Remote Sensing Applications: Society and Environment 36 (2024) 101316

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

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

Predictive modelling of mineral prospectivity using satellite remote sensing and machine learning algorithms

Muhammad Ahsan Mahboob [a, *], Turgay Celik [b], Bekir Genc [c]

[a] Sibanye-Stillwater Digital Mining Laboratory (DigiMine), Wits Mining Institute (WMI), University of the Witwatersrand, Johannesburg, South Africa
[b] School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa
[c] School of Mining Engineering, University of the Witwatersrand, Johannesburg, South Africa

ARTICLE INFO

Keywords:
Mineral prospectivity potential
Machine learning
Satellite remote sensing
Hydrothermal alterations
Copper mineral Pakistan
Deep learning
Convolutional neural networks
Support vector machine
Random forest

ABSTRACT

In today's world of falling returns on fixed exploration budgets, complex targets, and ever-increasing volumes of multi-parameter datasets, the effective management and integration of existing data are essential to any mineral exploration operation. Machine learning (ML) algorithms like Convolutional Neural Networks (CNN), Random Forest (RF), and Support Vector Machine (SVM) are powerful data-driven methods that are not implemented very often with remote sensing-derived hydrothermal alternation information and limited field datasets for mapping mineral prospectivity. The application of machine learning algorithms with satellite remote sensing data and limited field data, they have not been compared and evaluated together thoroughly in this field. A data science approach was applied to create nine predictor maps, incorporating limited field data and satellite remote sensing information. A confusion matrix, statistical measures, and a Receiver Operating Characteristic (ROC) curve were used to evaluate the prediction models efficacy on both the training and test datasets. The results suggested that the RF model exhibited the highest predictive accuracy, consistency and interpretability among the three ML models evaluated in this study. RF model also achieved the highest predictive efficiency in capturing known copper (Cu) deposits within a small prospective area. In comparison to the SVM and CNN models, the RF model outperformed them in terms of predictive accuracy and interpretability. These results imply that the RF model is the most suitable for Cu potential mapping in the Pakistan's North Waziristan region. Consequently, all the models including the RF model were used to generate a prospectivity map, which contained low to very-high potential zones, to support further exploration in the region. The newly discovered deposit inside the predicted prospective areas demonstrates the robustness and efficacy of the prospectivity modelling approach as proposed in this research for generating exploration targets.

1. Introduction

Mineral prospectivity mapping or modelling (MPM) is a technique for identifying and ranking areas which are likely to contain undiscovered mineral deposits of a particular type. This is accomplished by establishing a correlation between geological features (input variables) and the presence of the target mineral deposits (output variables) (Yin and Li, 2022). This technique typically involves these steps: (i) obtaining geospatial information from various geoscience data sources; (ii) determining exploration criteria that effectively depict processes essential for the formation of the desired deposit type; (iii) creating predictor maps from spatial datasets based

* Corresponding author.
E-mail address: mahsan.mahboob@wits.ac.za (M.A. Mahboob).

https://doi.org/10.1016/j.rsase.2024.101316
Received 5 February 2024; Received in revised form 2 July 2024; Accepted 3 August 2024
Available online 5 August 2024

