Physicochem. Probl. Miner. Process., 60(6), 2024, 196385

http://www.journalssystem.com/ppmp

Physicochemical Problems of Mineral Processing

ISSN 1643-1049

Evaluation of parametric effects of coal flotation based on boosting modeling method

Yuhua Ji [1], Liang Wang [2], Guangdong Ren [2], Tongjin Xu [2], Xingwang Li [2], Xiaoxia Chen [1], Xianging Bu [3], Jie Sha [3]

[1] Zaozhuang Mining Group Co., Ltd, Zaozhuang 277000, China

[2] Chaili Coal Mine Co., Ltd, Zaozhuang 277000, China

[3] Key Laboratory of Coal Processing and Efficient Utilization (Ministry of Education), School of Chemical Engineering and Technology, China University of Mining and Technology, Xuzhou 221116, China

Corresponding author: xiangning.bu@foxmail.com (Xianging Bu), shajie@cumt.edu.cn (Jie Sha)

Abstract: Accurate assessment of the effects of parameters on the flotation process is important for understanding the complex flotation mechanisms. To address the problem of unsatisfactory prediction of large sample flotation data (641 sets) by traditional machine learning algorithms, four advanced algorithms (GBDT, CatBoost, LightBGM and XGBoost) are used in this paper to investigate the effects of feed properties and flotation conditions on the effectiveness of coal flotation. It was found that the data at flotation recoveries below <40% were difficult to predict effectively by machine learning algorithms due to abnormal flotation results caused by lower flotation reagent dosages. An importance analysis of flotation parameters and prediction of flotation results were carried out based on the reordered data. The results showed that the fraction and ash content of -74 um in the feed are the main factors affecting concentrate yield and ash content. The XGBoost model also achieved the best prediction results compared to other models, and the prediction coefficient of determination R2 reached 0.877 and 0.971 for concentrate yield and ash content, respectively. The results are expected to provide a reference for the intelligent control of coal beneficiation plant by machine learning technology in the future.

Keywords: machine learning, coal flotation, prediction, XGBoost model, flotation parameters

1. Introduction

The overall sophistication of flotation intelligence within coal preparation plants remains inadequately low, representing a significant bottleneck to the advancement of intelligent production and management practices in this sector (Flores et al., 2024). A transformative approach to the flotation process is urgently needed to improve mineral processing efficiency, reduce production costs, and ensure operational safety. In this context, deep learning technology is emerging as a key avenue for facilitating intelligent coal slurry flotation in the foreseeable future (Abkhoshk et al., 2010; Meng et al., 2022). The application of deep learning enables real-time monitoring, intelligent control, and optimal scheduling of the flotation process, which is expected to significantly improve both flotation efficiency and environmental performance. An intelligent coal slurry flotation system can achieve precise control and optimization by integrating state-of-the-art control systems, sophisticated sensors, and advanced data processing technologies (Zhao et al., 2022). This system requires real-time and accurate monitoring of numerous parameters, including chemical dosing, slurry concentration, flow rate, liquid level height and product ash content. Building on this foundation, advanced machine learning techniques can be used to predict sorting results and adjust parameters accordingly, ultimately maximizing flotation yield and improving economics (Ali et al., 2018; Chelgani et al., 2024; Meng et al., 2022).

The parameters affecting the flotation process can be categorized into four primary dimensions: liquid variables, reagent variables, gas variables and solids variables (Sun et al., 2023; Vinnett et al., 2023). Traditionally, single-factor experimental designs have been used to evaluate the effect of individual parameters on the flotation process. However, a number of advanced experimental methods

DOI: 10.37190/ppmp/196385

have been used to optimize flotation processes and investigate the interactions between different influencing factors, including the Taguchi method (du Plessis and de Villiers, 2007; Lubisi et al., 2018; Sachinraj et al., 2022) and response surface methodology (Arancibia-Bravo et al., 2022; Bu et al., 2016; Wang et al., 2016; Wang et al., 2021). Multiple nonlinear regression is often used to predict flotation responses, both in the Taguchi method and in response surface methodology. However, multivariable linear regression typically provides modest predictive performance, with an R^2 value of approximately 0.8. To improve prediction accuracy, various machine learning techniques have been incorporated into flotation process prediction. Gomez-Flores et al. (2022) used multivariate linear regression, k-nearest neighbours, decision trees and random forests to model flotation grade and recovery based on physicochemical and operational parameters and found that random forests showed superior predictive performance for flotation concentrate grade and recovery. Guner et al. (2024) found that genetic programming (GP) with novel data demonstrated greater accuracy in predicting grade, while random forests excelled in predicting recovery. Furthermore, Ali et al. (2018) conducted a comparative analysis of the predictive behaviour in fine high-ash coal flotation using random forests (RF), artificial neural networks (ANN), adaptive neuro-fuzzy inference systems (ANFIS), Mamdani fuzzy logic (MFL) and a hybrid neural-fuzzy inference system (HyFIS), and concluded that the MFL model provided the most favourable performance. A number of studies have successfully implemented intelligent soft computing (IS) methods, including various types of artificial neural networks (ANNs) and the adaptive neuro-fuzzy inference system (ANFIS), to model mineral flotation responses dependent on process and sample conditions. Although these models are recognised for their reliability in predicting coal responses, they often fall short in assessing the correlations between flotation results and operating conditions (Bu et al., 2021). In this context, random forests (RF), as a sophisticated machine learning tool, not only facilitates the ranking of input variables based on their importance in influencing outputs, but also competently addresses linear and non-linear challenges (Chehreh Chelgani et al., 2016; Chelgani and Matin, 2018). In the field of mineral flotation, random forests have been applied to evaluate the effects of different parameters on coarse particle flotation recovery (Nazari et al., 2019), model coal flotation responses under different operating conditions (Bu et al., 2021), predict copper flotation recovery (Flores et al., 2024), and predict froth flotation responses influenced by different conditioning parameters (Shahbazi et al., 2017).

Gradient boosting trees are an advanced ensemble learning technique within the field of boosting methods (Zhang et al., 2019). Adaboost, for example, uses the error rates of weak learners from previous iterations to strategically adjust the weights of the training dataset, facilitating a sequential refinement process. In contrast, Gradient Boosting Decision Trees (GBDT) use a forward stepwise approach where the weak learners are constrained to the CART (Classification and Regression Trees) regression tree model (Zhang and Jánošík, 2024). The goal of each iteration is to identify a CART weak learner that minimises the loss function. As the algorithm has evolved, GBDT has given rise to several prominent implementations, most notably XGBoost, CatBoost and LightGBM. The proliferation of machine learning methods and the growing volume of data have spurred continuous advances in gradient boosting algorithms (Ma et al., 2018). For example, XGBoost (eXtreme Gradient Boosting) is a highly efficient boosting algorithm that uses second-order derivative information to optimise the loss function. It also incorporates features such as feature selection and tree pruning to reduce the risk of overfittingnm (Carmona et al., 2019). LightGBM, developed by Microsoft on top of the gradient boosting framework, uses a histogram-based decision tree algorithm, which allows it to effectively handle large datasets and high-dimensional feature spaces (Sun et al., 2020). Meanwhile, the CatBoost algorithm, innovated by Yandex, has features such as automatic missing value handling and support for GPU acceleration, which significantly improves its performance when dealing with datasets characterised by many categorical features. As a result, CatBoost has demonstrated remarkable effectiveness in various real-world applications (Chelgani et al., 2024).

