USING GIS DATA AND MACHINE LEARNING FOR MINERAL MAPPING. STUDY CASE, BOU SKOUR EASTERN ANTI-ATLAS, MOROCCO

Nouriddine HOURAN[1, *], Hicham AIT RAOUI[1], Mehdi MANAAN[1], Ayoub AABI[2], Mohamed Rabii SIMOU[1], Hassan RHINANE[1]

[1] Geoscience laboratory, Department of geology, Faculty of Sciences Ain Chock, University Hassan II Casablanca, BP 5366 Maarif Casablanca, Morocco

[2] Department of geology, Ecole Normale Supérieure, Mohamed-V University, Rabat, Morocco

KEY WORDS: Deposit, Prediction, Geology, Machine learning, Mineral mapping, Random Forest, Artificial Neural Network.

ABSTRACT:

The continued demand for mineral deposits in recent years has led exploration geologists for each stage of mineral exploration; find more effective and innovative ways of processing different data types. The use of Geographic Information Systems (GIS) allows various features, such as elevation, slope, tectonic structures, lithological units and indicator minerals of Bou Skour region, Eastern Anti-Atlas, Morocco to be mapped making targeted mining decisions easier. In this paper, a methodology was developed to enable the automated mapping of mineral using machine learning methods such Random Forest (RF) and Artificial Neural Network (ANN) achieves approximately 98% classification accuracy on a single Intel® Core™ i5-5300U CPU core with 16GB of memory, and come up with predictive maps representing the probable potentially mineralized areas.

1. INTRODUCTION

Mineral exploration involves providing and analysing geological maps in an attempt to locate the geological features related to target mineralization (Shirmard et al., 2022). These last include diverse features such as lithological units, tectonic structures, and indicator minerals. Typically some classic methods are used to mineral mapping such as Remote Sensing (RS) data that can be used to define significant zone of alteration, marked lithological connections, aeromagnetic data that can used to extract subsurface faults, and also geological, geochemical and field data to constrain the results, (Abdelkareem et al., 2018). Actively, geological mapping methods have progressed; and at the present time, the coupling GIS data and innovative data analytics such as machine learning is gaining considerable devotion. This amalgamation aids geologists stunned mutual challenges of old-style approaches such as independent decision that can offer consistent maps and avoid wasting money on prospecting for sterile regions (Shirmard et al., 2022).

Developing a suitable algorithm or for processing, analyzing and integrating various geospatial dataset (e.g., Geology, Topography, and GIS) is highly necessary for obtaining an efficient mineral map in order to visualize areas with a high favorability to be discovered further (Daviran et al., 2021). In this kind of modeling, various ranges of mathematical methods can be used for quantifying the spatial association between different evidential features and training locations,(Daviran et al., 2021). Recently, machine learning algorithms (MLAs), e.g., Random Forest (RF), and Artificial Neural Networks (ANNs) have gained much reputation and popularity in Mineral Mapping (MM), because of not requiring conditional independence of input features as well as ability to handle nonlinear correlations between known mineral deposits and spatial evidential features. (Park et al., 2021, Li et al., 2021)). In this paper, both classifying-models are used for the purpose of identifying different areas presenting the mineralization by treating the features as a pixel-level classification task (classify each pixel into each feature). The procedure of determining the class label is to superpose all the features in a way each pixel in the study area represents term information lithology, distance from faults, elevation and slope. Hence, the use of Random Forest (RF) and Artificial Neural Networks (ANNs) in Machine Learning results in splitting, training, and testing the data, which gives the accuracy of each model. Thus, using the power of python and its accompanying libraries (Geopandas, Rasterio…) Results can be shown and recorded with various extensions such as shapefile, TIF and csv.

2. STUDY AREA

Bou Skour is located in the south-eastern of Morocco, part of Draa-Tafilalet region territory. Consist of Idelsane commune, Ouarzazate province and Ait Sedrate sahl North, Ait Sedrate sahl South communes, part of Tinghir province, (Figure.1). The study region covers an area of 200396 hectares which is known for its mineral enrichment as deposits that are in operation which makes it a destination of geologists to raise the yield of minerals, as well as the events related to their positioning in place from sedimentation and deposition medium, magmatic events, tectonic structures and the geological phenomena responsible for their establishment.

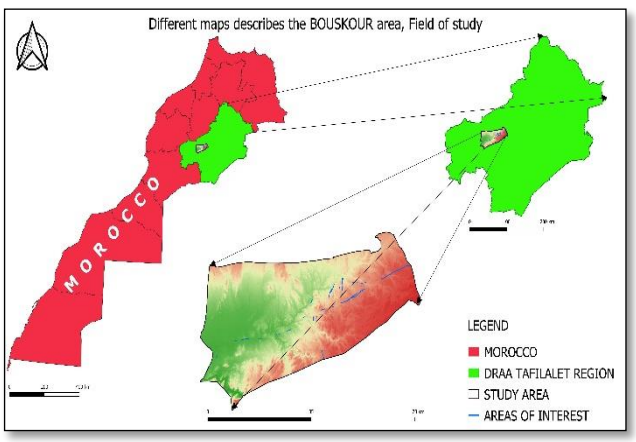


Figure 1: Overview map of Bou Skour study area

3. GEOLOGICAL SETTINGS

As part of the central Saghro massif, the Bou Skour area, the subject of this study, is located about 60 km east of Ouarzazate city on the southern side of the Sidi Flah inlier, located within precambrian extrusive and intrusive igneous rock units. As the oldest rocks in the prospect area, the extrusive rocks are composed of early Ediacaran andesitic-basalt rocks. During the last Pan-African event (Cadomian phase), these rocks underwent ductile-brittle deformation. This could be seen in schistosity metamorphism, and less developed orthogonal cleavages. Subsequently, various Pan-African plutons and dykes intrude on the metamorphosed andesitic-basaltic rocks of the Saghro Group, (Figure.2), (Aabi et al., 2021).

