Ore Geology Reviews 174 (2024) 106270

Contents lists available at ScienceDirect

Ore Geology Reviews

journal homepage: www.elsevier.com/locate/oregeorev

Review

Addressing imbalanced data for machine learning based mineral prospectivity mapping

Fahimeh Farahnakian [a,b], Javad Sheikh [a,*], Luca Zelioli [a], Dipak Nidhi [a], Iiro Seppä [a], Rami Ilo [a], Paavo Nevalainen [a], Jukka Heikkonen [a]

[a] University of Turku, Department of Computing, 20014, Finland
[b] Geological Survey of Finland, 02151, Finland

ARTICLE INFO

Keywords:
Mineral prospectivity mapping
Imbalanced data
Random forest
Multi-perceptron
Logistic regression
Decision tree
EIS toolkit

ABSTRACT

Effective Mineral Prospectivity Mapping (MPM) relies on the ability of Machine Learning (ML) models to extract meaningful patterns from geophysical data. However, in mineral exploration, identifying the presence of mineral deposits is often a rare event compared with the overall geological landscape. This rarity leads to a highly imbalanced dataset, where positive instances (mineralized samples) are considerably less frequent than negative instances (non-mineralized samples). Imbalanced data can potentially bias ML models towards the majority class, leading to inaccurate predictions for the minority class (mineralized samples) which are of primary interest. To address this challenge, we proposed two-level methods in this study. At the data level, we employed imbalanced data handling techniques that operate on the training dataset and change the class distribution. At the algorithmic level, we adjusted the decision threshold of a model to balance the trade-off between false positives and false negatives. Experimental results are collected on a geophysical data from Lapland, Finland. The dataset exhibits a significant class imbalance, comprising 17 positive samples contrasted with 1.84×10^6 negative samples. We investigate the effect of handling imbalanced data on the performance of four ML models including Multi-Layer Perceptron (MLP), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR). From the results, we found that the MLP model achieved the best overall performance, with total accuracy of 97.13% on balanced data using synthetic minority oversampling method. Random forest and DT also performed well, with accuracies of 88.34% and 89.35%, respectively. The implemented methodology of this work is integrated in QGIS as a new toolkit which is called EIS Toolkit[1] for MPM.

Contents

* Corresponding author.
E-mail address: javad.sheikh@utu.fi (J. Sheikh).

[1] https://github.com/GispoCoding/eis_toolkit

https://doi.org/10.1016/j.oregeorev.2024.106270
Received 31 March 2024; Received in revised form 7 August 2024; Accepted 5 October 2024
Available online 21 October 2024

1. Introduction

Mineral Prospectivity Mapping (MPM) aims to predict the likelihood of finding specific types of mineral deposits for mineral exploration, and resource assessment. Traditionally, MPM can be carried out by analyzing geophysical, geochemical, and spatial datasets to create maps highlighting areas with a higher potential for hosting particular mineral resources (Zuo, 2020; Chudasama et al., 2022). Over the past decade, Machine Learning (ML) methods have demonstrated their ability to improve the accuracy of MPM models, leading to more efficient and effective mineral exploration strategies (Brandmeier et al., 2019; Sun et al., 2020). In addition, they outperform the traditional statistical techniques and empirical explorative models in mineral prospectivity prediction, particularly when the input data is complex and its relationship with mineralization is nonlinear (Zhang et al., 2018, 2015; Carranza and Laborte, 2015). Some of the most commonly used ML methods include Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Boosting algorithms, and Artificial Neural Network (ANN) (Chudasama et al., 2022; Sun et al., 2020; Jung and Choi, 2021). However, the supervised ML algorithms require an appropriate and representative labeled dataset for training in order to identify patterns and relationships that can be used to predict the likelihood of mineralization efficiently.

Creating labeled data is usually expensive, resource-intensive, and time-consuming in numerous real-world applications, especially in domains like geophysics, geochemistry, and mineral exploration. In these fields, the problem is exacerbated by at least two major factors. Firstly, rock samples, usually gathered from boreholes, are the most precise geological data type, which allow for determining the exact chemical composition of the sample. However, drilling is prohibitively expensive, and can often only be used to confirm a potential mineral deposit. Secondly, mineralization is rare and happens in only spatially very constrained locations (Cheng, 2007). Therefore, these challenges lead to a highly imbalanced dataset characterized by a scarcity of known mineralized regions and a vast majority of non-mineralized locations. Scarcity of well-known locations makes it challenging to use deep machine learning approaches in these domains, as they tend to perform better with large amounts of data. In addition, training classical ML models on imbalanced datasets often tend to be biased towards the majority class, leading to poor generalization (Yadav and Bhole, 2020; Zou et al., 2016). As a result, models shows good accuracy on the majority class but poor accuracy on the minority class. However, high accuracy is crucial in MPM due to high costs associated with misclassifications, especially false positives (Xiong and Zuo, 2017).

The problem of imbalance has garnered significant attention in recent years (Yadav and Bhole, 2020; Spelmen and Porkodi, 2018; Prado et al., 2020; Ferreira da Silva et al., 2022). Generally, the imbalanced handling techniques can be divided into three main groups: (1) algorithm level, (2) data level and (3) hybrid methods (Yadav and Bhole, 2020). Data level methods aim to balance the data by adding more samples of the minority class such as over-sampling (Chawla et al., 2002) and under-sampling (Kotsiantis et al., 2006). For instance, in Prado et al. (2020) they used an over-sampling technique (SMOTE) to generate balanced data for modeling of Cu-Au prospectivity. Algorithm level approaches incorporate misclassification costs into the training process. It means higher costs are assigned to misclassifications of the minority class, reflecting the greater importance of correctly identifying these rare instances. Hybrid methods integrate data and algorithm level advantages.

Another approach for improving the ML model's performance of imbalanced datasets is thresholding (Buda et al., 2018). The model might consistently predict high probabilities for the majority class, even for borderline cases that should belong to the minority class. To address this problem, a decision threshold can be adjusted on the predicted probabilities of the minority class and classifying data points as belonging to the minority class only if the predicted probability exceeds the threshold. In this way, we can focus on identifying the minority class more accurately and control the balance between false positives and false negatives, thus potentially improving the model's performance on imbalanced datasets. In Xiong and Zuo (2018), authors proposed a rare event LR algorithm, which embeds sampling and decision threshold corrections into the original LR algorithm.

In this work, we address the mentioned challenges for developing ML models in a highly imbalanced distribution dataset in MPM. ML algorithms include Decision Tree (DT), RF, LR, and Multi-Layer Perceptron (MLP). The prediction results of models are ultimately determined according to prediction probabilities. In most classifiers, the default threshold is 0.5. However, this threshold does not work well for imbalanced classification prediction. For these reasons, we implement a nested cross-validation approach. Initially, we use Leave-One-Out Cross Validation (LOOCV) for reporting performance metrics. Within each LOOCV training fold, we further conduct a 4-fold Stratified Cross Validation (SCV) to optimize key parameters including oversampling, undersampling, decision threshold, and model hyperparameters. We chose a 4-fold approach as it provided the maximum fold size while maintaining a sufficient number of positive cases for effective evaluation, even with various levels of over and under sampling.

We evaluated the effectiveness of the imbalanced handling techniques on the performance of ML algorithms for MPM using a case study of mineralization in Lapland, Finland. In our highly imbalanced dataset, there are only 17 samples confirmed to contain mineralized samples, compared to approximately 1.84×10^6 samples that may or may not contain mineral deposits. For simplicity, we will use the term 'non-mineralized samples' throughout this paper.

The results show that balancing the data is the most effective for improving the performance of ML algorithms for MPM. We believe that our findings have important implications for the development of more accurate and efficient ML-based MPM models. When it comes to the quality of the positive samples, it seems that each and every sample is useful, and that for this particular problem more positive samples would be helpful. Based on our knowledge, there is no published research that uses data balancing techniques with decision threshold optimization for improving ML algorithms in MPM such as this study.

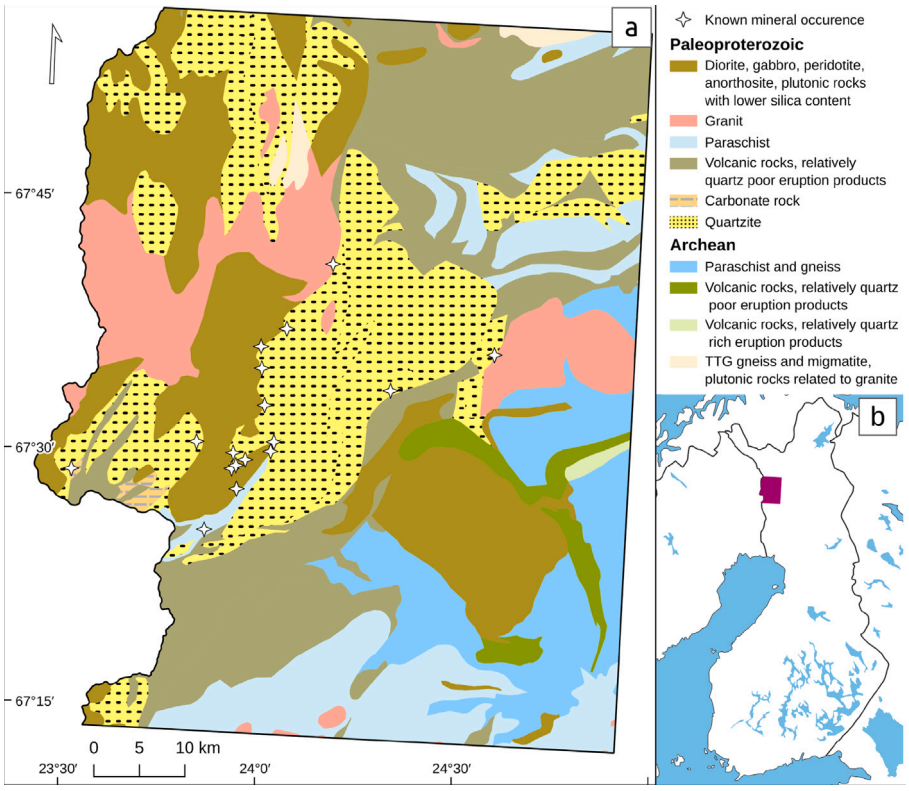

Fig. 1. (a) The generalized geology and location of the study area (b) The study area is located at western Finnish Lapland.

Table 1
Data Description.

Data Type	Used features	Count of Features
AEM	Aeroelectromagnetic Inphase and Quadrature components, Apparent resistivity, and electromagnetic ratio	4
Magnetic	DGRF65 Anomaly: Analytical Signal of magnetic anomaly; Directional cosines of magnetic anomaly in directions 45, 90, 135, and 180 degrees; X, Y, Z, Horizontal derivative of the tilt derivative and tilt derivative of the magnetic field, Pseudogravity derived from magnetism	12
Radiometric	Gamma ray radioactivity of elements K, U, and Th and total Gamma ray radioactivity	4
Total		20

Existing GIS tools often lack built-in ML models capable of handling imbalanced data specific to MPM. This work addresses this gap by introducing a novel methodology tailored for MPM tasks. The designed methodology is integrated into a popular GIS tool like QGIS as a part of a Horizon European project which is called Exploration Information System (EIS). However, this paper focuses on describing the methodology rather than the toolkit design.

The remainder of this paper is organized as follows. We describe the study areas and data in Section 2. The proposed methods for handling imbalanced data are described in Section 3.1. The evaluated ML methods are discussed in Section 3.2. Sections 4 and 5 provides the experimental design and results, respectively. The discussion and conclusions are drawn in Section 6.

2. Study area and dataset

Our study area is located in municipalities of Kittilä, Kolari, and Muonio, Lapland, Finland (Fig. 1). The study area contains 17 known Iron oxide–copper–gold (IOCG) deposits, which are mostly situated at faults that cross the boundary between synorogenic monzonite and its country rock (Niiranen, 2005). The occurrence data is from mineral deposit database of Geological Survey of Finland (GTK).[2]

Geophysical data serves as a valuable tool for MPM, aiding in the mapping of subsurface features and enhancing comprehension of the geological structure within a search area, particularly when surface outcrop is limited (Kreuzer et al., 2020). For this reason, we used different geophysical datasets used in the study which were all provided by the GTK.[3] A summary of the used data for this study is provided in Table 1. All the used raster data originate from aerogeophysical low altitude surveys conducted by GTK between 1972–2007. All rasters have 50 × 50 m^2 spatial resolution. Totally, we have three type of data as follow:

- Airborne Electromagnetic (AEM) includes measurements of the electrical conductivity of the earth's subsurface.[4] The predictor maps of these features are shown in Appendix A (Fig. A.11).