on the chosen exploration criteria; and (iv) combining predictor maps to generate a predictive model for locating exploration targets in unexplored regions (Zuo, 2020). Geospatial information is often gathered through field surveys, which are not only costly but also time intensive. Another important issue is the lack of access to unreachable locations and limited field samples. Satellite remote sensing, on the other hand, has proved its effectiveness and importance in identifying mineral deposits and associations by detecting spectral anomalies (Mahboob et al., 2019; Benaissi et al., 2022; Rajan et al., 2019; Rajesh, 2004; Shirmard et al., 2022; Fu et al., 2023). Satellite remote sensing is a powerful technology for geologists and other mining professionals to identify hydrothermally altered rocks, structures, lineaments, lithological units, vegetation, and other valuable data. Generally, the MPM models can be distinguished based on the approach taken to assign evidential weights: (i) knowledge-driven models that use an expert-based approach to assign weights to the evidential maps, and (ii) data-driven models which use the spatial relationship between the evidential maps and known mineral deposits to calculate the weights empirically (Carranza, 2008). Over the last few decades, various data-driven modelling approaches have enabled progress in the field of MPM. Weights of evidence and logistic regression are two probabilistic methods that have become widely used due to their easy to understand and interpretation approach. Additionally, in the past few years, ML algorithms, developed mainly by computer experts for solving multi-field and multi-dimensional problems of classification and pattern recognition, have been seen as promising tools for creating predictive mineral prospectivity modelling and mapping (Shirmard et al., 2022; Mahboob et al., 2022; Jung and Choi, 2021; Sun et al., 2020; Köhler et al., 2021; Rodriguez-et al., 2015). ML algorithms such as SVM and RF have demonstrated superior predictive performance when compared to traditional statistical techniques or empirical models, especially when dealing with complexly distributed input features and nonlinear mineralization associations (Sothe et al., 2020; Santos et al., 2022; Geranian et al., 2016). Recently, deep learning, a powerful branch of ML, has been widely successful across many scientific domains. By learning hierarchical representations of input data, deep learning can recognize complex patterns and accurately classify them. Several geoscience applications, such as land subsidence susceptibility mapping, geochemical mapping, meteorological modelling, and environmental degradation modelling have been using these advantageous algorithms (Mahboob et al., 2022). The research conducted by Xiong and Zuo (2020) applied the combination of SVM and deep learning models for accurate mineral exploration using geochemical field data. Another research conducted by Kong et al. (Kong et al., 2022) applied Student Teacher Ore-induced Anomaly Detection (STOAD) deep learning model for mineral predictions mapping. The study concluded that the model performed exceptionally well and accurately predicted the location of target deposits. The research conducted by da Silva et al. (da Silva et al., 2022) showed the benefits of applying machine learning model RF on the drillhole data for accurate gold predictions in unexplored regions.

However, deep learning has been rarely implemented in the domain of MPM. Numerous algorithms have been proposed for MPM, but none of them is superior in all situations. Thus, a comparative analysis of multiple predictive algorithms is also necessary for an effective data-driven MPM.

The North Waziristan district, part of the Khyber Pakhtunkhwa province of Pakistan, was selected as a case study area for the ML based data-driven MPM. The region, which once comprised the Federally Administrated Tribal Areas (FATA), is a mountainous territory bordered with Afghanistan and located south and west of Islamabad, Pakistan. It is well-known for its natural resources and minerals, specifically its abundant Cu deposits (Mehsud, 2012). Since 2001, the region has been facing a severe law and order situation mainly due to geopolitical instability (Spychała-Kij, 2020). In recent years the country's geological survey department undertook geological mapping along with historical records, and it was discovered that huge amounts of Cu are located at Shinkai and Degan in North Waziristan. Estimates of these Cu ore-reserves are 122.71 million tons with a Cu content ranging from 0.3865% to a maximum of 2% (Khan, 2000). However, due to the poor law and order situation in the region, there were very limited efforts made to fully explore and extract the mineral resources. A comprehensive and reliable map of potential mineralization in the area is still needed to help the government achieve their exploration goals through drilling.

2. Materials and methods

2.1. Study area

The study area for this research is in the Miran Shah tehsil of North Waziristan region of Pakistan and near the Afghanistan border (38 km), as shown in Fig. 1. The region was previously known as the FATA of Pakistan and in May 2018 it has became part of the Khyber Pakhtunkhwa Province of Pakistan. In the past limited research or field surveys were conducted in the Miran Shah due to geopolitical instability in the region (Khan et al., 2022). Due to these geopolitical challenges, there is very limited information available on the geographical, geological and economical aspects of features located in the region. After military operations, the government rebuilt the region and provided economic means to the people of Waziristan.

