Journal of Kryvyi Rih National University

Journal homepage: https://journal-knu.com.ua/en

Vol. 22, No. 2, 2024

Article's History:
Received: 06.08.2024
Revised: 17.11.2024
Accepted: 24.12.2024

UDC 620.97

DOI: 10.31721/2306-5451-2024-2-22-10-27

Comprehensive methodology for creating enhanced datasets for modelling the process of magnetic separation of iron ore

Oleksandr Volovetskyi*

Postgraduate Student
Kryvyi Rih National University
50027, 11 Vitalii Matusevich Str., Kryvyi Rih, Ukraine
https://orcid.org/0009-0003-1703-387X

Abstract. The study presents an innovative approach to creating extended datasets for modelling magnetic separation of iron ore, which is crucial for enhancing efficiency and automating enrichment processes in the mining industry. The aim of the research was to develop a methodology for creating extended datasets for modelling magnetic separation of iron ore that takes into account the specifics of Ukrainian deposits and allows for the generation of representative data in conditions of limited real production data by integrating physical modelling with machine learning methods. Research methods: modelling using mathematical learning, simulation based on physical processes, statistical analysis. The study examined the use of the USIM PAC simulator for modelling the iron ore enrichment system and adapting data for magnetic enrichment, ensuring the accuracy of modelling technological enrichment processes. The simulator was used to obtain a dataset from physical modelling of part of the enrichment process based on data from the Valyavkinske deposit. Primary modelling of the dataset was analysed, including statistical characteristics, distribution shape, and normality tests to identify fields requiring correction. Based on the analysis results, specific requirements for data distribution in the new dataset to be formed for further use were established. In accordance with these requirements, several mathematical models were implemented to reproduce the specified criteria and parameters. For each data field, the best model was carefully selected, and the dataset was corrected based on its data to bring the distribution as close as possible to the desired one. Comprehensive validation of the resulting corrected data was conducted, emphasising the preservation of the physical validity of the data and their correspondence to real enrichment processes. A detailed analysis of the corrected data was performed, as well as the statistical characteristics of the resulting dataset, confirming the effectiveness of the developed comprehensive methodology for modelling and adapting data for magnetic enrichment of iron ore. The methodology holds practical value due to its innovative approach to creating extended datasets for modelling magnetic separation of iron ore, enhancing the efficiency and automation of enrichment processes while considering the specifics of deposits and generating representative data in conditions of limited real data

Keywords: nonlinear modelling of enrichment; separation control; machine learning in enrichment; automation of enrichment processes; simulation of technological parameters

Introduction

The relevance of this work is determined by the necessity to improve the processes of modelling magnetic separation of iron ore under conditions of limited real data, especially for Ukrainian deposits. The specifics of Ukrainian iron ore deposits, particularly in the Kryvyi Rih basin, require the development of special

Suggested Citation:
Volovetskyi, O. (2024). Comprehensive methodology for creating enhanced datasets for modelling the process of magnetic separation of iron ore. *Journal of Kryvyi Rih National University*, 22(2), 10-27. doi: 10.31721/2306-5451-2024-2-22-10-27.

*Corresponding author

extraction and processing technologies. The development of a comprehensive methodology for creating extended datasets will help overcome these limitations, taking into account local geological conditions and ensuring more effective modelling of the magnetic separation process. This will contribute to the optimisation of enrichment processes, improvement of concentrate quality, and reduction of energy consumption in the mining industry of Ukraine.

Current trends in modelling magnetic separation are characterised by a comprehensive approach that integrates various computational and analytical methods. These approaches encompass a wide range from classical numerical methods to advanced techniques in machine learning and multiphysics modelling. Significant progress has been made in developing methods that allow for the simultaneous consideration of complex interactions between different physical processes inherent in magnetic separation. Concurrently, optimisation methods for control are being developed, aimed at enhancing the efficiency of enrichment processes. The integration of these diverse approaches creates a powerful foundation for developing adaptive and high-precision models capable of functioning under conditions of limited experimental data and accounting for the specifics of local conditions.

In the study by V.V. Shenoy *et al*. (2024), the influence of the magnetic field on flow behaviour in a step geometry is examined. Using modern computational fluid dynamics methods, particularly the open-source package OpenFOAM, the authors investigated the interaction between magnetic forces and geometric factors affecting flow characteristics. The study revealed important patterns in flow behaviour under the influence of the magnetic field and geometry. The proposed mathematical models allow for the prediction of key flow parameters under various conditions. The results of this CFD work have the potential for application in a wide range of engineering tasks related to magnetohydrodynamic flows and boundary layer control.

In the work by R. Chowdhury *et al*. (2024), a comprehensive method for optimising medium parameters for effective material separation in a hydrocyclone separator is proposed, combining theoretical approaches and CFD modelling. The use of CFD allowed for a detailed analysis of the impact of medium density on the separation of PVC and PET particles, visualising and assessing key process parameters. Although the study focused on plastics, the methodology can be applied in various fields, including magnetic enrichment, where consideration of medium and particle properties is critical for optimising separation.

M.E. Kinaci *et al*. (2020) investigated the process of indirect reduction of iron ore in fluidised beds using the discrete element method (DEM) in conjunction with computational fluid dynamics. The developed model can be applied to simulate the processes of iron ore reduction in industrial reactors, optimising process parameters and developing new iron production technologies. J. Liu *et al*. (2021) studied the magnetic separation process in an aerodynamic drum magnetic separator (ADMS) using the finite element method and multiphysics modelling in COMSOL Multiphysics software. The modelling of the magnetic field, airflow, and particle movement in the separator was conducted. The influence of various parameters (air velocity, magnetic field intensity, positioning of magnetic poles) on the separation efficiency of magnetic and non-magnetic particles was demonstrated. The simulation results were verified through experimental measurements and calculations. The proposed model allows for the prediction of particle trajectories and extraction probabilities under different conditions, which can be useful for precise control of the magnetic separation process using combined force fields.

Moreover, there is a growing interest in the application of machine learning methods, particularly convolutional neural networks (CNNs), which open new opportunities for predicting separation efficiency under complex conditions. Research conducted by Y. Li *et al*. (2022) demonstrates the successful use of CNNs for modelling grinding processes in ball mills, based on externally measured process variables. These approaches can be adapted for magnetic separation, enhancing prediction accuracy and reducing the need for large volumes of experimental data. The implementation of machine learning fosters the development of hybrid models that combine theoretical knowledge with data from discrete element method simulations, providing more effective and rapid modelling of complex systems.

N. Yang *et al*. (2022) analysed the development of modelling methods for mineral deposits, emphasising the transition to three-dimensional digital models and the importance of understanding ore formation processes. The authors highlighted the application of machine learning methods, particularly convolutional neural networks, for predicting hidden deposits. They stressed the issue of data scarcity and proposed the use of advanced machine learning techniques to process incomplete data, underscoring the importance of integrating expert knowledge. Despite progress in modelling magnetic separation, there remains a need for the development of comprehensive methodologies for creating accurate models under conditions of limited real data.

V. Morkun *et al*. (2020) investigated the identification of nonlinear dynamic enrichment objects using a second-order Volterra model and its projection onto orthonormal Laguerre basis functions. This potentially impacts the improvement of modelling accuracy for iron ore enrichment processes, reducing model complexity and sensitivity to noise. O. Porkuian *et al*. (2019) considered the development of a predictive control system for the iron ore enrichment process based on a hybrid Hammerstein model. The model combines a

fuzzy nonlinear block and a crisp linear dynamic block for effective approximation of nonlinear, dynamic, and non-stationary properties of enrichment line objects. The proposed algorithms ensure rapid real-time identification and optimal control considering constraints, leading to improved concentrate quality and reduced energy consumption.

S. Rajendran & C.V.G.K. Murty (2023) reviewed modern approaches to numerical modelling of enrichment processes for coal, iron ore, chromite, and bauxite. This allows for a better understanding of key process variables affecting the efficiency of enrichment equipment and the potential for optimising technological operations. The work provides tools for predicting the behaviour of complex mineral enrichment systems, contributing to the development of more effective mineral processing methods.

The aim of this research was to develop an innovative methodology for creating extended datasets for modelling magnetic separation of iron ore, taking into account the specifics of Ukrainian deposits and the limitations of available information.

Materials and Methods

Justification for the choice of modelling method. An analysis of existing methods for modelling the magnetic separation process revealed the necessity of applying a comprehensive approach to address the task at hand. Considering the complexity and non-linearity of the magnetic separation process, as well as the specifics of the available data and tools, particularly USIM PAC – a commercial simulator for technological processes developed by CASPEO (Brochot *et al.*, 1995) – and the Python Spyder IDE development environment (n.d.), the decision was made to employ a hybrid modelling method (McCoy & Auret, 2019). The main arguments in favour of choosing the hybrid method are as follows:

1. Complexity of data processing. The proposed hybrid method combines physical modelling of magnetic separation with machine learning techniques. Initially, data is obtained from a model built on physical principles using the USIM PAC technological process simulator. This model is based on fundamental physical laws and empirical relationships that describe the magnetic separation process. Subsequently, this data is sequentially expanded and restructured using machine learning algorithms. In particular, neural networks are employed to uncover hidden patterns, clustering methods are used to group similar results, and regression algorithms are applied to predict process efficiency under various conditions. This combination of physical modelling and machine learning methods allows for effective processing of complex, non-linear relationships in magnetic separation data, significantly enhancing the capabilities of the initial physical model.

