Enhancing Minerals Prospects Mapping with Machine Learning: Addressing Imbalanced Geophysical Datasets and Data Visualization Approaches

Dipak Kumar Nidhi, Iiro Seppä, Fahimeh Farahnakian, Luca Zelioli, Jukka Heikkonen, Rajeev Kanth
University of Turku, Savonia University of Applied Sciences Kuopio
Turku and Kuopio, Finland
{dknidh, iielse, fahimeh.farahnakian, luzeli, jukhei}@utu.fi, rajeev.kanth@savonia.fi

***Abstract*—Minerals prospects mapping plays a pivotal role in the sustainable development of mineral resources, offering critical insights into subsurface geology and mineral potential. Traditional geological methods are often labor-intensive and time-consuming. In contrast, machine learning (ML) techniques have emerged as a powerful tool for accelerating and improving the accuracy of mineral prospect mapping. This article explores an innovative approach to enhance the performance of supervised ML models, specifically logistic regression and multilayer perceptron. One of the primary challenges in mineral mapping is dealing with imbalanced geophysical datasets, where positive samples (indicating mineral occurrences) are vastly outnumbered by negative samples (non-mineral areas). This imbalance can lead to biased model predictions, favoring the majority class while neglecting the minority class. To address this issue, we propose a novel oversampling technique that generates synthetic samples for the minority class, effectively rebalancing the dataset. By introducing diversity to the training data, our approach mitigates the bias and enhances the models' ability to identify mineral prospects accurately. The proposed approach empowers ML models to discriminate between mineral-rich and non-mineral areas with unprecedented precision, facilitating more informed decision-making for resource exploration and extraction. Ultimately, the integration of imbalanced dataset handling and data visualization techniques offers a robust framework for harnessing the potential of machine learning in mineral prospect mapping.**

I. Introduction

Mineral prospectivity mapping is crucial in providing a fundamental framework for understanding the complex nature of the Earth's underlying geology and the potential treasures it may contain [1] [2] [3] [4]. Fundamentally, this technology identifies areas likely to have significant economic value from mineral reserves. Consequently, it offers exploration teams important information on prospective target places to allocate their efforts. In addition to their inherent appeal in terms of financial prosperity, natural resources possess a profound importance that transcends just economic considerations. These resources are pivotal in driving technological advancements, reinforcing critical infrastructure, and serving as a fundamental foundation for several indispensable aspects of modern life [5].

Minerals, serving as essential primary resources, play a crucial role in driving diverse sectors and facilitating the progress and prosperity of nations [6]. The fundamental significance of their existence resonates profoundly inside every aspect of contemporary society, including both commonplace technological devices and grand-scale infrastructural developments. The growing global need for resources, driven by population growth and technological advancements, highlights the importance of accurately identifying and efficiently exploiting mineral-rich areas. Despite the widespread distribution of mineral resources in several countries, their exact locations often need to be clarified [7]. In this particular scenario, a tool that demonstrates the ability to assess the probability of mineral existence transcends its ordinary value and becomes a critical role. Therefore, improving mineral prospect mapping is both a technological challenge and a socioeconomic requirement.

Although drilling approaches have the potential to provide precise outcomes, their implementation requires substantial time and expense commitments, resulting in a restricted availability of data [8]. On the other hand, airborne geophysical techniques, including magnetic, gravitational, and electromagnetic field measurements, provide extensive spatial coverage [9]. However, these methods frequently compromise resolution and precision. The previously mentioned contradictions highlight the need to use visualization and classification methodologies. Visualization methods illustrate complex patterns and relationships within data that may otherwise stay obscured [10]. Simultaneously, classification aims to develop prediction algorithms to identify potential mineral locations using existing knowledge about deposit characteristics [11].

The concept of prospectivity is of great significance in identifying and exploring mineral deposits in diverse geological landscapes. This methodology assesses the capacity to extract commercially viable minerals from specified regions. Prospectivity comprises two basic categories [12]:

- **Knowledge-driven:** It is based on the theoretical background of how mineral deposits form.
- **Data-driven:** In this approach, various data-analysis methods are used to explore spatial data and identify patterns associated with mineral deposits.

The use of modern computational techniques in mineral prospectivity mapping represents a significant change in approach, considering the inherent trade-offs associated with geological methodologies [13]. The deployment of machine learning and powerful data visualization techniques has sparked a fresh drive to automate and enhance interpreting information [14]. Machine learning, characterized by its diverse range of algorithms, acquires knowledge from existing data, providing valuable insights into previously unexplored domains [13] [15].

Despite the ability of machine learning in this domain, there are fundamental obstacles. Mineral-rich zones are minimal, resulting in datasets exhibiting a disproportionate abundance of negative samples (areas without mineral occurrences) compared to positive examples [16]. Such biases may lead to the distortion of machine learning models, resulting in a tendency to inaccurately forecast regions as deficient in minerals inaccurately, even when this is not the actual scenario [17]. Conventional two-dimensional mapping visualizations must be more comprehensive in capturing the complexities and interdependencies of geophysical attributes [18].

Mapping mineral prospectivity is crucial for exploring resources. Nevertheless, the current study has highlighted some issues that need addressing. The conventional approaches used in mineral mapping, even though reliable, require considerable time and effort, highlighting the need for improved and efficient methodologies. There is a notable disparity in the distribution of data, with a substantial number of locations without minerals compared to those high in minerals. This imbalance can result in biased outcomes, an issue that needs to be better addressed in several research investigations. While some research emphasizes using visualization tools like Principal Component Analysis (PCA) and Self-Organizing Maps (SOM) with machine learning, a comprehensive exploration of their combined efficiency is lacking. There are further concerns about the efficacy of machine learning when dealing with imbalances in data. Furthermore, most research endeavors concentrate on either data visualization or classification individually, with little effort to combine both methodologies.

This work aims to develop a comprehensive methodology for mineral prospectivity mapping by integrating powerful machine learning algorithms with state-of-the-art data visualization methods. The central goal is to enhance the accuracy of mineral location predictions, improve the interpretation of geophysical data, and update traditional methodologies. The basic theory suggests that integrating this technology will significantly improve the accuracy and effectiveness of mineral analysis, thereby promoting more sustainable resource exploitation.