The area is enriched with several precious metals and minerals due to the geological conditions, but the details of these mineralizations were unknown. In 2015, the geological filed visits were conducted to map the geology of the study area. The region is part of the larger geological framework of the Indian Plate's collision with the Eurasian Plate, which has resulted in a complicated tectonic history and a wide range of rock formations. The geology of Miran Shah comprises mainly complex ophiolite igneous rocks which are intensely folded, faulted, fractured and brecciated in places (Fig. 2). The ophiolite present in the study area is the part of the disjointed Tethyan ophiolites which is formed as back-arc basin and differs from mid-ocean ridges settings. The belt of the igneous rocks present in the study area extends south-west to north-east and is mainly composed of ultramafic, harzburgite, pyroxenites, quartz-diorites, micro-quartz diorites, granodiorites, dolerites and gabbros (Sothe et al., 2020). The Sulaiman Fold Belt is distinguished by substantial folding and thrusting of sedimentary strata. The deformation zone where the Indian Plate collided with and was forced over the Eurasian Plate is represented by this belt. The volcanic fine-grained porphyritic pillow basalts and andesites with secondary breccias and agglomerates are also present. The variety of sedimentary rocks that can be found, mainly consists of Jurassic to Eocene

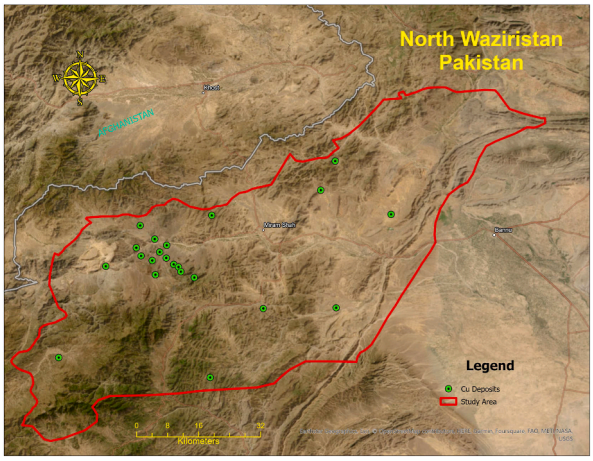

Fig. 1. The map of the study area.

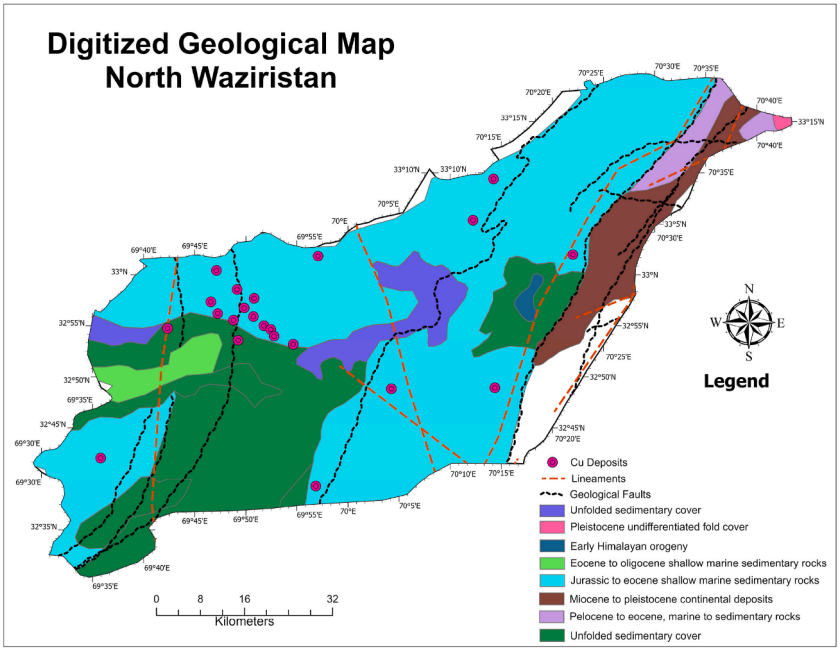

Fig. 2. Simplified Geological map of the study area along with lithological and structural features.