[2] https://tupa.gtk.fi/paikkatieto/meta/mineral_deposits.html
[3] https://www.gtk.fi/
[4] https://tupa.gtk.fi/paikkatieto/meta/aeroelectromagnetic_raster_data_of_finland.html

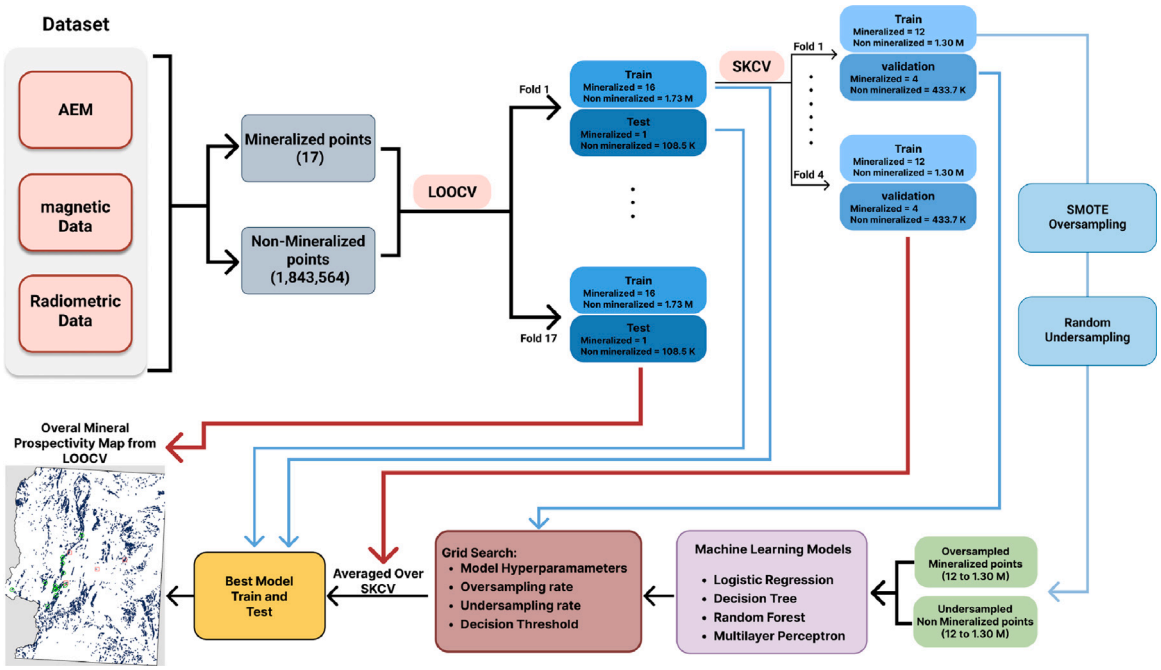

Fig. 2. Workflow.

- Magnetic data used to identify variations in the Earth's magnetic field caused by the magnetic properties of subsurface rocks.[5] The predictor maps of these features are shown in Appendix A (Fig. A.12, A.13 and A.14).
- Radiometric data gives the measurement of natural gamma radiation emitted from the earth's surface to infer the concentration of radio element isotopes such as uranium (U), thorium (Th), and potassium (K).[6] The predictor maps of these features are shown in Appendix A (Fig. A.15).

The final dataset is a two-class imbalanced dataset which contains 17 samples of known IOCG mineralized (positive) and 1.84×10^6 samples of non-mineralized (negative) samples, with each sample having 20 different features from the aerogeophysical data.

3. Methodology

The aim of our work is to improve the performance of the proposed ML techniques which are able to predict deposits and produce an accurate MPM based on a highly imbalanced dataset. The methodology implemented based on the following steps as shown in Fig. 2.

- **Data preprocessing:** The first step in the workflow involves preparing data for modeling. This includes cleaning and transforming the data to ensure it is clean, consistent, and suitable for ML algorithms such as removing missing values and normalization.
- **Cross-Validation Strategy:** We use LOOCV, where each fold corresponds to one deposit. Therefore, with 17 deposits in our dataset, the cross-validation naturally involves 17 folds. This approach ensures rigorous evaluation of our models by maximizing the use of available data while preventing data leakage between training and testing sets. Additionally, for each training dataset, we apply 4-fold SCV to assess the robustness of our models.
- **Imbalanced data handling:** Since the original dataset is highly imbalanced, where the minority class (mineralized regions) is significantly smaller than the majority class (non-mineralized regions), it is crucial to address this imbalance to prevent the model from overfitting to the majority class and neglecting the minority class. For this purpose, we utilized SMOTE for oversampling minority points and random undersampling techniques to create a balanced dataset, as detailed in Sub- Section 3.1.
- **ML modeling:** Based on the preprocessed and imbalanced-handled data, supervised ML models are trained to predict mineral occurrences (Sub- Section 3.2). We also performed hyperparameter tuning, decision threshold adjusting to optimize the performance of each algorithm.
- **Model evaluation and map generation:** The performance of each ML model is evaluated using appropriate metrics, such as accuracy, F1 score, sensitivity and specificity, to assess their ability to correctly identify mineral occurrences. In addition, we investigate the effect of the number of oversampled and undersampled data on ML's performance. To evaluate the model's performance in the context of mineral exploration, we used prediction–area and success-rate curves. Uncertainty estimation is shown using the accuracy-rejection curve to assess model performance with synthetic positive samples. Finally, classification maps are generated to classify each pixel as mineralized or non-mineralized.

3.1. Imbalanced data handling techniques

In this section, we describe the methods which is used in this paper for addressing the challenge of imbalanced data in MPM. The class imbalance happens when one class contains fewer samples (e.g. mineralized samples) than others (e.g. non-mineralized samples).

Random under-sampling (Kotsiantis and Pintelas, 2004) is a non-heuristic method that removes samples from the majority class randomly to balance class distribution. The major weakness of random under-sampling is that this method may eliminate potentially valuable data that could be crucial for the learning process (Kotsiantis et al., 2006). Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) is a common over-sampling methods in MPM (Prado et al.,

[5] https://tupa.gtk.fi/paikkatieto/meta/aeromagnetic_raster_data_of_finland.html

[6] https://tupa.gtk.fi/paikkatieto/meta/aeroradiometric_raster_data_of_finland.html

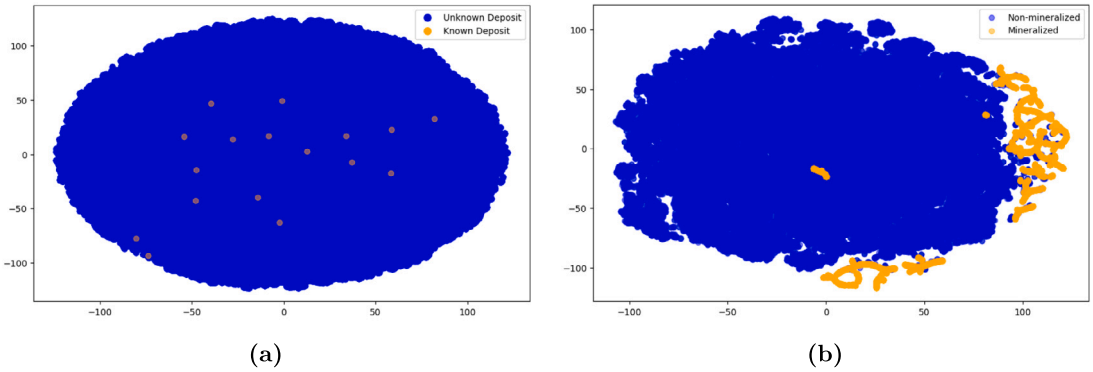

Fig. 3. Data distribution using t-SNE (a) original imbalanced dataset, (b) with 5000 over-sampled known mineralized samples and 100,000 under-sampled non-mineralized samples by applying SMOTE and random under-sampling techniques, respectively.

2020) which can create synthetic minority class data points by measuring the similarity between data points based on distance metrics. It first identifies k nearest neighbors for each minority class data point. Then, it randomly selects one of these neighbors and generates a new synthetic data point along the line segment connecting the minority class data point and its randomly selected neighbor. The new synthetic data point is assigned the minority class label. In this study, we employed a hybrid method that incorporates both over-sampling and under-sampling techniques to achieve a more balanced dataset. To visualize the distribution of data across mineralized and non-mineralized samples before and after applying balancing methods, we plot the dataset in reduced dimensions using t-SNE. Fig. 3(a) visualizes the original data distribution and Fig. 3(b) an example after applying SMOTE to generate balanced dataset. This visualization also demonstrates how the original dataset is highly imbalanced and then how it is balanced with the oversampled mineralized points and undersampled of non-mineralized points.

3.2. Machine learning methods

In the following subsections, we will give brief introductions to the machine learning methods used in this study.

3.2.1. Multi-layer perceptron (MLP)

MLP (Jiang et al., 2018) is a type of a feedforward artificial neural network with a layered structure. It consists of an input layer, one or more hidden layers, and an output layer. Each layer is made up of interconnected nodes, or neurons, which use nonlinear activation functions, that enable the network to learn complex patterns in the data. MLP's input layer receives the initial data, which then passes through the hidden layers, where most of the computation takes place. Each neuron in these layers processes the input by performing a weighted sum followed by a nonlinear transformation. The weights and biases in these neurons are adjusted during the training process using algorithms like backpropagation, which effectively tunes the network to produce the desired output.

One of the key strengths of MLPs is their ability to approximate virtually any continuous function, given sufficient neurons in the hidden layers. This attribute, known as universal approximation capability, makes MLPs highly versatile for a wide range of tasks, from simple regression problems to complex classification tasks to identify patterns in high-dimensional data. However, MLPs are prone to overfitting, especially when the network has too many layers or neurons relative to the amount of available training data. To counter this, techniques such as dropout and regularization are often used. The selection of the appropriate network architecture, including the number of layers and the number of neurons in each layer is also a possible pitfall of MLP, as they often require experimentation and validation against a held-out dataset to find the optimal architecture.

3.2.2. Random Forest (RF)

RF (Genuer et al., 2008) is an ensemble learning method known for its robustness and accuracy in various applications. RF can handle large datasets with high dimensionality, while providing estimates of feature importance. RF utilizes numerous decision trees during the training phase, each of which is independently developed, with the final output being determined by aggregating their individual predictions. In classification tasks, the aggregation is typically done by taking the mode of the classes predicted by each tree, ultimately allowing a 'majority vote' or the most prevalent result to decide the outcome. In regression tasks, RF predicts an outcome based on the mean of the predictions from all the trees, ensuring a balanced approach that mitigates individual tree biases. Most recent advances include methods in order to reduce overfitting and improve computational efficiency.

One of the benefits of RF is its good capability to process large datasets, even those with high dimensionality, making it a viable choice for scenarios where the data involves a vast number of features (Xu et al., 2012). Additionally, RF provides insightful estimates of feature importance, which is valuable for both predictive analytics, where the goal is to forecast future outcomes, and descriptive tasks, which aim to understand patterns within the data. By creating multiple trees and using their collective decision, RF ensures that the model is not overly complex and customized the specific details of the training data, thereby enhancing its generalization capabilities and reducing overfitting.

3.2.3. Decision tree (DT)

DT (Quinlan, 1986) is a non-linear predictive modeling tool organized as a tree, with nodes representing attribute tests, branching corresponding to test results, and leaf nodes containing the decision or conclusion. The root node represents the whole dataset, which is subsequently divided into subgroups depending on the attribute values that produce the greatest reduction in heterogeneity or impurity. This process, known as recursive partitioning, continues until each leaf node is pure (in classification) or additional splitting is neither conceivable nor practicable (in regression).

One of the most appealing aspects of DTs is their interpretability and simplicity. Unlike more complex models, DTs can be visualized and understood easily, making them an excellent tool for decision-making processes where the reasoning behind predictions needs to be explained. DTs can be sensitive to small changes in the data, leading to different splitting paths. This instability is often mitigated by ensemble methods like Random Forests, which incorporate multiple DTs to increase robustness and accuracy. Despite these challenges, the simplicity, interpretability, and versatility of DTs make them a valuable tool in the arsenal of machine learning methodologies.

Table 2
Main hyper-parameter of ML models and their best value which is selected by GFS method.