II. Related work

Integrating data science and mineral prospectivity mapping (MPM) has established a multidisciplinary linkage, enhancing the geoscience field's capabilities and accuracy [19] [20] [21]. As the field of geoscience undergoes this significant shift, it is crucial to thoroughly examine the progress made so far and the obstacles that remain to have a comprehensive understanding. This literature review aims to understand the dynamics stated above thoroughly.

Usually, the field of MPM has been predominantly influenced by two fundamental paradigms: the knowledge-driven paradigm and the data-driven paradigm [12]. The knowledge-centric methodology, derived from extensive geological expertise over many decades, provides a profound and invaluable viewpoint that has played a pivotal role in mineral exploration through several generations [22]. The successful interpretation of geological processes and their consequential mineral deposits is based upon a thorough understanding of the subject matter.

The data-driven approach stands in significant contrast to traditional methods, a characteristic of the contemporary period, driven by notable progress in computer capabilities and data analytics [23]. The approach advocated by researchers prioritizes the meticulous examination of extensive geophysical data. The growing availability of geophysical data in terms of quantity and level of detail has created a need for reliable analytical approaches. As a result, the data-driven approach has become more relevant.

The complicated and multi-dimensional nature of geoscientific data poses a distinct problem involving translating this complex information into understandable and significant forms. Data visualization approaches such as Self-Organizing Maps (SOM), Parallel Coordinate Plots (PCP), and Principal Component Analysis (PCA) are now exhibiting a significant impact on several domains, facilitating revolutionary outcomes [24] [25] [26].

Self-organizing maps (SOM), a kind of unsupervised learning, effectively decrease data's dimensionality while improving visual representation. The significance of SOM in the contemporary machine learning domain becomes apparent when integrated with algorithms such as Support Vector Machines [27]. On the other hand, Principal Component Analysis (PCA) facilitates the comprehension of information with many dimensions by representing them in more straightforward and lower-dimensional perspectives [26]. Furthermore, parallel coordinate plots enable the analysis of datasets with many dimensions by arranging them along parallel axes. This configuration reveals patterns and interconnections among distinct variables [25].

The use of machine learning in MPM is rapidly expanding. The current development is distinguished not only by the rising use of algorithms such as logistic regression and multilayer perceptrons but also by the innovative approaches employed to modify and enhance these methods to address issues particular to MPM.

Machine learning techniques, such as logistic regression and the multilayer perceptron (MLP), are often used in mineral prospectivity mapping [28] [29]. These methodologies have shown efficacy in identifying regions with high gold mineral potential via geological and geochemical data analysis. Logistic regression uses the synthetic minority oversampling method (SMOTE) to achieve a balanced distribution of classes

within a dataset. The level of balance achieved contributes to the improvement of classification accuracy. On the other hand, MLP has shown efficacy in identifying areas exhibiting mineral potential by using a comprehensive amalgamation of geological, geochemical, and geophysical data.

The literature on mineral prospectivity mapping using machine learning techniques has indicated several areas for improvement. Firstly, it is essential to note that regions abundant in minerals are not often found, leading to datasets that exhibit an imbalanced distribution of negative instances (areas lacking mineral occurrences) compared to positive cases. This phenomenon might result in the development of biased machine-learning models, leading to incorrect predictions of mineral deficiencies in some places. Secondly, conventional two-dimensional mapping visualizations cannot comprehensively depict the intricacies and interconnections of geophysical features, which can limit the interpretability and accuracy of mineral prospectivity maps. Thirdly, the majority of existing research on the use of machine learning in mineral prospectivity mapping mainly concentrates on either data visualization or classification as separate entities. There is a need to develop more comprehensive procedures that integrate both methods to improve the accuracy and comprehensibility of mineral prospectivity maps.

These gaps are critical because they can lead to inaccurate mineral prospectivity maps. Machine learning algorithms that exhibit bias may lead to incorrect predictions about the presence of minerals in some places, which can result in the inefficient allocation of time and resources towards exploration efforts in areas unlikely to result in mineral deposits. Better visualization methods can enhance the interpretability and accuracy of mineral prospectivity maps by facilitating users' comprehension of the underlying geological mechanisms dealing with mineralization. Integrating several methodologies can boost the accuracy and comprehensibility of mineral prospectivity maps by incorporating spatial patterns derived from geophysical data and including known mineral deposit sites. Addressing deficiencies in the existing research on mineral mapping will enable the development of more efficient, effective, and interpretable methodologies for exploring mineral resources more quickly and cheaply, which is essential for meeting the growing demand.

III. Dataset

The represented research area, as seen in Figure 1, represents a specific region located in Finland. The data is collected by the Geological Survey of Finland (GTK), as stated on their official website (www.gtk.fi). The principal objective within this domain is to get a comprehensive understanding of the underlying geology. The amalgamation of geological and geophysical data is used to accomplish this task. The geographical area is represented using a raster grid, where each cell has a spatial resolution of 50x50 square meters.

Geophysical data provides an in-depth description of the fundamental characteristics of the Earth, including gravitational forces, magnetic fields, and electrical resistivity. Identifying regions characterized by substantial concentrations of these attributes is of utmost importance since they may serve as indicators of significant mineral deposits. The known mineral types in this study include gold, iron, and copper. On the other hand, geological data offers an improved understanding of the underlying layers and structures, proving advantageous in identifying areas with similar geological characteristics found in well-documented mineral-rich regions. To provide a more comprehensive understanding of the data, it is interesting to note that the dataset has 17 known mineral deposit sites. Remarkably, there remains a substantial number of 1,843,564 locations that have yet to be classified as either abundant or lacking in the minerals above.