shallow marine continental and unfolded types. Sandstone, shale, limestone, and conglomerate are examples of these rocks. North Waziristan contains sedimentary sequences such as the Datta Formation, Lockhart Limestone, Lumshiwal Formation, and Tochi Formation. In addition to sedimentary rocks, the area has metamorphic rocks formed by tectonic activity's strong pressure and heat. Phyllite, schist, and gneiss are examples of metamorphic rocks. Some of the widely distributed sedimentary rocks are quite favorable for hosting economic Cu orebodies in the region and the most prominent outcropping structures are the North-South trending geological folds (Khan et al., 2007). Large porphyry intrusions formed in the early Cenozoic (Eocene-Oligocene) zones as a result of the convergence of the Waziristan Thrust and the subsequent subduction due to geotectonic settings. Significant porphyry copper and associated hydrothermal mineral deposits are found within these porphyry intrusions. The study area experienced mineralization processes during the Jurassic to Tertiary phases of magmatism, resulting in the formation of diverse porphyry Cu and Cu–Au deposits, low and high sulfidation epithermal Au deposits, and Cu–Pb–Zn vein-type deposits. The porphyry copper occurrences show hydrothermal alteration, which comprises of various regions including the potassium silicate alteration zone (K-zone), the propylitic alteration zones, as well as irregular Argillic and regular Phyllic alteration zones (Malkani et al., 2017).

The faults and folds in this area can be grouped as North-South and North-West trending basement faults, and North-East trending cap rock faults and folds.

2.2. Machine learning algorithms

In this research, Convolutional Neural Networks (CNN), Random Forest (RF), and Support Vector Machine (SVM) was applied mainly due to their effectiveness in geospatial and remote sensing applications, as well as the appropriate balance between interpretability and performance. CNNs are particularly well-suited for the processing of spatial patterns in remote sensing imagery, while RFs offer robustness and simplicity of interpretation. SVMs are particularly adept at managing high-dimensional data. The detailed explanation of each model is given in following sections.

2.2.1. Support Vector Machine (SVM)

The SVM is a powerful ML model used to solve the classification problems. The SVM was initially introduced by Cortes and Vapnik (1995), and because of its great predictive capability, adaptability, and robustness, it has been proven useful for identifying hidden patterns in data (Lim et al., 2002). Several researchers have applied SVM in geosciences, e.g., for land cover classification (Xu et al., 2019a), Geological 3D modelling using limited data from a variety of sources (Smirnoff et al., 2008), porosity prediction in a heterogeneous reservoir (Al- et al., 2010), consistency analysis of geophysical datasets (Turlapaty et al., 2010) and for mineral exploration and prospectivity mapping (Abedi et al., 2012). In the SVM algorithm, all data is expressed as points in an n-dimensional space, where n is the total number of attributes in the data. The aim is to identify the hyperplane that clearly differentiates the two feature groups by classifying the n-dimensional space, while maximizing the margin between classes.

For a given dataset of training which can be separated linearly, SVM categorizes the data α in the input space S into a high dimension space $D—\alpha \in \mathbb{R}^S \mapsto \Psi(\alpha) \in \mathbb{R}^D$ with a kernel function Ψ(α) to search for a separating hyperplane. SVM was originally established for binary classification problems, but later on it has been applied to multiclass problems. In multiclass, SVM starts with a single class and separates from the remaining classes one at a time. Each SVM is trained to differentiate all occurrences of a single class from the occurrences of all other classes. In testing, the class label y of a class pattern x is given by equation (1):

$$y = \begin{cases} m, if\ d_m\ (x) + t_h\ > 0 \\ 0, if\ d_m\ (x) + t_h\ \leq 0 \end{cases} \quad (1)$$

Where $d_m\ (x) = max\ \{d_c\ (x)\}_{c=1}^{N_h}$, $d_c\ (x)$ is the distance from x to the hyperplane corresponding to the class c, range from 1 to N_h and t_h is the classification threshold. The two main parameters that should be considered for the optimization of SVM results are regularization and gamma (Namdeo and Singh, 2021). The regularization usually denoted by C is the factor that decides the SVM optimization regarding the avoiding of misclassification in a class. The higher value of C will cause the optimizer to select a hyperplane with a small margin and higher misclassification, whereas a low value of C will result in a high-margin hyperplane with low misclassification. The gamma factor decides the influence of single class values on the selection of hyperplane (Tuba et al., 2017). Usually, the low gamma caused the data points far away from possible hyperplane are considered, whereas the high gamma means the points close to the possible hyperplane line are considered in the calculation.