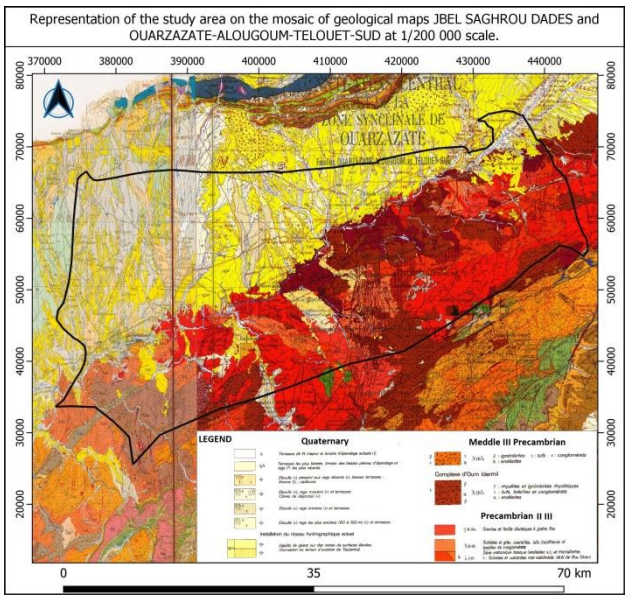


Figure 2: Geological map of the study area.

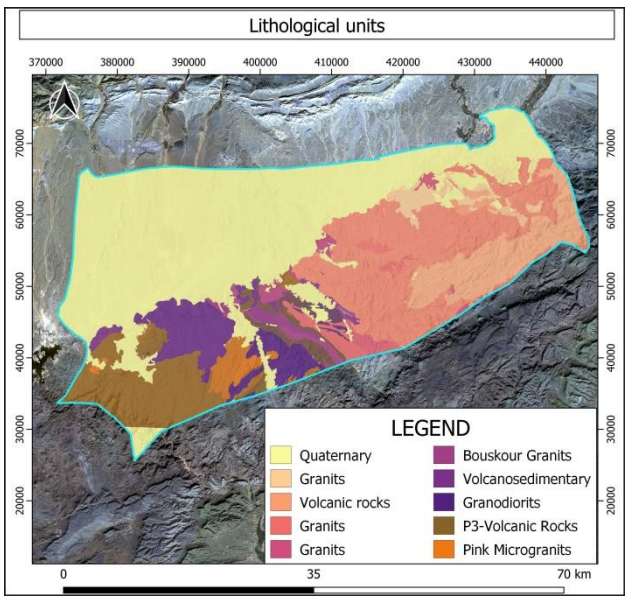


Figure 3: Lithological map of the study area. The lithological units at each pixel were used as the labelled data for the models.

4. DATA AND METHODS

4.1 Data collection

4.1.1 Images acquisition

The feed-in data was extracted from the Digital Elevation Model (DEM) that gave the elevation (Figure.4), and slope (Figure.5), values of each part on the area.

The DEM was extracted with a resolution of **30 meters** with a GeoTiff extension, that is to say it is georeferenced in the geographical system (WGS1984), to project it afterwards according to the project reference system that is **ESRI: 102191 - North_ Morocco_Degree** and to crop it on the study according to the dimensions, width: **2591** and height: **1719,** with a pixel size of **28.75.**

Then the geological map which includes the lithological units (Figure.3) as the tectonic structures that gives the Distance from faults feature (Figure.6).

The geological maps have been recovered as a **jpeg** images, of two zones, OUARZAZATE-ALGOUM-TELOUET_SUD and JBEL SAGHROU DADES in a scale of 1/200 000, the thing that makes them objects that has no reference, the thing that comes then is to give them one by georeferencing using GIS and return them to GeoTiff extension, giving the same reference system, and combine them to seek a map that represent the whole study area, hence, processing this map by the same method as DEM with the identical dimensions and pixel size. Moreover, to extract the lithology and the faulted zone in an exact and exploitable way, the only manner is the digitization of the formations and structures to make them hand and prepared for the steps that come after.

The fundamental step of data collection is acquiring these element images and treating them based on GIS software giving the same dimensions to each image in a way that every pixel on a single image is identical with others, then converting them to vector points that represent the pixel values for each one of them to get the inputs features in the machine learning algorithms.

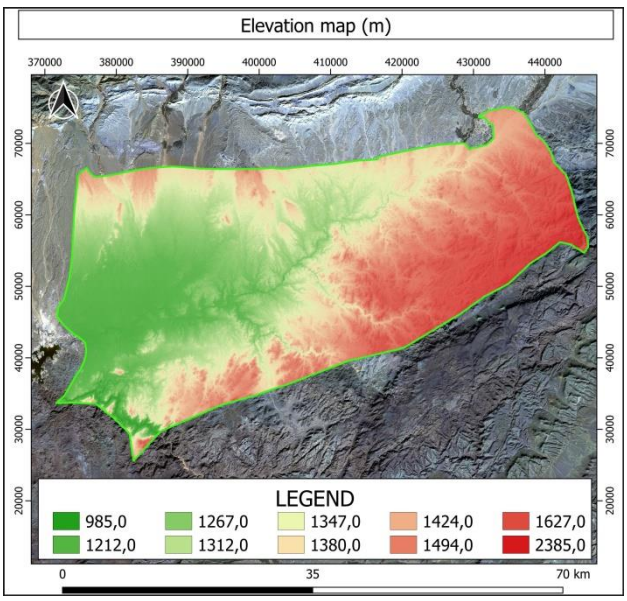


Figure 4: Elevation map of the study area. The elevation value at each pixel was used as the labeled data for the models.

