minerals

MDPI

Article

Machine Learning-Based Uranium Prospectivity Mapping and Model Explainability Research

Weihao Kong [1,2,3], Jianping Chen [1,2,*] and Pengfei Zhu [3]

[1] School of Earth Sciences and Resources, China University of Geosciences (Beijing), Beijing 100083, China; 3001190111@email.cugb.edu.cn
[2] Beijing Key Laboratory of Land and Resources Information Research and Development, Beijing 100083, China
[3] CNNC Key Laboratory of Uranium Resources Exploration and Evaluation Technology, Beijing Research Institute of Uranium Geology, Beijing 100029, China; zpfjpu@126.com
* Correspondence: 3s@cugb.edu.cn

Abstract: Sandstone-hosted uranium deposits are indeed significant sources of uranium resources globally. They are typically found in sedimentary basins and have been extensively explored and exploited in various countries. They play a significant role in meeting global uranium demand and are considered important resources for nuclear energy production. Erlian Basin, as one of the sedimentary basins in northern China, is known for its uranium mineralization hosted within sandstone formations. In this research, machine learning (ML) methodology was applied to mineral prospectivity mapping (MPM) of the metallogenic zone in the Manite depression of the Erlian Basin. An ML model of 92% accuracy was implemented with the random forest algorithm. Additionally, the confusion matrix and receiver operating characteristic curve were used as model evaluation indicators. Furthermore, the model explainability research with post hoc interpretability algorithms bridged the gap between complex opaque (black-box) models and geological cognition, enabling the effective and responsible use of AI technologies. The MPM results shown in QGIS provided vivid geological insights for ML-based metallogenic prediction. With the favorable prospective targets delineated, geologists can make decisions for further uranium exploration.

Keywords: MPM; post hoc; machine learning; explainability; random forest; sandstone-hosted uranium

Citation: Kong, W.; Chen, J.; Zhu, P. Machine Learning-Based Uranium Prospectivity Mapping and Model Explainability Research. *Minerals* **2024**, *14*, 128. https://doi.org/10.3390/min14020128

Academic Editor: José António de Almeida

Received: 25 December 2023
Revised: 17 January 2024
Accepted: 22 January 2024
Published: 24 January 2024

1. Introduction

The International Atomic Energy Agency (IAEA) has developed a descriptive classification system for uranium deposits. According to this classification, there are 13 types of uranium deposits recognized worldwide [1]. According to the statistical analysis conducted by the Nuclear Energy Agency of the Organization for Economic Cooperation and Development (OECD/NEA) and IAEA, there are 1430 sandstone-hosted uranium deposits globally. These deposits account for 39.6% of the total number of 3610 uranium deposits in the world, making them the most abundant type of uranium deposit. According to the information provided, sandstone-hosted uranium deposits account for 7.9% of the world's total uranium resources. This places them in the fourth position in terms of quantity, following black shale-hosted deposits (35.2%), phosphorite-hosted deposits (22.6%), and lignite coal-hosted deposits (11.4%). Sandstone-hosted deposits have a total resource of 5,095,214 tU of uranium.

Sandstone-hosted uranium deposits are considered an important uranium source globally and nationally, with 41 sandstone-hosted deposits in China accounting for 28% of the total in China. Sandstone-hosted uranium deposits are particularly significant in northern China due to their shallow burial depth and low exploitation cost. Mineral prospectivity mapping (MPM) with interdisciplinary knowledge has been consistently used to support decision making in sandstone-hosted uranium exploration in China. MPM geologists

propose the comprehensive utilization of multiple kinds of geoscience information to establish quantitative prediction models [2–6]. These models extract significant mineralization indicators and ore-controlling features from different types of deposits and metallogenic processes [7–10].

Machine learning, as an AI approach, has been a core area of scientific research since the 20th century. Various algorithms, such as the Hebbian learning rule, perceptron, back propagation (BP), decision tree, support vector machine (SVM), random forest (RF), and deep learning, have been developed and used in best practices over time [11–24]. These algorithms have been applied in different fields in the past two decades, including statistics, engineering, signal processing, and computer vision. Specifically, the application of machine learning in MPM has gained significant attention [25–28]. Machine learning has the capacity to handle a large volume of evidence characteristic layers associated with mineralization. Moreover, they have the potential to identify nonlinear relationships between known deposits and evidence layers. With advancements in technology, such as faster computers and larger datasets, machine learning algorithms have shown their full potential [29,30]. It is now recognized that a good representation of data and the availability of large amounts of example data are crucial for successful machine learning [31,32]. ML has been actively conducted in the mining industry since 2018, primarily for mineral exploration [33]. And it has been widely utilized in various applications within the mining and mineral industry [34]. When the fourth wave of data-driven science, known as a paradigm shift, emerged, the tools used by MPM transitioned from GIS to AI. Geologists have shown optimism toward this transition, but they have also expressed skepticism regarding the reliability of AI's results. This skepticism arises from the limited utilization of AI's explanatory capabilities, which differ from the comprehensibility of traditional methods like weight of evidence (WOE). The paper presents a comprehensive machine learning workflow for uranium prospectivity mapping. It focuses on the use of post hoc interpretability algorithms to ensure transparency, gain insights into geological processes, assess risks, and ensure regulatory compliance.

2. Study Area and Mineral Prospectivity Model

2.1. Geological Setting

The Erlian Basin is a large Meso-Cenozoic fault-depression composite basin located at the tectonic position of the suture between the Siberian plate and the North China plate [35]. It is developed on the Xingmeng Hercynian folded basement and consists of six geotectonic units: Wulanchabu depression, Chuanjing depression, Manite depression, Tenggeer depression, Wunite depression, and Sunite Uplift. Additionally, there are 53 depressions and 22 uplifts within the basin's internal tertiary geotectonic units [36]. These geological characteristics are essential for understanding the metallogenic processes of the Erlian Basin [37].

The Manite depression, with the study area shown in Figure 1, is a zonal valley that was formed via tectonic activity. It exhibits the characteristics of an ancient valley tectonic formation and was subsequently filled with alluvium and lacustrine sediments [38]. The strata in this area consist of the Lower Cretaceous Saihan lower group (K_1s^1), Saihan upper group (K_1s^2), the Paleogene Ildinmanha Group (E_2y), and the Quaternary (Q). The Saihan upper group (K_1s^2) is widely distributed in the study area and has a sedimentary thickness ranging from 50 to 400 m. The redox zone in this sedimentary stratum exhibits different colors due to varying degrees of oxidation and is considered the main target for prospecting sandstone-hosted uranium deposits. The main lithologies found in this group include sandstone, pebbly sandstone, argillaceous sandstone, and argillaceous siltstone. These lithologies display various colors such as grey, grey-green, and yellow, which can be attributed to different levels of oxidation.

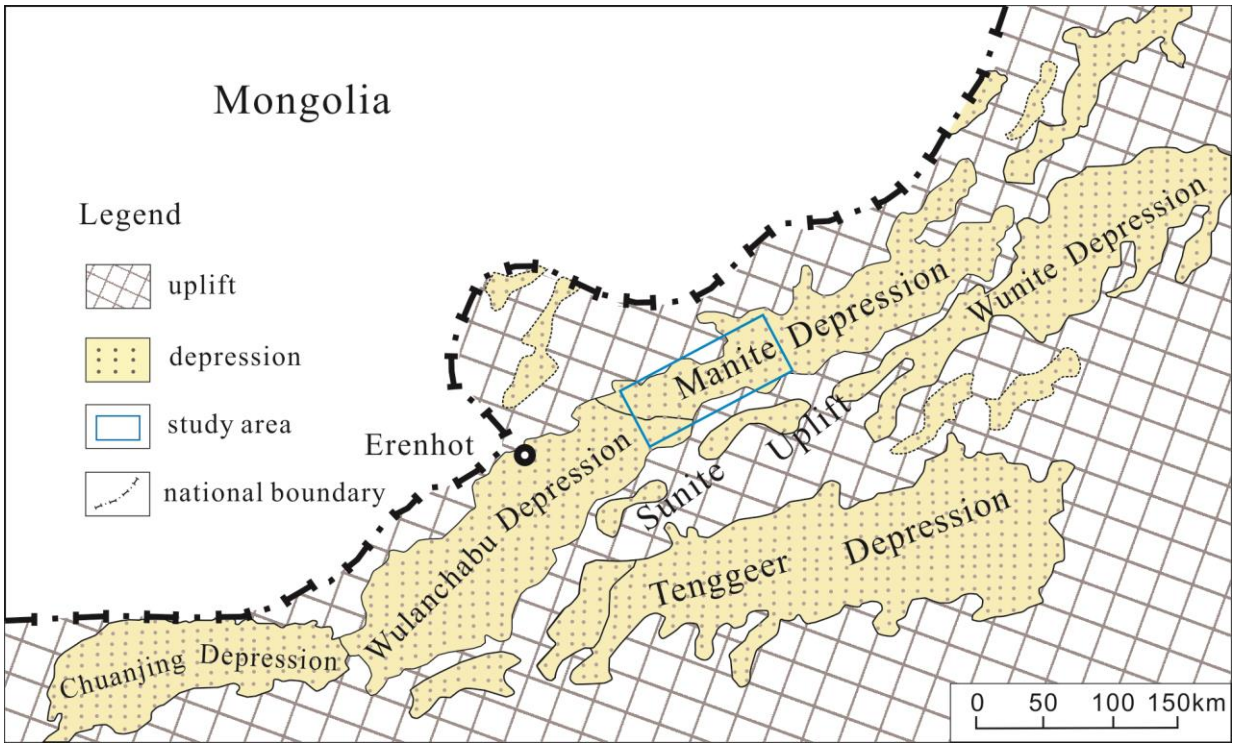

Figure 1. Geological location of study area.

2.2. Prospectivity Model Establishment

The mineral prospectivity model is a crucial component of MPM. It is mainly based on the study of the metallogenic geological setting and metallogenic period [39]. The model aims to establish relationships between geological features of deposit formation and anomalies observed in geoscience data. By summarizing prospectivity criteria, methods, and approaches, the model provides a qualitative description of the geological settings and quantitative features of the deposit. The accuracy of the model can vary depending on the different geological exploration phases [40].

The mineral prospectivity model for the study area is constructed based on the uranium metallogenic theory and the research on the genesis and metallogenic patterns of sandstone-hosted uranium deposits. The model considers the combination of strata, paleochannel, longitudinal bar, and sand body as the typical metallogenic geological conditions and ore-bearing patterns in the area. Moreover, inspired by the big data machine learning prospecting model [41,42], a mineral prospectivity model suitable for MPM in this study area has been developed. The model list in Table 1 provides a scientific guide for the quantitative prediction of uranium resources in the study area.

