JOURNAL AFRICAIN DES SCIENCES
VOL. 3, **3** (2026) 61-72

https://jafrisci.org

ISSN (online) : 3107-0949 ; ISSN (print) : 3107-0930

OPEN ACCESS

RESEARCH ARTICLE

Article Info.:
Received : 18th February 2026
Accepted : 30th March 2026
Published : 03nd April 2026

GEOSCIENTIFIC DATA FUSION AND MACHINE LEARNING FOR MINERAL PROSPECTIVITY MAPPING: A CASE STUDY OF BROKEN HILL-TYPE DEPOSITS

Michel Tshimanga Ntumba [1,*], Gedeon Muhenge Kapena [1], Celestin Tshimanga Kamanga [1], Rostin Mabela Makengo [1]

https://doi.org/10.70237/jafrisci.2026.v3.i3.07

Résumé

La cartographie du potentiel minéral de Broken Hill (Australie) a été réalisée par l'intégration de données géologiques, géophysiques et géochimiques au sein d'un Système d'Information Géographique couplé à des méthodes de Machine Learning. Les données lithologiques, structurales, gravimétriques, radiométriques et géochimiques (zinc, plomb, argent, fer, manganèse, sodium et potassium) ont été harmonisées spatialement, rasterisées et fusionnées avec des indicateurs minéraux Pb–Zn–Ag servant de couche cible. Un maillage régulier a permis d'extraire des caractéristiques par cellule, notamment la distance aux failles, la densité de linéaments et la proportion de roches porteuses de minéraux. Deux modèles supervisés, Random Forest et XGBoost, ont été entraînés, comparés et évalués à l'aide de matrices de confusion et d'analyses d'importance des variables. Les résultats montrent une meilleure performance du Random Forest et mettent en évidence le rôle déterminant de la distance aux failles et des lithologies riches en gahnite et grenat. Le modèle Random Forest a atteint une précision de 0,892, tandis que XGBoost a atteint 0,876. Le RF a démontré une précision supérieure, tandis que le XGBoost a montré une meilleure rappel, soulignant un compromis entre fiabilité et capacité de détection. La carte finale de probabilité minérale identifie des zones prioritaires cohérentes avec le cadre métallogénique régional, démontrant l'intérêt de cette approche pour optimiser le ciblage minier.

Mots-clés: Cartographie du potentiel minéral, systèmes d'information géographique, apprentissage automatique, Random Forest, XGBoost, Broken Hill

Abstract

Mineral prospectivity mapping in Broken Hill (Australia) was conducted by integrating geological, geophysical, and geochemical data within a Geographic Information System coupled with Machine Learning techniques. Lithological, structural, gravity, radiometric, and geochemical datasets (zinc, lead, silver, iron, manganese, sodium, and potassium) were spatially harmonized, rasterized, and merged with Pb–Zn–Ag indicator minerals used as the target layer. A regular grid enabled feature extraction per cell, including distance to faults, lineament density, and proportion of mineral-bearing rocks. Two supervised models, Random Forest and XGBoost, were trained, compared, and evaluated using confusion matrices and feature importance analysis. Results show better performance of Random Forest and highlight the dominant role of distance to faults and gahnite- and garnet-rich lithologies. The Random Forest model achieved an accuracy of 0.892, while XGBoost reached 0.876. RF demonstrated higher precision, whereas XGBoost showed better recall, highlighting a trade-off between reliability and detection capability. The final mineral probability map delineates priority targets consistent with the regional metallogenic framework, demonstrating the effectiveness of this approach for optimized mineral exploration.

Keywords: Mineral prospectivity mapping, geographic information systems, machine learning, Random Forest, XGBoost, Broken Hill

1. Introduction

1.1 General context

Mineral exploration is currently undergoing a paradigmatic transformation driven by the combined effects of the digitalization of geosciences, the increasing availability of geospatial data, and the rapid advancements in artificial intelligence (AI). This evolution is taking place within a global

Corresponding author : *michel.tshimanga@unikin.ac.cd (**M. T. Ntumba**)*

Copyright : © The Author(s) Published under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0)

[1] *Department of Mathematics, Statistics and Computer Science, University of Kinshasa, Dem. Rep. of Congo*

context of energy transition characterized by a growing demand for critical metals such as copper, cobalt, lithium, zinc, lead, silver, and rare earth elements, which are essential for the development of low-carbon technologies, electric batteries, smart energy grids, and advanced digital infrastructures (Lou & Liu, 2022; Shirmard et al., 2022).

At the same time, outcropping and easily accessible deposits are progressively being depleted, forcing exploration activities to target deeper, geologically complex, and spatially diffuse mineralizations. This shift leads to a significant increase in geological uncertainty, operational costs, and financial risks associated with exploration campaigns, while simultaneously reinforcing environmental and societal requirements related to natural resource exploitation (Zhang et al., 2025).

In this context, Mineral Prospectivity Mapping (MPM) emerges as a strategic decision-support tool for identifying the most favorable areas for mineralization through the integration of multisource geoscientific data. By combining geological, geochemical, geophysical, and structural information within a coherent geospatial framework, MPM helps reduce exploration uncertainty and optimize resource allocation in mineral exploration programs.

Furthermore, resource-rich countries such as the Democratic Republic of the Congo (DRC), which hosts some of the world's largest reserves of copper and cobalt, are at the center of these challenges. Despite this strategic mineral potential, the management of geoscientific data and the reduction of exploration-related risks remain major challenges. These issues highlight the need for advanced quantitative approaches capable of effectively integrating multisource data and producing scientifically robust and reproducible predictions (Farahnakian et al., 2024).

Despite recent advances in machine learning for mineral prospectivity mapping, several limitations remain, particularly regarding the integration of heterogeneous geoscientific datasets into unified predictive frameworks, the reproducibility of workflows, and the interpretability of model outputs in complex geological environments. This study addresses these challenges by proposing a structured GIS–ML framework applied to the Broken Hill district

Although this study focuses on Broken Hill, the methodological framework is designed to be transferable to data-rich and data-scarce metallogenic contexts, including those of Central Africa.

1.2 State of the art of mineral potential mapping

Historically, mineral exploration primarily relied on qualitative approaches based on the expertise of geologists, visual interpretation of geological maps, and the analysis of geoscientific data using simplified statistical methods. Although these approaches have led to numerous significant discoveries, they present notable limitations in terms of subjectivity, reproducibility, and their ability to model complex nonlinear relationships between geological, geochemical, and geophysical variables.

Traditional mineral prospectivity mapping methods, such as Weights of Evidence, logistic regression, and multicriteria approaches, have long served as important tools for spatial analysis in mineral exploration. However, these approaches often rely on simplifying assumptions and struggle to capture multi-scale spatial interactions and nonlinear dependencies characteristic of mineralized systems (Sun et al., 2019).

With the rise of Big Data and artificial intelligence technologies, data-driven approaches have experienced rapid development in geosciences. The integration of Geographic Information Systems (GIS), remote sensing, and Machine Learning (ML) techniques has profoundly transformed practices in mineral exploration. GIS enables the coherent integration of heterogeneous data from different disciplines, while Machine Learning algorithms facilitate the identification of complex spatial patterns and the probabilistic prediction of mineralization at high resolution (Lou & Liu, 2022).

Algorithms such as Random Forest, XGBoost, Support Vector Machines, and neural networks have demonstrated superior capability in modeling nonlinear relationships and handling large and noisy datasets, thereby improving the performance of predictive models in mineral prospectivity mapping (Farahnakian et al., 2024).