Recently, Chelgani et al. (2024) conducted a comparative analysis of the predictive effectiveness of various machine learning models, including Catboost, Random Forest, Support Vector Regression, Extreme Gradient Boosting and Convolutional Neural Networks. Their results showed that Catboost outperformed the other models, achieving an impressive accuracy (R^2: 0.90) in predicting the metallurgical responses of copper flotation, particularly in terms of grade and recovery. In the context

of coal flotation, Bu et al. (2024) collected a dataset of 641 valid samples for model training and validation. They developed an enhanced deep neural network (DNN) architecture to predict the quality of flotation products. Their results showed that the proposed DNN model yielded superior R^2 values - 0.71 for concentrate yield and 0.87 for concentrate ash content - compared to Random Forest. However, it is noteworthy that the permutation feature importance analysis could not be performed for all input features in the DNN model proposed by Bu et al. (2024). This limitation highlights the urgent need to develop highly efficient and accurate models that can provide robust predictive performance across the full set of 641 datasets.

In this study, a comparative analysis of the predictive performance of GBDT, CatBoost, LightGBM, and XGBoost in the context of coal flotation was conducted. The flotation data were categorized into two different groups based on concentrate levels and were used to explore the reasons for the low prediction accuracy of flotation results. Also, the feature importance has been analyzed using different datasets.

2. Materials and methods

2.1. Data collection

A comprehensive dataset of 641 instances of laboratory unit flotation data for coal samples from different regions was curated from the publicly available supplementary material detailed in the literature Bu et al. (2024). The dataset encompasses eight input parameters alongside four output parameters, as delineated in Table 1. Our analysis concentrates exclusively on the modeling and prediction of concentrate yield and ash content. Fig. 1 elucidates the complex interrelationships among concentrate yield, ash content, and tailings. The concentrate yields display a remarkable range, oscillating between 1.3% and 97.1%, while the ash content spans from 2.97% to 33.2%. Tailings are similarly observed to fluctuate between 17.3% and 84.5%. This considerable variability in concentrate yield and ash content underscores the dataset's ability to encapsulate a diverse spectrum of feed characteristics and flotation conditions.

Table 1. Input and output parameters in the dataset

Parameter types		Parameters
Input parameters	Feed properties	Ash content of feed
		Pulp density
		Fraction of -74 μm in feed
		Ash content of -74 μm in feed
	Flotation conditions	Collector dosage
		Frother dosage
		Aeration
		Rotation speed
Output parameters	Concentrate properties	Concentrate yield
		Concentrate ash content
	Tailing properties	Tailing yield
		Tailing ash content

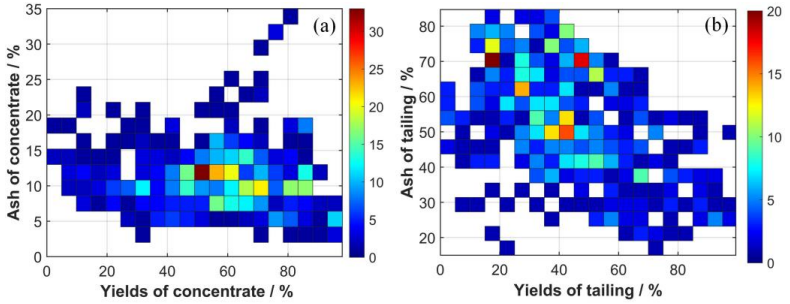

Fig. 1. Relationship between yield and ash content of flotation concentrate (a) and tailing (b) in the full dataset

2.2. Regression prediction

Model training and data analysis were performed using the SPSSPRO software suite. The dataset, consisting of 641 instances of flotation data, was divided into training and test sets in an 8:2 ratio. This study is based on a decision tree framework inspired by the boosting methodology, specifically using the gradient boosting algorithm for our investigations. We selected four widely used gradient boosting decision tree models recognized for their superior predictive performance: GBDT, CatBoost, LightGBM and XGBoost. Compared to other machine learning models, such as Support Vector Machine (SVM), Random Forests (RF), and Artificial Neural Networks(ANN), the four algorithms analyzed in this paper offer several advantages, including high robustness, strong generalization capabilities, and rapid training times. Additionally, these algorithms can effectively assess the importance of various features, providing valuable insights for understanding and evaluating the models. Using these selected models, we performed regression analysis to predict both coal yield and ash content.

2.3. Model optimization

Hyperparameters are parameters that need to be manually set before training a machine learning model, such as learning rate and tree depth. The choice of hyperparameters can have a large impact on the performance of the model, so hyperparameter optimization is needed to find the best combination of hyperparameters to improve the performance and generalization of the model. The Bayesian algorithm is used to optimize the model in this model building process. The hyperparameters for the four models to be optimized are shown in Table 2.

Table 2. Hyperparameters optimized for different models

GBDT	CatBoost	LightGBM	XGBoost
Loss function	Number of iterations	Base learner	Base learner
Node split criterion	Learning rate	Number of base learners	Number of base learners
Number of base learners	L2 regularization term	Learning rate	Learning rate
Learning rate	Maximum depth of tree	L1 regularization term	L1 regularization term
Sampling ratio without replacement	Overfitting detection threshold	L2 regularization term	L2 regularization term
Maximum features ratio considered for splitting	Number of iterations after convergence	Sample feature sampling rate	Sample feature sampling rate
Minimum samples for split in internal nodes	-	Tree feature sampling rate	Tree feature sampling rate
Minimum samples for leaf node	-	Node split threshold	Node split threshold
Minimum weight of samples in leaf node	-	Minimum weight of samples in a leaf node	Minimum weight of samples in a leaf node

3. Results and discussion

3.1. Regression prediction for the all datasets

Four regression models, GBDT, CatBoost, LightBGM and XGBoost, were used to learn the training set (512 records) and predict the test set (129 records) respectively. The training and test sets are the same for all four models. The hyperparameters are optimised using a Bayesian approach to obtain the optimal prediction model. The relationship between the real and predicted data for refined mineral yield and ash content after optimisation of the different models is shown in Fig. 2 and Fig. 3. The coefficient of determination (R^2) of the four regression models, GBDT, CatBoost, LightBGM and XGBoost, for the prediction of concentrate yield were 0.764, 0.827, 0.800, 0.836 and the R^2 for the prediction of ash content were 0.688, 0.829, 0.817 and 0.853 respectively. The results indicate that the XGBoost model has the

highest accuracy for both yield and ash prediction. However, the prediction accuracy of all four models was insufficient for concentrate yield and ash. As can be seen from the offset law between the predicted and actual values in Fig. 2, the prediction accuracy for concentrate yield below 40% is significantly lower than that for concentrate yield above 40%.