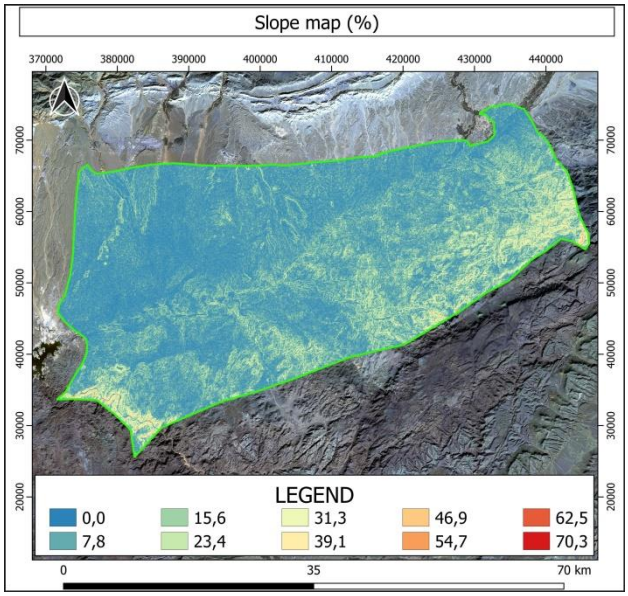


Figure 5: Slope map of the study area. The slope value at each pixel was used as the labeled data for the models. The minimal values are 0% and the maximal values are 70, 3%.

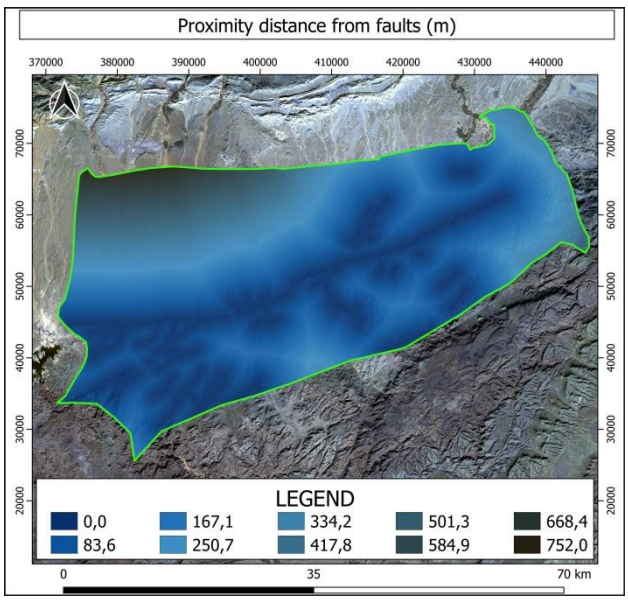


Figure 6: Proximity distance from faults map. In the models, labeled data was taken from the distance values at each pixel. 0 (meters) is the minimal distance, and 752 (meters) is the maximal distance.

4.1.2 Indicator minerals

The labeled ground truth data or indicator minerals were extracted from the mineralogy of pre-existing sites with mineral potential (Figure.7). This construction is generally relied to tectonic structures (faults), the thing that makes us work on faults mineralization and puts us in a situation where we are talking about tectonic-related minerals. Based on some present studies in the same area, this simple was mined, digitized, resized and pixelated in the same way as other features, hence, giving all the pixels varied values between 0 and 1 which the Zeros represent the vacant areas and Ones represent the areas with mineral existing to finally get the target layer used to train the mineral mapping model.

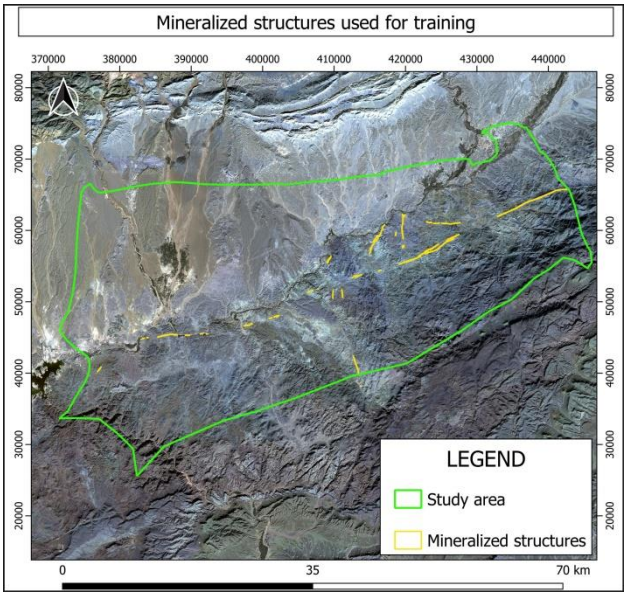


Figure 7: Map of pre-existing Indicator minerals in the study area. Each pixel in the Mineralized structures takes a value of 1 and the rest pixel values takes 0. Then each pixel was used as the labeled data for the models.

4.2 Data gathering

Gathering the dataset in one table (Table.1) is the final step of preparations using Geographic Information Systems (GIS), collecting all the feature values to facilitate the machine learning process and make it meaningful. Using some plugins and field calculator as some spatial queries, the table was successfully generated and assembled all the data needs to modeling the mineral mapping process in machine learning part which in turn provides some algorithms to visualize the data and see how it is consistent and coherent whether there are null values and other stuffs to begin the analysis, making sure that the training data are really clean and ready to go into the machine learning models, using Random forest (RF) and Artificial Neural Networks (ANN).