2.2.2. Convolutional neural network (CNN)

A CNN is a form of deep learning neural network comprising of an input layer, one or multiple convolutional layers, pooling layers, fully connected layers, and an output layer. Input data, $X = (x_1,x_2,\ldots,x_n)$, is fed into a convolutional layer, which applies a randomly initialized filter (or kernel) to execute a convolution operation. Fig. 3 demonstrates a general convolution, where a filter is slid across the input data, calculating the dot products between the filter and the input, which results in a new feature dataset (a feature map).

Applying different $\sum i$ representing kernels yields multiple convolutional outputs, resulting in a set of feature maps that represent the input vectors in a hierarchy as shown by equation (2) derived by Imamverdiyev and Sukhostat (2019).

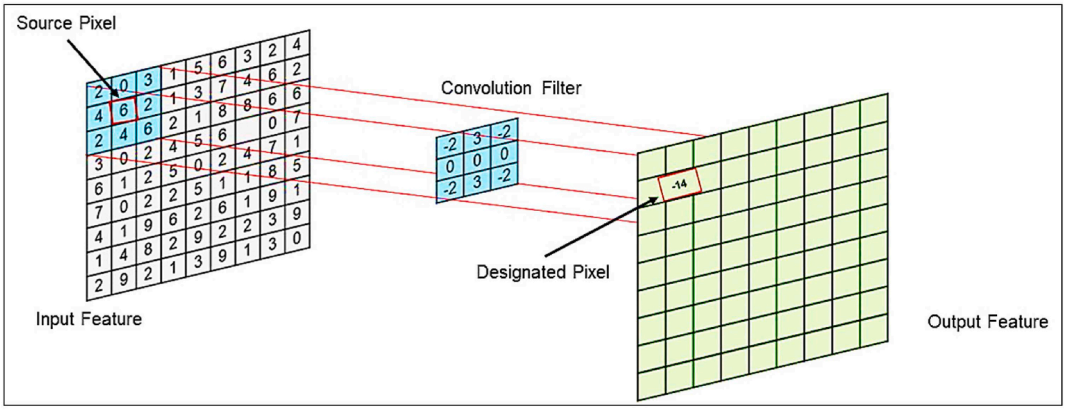

Fig. 3. General simplified architecture of convolutional neural network (conceptualized from source Robinson, 2018).

$$C_i = (c_1, c_2, \ldots, c_n) = k\,(X \cdot \Sigma_i) \tag{2}$$

Whereas C_i represents the ith convulational output with c_1,c_2,…,c_n as this elements. X is the input data and k is a scaling factor applied to the convolution operation. Using the activation function k, nonlinear amplification of convolutional results is achieved (Ghorbanzadeh et al., 2019). The most widely used activation function in CNNs is the Rectified Linear Unit (ReLU), expressed as equation (3).

$$k\,(x) = \max\,(0, x) \tag{3}$$

A pooling layer receives the outputs of one or more convolutional layers. Pooling is a filter-based technique for obtaining the maximum (max-pooling) or average (average-pooling) value from feature maps. This research uses max-pooling to downsample the data and reduce network dimensions, hence limiting overfitting and lowering computing costs. Features are extracted after the convolutional and pooling layers have processed the input data. Ultimately, the output of the fully connected layers is transmitted to the final layer, which employs a softmax activation function to compute the probability of each class according to the formula given in equation (4) (Adem, 2022).

$$P\,(l = m|x) = \frac{exp\,(w_m x + b_m)}{\sum_{i=1}^{2} exp\,(w_i x + b_i)} \tag{4}$$

The predicted class l is determined by the weights and bias terms w and b, respectively. The m represents the class lables and x is the input data. The input data of ML in MPM can be considered as an image, with each pixel representing a measurement of a certain attribute. Consequently, each grid cell is represented by a column vector whose length is equal to the number of features discernible, as a result, CNN with one-dimensional data representation was also used in this research (Krupski et al., 2021). The final layers of a CNN are fully connected layers that can transform feature maps into a column vector and classify them into categories or use them as a feature vector for further processing.