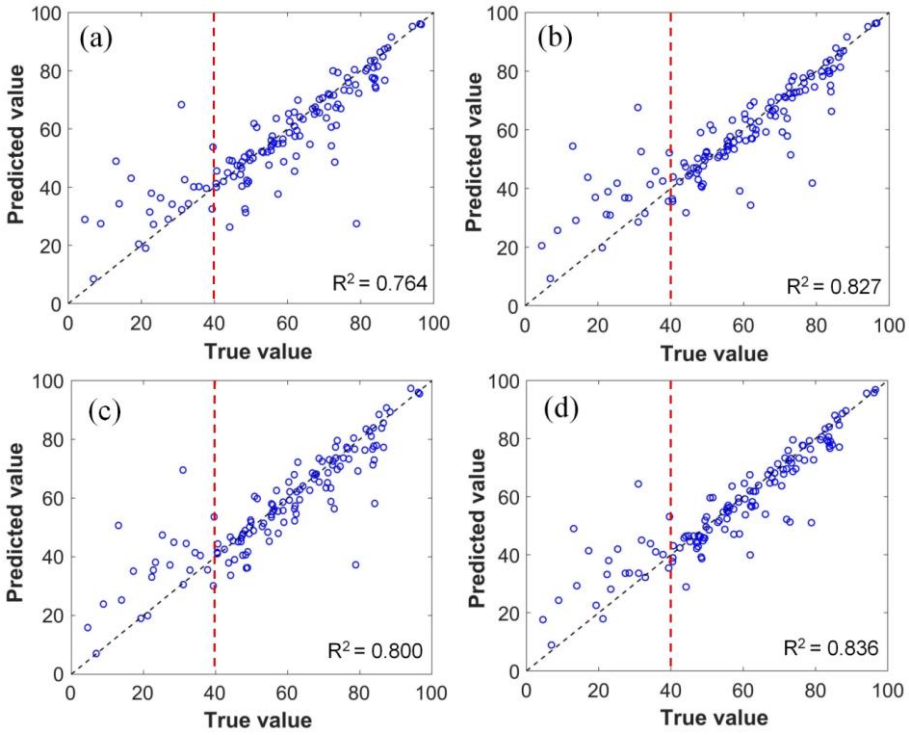

Fig. 2. Test set prediction results of (a) GBDT, (b) CatBoost, (c) LightBGM, and (d) XGBoost model for the flotation concentrate yield

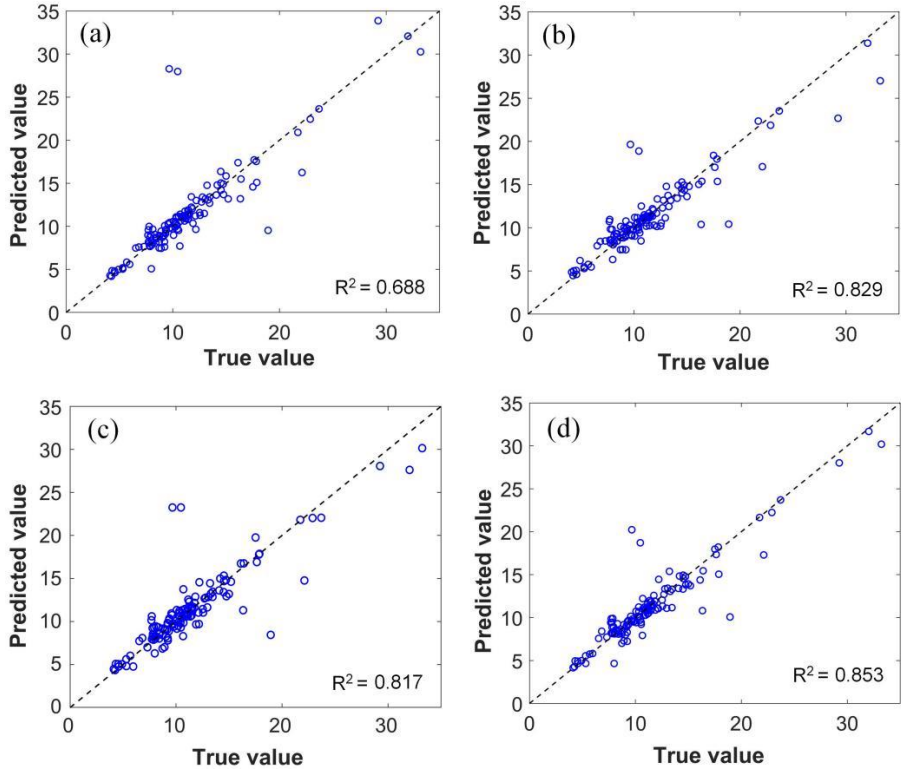

Fig. 3. Test set prediction results of (a) GBDT, (b) CatBoost, (c) LightBGM, and (d) XGBoost model for the ash content of flotation concentrate

Two representative sets of flotation data were selected from the dataset to determine the reasons for the large difference in prediction accuracy for concentrate yields less than 40% and greater than 40%. The results of the variation of flotation concentrate yield and ash with collector dosage (the ratio of collector to frother dosage is 3:1) are shown in Fig. 4. Data 1 and Data 2 show the flotation results of two coal pulps with different feed characteristics. The results of data 1 show that the concentrate yield and ash content are about 15% and 10%, respectively, when the collector dosage is less than 400 g/t. However, when the collector dosage is more than 600 g/t, the concentrate yield and ash content increase slowly after a sharp increase, and the yield and ash content are more than 50% and 20%, respectively. The results of data 2 show that the concentrate yield is 32% when the collector dosage is 100g/t, and when the collector dosage is more than 200g/t, the concentrate yield increases significantly to more than 60% and then increases slowly. It can be seen that the concentrate yield at low trapping agent dosage is significantly lower than the flotation yield at normal or excessive trapping agent dosage, thus showing different laws. Therefore, the use of the full dataset for machine learning may be the main reason for the low prediction accuracy in Fig. 2 and Fig. 3.