1.3. Scientific Gap

Despite these advances, existing studies still face limitations in integrating heterogeneous datasets within a unified spatial framework, ensuring reproducibility, and improving model interpretability in complex geological settings

Study	Method	Data Type	Limitation
Sun et al. (2019)	*Logistic Regression*	*Geochemical*	*Linear assumption*
Farahnakian et al. (2024)	*RF / SVM*	*Multi-source*	*Limited interpretability*
Lou & Liu (2022)	*ML + GIS*	*Remote sensing*	*Weak structural integration*

Unlike previous studies, this work emphasizes the integration of structural, lithological, geochemical, and geophysical variables within a unified spatial framework and evaluates multiple supervised learning algorithms under consistent preprocessing conditions

1.4. Rationale for Study Site Selection

The Broken Hill mining district, located in the Curnamona metallogenic province in New South Wales (Australia), is one of the most important polymetallic deposits in the world. This district is particularly known for its Broken Hill-type (BHT) mineralization, characterized by significant concentrations of lead, zinc, and silver (Pb–Zn–Ag) associated with Proterozoic metamorphic rocks.

Since its discovery in the late 19th century, Broken Hill has undergone intensive exploitation and extensive geological exploration. Despite the abundance of available geological, geochemical, and geophysical data, the potential for identifying satellite deposits or deep extensions of mineralization remains partially underexplored using traditional methods.

The availability of detailed geoscientific databases therefore makes Broken Hill a particularly relevant natural laboratory for testing and evaluating the effectiveness of artificial intelligence-based predictive mineral prospectivity mapping approaches.

1.5. Problem and objectives of the study

Despite recent methodological advances, several scientific challenges persist in mineral prospectivity mapping. One of the main issues concerns the effective integration of heterogeneous geoscientific data particularly geological, geochemical, and geophysical data into a unified analytical framework capable of capturing the complex relationships controlling mineralization distribution.

In this context, the following research questions arise:

- How can heterogeneous data (geology, geochemistry, geophysics) be effectively integrated into a unified analytical framework that enables robust modeling of mineral prospectivity ?
- Can supervised learning models, particularly Random Forest and XGBoost, outperform traditional methods in predicting the spatial distribution of mineral prospectivity in a complex geological setting such as Broken Hill ?

To address these questions, the main objective of this study is to develop and compare predictive models based on machine learning in order to produce a new high-resolution mineral prospectivity map and to identify areas favorable for exploration.

More specifically, this research aims to integrate geological, geochemical, and geophysical data within a unified GIS environment, to train and compare the performance of Random Forest and XGBoost algorithms, and to generate a probabilistic mineral prospectivity map that enables the identification of new priority targets for future exploration. This methodological approach also aims to propose a reproducible and transferable framework applicable to other metallogenic contexts, particularly in the Democratic Republic of the Congo (Lou & Liu, 2022; Farahnakian et al., 2024).

2. Study area and geological setting

2.1. Study area

The study area corresponds to the Broken Hill mining district, located in southwestern New South Wales (Australia), within the Curnamona metallogenic province.

This district is globally recognized for its major polymetallic deposits of lead (Pb), zinc (Zn), and silver (Ag), which have been exploited since 1888. It is part of the Willyama Supergroup, a Paleoproterozoic basin that has undergone intense deformation and metamorphism (Figure 1).

The semi-arid climate, characterized by low precipitation and sparse vegetation, favors the use of remote sensing and geospatial analyses. The relatively flat geomorphology and widespread rock outcrops facilitate the integration of geological, structural, and satellite data.

The availability of detailed geospatial databases makes Broken Hill an ideal test site for evaluating artificial intelligence models applied to predictive mineral prospectivity mapping.

Figure 1. *General map of the Broken Hill study area*

2.2. Geological Context

The Broken Hill deposit is hosted within the upper part of the Broken Hill Group of the Paleoproterozoic Willyama Supergroup. It corresponds to a metamorphosed SEDEX deposit that was extensively reworked during the Olarian Orogeny (approximately 1600–1500 Ma). Figures 2 and 3 present the regional geological maps of the Broken Hill district as well as the stratigraphic column of the Willyama Supergroup.

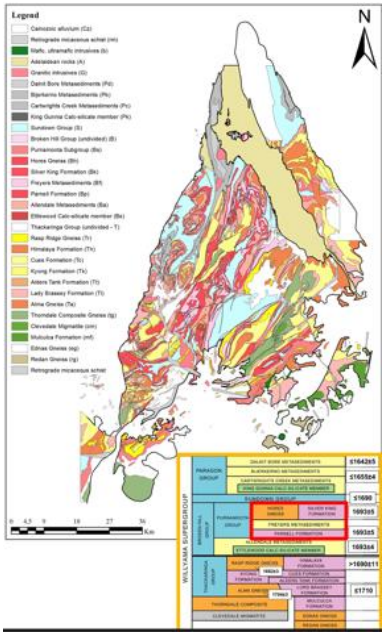

Figure 2. *Regional Geologic Map of the Broken Hill District*

The geological evolution comprises three main phases: (1) a pre-Olarian phase marked by rifting and sedimentation within a back-arc half-graben, associated with bimodal volcanism and synsedimentary metal concentration; (2) an Olarian phase characterized by regional folding, shearing, and granulite-facies metamorphism, responsible for metal redistribution; and (3) a post-Olarian phase involving local remobilization and the formation of mineralized lenses.

Lithologically, the mineralization is associated with gneisses, quartzites, and pelitic schists (Hores Gneiss, Silver King, Freyers, Parnell). Structurally, folds and NE–SW-trending shear zones control the distribution of ore bodies.

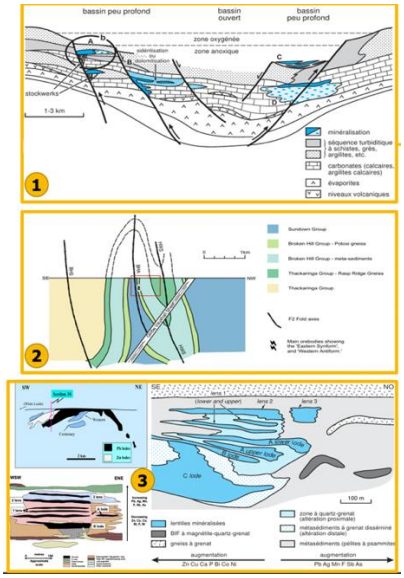

Figure 3. *Geological setting and mineralization model of the Broken Hill District*

2.2.1. Regional lithostratigraphic framework

The regional lithostratigraphic framework of Broken Hill is part of the upper section of the Willyama Supergroup, dominated by the Broken Hill Group and its associated units. The stratigraphy is structured into several high-grade metamorphic formations, including the Hores Gneiss, the Silver King Formation, the Freyers Metasediments, and the Parnell Formation (Figure 4).

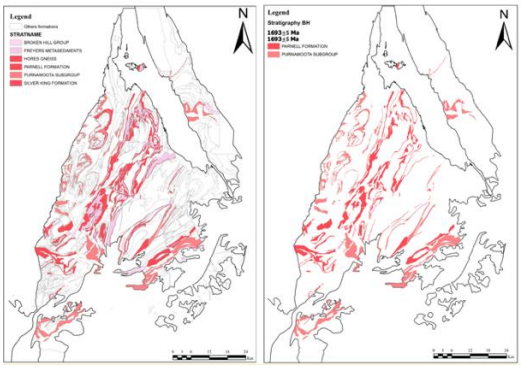

Figure 4. *Lithostratigraphic Mapping of Broken Hill*

These units are primarily composed of gneisses, quartzites, and pelitic metasediments, which are closely associated with Pb–Zn–Ag mineralization zones. The corresponding figure highlights the spatial distribution of these formations and their structural organization into elongated bands, controlled by major regional folds and shear zones associated with the Olarian orogeny.