2.2.3. Random Forest (RF)

Breiman (2001) developed RF ML algorithm which makes repeated predictions of a single phenomenon by combining multiple decision trees. Typically, the RF generates a huge number of decision trees that are made up of several subsets of the original training dataset based on the bootstrap aggregating (or bagging) sampling method (Wang et al., 2016). After that, the selection and implementation of evidential layers will usually be done at each node of the tree to diversify and keep the forest growing. In this research, the root node was divided into several leaf nodes in each decision tree and finds the best one which contributes to the most accurate of the resultant trees. Several parameters can be considered in RF to measure the accuracy of the resultant trees, the most common ones are Gini index (I_G), gain ratio and Chi-square (Wijaya et al., 2022). In this research, the I_G was applied to calculate the purity of the leaf nodes compared to their root nodes as per the following equation (5).

$$I_G f = \sum_{i=1}^{n} f_i\ (\,1 - f_i) \tag{5}$$

Where f_i can be described as the probability of class i at node n, and can be calculated as equation (6):

$$f_i = \frac{m_j}{m} \tag{6}$$

Where the m_j is the number of samples related to class j, and m is the number of total samples present in a particular node.

2.3. Data preparation

2.3.1. Predictor maps

The predictor maps were developed from the spatial data collected through the field survey, published reports, maps and remotely sensed satellite imagery. A total of nine predictor maps were developed as the input datasets to ML models. In the predictive modelling process, integrating satellite remote sensing data with field data is crucial for enhancing model accuracy and reliability. Both datasets were georeferenced to a common coordinate system using control points and GPS coordinates to ensure precise spatial alignment. To avoid temporal discrepancies, satellite images and field data were collected within the same time frame, ensuring the field data was current and relevant to the satellite data acquisition period. The details are given in Table 1.

An understanding of complex mineral systems was used for the generation of predictive maps based on publicly available datasets of geologic features and satellite-derived hydrothermal alteration. These datasets are used as spatial approximations to represent the source, transport, sink, and depositional processes essential to Cu ore formation. The geology of the area (1:2,000,000 and 1:2,040,000 geological maps collected by the Geological Survey of Pakistan), is used to study various geological features such as structural zones, bedrock faults, lineaments formation and blanket fault intersections as shown in Fig. 2.

A multizone buffer analysis was performed to obtain numerical representations of the proximity of known Cu deposits to intrusive contacts (Fig. 4).

Table 1
The evidential features applied in the research for the mineral prospectivity mapping.

Serial. No	Evidential features	Description	Rationale
1	Geological Map	Delineation of suitable lithological contacts	The lithological units and contacts have significance for interpreting geological settings and pathways through which mineral-rich fluids flow, which improves the potential for mineral exploration
2	Geological Lineaments	Proximity to faults and folds	Faults and folds act as pathways for hydrothermal fluids, which are essential for the deposition of minerals. Their significance is well established in numerous studies emphasizing their role in increasing permeability and fluid flow.
3	Intrusive Contacts	Proximity to inferred intrusive contacts identified based on the resistivity anomalies	Intrusive rocks frequently have an association with mineralization because of their function in supplying the required heat and compounds for hydrothermal processes. The close proximity to these contacts serves as a robust indicate of possible mineralization.
4	Iron-oxide Alteration	Proximity to iron-oxide anomalies delineated from Sentinal-2 Satellite data	Iron-oxide alteration zones are an evident indication of hydrothermal processes, which are usually linked to the presence of copper mineralization. These regions can be identified using remote sensing and are crucial for identifying areas of significance.
5	Argillic Alteration	Proximity to argillic anomalies delineated from Sentinal-2 Satellite data	This alteration can be detected by the occurrence of clay minerals that are formed via hydrothermal processes and linked to Cu mineralization.
6	Phyllic Alteration	Proximity to phyllic anomalies delineated from Sentinal-2 Satellite data	This alteration is characterized by the creation of minerals such as sericite and quartz, has been repeatedly linked to porphyry copper systems, making it an important indicator of mineralization.
7	Propylitic Alteration	Proximity to propylitic anomalies delineated from Sentinal-2 Satellite data	The alteration zone is commonly located at the fringes of mineralized systems, serving as an indication of the boundaries of hydrothermal alteration. Delineating prospective mineralized zones is of the greatest significance.
8	Topographic Elevation	Surface height of the deposits from the mean sea level extracted from space based digital elevation model	The topography of an area can impacts the processes of mineral deposition by influencing the movement of fluids and patterns of erosion. Elevated regions could reveal mineral-rich areas because of erosion.
9	Topographic Slope	Steepness of deposits relative to the horizontal plane delineated through space based digital elevation model	The slope of the ground has a significant impact on the process of drainage and erosion, which in turn plays a crucial role in the identification and exposure of mineralized zones. Areas with steeper slopes may suggest higher levels of erosion, which could lead to the exposure of valuable mineral reserves below the surface.