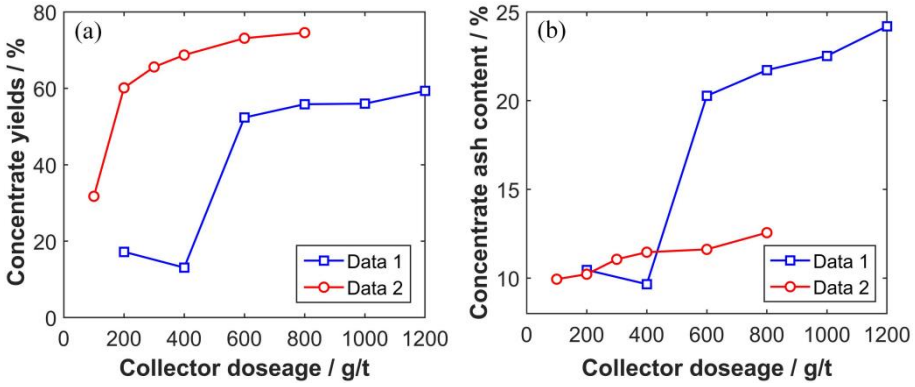

Fig. 4. (a) yield and (b) ash content of the flotation concentrate using different data

3.2. Regression prediction for the categorical datasets

The full dataset used for machine learning was categorised. Data with less than 40% concentrate recoveries were filtered into Dataset I (116 records), while data with recoveries greater than 40% were grouped into Dataset II (525 records). Dataset I consists mainly of the results when the flotation chemicals are used at low doses, while Dataset II consists mainly of the results when the flotation chemicals are used at normal or excessive doses. These two datasets were randomised separately and then 80% of the data was used as the training set and the remaining 20% as the test set. Four machine learning methods, GBDT, CatBoost, LightBGM and XGBoost, were used to train modelling and regression prediction respectively on the two datasets. The comparison of predicted and true values for Dataset I and Dataset II is shown in Fig. 5 and Fig. 6.

The prediction accuracy of the four machine learning methods is evaluated by three metrics, Root Mean Squard Error (RMSE), Mean Absolute Error (MAE) and R^2, and the results are shown in Fig. 7. The smaller the calculation results of MAE and RMSE, the smaller the error value between the predicted value and the actual value, and the model predicted value has better reliability. The value of R^2 is between 0 and 1, and the closer R^2 is to 1, the better the fitting effect works. The prediction models built by the four machine learning methods for Dataset I have high MAE and RMSE and very low R^2 (below 0.4). This indicates that it is difficult for all four machine learning methods to build accurate predictive models for Dataset I. When the flotation reagent dosage is low, the flotation concentrate yield and ash content have a certain degree of randomness and poor regularity. This leads to the difficulty of regression prediction of this part of the data using machine learning methods. The accuracy of the prediction models built by the four models for Dataset II was significantly improved when the data with flotation concentrate yield less than 40% in the original dataset were removed. Among the four machine learning methods, the accuracy of the prediction model XGBoost>LightBGM>CatBoost>GBDT was the best. Taking the prediction model established by XGBoost as an example, compared to the full dataset, the RMSE, MAE and R2 metrics of the concentrate yield prediction model established for

Dataset I were improved from 8.37, 5. 28 and 0.836 to 5.08, 3.56 and 0.877 respectively, and those of the concentrate ash content prediction model were improved from 1.83, 0.96 and 0.853 to 0.85, 0.57 and 0.971 respectively. Compared to the whole Dataset and Dataset I, the reliability and fit of the prediction model for Dataset II were significantly improved. From this result, it can be speculated that machine

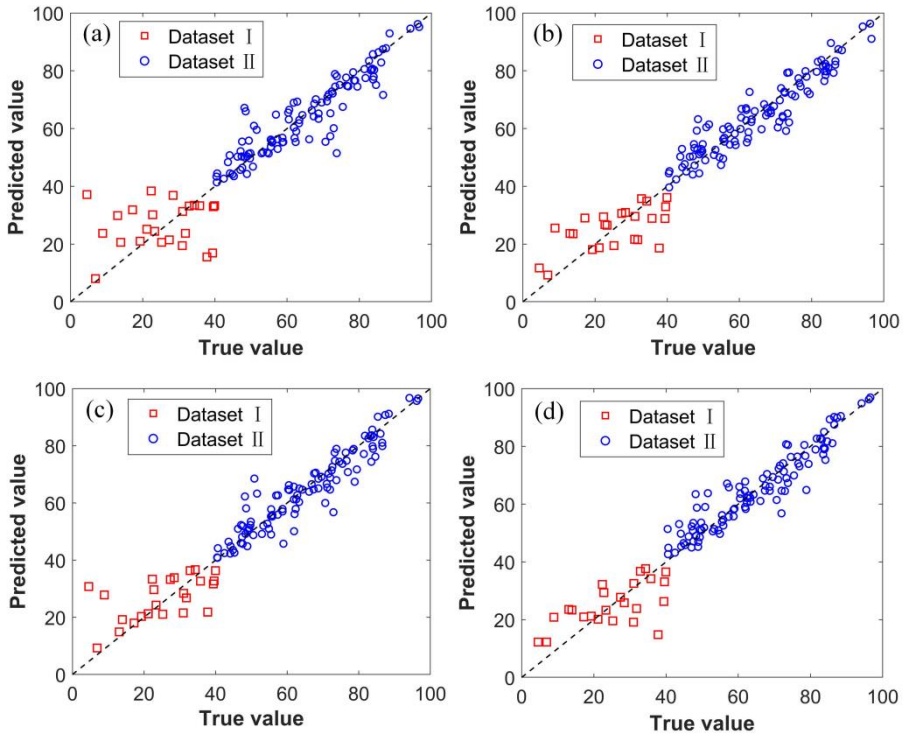

Fig. 5. Test set prediction results of (a) GBDT, (b) CatBoost, (c) LightBGM, and (d) XGBoost models for the flotation concentrate yield using different dataset