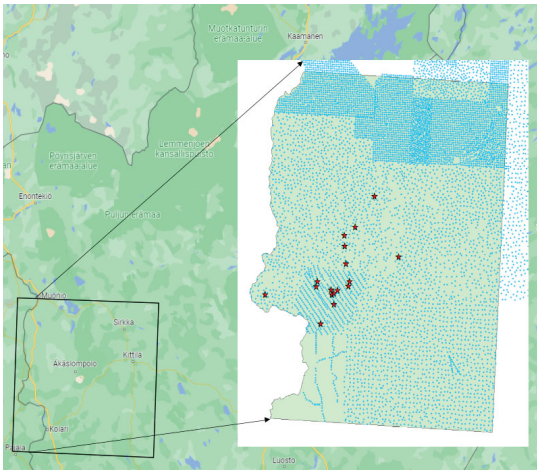

Fig. 1. Area of study (GTK, Finland)

IV. Methodology

Mineral prospectivity mapping is an essential procedure that helps to identify regions with significant economic value from mineral reserves. Prospectivity maps are often generated using geological and geophysical data, including structure and topographical maps, airborne electromagnetic, gravity, and radiometric imaging. Prospectivity mapping comprises two principal methodologies: data-driven and knowledge-driven strategies. The former employs diverse data analysis techniques to investigate geographical data and identify patterns associated with mineral deposits. At the same time, the latter is based on the theoretical framework of the formation of mineral deposits. The ultimate objective of this study is to combine sophisticated machine learning algorithms with state-of-the-art visualization methods to provide a comprehensive methodology for mineral prospectivity mapping.

This study used a mixed-method approach, which incorporates visualization and classification techniques. Visualization techniques, such as PCA, SOM, and PCP, depict intricate patterns and interconnections within data that could otherwise remain obscured. Classification techniques, such as Logistic Regression and MLP, are implemented to develop prediction

algorithms to identify potential mineral locations using knowledge about mineral deposit characteristics.

The mixed methods approach was selected due to its capacity to provide a thorough knowledge of the intricate issue of mineral prospectivity mapping through statistical analysis and visualization approaches. Moreover, using a mixed-method approach has exhibited greater efficiency than traditional procedures, which require substantial time and effort.

The following research tools and procedures were used in this study:

A. VISUALIZATION

1) SELF-ORGANIZING MAPS (SOM): The Self-Organizing Map (SOM) is an unsupervised machine-learning technique that learns how to create a low-dimensional grid structure from a high-dimensional dataset [24]. The input is connected with each unit of the lattice(map). The SOM is competitive-based learning, where each neuron competes against each other to represent various regions of the given input data space. In other words, the units called neurons in a SOM are very competitive in responding to a given input pattern. In this competition, the neuron that wins the race is the nearest to the input pattern in terms of Euclidean distance. As this algorithm is trained, the weights of the units are adjusted so that they become very similar to the inputs, and this operation of weight adjustment is iterated till the SOM has learned to represent the dataset in a low-dimensional space. In cases when there are nonlinear interactions between the features, SOM is a good alternative for dimensionality reduction.

2) PARALLEL COORDINATES PLOT: Parallel coordinates plot is a technique to display and analyze high-dimensional data that contains many variables or features [25]. It represents each observation in the data as a multiline that connects the values of its variables along a set of parallel lines. Parallel coordinate plots incorporate data preparation, scaling, plotting, interaction, and interpretation. Furthermore, histograms for each variable in the dataset are calculated for a specific bin size and plotted along the vertical axes of the parallel lines. These histograms aid the additional information about the distribution of variables along the axes and their potential impact on the overall patterns.

3) PRINCIPAL COMPONENT ANALYSIS (PCA): Principal Component Analysis is a widely used technique that reduces a dataset's dimensionality while retaining the data's variability [26]. The method involves determining the principal components, which are linear combinations of the original variables and account for the majority of the variance in the data, and then mapping the data onto these components. PCA involves standardizing the data, computing the covariance matrix of the features, performing eigendecomposition of the covariance matrix, ordering the eigenvectors in decreasing order based on the magnitude of their corresponding eigenvalues, selecting the principal components, and projecting the data onto the selected components. Additionally, PCA can be used to identify the essential features in the dataset, i.e., the features that account for most of the variation in the data correspond to the principal components with the most significant eigenvalues. When the correlations between the features are linear, PCA is a practical choice for dimensionality reduction for visualization.

B. CLASSIFICATION

1) LOGISTIC REGRESSION: Logistic regression is a descriptive statistical model used in supervised machine learning classification problems [28]. It is mainly used when the dependent (target) variable is categorical. It explains the relationship between one dependent binary variable and one or more independent variables.

We know that a simple equation for linear regression is given by

$$Y = b_0 + b_1 x \tag{1}$$

where b_0 and b_1 are the intercept and slope coefficients, respectively, x is the independent variable, and Y is the dependent variable.

Here, our goal is to model the probability of an event; we replace Y with p. As we know, the probability range varies between 0 and 1, but in this case, the range may exceed. We take the "odds" of p, the ratio of the probability of success and the probability of failure, to overcome this problem.

$$\frac{p}{1-p} = b_0 + b_1 x \tag{2}$$

Here, the range of odds varies from 0 to ∞. We can see that this range is restricted, which will limit the data points and ultimately decrease our correlation. To overcome this shortcoming, we will deploy the log of odds

$$log(\frac{p}{1-p}) = b_0 + b_1 x \tag{3}$$

Now, after some mathematical calculations, we will get

$$p = \frac{1}{1 + e^{-}(b_0 + b_1 x)} \tag{4}$$

The above equation is a logistic function, also called a sigmoidal function.

2) MULTILAYER PERCEPTRON: A multilayer perceptron is an artificial deep neural network with more than one perception [29]. It is comprised of an input layer to receive the signal, an Output layer to make a prediction, and in between these two layers, an arbitrary number of layers called hidden layers that are used for the actual computation of this multilayer perceptron.

The general block diagram of multilayer perceptron can be shown in Figure 2. In forward propagation, the signal flows from the input to the output layer through the hidden layers, and the prediction of the output layer is measured against the actual true value called error. During backpropagation, the calculated error is backpropagated to the layers of the multilayer perceptron. Using the chain rule, the gradient of the cost function is calculated concerning weights and biases.