	X_coord	Y_coord	Lithology	Slope	DisToFault	Elevation	Mineral
0	371567	33679	97.0	8.898444	79.0	1178.0	0
1	371595	33679	97.0	8.898444	78.0	1175.0	0
2	371624	33707	97.0	8.898444	77.0	1170.0	0
3	371624	33679	97.0	8.898444	77.0	1169.0	0
4	371653	33736	97.0	8.898444	76.0	1144.0	0
...	...	...	...	...	...	...	...
2428428	411191	50730	29.0	7.000000	55.0	1427.0	1
2428429	411191	50701	29.0	9.000000	56.0	1434.0	1
2428430	411191	50673	29.0	12.000000	56.0	1434.0	1
2428431	411191	50644	29.0	14.000000	57.0	1434.0	1
2428432	411191	50615	29.0	18.000000	58.0	1427.0	1

2428433 rows × 7 columns

Table 1: Overview of the dataset

4.3 Workflow

The mineral mapping modeling based on GIS and machine learning is mainly divided into four principal parts (Figure.8). The first step is data collection that contains the geological map where lithological units and tectonic structures was extracted, the

DEM giving the slope and elevation for each part in the study region, and indicator minerals as Areas of interest that will be the targeting of our model next. The following step is to extract the features and the target of the model in a table where every row combines the pixel values for each feature correspond to given X and Y coordinate represented in the coordinate reference system (SCR) **ESRI: 102191 - North_ Morocco_Degree - Projected** (Table.1). The third step is to train machine learning models using the known samples and then use the trained model to predict mineral probability for the whole area. The fourth step is the comparison and evaluation of the predictive results. According to the metallogenic background and model, as well as previous experiences and knowledge, the prediction results can be evaluated and screened. Then, the most mineralized parts of the target area can be delineated.

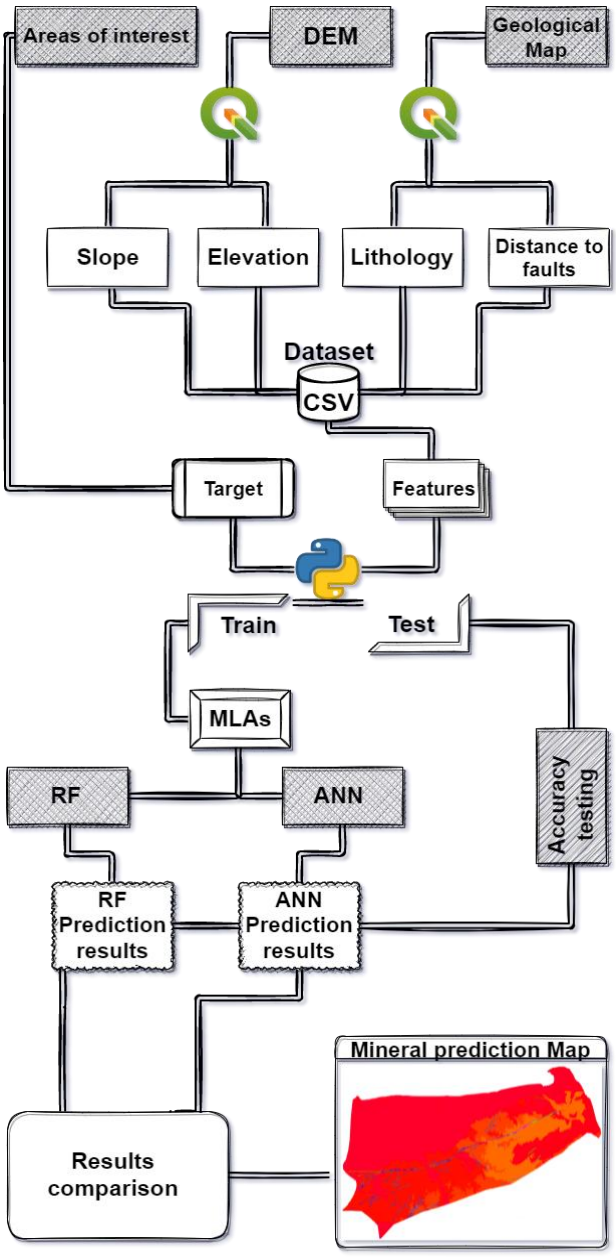


Figure 8: Workflow for Bou Skour mineral mapping based on GIS and machine learning. (Designed in: https://app.diagrams.net/)

4.4 Algorithms

4.4.1 RF

Random Forest works well on a wide range of problems. Basically, multiple decision trees are used for this (Li et al., 2021). The idea is to solve the issue that an individual decision tree may be disposed to over-fit a portion of the data.

By combining different individual decision trees into an ensemble (Figure.9), a random forest can average out the individual mistakes to reduce the risk of over-fitting. The random forest offers the advantage of not requiring pre-processing the data.

However, to achieve good performance it is critical to realize the important hyper-parameters needs to be tuned which include the extreme depth of the trees and the maximum number of features, (Li et al., 2021).

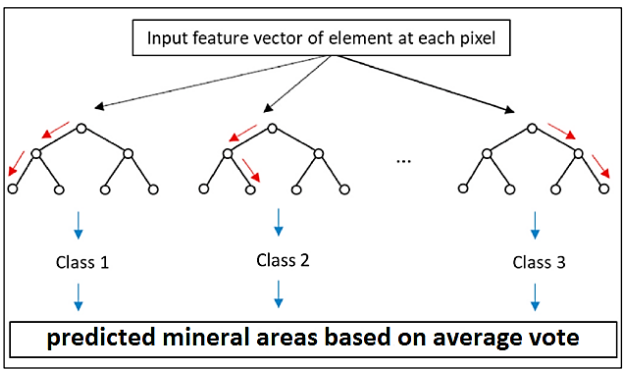


Figure 9: Illustration of the random forest for mineral mapping

4.4.2 ANN

A number of different machine learning algorithms exist, but one of the most popular technologies is Artificial Neural Networks (ANN) (Figure.10). Based on a large amount of training data, this algorithm can extract both implicit and complex correlations. ANN algorithms are based on a series of layers. Each layer contains a number of “neuron” units and carries a calculation of weighted input plus a bias term followed by a non-linear transformation. The results gotten by the above events are then fed into the next layer. The training process includes reducing the variance between the true values and predicted values. Throughout the process, the weights and bias of each layer are iteratively efficient by back broadcast algorithms. Due to the nonlinear activation function and hidden neurons, deep neural networks are established to deal with situations where input-output mappings are extensively complex (Li et al., 2021)