ML Model	Hyper-parameter	Search space	Optimal Hyper-parameter Value
MLP	solver	lbfgs, sgd, adam	adam
	hidden_layer_sizes	[(2), (4), (8)]	(16, 2)
	learning_rate_init	0.001, 0.005, 0.01, 0.05, 0.1, 0.2	0.01
RF	n_estimators	2, 4, 8, 16	8
	max_depth	10, 20, 30, 50	50
	min_samples_split	2, 5	2
	max_features	auto, sqrt	auto
DT	criterion	gini, entropy, log_loss	entropy
	splitter	best, random	random
	max_depth	2, 4, 6, 8	6
LR	penalty	l1,l2	l2
	Inverse of Regularization Strength	0.001, 0.01, 0.1, 1, 10, 100	100
	max_iter	1000, 2000, 3000	1000

3.2.4. Logistic regression (LR)

LR (Foster et al., 2018) is a statistical model used primarily for binary classification tasks, categorizing data points into one of the two groups based on the values of one or more independent variables. The categorization is achieved through a logistic function, a special S-shaped curve that transforms any input into a value between 0 and 1 (probability). One of the most important advantages of LR is its interpretability. Unlike some black box-like models where the decision-making can be ambiguous, LR provides clear insights into how each predictor influences the probability of the outcome. This transparency is a major reason why LR is so widely used in fields where understanding the impact of variables is crucial, such as in medicine, where it might be used to predict the likelihood of a patient having a particular disease based on symptoms and test results, or in social sciences to understand the factors influencing a particular social behavior.

LR is quite straightforward and less resource-intensive compared with more complex alternatives, making it a practical choice for both small and large-scale applications. Its method of handling probabilities is direct and uncomplicated, which simplifies the process of model training and prediction.

Regularization methods (Nusrat and Jang, 2018), such as L1 or L2 regularization, are one of the improvements that can incorporate to LR by helping to prevent overfitting, i.e. when a model performs well on training data but poorly on unseen data. By penalizing the magnitude of the coefficients, regularization ensures that the model remains in general domain and is resistant to the specifics of the training data, enhancing its performance and reliability on more complex or varied datasets.

4. Experimental setup

4.1. Model training

In order to evaluate the performance of ML models, improve model generalizability and select optimal hyperparameters, we first used LOOCV to divide the data into 17 folds. Each fold is used as a test set for a different model, trained on the remaining folds. The model performance metrics are estimated based on the test sets. Within each training set from the outer loop, a second layer of cross-validation (SCV) is conducted to optimize the model's hyperparameters. This includes tuning three specific parameters: the decision threshold, the optimal rates for oversampling and undersampling, and the ML models hyperparameters. The results of SCV are more reliable than the results of traditional cross-validation because they are less likely to be affected by data leakage (Wainer and Cawley, 2018)

In addition, during the inner loop, the data balancing methods are applied to the training data to generate different numbers of samples for the minority class. For each hyperparameter setting, the imbalanced method is used to balance the class distribution in the training set, and the resulting dataset is used to train the ML model. To find the best model and evaluate its performance in cross-validation (CV), we employed the geometric mean (G_{mean}) as a loss function. Unlike traditional classification loss functions, which primarily focus on minimizing misclassifications, G_{mean} considers the distance between sample features. By taking into account both sensitivity (recall) and specificity, G_{mean} provides a balanced evaluation metric that is robust to class imbalance (Hastie et al., 2009).

$$G_{mean} = \sqrt{Sensitivity \times Specificity} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

where TP, FP, TN and FN indicate the total number of true positive, false positive, true negative, and false negative pixels, respectively.

Table 2 lists the key ML parameters and their best values. The parameter names at the table align with the standard parameter names used in the Scikit-learn Python package.

4.2. Evaluation metrics

The common metrics for evaluating the trained ML models are assumed including classification accuracy, sensitivity, specificity, F1-score, ROC-AUC. The formula used are:

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \tag{4}$$

$$F1 - score = 2 \times \frac{precision \times Sensitivity}{precision + Sensitivity} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

In this work, we used a variant of F1-score to handle imbalanced dataset which is called weighted F1-score. The weighted F1-score computes the average F1-score weighted by the number of true instances for each class as follows:

$$\text{Weighted F1-score} = \frac{\sum_{i=1}^{n} w_i \times \text{F1-score}_i}{\sum_{i=1}^{n} w_i} \tag{7}$$

where n is the number of classes and w_i is the weight assigned to class i, typically equal to the proportion of true instances for class i in the dataset. F1-score$_i$ is the F1-score for class i.

5. Results and discussion

5.1. Comparison of ML methods

Table 3 presents the performance metrics from 4-fold cross validation on the test dataset for ML methods with and without an imbalanced data handling method. This result demonstrates the superiority of

Table 3
The classification performance metrics (%) from nested cross validation on the test dataset with and without imbalanced data handling.

Models	MLP		RF		DT		LR	
Imbalanced Data Handling	With	Without	With	Without	With	Without	With	Without
Accuracy	97.13	90.90	88.34	**99.99**	89.35	99.20	79.29	78.23
Weighted F1-Score	98.52	92.40	93.69	**99.99**	91.55	99.60	84.75	80.18
Sensitivity	47.05	35.29	**88.23**	0.00	58.82	41.17	82.35	58.82
Specificity	97.13	90.90	88.34	**99.99**	89.35	99.20	79.29	78.23

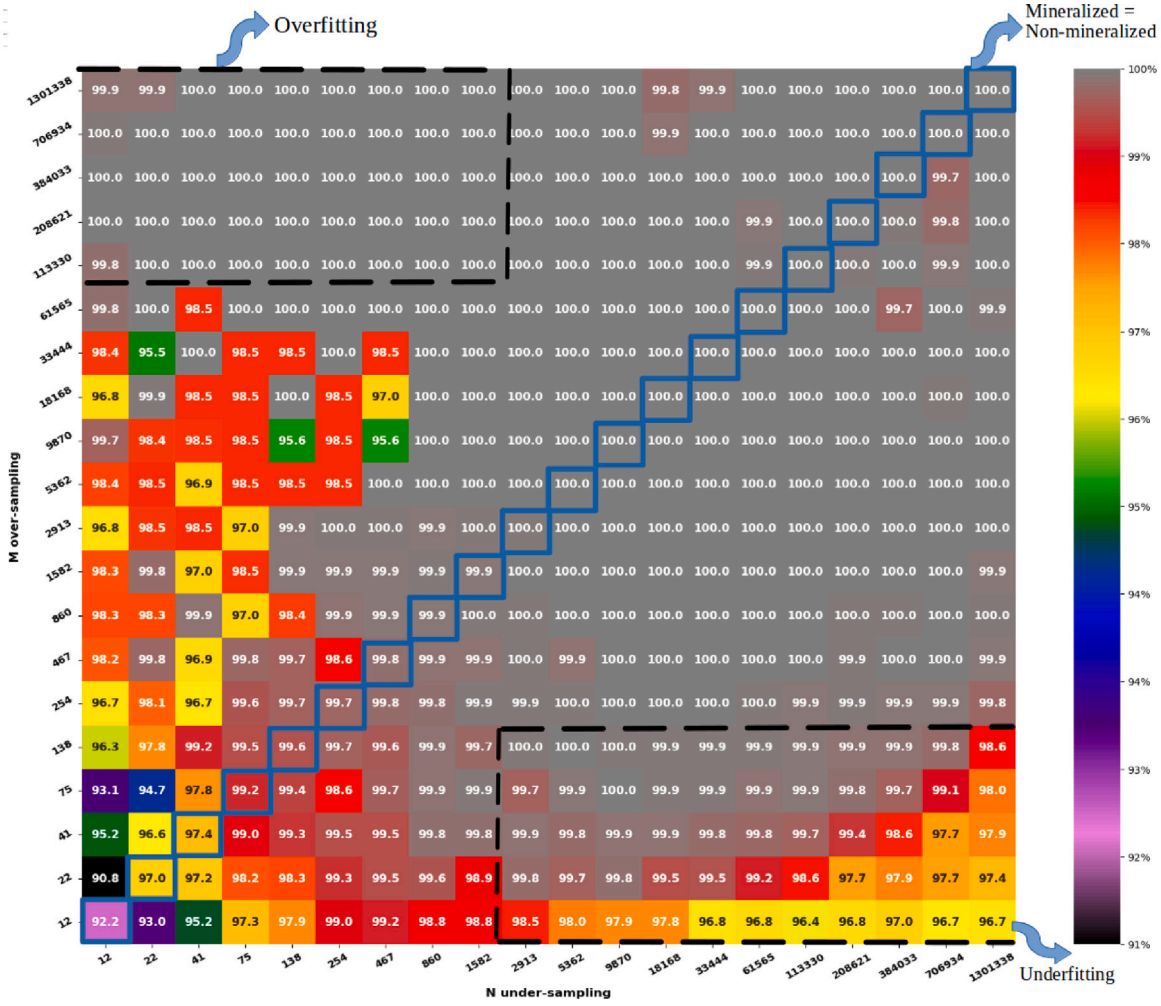

Fig. 4. The weighted F1-score (%) for the training dataset based on different number of over-sampling mineralized samples N and under-sampling non-mineralized samples M for the best model (MLP).

data augmentation-based framework. It is evident that in all models without imbalanced data handling, the models become biased towards the majority class. Our results show that models trained with augmented data outperform those trained without augmentation across all evaluated metrics. This highlights the effectiveness of our data augmentation-based framework.

DT and LR show better sensitivity, which suggests that simpler models have less capacity to become biased and overfitted on the majority class. The results of the models with imbalanced data handling indicate that the MLP model achieved the highest accuracy of 97.13%. Nevertheless, the accuracy of the RF model remains sufficiently high for classification purposes. Notably, its sensitivity is 41.17% points higher than that of the MLP model, making it significantly more sensitive in handling imbalanced data. Even with imbalanced data handling the MLP and DT got biased towards the majority class.

5.2. Effectiveness of data balancing techniques

We evaluate the proposed ML models based on the different number of over-sampling N and under-sampling M by using nested cross validation as these parameters are crucial for achieving a balanced representation of classes in order to avoid overfitting and underfitting. Generating too few samples may not effectively capture the underlying distribution of the minority class, while generating too many samples may lead to overfitting or poor generalization.

Considering the overall robustness and classification accuracy of the two ML models, MLP was chosen as the best model to generate the result in Fig. 4. This result shows the weighted F1-scores which is obtained by different values of N and M to determine the effect of over and under-sampling on MLP models' performance. We leverage the logarithmic space to generate a range of values for both over-sampling (N) and under-sampling (M).

Fig. 4 highlights the model's performance with varying over-sampling and under-sampling rates for generating the training

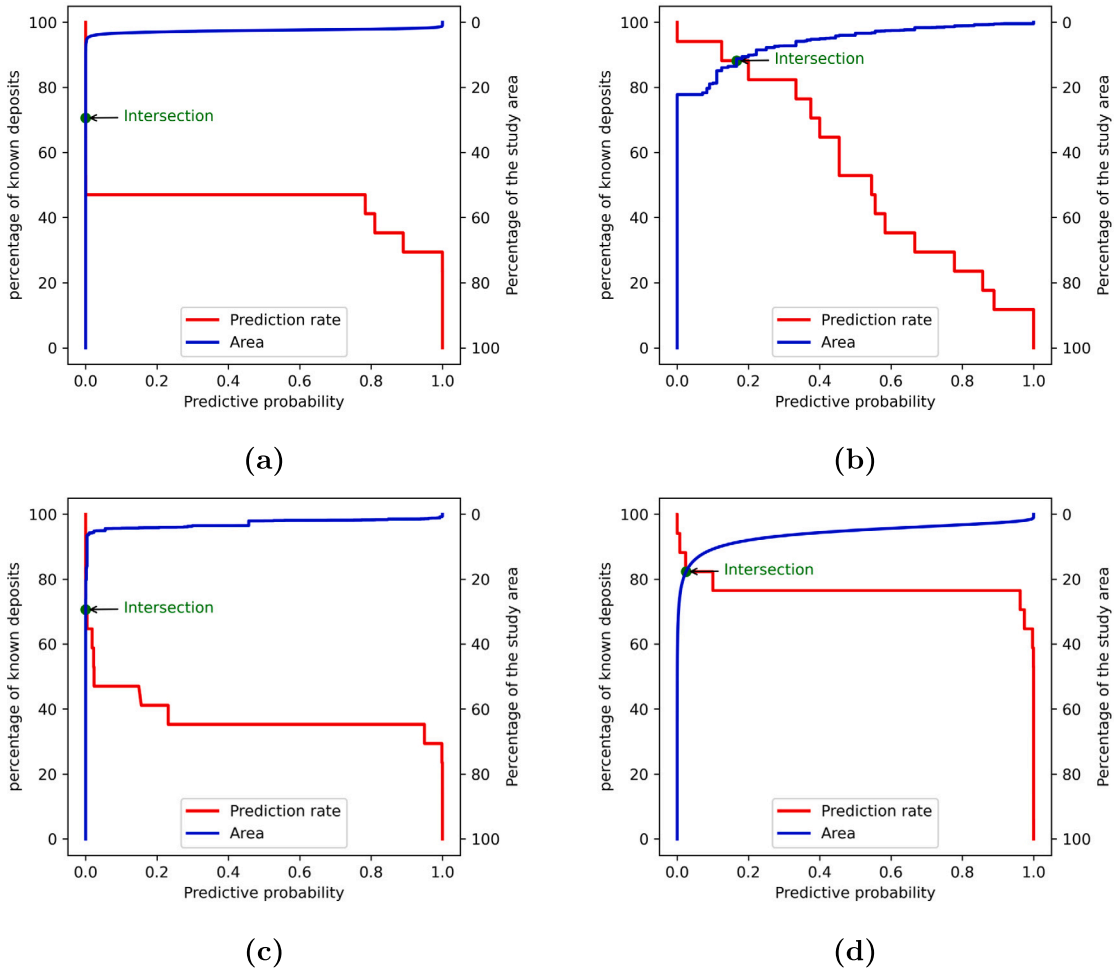

Fig. 5. The P–A plots for each model: (a) MLP, (b) RF, (c) DT, and (d) LR.

dataset. Models with the highest F1 scores were trained with low mineralized over-sampling rates and high non-mineralized under-sampling rates (bottom right of Fig. 4). Similarly, models with high mineralized over-sampling rates and low non-mineralized under-sampling rates (top left of Fig. 4) also achieved high F1 scores. This indicates that models trained with imbalanced datasets, having a high ratio of mineralized to non-mineralized samples, tend to overfit.