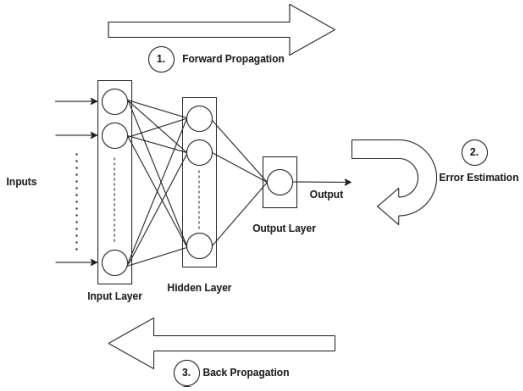

Fig. 2. Representation of Multilayer Perceptron

C. CLASS IMBALANCE HANDLING

To mitigate the problem of class imbalance, several techniques such as L2 regularisation, class weight balancing, random undersampling using Multilayer Perceptron (MLP), and Synthetic Minority Over-sampling Technique (SMOTE) with Logistic Regression have been used.

L2 regularisation is a method that can improve data balance by discouraging the model from giving excessive weight to features present in a limited subset of the data [30]. The reason for this is that L2 regularization imposes a penalty on the model on the model for assigning large weights to any given feature. This method promotes the model to allocate equal importance to each feature, improving the model's performance while dealing with unbalanced data.

Class weight balancing is a machine learning technique applied to address the problem of class imbalance. During the training phase, the model assigns individual weights to each class to contribute equally to each category [31]. The allocation of weights to individual classes is often determined in an inverse relationship to their relative frequencies within the dataset.

The Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling technique used to address class imbalance in datasets. It does this by generating synthetic instances of the minority class by interpolation between actual minority class samples. This method of balancing the dataset is shown to have a positive impact on enhancing the performance of machine learning models when dealing with imbalanced data sets. The combination of SMOTE and Logistic Regression enhances the efficacy of Logistic Regression models while dealing with unbalanced datasets [32] [33].

Random undersampling is a practical approach to address the issue of unbalanced datasets. This technique randomly eliminates instances from the majority class to get a more equal distribution of class labels [33]. This means that samples from the majority class are selected randomly, ensuring that the number of desired cases matches that of the minority class. This approach can potentially enhance the efficacy of machine learning models when dealing with unbalanced datasets, particularly in the context of MLP models.

V. IMPLEMENTATION

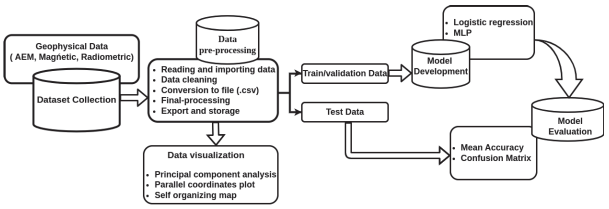

Fig. 3. The Overall proposed system architecture of mineral prospectivity mapping

The overall system for mineral prospectivity mapping is illustrated in Figure 3. The geographical data, which consisted of Airborne Electromagnetic (AEM), Magnetic, and Radiometric data, were converted into a point format and passed through the preprocessing pipeline. AEM provides information about the conductivity of the ground, which gives valuable information about the potential mineral deposits. On the other hand, Magnetic data reveal information regarding magnetic minerals. Similarly, Radiometric or gamma-ray spectrometry captures the radiation emanating from Earth's crust, which is vital in identifying a specific type of mineral deposit like potassium. These geographic data sets, when combined, provide a complete perspective of the subsurface, making it easier to map the area and potentially find mineral deposits.

A. DATA PREPROCESSING

The data processing for mineral prospectivity mapping using ArcGIS was carried out in many steps. Initially, the AEM, magnetic, and radiometric data in .tif format were imported into ArcGIS. Then, the data was cleaned to eliminate mistakes and inconsistencies, using various data analysis and cleaning techniques that were readily available. Subsequently, the data was transformed into the .csv format, a standard file format for machine learning and statistical software.

The following actions were executed to complete the final preprocessing for mineral prospectivity mapping. First of all, the identification of mineral-containing locations was conducted by manually digitizing known mineral deposits. In the meantime, class 1 was assigned to those points that contained minerals, whereas class 0 was allocated to all remaining points. Once the data had been preprocessed, it was exported into a format easily stored and accessed.

B. VISUALIZATION

Data visualization is a technique used to effectively convey information, enhance comprehension, identify patterns and trends, and support decision-making. Principal Component Analysis (PCA), Parallel Coordinate Plot (PCP), and Self-Organizing Maps (SOM) were three data visualization approaches used to extract meaningful insights from the data.

1) SELF-ORGANIZING MAP (SOM): The raster data was preprocessed using winsorization, a method for reducing the impact of extreme data values with specific percentiles to reduce their influence, and normalization to scale the data

to a standard range. The preprocessed data was then given into a SOM with a grid configuration 20x20 for training. K-means clustering facilitated partitioning the SOM into 12 distinct clusters. During this period, Best Matching Units (BMUs) were employed to allocate established mineral deposit locations to the SOM grid based on the degree of similarity with the input data points. The SOM grid is a topographical map that retains the spatial relationships in the input data. The BMU for each input data point was determined by identifying the neuron in the SOM grid with the weight vector closest to the input data point inside the feature space.

2) PARALLEL COORDINATES PLOT: In a parallel coordinates plot, each feature has its axis, which are all parallel to one another. Since each variable uses a separate unit of measurement, each axis may have a different scale. Thus, all the axes can be normalized to maintain consistency of scale. Values are plotted on a graph as a sequence of lines linking all axes. The color of the line is determined by the label of the data point, which is red for known deposit points and black for unknown mineral deposit points. The integration of histograms with bin sizes of 30 on each axis enables the user to quickly review the distribution of data points at the beginning of the process.

3) PRINCIPAL COMPONENT ANALYSIS: Principal component analysis (PCA) was used to calculate the principal components responsible for capturing the fundamental patterns in the given datasets. In this implementation, the first three principal components were computed on the provided dataset. The principal component values were rescaled and converted into the RGB format to visually depict the principal components inside the image to ensure the proper representation of color intensity. The RGB value for each pixel was calculated by computing the principal components and visualizing them. Nevertheless, the primary components' actual values were not altered or normalized throughout this procedure. This methodology facilitated the identification of intricate structures and essential patterns within the data via the visual representation of the principal components as images.