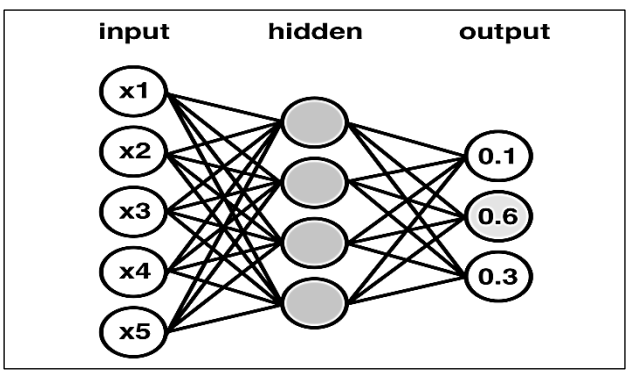


Figure 10: Illustration of the Artificial Neural Networks algorithm.

4.5 Data pre-processing and characteristic analysis

Through the data acquisition procedure illustrated in Figures 3-4-5-6-7 and assembled in Table.1 the **2428433** data were prepared for pre-processing, as shown in (Figure.8) goes into analysis before processing in order to have a global vision in term of variables frequency distribution (Figure.12), the plot A shows the frequent distances are intersected with the faults zone represented by a distance of 0 meter with 250 000 data points. The plot B indicates that the dominant lithological unit is the once labelled with the 1 value representing the Quaternary formations. The plot C directs to the recurrent elevation values in the area of study so that the high altitudes vary between 1200 and 1500 meters. The plot D designates that the dominant slopes range in values from 0% to 10%.

Concerning the correlations between the variables, the (Figure.11), demonstrates that the features representing a high and positive correlation are slope and elevation in the first range, then comes lithology and slope with low and positive correlation. In the other hand, the distance to faults and lithology represents a high and negative correlation. For the other variables the correlation is weaker.

Pre-processing the available data divided into two categories: training, and test datasets. After randomly assigning 25% of the data as test, the remaining 75% are randomly assigned to training data. Therefore, before the launch of the algorithms, it is essential to tuning the hyper- parameters in order to make the model more efficient and more accurate as they provide guidelines to prevent the over-fitting problems and trained model with the lowest error. After the execution of each algorithm, it is obligatory to test its accuracy using the test data, as and other supplementary metrics as receiver operating characteristic and features importance to validate the model and well understanding the provided results.

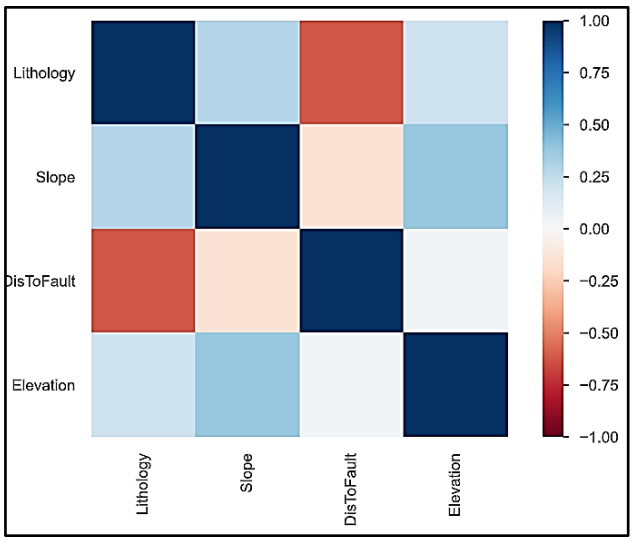
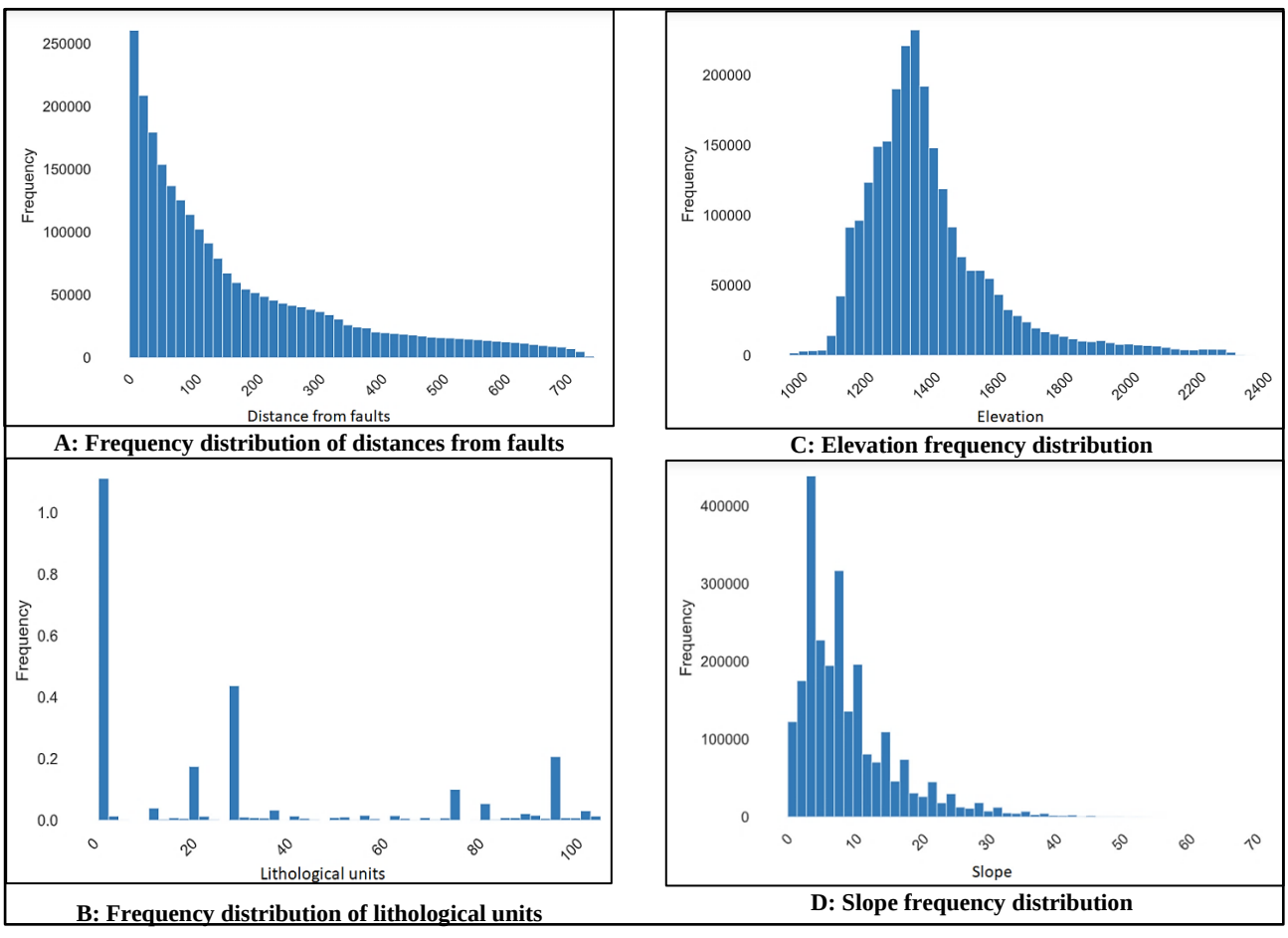