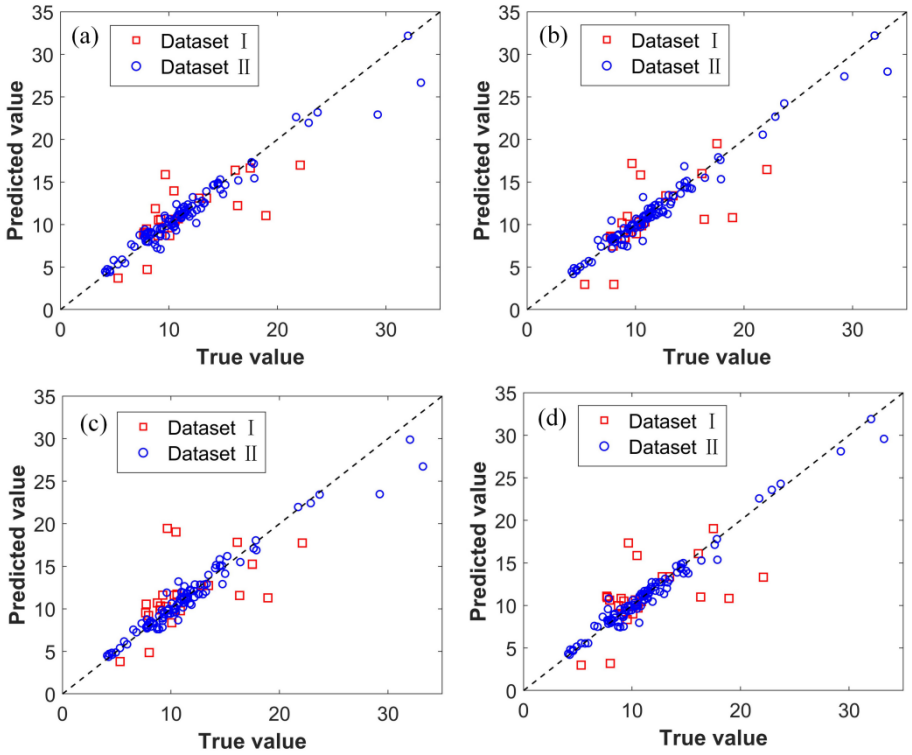

Fig. 6. Test set prediction results of (a) GBDT, (b) CatBoost, (c) LightBGM, and (d) XGBoost models for the ash content the flotation concentrate using different dataset

learning modelling using data results of appropriate flotation chemical dosing may be more suitable for online prediction of concentrate yield and ash content in future flotation plants.

According to the collected flotation data (Bu et al., 2024), it can be seen that the corresponding flotation reagent dosage is usually lower when the flotation yield is below 40%. The basic requirement for a high flotation selectivity is the existence of a significant difference in hydrophobicity between coal and gangue (Ramudzwagi et al., 2020). When the amount of collector is low, the difference in hydrophobicity between coal and gangue is too small, making it difficult for coal particles to adhere to the surface of bubbles and enter the concentrate product (Bu et al., 2020). In addition, when the amount of frother is too low, it is difficult to form a stable foam layer in flotation, resulting in that hydrophobic coal particles adhered to the bubble surface cannot be recycled into concentrate products through the foam area (Pawliszak et al., 2024). The lower dosage of flotation reagents results in very poor selectivity of flotation results, leading to high randomness and poor predictability of flotation results. Therefore, appropriate screening of collected flotation data is an important step in ensuring the accuracy and effectiveness of flotation prediction.

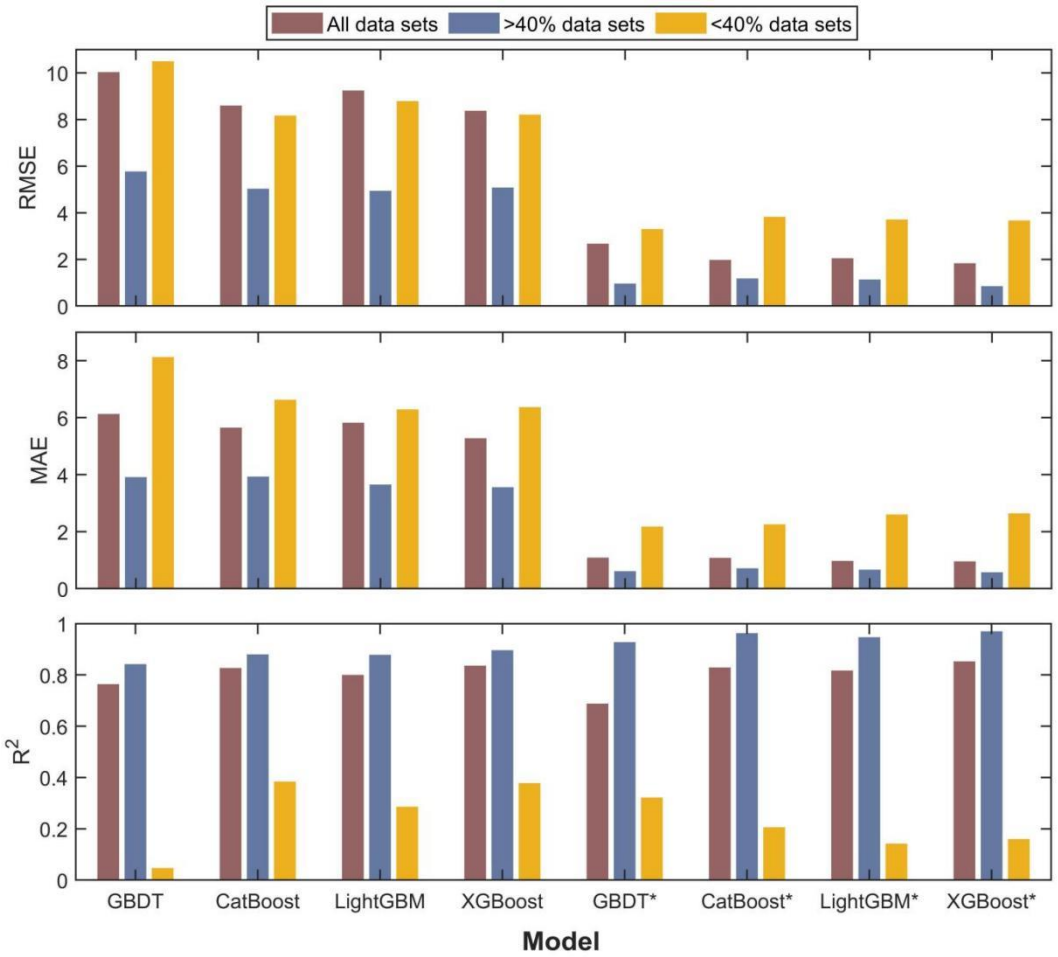

Fig. 7. Comparison of the accuracy of the prediction models of the four machine learning methods

3.3. Feature importance analysis

The results of the importance analyses of the prediction models for the full dataset are shown in the Supporting Information (**Fig. R1** and **Fig. R2**). The results of the importance analysis of concentrate yield show that the importance of frother dosing is always in the top two in the order of importance obtained by all models. The presence of a frother reduces the surface tension, which is beneficial for inhibiting bubble incorporation and maintaining the stability of the foam layer. Therefore, frother dosage is a key factor influencing concentrate yield. The results of the importance analysis of the

prediction model for Dataset II are shown in Fig. 8 and Fig. 9. When the part of the dataset with a yield of <40% is removed, the importance of the frother for the concentrate yield decreases significantly. This may be due to the fact that with a sufficient amount of frother, any further increase in the amount of frother will not significantly improve the concentrate yield. Meanwhile, the results also show that the increase in frother dosage is the main reason for the apparent increase in concentrate yield in Fig. 4(a), which is also an important reason why frother dosage has a higher degree of importance in the prediction model for the full dataset. The relationship between the importance level of the full dataset and the other factors in Dataset II did not change significantly except for frother dosage.