C. MACHINE LEARNING MODELS

To classify mineral deposit points, two machine learning models, namely logistic regression and multilayer perceptron, were trained using a dataset that exhibited a significant disparity between the number of known (17) and unknown (1,843,564) deposit points. To mitigate the class imbalance problem, the Synthetic Minority Over-sampling Technique (SMOTE) was used to produce synthetic data points for the logistic regression model. Furthermore, the multilayer perception model implemented the random undersampling method to balance the data from both classes. Moreover, to address the issue of class imbalance, both models used L2 regularization and balanced class weight to reduce the bias towards the majority class and improve the performance of the minority class.

1) LOGISTIC REGRESSION: The given dataset had a significant disparity in class distribution, as the minority class (known deposits) consisted of just 17 data points, while the majority class (unknown deposits) included a substantial 1,843,564 data points. To address this issue, the SMOTE technique was used to generate synthetic instances for the minority class. The Synthetic Minority oversampling technique (SMOTE) is used for oversampling in machine learning. It involves the creation of synthetic data by interpolating existing samples from the minority class. This was used to balance the dataset and make it more representative of the real world. The dataset was then divided into training and testing sets at a ratio of 0.8 to 0.2. The model performed training with 80% of the available data, while the remaining 20% were used for testing.

The logistic regression method used the balanced class weight parameter to address the class imbalance problem during training. This parameter was utilized to assign a weight to each class, facilitating the algorithm's assigning more significance to the minority class throughout the training process. Furthermore, L2 regularization was used to mitigate the overfitting of the model to the training data and enhance its ability to generalize to unseen data.

During the training process, the model iteratively updated its coefficients to minimize the logistic loss function, measuring how well the model could predict the appropriate class for a given data point. After the completion of the training process, the model was used to predict the test dataset. The model's accuracy was evaluated by comparing the predicted and actual values in the test set.

2) MULTILAYER PERCEPTRON: To use a multilayer perceptron (MLP) model for binary classification, the initial step involved dividing unknown mineral data points into 108,445 distinct clusters. Each cluster was comprised of 17 mineral points that were randomly chosen. During each training iteration, a set of 17 points was selected from the class of known mineral deposits. One cluster of 17 randomly chosen points was also selected from the 108,445 unknown mineral deposit groups. This process created a balanced training set for each iteration to address the class imbalance issue.

To address the issue of class imbalance, the MLP model included balanced class weights. This meant more weight was given to the minority class (known mineral deposits) during training. This helped to ensure that the model did not overfit to the majority class (unidentified mineral deposits). Furthermore, the implementation employed L2 regularisation, which imposes a penalty on the model for using large weights. This approach helped in mitigating the issue of model complexity and overfitting to the training data.

The MLP model's performance was evaluated using leave-one-out cross-validation (LOOCV). In the Leave-One-Out Cross-Validation (LOOCV) technique, each data point was used once as the test set, while the remaining data points were utilized for training purposes. The process was repeated for each data point, and the mean performance accuracy was computed across all iterations. The hyperparameter tuning process was performed to identify the optimal values for the hyperparameters of the MLP model. Three essential hyperparameters were identified: the number of neurons, the L2

regularisation value, and the learning rate. The hyperparameter space was established, and afterward, the model was trained and assessed for every possible combination of hyperparameters. The hyperparameter combination that resulted in the best performance on the leave-one-out cross-validation (LOOCV) set was identified as the most optimum.

VI. Results and Discussion

The main objective of this study was to explore the implementation of visualization and classification techniques in the context of mineral prospectivity mapping. Identifying the complex patterns and relationships within geological and geophysical data can result in more effective identification of areas with a high probability of mineral potential.

The results obtained from this study not only effectively answered our original research questions but also had notable implications for geophysics. The research found that visualization techniques can be used to identify trends in data that were previously unnoticed, while classification methods may be used to estimate the possibility of minerals in particular locations. The findings of this study can significantly transform the methodology used in mineral prospectivity mapping, perhaps resulting in the identification of previously unnoticed mineral deposits.

A. VISUALIZATION

The significance of visualization in the study of geophysical data is pivotal. The function of this entity is to serve as a conduit, facilitating the transformation of complex data into valuable information. The findings of this investigation clearly demonstrated that the visualization approaches used effectively revealed major trends and relationships. These methods enabled an in-depth comprehension of the data and helped the reduction of its dimensionality, retaining the most relevant features. Furthermore, these techniques helped us to visualize multiple data points across various dimensions, which unraveled their complex relationships.

1) SELF-ORGANIZING MAP (SOM): The geophysical data shown in Figure 4 was subjected to the combined SOM and K-means clustering technique. The data were presented in a two-dimensional grid, whereby each point corresponded to a specific data point. The circles on the map denote the known occurrences of minerals.

The SOM method first transformed the data, resulting in a two-dimensional representation. This was done by building a map of neurons, where each neuron was associated with a specific data region. The neurons were then structured in a two-dimensional grid, whereby neurons showing similarities were positioned close to one another. The K-means algorithm then further clustered the data into clearly defined categories. To do this, the data points were randomly assigned to one of the K clusters. The clusters were further modified via an iterative process to maximize the similarity between data points inside each group.

The reference figure illustrates the potential of using the combined SOM and K-means clustering technique to conduct mineral prospectivity mapping effectively. The methodology effectively identified clusters of data points associated with known mineral occurrences, indicating its potential applicability in identifying regions with a high mineralization probability.

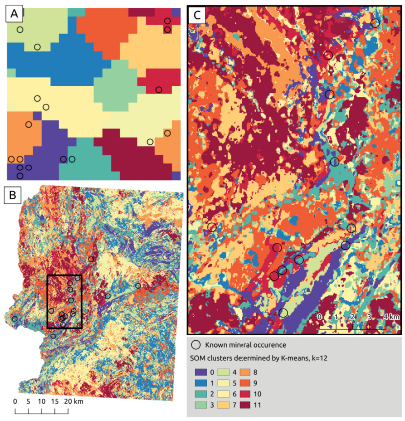

Fig. 4. Self-Organizing Map (SOM), and K-means Clustering, Highlighting the Role of Best Matching Units (BMUs) in Mapping Known Mineral Deposits

2) PRINCIPAL COMPONENT ANALYSIS (PCA): Principal component analysis (PCA) was conducted on a set of ten distinct raster layers, whereby each layer comprised information on the radioactivity, electrical characteristics, and magnetic within the specified region. The resulting raster is depicted in Figure 5. The first principal component (PC1) was assigned to the red band, the second principal component (PC2) to the green band, and the third component (PC3) to the blue band. Since the principal components lack correlation, the resulting raster is very colorful. This indicated that each principal component uniquely captures a particular aspect of the data.