2. Adaptability to different conditions. In the study, data from the Valyavkinske deposit (Bogdanov, 1984) was used as an example for initial modelling. The deposit was chosen due to its typical characteristics, which well represent the general conditions of iron ore deposits in Ukraine. However, the developed approach aims to create a general model of the enrichment system that can be adapted to various mining and processing plants. The hybrid method provides the necessary flexibility for such adaptation, allowing the model to be tailored to the specific conditions of other deposits and mining and processing plants (MPP). A key aspect of this adaptability is the ability to replace the technological parameter data of MPPs and deposits. This allows for the modelling results to be aligned with the conditions of different MPPs. For example, by changing parameters such as ore characteristics, equipment configuration, or operating modes, the model can be adapted to the specifics of a particular plant. Such flexibility is especially useful when optimising processes at new deposits or modernising existing MPPs. The hybrid method, by combining physical modelling with machine learning techniques, allows for rapid retraining of the model on new data while maintaining a fundamental understanding of the physical enrichment processes.

3. Working with limited data. In conditions of limited access to real production data, the hybrid method allows for the effective use of artificially generated data while preserving the physical validity of the model through the use of USIM PAC. The effectiveness of this approach is supported by general principles of using simulators in modelling enrichment processes, as detailed in the work of A. Karpatne *et al.* (2017), where the authors emphasise the importance of integrating physical models and machine learning methods to enhance prediction accuracy in complex systems. However, further validation on real production data remains an important step for fully confirming the accuracy and reliability of the developed model.

4. Preparation for the development of a control system model. Based on a limited initial dataset obtained from the operation of USIM PAC, an expanded dataset is created. This expanded dataset is characterised by a significantly larger volume while preserving key relationships between fields that correspond to the mathematical dependencies of the USIM PAC simulation system. Such an approach potentially allows for the generation of a more diverse data sample, which can serve as a foundation for the further development of a predictive automated control system for the non-linear iron ore enrichment system. However, to confirm the effectiveness of this approach, thorough validation of the expanded dataset is necessary. As noted by T. Hastie *et al.* (2009) in their foundational work on statistical learning, it is important to conduct comprehensive statistical analysis to verify the preservation of key relationships and to perform testing on real data where possible. This ensures the reliability and practical applicability of the developed model.

5. Potential for further development. The hybrid method leaves room for the integration of additional modelling methods in the future, which may be beneficial for further research and improvement of the system.

Characteristics and structure of the enrichment system model. The enrichment system model is based on geological and mineralogical data from the Valyavkinske deposit of iron quartzites (Bogdanov, 1984; Kupin, 2008). This data includes ore characteristics such as iron content, mineral composition, textural-structural features, and physical properties, which are crucial for designing the enrichment process. Although the data is derived directly from the deposit, it has been adapted for modelling the first stage of magnetic enrichment, typical of most Ukrainian MPPs (Sokur *et al.*, 2022). This allows for the creation of a model that reflects the typical conditions for enriching iron quartzites in Ukraine. The overall structure of the studied part of the iron ore enrichment system is presented in Figure 1.

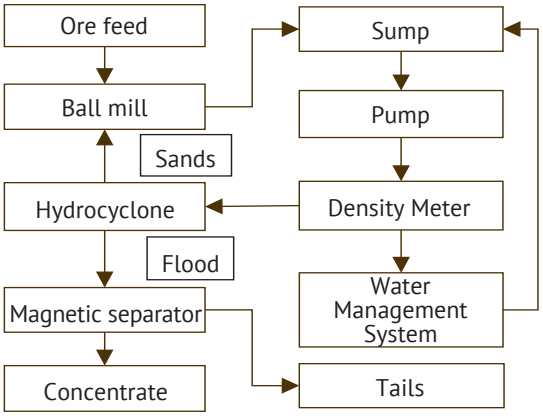

Figure 1. Technological scheme of iron ore enrichment with control of the solid density in the hydrocyclone
Source: developed by the author based on typical technological process schemes presented in M. Sokur *et al.* (2022)

The key input parameters of the model include the percentage of solids entering the hydrocyclone (25-35%), the flow rate of additional water (180-393 m^3/h), and the iron content in the incoming ore (36-38%). The output parameters encompass the iron content in the concentrate (52.5-55.5%) and tails (12.6-12.7%), as well as the mass flow rate of the concentrate (55-60 t/h) and tails (40-45 t/h). The selected characteristics of the deposit include an average rock density of 3.2-3.4 t/m^3, the grain size of magnetite inclusions of 0.074-0.044 mm, and the ratio of magnetic to non-magnetic minerals of 45-55.

The technological process, illustrated in Figure 1, consists of the following operations: the feed ore is supplied by a feeder to a ball mill for fine grinding. The resulting ore pulp is directed to the hydrocyclone for hydraulic classification by size. The overflow from the hydrocyclone is sent to a magnetic separator, where the material is separated based on magnetic properties into a magnetic product (concentrate) and a non-magnetic fraction (tails). The sands from the hydrocyclone are returned for regrinding in the ball mill, forming a closed grinding circuit. Control is achieved by adjusting the percentage of solids entering the hydrocyclone within the range of 25-35%, which affects the flow rate of additional water and ensures the quality of the concentrate (specifically, the iron content) in accordance with the target indicators established in the technological maps of the MPPs. This model provides a foundation for the development of automated control systems tailored to the specifics of Ukrainian iron ore deposits.

The use of the USIM PAC simulator. The technological process simulator USIM PAC from CASPEO was chosen to create the initial model of the iron ore beneficiation system. USIM PAC stands out among alternatives due to its greater number of equipment prototypes, the ability to use different models for circuit elements, and advanced result analysis. Its reliability is confirmed by widespread application in metallurgy and chemistry (Brochot *et al.*, 2002). For modelling the iron ore beneficiation process, the internal base Model 140 – "Feed Liberation" was utilised, which is an integral part of the commercial USIM PAC simulator (Fig. 2).

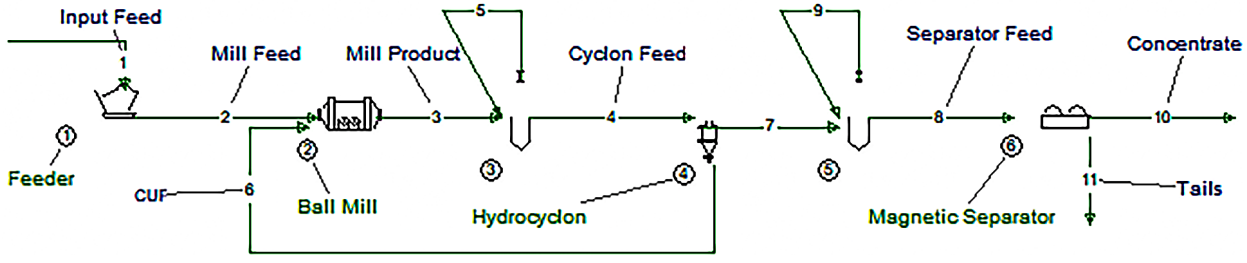

Figure 2. Iron ore enrichment system based on the Model 140 USIM PAC
Source: image of a typical educational Model 140 that is part of the USIM PAC

This model was chosen for its effectiveness in simulating the early stages of ore processing, making it particularly valuable for modelling the enrichment of iron ores from Ukrainian deposits. Model 140 is based on the method of R.L. Wiegel (1975) and the liberation model of A.M. Gaudin (1939), which allows for an

accurate description of the mineral liberation process during grinding. It operates with key parameters: the dilution factor with waste rock, the content of the valuable mineral, and the effective grain size of the mineral. An important advantage of Model 140 is that the composition by liberation classes does not depend on the size distribution, making it especially useful at the beginning of the technological scheme. This corresponds to the actual operating conditions of Ukrainian MPPs. When applying Model 140 to the Valyavkinske deposit, specific characteristics of the local ores were taken into account, particularly the average magnetite inclusions and the typical mineralogical composition.

Description of model parameters. The modelling of the iron ore enrichment process was carried out based on data obtained using USIM PAC. Key model parameters, including input and output variables, their ranges, and units of measurement, are presented in Table 1.

Table 1. Key parameters of the iron ore enrichment process model

Parameter	Description	Range of values	Units of measurement
	Input parameters		
solid_feed_percent	Percentage of solids at the hydrocyclone inlet	25-35	%
water_add_mass_flow*	Additional water flow rate	180-393	m^3/h
feed_fe_percent	Iron (Fe) content in the feed ore	36-38	%
	Output parameters		
concentrate_fe_percent	Iron (Fe) content in the concentrate	52.5-55.5	%
tailings_fe_percent	Iron (Fe) content in the tailings	12.6-12.7	%
concentrate_mass_flow	Mass flow rate of the concentrate	55-60	t/h
tailings_mass_flow	Mass flow rate of the tailings	40-45	t/h

Notes: * – in fact, at this stage, the flow rate of additional water, despite being an input parameter, should be calculated
Source: developed by the author based on the data from O. Bogdanov (1984)

Modelling the distribution of iron content in incoming ore. The initial assumption for modelling the distribution of iron content in incoming ore was based on a normal distribution. This assumption is supported by the research of J.C. Davis (2002), who demonstrated that the natural variability of geological processes and the effects of ore mixing during extraction and transportation contribute to the formation of a normal distribution of valuable component content. This approach is also reinforced by the central limit theorem, which is relevant for many geostatistical processes. Thus, adopting this assumption is justified and beneficial for modelling iron ore enrichment processes.

To improve the fit of the data to a normal distribution, various transformation methods were explored. Among them, mathematical transformations (power, logarithmic, exponential) were applied, as well as statistical transformations such as the Box-Cox method (Box & Cox, 1964) and Yeo-Johnson method (Yeo & Johnson, 2000). Additionally, data processing methods were utilised, including outlier removal and the calculation of moving averages, as well as more complex approaches such as kernel density estimation (Silverman, 1986), principal component analysis (PCA), and rank normalisation. The chosen transformation method was applied to create a dataset with iron content distribution that closely aligns with a normal distribution. This provided the necessary foundation for further modelling of iron ore enrichment processes.