Conversely, models trained with over-sampling below 860 mineralized samples exhibited the lowest training F1 scores (bottom left of Fig. 4). Some of these models had testing F1 scores below 98, indicating that imbalanced datasets with low mineralized to non-mineralized ratios tend to underfit. MLP model trained with balanced datasets, having equal numbers of mineralized and non-mineralized samples (highlighted diagonal of Fig. 4), shows F1 scores ranging from 92.2 to 100. The F1 scores for these models consistently increased with higher over-sampling rates of the mineralized class and lower under-sampling rates of the non-mineralized class, resulting in a greater total number of training samples (from bottom left to top right of Fig. 4). This suggests that models trained with balanced datasets tend to be more stable and are less prone to overfitting or underfitting. This result for RF, DT, and LR are also presented in Appendices Appendix B (Fig. B.16), C (C.17), and D (Fig. D.18) respectively.

5.3. Predictive efficiency

To effectively evaluate the predictive performance of our models in the context of mineral exploration, we utilized Prediction–Area (P–A) plots and success-rate curves. The P–A plot (Yousefi and Carranza, 2015), and the success-rate curve (Chung and Fabbri, 1999), are widely accepted methods for assessing the predictive power of spatial prediction models. These tools are crucial for linking modeling results to practical applications by focusing on predictive efficiency, specifically the ability to capture more deposits within smaller target areas.

The P–A plot is used to visualize the relationship between the cumulative percentage of predicted mineral deposits and the cumulative percentage of the area. P–A plots for the proposed ML models are shown in Fig. 5. For MLP model (Fig. 5(a)), we can see the prediction rate (red line) is stable across all probabilities, indicating consistent predictive accuracy. The area (blue) decreases sharply at a low probability, showing that MLP effectively concentrates high-probability predictions in a smaller area. RF Model (Fig. 5(b)) shows a moderate intersection between prediction rate and area, suggesting it covers more area to achieve similar deposit predictions compared to MLP, indicating a dispersed prediction pattern. DT Model (Fig. 5(c))has a steep area decline at higher probabilities with a significant drop in prediction rate beyond the 0.4 threshold, implying effectiveness at higher probabilities but limited coverage at lower ones. LR Model (Fig. 5(d)) quickly focuses on very high-probability zones, as shown by the sharp area drop at low probabilities. The high prediction rate across the curve indicates strong accuracy in targeted high-probability areas.

The success-rate curves further complement the P–A plots by plotting the cumulative percentage of correctly predicted deposits against the cumulative percentage of the area, thereby providing a quantitative measure of the model's predictive performance. The success-rate curves for all four models are depicted in Fig. 6. From the results, we can see the RF model shows the highest efficiency, with its curve positioned

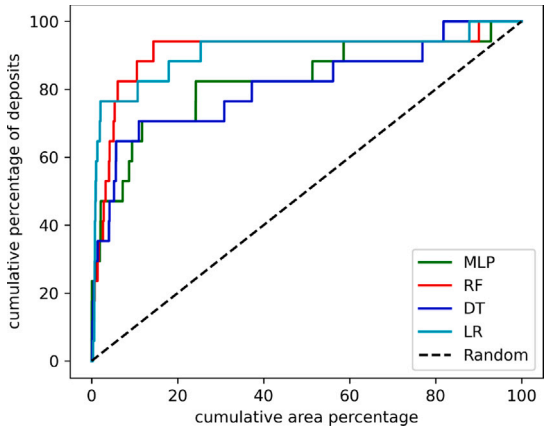

Fig. 6. Success-rate curves for the machine learning models.

highest, indicating superior performance in identifying deposits over smaller areas. The DT also performs robustly, though slightly below RF, showcasing its capability to identify deposits effectively over moderate areas. The MLP model follows, presenting a balanced identification rate across the examined area, while LR is the least efficient among the specific models, yet still performs better than random guessing, depicted by the Random model's baseline curve.

In conclusion, the P–A plots and success-rate curves provide a comprehensive evaluation of the predictive efficiency of our models. The MLP and RF models exhibit superior performance in identifying mineralized areas within a smaller target region, thereby offering more practical value for mineral exploration applications.

5.4. Confusion matrix

To find more details about the performance of models for two classes (mineralized and non-mineralized), we interpreted the confusion matrices. The rows in confusion matrix represent the actual class labels, and the columns represent the predicted class labels. From confusion matrix, four outcomes of a classification results were summarized including These terms can be defined in our application as follows:

1. True Positive (TP): Pixels classified as "mineralized" that are indeed part of a mineralized region in the ground truth (actual positive instances).
2. False Positive (FP): Pixels classified as "mineralized" that are actually part of a non-mineralized region in the ground truth (incorrectly classified as positive instances).
3. True Negative (TN): Pixels classified as "non-mineralized" that are indeed part of a non-mineralized region in the ground truth (actual negative instances).
4. False Negative (FN): Pixels classified as "non-mineralized" that are actually part of a mineralized region in the ground truth (incorrectly classified as negative instances).

The result which is depicted in Fig. 7(a) shows that MLP can get the maximum correct observations belongs to class "non-mineralized". For RF, as illustrated in Fig. 7(b), the highest correct observations belong to the classes "mineralized" which is 88.24%.

5.5. Decision threshold

Thresholding allows model to fine-tune the classifier's decision boundary to better account for the class imbalance. Fig. 8(a) shows

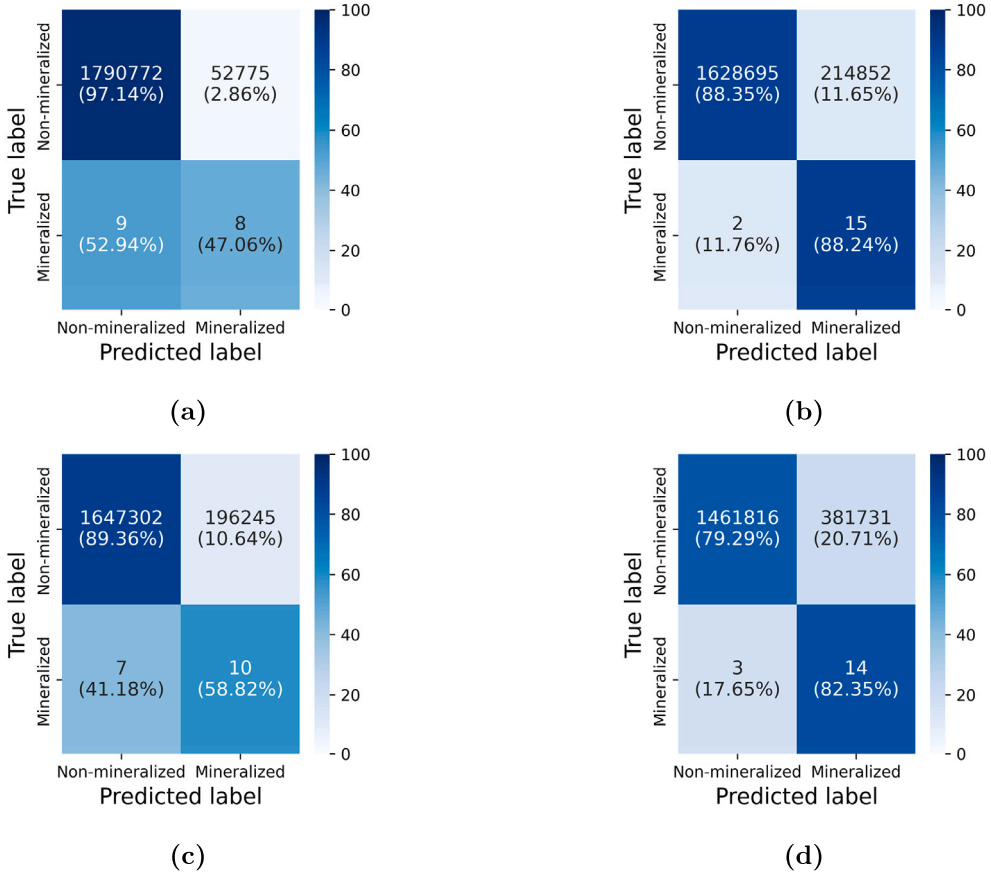

Fig. 7. The confusion matrix of (a) MLP, (b) RF, (c) DT, and (d) LR on the test dataset.

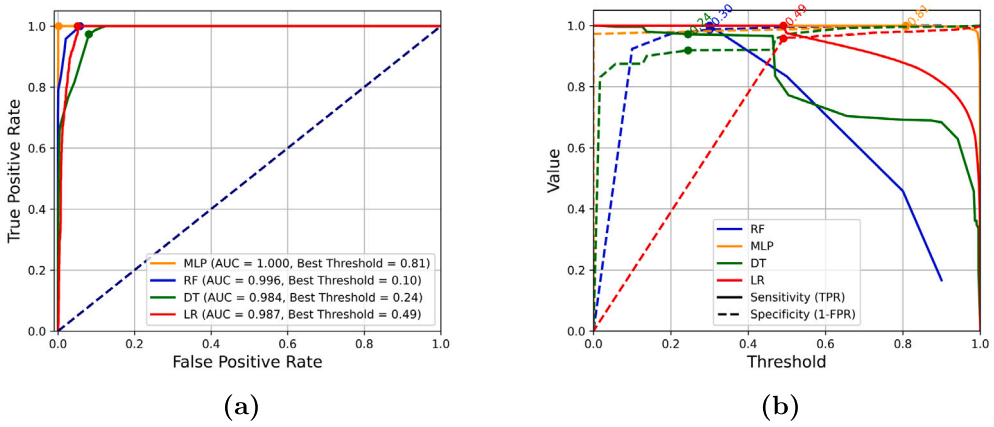

Fig. 8. (a) ROC and (b) sensitivity-specificity curves of proposed models using test dataset based on different thresholds.

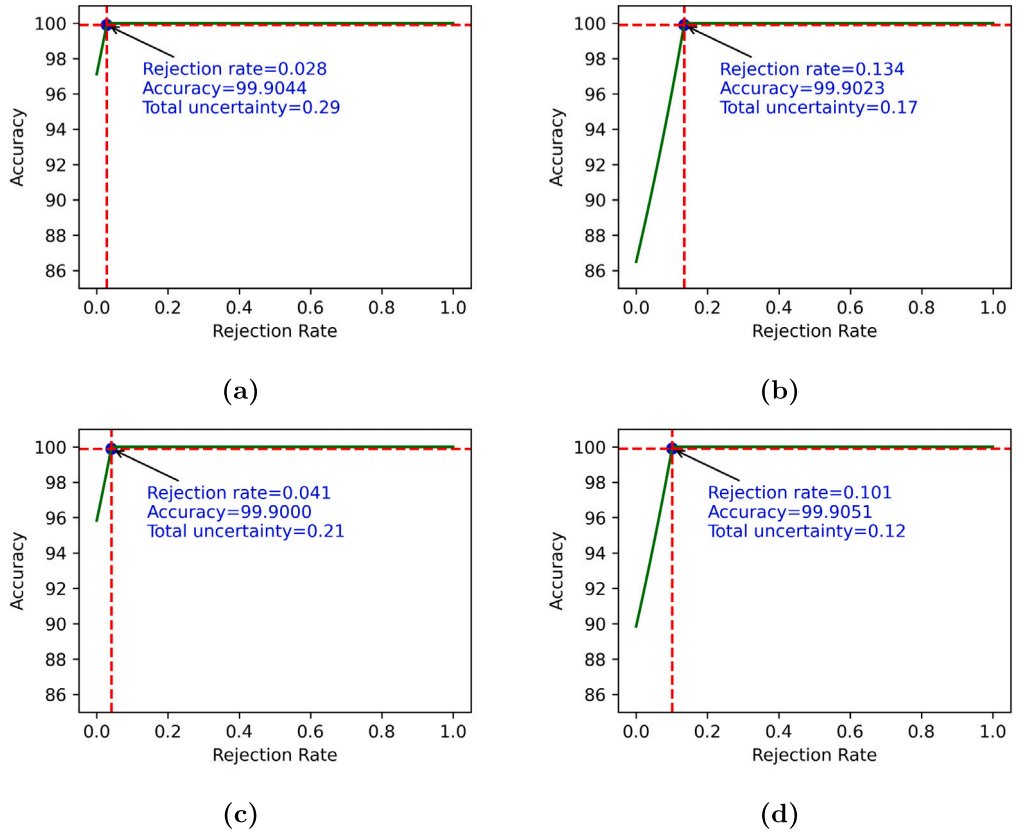

Fig. 9. The accuracy-rejection curves of (a) MLP, (b) RF, (c) DT, and (d) LR.