Using color to represent the data makes previously hidden relationships and patterns easier to spot. The known mineral reserves, for instance, are shown with white circles. These rings are situated in an area of the raster where PC1 and PC2 have high values. This implies that the underlying geology of these ore deposits exhibits specific patterns of variability. The PCA outcomes were used to identify regions with a high mineralization probability. Regions exhibiting high values of both PC1 and PC2 indicated potential areas for the occurrence of ore resources. Nevertheless, it is crucial to acknowledge that PCA is a statistical methodology that does not guarantee the existence of mineralization. Additional investigation and examination are needed to validate the existence of mineral resources.

3) PARALLEL COORDINATES PLOT (PCP): The data in Figure 6 was visualized using a parallel coordinates plot (PCP). A PCP is a graphical representation that visually displays data points in a multidimensional space, with each dimension corresponding to a distinct axis. The data points are assigned distinct colors, namely black to represent deposits of unknown deposits and red to represent known deposits.

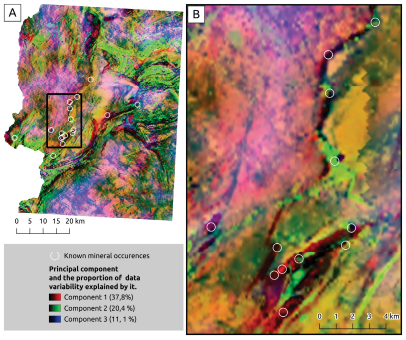

Fig. 5. PCA visualization with geophysical data

The PCP demonstrates a statistically significant correlation between the two dimensions. The known deposits have been located in the upper-right region of the map, whereas unknown deposits are mostly grouped in the lower-right region. This observation implies that the two categories of deposits exhibit distinct features about their values along the two axes. The histograms shown along the axes provide further insights into the distribution patterns of the features along the vertical axes.

Overall, the PCP offers a technique for visualizing the data and identifying relationships between the various dimensions. To facilitate further analysis, the histograms along the vertical axes provide extra information about the distribution of the features.

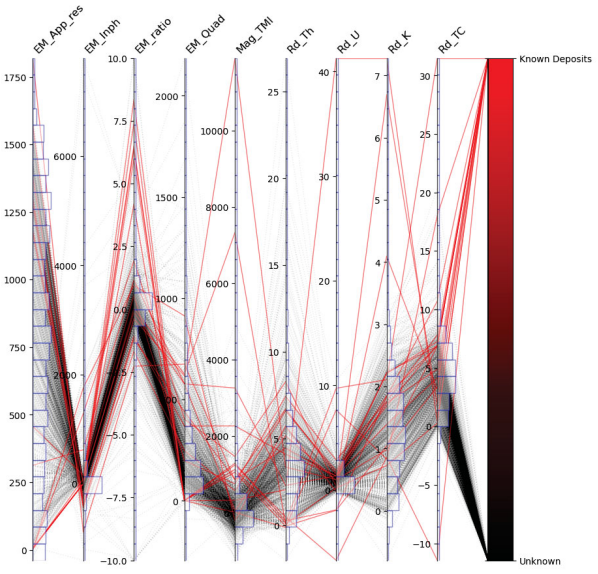

Fig. 6. Histograms are placed on top of the parallel coordinates plot to display the item distribution along each axis

B. CLASSIFICATION

1) LOGISTIC REGRESSION: A logistic regression algorithm was used to classify the data points into two classes: unknown mineral deposits (class 0) and known mineral deposits (class 1). The collection consisted of 17 known mineral deposits and 1,843,564 unknown mineral deposits. The logistic regression model was first trained on the unbalanced dataset. The dataset was divided into training and testing sets, following a ratio of 80:20. A subset comprising 1,474,864 data points, accounting for 80% of the original dataset, was used for training. The remaining 20% of the total dataset, specifically 368,717 data points, were reserved for testing.

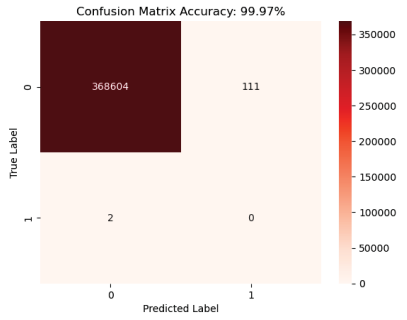

Fig. 7. Consusion matrix for logistic regression on imbalanced dataset

TABLE I. PERFORMANCE EVALUATION METRICS OF LOGISTIC REGRESSION ON IMBALANCED DATASET

Class	Precision	Recall	F1 Score
0 (Unknown Deposits)	1	1	1
1 (Known Deposits)	0	0	0

Figure 7 shows the confusion matrix for the logistic regression model on the unbalanced dataset. The confusion matrix shows that the model has an accuracy of 99.7% but misclassifies the minority class. Table 1 illustrates the metrics used for evaluating the efficacy of logistic regression on an imbalanced dataset. The table consists of two separate classes: unknown deposits (class 0) and known deposits (class 1). The precision, recall, and F1-score for class 0 in the table are all 1, although all the values for class 0 are 0. This is because the logistic regression model is biased towards the majority class (unknown mineral deposits).

To address this class imbalance issue, the Synthetic Minority Over-sampling approach (SMOTE), a data-balancing approach, was used before the model's training. SMOTE generates synthetic data points that resemble those from the minority class. This approach proves useful in achieving a more balanced distribution within the dataset. Implementing the Synthetic Minority Over-sampling Technique (SMOTE) demonstrated its efficacy, leading to a substantial increase in the dataset size to 3,687,120 instances. Consequently, both the known and unknown mineral deposit classes now possess an equal distribution of 1,843,560 instances each. After using SMOTE to balance the dataset, it was divided into training and testing subsets with proportions of 80% and 20%, respectively. The objective of this technique was to boost the performance of the logistic regression model by exploiting a balanced dataset.