Table 1. Mineral prospectivity model.

Feature Category	Feature Type	Ore Controlling Features	Feature Description
Metallogenic geological setting	Geotectonic setting	Depression	Manite depression
	Formation	Saihan upper group	Prospecting stratum
Mineralization period	Ore-bearing rock	Clastic rock	Grey sand body
	Sedimentary system	Braided fluvial facies	Longitudinal bar
	Migration conditions	River centerline	The metallogenic position is within the range of paleochannel.
	Characteristics of ore body	The intersection of rivers	The scale of ore body at the intersection of river courses increases.

3. Methodology

3.1. Dataset Establishment

The dataset used in the study was derived from the quantitative features of the mineral prospectivity model. The data in the dataset should be numerical or able to be transformed into numerical data. There are four features of the Saihan upper group that make up the dataset in the study based on inductive bias from the experts, namely sedimentary facies, sand thickness, sand rate, and grey sand rate. Sedimentary facies (FAS) as discrete data were cataloged into three facies. The others as continuous data were kept in the original format. A total of 970 borehole data labeled as 0 or 1 were used to perform supervised machine learning, as shown in Table 2. The dataset used for training and testing the machine learning model consisted of 506 negative samples labeled as 0 and 464 positive samples labeled as 1. The dataset was roughly balanced, with a similar number of samples for both classes.

Table 2. Feature description of ML dataset.

Feature	FAS	ST	SR	GR	Target
Full Name	Sedimentary Facies	Sand Thickness	Sand Rate	Grey Sand Rate	Mineralized Borehole
Variable Types	Categorical	Numeric	Numeric	Numeric	Categorical
Value Domain	4, 5, 6	9.14 m~106.43 m	20%~89.4%	9.77%~82.99%	0, 1
Geological Description	4—Channel 5—Longitudinal Bar 6—Floodplain				0—No mineralization 1—Mineralization

In the study, the vectorized feature maps of sedimentary facies, sand thickness, sand rate, and grey sand rate and borehole data were converted into rasterized grids with an equal resolution of 200 m. Each grid cell extracted from the four feature maps represents 1 of 100,388 records from the dataset of the study area. Spatial analysis tools in QGIS were used to assist in this process. The resulting dataset for machine learning consisted of four attributes derived from the feature maps, assigned with a corresponding label from borehole data based on the geographical location, as shown in Figure 2.

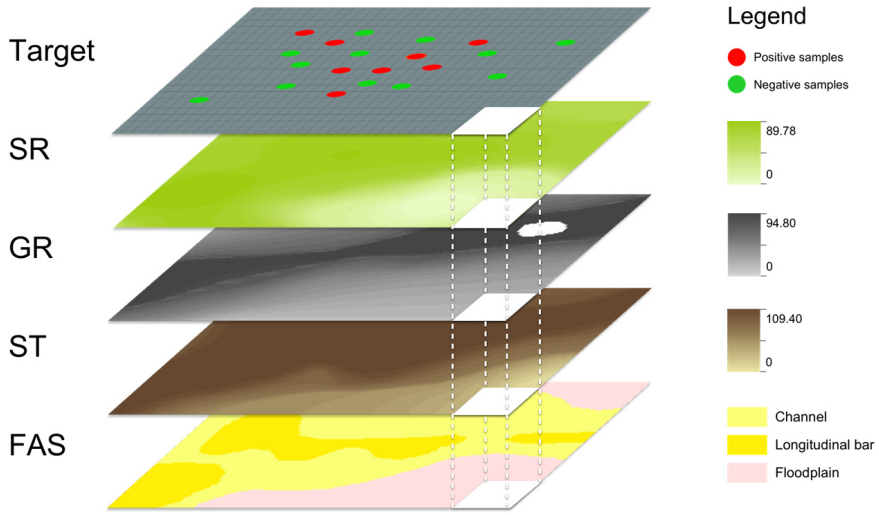

Figure 2. Preprocessing diagram of geodata for ML.

3.2. Explorational Data Analysis

Exploratory data analysis (EDA) is a data analysis approach and methodology used to explore the internal structure and patterns of data through various technical means, primarily data visualization [43]. EDA can be used to acquire geoscience insights by analyzing data. It aids in comprehending the relationships between geological features in both green-

fields (undeveloped land) and brownfields (previously developed land that may require redevelopment). It involves consulting experts to acquire experiences and knowledge, as well as extracting important features, detecting outliers, testing basic assumptions, and establishing preliminary models by using scientific data insights.

The Pearson product-moment correlation coefficient (r) is commonly used for exploratory analysis of data to measure the strength and direction of the linear correlation between two features [44]. In this study, the correlation between different geological features in the context of uranium mineralization, as shown in Figure 3, can be summarized as follows: This relationship was analyzed using statistical methods to understand the connection between geological features and their potential impact on uranium mineralization. There was a strong positive correlation between SR and ST. This indicates that as the sand rate increases, the sand thickness also increases, potentially favoring uranium mineralization. There was also a strong positive correlation between GR and ST. There was a moderate positive correlation between SR and GR. This implies that as the sand rate increases, there is a moderate increase in the grey sand rate, which may be related to uranium mineralization.

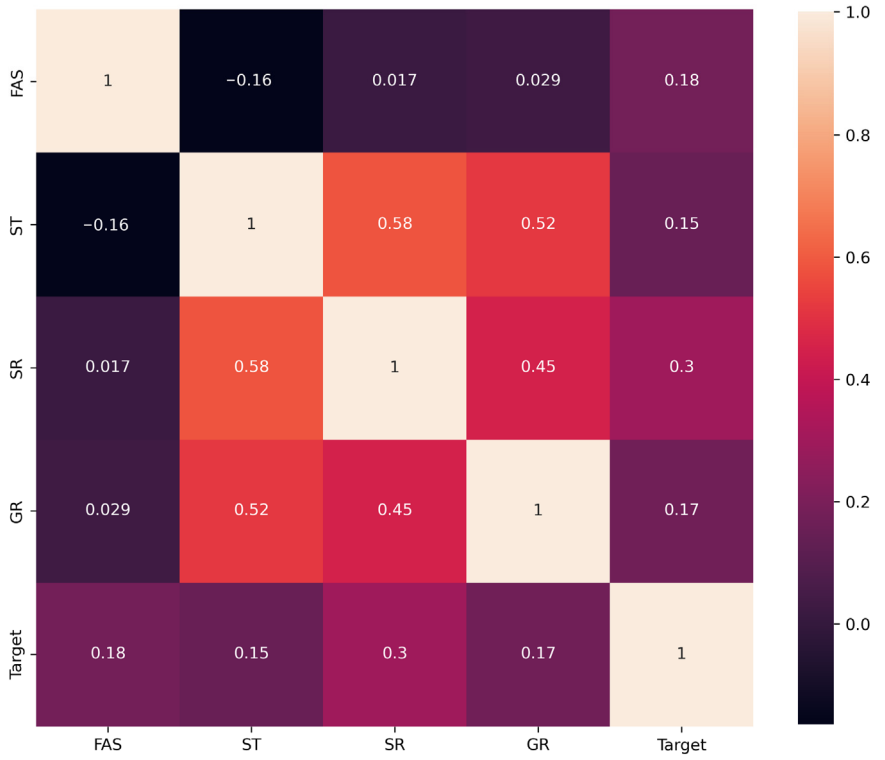

Figure 3. Pearson's correlation coefficient (r) of features.

Based on the geological study of the area, the uranium mineralization in the Saihan upper group is primarily of the paleochannel type. It is mainly concentrated in the longitudinal bar and sand within the channel during the I and II cycles of the Saihan upper group [45]. According to the correlation analysis between FAS and Target [46], the face of the longitudinal bar shows a higher number of mineralized samples compared to the other facies. This suggests that the face of a longitudinal bar is more conducive to mineralization (Figure 4). And in order to investigate the relationship between mineralization and sedimentary facies, we proceed to the target plot for the face of the longitudinal bar [47] in Figure 5. Out of the total 970 samples, 638 samples were in the face of the longitudinal bar. The average target value of mineralization in this area was 57.4%. This indicates that, on average, there was a 57.4% probability of finding uranium mineralization in the face of the longitudinal bar in this study.

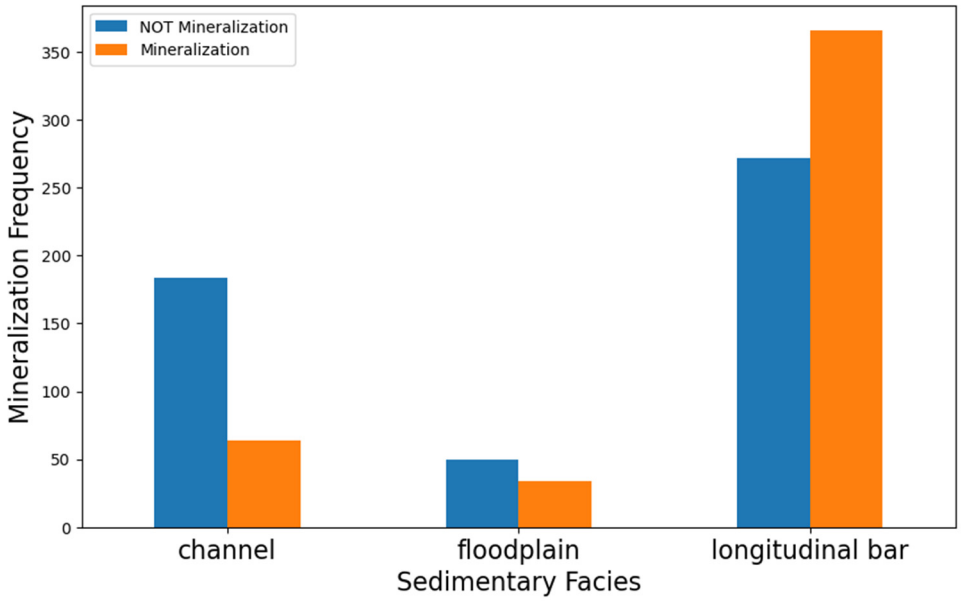

Figure 4. Mineralization frequency for sedimentary facies.