The fraction and ash content of -74 um in the feed are the two factors that rank in the top three in the importance results of most models. Flotation is a method of mineral separation based on the differential adhesion of hydrophobic and hydrophilic minerals to flotation bubbles. The hydrophobic mineral particles adhere to the bubbles and move from the slurry zone to the froth zone, eventually becoming the concentrate. It is well known that fine particles (-74 um) have a low probability of collision with flotation bubbles due to their fine size and low inertia. In addition, the high specific surface area of the fine particles will consume a large amount of the limited flotation chemicals, leaving the coarse particles unable to obtain sufficient flotation reagent to mineralise with the bubbles. With the increase of the -74 um fraction in the feed, the limited amount of reagent has been consumed in large quantities. This results in a large loss of coarse particles with good floatability in the flotation concentrate and a significant reduction in concentrate yield. In addition, as the ash content of -74 μm increases, a large number of fine vein particles present in the slurry will cover the fine coal particles. As a result, the fine coal particles are unable to interact with the flotation chemicals and bubbles and are eventually lost in the tailings.

Compared to the amount of foaming agent, the amount of trapping agent was significantly less important for concentrate yield and ash, both ranking fifth in all models. This is due to the fact that coal particles tend to float naturally and do not require excessive amounts of foaming agent. Feed ash content, -74 um content and ash content all reflect to some extent the degree of floatability of the feed. Therefore, the characteristics of the feed have a greater influence on the flotation results than the dosage of the collector. Also, slurry concentration, agitation rate and aeration were the three factors that had the least effect on concentrate yield of all the predictive models, which may be due to the fact that these three factors are less controllable in flotation experiments.

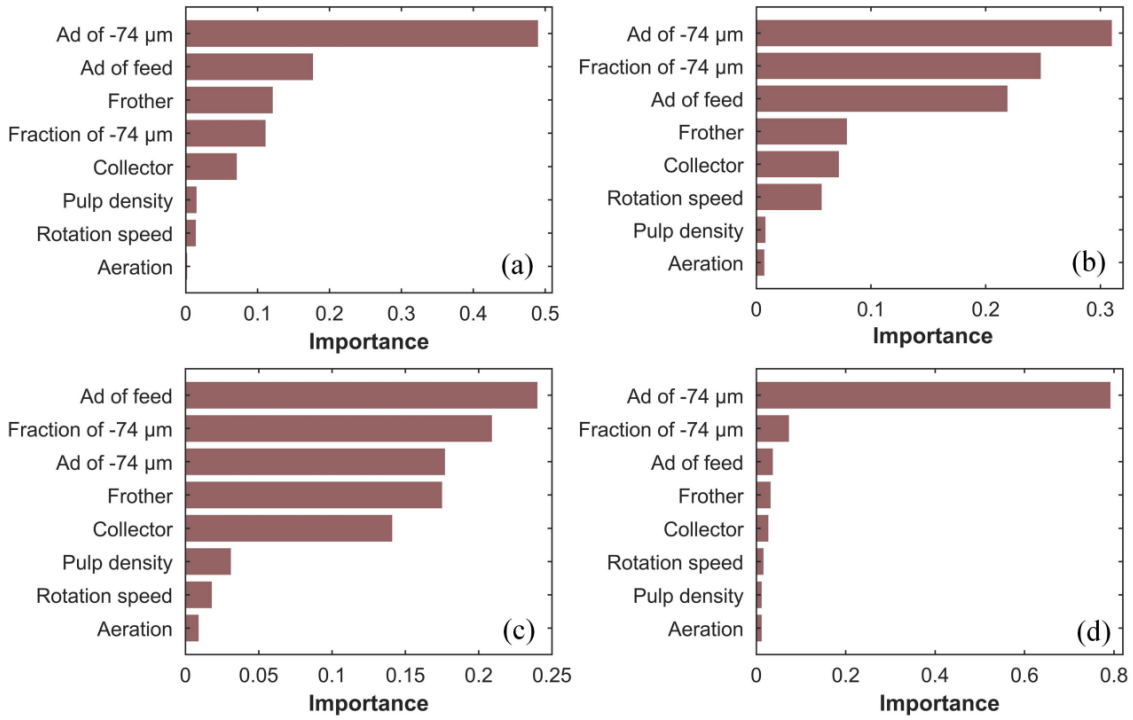

Fig. 8. Feature importance results for the flotation concentrate yield using (a) GBDT, (b) CatBoost, (c) LightBGM, and (d) XGBoost models

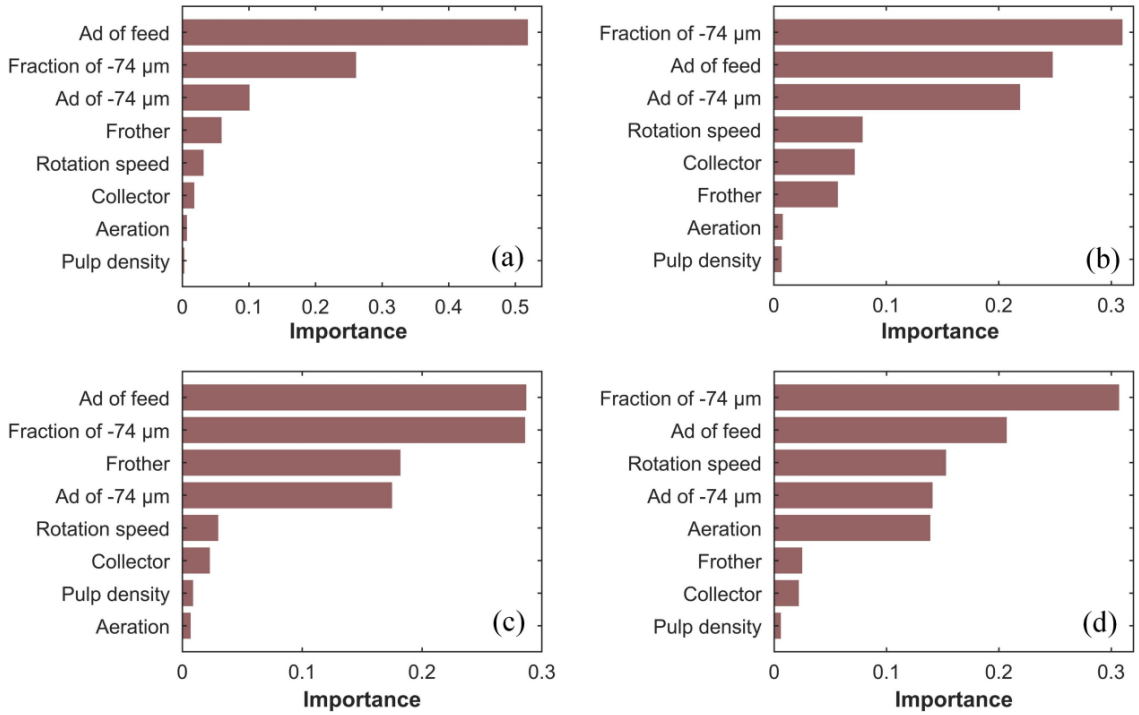

Fig. 9. Feature importance results for the ash content of the flotation concentrate using (a) GBDT, (b) CatBoost, (c) LightBGM, and (d) XGBoost models