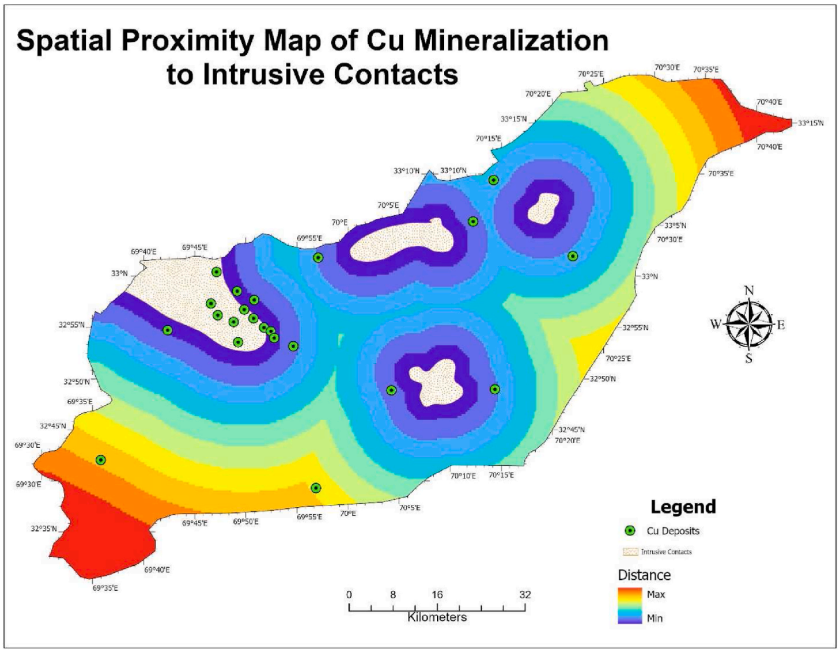

Fig. 4. Spatial proximity map of Cu mineralization to intrusive contacts.

Fig. 5 shows a fault density map that was produced to highlight the wide dispersion of caprock faults, North-South trending faults and lineaments intersections.

In addition to field geological data, remote sensing satellite imagery from Landsat-8 (acquired in October 2019) and Sentinel-2 (acquired in October 2019) were utilized to identify and map hydrothermal alteration zones (Sekandari et al., 2020; Tompolidi et al., 2020; Van der et al., 2016) and lithological components related to Cu mineralization in Waziristan, on regional, local, and district scales. The ASTER satellite has become known as a prominent tool in the field of satellite remote sensing for mineral exploration, par-