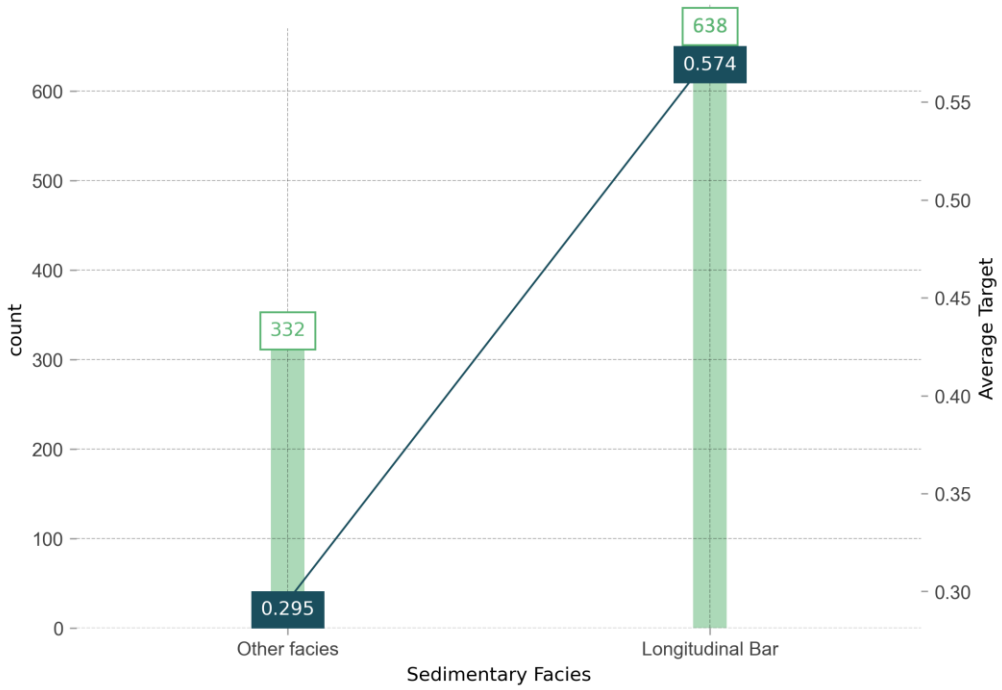

Figure 5. Target plot for longitudinal bar.

The violin plot is a type of data visualization technique created by integrating the box plot and kernel density plot. It is commonly used to represent continuous data, such as sand rate, sand thickness, and grey sand rate in the study, to provide a compact and attractive visualization of the data [48]. In a violin plot, the central part represents the box plot, showing the median and interquartile range of the data statistically. The width of this central part indicates the density or frequency of the data at different values.

The violin plot of sand rate shows a relatively even distribution for non-mineralized samples. Mineralized samples, on the other hand, have fewer occurrences of low values and relatively more occurrences of high values. The distribution of mineralized samples is most concentrated around the median of 70%, and 50% of the data fall between 62% and 78%, as shown in Figure 6. Similarly, in the violin plot of grey sand rate in Figure 7,

the mineralized samples are most concentrated around a median of 60%. However, in the violin plot of sand thickness, both non-mineralized and mineralized samples have a similar distribution pattern, indicating that this feature does not have a significant impact on mineralization (Figure 8).

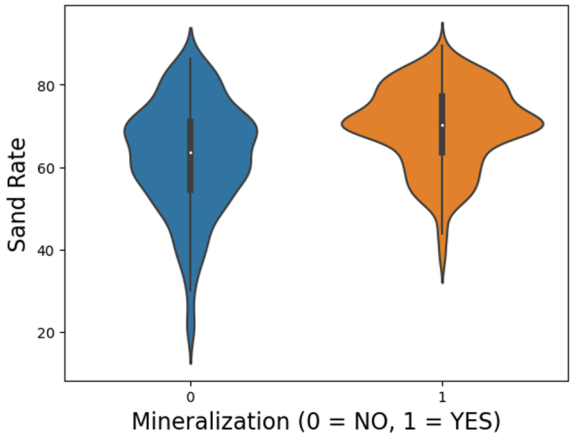

Figure 6. Violin plot of sand rate.

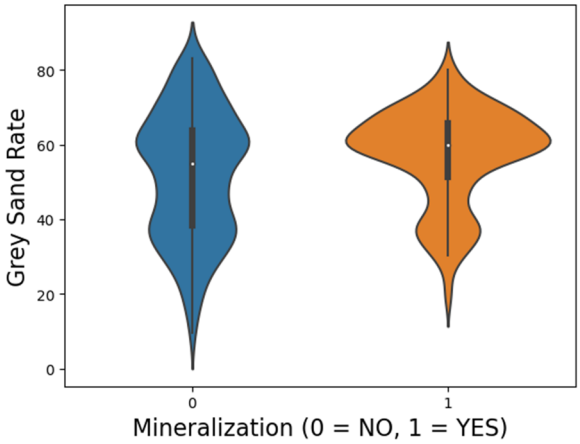

Figure 7. Violin plot of grey sand rate.

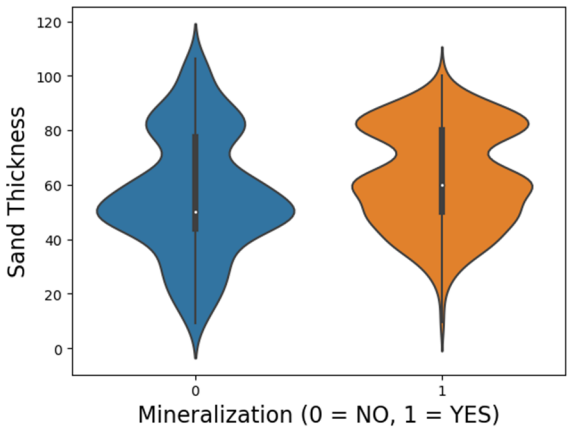

Figure 8. Violin plot of sand thickness.

3.3. Random Forest Classification

The "No Free Lunch" theorem suggests that there is no universally optimal algorithm for machine learning [49]. The performance of an algorithm depends on the specific

problem at hand. Different algorithms may perform well on certain problems but poorly on others [50]. Therefore, it is important to carefully choose an algorithm that is well suited to the classification problem within MPM in order to achieve optimal results.

Random forest is a popular choice for mineral prospectivity mapping due to its excellent performance and effectiveness in handling imbalanced data; it is an ensemble method that combines multiple decision trees to make predictions [51]. After the ML dataset was randomly split into 70% for training and 30% for testing, which is based on hold-out methodology, a Python function named *RandomForestClassifier* in the scikit-learn ensemble module was used to implement machine learning in this study. It creates a collection of decision trees, and each decision tree is built using a random subset of the training data. This randomness helps to reduce overfitting and improve the model's generalization ability. During training, each decision tree tries to find the best splits in the data based on the features and their values. The splitting process continues recursively until a criterion is met, which is the GINI index. The optimal partition is established when the smallest Gini index is calculated, and the calculation formula is shown in Equation (1). Once all of the decision trees are trained, each tree votes for a class, and the class with the most votes becomes the final prediction.

$$\text{Gini\_index}(\mathrm{D}, a) = \sum_{v=1}^{V} \frac{|\mathrm{D}^v|}{|\mathrm{D}|} \text{Gini}(\mathrm{D}^v) \quad (1)$$

where D is the dataset, a is the feature, and V is the number of all of the values of a.

3.4. Hyperparameter Tuning

In random forest, each decision tree is trained using a bootstrap sample of the original dataset, which means that some samples are left out "out-of-bag" (OOB). It helps to introduce randomness and diversity in the training process, which improves the overall performance and robustness of the random forest model [52]. As a result of bootstrapping, approximately 36.8% of the dataset is not used as the training dataset for each tree in the random forest, and the formula is as shown in Equation (2). The OOB estimate as the indicator of evaluation is then calculated by averaging the prediction errors of all of the out-of-bag samples. This estimate provides an unbiased evaluation of the model's performance because it is based on samples that were not used in the training process.

$$\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n \rightarrow \frac{1}{e} \approx 0.368 \quad (2)$$

Hyperparameter tuning is a necessary step to optimize the hyperparameters of the random forest algorithm to improve the model's performance. There are several methods available for hyperparameter tuning, and one commonly used method is *GridSearchCV* from the scikit-learn module [53]. In the study, *GridSearchCV* traverses all specified parameter values and evaluates the model's performance using 10-fold cross-validation. It then determines the optimal values for the hyperparameters based on the score of the OOB estimate, and the hyperparameters include the total number of trees, maximum depth of trees, and minimum number of samples required to split a node. The RF classification model was established subsequently based on parameter optimization. Table 3 lists the adjusted parameters.

3.5. Model Evaluation

3.5.1. Confusion Matrix

The performance of a random forest classification model is evaluated by using a confusion matrix based on the testing dataset, which consists of 30% of the entire ML dataset and is not used in the training process. It provides a tabular representation of the model's predictions compared to the actual values. In the confusion matrix in Figure 9, the model correctly predicted 131 instances as negative (true negatives) and 126 instances as positive (true positives). However, it incorrectly predicted 10 instances as positive

when they were actually negative (false positives) and 24 instances as negative when they were positive (false negatives). The confusion matrix allows for the calculation of various evaluation metrics in classification tasks, such as accuracy, precision, recall, and F1 score. These metrics are commonly used together to evaluate the performance of a classification model and provide insights into its effectiveness in correctly classifying instances (Table 4). Statistically speaking, accuracy provides an overall assessment of the model's performance but may not be suitable for imbalanced datasets. Precision indicates how well the model identifies true positives and avoids false positives. Recall indicates how well the model captures all positive instances and avoids false negatives. The F1 score indicates the harmonic mean of precision and recall, which provides a balanced measure of the model's performance. Macro average is calculated by taking the average of the metric calculated for each class individually (not mineralization and mineralization). It treats each class equally and does not consider class imbalance. On the contrary, weighted average considers class imbalance and gives more weight to classes with more samples. The two metrics are very much the same due to the roughly balanced set of positive and negative datasets.

Table 3. Adjusted parameters of RF.