ROC (Receiver Operating Characteristic) curves plotted for threshold values, showing the trade-off between true positive rate (sensitivity) and false positive rate as the number of samples generated varies. This can help in assessing the impact of sampling on the model's ability to discriminate between classes for MPM (Nykänen et al., 2015). The ROC curves show that the AUC value of the result obtained by MLP is the highest followed by RF, LR, and DT with balanced data. In addition, it shows that the best threshold value for all the 4 models. Fig. 8(b) plots the sensitivity-specificity curve and consider adjusting the decision threshold based on the trade-off between sensitivity and sensitivity. Each color represents a different model, with the solid lines indicating the sensitivity and the dashed lines the specificity at various thresholds. MLP achieve high sensitivity without a significant loss in specificity which highlights its robustness.

5.6. Uncertainty estimation

In this study, accuracy-rejection curves (Nadeem et al., 2009) with total uncertainty were utilized to evaluate the performance of the models under conditions of synthetic positive samples. Fig. 9 shows

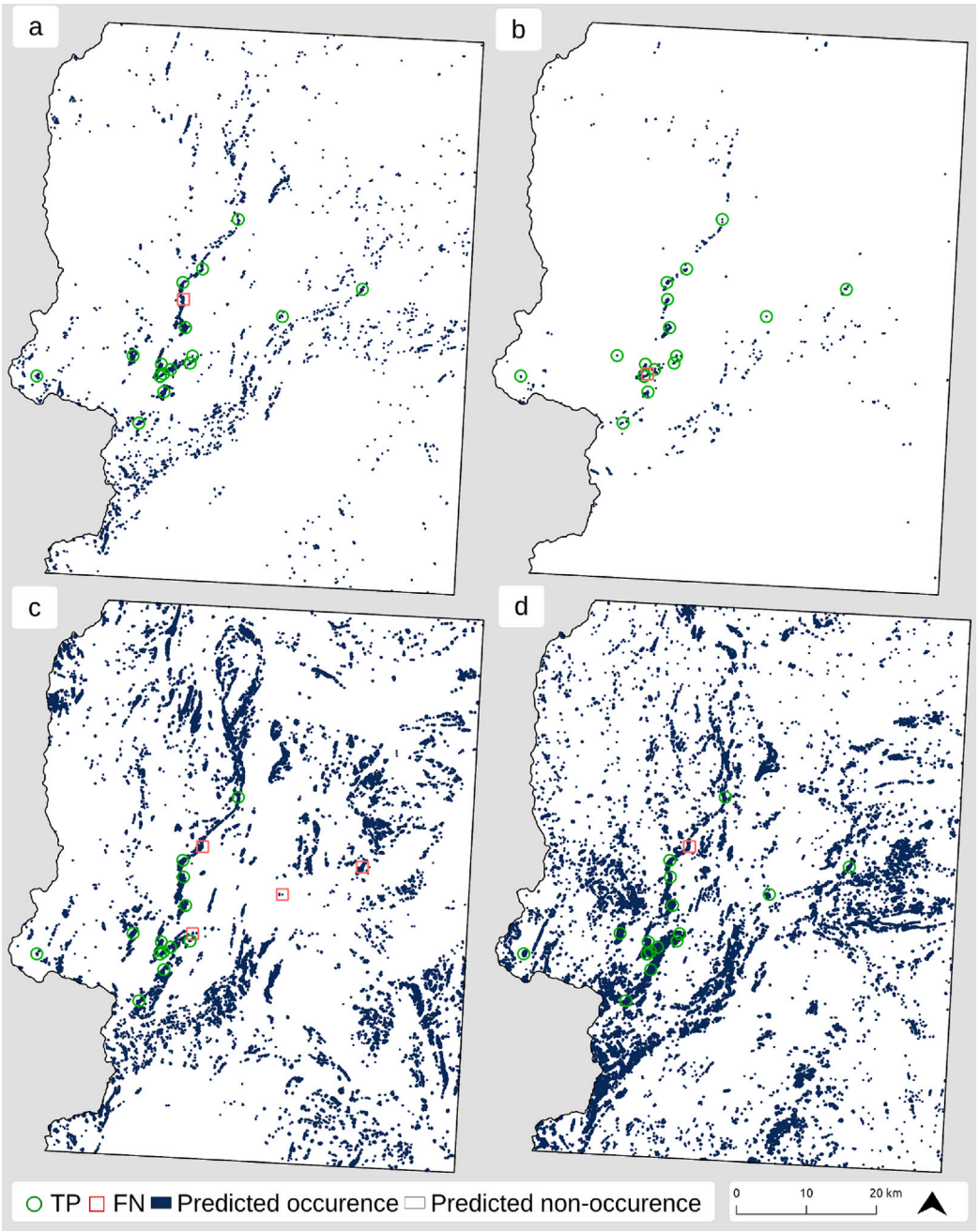

Fig. 10. Prospectivity mineral maps for (a) MLP, (b) RF, (c) DT, and (d) LR models. The occurrence sizes have been visually exaggerated to make them visible.

that all models achieved near-perfect accuracy at very low rejection rates, highlighting the robustness of the predictions. The total uncertainty, which encompasses both aleatoric and epistemic uncertainties, remained significantly low across all models, even with the injection of synthetic positives. For instance, in MLP, the total uncertainty is low (0.24), indicating that the model has high confidence in its predictions. In addition, the rejection rate is extremely low (0.029), suggesting that very few predictions are rejected.

This outcome underscores the effectiveness of the uncertainty quantification mechanisms embedded within our approach. The consistency of low uncertainty across various computational models validates the reliability of our predictions and supports the decision-making process in the exploration of mineral resources. By incorporating synthetic positives, our approach not only maintains high accuracy but also ensures that the model's integrity is not compromised, thereby bolstering the confidence in the predictive capabilities of our framework.

5.7. Mineral prospectivity mapping

Fig. 10 showcases the mineral prospectivity maps produced by all four models. For generating these maps, we utilized the entire dataset

for training the models based on the optimal values of model hyper-parameters, decision threshold, and the number of over-sampling and under-sampling instances. We also visualize the 17 mineralized samples on the maps and indicate TP and FN samples with different colors to interpret how the ML models perform for MPM. To enhance the interpretability of these occurrence points on the maps, their representation has been exaggerated. This deliberate visual adjustment ensures that even on a broad geographical scale, where actual mineralized sizes would be visually indiscernible, the analysis of model performance remains straightforward. The color-coding effectively conveys the accuracy of each model in spatially discriminating between zones with high mineralization potential from less prospective areas, thereby demonstrating the models' utility in guiding exploratory endeavors in geoscience.

Geological insights provided by experienced geologists confirm that all four maps are geologically sensible, with predicted occurrences aligning with potential mineralized sites based on the region's geophysical characteristics, particularly magnetic anomalies. Maps A and B are noted for their accuracy in prediction without an excess of predicted sites, making them particularly valuable for field geologists. Map A, in particular, demonstrates an impressive capacity to delineate separate ore bodies within a single mineralized, a feat likely aided by the magnetic data employed. Despite the general bias associated with using aeromagnetic anomaly maps and their derivatives in modeling, our methods have managed to produce meaningful results, as evidenced in the geologically informed predictions across all maps. However, it is recommended to use multiple points to represent the true spatial extent of mineralized samples for future modeling to enhance the accuracy and utility of these predictions.

6. Conclusion

This paper presents an innovative approach based on machine learning methods, including random forest (RF), dicision tree (DT), logistic regression (LR), and multi-layer perceptron (MLP), to predict mineral prospectivity in Lapland, Finland. The study utilizes multi-source geo-information from a dataset with $n_+ = 17$ known mineralized and $n = 1.84 \times 10^6$ non-mineralized. The data is so imbalanced and only there is a few labeled data. To address this problem, we proposed different approaches to generate balanced data before applying ML methods. By systematically addressing class imbalances, incorporating a decision threshold, and employing rigorous validation techniques, our training methodology ensures the robustness and generalizability of machine learning models for mineral prospectively mapping (see Fig. B.16). The results demonstrate that the MLP model exhibits the best classification performance with an accuracy of 97.13%. This high accuracy is complemented by an impressive F1-score of 98.52%, suggesting a robust balance between precision and recall. The Random Forest RF model, while trailing behind the MLP in terms of accuracy with an 88.34% score, demonstrates exceptional performance in sensitivity, achieving 88.23%. This highlights the RF model's strength in correctly identifying positive cases.

What comes to the positive set itself, our experiments with affine models based on subsets of positive samples indicate that each sample is useful contributor for model performance and the prediction would benefit from having more positive samples. The paper also discusses the challenges and considerations involved in applying these techniques to real-world geophysical datasets. These challenges include the availability and quality of data, the choice of appropriate evaluation metrics, and the interpretation of model results.

The findings of the study demonstrate that innovative machine learning algorithms can significantly enhance the predictive accuracy of mineral prospectivity modeling and contribute to the identification of undiscovered mineralization potential.g. In addition, this paper provides an outlook on future research directions in this area, including the development of novel techniques for handling complex and heterogeneous geophysical data, and the exploration of ML approaches for mineral prospectivity mapping (see Figs. C.17 and D.18).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jukka Heikkonen reports financial support was provided by Horizon Europe. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The compilation of the presented work is supported by funds from the Horizon Europe research and innovation program under Grant Agreement number 101057357, EIS – Exploration Information System (https://eis-he.eu). We extend our special thanks to Tero Niiranen for his valuable geological insights and evaluations that significantly enhanced the interpretability and accuracy of our mineral prospectivity maps.

Appendix A. Spatial data inputs used in modeling

See Figs. A.11–A.15.

Appendix B. Effect of balancing techniques on RF

See Fig. B.16.

Appendix C. Effect of balancing techniques on DT

See Fig. C.17.

Appendix D. Effect of balancing techniques on LR

See Fig. D.18.

Data availability

The data that has been used is confidential.