Generation of solid percentage values at the hydrocyclone inlet. The solid percentage at the hydrocyclone inlet (solid_feed_percent) is a key control parameter in the developed model. A comprehensive methodological approach was employed for its analysis and generation. Initially, a statistical investigation of the distribution of solid_feed_percent values in the primary dataset was conducted. The Kolmogorov-Smirnov tests (Massey, 1951) and Shapiro-Wilk tests (Shapiro & Wilk, 1965) were used to verify the normality of the distribution. The coefficient of variation, skewness, and kurtosis were also calculated to characterise the shape of the distribution, using methods described by T. Hastie *et al*. (2009).

To fill in missing data, two methods were developed and compared: kernel density estimation (KDE) and random filling within quantile constraints (RFQL). The KDE method, described by B.W. Silverman (1986), uses kernel density estimation to model the distribution of existing data. This method fills in gaps with random values within defined quantile constraints, preserving the statistical structure of the data while filling in the gaps. After generating data using both methods, a comparative analysis of their statistical characteristics was conducted. Mean values, data dispersion, distribution shape, and the presence of outliers were assessed using methods described in T. Hastie *et al*. (2009). This analysis allowed for the identification of the most suitable method for filling in missing solid_feed_percent values, ensuring the accuracy and representativeness of the data for further modelling. The chosen method was applied to create an extended dataset that includes both original and generated solid_feed_percent values. This approach ensures the preservation of the statistical structure of the original

data while simultaneously expanding the dataset for more accurate modelling of the enrichment process.

Determining additional water flow values. In the model structure, the parameters of solid percentage in the hydrocyclone (solid_feed_percent) and additional water flow (water_add_mass_flow) have a close yet nonlinear relationship. To determine this relationship and fill in missing water_add_mass_flow values, the following methodology was applied. Initially, an analysis of the primary dataset was conducted to study the nature of the relationship between solid_feed_percent and water_add_mass_flow. It was established that this relationship is most accurately described by a second-degree polynomial dependence. A subset of records with incomplete data containing values for both parameters was extracted from the full dataset. Four methods were chosen for training and prediction: Gradient Boosting (Friedman, 2001), Random Forest (Breiman, 2001), Linear Regression (Hastie *et al.*, 2009), and Ridge Regression (Hoerl & Kennard, 1970). The selected methods provide a variety of approaches to data modelling, allowing for the capture of both complex nonlinear and linear dependencies. The models were trained on a sample of complete records. Quality metrics were calculated for each model, enabling a comparative analysis of the methods' effectiveness. This approach allows for the identification of the most accurate method for filling in missing water_add_mass_flow values and ensures data integrity for further analysis of the enrichment process.

Determining dependant parameters. To determine the iron content in the concentrate and tails, as well as the mass flow rates of the concentrate and tails for incomplete records in the dataset, the following methodology was applied. Initially, models were trained based on the complete dataset to fill in missing values in incomplete records. Six machine learning methods were used for this purpose: eXtreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016), Support Vector Machines (SVR) with a Laplace kernel (Cortes & Vapnik, 1995), Random Forest (Breiman, 2001), Multilayer Perceptron (MLP) (Rumelhart *et al.*, 1986), Ridge Regression (RR) (Hoerl & Kennard, 1970), and k-Nearest Neighbours Regression (kNN) (Altman, 1992). The choice of these methods is due to their ability to effectively work with multiple input/output (MIMO) models and address approximation tasks. Subsequently, the parameters of each model were optimised to enhance its performance. Based on the optimised parameters, final models were formed for further use in the enrichment process. This approach ensures the creation of an extended dataset with complete data for further analysis and modelling of the enrichment process.

Results and Discussion

Development of the functional diagram. To better understand the relationships between the model parameters and their roles in the iron ore beneficiation process, a functional diagram has been developed (Fig. 3). This diagram is based on the technological scheme of iron ore beneficiation with solid density control in the hydrocyclone (Fig. 1) and the iron ore beneficiation system based on Model 140 USIM PAC (Fig. 2). It visualises the main input and output variables, as well as their impact on various stages of the beneficiation process, integrating information from the previous diagrams into a more detailed functional model.

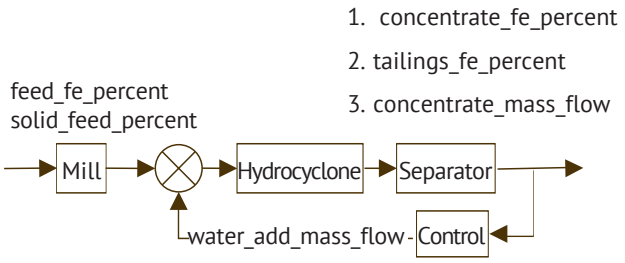

Figure 3. Functional diagram of the relationships between the parameters of the iron ore enrichment process model

Source: author's own development after processing the Model 140 USIM PAC

The functional diagram (Fig. 3) illustrates the key variables of the model and their interrelationships, which are crucial for understanding the iron ore beneficiation process. It demonstrates how the percentage of solids at the inlet of the hydrocyclone and the flow rate of additional water affect its operation, which in turn impacts the efficiency of magnetic separation. J. Svoboda (2004) notes that these factors are critical for achieving optimal results in the separation process, as they determine how effectively the useful components are separated.

The iron content in the incoming ore directly influences the quality of the obtained concentrate and tails, which is an important aspect for assessing the economic efficiency of beneficiation. This functional diagram not only summarises information from previous figures but also expands it by showing detailed interrelationships between parameters and their impact on each stage of the beneficiation process. As a result, it allows for a better understanding of how changes in one parameter can affect other aspects of the process. This knowledge is critically important for optimising the operation of the iron ore beneficiation system, as emphasised by T.J. Napier-Munn *et al.* (2014). Thus, the functional diagram serves as an important tool for analysing and improving technological processes in the mining industry.

The interrelationship between parameters is key to understanding the dynamics of the iron ore beneficiation system. The parameter feed_fe_percent reflects the percentage of iron content in the ore being fed into the system. When this value is constant, the system remains stable, but output parameters, such as concentrate quality and product yield, which constitute the

objective function, do not reach their optimal values. To optimise the process, it is necessary to adjust the parameter solid_feed_percent, which represents the percentage of solid material in the pulp. This parameter is regulated through water_add_mass_flow, i.e., the mass flow rate of water added to the system. Increasing or decreasing the water supply alters the pulp density, which directly affects the efficiency of the beneficiation process. Thus, the correct adjustment of solid_feed_percent allows the system to achieve an optimal state, maximising concentrate quality while maintaining a high product yield. In the modelling process, solid_feed_percent and feed_fe_percent are set within defined constraints, while other parameters are calculated based on the mathematical model of the process.

The choice of these parameters is driven by the primary objective of the work – generating a dataset that describes the nonlinear process of iron ore beneficiation for further use in creating a predictive control system (Hodouin, 2009). The selected set of input, output, and influencing parameters provides an adequate description of the process to achieve this goal. Determining the optimal size of the expanded dataset is a key stage in modelling iron ore beneficiation, ensuring a balance between data representativeness and computational efficiency. This is critically important for the accuracy of the model, avoiding overfitting, and effectively utilising resources.

As noted by B.A. Wills & J.A. Finch (2015), optimising the dataset size is an important aspect for achieving high model accuracy and preventing overfitting. This optimisation requires consideration of the specifics of the iron ore beneficiation process, particularly the nonlinear relationships between parameters and the variability of process conditions. Methods for determining the optimal size may include learning curve analysis, cross-validation, and assessing the statistical significance of sample size increases, as described in the work of G. James *et al.* (2021). The application of these methods allows for the determination of the optimal dataset size that provides sufficient data representativeness for accurate modelling of the nonlinear iron ore beneficiation process while maintaining computational efficiency.

Determining the optimal size is an iterative process that requires constant balancing between accuracy and efficiency. Evaluation criteria may include model quality metrics (e.g., RMSE, R^2) and computational costs. A typical dataset size for modelling beneficiation processes can range from several thousand to hundreds of thousands of samples, depending on the complexity of the process and accuracy requirements (Napier-Munn *et al.*, 2014). It is also important to consider specific challenges associated with iron ore beneficiation data, such as the uneven distribution of ore quality classes and the potential presence of outliers, which may affect the representativeness of the sample.

Analysis of the results of the initial modelling and characterisation of the generated dataset. As a result of the preliminary modelling using the commercial software USIM PAC (Brochot *et al.*, 1995), 915 data records were generated. The analysis of the statistical parameters of the primary dataset demonstrates the following features (Table 2).

Table 2. Statistical characteristics of the initial dataset

	solid_feed_ percent	water_add_ mass_flow	feed_fe_ percent	concentrate_ fe_percent	tailings_fe_ percent	concentrate_ mass_flow	tailings_ mass_flow
Key indicators of central tendency							
Mean	29.76	272.82	36.65	54.29	12.65	57.62	42.37
Med	29.53	267.84	36.63	54.31	12.65	57.60	42.42
Dispersion indicators							
Std dev	2.97	62.26	0.79	0.70	0.03	1.21	1.18
Min	25.01	180.44	35.30	52.67	12.60	54.88	39.87
Max	34.99	393.80	38.02	55.86	12.70	60.18	44.97
CV	0.1000	0.2282	0.0216	0.0128	0.0020	0.0210	0.0279
Distribution shape indicators							
Kurtosis	-1.24	-1.13	-1.24	-0.70	-1.18	-0.82	-0.86
Skewness	0.09	0.28	0.06	-0.01	0.05	0.01	-0.02
Normality tests							
Shapiro-Wilk	2.21E-17	2.72E-18	6.20E-17	2.12E-06	4.25E-15	3.49E-08	6.04E-09

Notes: the most significant indicators were taken for the fields
Source: author's own calculations when processing the data

The solid phase content in the hydrocyclone liquid (solid_feed_percent) is characterised by a mean value of 29.76% and a median of 29.53%, indicating a typical level of solid phase content and a relatively symmetrical distribution of the data. The standard deviation of 2.97 and the coefficient of variation of 0.1000 indicate moderate variability of the parameter. The range of values from 25.01 to 34.99% demonstrates

significant amplitude of fluctuations. The skewness coefficient of 0.09 indicates slight right-side skewness, while the kurtosis coefficient of -1.24 suggests a flatter distribution compared to normal. These characteristics indicate a stable, yet not static, process of solid phase feeding, with certain distribution peculiarities that should be considered in further analysis and modelling.