4. Conclusions

This study effectively evaluated the predictive performance of GBDT, CatBoost, LightGBM, and XGBoost for coal flotation results. The main findings are as follows:

1. The XGBoost model outperformed the others, achieving high prediction accuracy with R^2 values of 0.877 and 0.971 for concentrate yield and ash content, respectively.
2. The low prediction accuracy for concentrate yield below 40% was attributed to the abnormal flotation results caused by low flotation reagent dosages. Filtering out this data subset significantly improved the model accuracy.
3. The fraction and ash content of -74 μm particles in the feed were identified as the key factors affecting flotation results, emphasizing the importance of feed properties over flotation reagent dosages.
4. The results indicate that machine learning models based on appropriately dosed flotation chemicals could potentially provide accurate online predictions for coal flotation plants, facilitating intelligent control and management.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (52204296).

References

ABKHOSHK, E., KOR, M., REZAI, B., 2010. *A study on the effect of particle size on coal flotation kinetics using fuzzy logic*. Expert Syst. Appl. 37, 5201-5207.

ALI, D., HAYAT, M.B., ALAGHA, L., MOLATLHEGI, O.K., 2018. *An evaluation of machine learning and artificial intelligence models for predicting the flotation behavior of fine high-ash coal*. Adv. Powder Technol. 29, 3493-3506.

ARANCIBIA-BRAVO, M.P., LUCAY, F.A., SEPúLVEDA, F.D., CORTéS, L., CISTERNAS, L.A., 2022. *Response Surface Methodology for Copper Flotation Optimization in Saline Systems*. Minerals 12, 1131.

BU, X., VAHED, A.T., GHASSA, S., CHELGANI, S.C., 2021. *Modelling of coal flotation responses based on operational conditions by random forest*. Int. J. Oil Gas Coal Technol. 27, 457-468.

BU, X., XIE, G., PENG, Y., CHEN, Y., 2016. *Kinetic modeling and optimization of flotation process in a cyclonic microbubble flotation column using composite central design methodology*. Int. J. Miner. Process. 157, 175-183.

BU, X., ZHANG, T., CHEN, Y., XIE, G., PENG, Y., 2020. *Comparative study of conventional cell and cyclonic microbubble flotation column for upgrading a difficult-to-float Chinese coking coal using statistical evaluation*. Int. J. Coal Prep. Util. 40, 359-375.

BU, X., ZHOU, S., DANSTAN, J.K., BILAL, M., UL HASSAN, F., CHAO, N., 2024. *Prediction of coal flotation performance using a modified deep neural network model including three input parameters from feed*. Energy Sources Part A-Recovery Util. Environ. Eff. 10.1080/15567036.2022.2036272, 1-13.

CARMONA, P., CLIMENT, F., MOMPARLER, A., 2019. *Predicting failure in the U.S. banking sector: An extreme gradient boosting approach*. Int. Rev. Econ. Financ. 61, 304-323.

CHEHREH CHELGANI, S., MATIN, S.S., MAKAREMI, S., 2016. *Modeling of free swelling index based on variable importance measurements of parent coal properties by random forest method*. Measurement 94, 416-422.

CHELGANI, S.C., HOMAFAR, A., NASIRI, H., LAKSAR, M.R., 2024. *CatBoost-SHAP for modeling industrial operational flotation variables - A "conscious lab" approach*. Miner. Eng. 213, 108754.

CHELGANI, S.C., MATIN, S.S., 2018. *Study the relationship between coal properties with Gieseler plasticity parameters by random forest*. Int. J. Oil Gas Coal Technol. 17, 113-127.

DU PLESSIS, B.J., DE VILLIERS, G.H., 2007. *The application of the Taguchi method in the evaluation of mechanical flotation in waste activated sludge thickening*. Resour. Conserv. Recycl. 50, 202-210.

FLORES, V., HENRíQUEZ, N., ORTIZ, E., MARTINEZ, R., LEIVA, C., 2024. *Random forest for generating recommendations for predicting copper recovery by flotation*. IEEE Latin Am. Trans. 22, 443-450.

GOMEZ-FLORES, A., HEYES, G.W., ILYAS, S., KIM, H., 2022. *Prediction of grade and recovery in flotation from physicochemical and operational aspects using machine learning models*. Miner. Eng. 183, 107627.

GUNER, M., AKYILDIZ, O., BASARIR, H., KOWALCZUK, P., 2024. *Exploring the impact of thiol collectors system on copper sulfide flotation through machine learning-driven modeling*. Physicochem. Probl. Mineral Pro. 60, 191709.

LUBISI, T.P., NHETA, W., NTULI, F., 2018. *Optimization of Reverse Cationic Flotation of Low-Grade Iron Oxide from Fluorspar Tails Using Taguchi Method*. Arab. J. Sci. Eng. 43, 2403-2412.

MA, X., SHA, J., WANG, D., YU, Y., YANG, Q., NIU, X., 2018. *Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning*. Electron. Commer. Res. Appl. 31, 24-39.

MENG, S., WEN, S., HAN, G., WANG, X., FENG, Q., 2022. *Wastewater Treatment in Mineral Processing of Non-Ferrous Metal Resources: A Review*. Water 14, 726.

NAZARI, S., CHEHREH CHELGANI, S., SHAFAEI, S.Z., SHAHBAZI, B., MATIN, S.S., GHARABAGHI, M., 2019. *Flotation of coarse particles by hydrodynamic cavitation generated in the presence of conventional reagents*. Sep. Purif. Technol. 220, 61-68.

PAWLISZAK, P., BRADSHAW-HAJEK, B.H., SKINNER, W., BEATTIE, D.A., KRASOWSKA, M., 2024. *Frothers in flotation: A review of performance and function in the context of chemical classification*. Miner. Eng. 207, 108567.

RAMUDZWAGI, M., TSHIONGO-MAKGWE, N., NHETA, W., 2020. *Recent developments in beneficiation of fine and ultra-fine coal -review paper*. J. Clean Prod. 276, 122693.

SACHINRAJ, D., KOPPARTHI, P., SAMANTA, P., MUKHERJEE, A.K., 2022. *Optimization of Column Flotation for Fine Coal Using Taguchi Method*. Trans. Indian Inst. Met. 75, 1255-1267.

SHAHBAZI, B., CHEHREH, C.S., MATIN, S.S., 2017. *Prediction of froth flotation responses based on various conditioning parameters by Random Forest method*. Colloid Surf. A-Physicochem. Eng. Asp. 529, 936-941.