2.2.2. Structural framework

The structural framework of Broken Hill is dominated by a complex system of folds, faults, and shear zones predominantly oriented NE–SW, in relation to regional deformation associated with the Olarian orogeny.

The main structures include large-scale synforms and antiforms, as well as normal and reverse faults inherited from successive tectonic phases.

These structures play a fundamental role in the localization and redistribution of mineralized bodies.

Figure 5 clearly illustrates the density of faults, the orientation of folds, and their spatial relationship with mineralized lithological units, confirming the major structural control on the distribution of Pb–Zn–Ag lenses.

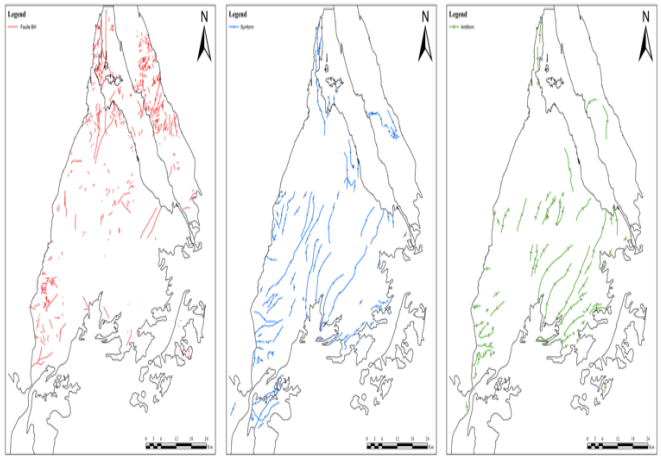

Figure 5. *Map of Broken Hill faults, synforms and antiforms*

2.2.3. Geological interpretations

Geological interpretations of the Broken Hill district are based on the integration of lithostratigraphic, structural, and mineralogical data to identify potential exploration targets. The current position of the district on the overturned limb of a D2 fold, located between major synforms and antiforms, suggests a strong structural control on the distribution of mineralization.

Zones enriched in garnet-bearing quartzite, amphibolite, and gahnite-bearing rocks correspond to favorable corridors for metal concentration, due to their role in channeling hydrothermal fluids during and after the Olarian orogeny. The combined analysis of faults, folds, and lithological units shows that mineralized lenses are preferentially located along shear zones and reactivated synsedimentary faults.

These structures have facilitated the remobilization of metals and their trapping within specific lithological levels. Distribution maps of garnet-rich and amphibolite-bearing rocks enable the identification of geological domains with high mineral potential, which coincide with the main regional structures.

Figure 6 synthesizes these interpretations by superimposing the main lithological units, tectonic structures, and preliminary exploration targets. It highlights several priority areas for exploration, located at the intersection of major structural corridors and favorable lithological formations. These interpretations therefore provide a solid foundation for guiding data selection and the development of artificial intelligence models in the subsequent phase of the study.

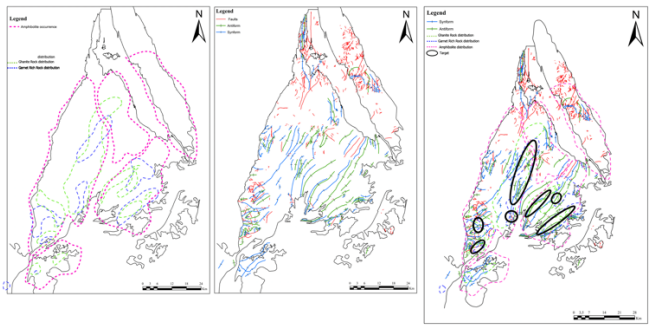

Figure 6. *Potential targets generated by structural and lithological Analysis*

2.3. Geochemical Context

The geochemical analysis conducted in the Broken Hill district highlights elemental signatures consistent with a polymetallic SEDEX-type mineralization that has been modified by metamorphism and structural remobilization. Whole-rock geochemical analyses reveal significant enrichments in zinc (Zn), lead (Pb), and silver (Ag), associated with notable concentrations of sodium (Na), lead-related potassium (K-Pb), as well as iron and manganese (Fe–Mn).

Distribution diagrams show a strong correlation between Fe, Mn, and Pb–Zn-enriched zones, suggesting that mineralizing fluids preferentially circulated along synsedimentary faults and shear zones. Spatial distribution maps of Zn, Pb, and Ag concentrations indicate high values within garnet-rich metasedimentary formations and quartzites, as well as in amphibolitic and gahnite-bearing units.

Silver occurs in lower and more variable proportions, often associated with proximal alteration zones and the margins of mineralized lenses. Figure 7 illustrates the geochemical distribution of the main metals (Zn, Pb, and Ag) and their relationships with associated elements within the Broken Hill district.

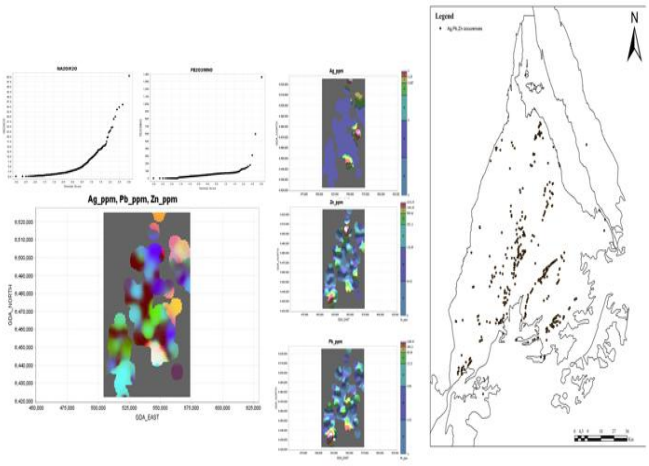

Figure 7. *Geochemical map and spatial distribution of metallic elements*

The spatial organization of geochemical anomalies closely coincides with major structures (NE–SW-trending folds and faults), confirming the decisive role of Olarian deformation in the remobilization and concentration of metals. These results support the hypothesis of a multi-stage mineralizing system, combining an initial exhalative-sedimentary deposit with subsequent redistribution controlled by metamorphism and tectonics.

2.4. Geophysical Context

The geophysical analysis of the Broken Hill area is based on the integration of three main types of data: gravimetry, radiometry, and structural lineaments, in order to characterize mineralized bodies, their spatial distribution, and the structures controlling metal remobilization.

The gravity map highlights high-density anomalies, primarily associated with lenticular and vein-type mineralized bodies characteristic of the Broken Hill Type (BHT) model. These anomalies suggest the presence of sulfide-rich masses (Pb–Zn–Ag), whose high density contrasts with the surrounding metasedimentary and gneissic host rocks.

Radiometric data, combined into RGB imagery, reveal spectral signatures consistent with thorium-enriched (Th) zones, generally associated with hydrothermal alteration and post-orogenic remobilization processes.

These radiometric zones partly coincide with gravity anomalies, reinforcing the hypothesis of mineralization controlled by late metamorphic fluids.

Figure 8 presents the main results of the integrated geophysical analysis (gravimetry, radiometry, and structural lineaments) of the Broken Hill district.