The water flow rate (water_add_mass_flow) has a mean value of 272.82 and a median of 267.84, indicating slight right-side skewness of the distribution. The high standard deviation of 62.26 and the coefficient of variation of 0.2282 indicate significant variability of this parameter. The wide range from 180.44 to 393.80 demonstrates substantial fluctuations in water flow, which may be related to different operating modes of the hydrocyclone or changes in the input raw material.

Iron content indicators (feed_fe_percent, concentrate_fe_percent, tailings_fe_percent) demonstrate high stability. Low coefficients of variation (0.0216, 0.0128, 0.0020 respectively) and narrow ranges of values indicate the stability of the enrichment process and the effectiveness of separating iron-containing components. The closeness of the mean values and medians for these indicators suggests the symmetry of their distributions, which is a sign of a stable technological process.

The mass flows of concentrate and tailings (concentrate_mass_flow, tailings_mass_flow) are characterised by low coefficients of variation (0.0210 and 0.0279 respectively), indicating the stability of the separation process. The proximity of the mean values and medians, as well as relatively narrow ranges of values, confirm the stability of mass flows, which is an important indicator of the hydrocyclone's operational efficiency. The analysis of the distribution shape shows that all variables have negative kurtosis coefficients (ranging from -0.70 to -1.24), indicating a platykurtic distribution. This means that the distributions have a flatter shape compared to a normal distribution, which may indicate greater uniformity of values in the central part of the distribution. The skewness coefficients are close to zero (ranging from -0.02 to 0.28), indicating relatively symmetrical distributions for all parameters.

The results of the Shapiro-Wilk normality tests (Shapiro & Wilk, 1965) show very low P-values for all variables. This indicates a statistically significant deviation from normal distribution for all studied parameters. Such results may be a consequence of the specifics of the technological process or the presence of certain constraints or controls over the parameters.

Overall, the analysis of statistical characteristics demonstrates a stable distribution of most studied indicators with moderate variability. Such results correspond to typical observations for technological processes, as noted by D.C. Montgomery (2021) in his work on statistical methods analysis in industry. Deviations from normal distribution and platykurticity are important features that need to be considered in further analysis and modelling of the data. These characteristics may influence the choice of statistical analysis and modelling methods, as well as the interpretation of the results of iron ore enrichment process studies. A visual analysis of the statistical parameters of the primary dataset, presented through histograms and distribution density curves, is shown in Figure 4.

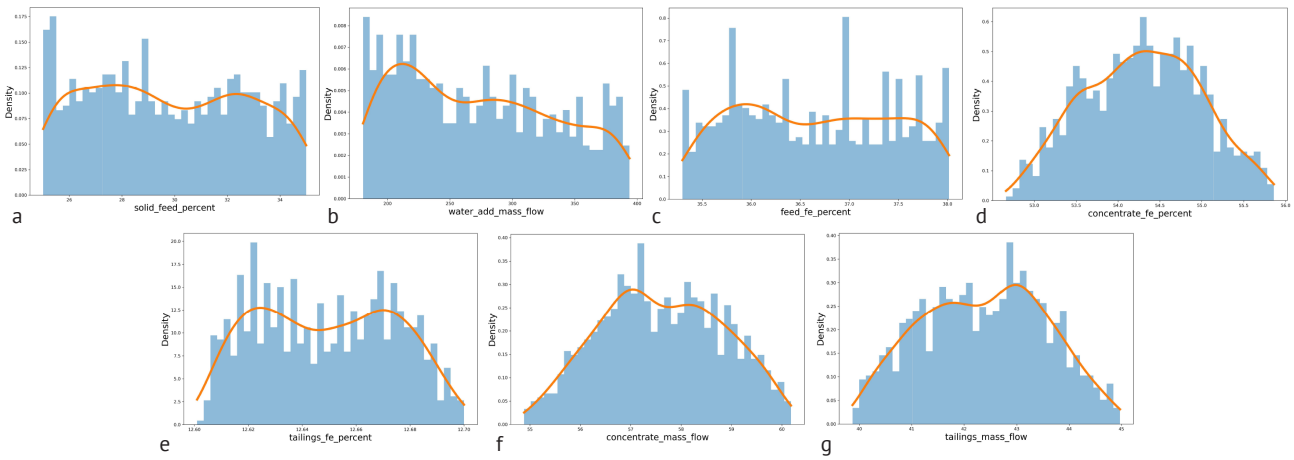

Figure 4. Histograms and density distribution curves of the main parameters of the primary dataset

Notes: a – solid phase content in the hydrocyclone liquid; b – water flow rate; c – iron content in the feed ore; d – iron content in the concentrate; e – iron content in the tailings; f – mass flow rate of the concentrate; g – mass flow rate of the tailings Bar chart – histogram, line graph – density curve of the corresponding parameter

Source: author's own development based on conducted calculations

Modelling the distribution of iron content in the incoming ore. In the process of modelling the distribution of iron content in the incoming ore, an initial assumption was made regarding the normal distribution of the data. This assumption was based on theoretical considerations and widely accepted practises in the field of

ore enrichment. However, as previously demonstrated (Table 2; Fig. 4), the analysis of the primary model data revealed significant deviations from the expected normal distribution. The primary cause of this deviation was identified as the manual entry of data for the relevant variable, which led to an uneven distribution. Consequently, the task arose to adjust the distribution of this field to normal, which is critically important for the accuracy of further modelling of the enrichment process.

To address this issue, a number of distribution correction methods were proposed and analysed. Among them were: logarithmic transformation, exponential transformation, the Box-Cox method (Box & Cox, 1964), the moving average (MA) method, Kernel Density Estimation (Rosenblatt, 1956), and Principal Component Analysis (PCA) (Pearson, 1901). The results of applying these methods are presented in Figure 5.

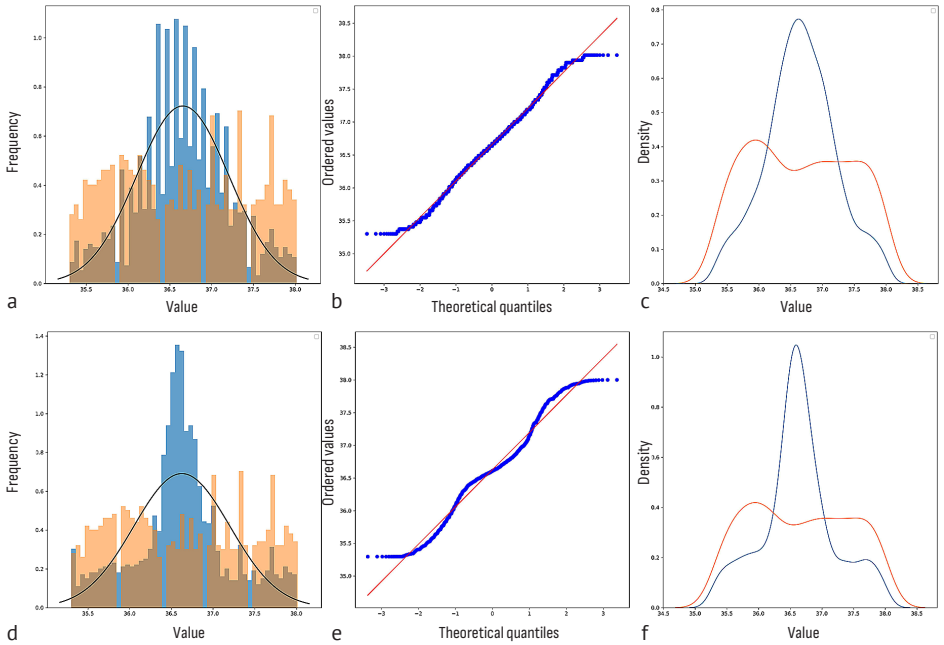

Figure 5. Results of transformation using different methods

Notes: a – Log; b – Exp; c – Box-Cox; d – MA; e – KDE; f – PCA. Distribution of diagrams in the group from left to right, top to bottom: 1. Histograms of the original (orange) and modified (blue) distributions; 2. Q-Q plot after modification; 3. Density curves of the original (orange) and modified (blue) distributions

Source: author's own development based on conducted calculations

Based on a comparative analysis of quality metrics (Table 3), the moving average method with automatic parameter optimisation was selected. This method provided the best balance between achieving normality of the distribution and preserving key characteristics of the original data. Specifically, the MA method showed optimal results across four of the five key criteria: multiplier, skewness, kurtosis, and preservation of the original data. Although the method did not achieve optimal results for P-value, automatic optimisation helped minimise undesirable effects of the transformation.