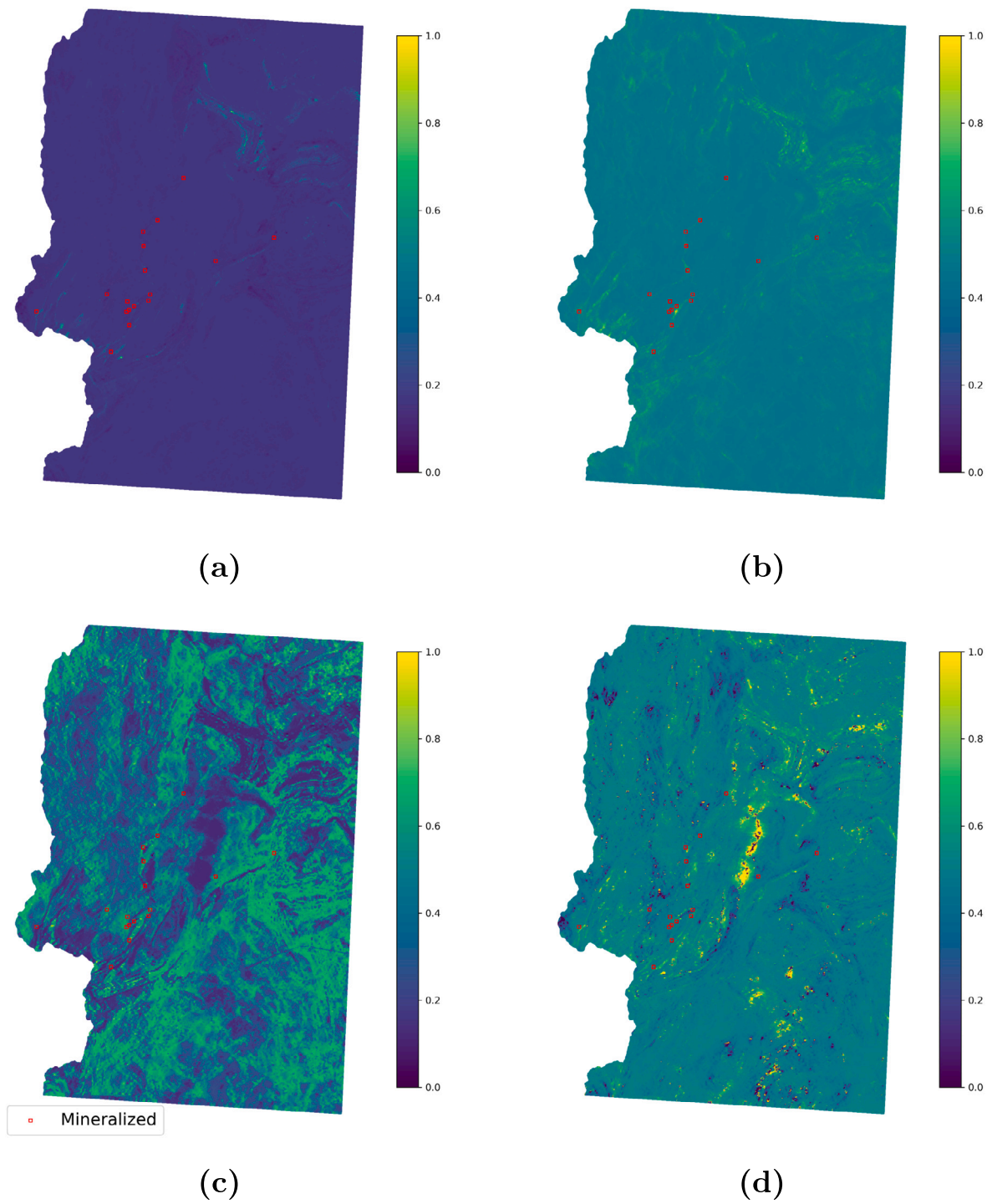

Fig. A.11. Predictor maps of AEM features including (a) Aeroelectromagnetic Inphase components (b) Aeroelectromagnetic Quadrature components (c) Aeroelectromagnetic Apparent resistivity (d) electromagnetic ratio.

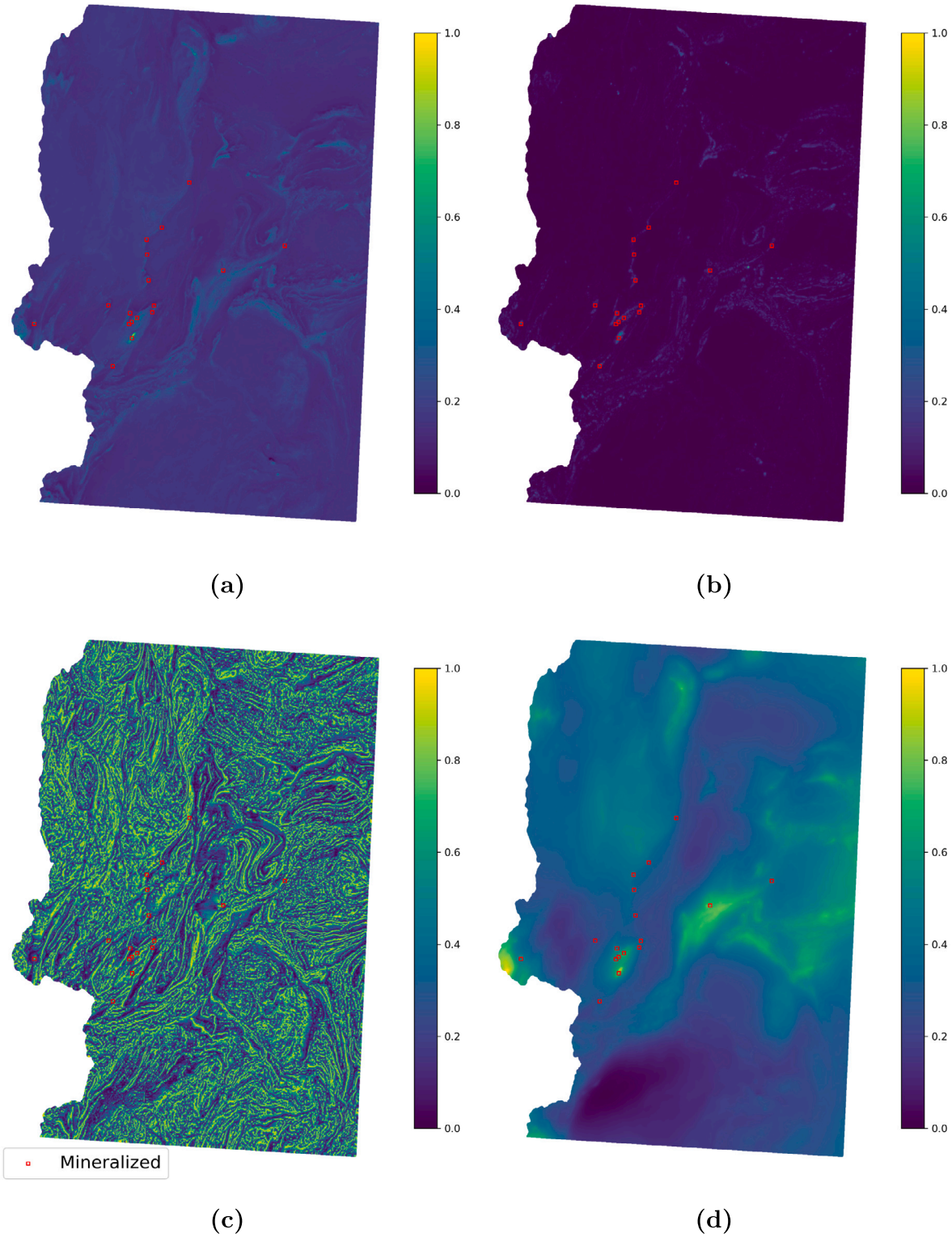

Fig. A.12. Predictor maps of Magnetic features including (a) DGRF65 Anomaly, (b) analytical Signal of magnetic anomaly, (c) tilt derivative of the magnetic field, and (d) Pseudogravity derived from magnetism.

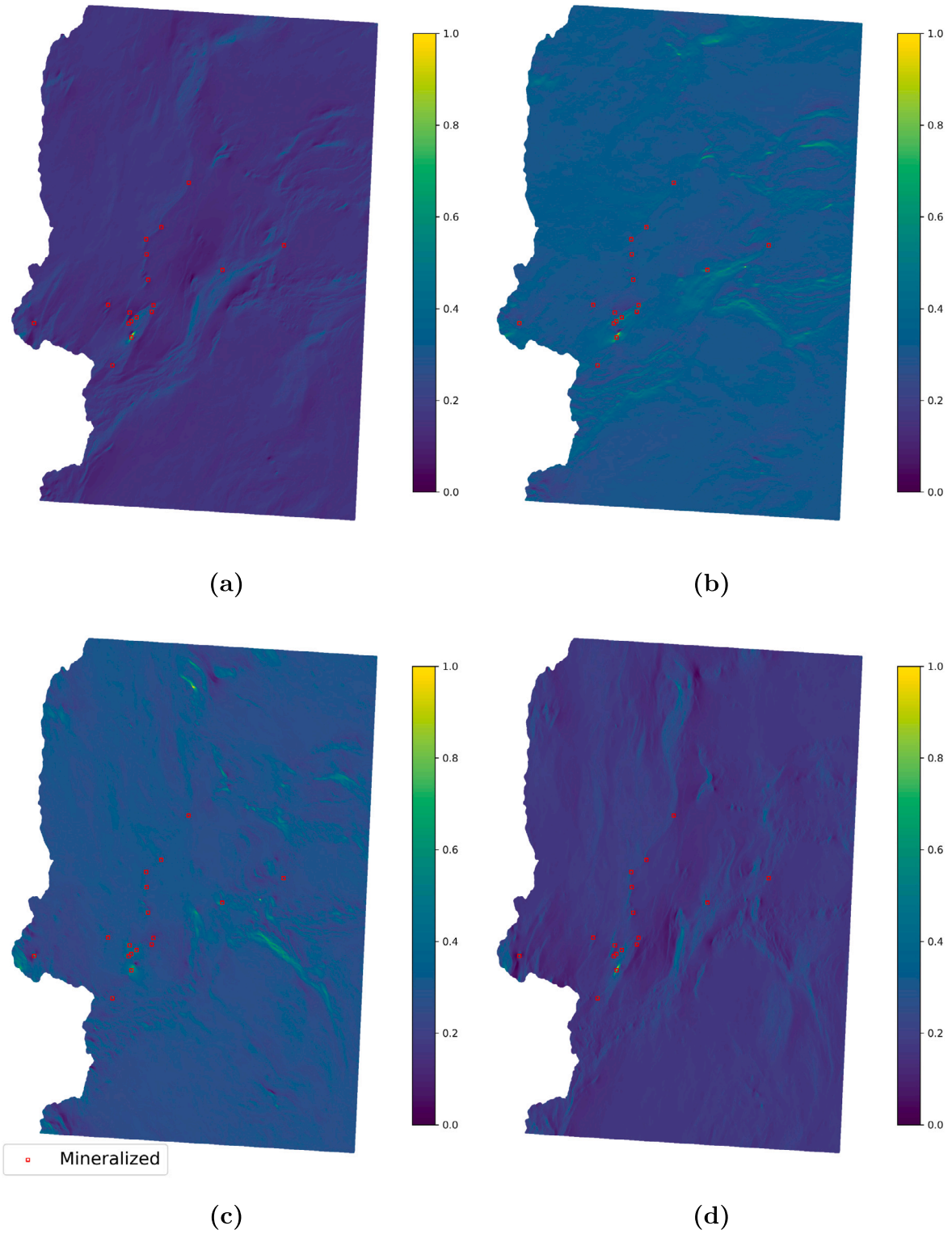

Fig. A.13. Predictor maps of Magnetic features for directional cosines of magnetic anomaly in directions (a) 45, (b) 90, (c) 135, and (d) 180.

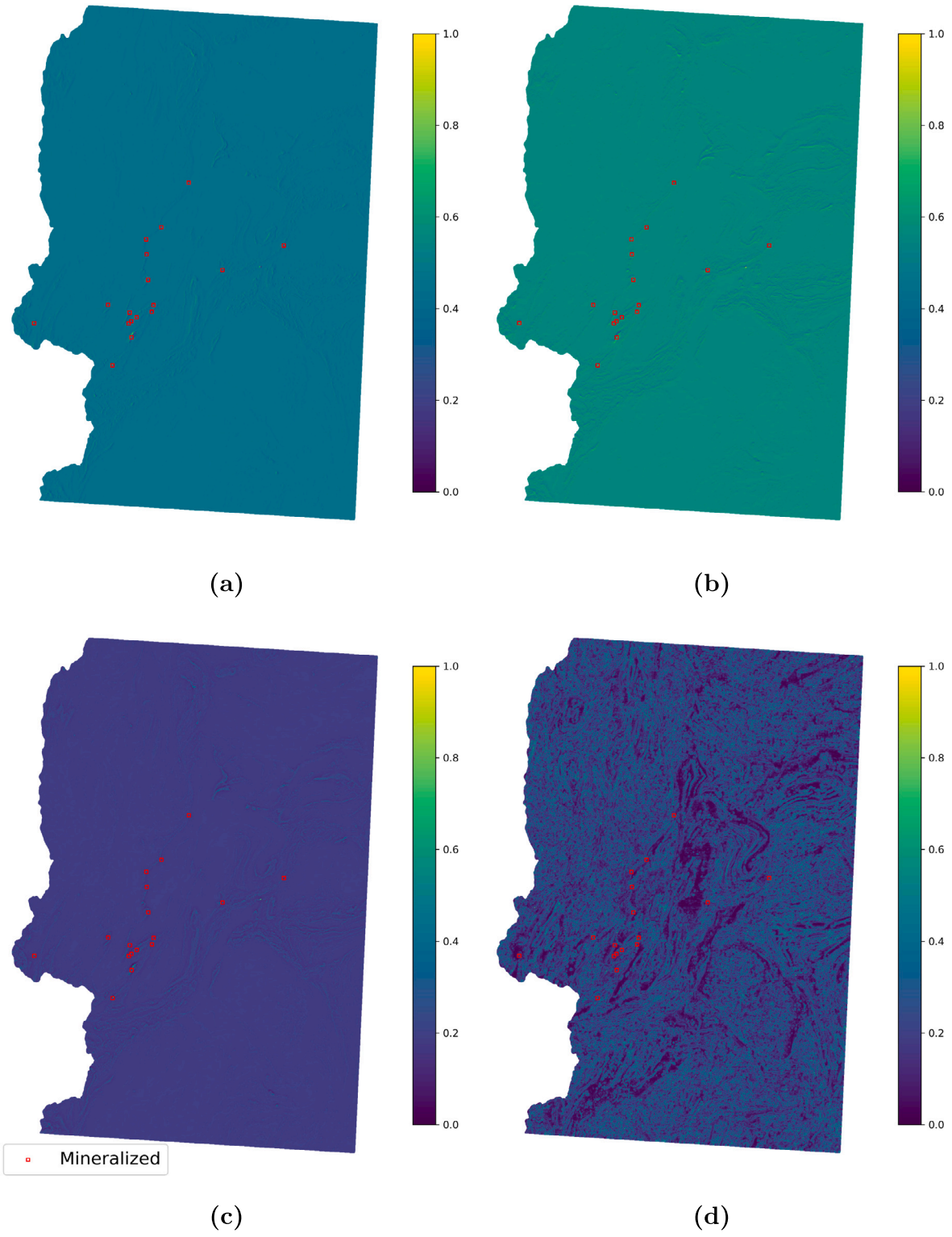

Fig. A.14. Predictor maps of Magnetic features including (a) X, (b) Y, (c) Z, and (d) Horizontal derivative of the tilt derivative.