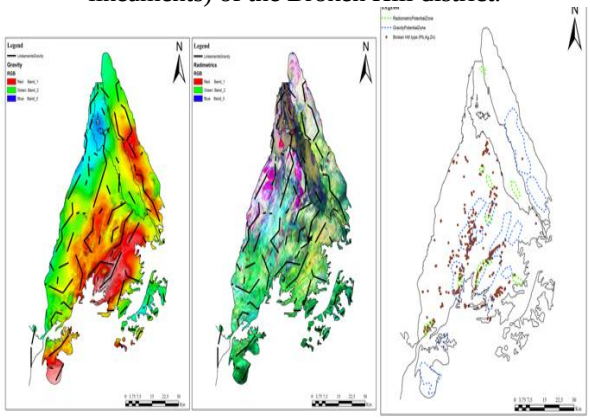

Figure 8. *Integrated geophysical data of the Broken Hill District.*

The extraction of structural lineaments reveals a predominance of NE–SW orientations, associated with synsedimentary faults and deformation related to the Olarian orogeny. These structures play a key role in the migration of metalliferous fluids and the concentration of metals.
The integration of these three approaches confirms that mineralization at Broken Hill is strongly controlled by the density of mineralized bodies, radiogenic alteration, and a complex structural network that facilitates metal remobilization.

2.5. Mineral Indicators

The reference data (ground truth) related to mineralogical indicators were extracted from existing mineralized sites in the Broken Hill district, as documented in official geological

databases and previous scientific literature. These known occurrences of Pb–Zn–Ag mineralization served as the basis for constructing the target layer used to train supervised learning models in predictive mineral prospectivity mapping.

The spatial distribution of these mineral indicators is closely associated with major tectonic structures, particularly NE–SW-trending faults and shear zones, confirming the structural control of mineralization in the region. To ensure compatibility with other geospatial layers (geological, geochemical, geophysical, and remote sensing data), the mineral occurrence points were digitized, georeferenced, resampled, and rasterized using a uniform spatial resolution.

Each pixel of the target layer was thus encoded with a value ranging from 0 to 1, where 0 represents areas with no known mineralization and 1 corresponds to areas containing mineralized occurrences or deposits (Pb–Zn–Ag). This binary layer constitutes the dependent variable of the model.

3. Data and Preprocessing

3.1. Data sources

The consolidation of the different data layers into a single table represents the final step in dataset preparation within the Geographic Information Systems (GIS) environment before applying machine learning algorithms for mineral prospectivity mapping.

This step aims to harmonize, structure, and standardize all spatial and non-spatial variables required for predictive modeling. In this study, several heterogeneous datasets were collected, preprocessed, and integrated, including : the study area boundary (perimetre_prepared_for_ML.csv), structural lineaments (Lineaments_prepared_for_ML.csv), mineralogical occurrences (Garnet_GDA.csv, Amphib_GDA.csv, Gahnite_GDA.csv), geochemical data (geochimie_prepared.csv), as well as structural elements such as faults, synforms, and antiforms (Structural_to_Antiform.csv, Structural_to_SyncInform.csv ,Faults_n_ready.csv).

In addition to these vector datasets, geophysical measurements from the GravityAS.xlsx file were incorporated, containing gravity anomaly (AS) values associated with X and Y coordinates.

The integration of these different data sources was carried out using GIS tools and spatial queries (spatial joins, interpolation, and extraction of values by location). Field calculators were used to generate new derived variables (distance to faults, lineament density, proximity to mineralized zones, etc.).
All variables were then assembled into a final structured table, ready to be used for training Random Forest (RF) and XGBoost.

Prior to the modeling phase, a systematic data quality assessment was conducted to identify and correct missing, anomalous, or inconsistent values, thereby ensuring a clean and reliable dataset for predictive mineral prospectivity analysis.
The study area was discretized into a regular grid with a spatial resolution of 100 meters, generating several thousand spatially distributed samples. The dataset includes both mineralized and non-mineralized instances, with an intrinsic imbalance reflecting the rarity of mineral occurrences. In order to correct this imbalance, a class weighting was mainly applied during learning in order to improve the sensitivity of the model to the minority class. In addition, the SMOTE (Synthetic Minority Over-sampling Technique) technique was explored to improve the representation of the minority class, while ensuring that geological coherence was preserved. This approach ensures robust learning and reliable predictive performance.

3.2. Data Preprocessing

The data preprocessing phase constitutes a critical step to ensure the quality, consistency, and usability of geospatial information in mineral prospectivity modeling using machine learning.

All vector layers (faults, folds, lineaments, lithologies, and geochemical data) were first imported using an automated procedure based on GeoPandas. This procedure corrects invalid geometries using the make_valid function and removes those that remain inconsistent or empty, thereby ensuring reliable spatial analyses.

All data were then reprojected into a unified metric coordinate system to enable accurate calculations of distances, areas, and densities. A regular grid covering the entire study area was generated; each cell represents a spatial unit of analysis, and its centroid serves as a reference point for spatial computations.
Lithological information was integrated by calculating, for each cell, the proportion of surface occupied by mineralogical units (gahnite and garnet).

For tectonic structures, two indicators were extracted: the minimum distance to faults and the local linear density, reflecting the structural control of Pb–Zn–Ag mineralization.
Point-based geochemical data (Zn, Pb, Ag, Fe, Mn, Na, K) were aggregated at the cell level as spatial averages. Finally, all these variables were merged into a single table exported in CSV, Parquet, and GeoPackage formats, forming a structured dataset ready for training Random Forest and XGBoost models.

3.2.1 Statistical Standardization of Geospatial Variables

Within the framework of multi-source data integration, variables derived from different layers (geology, geochemistry, geophysics, and remote sensing) exhibit heterogeneous orders of magnitude.

For example, geochemical values are expressed in ppm, magnetic anomalies in nT, and gravity indices in mGal.

Without prior normalization, Machine Learning algorithms may be biased by assigning disproportionate weight to variables with larger numerical magnitudes.
The variables were standardized using the z-score normalization (1), defined as:

$$Z = \frac{X-\mu}{\sigma} \qquad (1)$$

where X represents the raw variable, μ its mean, and σ its standard deviation.

This transformation ensures statistical comparability across the different thematic layers and improves the numerical stability of Random Forest and XGBoost models. It also reduces the

effects of excessive collinearity and facilitates the learning of underlying nonlinear relationships between the variables controlling Pb–Zn–Ag mineralization at Broken Hill.

3.2.2 Distance to faults : key structural parameter

Tectonic structures play a major role in controlling mineralization by acting as conduits for hydrothermal fluids. To quantify this structural control, the minimum Euclidean distance between each pixel and the main mapped faults was calculated:

$$D_f(x,y) = \min\sqrt{(x-x_f)^2 + (y-y_f)^2} \quad (2)$$

where (x, y) denotes the centroid of the analyzed pixel and x_f, y_f corresponds to the coordinates of the fault segments (2). The derived variable allows for the explicit incorporation of spatial geological information into the predictive model. Previous studies have shown that proximity to faults is strongly correlated with mineralization zones in Broken Hill-type Pb–Zn–Ag systems. Its integration improves the geological consistency of predictions and reduces the occurrence of false positives in structurally unfavorable areas.

3.3. Methodological workflow

The first stage concerns the extraction and preparation of multi-source data. It relies on the acquisition and preprocessing of several types of geospatial information, including:

(i) the Digital Elevation Model (DEM), from which slope and elevation are derived;

(ii) the regional geological map of Broken Hill, used to extract lithology and distance to faults;

(iii) geochemical data (Zn, Pb, Ag, Fe, Mn, Na, K) interpolated into continuous raster maps;

(iv) geophysical data, including gravity anomalies and structural lineaments;

(v) mineral indicators (Areas of Interest) corresponding to known occurrences of Pb–Zn–Ag mineralization, used as the target layer.