Table 3. Comparison of data transformation methods

Method	Multiplier	P-value	Skewness	Kurtosis	Original data percentage
Optimal	⩽10	0.05	0.5	0.5	⩾10%
Log	3	0.0950	0.0309	0.1721	25.03%
Exp	1	0.2364	-0.0065	0.2006	50%
Box-Cox	8	0.8050	-0.0009	-0.1157	11.11%
MA	2	0.1725	0.0265	0.1731	33.38%

Notes: the first line is the optimal indicators that needed to be achieved during the automatic search for parameters

Source: author's own calculations when processing the data

The application of the chosen method allowed for the creation of a new dataset, which includes 730 complete and 2,026 partially filled records, demonstrating an approximate Gaussian distribution of iron content (the filled data column feed_fe_percent with 2,756 values). This created a reliable foundation for further analysis and modelling of the enrichment process. It is important to note that correcting the distribution of iron content in the incoming ore is critical for the accuracy of the entire enrichment model. This enables more accurate forecasting of process outcomes and optimisation of control parameters, ultimately enhancing the efficiency of the entire iron ore enrichment process.

Generation of solid feed percentage values at the hydrocyclone inlet. The solid feed percentage at the hydrocyclone inlet (solid_feed_percent) is a key control parameter in automated iron ore beneficiation systems. B.A. Wills & J.A. Finch (2015) emphasise that this parameter significantly affects the efficiency of the beneficiation process and the optimisation of target indicators, such as concentrate quality and product yield. A statistical analysis of the primary dataset revealed that the distribution of solid_feed_percent has a flat structure. This feature creates favourable conditions for exploring various operating modes of the system (Fig. 4). Two methods were employed to fill in the missing data: Kernel Density Estimation (KDE) (Rosenblatt, 1956) and Random Filling with Quantile Limits (RFQL) (Hastie *et al.*, 2009). The results of the transformation of the solid_feed_percent distribution using these methods are presented in Figure 6, which visually demonstrates the differences between the KDE and RFQL methods: the KDE method provides a smoother distribution, while RFQL better preserves the structure of the original data. This visual comparison is complemented by a detailed analysis of statistical indicators presented in Table 4.

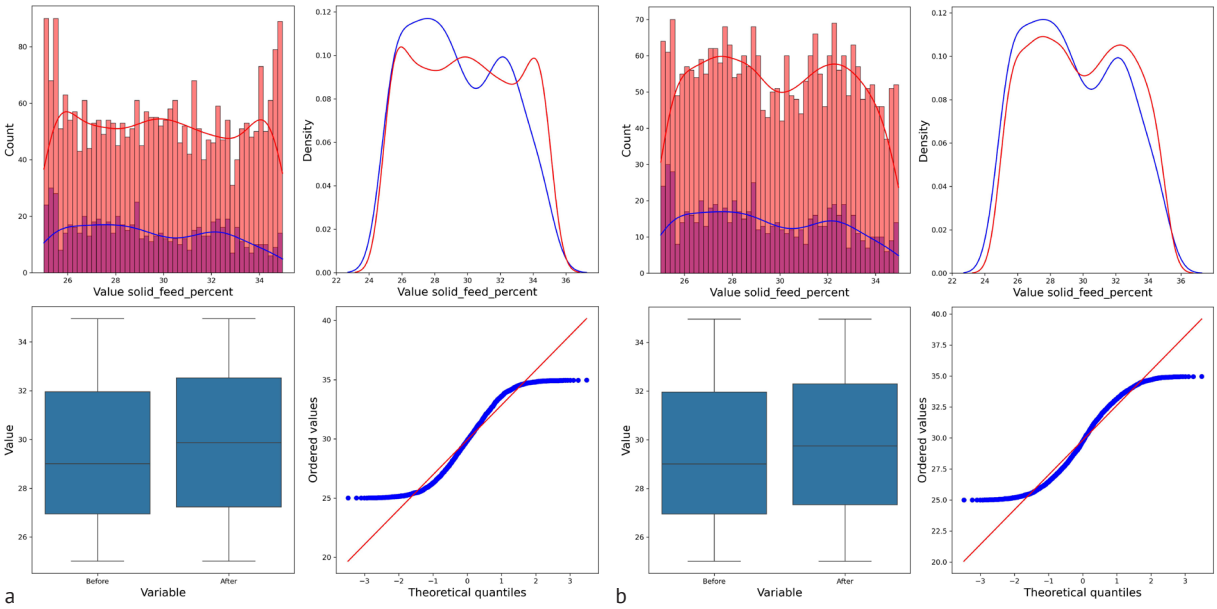

Figure 6. Transformation of the solid_feed_percent distribution due to gap filling

Notes: a – use of the KDE method; b – use of the RFQL method. Distribution of diagrams in the group from left to right, top to bottom: 1. Histograms of the original (red) and modified (orange) distributions; 2. Density curves of the original (blue) and modified (red) distributions; 3. Box plot of the value distribution; 4. QQ plot of the residuals

Source: author's own development based on conducted calculations

Table 4. Comparison of statistical indicators of missing data imputation methods

Metric	RFQL	KDE	Difference
Mean	29.8897	29.8905	0.0008
Std dev	2.8954	3.0256	0.1303
Med (50%)	29.8870	29.8740	-0.0130
25th percentile	27.3300	27.2000	-0.1300
75th percentile	32.3650	32.5003	0.1352
Skewness	0.0329	0.0472	0.0143
Kurtosis	-1.1982	-1.2442	-0.0460

Source: author's own calculations in data processing

The data analysis in Table 4 shows that both methods demonstrate similar results regarding means, dispersion, and skewness. However, the RFQL method proved to be more stable, with a lower tendency to create outliers. While KDE offers a broader coverage of possible values, RFQL better maintains the realistic characteristics of the process, which is critical for the accuracy of the model, as discussed in the work of T. Hastie *et al.* (2009). The choice of the RFQL method for further work is justified by its ability to preserve the statistical structure of the original data, which is particularly important for modelling complex technological processes. This method allows for the generation of data that not only fills in gaps but also retains the characteristics of the actual beneficiation process, as noted by A. Gelman & J. Hill (2006). It is important to note that changes in the solid percentage significantly impact the efficiency of the beneficiation process, and optimal control of this parameter can lead to improved concentrate quality and reduced losses of valuable components in the tails.

Determining the values of additional water consumption. The parameters of solid feed percentage (solid_feed_percent) and additional water flow rate (water_add_mass_flow) were found to be non-linearly interrelated. Solid_feed_percent serves as an indicator of the system's control mode, while water_add_mass_flow regulates this mode (Wills & Finch, 2015). An analysis of the initial dataset confirmed the non-linearity of the relationship between solid_feed_percent and water_add_mass_flow, which is most accurately described by a second-degree polynomial dependence. From 730 complete records containing values for both parameters, a model was developed to predict water consumption for 2,026 records lacking this value. Four machine learning methods were applied for modelling: Gradient Boosting (Friedman, 2001), Random Forest (Breiman, 2001), Linear Regression, and Ridge Regression (Hastie *et al.*, 2009), providing a variety of approaches to data modelling. The effectiveness of each method was evaluated using key metrics (Table 5) and visualised in Figure 7.

Table 5. Comparison of the effectiveness of machine learning methods for modelling

Method	MSE	MAE	R^2
Gradient Boosting	0.9691	0.7396	0.9997
Random Forest	0.7237	0.6129	0.9998
Linear Regression	1.6805	1.0276	0.9996
Ridge	4.1035	1.5272	0.9989

Source: author's own calculations in data processing

The analysis of results showed that the Random Forest method demonstrated the best performance with the lowest MSE and MAE values, as well as the highest R^2 (James *et al.*, 2021). This indicates its high accuracy and ability to effectively model complex non-linear dependencies between the parameters of the enrichment process. The obtained results have significant practical implications for optimising the enrichment process. They allow for more accurate forecasting and control of the solid percentage at the inlet of the hydrocyclone, which is crucial for enhancing the efficiency of the entire iron ore enrichment process. B.A. Wills & J.A. Finch (2015) emphasise in their work that precise control of this parameter can significantly impact the quality of the final product and reduce processing costs.

Definition of dependant parameters. To search for missing values in the fields of iron content in the concentrate, iron content in the tails, mass flow rate of the concentrate, and mass flow rate of the tails, an approach is employed that utilises a complete dataset to train machine learning algorithms that fill in the missing values. This enhances the integrity of the extended dataset, which is critically important for further analysis and modelling of enrichment processes. Six machine learning methods were selected for this purpose: eXtreme Gradient Boosting (Chen & Guestrin, 2016), Support Vector Machines with Laplace kernel (Cortes & Vapnik, 1995), Random Forest (Breiman, 2001), Multilayer Perceptron (Goodfellow *et al.*, 2016), Ridge Regression (Hoerl & Kennard, 1970), and k-Nearest Neighbours Regression (Altman, 1992). These methods are distinguished by their ability to effectively solve approximation tasks using MIMO models. Each of these methods was optimised to ensure maximum efficiency, allowing for the creation of models for accurate modelling of technological enrichment processes. The analysis results presented in Table 6 show that the Multilayer Perceptron demonstrates the best performance. This model has the lowest error values and the highest coefficient of determination R^2, indicating its high accuracy and effectiveness in generalising data, as detailed by I. Goodfellow *et al.* (2016) in their work on deep learning.