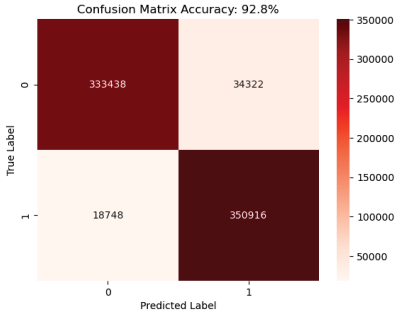

Fig. 8. Confusion matrix for logistic regression on balanced (SMOTE) dataset

TABLE II. PERFORMANCE EVALUATION METRICS OF LOGISTIC REGRESSION ON BALANCED (SMOTE) DATASET

Class	Precision	Recall	F1 Score
0 (Unknown Deposits)	0.95	0.91	0.93
1 (Known Deposits)	0.91	0.95	0.93

The confusion matrix for the logistic regression model that was trained using balanced data is shown in Figure 8 with an accuracy of 92.8%. Table 2 provides the precision, recall, and f1-score metrics for class 0 and class 1. In the context of class 0, the accuracy, recall, and f1-score values are 0.95, 0.91, and 0.93, respectively. For class 1, the precision, recall, and f1-score values are 0.91, 0.95, and 0.93, accordingly. The results indicate that the logistic regression model improved performance after using the SMOTE technique to balance the dataset.

In mineral prospectivity mapping, we are mainly concerned with the points that contain a high probability of minerals. It is widely known that mineral deposits are found at recognized deposit locations due to extensive exploration and drilling activities conducted in certain areas. The mineral composition of unknown deposit locations remains to be determined due to the lack of exploration activities undertaken so far. Hence, this study focuses on identifying false positive data points that indicate a high probability of mineral occurrence. When trained on the balanced dataset, the logistic regression model exhibits a high precision, indicating a lower chance of predicting a false positive. This characteristic makes it a good choice for use in mineral prospectivity mapping.

2) MULTILAYER PERCEPTRON: MLP analysis was used to understand its strengths and weaknesses and its potential application in mineral prospectivity analysis. Initially, only the original data was utilized, addressing the class imbalance issue within the dataset using the random undersampling technique. In the meantime, a multilayer perceptron (MLP) was used for classification prediction in two different settings:

- Without balanced class weights and L2 regularization: The model acts as a baseline model, which did not account for the class imbalance problem within the dataset as well as the potential danger of overfitting. This model was used as a reference model to compare the performance of other models.
- With balanced class weights and L2 regularization: The proposed approach aims to mitigate the issue of class imbalance by providing higher weights to the minority classes. Additionally, L2 regularisation is used to mitigate the problem of overfitting. It is anticipated that this model would exhibit superior performance compared to the baseline model, particularly in the minority classes.

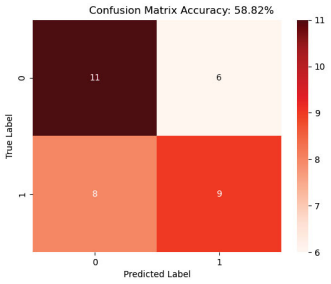

Fig. 9. Confusion matrix for MLP without balanced class weights and L2 regularization

The performance of the MLP was evaluated using the leave-one-out cross-validation technique. The mean accuracy of the model without class weights and L2 regularization is 58.82%, as shown in Figure 9.

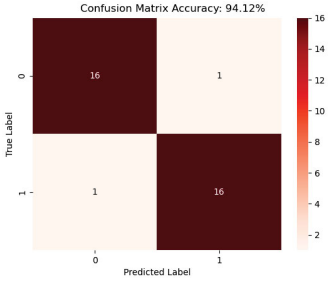

Fig. 10. Confusion matrix for MLP with balanced class weights and L2 regularization

On the other hand, the average accuracy of the algorithm with L2 regularization and balanced class weights was 94.12%, as shown in Figure 10. The findings indicated that the MLP model with the balanced class weights and L2 regularisation showed superior performance compared to the method that did not use balanced class weights and L2 regularisation.

Hyperparameter tuning has been performed to identify the optimum values of the number of neurons, the L2 regularisation, and the learning rate. Figure 11 shows that the maximum accuracy of 94.12% was achieved when the learning rate was 0.01, the number of neurons was 16, and the L2 regularization value was 0.001. These outcomes indicate that the use of class weights and L2 regularisation in the MLP model has significant potential as an effective method for mineral prospectivity mapping.

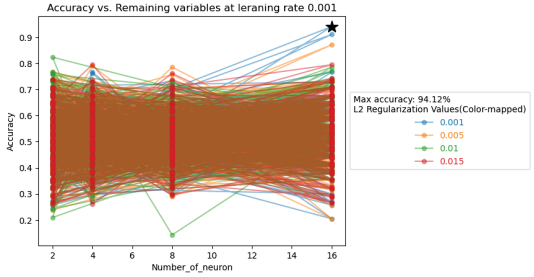

Fig. 11. Accuracy of MLP model with varying number of neurons and L2 regularization for given learning rate

VII. Conclusion

In conclusion, this study included a multi-technique approach for geological data analysis, focusing on the visualization and classification of geophysical data to identify the likelihood of potential mineral deposits.

The outcomes illustrated the significance of visualization methods in many respects. At first, the visualization helped to understand trends and patterns in the data to identify potential mineral trends and geological structure. Next, by reducing the dimensionality of the data, the visualization approach presented a clear interpretation of the complex geological attributes. Finally, visualizing data along the multiple dimensions depicted a clear understanding of complex relationships between geological parameters.

One vital aspect of geological data analysis is the handling of data imbalance, which is common in mineral prospectivity mapping. SMOTE and random undersampling strategies were implemented effectively to address this issue. The logistic regression model's performance on highly imbalanced data improved significantly by implementing these techniques, balance class weights, and stratified cross-validation. Moreover, the study further addressed the data imbalance problem by deploying the multilayer perceptron (MLP) method with L2 regularization and class weights.