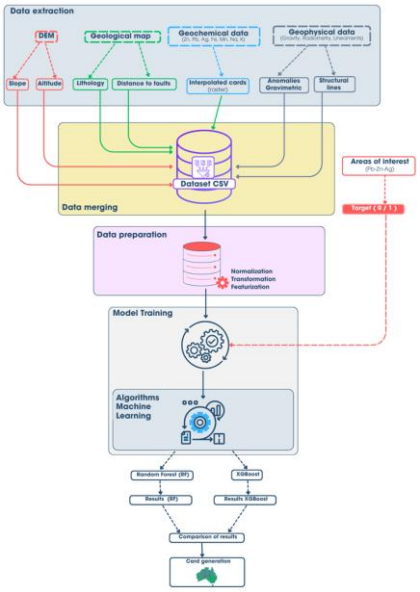

Figure 9. *Workflow for Machine Learning-Based Broken Hill Mineral Mapping*

Figure 9 illustrates the complete workflow from data acquisition and preprocessing to machine learning modeling and final mineral prospectivity mapping.

The second stage consists of merging the data into a unified database (CSV dataset). The different raster and vector layers are normalized, resampled, and spatially aligned to a common resolution. For each pixel or point (X, Y), values derived from geological, geochemical, geophysical, and topographic variables are integrated into a structured attribute table, thereby forming the training dataset. In parallel, the target layer is binarized (0 = absence of mineralization, 1 = presence of Pb–Zn–Ag occurrences).

The third stage is dedicated to the training of Machine Learning models. The dataset is divided into training and testing subsets. The dataset was split into 70% training and 30% testing subsets using random sampling. To mitigate class imbalance, class weighting was incorporated during model training, allowing the algorithm to appropriately penalize misclassification of the minority class and improve sensitivity to mineralized zones.
Two supervised algorithms are then applied: Random Forest (RF) and XGBoost, selected for their robustness to heterogeneous data and their ability to model complex nonlinear relationships. These models learn the relationships between predictive variables (features) and the presence of mineralization (target).

The fourth stage focuses on the validation, comparison, and interpretation of results. The outputs of the RF and XGBoost models are evaluated using performance metrics and compared with existing metallogenic knowledge.
The resulting mineral probability maps are analyzed to identify areas of high predictive convergence. These areas are subsequently delineated as priority targets for future mineral exploration campaigns.
It is important to note that spatial autocorrelation between neighboring samples may introduce bias in model evaluation. Future work should incorporate spatial cross-validation to ensure more robust generalization

4. Discussion and Results

4.1. Random Forest (RF) Classifier Analysis

Random Forest is based on an ensemble learning approach using independent decision trees, constructed from bootstrap samples of the training data.
The final prediction is obtained through averaging :

$$\hat{Y} = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \quad (3)$$

where $T_i(x)$represents the prediction of the i-th tree and N is the total number of trees (3).

This aggregation mechanism significantly reduces model variance and limits overfitting, a common issue in high-dimensional geospatial applications. It also enhances robustness to noisy data, which is particularly relevant in mineral exploration where geophysical and geochemical measurements may contain significant uncertainties.

Moreover, Random Forest provides a degree of interpretability through variable importance measures.

The evaluation of the Random Forest (RF) model was

conducted using the confusion matrix and key classification metrics (Accuracy, Precision, and Recall). In addition, F1-score and ROC-AUC were computed to provide a more comprehensive evaluation, particularly in the presence of class imbalance. On the test dataset, the model achieved an accuracy of 0.892, indicating that 89.2% of the pixels were correctly classified as mineralized or non-mineralized. This performance reflects the model's strong ability to discriminate between favorable and unfavorable zones for Pb–Zn–Ag mineralization (Figure 10).

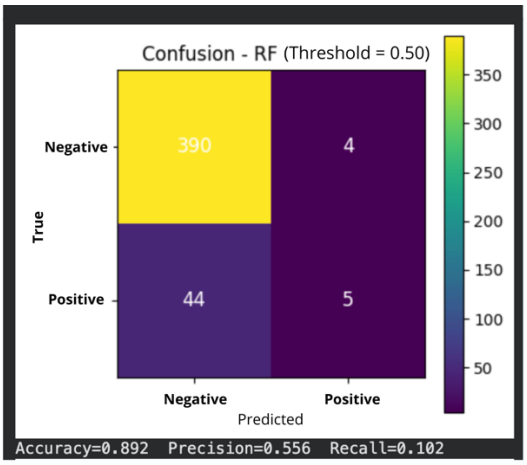

Figure 10. *Confusion Matrix of the Random Forest Model*

The model performance was evaluated using the following standard metrics:

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

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

$$Recall = \frac{TP}{TP + FN}$$

where TP denotes true positives, TN true negatives, FP false positives, and FN false negatives.

A detailed examination of the confusion matrix reveals that the model correctly identified 390 truly negative pixels (True Negatives) and 5 truly positive pixels (True Positives). However, it produced 44 false negatives, corresponding to mineralized zones incorrectly classified as non-mineralized, and 4 false positives, i.e., non-mineralized zones incorrectly predicted as mineralized. This distribution indicates that the RF model is highly effective in recognizing barren zones but less efficient in capturing all actual mineral occurrences.

In terms of complementary metrics, the model achieved a Precision of 0.556, meaning that among the pixels predicted as mineralized, 55.6% actually correspond to mineralized zones. In contrast, the Recall of 0.102 remains low, reflecting a limited ability of the model to detect all known mineralized zones. This low recall is mainly due to the imbalance of the dataset, where mineralized pixels are largely underrepresented compared to non-mineralized pixels.

To optimize RF performance, several hyperparameter configurations were tested, including the number of trees (n_estimators), the splitting criterion (entropy vs. gini), and the use of bootstrap sampling.

The best results were obtained with 100 trees, bootstrap enabled, and entropy as the decision criterion, selected through a Grid Search procedure. This configuration improved the model's stability and its capacity for spatial generalization across the entire study area.

Despite its limitations in terms of Recall, the RF model provides a robust initial framework for mineral prospectivity mapping and delivers a reliable estimation of favorable zones, serving as a reference for comparison with the XGBoost model presented below.

4.2. Analysis of the XGBoost (XGB) classifier

XGBoost optimizes a regularized loss function defined as:

$$L = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega\,(f_k) \quad (4)$$

where $l(y_i, \hat{y}_i)$measures the error between observed and predicted values, and $\Omega\,(f_k)$ is a regularization term controlling the complexity of the trees. (4)

Unlike Random Forest, XGBoost explicitly incorporates mathematical regularization that penalizes overly complex models. This promotes better spatial generalization, particularly useful when the model is applied to sparsely sampled areas.

The XGBoost (XGB) model was evaluated using the same methodology as Random Forest, based on the confusion matrix and classification metrics on the test dataset. The model achieved an accuracy of 0.876, slightly lower than that of RF, indicating that 87.6% of the pixels were correctly classified. This value remains high and confirms the relevance of XGBoost for predictive mineral prospectivity mapping.

Analysis of the confusion matrix shows that XGBoost correctly identified 377 truly negative pixels (True Negatives) and 11 truly positive pixels (True Positives). Compared to RF, the XGB model therefore detected a greater number of actual mineralized zones. However, it also produced 38 false negatives and 17 false positives, indicating a stronger tendency to classify certain non-mineralized areas as favorable.