Table 6. Comparison of metrics for mathematical learning systems

Method	MSE	RMSE	MAE	R^2
XGB	0.0021	0.0460	0.0288	0.9977
SVR	0.0015	0.0383	0.0188	0.9989
RF	0.0032	0.0564	0.0336	0.9973
MLP	0.0013	0.0361	0.0208	0.9990
RR	0.0016	0.0396	0.0247	0.9986
kNN	0.0030	0.0540	0.0329	0.9972

Source: author's own development based on the calculations performed

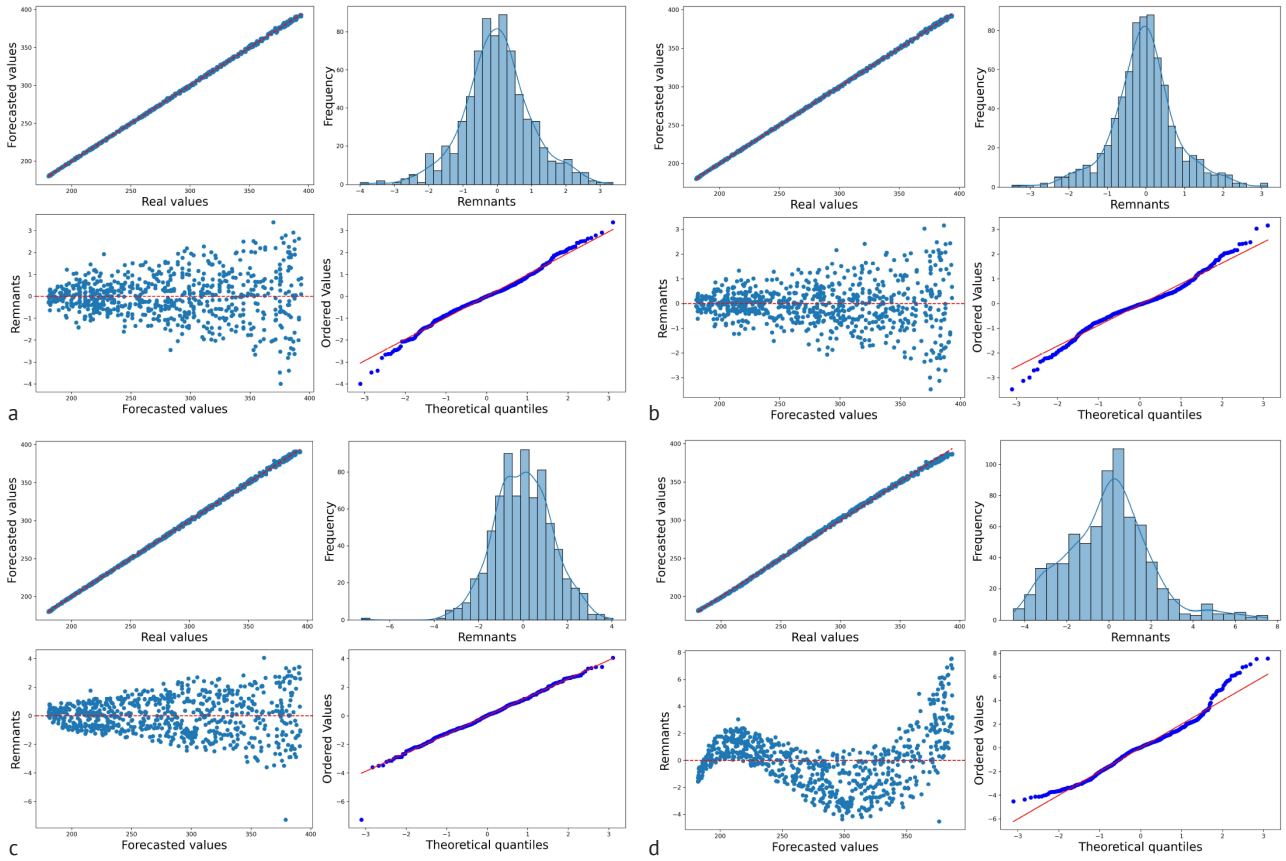

Figure 7. Quality metrics for forecasting using different forecasting methods

Notes: a – Gradient Boosting; b – Random Forest; c – Linear Regression; d – Ridge. Distribution of diagrams in the group from left to right, top to bottom: 1. Actual and forecasted values; 2. Distribution of residuals; 3. Residuals of forecasted vs. values; 4. QQ plot of residuals

Source: author's own development based on the calculations performed

Additionally, the SVR method with Laplace kernel also showed competitive results. With a high R^2 value and low error values, SVR is a reliable alternative for modelling, especially when neural networks are overly complex or resource-intensive. This method offers a balanced solution between model complexity and accuracy, making it very useful for real production conditions. Figure 8 illustrates the residuals when using machine learning methods. MLP demonstrates the most consistent results without significant deviations, while SVR also proved to be stable, confirming its reliability. MLP is the optimal choice for high-precision solutions to complex enrichment technological tasks, while SVR with Laplace kernel can be a practical option for situations where a combination of efficiency and simplicity is required.

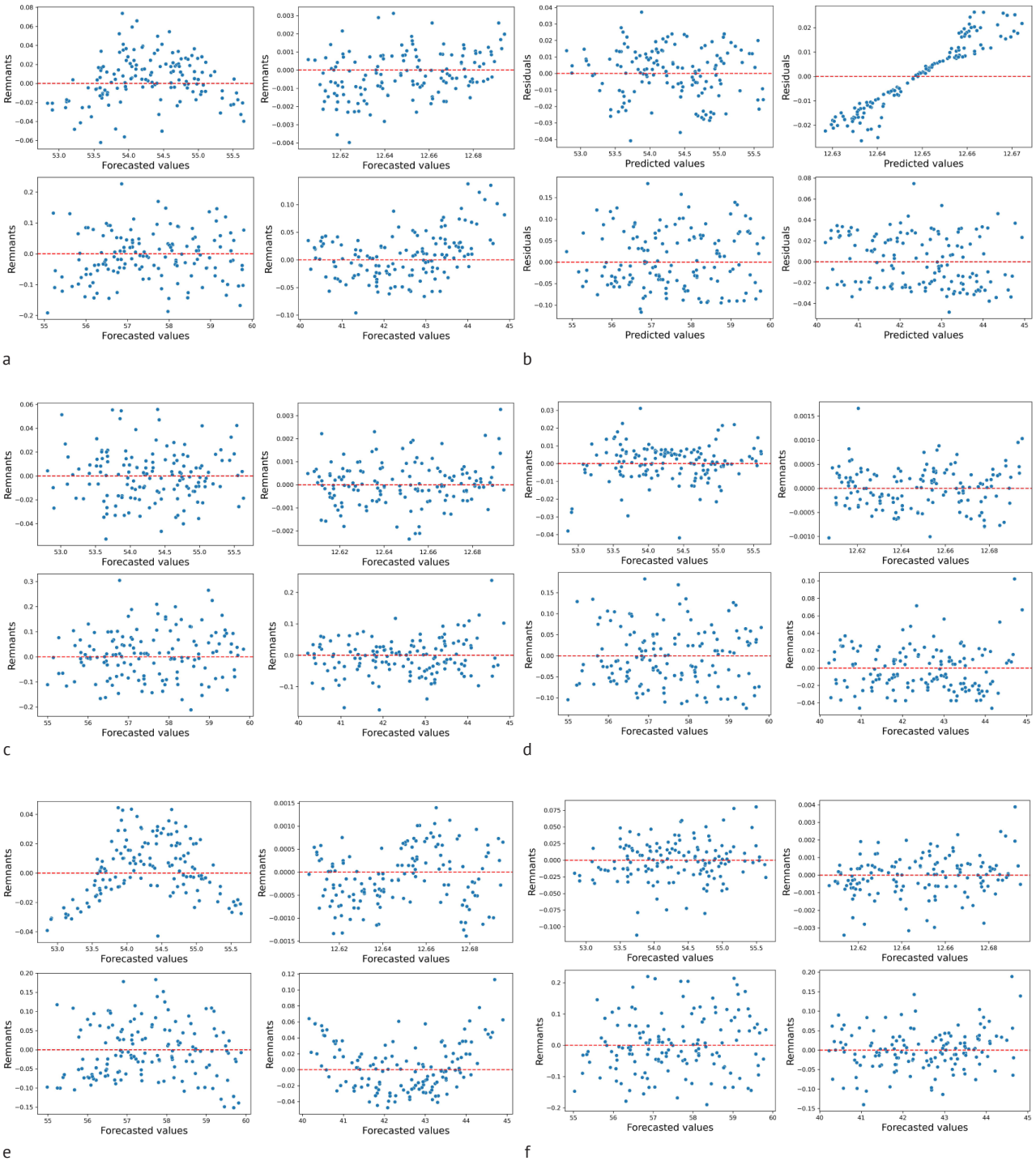

Figure 8. Visualisation of residuals when using different learning methods

Notes: a – XGBoost; b – Support Vector Regression; c – Random Forest; d – Perceptron Neural Network; e – Ridge Regression; f – k-Nearest Neighbors Regression. Fields in the group from left to right, top to bottom: concentrate_fe_percent, tailing_fe_percent, concentrate_mass_flow, tailing_mass_flow

Source: author's own development based on the calculations performed

Analysis of extended data. As a result of working with the data, an extended dataset was obtained with several modified distributions. The total number of records in the new dataset amounted to 2,756 records, which was determined by the initial requirements for automatic parameter selection when working with the feed_fe_percent field. Further expansion of the data array can be conducted through an iterative cycle according to the developed methodology. The statistical indicators of the new dataset are presented in Table 7.