Figure 11: Feature correlations



A: Frequency distribution of distances from faults

C: Elevation frequency distribution

B: Frequency distribution of lithological units

D: Slope frequency distribution

Figure 12: (A, B, C, D) Features frequency distribution.

5. RESULTS AND DISCUSSION

5.1 RF Classifier

In this model, the performance of the RF classifier was tested on training dataset by the accuracy (98, 36%) and Receiver operating characteristic (Roc) with a value of (0.96). The RF classifiers were chosen in this study. Additionally, the effect of the number of trees in the forest (n-estimator) on the classifier performance was evaluated as well. The best performance for the training was done with 100 trees turning the Bootstrap to True and using entropy as criterion. Those hyper-parameters were chosen aids of grid search module are the best tunings to get the best performance as shows the results in the Figures.13-14.

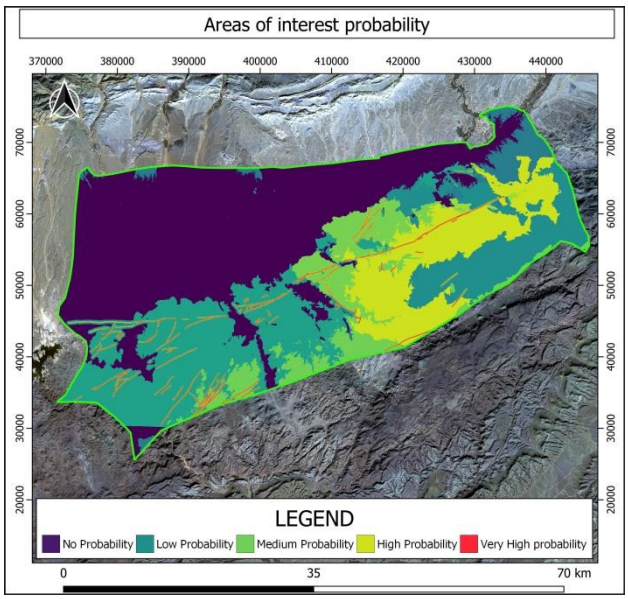


Figure 13: Mineral map predicted using Random Forest (Generated and reclassed in GIS Software).

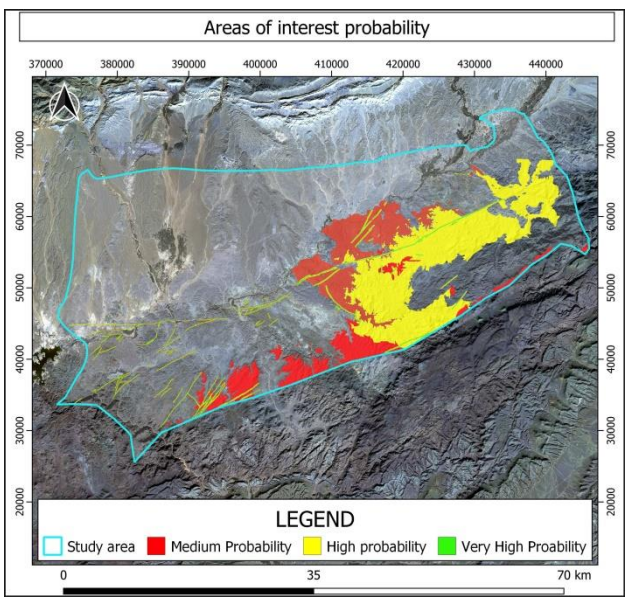


Figure 14: Mineral map predicted using Random Forest (Generated and reclassed in GIS Software).

5.2 ANN Classifier

The architecture of ANN was exposed to affect the prediction performance over the prediction investigates, it was observed that when ANN was applied with two or more hidden layers the performance did not increase over the use of a single hidden layer (Li et al., 2021). Therefore, in this study, ANN with one hidden layer with four nodes was used, gives a measurement accuracy of (97, 12%) and Receiver operating characteristic
(Roc) with a value of (0.83) were adopted to evaluate the performance of the model. As results of the Artificial Neural Network algorithm, the prediction at each pixel or cell is represent a probability value ranging from 0 to 1, which indicates the probability that it contains minerals occurrence in the study area, generally the model provides a moderate result (Figure.15), which are not very satisfied in term of performance.