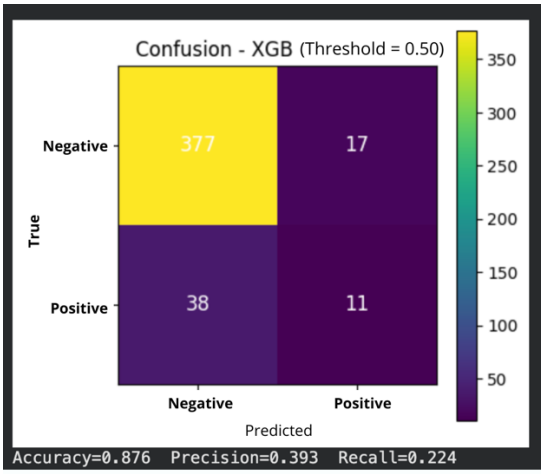

Figure 11. *Xgboost Model Confusion Matrix*

In terms of metrics, XGBoost achieved a Precision of 0.393, lower than that of RF, indicating that a larger proportion of pixels predicted as mineralized were actually barren. In contrast, its Recall of 0.224 is significantly higher than that of RF, reflecting a better ability to capture known mineralized zones. This difference highlights a classic trade-off between precision and recall: XGBoost favors the detection of mineralized zones at the cost of a higher rate of false positives. The optimization of the XGBoost model focused on key parameters such as the learning rate, maximum tree depth (max_depth), the number of estimators, and regularization. These hyperparameters were tuned using cross-validation to improve the balance between bias and variance.
Overall, XGBoost proved to be more sensitive to diffuse mineralization signals, which can be advantageous for preliminary exploration. However, its tendency to generate more false positives requires additional validation using field data or supplementary metallogenic criteria.

4.3. Data imbalance and remedial strategies

A recurring issue in Mineral Prospectivity Mapping (MPM) is class imbalance: mineralized occurrences typically represent less than 1% of the study area, which biases models toward the majority class (non-mineralized). To mitigate this bias, several strategies have been proposed:
- SMOTE to artificially generate mineralized samples; However, due to the spatial nature of geoscientific data, the use of SMOTE may introduce synthetic samples that do not fully preserve geological continuity. Therefore, class weighting remains a more reliable strategy in spatial modeling contexts.

- PU-learning (Positive–Unlabeled Learning);
- Class weighting and adjustment of decision thresholds.
Farahnakian et al. (2024) demonstrate that these techniques significantly improve the ability of RF, SVM, and logistic regression models to detect mineralized zones by reducing the false negative rate.

To address class imbalance, we applied SMOTE:

$$X_{synthetic} = X_i + \lambda\,(X_j - X_i) \quad (5)$$

where X_i is a minority sample, X_j its nearest neighbor, and $\lambda \in [0,1]$ is a random coefficient.(5)

4.4. Uncertainty and Explainability of ML Models in MPM

Uncertainty remains a major issue in predictive mapping. Zhang et al. (2025) proposed an approach based on sets of uncorrelated models to simultaneously produce a consensus map and an uncertainty map, thus strengthening the scientific credibility of the results.
In addition, the interpretability of deep learning models has improved thanks to tools such as SHAP (Shapley Additive Explanations), which make it possible to identify the most influential geological variables in predictions. In the case of uranium, these approaches have made it possible to quantitatively link the outputs of the ML models to known geological processes.

4.5. Comparison of Random Forest and XGBoost models

Comparison of the performance of the Random Forest (RF) and XGBoost (XGB) models highlights significant differences in their ability to predict the mineral potential of the Broken Hill Zone, reflecting distinct trade-offs between accuracy and sensitivity to mineralization.

Overall, the RF model has a **better accuracy (Accuracy = 0.892)** compared to XGBoost (**Accuracy = 0.876**). This indicates that RF is more reliable in correctly classifying the set of pixels, especially those corresponding to non-mineralized zones. This robustness is confirmed by the low number of false positives, which reflects a more conservative prediction strategy that is better suited to reducing costly errors in mineral exploration.

However, the analysis of the **Recall** reveals a major limitation of the RF (**0.102**), showing that this model hardly detects a large part of the truly mineralized zones. Conversely, XGBoost obtains a **higher Recall (0.224),** reflecting a better ability to identify known mineral occurrences. This difference suggests that XGBoost is more sensitive to the geological, geochemical and geophysical signatures associated with Pb–Zn–Ag mineralization, even when these are spatially diffuse or masked. On the other hand, this increased sensitivity of XGBoost is accompanied by a **decrease in accuracy (0.393 compared to 0.556 for RF)** and a higher number of false positives (17). This means that XGBoost tends to overestimate some areas as favourable, which could lead to additional exploration costs if the model was used alone. While Random Forest provides higher precision and reduces false positives, XGBoost shows higher recall, making it more effective for detecting mineralized zones. Therefore, model selection depends on exploration priorities.

From a metallogenic point of view, RF appears more suitable for the delineation of priority targets with a high probability of mineralization, while XGBoost is more relevant for preliminary regional exploration aimed at maximizing the detection of potential targets, even at the cost of greater spatial noise.
Thus, a **combined approach** is recommended:

- Use **RF for final target prioritization** and false positive reduction.
- Use **XGBoost for broad exploration and hypothesis generation** on new favorable areas.

This complementarity of the two models reinforces the robustness of the workflow and improves the reliability of the final mineral potential map produced for Broken Hill.

4.6. Importance of explanatory variables

The analysis of the importance of the variables highlights the factors that most control the prediction of mineral potential at Broken Hill. For the Random Forest model, the dominant variable is rock_area_frac_gahnite, highlighting the key role of lithology. Then come the distance to the lineaments and the distance to the faults, confirming the structural control of the mineralization. Structural density variables have a weaker influence.

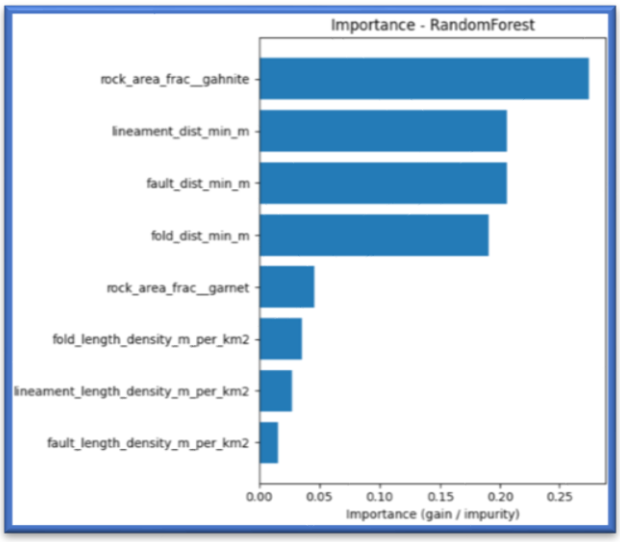

Figure 12. *Important features with RF*

For XGBoost, rock_area_frac_garnet becomes the most important variable, followed closely by rock_area_frac_gahnite. As with RF, proximity to faults and lineaments remains decisive, but XGBoost gives more weight to structural densities.