Table 7. Statistical characteristics of the resulting dataset

	solid_feed_ percent	water_add_ mass_flow	feed_fe_ percent	concentrate_ fe_percent	tailings_fe_ percent	concentrate_ mass_flow	tailings_ mass_flow
Key indicators of central tendency							
Mean	29.86	270.02	36.65	54.28	12.65	57.65	42.34
Med	29.82	261.65	36.66	54.28	12.65	57.7	42.32
Dispersion indicators							
Std dev	2.86	59.66	0.54	0.55	0.02	0.91	0.87
Min	25.01	180.58	35.3	52.66	12.6	54.88	39.98
Max	34.96	393.8	38.02	55.86	12.7	60.09	44.96
CV	0.0956	0.221	0.0148	0.0101	0.0014	0.0157	0.0206
Distribution shape indicators							
Kurtosis	-1.18	-1.03	0.01	-0.27	-0.04	-0.16	-0.1
Skewness	0.04	0.35	0.04	0.05	0.06	-0.14	0.11
Normality tests							
Shapiro-Wilk	7.51E-28	1.82E-30	5.55E-10	7.06E-03	4.82E-07	1.64E-04	1.06E-03

Source: author's own development based on the calculations performed

The main indicators of central tendency, such as the mean and median, remained virtually unchanged for most indicators, indicating the preservation of the overall data structure. However, slight changes were observed in the indicators of water_add_mass_flow and concentrate_mass_flow, which may be related to the modification of the distribution of these fields. The analysis of dispersion indicators revealed that the standard deviation decreased for most indicators, indicating a reduction in data spread. The coefficients of variation also decreased, suggesting an increase in data homogeneity.

The study of distribution shape indicators demonstrated that the coefficients of excess and skewness underwent slight changes, indicating the preservation of the overall shape of the distribution. However, for some indicators, such as feed_fe_percent, concentrate_fe_percent, and tailings_fe_percent, a decrease in excess was observed, which may indicate a convergence towards a normal distribution. The results of normality tests, particularly the Shapiro-Wilk test, indicate that the distribution of most indicators remains non-normal. However, for some indicators (concentrate_fe_percent, tailings_fe_percent), a slight approach to normality is observed.

In general, it can be concluded that the modification of the distribution of certain fields led to minor changes in the statistical characteristics of the dataset. The main indicators of central tendency remained virtually unchanged, while the dispersion and shape indicators experienced slight improvements. This suggests that the overall data structure has been preserved, but their homogeneity and approach to normal distribution have somewhat increased. The visual distribution of the resulting dataset, presented through histograms and density distribution curves, is shown in Figure 9.

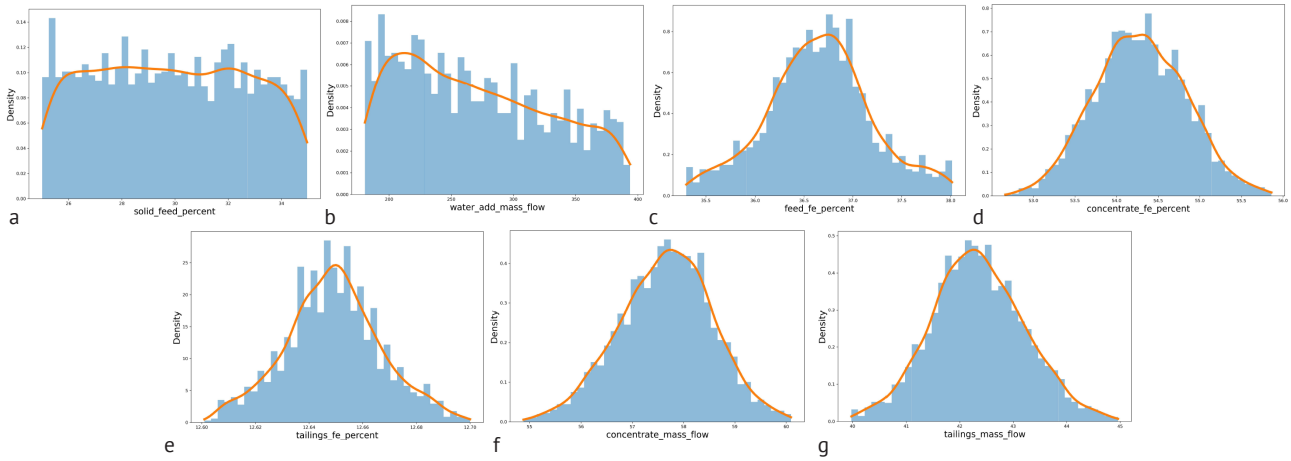

Figure 9. Histograms and density curves of the data distribution of the main parameters of the primary dataset

Notes: a – solid phase content in the hydrocyclone liquid; b – water flow rate; c – iron content in the feed ore; d – iron content in the concentrate; e – iron content in the tailings; f – mass flow rate of the concentrate; g – mass flow rate of the tailings. Bar chart – histogram, line graph – density curve of the corresponding parameter

Source: author's own development based on the calculations performed

As a result of the conducted analysis, an extended dataset was formed while preserving the overall structure of the original data. The modification of the distributions of certain parameters led to minor changes in the indicators of central tendency and an improvement in the homogeneity of the data. The presented statistical characteristics and visualisations confirm the increased proximity of the distributions to normality. The developed methodology for creating extended datasets for modelling magnetic separation of iron ore demonstrates significant potential for enhancing enrichment processes in the mining industry. The hybrid method used in the study has the potential to integrate aspects such as modularity, which, as defined by S. Shalev-Shwartz & S. Ben-David (2014), allows for the updating of components without the need to rebuild the system, ensuring adaptation to new requirements.

The use of the USIM PAC simulator for generating the primary dataset aligns with the approach described in the work of T.J. Napier-Munn *et al*. (2014), who emphasise the importance of applying specialised simulators for modelling enrichment processes. However, unlike their study, this work identified deviations from the expected Gaussian distribution, highlighting the necessity of validating theoretical models against real data, as noted by B.A. Wills & J.A. Finch (2015). The application of the moving average method with automatic parameter optimisation for data distribution correction is an innovative approach that has not been widely covered in previous research. This method demonstrated better results compared to traditional data transformation methods, such as logarithmic and exponential transformations, as described in the work of T. Hastie *et al*. (2009), which discusses various data processing techniques to improve their quality.

The use of the Random Forest with Quantile Limits method for filling in data gaps shows similarities to the approach proposed by L. Breiman (2001), but with additional constraints to ensure the physical validity of the data. This enhancement allows for better preservation of the characteristics of the actual enrichment process, which is a critical aspect emphasised by A. Gelman & J. Hill (2006). A comparison of different machine learning methods for modelling the relationships between parameters of the enrichment process revealed the superiority of the multilayer perceptron over other methods. This aligns with the findings of A. Karpatne *et al*. (2017), whose authors also noted the effectiveness of neural networks for modelling complex nonlinear processes in the mining industry. However, unlike their work, this study also found high effectiveness in the SVR method with a Laplace kernel, which may serve as a useful alternative in conditions of limited computational resources.

The developed methodology for creating extended datasets corresponds to the current trends of Industry 4.0, as mentioned in the research by H. Lasi *et al*. (2014). It provides modularity in the approach, allowing for the integration of new methods and data sources (Khaleghi *et al.*, 2013). It is important to note that this research focuses on the specifics of Ukrainian iron ore deposits, particularly the Kryvyi Rih basin, which distinguishes it from many international studies. This allows for consideration of local geological conditions and technological features, which are critical for the practical application of the results. Overall, the obtained results lay the foundation for further development of automated control systems for enrichment processes, aligning with the research directions outlined by P. Kadlec *et al*. (2009) and I.E. Grossmann & G. Guillén-Gosálbez (2010), who detail the importance of automation in managing technological processes and its impact on production efficiency. In particular, a promising direction is the integration of the developed methodology with decision-making systems and energy consumption optimisation.

Compared to existing studies, the developed methodology offers a comprehensive approach that combines physical modelling, statistical methods, and machine learning. This allows for overcoming the limitations associated with the lack of real production data while maintaining the physical validity of the model. Such an approach opens new opportunities for optimising iron ore enrichment processes and enhancing production efficiency in the context of Ukrainian mining and enrichment plants.

Conclusions

As a result of the research, a comprehensive methodology for creating extended datasets for modelling the magnetic separation process of iron ore has been developed, taking into account the specifics of Ukrainian deposits and the limitations of available information. Key achievements include: the creation and validation of an extended dataset based on the technological simulation USIM PAC; the development of a method for correcting data distribution with automatic parameter optimisation; and a comparative analysis of machine learning methods, where the multilayer perceptron demonstrated the highest prediction accuracy. The scientific novelty of the research lies in the development of an innovative methodology that combines technological simulation, statistical data correction methods, and modern machine learning algorithms for modelling the processes of magnetic separation of iron ore. This approach allows overcoming the limitations associated with the lack of real production data while maintaining the physical validity of the models.

The work provides opportunities to enhance production efficiency and product quality at Ukrainian mining and beneficiation plants. The developed methodology creates conditions for more precise tuning of technological processes, which is particularly important in the early stages of design and in conditions of limited access to technological data. Furthermore,

this methodology represents a significant step towards improving the efficiency and competitiveness of Ukrainian MPPs. This innovative approach opens new possibilities for optimising production and enhancing product quality in the field of iron ore beneficiation, contributing to the overall progress of the industry and strengthening Ukraine's position in the global iron ore raw materials market.

Nevertheless, an important direction for further research is the integration of the developed models into comprehensive automated control systems for technological processes, which will promote an increase in the level of automation and optimisation of management in enterprises. The application of deep learning methods to improve the accuracy of predicting beneficiation process parameters will enable the creation of more precise and reliable models for decision-making, while the development of adaptive control algorithms and optimisation of energy consumption will contribute to cost reduction and enhanced environmental safety in production.

Acknowledgements

The author expresses sincere gratitude to CASPEO for the opportunity to use the USIM PAC simulator in conducting this research. Special thanks to Gwenaëlle Larousse for her support and assistance in working with the software. This technical help was invaluable for the successful completion of the study and for obtaining reliable results in modelling the magnetic separation process of iron ore.

Conflict of Interest

None.