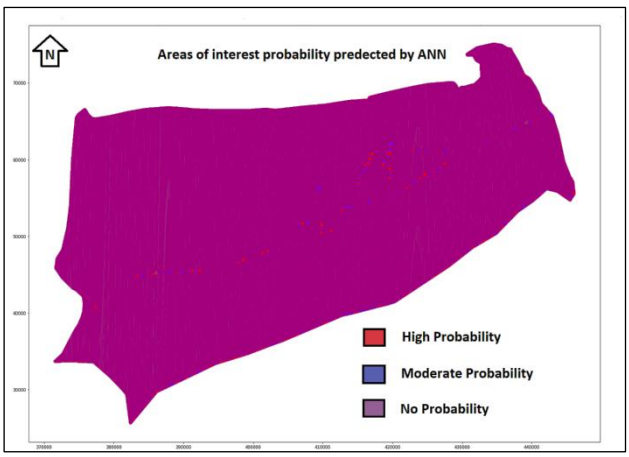


Figure 15: Mineral map predicted using Artificial Neural Network.

5.3 Prediction results comparison

For the shallow machine learning models, the validation dataset was used to measure the prediction performance of each classifier.
Performance results obtained by the accuracy are very similar for the two algorithms RF and ANN with an approximate value
(98%), then classification report shown in (Table.2, Table,3) as the Receiver operating characteristic (Roc) (Figure.16, Figure.17) are totally different and shows the performance of Random forest classifier with a macro F1 score of 0.83 performed best more than the Artificial Neural Network, Moreover, the scores calculated by different averaging strategy show slight differences, meaning that models trained by a balanced dataset can perform well regardless of the distribution of mineral indicators.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	605736
1	0.64	0.68	0.66	1373
accuracy			1.00	607109
macro avg	0.82	0.84	0.83	607109
weighted avg	1.00	1.00	1.00	607109

Table 2: Random Forest classification report.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	607109
1	0.00	0.00	0.00	0
accuracy			1.00	607109
macro avg	0.50	0.50	0.50	607109
weighted avg	1.00	1.00	1.00	607109

Table 3: Artificial Neural Networks classification report.

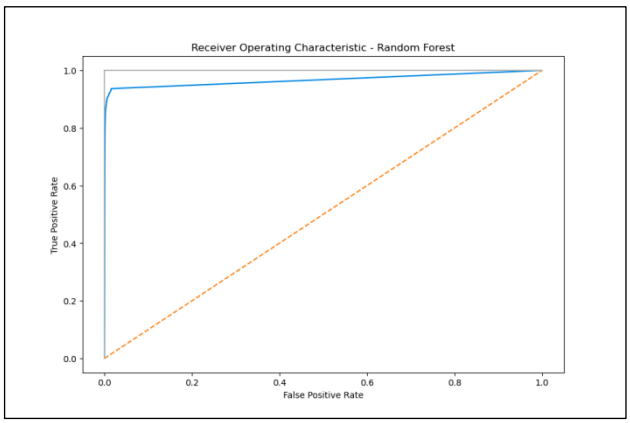


Figure 16: Random Forest Receiver operating characteristic.

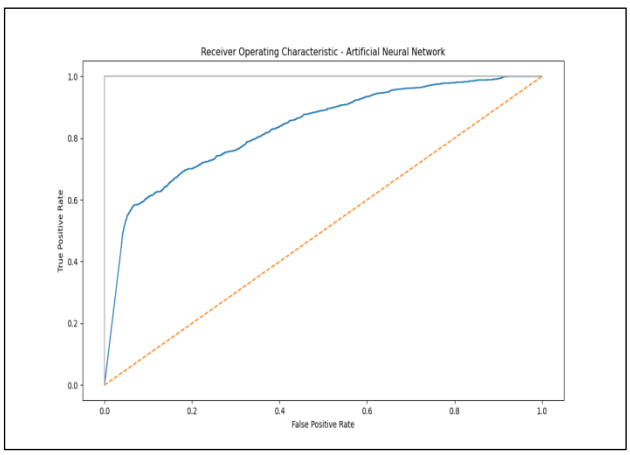


Figure 17: Artificial Neural Network Receiver operating characteristic.

5.4 Features importance

Calculating the importance of each factor in the training models, and indicate those that redirect and impact the prediction, The graph (Figure.18) indicate that the distance from faults factor with a percentage of (66%) control mostly the training models, that can be validated with the logical relation between mineralization and faults since the simples are highly concentrated on the faults zone. The Second feature affected the prediction represented as lithology with moderate percentage (29%) which is also related to the mineralization, highlighting the importance of this later highlight the importance of the latter in the deposition of the mineralization and also the ages, events and structural domains. to comes the elevation and slope with very low importance.

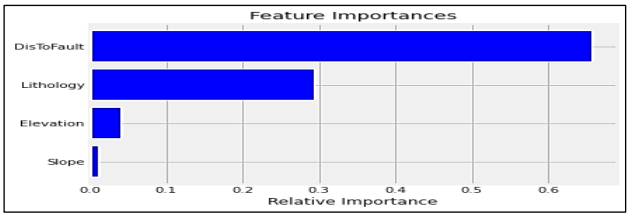


Figure 18: Features importance through the machine learning modelling.

6. ISSUES TO CONSIDER

6.1 Input Data Issues

The training data is highly concentrated in the central part which has a bias towards due to the involvement of fault structures in the study area. More data are needed to correct this imbalance.