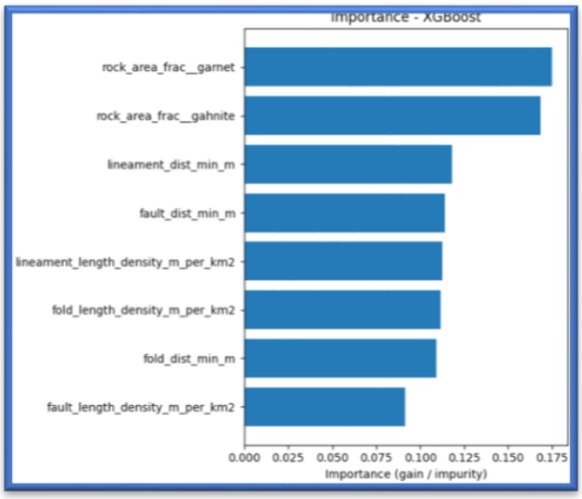

Figure 13. *Important features with XGBoost*

Overall, both models show that mineralization is primarily controlled by lithology and tectonic structures.
The importance of distance to faults confirms the structural control of mineralization, as faults act as pathways for hydrothermal fluids. Similarly, the significance of gahnite- and garnet-bearing lithologies reflects favorable conditions for Pb–Zn–Ag mineralization.

4.7. Final Mineral Potential Map

The final probability of mineralization is estimated as:

$$P\,(M|X) = \frac{1}{N}\sum_{i=1}^{N} T_i(X) \quad (6)$$

This probabilistic formulation makes it possible to produce a continuous map of mineral potential rather than a rigid binary classification (6). Areas with a probability greater than 0.7 are considered priority targets for exploration. This approach facilitates decision-making by explicitly incorporating predictive uncertainty and prioritizing areas of interest.

The final prospectivity map was generated using the Random Forest model due to its higher precision and interpretability.

White areas correspond to areas excluded from the model or without data. The black outline delimits the study area. The topographic background is derived from the DEM and the predictions are made at the pixel scale (resolution of the mesh used).

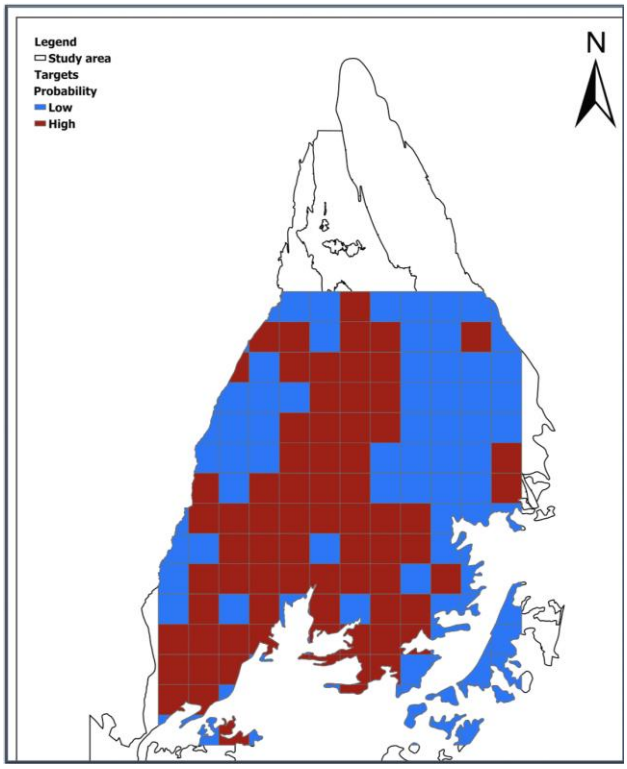

Figure 14. *Final Broken Hill mineral potential map from the ML Model (RF-XGBoost)*

5. Limitations

This study presents several limitations. First, class imbalance may affect sensitivity. Second, results depend on the quality of input data. Third, the absence of spatial cross-validation may lead to optimistic performance estimates due to spatial autocorrelation between neighboring samples.

This limitation may lead to spatial overfitting due to autocorrelation between neighboring samples, potentially inflating model performance metrics.

6. Conclusion

This work developed and adapted Geographic Information Systems (GIS)-integrated machine learning approaches for mapping mineral potential in Broken Hill. The combined use of geological, geophysical, geochemical and structural data has made it possible to build a robust predictive model, capable of identifying zones favourable to Pb–Zn–Ag mineralization with fine spatial resolution.

GIS datasets have been an essential resource for overcoming the limitations of geological interpretation based solely on field

observations. As classification and prediction tools, the Random Forest and the XGBoost have demonstrated their effectiveness in the spatial analysis of complex data, integrating lithological, structural and environmental information. These methods have made it possible to process large volumes of heterogeneous data, while reducing the impact of noise and uncertainties inherent in geospatial measurements.

The results obtained were compared to existing metallogenic knowledge and known mineral occurrences in the Broken Hill area. This comparison showed good consistency between the zones predicted as favourable and the historically mineralized areas, confirming the relevance of the model for mineral exploration. The analysis of the importance of the variables also highlighted the decisive role of tectonic structures and lithology in the control of mineralization.

Ultimately, while geology is not a perfectly deterministic science, the integration of GIS and machine learning is a major step forward in improving the accuracy of mineral prospecting models. This approach opens up new perspectives for the discovery of deposits, optimizing exploration strategies and reducing the risks associated with field campaigns.

This study demonstrates the effectiveness of integrating geoscientific data and machine learning for mineral prospectivity mapping.

The proposed workflow is reproducible and transferable to other geological contexts. Future work will focus on improving validation strategies and applying this approach to regions such as the Democratic Republic of the Congo.

7. References

[1] Zhang, S. E., Lawley, C. J. M., Bourdeau, J. E., Nwaila, G. T., & Ghorbani, Y. (2024). Workflow-Induced Uncertainty in Data-Driven Mineral Prospectivity Mapping. *Natural Resources Research*, *33*(3), 995–1023. https://doi.org/10.1007/s11053-024-10322-8

[2] Houran, N., Raoui, H. A., Manaan, M., Aabi, A., Simou, M. R., & Rhinane, H. (2023). USING GIS DATA AND MACHINE LEARNING FOR MINERAL MAPPING. STUDY CASE, BOU SKOUR EASTERN ANTI-ATLAS, MOROCCO. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *48*(4/W6-2022), 423–430. https://doi.org/10.5194/isprs-archives-XLVIII-4-W6-2022-423-2023

[3] Fu, Z., Zheng, X., Yan, Y., Xu, X., Zhou, F., Li, X., Zhou, Q., & Mai, W. (2025). The Evolution of Machine Learning in Large-Scale Mineral Prospectivity Prediction: A Decade of Innovation (2016–2025). In *Minerals* (Vol. 15, Issue 10). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/min15101042

[4] Wang, L., Li, T., Wu, S., Yang, J., Hu, Y., Hu, L., Chen, Y., Ge, Y., Chen, Y., Rao, C., & Du, Z. (2025). Geologically constrained data-driven modeling for mineral prospectivity mapping. *Geology*. https://doi.org/10.1130/g53947.1

[5] Forson, E. D., Amponsah, P. O., Wemegah, D. D., & Ahwireng, M. D. (2024). Random forest-based mineral prospectivity modelling over the Southern Kibi–Winneba belt of Ghana using geophysical and remote sensing techniques. *Applied Earth Science: Transactions of the Institute of Mining and Metallurgy*, *133*(1), 30–45. https://doi.org/10.1177/25726838231225055

[6] Seyedhamzeh, M., Maghsoudi, A., Ghezelbash, R., & Hajihosseinlou, M. (2024). *Uncertainty reduction with Hyperparameter Optimization in mineral prospectivity mapping: A Regularized Artificial Neural Network approach*. https://doi.org/10.21203/rs.3.rs-4760956/v1