SUN, X., LIU, M., SIMA, Z., 2020. *A novel cryptocurrency price trend forecasting model based on LightGBM*. Financ. Res. Lett. 32, 101084.

SUN, Y., BU, X., ULUSOY, U., GUVEN, O., VAZIRI HASSAS, B., DONG, X., 2023. *Effect of surface roughness on particle-bubble interaction: A critical review*. Miner. Eng. 201, 108223.

VINNETT, L., LEóN, R., MESA, D., 2023. *Artificial neural network (ANN) modelling to estimate bubble size from macroscopic image and object features*. Physicochem. Probl. Mineral Pro. 59, 185759.

WANG, C., WANG, H., LIU, Y., HUANG, L., 2016. *Optimization of surface treatment for flotation separation of polyvinyl chloride and polyethylene terephthalate waste plastics using response surface methodology*. J. Clean Prod. 139, 866-872.

WANG, X., BU, X., NI, C., ZHOU, S., YANG, X., ZHANG, J., ALHESHIBRI, M., PENG, Y., XIE, G., 2021. *Effect of scrubbing medium's particle size on scrubbing flotation performance and mineralogical characteristics of microcrystalline graphite*. Miner. Eng. 163, 106766.

ZHANG, C., ZHANG, Y., SHI, X., ALMPANIDIS, G., FAN, G., SHEN, X., 2019. *On Incremental Learning for Gradient Boosting Decision Trees*. Neural Process. Lett. 50, 957-987.

ZHANG, L., JáNOŠíK, D., 2024. *Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches*. Expert Syst. Appl. 241, 122686.

ZHAO, B., HU, S., ZHAO, X., ZHOU, B., LI, J., HUANG, W., CHEN, G., WU, C., LIU, K., 2022. *The application of machine learning models based on particles characteristics during coal slime flotation*. Adv. Powder Technol. 33, 103363.

Supporting material

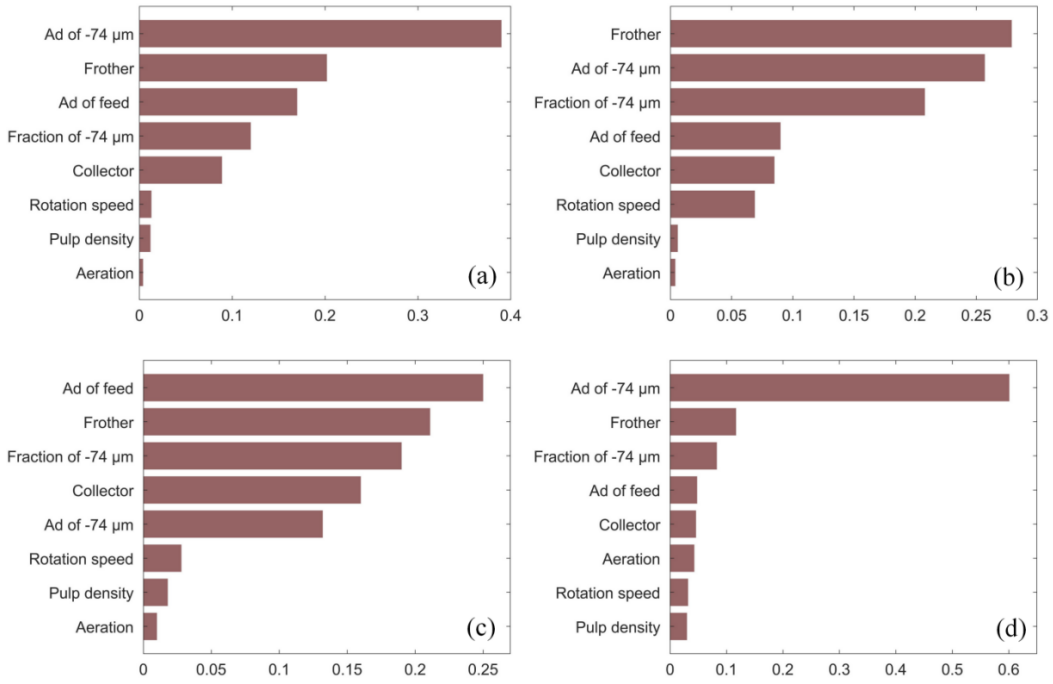

Fig. R1. Feature importance of concentrate yield prediction model of full dataset

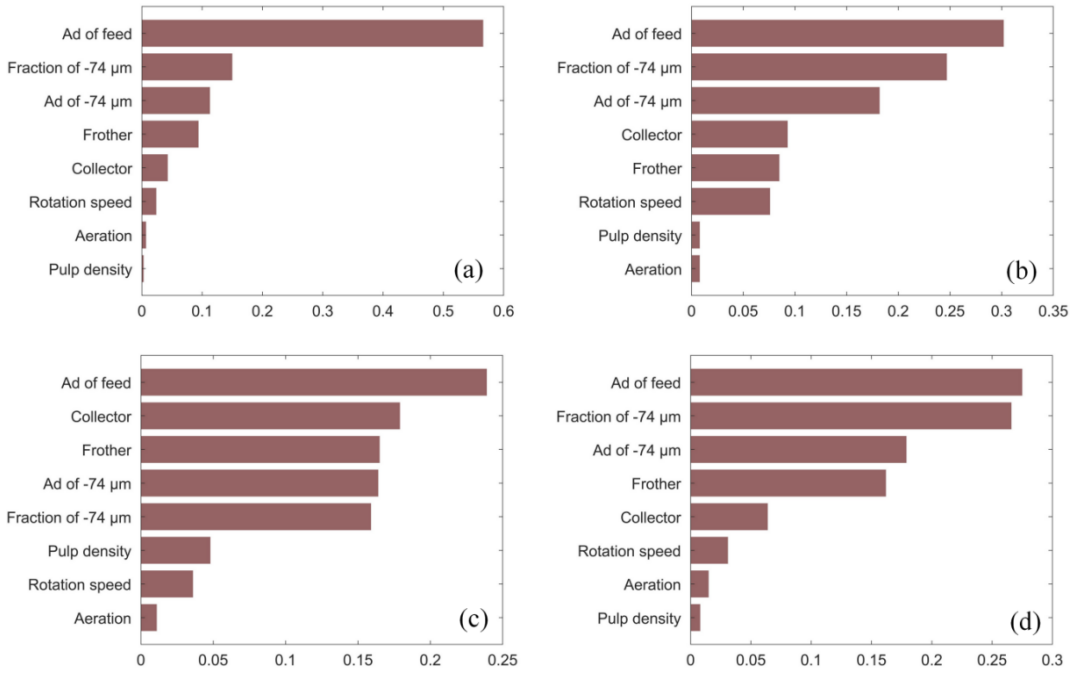

Fig. R2 Feature importance of concentrate ash content prediction model of full dataset