6.2 Geological Issues

Mineral deposits are more often than not structurally controlled, which is why the 'Distance from Fault' feature was incorporated into the model.
One thing that was not included in the thought process of this feature is what each fault represents. Are there different faulting events? What is the structure-mineralization relationship? Did some of this faulting occurred post mineralization? This is an example of how understanding all aspects of geology in the study area will greatly affect what can and should go into the model.

6.3 Model Issues

The Random Forest and Artificial Neural Network models returned high classification accuracy. If we think about the issues already mentioned, the accuracy realistically should not be that high. This is most likely due to over-fitting, where the model has learnt all the little variations/data noise in our training/test data and accommodated for it. A way of removing over-fitting is by altering model parameters or trying a multitude of different machine learning algorithms and finding out what one works best.
Checking if yes or not the models are under over-fitting, the learning curve method was used as shown in the plots (Figure.19, Figure.20), the training score indicates how well the model is fitting the training data, while the cross validation score indicates how well the model fits new data, resulting that the RF model (Figure.19), works best than the ANN model (Figure.20), that has a tangle of training and validation score curves which indicate that is under over-fitting.

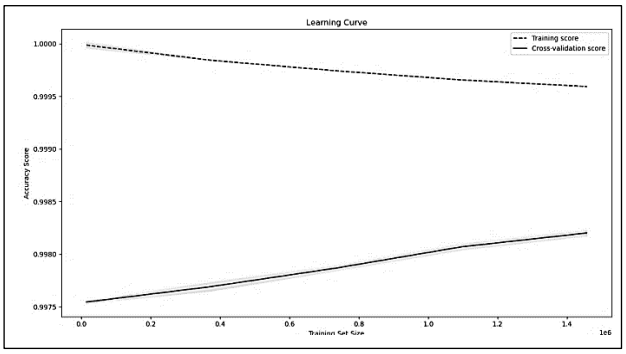


Figure 19: Random Forest Learning curve.

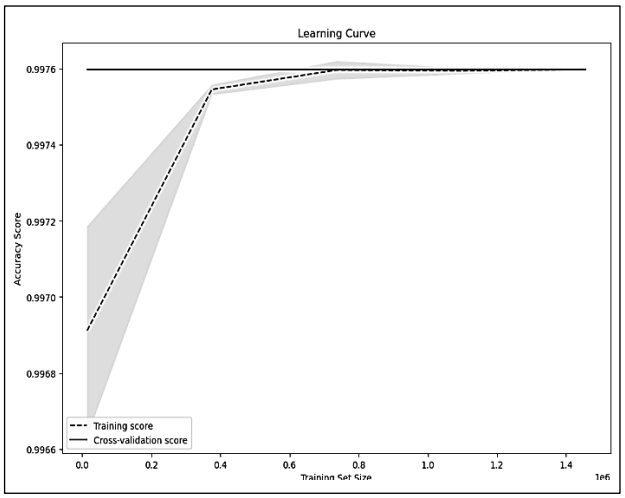


Figure 20: Artificial Neural Network Learning curve.

7. CONCLUSION

Developed and adapted some machines learning methods that have recently become popular and established for geographic information systems (GIS) data processing and investigated their applications for exploring different ore deposits.

GIS datasets have provided a new data resource to overcome problems associated with mapping geological features from field data alone. As a data-driven classification or prediction tool, Random Forests and neural networks have been widely applied in GIS data processing as well as a large number of research areas ranging from engineering and environmental science to physics and astronomy, (Shirmard et al., 2022).Recent advancements in machine learning methods have the potential to deal with large and complex data with features in processing ground truth measurements against noise and uncertainties.

Therefore, based on all the data mentioned, taking in consideration the problems and issues that occur, in this example, predict results were obtained and compared with the facts in the field as well as experts were consulted verifying that the potential sites for the presence of minerals are actually sites with properties and advantages that make the presence of minerals in them crucial.

To close, Geology has never been an exact science, but getting as close to exact as possible is crucial for the future of mineral deposit discovery.

8. REFERENCES

- Aabi, A., Baidder, L., Hejja, Y., Azmi, M. E., Bba, A. N., & Otmane, K. (2021). The Cu–Pb–Zn-bearing veins of the Bou Skour deposit (Eastern Anti-Atlas, Morocco): Structural control and tectonic evolution. *Comptes Rendus. Géoscience*, *353*(1), 81–99. https://doi.org/10.5802/crgeos.54

- Abdelkareem, M., Kamal El-Din, G. M., & Osman, I. (2018). An integrated approach for mapping mineral resources in the Eastern Desert of Egypt. *International Journal of Applied Earth Observation and Geoinformation*, *73*, 682–696. https://doi.org/10.1016/j.jag.2018.07.005

- Daviran, M., Maghsoudi, A., Ghezelbash, R., & Pradhan, B. (2021). A new strategy for spatial predictive mapping of mineral prospectivity: Automated hyperparameter tuning of random forest approach. *Computers & Geosciences*, *148*, 104688. https://doi.org/10.1016/j.cageo.2021.104688

- Li, C., Wang, D., & Kong, L. (2021). Application of Machine Learning Techniques in Mineral Classification for Scanning Electron Microscopy—Energy Dispersive X-Ray Spectroscopy (SEM-EDS) Images. *Journal of Petroleum Science and Engineering*, *200*, 108178. https://doi.org/10.1016/j.petrol.2020.108178

- Park, S. Y., Son, B.-K., Choi, J., Jin, H., & Lee, K. (2021). Application of machine learning to quantification of mineral composition on gas hydrate-bearing sediments, Ulleung Basin, Korea. *Journal of Petroleum Science and Engineering*, 109840. https://doi.org/10.1016/j.petrol.2021.109840

- Shirmard, H., Farahbakhsh, E., Müller, R. D., & Chandra, R. (2022). A review of machine learning in processing remote sensing data for mineral exploration. *Remote Sensing of Environment*, *268*, 112750. https://doi.org/10.1016/j.rse.2021.112750