[7] Fraser, S., Soto-Berelov, M., Holden, L., Hewson, R., Webb, J., & Jones, S. (n.d.). *Mapping stony rise landforms using a novel remote sensing, 2 geophysical, and machine learning approach*. https://ssrn.com/abstract=4562000

[8] Shaw, K. O., Goïta, K., & Germain, M. (2022). Prospectivity Mapping of Heavy Mineral Ore Deposits Based upon Machine-Learning Algorithms: Columbite-Tantalite Deposits in West- Central Côte d'Ivoire. *Minerals*, *12*(11). https://doi.org/10.3390/min12111453

[9] Abdulsalami, M., Omonile, J., Abubakar, F., Aliyu, A., & Yahaya, K. (2025). Multisource data fusion for enhanced gold mineral prospectivity mapping in Yagba West, Kogi State: a machine learning approach. *Proceedings of the Nigerian Society of Physical Sciences*, 182. https://doi.org/10.61298/pnspsc.2025.2.182

[10] Zhang, S. E., Lawley, C. J. M., Bourdeau, J. E., Parsa, M., Cumani, R., & Thompson, A. (2025). Mineral Prospectivity Modeling of Graphite Deposits and Occurrences in Canada. *Natural Resources Research*, *34*(3), 1243–1270. https://doi.org/10.1007/s11053-024-10451-0

[11] Ning, Y., Wang, Y., Lu, J., Tian, J., Wang, C., Sheng, S., Wen, S., Wang, S., & Dong, Y. (2025). Mineral prospectivity mapping for multi-source geoscience data: A novel unsupervised deep learning method. *Ore Geology Reviews*, *186*. https://doi.org/10.1016/j.oregeorev.2025.106866

[12] Kong, W., Chen, J., & Zhu, P. (2024). Machine Learning-Based Uranium Prospectivity Mapping and Model Explainability Research. *Minerals*, *14*(2). https://doi.org/10.3390/min14020128

[13] Johnson, S. J., & Murty, M. R. (2023). Machine Learning Approach to Improve Data Connectivity in Text-based Personality Prediction using Multiple Data Sources Mapping. *Journal of Scientific and Industrial Research*, *82*(1), 109–119. https://doi.org/10.56042/jsir.v82i1.70218

[14] Liu, Y., & Yonghang, L. (2022). *Mineral prospectivity mapping of tungsten polymetallic deposits using machine learning algorithms and comparison of their performance in the Gannan region, China*. https://doi.org/10.1002/essoar.10512400.1

[15] P, L., C, M., Mathew, A., & Shekar, P. R. (2025). Machine learning and deep learning-based landslide susceptibility mapping using geospatial techniques in Wayanad, Kerala state, India. *HydroResearch*, *8*, 113–126. https://doi.org/10.1016/j.hydres.2024.10.001

[16] Abedini, M., Ziaii, M., Timkin, T., & Pour, A. B. (2023). Machine Learning (ML)-Based Copper Mineralization Prospectivity Mapping (MPM) Using Mining Geochemistry Method and Remote Sensing Satellite Data. *Remote Sensing*, *15*(15). https://doi.org/10.3390/rs15153708

[17] Teknik, V., Monsef, I., Abdelnasser, A., & Ghods, A. (2025). *Machine learning (ML)-based mineral prospectivity mapping (MPM): Detecting Iranian plateau high-potential metallogenic zones using geospatial big data*. https://doi.org/10.21203/rs.3.rs-7730637/v1

[18] Tshanga Matthieu, M., Ncube, L., & Thamaga, K. H. (2026). Geological mapping of copper deposits in the democratic republic of Congo through remote sensing data and machine learning. *Frontiers in Remote Sensing*, *6*. https://doi.org/10.3389/frsen.2025.1678991

[19] Lasko, K., O'Neill, F. D., & Sava, E. (2024). Automated Mapping of Land Cover Type within International Heterogenous Landscapes Using Sentinel-2 Imagery with Ancillary Geospatial Data. *Sensors*, *24*(5). https://doi.org/10.3390/s24051587

[20] Guo, Y., Zhang, S., Fu, C., & Yang, Y. (2023). *GIS-based mineral prospectivity mapping: a systematic study on machine learning at Hezuo-Meiwu District, Gansu Province*. 10. https://doi.org/10.1117/12.2667272

[21] Riquelme, Á. I. (2025). Dual Random Fields and Their Application to Mineral Potential Mapping. *Mathematical Geosciences*, *57*(5), 845–881. https://doi.org/10.1007/s11004-025-10195-4

[22] Lachaud, A., Adam, M., & Mišković, I. (2023). Comparative Study of Random Forest and Support Vector Machine Algorithms in Mineral Prospectivity Mapping with Limited Training Data. *Minerals*, *13*(8). https://doi.org/10.3390/min13081073

[23] Saim, A. al, & Aly, M. (2024). *Enhancing Tree Species Mapping in Arkansas' Forests through Machine Learning and Satellite Data Fusion: A Google Earth Engine-Based Approach*. https://doi.org/10.21203/rs.3.rs-5385368/v1

[24] Farahbakhsh, E., Goel, D., Pimparkar, D., Müller, R. D., & Chandra, R. (2025). Convolutional Neural Networks for Mineral Prospecting Through Alteration Mapping with Remote Sensing Data. *PFG - Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, *93*(4), 379–400. https://doi.org/10.1007/s41064-025-00344-z

[25] Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., & Burnaev, E. (2024). Challenges in data-driven geospatial modeling for environmental research and practice. In *Nature Communications* (Vol. 15, Issue 1). Nature Research. https://doi.org/10.1038/s41467-024-55240-8

[26] Prospectors & Developers Association of Canada. Franck Arnott – Next Generation Explorers Award 2024/2025. Toronto : PDAC; 2025. Disponible sur : https://pdac.ca

[27] Farahnakian, F., Sheikh, J., Zelioli, L., Nidhi, D., Seppä, I., Ilo, R., Nevalainen, P., & Heikkonen, J. (2024). Addressing imbalanced data for machine learning based mineral prospectivity mapping. In *Ore Geology Reviews* (Vol. 174). Elsevier B.V. https://doi.org/10.1016/j.oregeorev.2024.106270

[28] Dillner, R. P., Wimmer, M. A., Porten, M., Udelhoven, T., & Retzlaff, R. (2025). Combining a Standardized Growth Class Assessment, UAV Sensor Data, GIS Processing, and Machine Learning Classification to Derive a Correlation with the Vigour and Canopy Volume of Grapevines. *Sensors*, *25*(2). https://doi.org/10.3390/s25020431

[29] Katiyar, N., Kulshreshtha, A., & Singh, P. (2024). A Review of Integration Techniques of Multi-Geoscience Data-Sets in Mineral Prospectivity Mapping. *Earth Sciences*, *13*(4), 127–140. https://doi.org/10.11648/j.earth.20241304.12

[30] Forson, E. D., Hagan, G., & Addo-Wuver. (n.d.). *Data-driven multi-index overlay gold prospectivity mapping using geophysical and remote sensing datasets*. www.usgs.gov/centers/eros

[31] Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. Systematic Reviews, 5(1), 210. https://doi.org/10.1186/s13643-016-0384-4.

[32] Harzing, A.W. (2025). Publish or Perish (Version [8.19.5300.9483 (2025.12.17.1530)]) [Software]. Available on https://harzing.com/pop.

[33] ©State of Queensland (Geological Survey of Queensland), 2025. Data sourced from the Department of Resources.

[34] Leyh, W.R., and Conor, C.H.H., 2000 *A redox boundary in the Broken Hill Block: implications for mineralisation. Australian Journal of Earth Sciences*, 47, pp. 123–134.