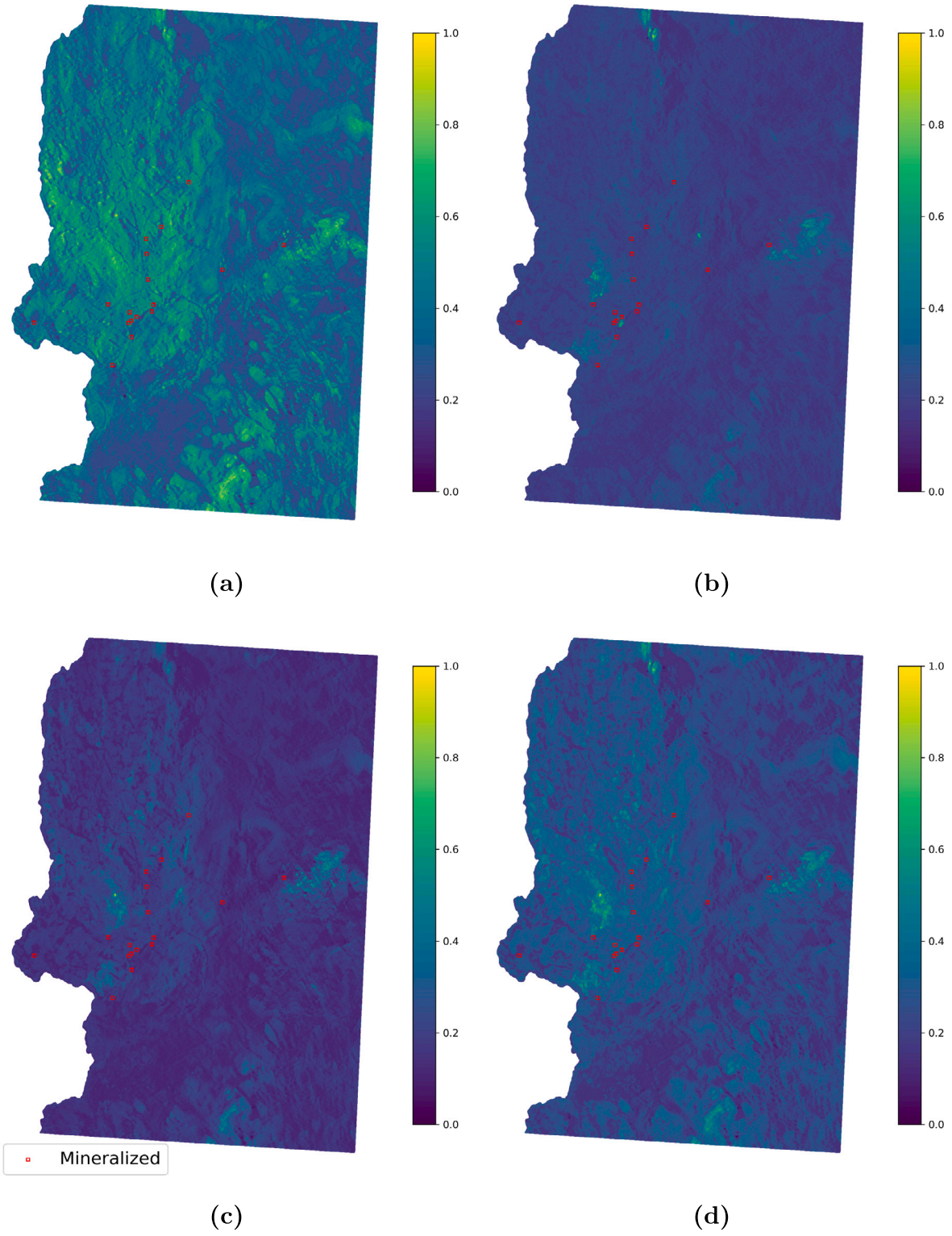

Fig. A.15. Predictor maps of Radiometric features gamma ray radioactivity of element including (a) K, (b) U, (c) Th, and (d) total Gamma ray radioactivity.

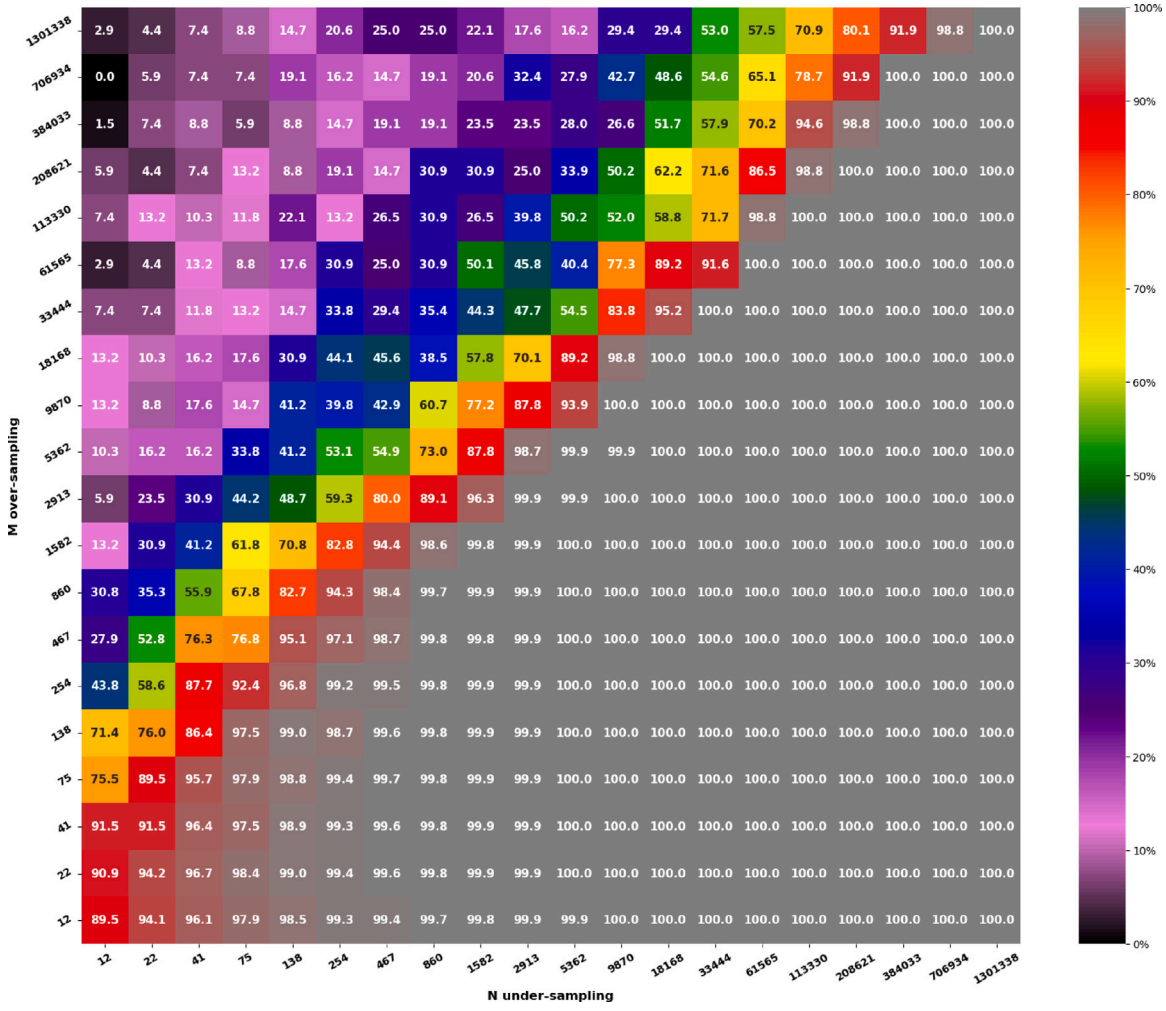

Fig. B.16. The weighted F1-score (%) for the training dataset based on different number of over-sampling N and under-sampling M for RF.

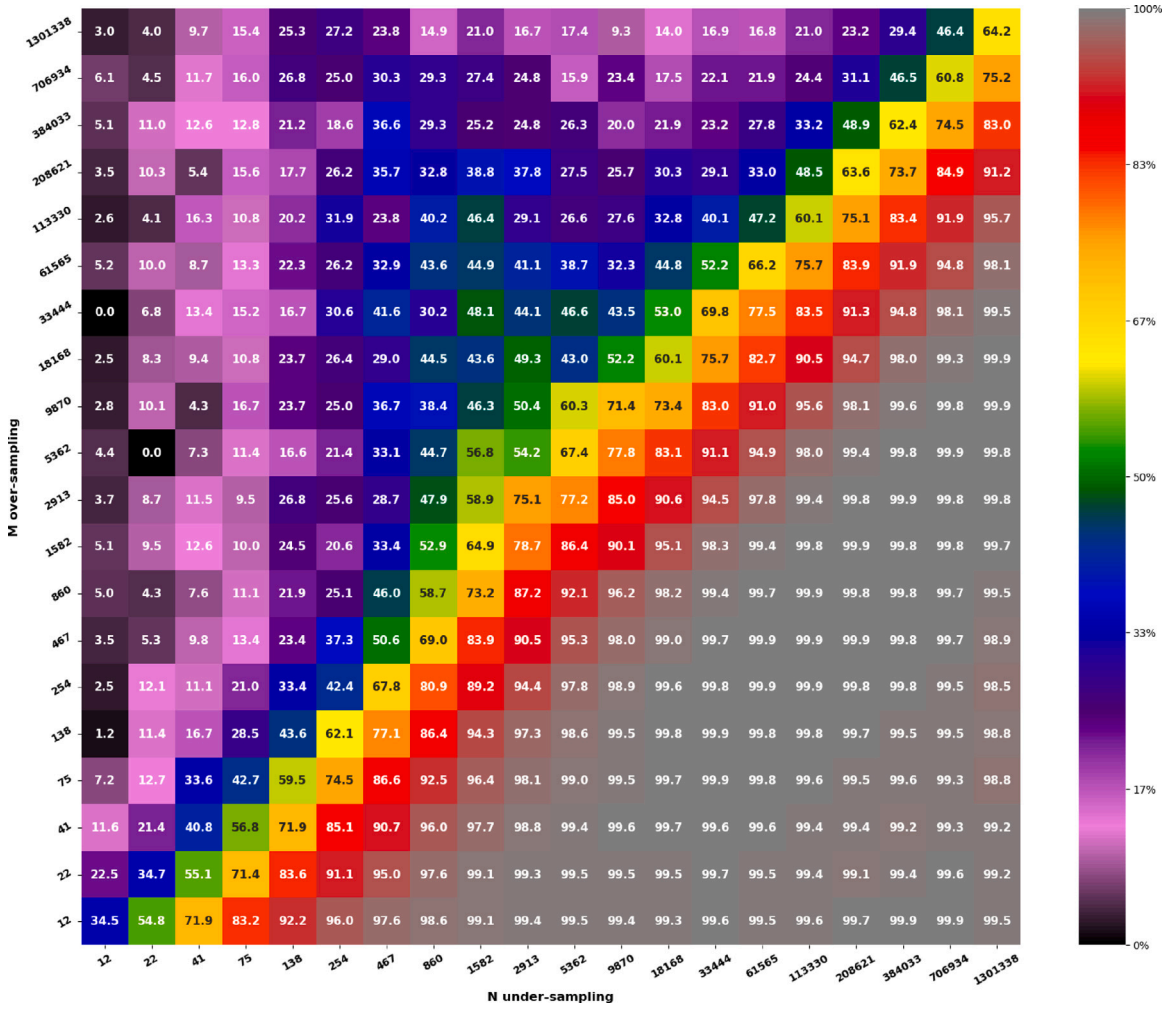

Fig. C.17. The weighted F1-score (%) for the training dataset based on different number of over-sampling N and under-sampling M for DT.