Parameters	Description	Optimized Value
n_estimators	The number of trees in the forest	108
criterion	The function to measure the quality of a split	"gini"
max_depth	The maximum depth of the tree	81
min_samples_split	The minimum number of samples required to split an internal node	2
max_features	The number of features to consider when looking for the best split	1
min_samples_leaf	The minimum number of samples required to be at a leaf node	1
oob_score	Whether to use out-of-bag samples to estimate the generalization score	True

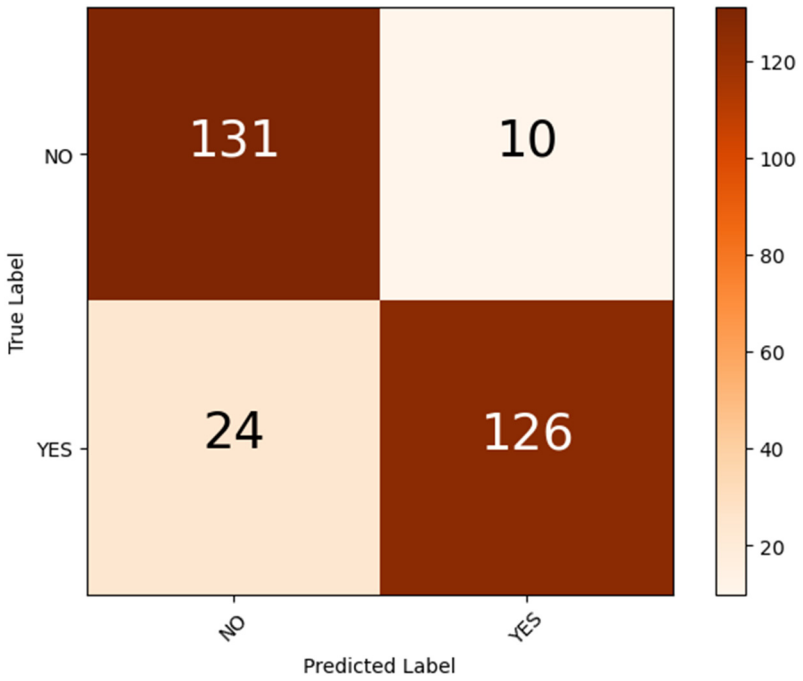

Figure 9. Confusion matrix of test datasets.

Table 4. Confusion matrix report.

	Precision	Recall	F1 Score	Support
Not mineralization	0.85	0.93	0.89	141
Mineralization	0.93	0.84	0.88	150
Accuracy			0.88	291
Macro average	0.89	0.88	0.88	291
Weighted average	0.89	0.88	0.88	291

3.5.2. Receiver Operating Characteristic Curve

The receiver operating characteristic (ROC) curve is a graphical representation of the performance of the random forest classification model at different thresholds. It is created by plotting the *true positive rate* (sensitivity) against the *false positive rate* (1—specificity) at various threshold settings [54,55]. The closer the ROC curve is to the top-left corner of the plot, the better the model's performance. The ROC curve can also provide an appropriate classification threshold that balances the trade-off between the *true positive rate* and *false positive rate*, depending on the relative importance of false positives and false negatives in the study of MPM application.

The area under the ROC curve (AUC) is a commonly used metric to represent the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance by the model. And the calculation formula is as shown in Equation (3). An AUC of 1 indicates a perfect classifier, and an AUC of 0.5 suggests a random classifier [56]. The implicit goal of AUC is to deal with situations where one has a skewed sample distribution and does not want to overfit to a single class, and AUC is a better measure than accuracy based on formal definitions of discriminancy and consistency [57]. In this study, the AUC of the random forest classification model was found to be 0.92, as shown in Figure 10. This indicates that the model has good discriminatory power and performs well in distinguishing between positive and negative instances.

$$\mathrm{AUC} = \frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i)\cdot(y_i + y_{i+1}) \tag{3}$$

where x and y are the coordinates and m is the max value of x.

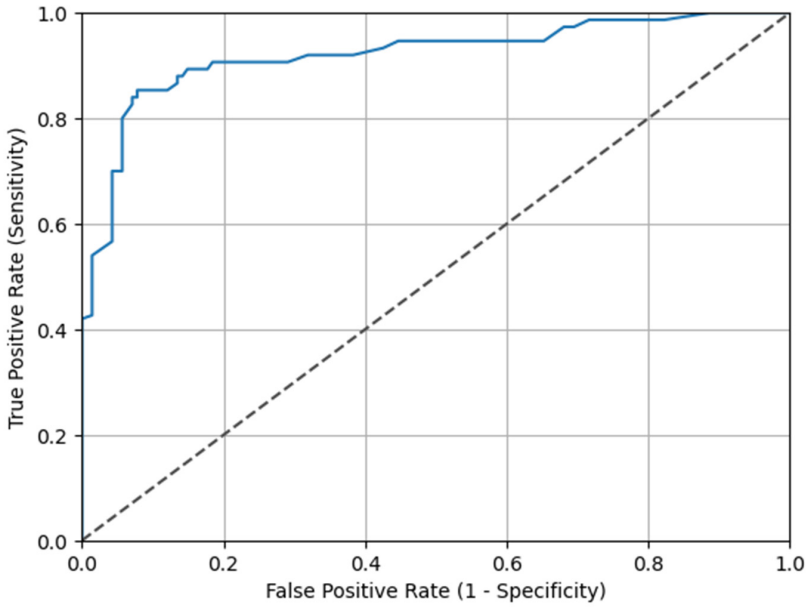

Figure 10. Receiver operating characteristic curve.

3.6. Model Explainability

There is no mathematical definition of explainability, which is the degree to which people are able to understand the reasons for a decision and are able to predict the outcome of a model consistently. It is important to have insights into how a model arrives at its predictions, especially in domains where interpretability is crucial, such as mineral prospectivity mapping. There are various techniques available for model explainability to make the model's behavior transparent, including simulatability, decomposability, and algorithmic transparency [58]. The more explainable an ML model, the easier it is for people to understand why certain decisions or predictions are made. If the decisions of one model are easier to understand than those of another, then it is more explanatory than another.

3.6.1. Features Importance

The *permutation_importance* function in the scikit-learn module is used to measure the importance of features. It calculates the decrease in model performance when the values of a particular feature are randomly shuffled. The larger the decrease in performance, the more important the feature is considered to be [59]. This information can be used for feature selection or understanding the underlying geological relationships in the dataset. The permutation importance ranks the results with the most important feature at the top for this study in Table 5. The first number in each row indicates how much the performance of the model has decayed, and the number with ± indicates the standard deviation.

Table 5. Permutation importance of features.

Serial Number	Weight Value	Feature
1	0.1527 ± 0.0421	Sand Rate
2	0.0182 ± 0.0381	Grey Sand Rate
3	0.0109 ± 0.0388	Longitudinal Bar
4	0.0018 ± 0.0073	Floodplain
5	−0.0109 ± 0.0178	Channel
6	−0.0291 ± 0.0313	Sand Thickness

3.6.2. Partial Dependence Plots

A partial dependence plot (PDP) shows the relationship between a feature and the model's predictions while holding other features constant. Comparing the findings from the PDP with metallogenic knowledge helps in understanding how the model's predictions change with variations in a specific feature [60].

The PDP of the sand rate exhibits a positive trend, indicating that as the sand rate increases above 70%, the prediction probability also significantly improves (Figure 11). The sand thickness variable has a negative effect on the prediction probability within specific ranges. Specifically, the negative effect is observed when the sand thickness is above 80 m in Figure 12. The grey sand rate has a significant impact on the prediction probability within certain ranges, indicating that higher values within the range of 45% to 75% approximately lead to improved prediction results, while values exceeding 75% lead to decreased prediction probability, as shown in Figure 13.

The relationship between sand rate, sand thickness, grey sand rate, and a target variable is probably not independent and has interactions. The PDP of multiple feature interactions on the target variable is different from the features' individual effects [61]. The 2D PDP diagram provides insights into the relationship between the input variables (sand rate, grey sand rate, and sand thickness) and the predicted outcome. Based on the diagram in Figures 14 and 15, the prediction is optimal when the sand rate is above approximately 70%, the grey sand rate is within the range of 54% to 68%, and the sand thickness is within the range of 60 m to 70 m.

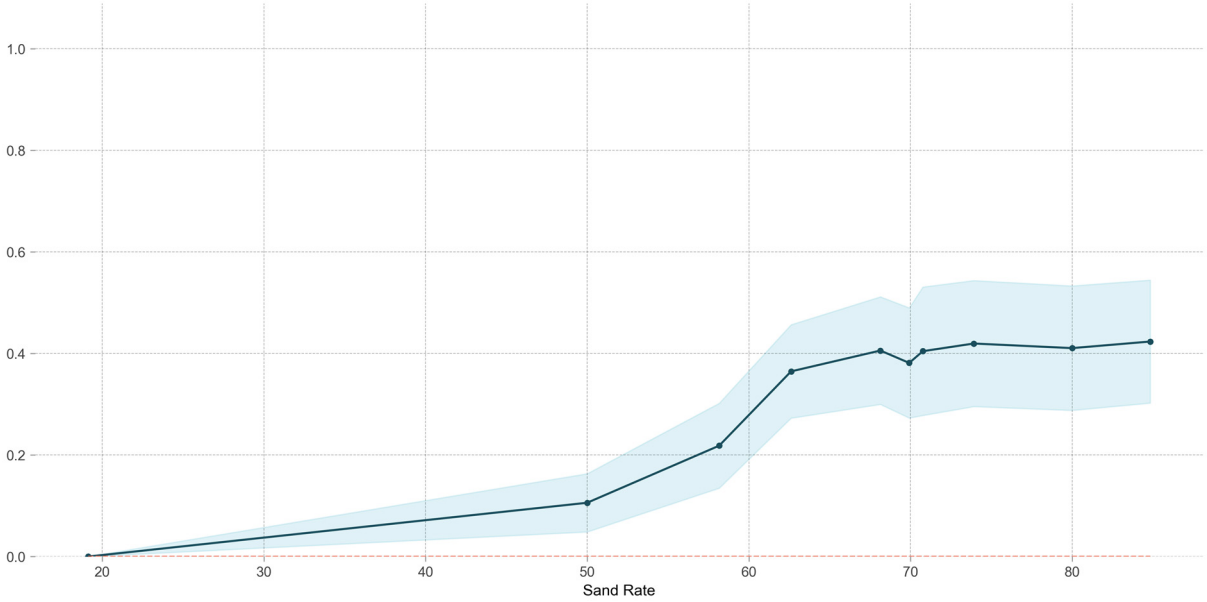

Figure 11. PDP for feature sand rate.