References

[1] Altman, N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175-185.
[2] Bogdanov, O. (Ed.). (1984). *Ore beneficiation handbook: Basic processes*. Moscow: Nedra.
[3] Box, G.E.P., & Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), 211-252.
[4] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. doi: 10.1023/A:1010933404324.
[5] Brochot, S., Durance, M.V., Fourniguet, G., Guillaneau, J.C., & Villeneuve, J. (1995). Modelling the minerals diversity: A challenge for ore processing simulation. In *EUROSIM 1995 congress on modelling and simulation*. Vienna: Federation of European Simulation Societies.
[6] Brochot, S., Villeneuve, J., Guillaneau, J.C., Durance, M.V., & Bourgeois, F. (2002). USIM PAC 3: Design and optimisation of mineral processing plants from crushing to refining. In *Mineral processing plant design, practice, and control* (pp. 479-494). Englewood: Society for Mining, Metallurgy, and Exploration.
[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785-794). New York: Association for Computing Machinery. doi: 10.1145/2939672.2939785.
[8] Chowdhury, R., Billah, M.M., Roy, S., Banik, S.C., & Barua, A. (2024). Analyzing the effect of the density of medium on efficiency of hydrocyclone separator in sorting of PVC and PET using CFD. *Results in Materials*, 21, article number 100497. doi: 10.1016/j.rinma.2023.100497.
[9] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. doi: 10.1007/BF00994018.
[10] Davis, J.C. (2002). *Statistics and data analysis in geology*. Hoboken: Wiley.
[11] Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232.
[12] Gaudin, A.M. (1939). *Principles of mineral dressing*. New York: McGraw-Hill.
[13] Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
[14] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge: MIT Press.
[15] Grossmann, I.E., & Guillén-Gosálbez, G. (2010). Scope for the application of mathematical programming techniques in the synthesis and planning of sustainable processes. *Computers & Chemical Engineering*, 34(9), 1365-1376. doi: 10.1016/j.compchemeng.2009.11.012.
[16] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer. doi: 10.1007/978-0-387-84858-7.
[17] Hodouin, D. (2009). Automatic control in mineral processing plants: An overview. *IFAC Proceedings Volumes*, 42(23), 1-12. doi: 10.3182/20091014-3-CL-4011.00003.
[18] Hoerl, A.E., & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.
[19] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. New York: Springer. doi: 10.1007/978-1-0716-1418-1.

[20] Kadlec, P., Gabrys, B., & Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 33(4), 795-814. doi: 10.1016/j.compchemeng.2008.12.012.
[21] Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10), 2318-2331. doi: 10.1109/TKDE.2017.2720168.
[22] Khaleghi, B., Khamis, A., Karray, F.O., & Razavi, S.N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1), 28-44. doi: 10.1016/j.inffus.2011.08.001.
[23] Kinaci, M.E., Lichtenegger, T., & Schneiderbauer, S. (2020). A CFD-DEM model for the simulation of direct reduction of iron-ore in fluidised beds. *Chemical Engineering Science*, 227, article number 115858. doi: 10.1016/j.ces.2020.115858.
[24] Kupin, A. (2008). *Intelligent identification and control in beneficiation processes*. Kyiv: Korniychuk.
[25] Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business & Information Systems Engineering*, 6(4), 239-242. doi: 10.1007/s12599-014-0334-4.
[26] Li, Y., Bao, J., Chen, T., Yu, A., & Yang, R. (2022). Prediction of ball milling performance by a convolutional neural network model and transfer learning. *Powder Technology*, 403, article number 117409. doi: 10.1016/j.powtec.2022.117409.
[27] Liu, J., Xue, Z., Dong, Z., Yang, X., Fu, Y., Man, X., & Lu, D. (2021). Multiphysics modelling simulation and optimisation of aerodynamic drum magnetic separator. *Minerals*, 11(7), article number 680. doi: 10.3390/min11070680.
[28] Massey, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253), 68-78. doi: 10.1080/01621459.1951.10500769.
[29] McCoy, J.T., & Auret, L. (2019). Machine learning applications in minerals processing: A review. *Minerals Engineering*, 132, 95-109. doi: 10.1016/j.mineng.2018.12.004.
[30] Montgomery, D.C. (2021). *Introduction to statistical quality control*. Hoboken: John Wiley & Sons.
[31] Morkun, V., Morkun, N., Tron, V., & Sulyma, T. (2020). Synthesizing models of nonlinear dynamic objects in concentration on the basis of Volterra-Laguerre structures. *Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu*, 2, 30-36. doi: 10.33271/nvngu/2020-2/030.
[32] Napier-Munn, T.J., Morrell, S., Morrison, R.D., & Kojovic, T. (2014). *Mineral comminution circuits: Their operation and optimisation*. Indooroopilly, Qld Australia: Julius Kruttschnitt Mineral Research Centre, University of Queensland.
[33] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559-572. doi: 10.1080/14786440109462720.
[34] Porkuian, O., Morkun, V., Morkun, N., & Serdyuk, O. (2019). Predictive control of the iron ore beneficiation process based on the Hammerstein hybrid model. *Acta Mechanica Et Automatica*, 13(4), 262-270. doi: 10.2478/ama-2019-0036.
[35] Python Spyder IDE. (n.d.). Retrieved from https://www.spyder-ide.org/.
[36] Rajendran, S., & Murty, C.V.G.K. (Eds.). (2023). *Mineral processing: Beneficiation operations and process optimisation through modelling*. Amsterdam: Elsevier. doi: 10.1016/C2019-0-03555-8.
[37] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3), 832-837. doi: 10.1214/aoms/1177728190.
[38] Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature*, 323, 533-536. doi: 10.1038/323533a0.
[39] Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge: Cambridge University Press.
[40] Shapiro, S.S., & Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4), 591-611. doi: 10.2307/2333709.
[41] Shenoy, V.V., Prasad, G.R., & Kumar, A.S. (2024). Effects of external magnetic field on flow separation, control and reattachment. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 46(7), article number 404. doi: 10.1007/s40430-024-04968-x.
[42] Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.
[43] Sokur, M., Argat, R., Biletskyi, V., & Ravinska, V. (2022). Selection of rational principle scheme of iron quartz enrichment. *Transactions of Kremenchuk Mykhailo Ostrohradskyi National University*, 1(132), 149-156. doi: 10.32782/1995-0519.2022.1.20.
[44] Svoboda, J. (2004). *Magnetic techniques for the treatment of materials*. Dordrecht: Springer. doi: 10.1007/1-4020-2107-0.
[45] Wiegel, R.L. (1975). Liberation in magnetite iron formations. *Transactions of the Society of Mining Engineers of AIME*, 258(3), 247-256.

[46] Wills, B.A., & Finch, J.A. (2015). *Wills' mineral processing technology: An introduction to the practical aspects of ore treatment and mineral recovery*. Oxford: Butterworth-Heinemann. doi: 10.1016/C2010-0-65478-2.
[47] Yang, N., Zhang, Z., Yang, J., & Hong, Z. (2022). Applications of data augmentation in mineral prospectivity prediction based on convolutional neural networks. *Computers & Geosciences*, 161, article number 105075. doi: 10.1016/j.cageo.2022.105075.
[48] Yeo, I.-K., & Johnson, R.A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4), 954-959.

Комплексна методологія створення розширених датасетів для моделювання процесу магнітної сепарації залізної руди

Олександр Воловецький

Аспірант
Криворізький національний університет
50027, вул. Віталія Матусевича, 11, м. Кривий Ріг, Україна
https://orcid.org/0009-0003-1703-387X

Анотація. Дослідження пропонує інноваційний підхід до створення розширених наборів даних для моделювання магнітної сепарації залізної руди, що є важливим для підвищення ефективності та автоматизації процесів збагачення в гірничодобувній промисловості. Мета дослідження полягала в розробці методології створення розширених наборів даних для моделювання магнітної сепарації залізної руди, яка враховує специфіку українських родовищ та дозволяє генерувати репрезентативні дані в умовах обмеженості реальних виробничих даних шляхом інтеграції фізичного моделювання з методами машинного навчання. Методи дослідження: моделювання з використанням математичного навчання, симуляція на основі фізичних процесів, статистичний аналіз. У дослідженні розглянуто використання симулятора USIM PAC для моделювання системи збагачення залізної руди та адаптацію даних для магнітного збагачення, що забезпечує точність моделювання технологічних процесів збагачення. Застосуванням симулятору отримано набір даних фізичного моделювання частини процесу збагачення на основі даних Валявкінського родовища. Проаналізовано первинне моделювання набору даних, включаючи статистичні характеристики, форму розподілу та тести на нормальність для виявлення полів, що потребують корекції. На основі результатів аналізу визначено конкретні вимоги до розподілу даних у новому датасеті, який має бути сформований для подальшого використання. Відповідно до цих вимог реалізовано декілька математичних моделей, що відтворюють задані критерії та параметри. Для кожного поля даних ретельно підібрано найкращу модель та виконано корекцію датасету за її даними, щоб максимально наблизити розподіл до бажаного. Для отриманих скоригованих даних проведена всебічна валідація результатів з акцентом на збереженні фізичної достовірності даних та їх відповідності реальним процесам збагачення. Проведено детальний аналіз відкоригованих даних, а також статистичні характеристики результуючого датасету, в результаті чого підтверджена ефективність розробленої комплексної методології моделювання та адаптації даних для магнітного збагачення залізної руди. Методологія має практичну цінність завдяки інноваційному підходу до створення розширених наборів даних для моделювання магнітної сепарації залізної руди, що підвищує ефективність і автоматизацію процесів збагачення, враховуючи специфіку родовищ та генеруючи репрезентативні дані в умовах обмеженості реальних даних

Ключові слова: нелінійне моделювання збагачення; керування сепарацією; машинне навчання в збагаченні; автоматизація збагачувальних процесів; симуляція технологічних параметрів