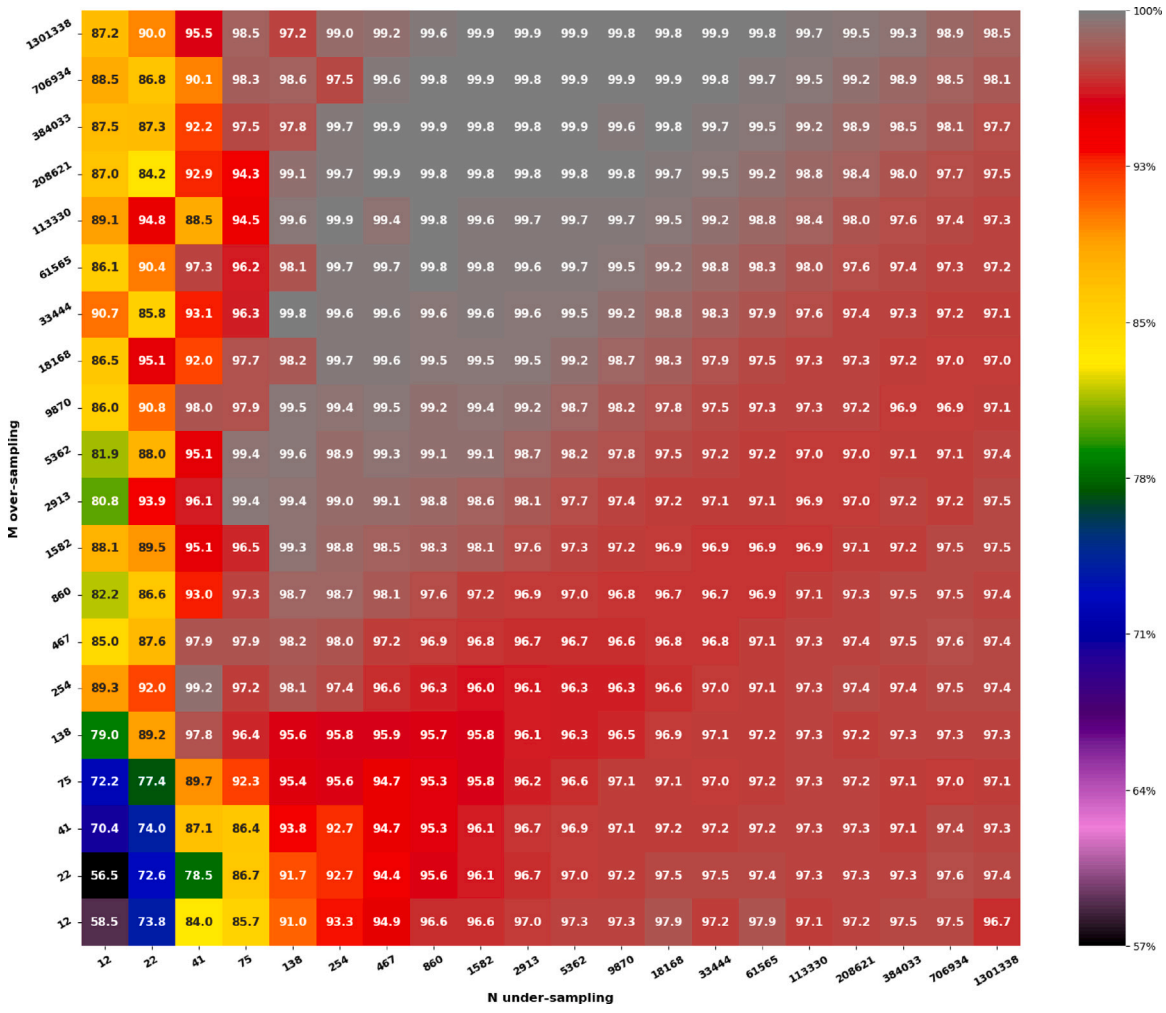

Fig. D.18. The weighted F1-score (%) for the training dataset based on different number of over-sampling N and under-sampling M for LR.

References

Brandmeier, M., Zamora, I.G.C., Nykänen, V., Middleton, M., 2019. Boosting for mineral prospectivity modeling: A new gis toolbox. Natural Resour. Res. 29, 71–88, URL https://api.semanticscholar.org/CorpusID:146011015.

Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 106, 249–259, URL https://www.sciencedirect.com/science/article/pii/S0893608018302107.

Carranza, E.J.M., Laborte, A.G., 2015. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in abra (Philippines). Comput. Geosci. 74, 60–70, URL https://www.sciencedirect.com/science/article/pii/S0098300414002271.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Cheng, Q., 2007. Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan province, China. Ore Geol. Rev. 32 (1), 314–324. http://dx.doi.org/10.1016/j.oregeorev.2006.10.002, URL https://www.sciencedirect.com/science/article/pii/S0169136806001119.

Chudasama, B., Torppa, J., Nykänen, V., Kinnunen, J., 2022. Target-scale prospectivity modeling for gold mineralization within the rajapalot au-co project area in northern Fennoscandian shield, Finland. part 2: Application of self-organizing maps and artificial neural networks for exploration targeting. Ore Geol. Rev. 147 (104), 936. http://dx.doi.org/10.1016/j.oregeorev.2022.104936, URL https://www.sciencedirect.com/science/article/pii/S016913682200244X.

Chung, C.J.F., Fabbri, A.G., 1999. Probabilistic prediction models for landslide hazard mapping. Photogrammetr. Eng. Remote Sens. 65 (12), 1389–1399.

Ferreira da Silva, G., Silva, A.M., Toledo, C.L.B., Chemale Junior, F., Klein, E.L., 2022. Predicting mineralization and targeting exploration criteria based on machine-learning in the serra de jacobina quartz-pebble-metaconglomerate au-(u) deposits, São Francisco Craton, Brazil. J. South Am. Earth Sci. 116 (103), 815. http://dx.doi.org/10.1016/j.jsames.2022.103815, URL https://www.sciencedirect.com/science/article/pii/S0895981122001067.

Foster, D.J., Kale, S., Luo, H., Mohri, M., Sridharan, K., 2018. Logistic regression: The importance of being improper.

Genuer, R., Poggi, J.M., Tuleau, C., 2008. Random forests: some methodological insights.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. In: Springer series in statistics, Springer.

Jiang, W., He, G., Long, T., Ni, Y., Liu, H., Peng, Y., Lv, K., Wang, G., 2018. Multilayer perceptron neural network for surface water extraction in landsat 8 oli satellite images. Remote Sens. 10 (5), http://dx.doi.org/10.3390/rs10050755, URL https://www.mdpi.com/2072-4292/10/5/755.

Jung, D., Choi, Y., 2021. Systematic review of machine learning applications in mining: Exploration, exploitation, and reclamation. Minerals 11, 148. http://dx.doi.org/10.3390/min11020148.

Kotsiantis, S.B., Kanellopoulos, D.N., Pintelas, P.E., 2006. Handling imbalanced datasets: A review. URL https://api.semanticscholar.org/CorpusID:14354369.

Kotsiantis, S., Pintelas, P., 2004. Mixture of expert agents for handling imbalanced data sets. Ann. Math. Comput. Teleinformat. 1, 46–55.

Kreuzer, O., Yousefi, M., Nykänen, V., 2020. Introduction to the special issue on spatial modelling and analysis of ore-forming processes in mineral exploration targeting. Ore Geol. Rev. 119 (103), 391. http://dx.doi.org/10.1016/j.oregeorev.2020.103391.

Nadeem, M.S.A., Zucker, J.-D., Hanczar, B., 2009. Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. In: Proceedings of the Third International Workshop on Machine Learning in Systems Biology, Volume 8 of Proceedings of Machine Learning Research. PMLR, Ljubljana, Slovenia, pp. 65–81, URL https://proceedings.mlr.press/v8/nadeem10a.html.

Niiranen, T., 2005. Iron oxide-copper-gold deposits in finland : case studies from the peräpohja schist belt and the central lapland greenstone belt. URL https://api.semanticscholar.org/CorpusID:55508395.

Nusrat, I., Jang, S.B., 2018. A comparison of regularization techniques in deep neural networks. Symmetry 10 (11), http://dx.doi.org/10.3390/sym10110648, URL https://www.mdpi.com/2073-8994/10/11/648.

Nykänen, V., Lahti, I., Niiranen, T., Korhonen, K., 2015. 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. 71, 853–860. http://dx.doi.org/10.1016/j.oregeorev.2014.09.007, URL https://www.sciencedirect.com/science/article/pii/S0169136814002091.

Prado, E.M.G., de Souza Filho, C.R., Carranza, E.J.M., Motta, J.G., 2020. Modeling of cu-au prospectivity in the Carajás mineral province (Brazil) through machine learning: Dealing with imbalanced training data. Ore Geol. Rev. 124 (103), 611. http://dx.doi.org/10.1016/j.oregeorev.2020.103611, URL https://www.sciencedirect.com/science/article/pii/S0169136819308819.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1 (1), 81–106. http://dx.doi.org/10.1007/BF00116251.

Spelmen, V.S., Porkodi, R., 2018. A review on handling imbalanced data. In: 2018 International Conference on Current Trends Towards Converging Technologies. ICCTCT, pp. 1–11. http://dx.doi.org/10.1109/ICCTCT.2018.8551020.

Sun, T., Li, H., Wu, K., Chen, F., Zhu, Z., Hu, Z., 2020. : Data-driven predictive modelling of mineral prospectivity using machine learning and deep learning methods: A case study from Southern Jiangxi Province, China. Minerals 10 (2), http://dx.doi.org/10.3390/min10020102, URL https://www.mdpi.com/2075-163X/10/2/102.

Wainer, J., Cawley, G.C., 2018. Nested cross-validation when selecting classifiers is overzealous for most practical applications. CoRR abs/1809.09446. URL http://arxiv.org/abs/1809.09446.

Xiong, Y., Zuo, R., 2017. Effects of misclassification costs on mapping mineral prospectivity. Ore Geol. Rev. 82, 1–9. http://dx.doi.org/10.1016/j.oregeorev.2016.11.014, URL https://www.sciencedirect.com/science/article/pii/S0169136816303924.

Xiong, Y., Zuo, R., 2018. Gis-based rare events logistic regression for mineral prospectivity mapping. Comput. Geosci. 111, 18–25. http://dx.doi.org/10.1016/j.cageo.2017.10.005, URL https://www.sciencedirect.com/science/article/pii/S0098300417305447.

Xu, B., Huang, J., Williams, G., Wang, Q., Ye, Y., 2012. : Classifying very high-dimensional data with random forests built from small subspaces. Int. J. Data Warehousing Mining 8, http://dx.doi.org/10.4018/jdwm.2012040103.

Yadav, S., Bhole, G.P., 2020. Handling imbalanced dataset classification in machine learning. In: 2020 IEEE Pune Section International Conference (PuneCon). pp. 38–43. http://dx.doi.org/10.1109/PuneCon50868.2020.9362471.

Yousefi, M., Carranza, E.J.M., 2015. Prediction–area (p–a) plot and c–a fractal analysis to classify and evaluate evidential maps for mineral prospectivity modeling. Comput. Geosci. 79, 69–81. http://dx.doi.org/10.1016/j.cageo.2015.03.007.

Zhang, N., Zhou, K., Li, D., 2018. Back-propagation neural network and support vector machines for gold mineral prospectivity mapping in the Hatu Region, Xinjiang, China. Earth Sci. Inform. 11, http://dx.doi.org/10.1007/s12145-018-0346-6.

Zhang, Z., Zuo, R., Xiong, Y., 2015. A comparative study of fuzzy weights of evidence and random forests for mapping mineral prospectivity for skarn-type fe deposits in the southwestern Fujian metallogenic belt, China. Sci. China Earth Sci. 59, http://dx.doi.org/10.1007/s11430-015-5178-3.

Zou, Q., Xie, S., Lin, Z., Wu, M., Ju, Y., 2016. Finding the best classification threshold in imbalanced classification. Big Data Res. 5, 2–8. http://dx.doi.org/10.1016/j.bdr.2015.12.001, URL https://www.sciencedirect.com/science/article/pii/S2214579615000611. Big data analytics and applications.

Zuo, R., 2020. Geodata science-based mineral prospectivity mapping: A review. Nat. Res. Res. 1–10, URL https://api.semanticscholar.org/CorpusID:218682801.