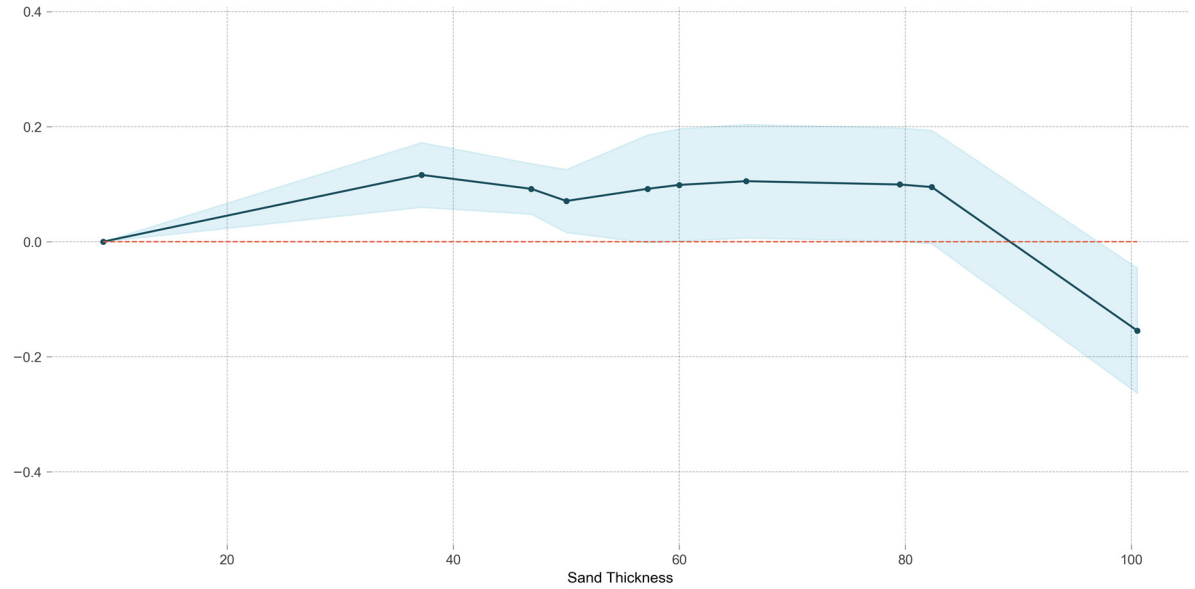

Figure 12. PDP for feature sand thickness.

3.6.3. Shapley Additive Explanations

Shapley additive explanations (SHAP) is a powerful method for interpreting model predictions and understanding the contribution of each feature to the model's output [62]. It is a post hoc method and can provide insights into the decision-making process of machine learning models and help in understanding the underlying mechanisms [63]. The core idea of SHAP is to calculate the marginal contribution of features to the model output and explain the predictions from both global and local perspectives. This method does not depend on the structure of any machine learning model and can consider the synergies between features.

The feature density shows a relationship between the feature value and SHAP value for each feature in Figures 16 and 17. The feature in each row is sorted in order of importance, and the wider the distribution, the greater the influence of the feature. Red represents a data point with a higher value of the feature, and blue represents a data point with a lower

value of the feature; the more right the point, the higher the positive effect of this feature on the prediction of mineralization. In this instance, the impact of SR was big and wide; meanwhile, the face of the floodplain had an impact on a small subset.

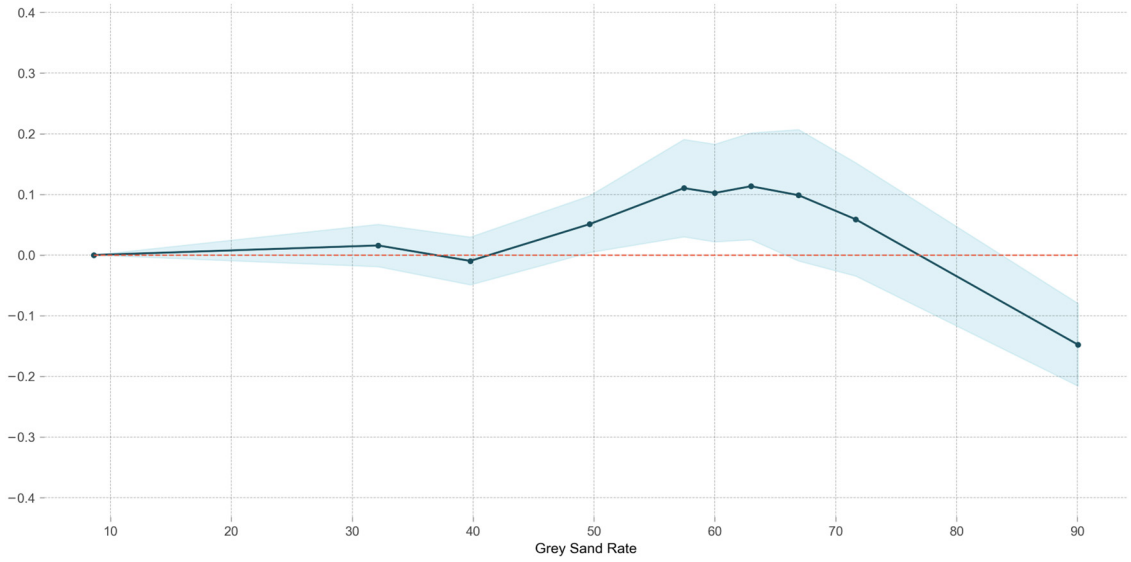

Figure 13. PDP for feature grey sand rate.

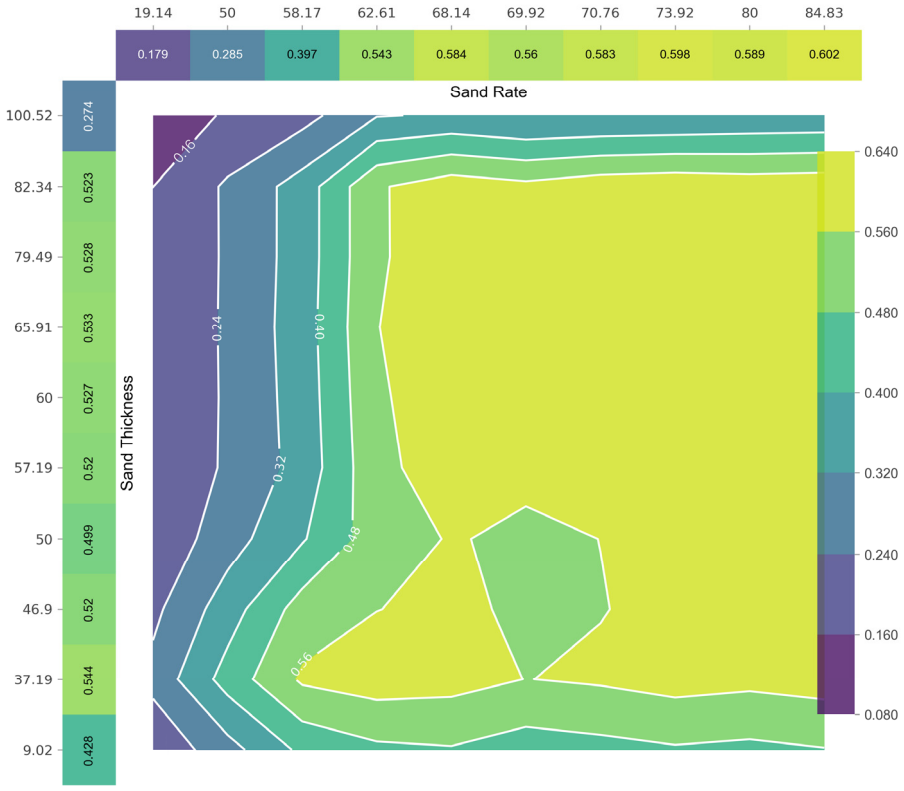

Figure 14. PDP interaction plot of sand rate and sand thickness.

A SHAP decision diagram shows how the model makes decisions, which is easy to interpret and helps easily identify the magnitude and direction of the primary impact. A sample was selected from the validation dataset to demonstrate the process from the base value to the final score of the model at the top of the diagram step by step, as shown in Figure 18.

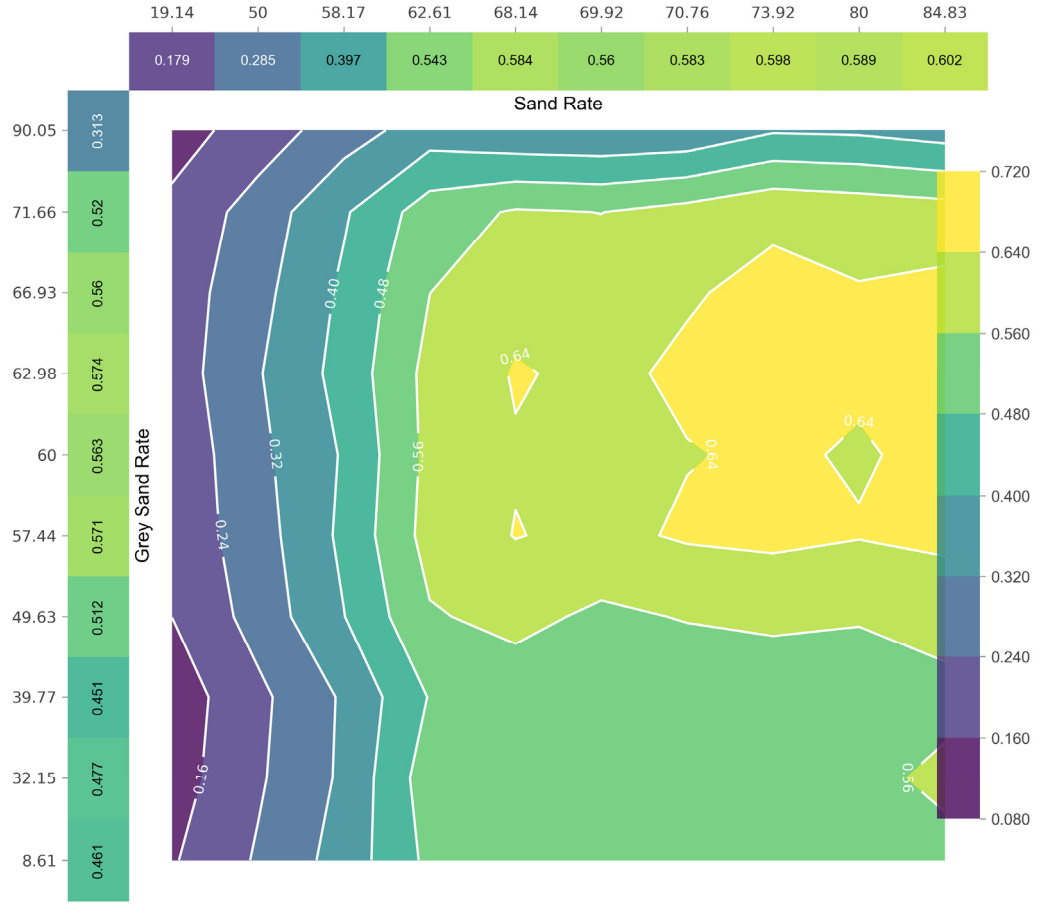

Figure 15. PDP interaction plot of sand rate and grey sand rate.