Overall, the findings highlight the importance of an integrated approach for practical geological data interpretation and exploration of mineral deposits, integrating specialized methodologies and data handling strategies. By utilizing these approaches, one may enhance the efficacy and efficiency of their exploration efforts, eventually leading to more successful mineral prospecting mapping.

Acknowledgment

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).

References

[1] He, Binbin, et al. "Mineral prospectivity mapping method integrating multi-sources geology spatial data sets and case-based reasoning." (2012).

[2] Knox-Robinson, C. M., and L. A. I. Wyborn. "Towards a holistic exploration strategy: using geographic information systems as a tool to enhance exploration." Australian journal of earth sciences 44.4 (1997): 453-463.

[3] Harris, J. R., et al. "Application of GIS processing techniques for producing mineral prospectivity maps—a case study: mesothermal Au in the Swayze Greenstone Belt, Ontario, Canada." Natural Resources Research 10 (2001): 91-124.

[4] Yousefi, Mahyar, et al. "Data analysis methods for prospectivity modelling as applied to mineral exploration targeting: State-of-the-art and outlook." Journal of Geochemical Exploration 229 (2021): 106839.

[5] Dewulf, Jo, et al. "Rethinking the area of protection "natural resources" in life cycle assessment." Environmental science and technology 49.9 (2015): 5310-5317.

[6] Christmann, Patrice. "Mineral resource governance in the 21st century and a sustainable European Union." Mineral Economics 34 (2021): 187-208. 161-184.

[7] Manning, D. A. C. "Assessment of mineral deposits." Introduction to Industrial Minerals (1995): 200-226.

[8] Liu, Cancan, et al. "An experimental investigation into the borehole drilling and strata characteristics." Plos one 16.7 (2021): e0253663.

[9] Siemon, Bernhard, et al. "Airborne electromagnetic, magnetic, and radiometric surveys at the German North Sea coast applied to groundwater and soil investigations." Remote Sensing 12.10 (2020): 1629.

[10] Post, Frits H., Gregory Nielson, and Georges-Pierre Bonneau, eds. "Data visualization: The state of the art." (2002).

[11] Nathwani, Chetan L., et al. "Machine learning for geochemical exploration: classifying metallogenic fertility in arc magmas and insights into porphyry copper deposit formation." Mineralium Deposita 57.7 (2022): 1143-1166.

[12] Harris, J. R., et al. "Data-and knowledge-driven mineral prospectivity maps for Canada's North." Ore Geology Reviews 71 (2015): 788-803.

[13] Shirmard, Hojat, et al. "A review of machine learning in processing remote sensing data for mineral exploration." Remote Sensing of Environment 268 (2022): 112750.

[14] Wang, Qianwen, et al. "A survey on ML4VIS: Applying machine learning Advances to data visualization." IEEE transactions on visualization and computer graphics 28.12 (2021): 5134-5153.

[15] Zuo, Renguang, and Emmanuel John M. Carranza. "Machine Learning-Based Mapping for Mineral Exploration." Mathematical Geosciences (2023): 1-5.

[16] Nolen, Matthew S., et al. "Predicting probability of occurrence and factors affecting distribution and abundance of three O zark endemic crayfish species at multiple spatial scales." Freshwater Biology 59.11 (2014): 2374-2389.

[17] Berrar, Daniel. "Cross-Validation." (2019): 542-545.

[18] Joly, Aurore, et al. "Mineral systems approach applied to GIS-based 2D-prospectivity modelling of geological regions: Insights from Western Australia." Ore Geology Reviews 71 (2015): 673-702.

[19] Sun, Tao, et al. "GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China." Ore Geology Reviews 109 (2019): 26-49.

[20] Xiong, Yihui, Renguang Zuo, and Emmanuel John M. Carranza. "Mapping mineral prospectivity through big data analytics and a deep learning algorithm." Ore Geology Reviews 102 (2018): 811-817.

[21] Ma, Xiaogang. "Data science for geoscience: Recent progress and future trends from the perspective of a data life cycle." (2023).

[22] Aljamel, Abduladem, et al. "Smart information retrieval: Domain knowledge centric optimization approach." IEEE Access 7 (2018): 4167-4183.

[23] Zhang, Steven E., et al. "Towards a fully data-driven prospectivity mapping methodology: A case study of the Southeastern Churchill Province, Québec and Labrador." Artificial Intelligence in Geosciences 2 (2021): 128-147.

[24] Bigdeli, Amirreza, Abbas Maghsoudi, and Reza Ghezelbash. "Application of self-organizing map (SOM) and K-means clustering algorithms for portraying geochemical anomaly patterns in Moalleman district, NE Iran." Journal of Geochemical Exploration 233 (2022): 106923.

[25] Heinrich, Julian, and Daniel Weiskopf. "State of the Art of Parallel Coordinates." Eurographics (state of the art reports) (2013): 95-116.

[26] Bro, Rasmus, and Age K. Smilde. "Principal component analysis." Analytical methods 6.9 (2014): 2812-2831.

[27] Ismail, S., A. Shabri, and R. Samsudin. "A hybrid model of self-organizing maps and least square support vector machine for river flow

forecasting." Hydrology and Earth system sciences 16.11 (2012): 4417-4433.
[28] Chang, Liheng. "Imbalance data logistic regression for mineral prospectivity mapping." AGU Fall Meeting Abstracts. Vol. 2019. 2019.
[29] Brown, Warick M., et al. "Artificial neural networks: a new method for mineral prospectivity mapping." Australian journal of earth sciences 47.4 (2000): 757-770.
[30] Kamalov, Firuz, and Ho Hon Leung. "Deep learning regularization in imbalanced data." 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI). IEEE, 2020.
[31] Ivanov, Ivan, Borislava Toleva, and Nikolay Netov. "Modified Training Models Based on Medical Data Sets." 2022 57th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST). IEEE, 2022.
[32] Rahim, A.H.A., Rashid, N.A., Nayan, A. and Ahmad, A.R., 2019. Smote approach to imbalanced dataset in logistic regression analysis. In Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017) Transcending Boundaries, Embracing Multidisciplinary Diversities (pp. 429-433). Springer Singapore.
[33] Elhassan, T., and M. Aljurf. "Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method." Global J Technol Optim S 1 (2016): 2016.