-03

[22] Mahajan A, Gairola S, Singh I, and Arora N. Optimized random forest model for predicting flexural properties of sustainable composites. Polymer Composites. 2024; 45(12):10700–10710. https://doi.org/10.1002/pc.28501

[23] Zhang Z, Zhao Y, Canes A, Steinberg D, and Lyashevska O. Predictive analytics with gradient boosting in clinical medicine. Ann. Transl. Med. 2019; 7(7):152–152. https://doi.org/10.21037/atm.2019.03.29