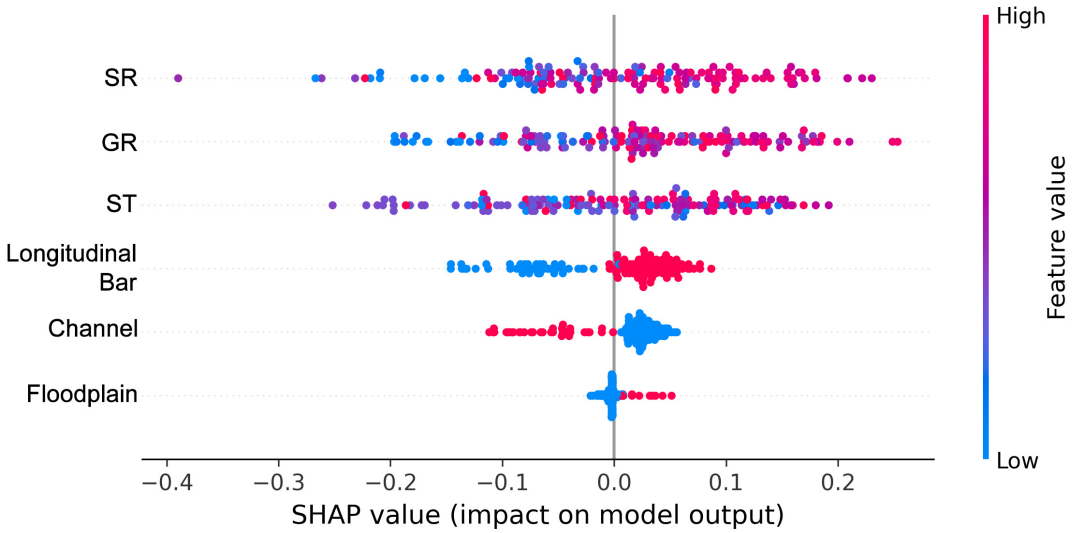

Figure 16. Feature density scatterplot.

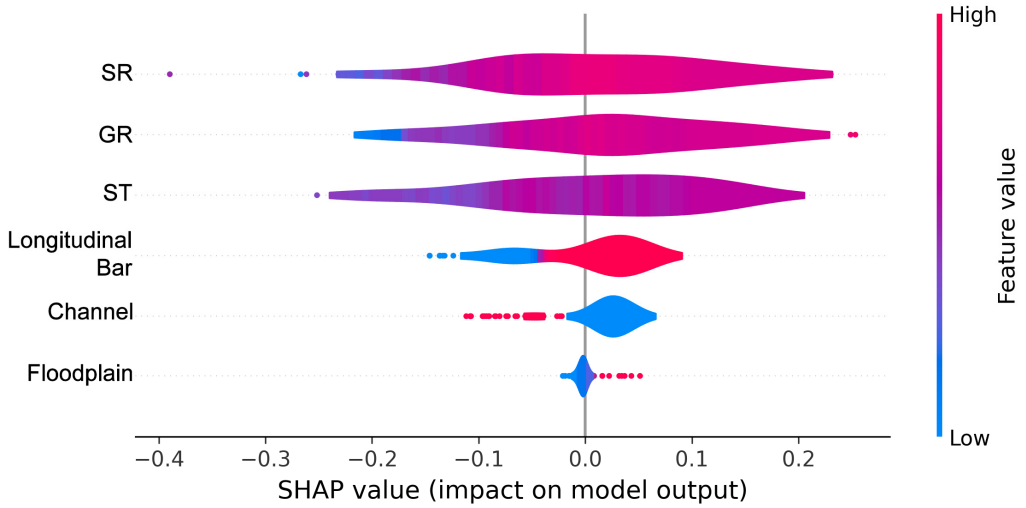

Figure 17. Feature density violin plot.

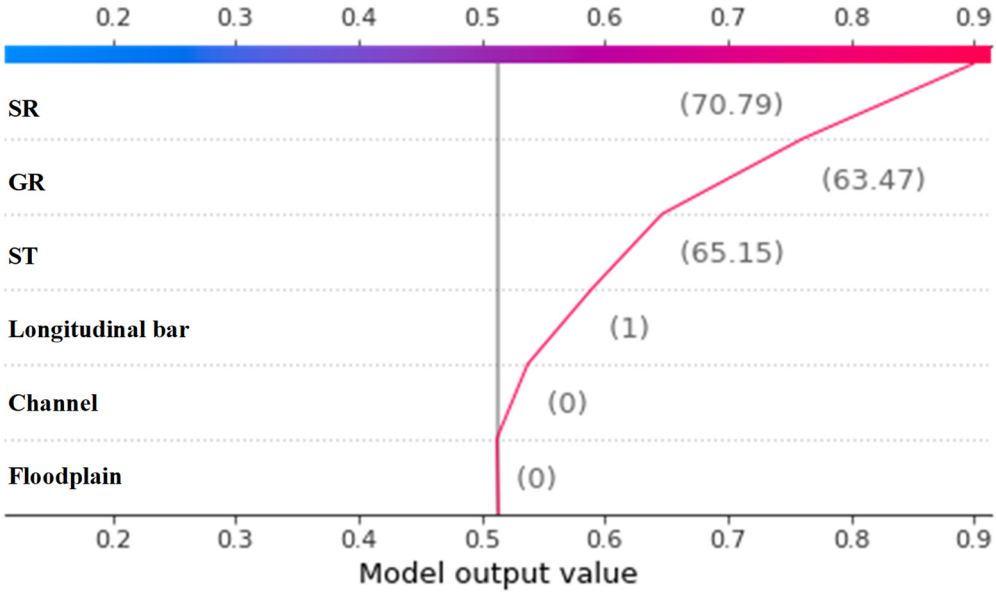

Figure 18. Decision plot of a sample.

4. Results and Discussion

4.1. MPM Results

The application of the random forest algorithm in the mineral prospectivity mapping (MPM) of regional uranium mineralization has been successfully completed. The model utilized 100,388 grid values as input and generated probabilities of uranium mineralization potentials ranging from 0 to 1. The visualization of these probabilities in QGIS demonstrated that areas with probabilities above 50% should be considered potential uranium mineralization zones based on a roughly balanced set of positive and negative datasets [64]. As the probability increased, the potential areas became more concentrated and smaller in size, as shown in Table 6.

From the visualization of MPM results in Figure 19, it can be observed that there is a relationship between the feature characteristic distribution and the prediction results for sedimentary facies, sand thickness, sand rate, and grey sand rate in the Saihan upper group formation. It is evident that the prediction results are mainly located in the channel face of the sedimentary facies and are closely related to the spatial distribution of the longitudinal

bar face. Most of the prediction results are found in the transitional zone between the channel face and the longitudinal bar face.

Table 6. Uranium mineralization potentials report.

Serial Number	Probability	Number of Grids	Percentage of Grids
1	>0.5	3958	3.94%
2	>0.6	2077	2.07%
3	>0.7	1004	1.00%
4	>0.8	353	0.35%
5	>0.9	77	0.08%

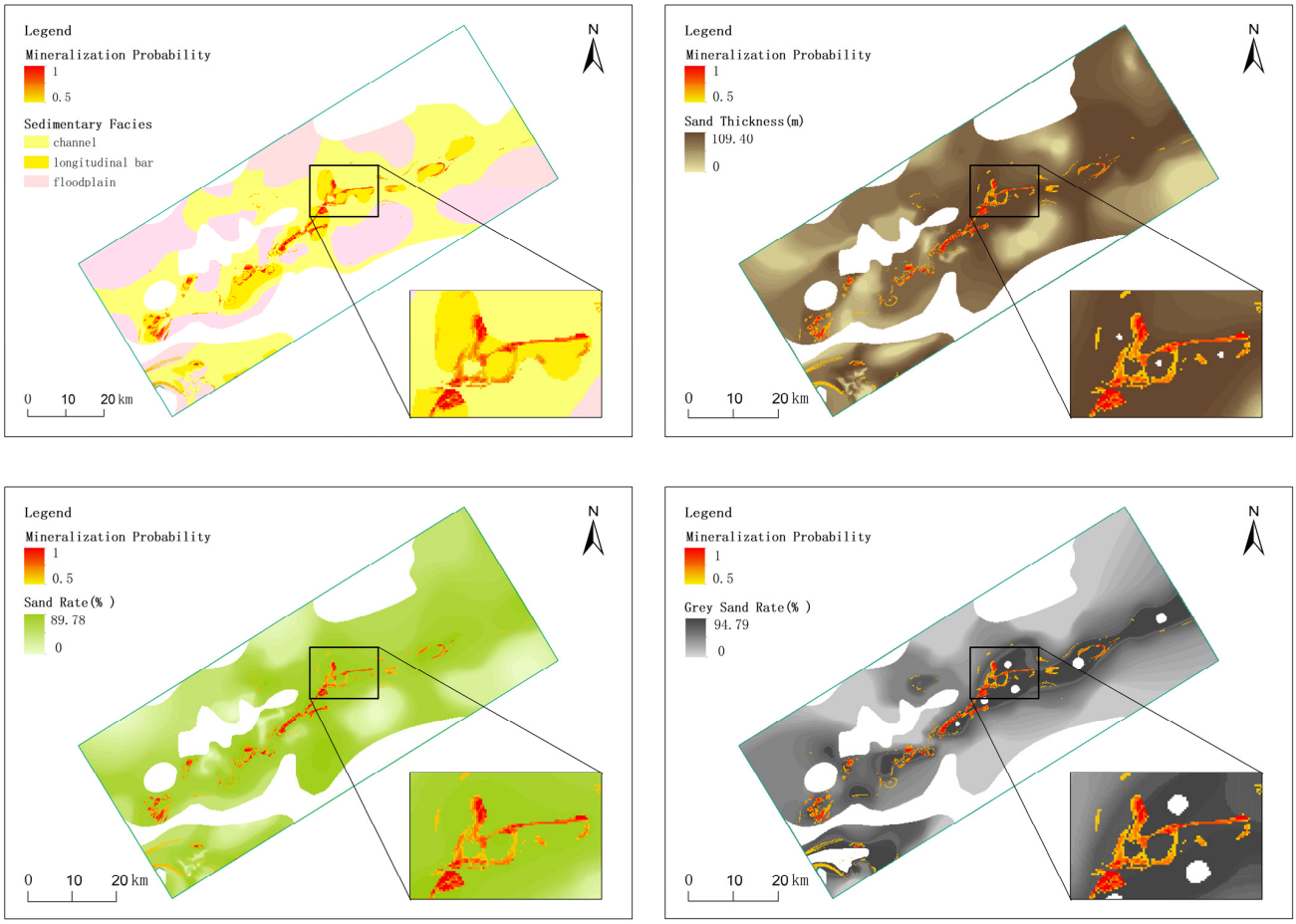

Figure 19. Superposition of prediction result and each feature layer (stratigraphic projection of Saihan upper group).

Moreover, the areas with small sand thickness are not favorable for mineralization. Based on the variation in sand thickness, it can be inferred that the water flow direction is from southwest to northeast along the paleochannels and the thickness of the sand body gradually increases due to sedimentation. The prediction results do not show a clear indication of sand thickness in the entire study area, but, in a small area, there is good agreement between the high-value area of sand thickness and the prediction results. So, it can be concluded that when the sedimentation reaches a greater thickness in a small area, it is favorable for mineralization. Additionally, according to the diffusion theory, in the downstream direction, the uranium-bearing material will decrease as it is transported and deposited, hence requiring a larger sand thickness to achieve mineralization compared to

areas with lower sand thickness upstream. The same applies to the sand rate and grey sand rate; the quantity of sand bodies has an impact on the mineralization potential.

4.2. Discussion

The ML-based MPM approach was applied to the Uranium Metallogenic zone in the Manite depression of the Erlian basin, China. The research utilized the random forest (RF) algorithm and data science tools to gain geological insights through explorational data analysis, ML model evaluation, and an explainability study. The findings indicate that features such as sand rate, grey sand rate, and the sedimentary face of the longitudinal bar have a positive impact on sandstone-hosted uranium mineralization. The MPM results were effectively visualized in QGIS using hierarchical probabilities, providing valuable information for future exploration and mining activities in the study area.

The input dataset's uncertainty can originate from various sources, including the inherent variability and complexity of real-world data, data collection issues, and human errors during data entry. In this article, the ML dataset was generated through data processing by geologists. Some data are obtained from geological logging, such as boreholes, while others require geological understanding and inference, which introduces uncertainty. This uncertainty can greatly affect the performance and reliability of machine learning models. To mitigate this uncertainty, several approaches can be employed, including data cleaning, feature engineering, data augmentation, ensemble methods, and robust machine learning techniques [65].

RF is a classifier that utilizes ensemble learning and is composed of multiple decision trees. During the training process, each decision tree in the random forest is constructed by randomly selecting a subset of the training samples through a process called bootstrapping. This means that each decision tree is trained on a fraction of the total samples. However, if the dataset has a severe class imbalance, where the number of positive and negative samples varies greatly, it can potentially affect the classification results. In such cases, the random forest may exhibit a bias toward classifying samples as negative and may overlook the positive samples due to the larger number of negative samples. Consequently, this can lead to a decrease in the accuracy of the classifier, particularly when dealing with a small number of sample categories. In this article, the positive and negative sample sets had a balanced ratio of approximately 1:1, and the potential impact of varying sample sizes was considered to be negligible.

5. Conclusions

The study utilized machine learning (ML) methodology, specifically the random forest algorithm, for mineral prospectivity mapping (MPM) in the Manite depression of the Erlian Basin. The data on sedimentary facies, sand thickness, sand rate, and grey sand rate of the Saihan upper group formation and borehole were used to establish the ML dataset. The random forest algorithm was used in the machine learning process after the hyperparameters were optimized and tuned. The ML model achieved an accuracy of 92% in uranium mineralization prediction in the study area, and it was evaluated using the confusion matrix and receiver operating characteristic curve. And feature importance, PDP analysis, and SHAP were put into practice to verify the reliability of the explainable artificial intelligence (XAI) in the domain of uranium MPM. The MPM results, visualized in QGIS, provided valuable geological insights and identified prospective targets for further uranium exploration.

This ML-based approach can assist uranium geologists in making informed decisions regarding the locations of future mineral exploration activities. However, ML algorithms cannot be fully understood due to human limitations, commercial barriers, data wildness, and algorithmic complexity [66–68]. The internals of machine learning are often considered opaque models, also known as a black-box, leading to distrust from geologists. The primary focus of this research was on post hoc interpretability, which plays a crucial role in XAI. The main objective was to address the issue of the black-box effect caused by opaque models,

which can be perplexing for readers. To illustrate the concept of explainability, the random forest algorithm was deliberately chosen as an example due to its strong interpretability and ease of comprehension. In the next stage of research on XAI for MPM of uranium resources, other algorithms such as support vector machines and neural networks will be explored to further enhance the overall XAI research technology system. Fortunately, there are existing studies that offer valuable insights and can serve as a source of inspiration for future research endeavors [69,70].

To address this issue, it is crucial to develop algorithms that prioritize the needs of humans and incorporate model explainability from the start. And also, post hoc interpretability algorithms offer a solution to understanding and explaining the decision-making process and predictions of machine learning models. These algorithms utilize various explainable approaches to provide insights into how the model generates predictions based on input data. This enables us to interpret the results, identify any biases or uncertainties in the model, and enhance its credibility and reliability. Ultimately, combining human-centered algorithms and post hoc interpretation empowers us to make informed decisions and applications by gaining a better understanding of the model's outputs.

Author Contributions: Conceptualization, W.K.; Methodology, W.K.; Software, W.K.; Data curation, P.Z.; Writing—original draft, W.K.; Writing—review & editing, J.C. and P.Z.; Supervision, J.C. and P.Z.; Project administration, J.C.; Funding acquisition, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology of the People's Republic of China [2023YFC2906404-01] and the APC was funded by the project of Quantitative Evaluation of Strategic Mineral Resource Potential in Key Metallogenic Zones (2023YFC2906404-01). The Quantitative Evaluation of Strategic Mineral Resource Potential in Key Metallogenic Zones [2023YFC2906404-01] is a project of Jianping Chen, which is acquired from Ministry of Science and Technology of the People's Republic of China.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Beijing Research Institute of Uranium Geology (BRIUG) and are available from the authors with the permission of BRIUG.

Acknowledgments: We are grateful for the editor and the constructive comments and suggestions from the peer reviewers.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. International Atomic Energy Agency. *World Uranium Geology, Exploration, Resources and Production*; Non-Serial Publications; IAEA: Vienna, Austria, 2020.
2. Cheng, M. Ideas and Methods for Mineral Resources Integrated Prediction in Covered Areas. *Earth Sci.* **2012**, *37*, 1109–1125.
3. Xiao, K.; Li, N.; Sun, L.; Zou, W.; Li, Y. Large scale 3D mineral prediction methods and channels based on 3D information technology. *J. Geol.* **2012**, *36*, 229–236.
4. Xiao, K.; Li, N.; Porwal, A.; Holden, E.-J.; Bagas, L.; Lu, Y. GIS-based 3D prospectivity mapping: A case study of Jiama copper-polymetallic deposit in Tibet, China. *Ore Geol. Rev.* **2015**, *71*, 611–632. [CrossRef]
5. Xiao, K.; Sun, L.; Zhu, Y. Theory of Mineral Resource Assessment. *Geol. Rev.* **2016**, *62*, 63–64.
6. Chen, J.; Yu, P.; Shi, R.; Yu, M.; Zhang, S. Research on three-dimensional quantitative prediction and evaluation methods of regional concealed ore bodies. *Earth Sci. Front.* **2014**, *21*, 211–220.
7. Wang, Y.; Chen, J.; Jia, D. Three-Dimensional Mineral Potential Mapping for Reducing Multiplicity and Uncertainty: Kaerqueka Polymetallic Deposit, QingHai Province, China. *Nat. Resour. Res.* **2019**, *29*, 365–393. [CrossRef]
8. Xiang, J.; Xiao, K.; Chen, J.; Li, S. 3D Metallogenic Prediction Based on Metallogenic System Analysis: A Case Study in the Lala Copper Mine of Sichuan Province. *Acta Geosci. Sin.* **2020**, *41*, 135–143.
9. Zuo, R.; Kreuzer, O.P.; Wang, J.; Xiong, Y.; Zhang, Z.; Wang, Z. Uncertainties in GIS-Based Mineral Prospectivity Mapping: Key Types, Potential Impacts and Possible Solutions. *Nat. Resour. Res.* **2021**, *30*, 3059–3079. [CrossRef]
10. Kong, W.; Xiao, K.; Chen, J.; Sun, L.; Li, N. A combined prediction method for reducing prediction uncertainty in the quantitative mineral resources prediction. *Earth Sci. Front.* **2021**, *28*, 128–138.
11. Morris, R.G.M. D.O. Hebb: The Organization of Behavior, Wiley: New York; 1949. *Brain Res Bull.* **1999**, *50*, 437. [CrossRef]

12. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408. [CrossRef]
13. Linnainmaa, S. The Representation of the Cumulative Rounding Error of an Algorithm as a Taylor Expansion of the Local Rounding Errors. Master's Thesis, University of Helsinki, Helsinki, Finland, 1970.
14. Werbos, P.J. Applications of advances in nonlinear sensitivity analysis. In *System Modeling and Optimization*; Drenick, R.F., Kozin, F., Eds.; Lecture Notes in Control and Information Sciences; Springer: Berlin/Heidelberg, Germany, 1982; Volume 38, pp. 762–770.
15. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. *Learning Internal Representations by Error Propagation*; Defense Technical Information Center: Fort Belvoir, VA, USA, 1985.
16. Hecht-Nielsen, R. Theory of the backpropagation neural network. In Proceedings of the International 1989 Joint Conference on Neural Networks, Washington, DC, USA, 18–22 June 1989; pp. 593–605.
17. Quinlan, J. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
18. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
19. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
20. Geoffrey, E.; Simon, O.; Yee-Whye, T. A fast-learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554.
21. Bengio, Y.; Lamblin, P.; Popovici, D.; Larochelle, H. Greedy Layer-Wise Training of Deep Networks. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference, Proceedings of the 19th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006*; MIT Press: Cambridge, MA, USA, 2007.
22. Marc, A.; Christopher, P.; Sumit, C.; Yang, L. Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference, Proceedings of the 19th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006*; MIT Press: Cambridge, MA, USA, 2007.
23. Yann, L.; Lenon, B.; Yoshua, B.; Patrick, H. Gradient-based learning applied to document recognition. *Proc. IEEE* **1989**, *86*, 2278–2324.
24. Yann, L.; Yoshua, B. Convolutional Networks for Images, Speech, and Time Series. In *The Handbook of Brain Theory and Neural Networks*; MIT Press: Cambridge, MA, USA, 1995; pp. 255–258.
25. Harris, D.P. *Mineral Resources Appraisal: Mineral Endowment, Resources, and Potential Supply: Concepts, Methods and Cases*; Oxford University Press: New York, NY, USA, 1984.
26. Agterberg, F.P.; Bonham-Carter, G.F. Measuring the performance of mineral-potential maps. *Nat. Resour. Res.* **2005**, *14*, 1–7. [CrossRef]
27. Li, S.; Chen, J.; Liu, C.; Wang, Y. Mineral Prospectivity Prediction via Convolutional Neural Networks Based on Geological Big Data. *J. Earth Sci.* **2021**, *32*, 327–347. [CrossRef]
28. Zhang, Q.; Chen, J.; Xu, H.; Jia, Y.; Chen, X.; Jia, Z.; Liu, H. Three-Dimensional Mineral Prospectivity Mapping by XGBoost Modeling: A Case Study of the Lannigou Gold Deposit, China. *Nat. Resour. Res.* **2022**, *31*, 1135–1156. [CrossRef]
29. Li, X.; Lin, W.; Guan, B. The impact of computing and machine learning on complex problem-solving. *Eng. Rep.* **2023**, *5*, e12702. [CrossRef]
30. Anmol, A.; Ananya, A. Machine learning models trained on synthetic datasets of multiple sample sizes for the use of predicting blood pressure from clinical data in a national dataset. *PLoS ONE* **2023**, *18*, e0283094.
31. Li, C.; Xiao, K.; Li, N.; Song, X.; Xhang, S.; Wang, K.; Chu, W.; Cao, R. A comparative study of support vector machine, random forest and artificial neural network machine learning algorithms in geochemical anomaly information extraction. *Acta Geosci. Sin.* **2020**, *41*, 309–319.
32. Hong, J.; Gan, C.; Liu, J. Prediction of REEs in OIB by major elements based on machine learning. *Earth Sci. Front.* **2019**, *26*, 45–54.
33. Jung, D.; Choi, Y. Systematic review of machine learning applications in mining: Exploration, exploitation, and reclamation. *Minerals* **2021**, *11*, 148. [CrossRef]
34. Jooshaki, M.; Nad, A.; Michaux, S. A systematic review on the application of machine learning in exploiting mineralogical data in mining and mineral industry. *Minerals* **2021**, *11*, 816. [CrossRef]
35. Guo, Z.; Ytai, Y.; Ren, W.; Qang, Z.; Feng, Z.; Chen, L.; Tanf, Z. Emplacement and episodic denudation of basement granites from the southern Jiergalangtu Sag, Erlian Basin and its tectonic implications. *Earth Sci. Front.* **2023**, *30*, 259–271.
36. Qi, J.; Zhao, X.; Li, X.; Yang, M.; Xiao, Y.; Yu, F.; Dong, Y. The distribution of Early Cretaceous faulted sags and their relationship with basement structure within Erlian Basin. *Earth Sci. Front.* **2015**, *22*, 118–128.
37. Han, X.; Wu, Z.; Lin, Z.; Jiang, J.; Hu, H.; Yin, D.; Qiao, H.; Li, Z. Constraints of Sedimentary Facies of the Targeting Layers on Sandstone-type Uranium Mineralization in Major Uranium-producing Basins in Northern China: A Brief Discussion. *Geotecton. Metallog.* **2020**, *44*, 697–709.
38. Kang, S.; Yang, J.; Liu, W.; Zhao, X.; Qiao, P.; Du, P.; Lv, Y. Mineralization Characteristics and Potential of Paleo-Valley Type Uranium Deposit in Central Erlian Basin, Inner Mongolia. *Uranium Geol.* **2017**, *33*, 206–214.
39. Chen, Y.; Zhu, Y. *Metallogenic Model of Chinese Ore Deposits*; Geological Press: Beijing, China, 1993.
40. Shi, J.; Tang, J.; Zhou, P.; Jin, Q.; Yang, Z.; Zhu, L.; Jin, X. A discussion on the exploration model. *Geol. Bull. China* **2011**, *30*, 1119–1125.
41. Li, S.; Chen, J.; Xiang, J. Classification and visualization of geoscience text big data based on convolutional neural network: A case study of Lala copper mine in Sichuan. In Proceedings of the 2018 Annual Meeting of Chinese Geoscience Union, Beijing, China, 19–23 October 2018.

42. Li, S.; Chen, J.; Xiang, J.; Zhang, Z.; Zhang, Y. Two-dimensional prospecting prediction based on AlexNet network: A case study of sedimentary Mn deposits in Songtao-Huayuan area. *Geol. Bull. China* **2019**, *38*, 2022–2032.
43. Pieter, V. Exploratory Analysis of Provenance Data Using R and the Provenance Package. *Minerals* **2023**, *13*, 375.
44. Xu, N.; Huang, B.; Li, Q.; Zhu, W.; Wang, Z.; Wang, R. Towards the study on the geochemistry through machine learning. *J. China Coal Soc.* **2022**, *47*, 1895–1907.
45. Liu, W.; Kuang, S.; Jiang, L.; Shi, Q.; Peng, C. Characteristics of Paleo-valley Sandstone-type Uranium Mineralization in the Middle of Erlian Basin. *Uranium Geol.* **2013**, *29*, 328–335.
46. Luo, S.; Luo, B. *Pandas Data Analysis Quickly Starts with 500 Moves*; Tsinghua University Press: Beijing, China, 2023.
47. Zhang, Z. Spatial and Temporal Characteristics of Air Quality and Its Influence Factors in Wuhan. Master's Thesis, Wuhan University of Science and Technology, Wuhan, China, 2020.
48. Cui, Z. On the Cover: Violin Plot. *Educ. Meas. Issues Pract.* **2020**, *39*, 7. [CrossRef]
49. Mohsen, R.; Matthias, G. No-Free-Lunch Theorems for Reliability Analysis. *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng.* **2023**, *9*, 04023019.
50. Zhou, Z. *Machine Learning*; Tsinghua University Press: Beijing, China, 2016.
51. Zhang, S.; Carranza, E.J.M.; Xiao, K.; Wei, H.; Yang, F.; Chen, Z.; Li, N.; Xiang, J. Mineral Prospectivity Mapping based on Isolation Forest and Random Forest: Implication for the Existence of Spatial Signature of Mineralization in Outliers. *Nat. Resour. Res.* **2022**, *31*, 1981–1999. [CrossRef]
52. Zhou, Z.; Li, N. *Ensemble Methods: Foundations and Algorithms*; Publishing House of Electronics Industry: Beijing, China, 2020.
53. Suroor, N.; Jaiswal, A.; Sachdeva, N. Stack Ensemble Oriented Parkinson Disease Prediction Using Machine Learning Approaches Utilizing GridSearchCV-Based Hyper Parameter Tuning. *Crit. Rev. Biomed. Eng.* **2022**, *50*, 39–58. [CrossRef] [PubMed]
54. Nykänen, V.; Lahti, I.; Niiranen, T.; Korhonen, K. Receiver operating characteristics (ROC) as validation tool for prospectivity models—A magmatic Ni–Cu case study from the Central Lapland Greenstone Belt, Northern Finland. *Ore Geol. Rev.* **2015**, *71*, 853–860. [CrossRef]
55. Sun, T.; Li, H.; Wu, K.; Chen, F.; Zhu, Z.; Hu, Z. Data-Driven Predictive Modelling of Mineral Prospectivity Using Machine Learning and Deep Learning Methods: A Case Study from Southern Jiangxi Province, China. *Minerals* **2020**, *10*, 102. [CrossRef]
56. Althouse, A.D. Statistical graphics in action: Making better sense of the ROC curve. *Int. J. Cardiol.* **2016**, *215*, 9–10. [CrossRef]
57. Ling, C.X.; Huang, J.; Zhang, H. AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. In *Advances in Artificial Intelligence. Canadian AI 2003, Proceedings of the Canadian AI 2003, Halifax, NS, Canada, 11–13 June 2003*; Xiang, Y., Chaib-draa, B., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2003; Volume 2671.
58. Lipton, Z.C. The Mythos of Model Interpretability. *arXiv* **2016**, arXiv:1606.03490.
59. Li, X.-M.; Zhang, Y.-X.; Li, Z.-K.; Zhao, X.-F.; Zuo, R.-G.; Fan, X.; Yi, Z. Discrimination of Pb-Zn deposit types using sphalerite geochemistry: New insights from machine learning algorithm. *Geosci. Front.* **2023**, *14*, 200–219. [CrossRef]
60. Zhang, S.; Xiao, K. Random forest-based mineralization prediction of the Lala-type Cu deposit in the Huili area, Sichuan Province. *Geol. Explor.* **2020**, *56*, 239–252.
61. Sun, D.; Chen, D.; Mi, C.; Chen, X.; Mi, S.; Li, X. Evaluation of landslide susceptibility in the gentle hill-valley areas based on the interpretable random forest-recursive feature elimination model. *J. Geomech.* **2023**, *29*, 202–219.
62. Ancona, M.; Öztireli, C.; Gross, M. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In Proceedings of the International Conference on Machine Learning 2019, Long Beach, CA, USA, 10–15 June 2019; pp. 272–281.
63. Luo, Z.; Zuo, R.; Xiong, Y.; Zhou, B. Metallogenic-Factor Variational Autoencoder for Geochemical Anomaly Detection by Ad-Hoc and Post-Hoc Interpretability Algorithms. *Nat. Resour. Res.* **2023**, *32*, 835–853. [CrossRef]
64. Liu, Z. Towards Versatile Class-Imbalanced Learning: Algorithm, Application, and Software Library. Master's Thesis, Jilin University, Changchun, China, 2022.
65. Burton, S.; Herd, B. Addressing uncertainty in the safety assurance of machine-learning. *Front. Comput. Sci.* **2023**, *5*, 1132580. [CrossRef]
66. Fan, F.; Wang, G. Learning from pseudo-randomness with an artificial neural network—Does God play pseudo-dice? *IEEE Access* **2018**, *6*, 22987–22992. [CrossRef]
67. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
68. van der Mass, H.L.; Verschure, P.F.M.J.; Molenaar, P.C.M. A note on chaotic behavior in simple neural networks. *Neural Netw.* **1990**, *3*, 119–122. [CrossRef]
69. Wang, Z.J.; Turko, R.; Shaikh, O.; Park, H.; Das, N.; Hohman, F.; Kahng, M.; Chau, D.H.P. CNN explainer: Learning convolutional neural networks with interactive visualization. *IEEE Trans. Vis. Comput. Graph.* **2020**, *27*, 1396–1406. [CrossRef] [PubMed]
70. Cantürk, S.; Singh, A.; St-Amant, P.; Behrmann, J. Machine-learning driven drug repurposing for COVID-19. *arXiv* **2020**, arXiv:2006.14707.