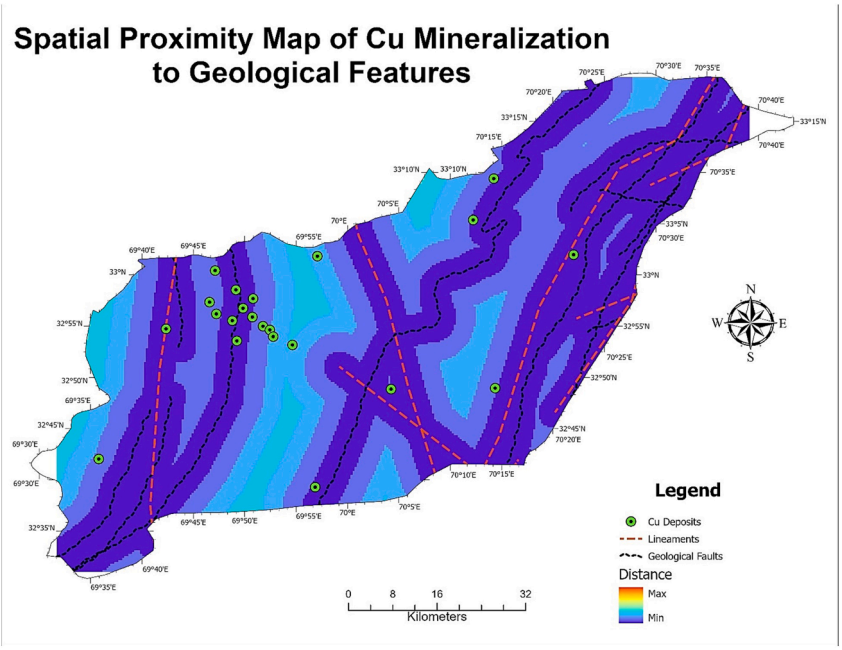

Fig. 5. Spatial proximity map of Cu mineralization to geological faults and folds.

ticularly in the domain of hydrothermal mineral mapping and exploitation. However, the SWIR sensors on ASTER had been dysfunctional since 2008 due to a fault in the cryocooler (Shimoda and Kimura, 2017). This particular situation has necessitated the emergence of alternate remote sensing data sources for mineral exploration, which are Landsat and Sentinel satellite series. The technical characteristics of the Landsat-8 and Sentinel-2 sensors are shown in Table 2. Both the Landsat-8 and Sentinel-2 satellites have prospective applications and benefits for mineral exploration research (Mahboob et al., 2019; Blandine et al., 2023; Chen et al., 2022). This relates mostly to their good multispectral imaging capabilities, sensitivity to many minerals, including copper minerals, reasonably high spatial resolution, worldwide coverage, and open data policy. They provide essential data for the detection and mapping of various mineral resources, including copper. Specific sections of the electromagnetic spectrum, such as (0.64–0.68 μm), are sensitive to the reflectance characteristics of iron oxide minerals including hematite or goethite, which have a strong association with copper presence. The near infrared band (0.84–0.881 μm) can also be used to differentiate between different rock types associated with copper deposits. Because of their mineralogical compositions, copper-bearing rocks frequently show distinct spectral signatures

Table 2
Spectral and spatial characteristics of Landsat-8 and Sentinel-2 satellite imagery.

LANDSAT – 8				SENTINEL-2			
Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS)				Multispectral Instrument (MSI)			
Band No	Band Name	Spectral Range (μm)	Spatial Resolution	Band No	Band Name	Spectral Range (μm)	Spatial Resolution
B1	Coastal aerosol	0.43–0.45	30	B1	Coastal aerosol	0.433–0.453	60
B2	Blue	0.45–0.51	30	B2	Blue	0.458–0.523	10
B3	Green	0.53–0.59	30	B3	Green	0.543–0.578	10
B4	Red	0.64–0.67	30	B4	Red	0.650–0.680	10
				B5	Vegetation red edge	0.698–0.713	20
				B6	Vegetation red edge	0.733–0.748	20
				B7	Vegetation red edge	0.773–0.908	20
B5	Near Infrared (NIR)	0.85–0.88		B8	Near Infrared (NIR)	0.848–0.881	10
			30	B8a	Narrow Near Infrared (NIR)	0.931–0.958	20
				B9	Water vapour	1.338–1.414	60
B9	Cirrus	1.36–1.38	30	B10	Cirrus	1.539–1.681	60
B6	Short wave infrared (SWIR 1)	1.57–1.65	30	B11	Short wave infrared (SWIR 1)	2.072–2.312	20
B7	Short wave infrared (SWIR 2)	2.11–2.29	30	B12	Short wave infrared (SWIR 2)	2.081–2.323	20
B8	Panchromatic	0.50–0.68	15				
B10	Thermal Infrared (TIRS) 1	10.6–11.19	100				
B11	Thermal Infrared (TIRS) 2	11.50–12.51	100				

in the NIR region. Minerals that contain copper, such as bornite or chalcopyrite, indicate unique absorption characteristics within the short-wave infrared (SWIR), which covers about 1.57–2.323 μm range. By analysing the reflectance values within a specific range of the electromagnetic spectrum, geologists are able to identify regions exhibiting increased absorption, which serves as an indicator of the existence of copper minerals (van et al., 2009).

This is because remote sensing data has been widely and efficiently used for mineral exploration (Mahboob et al., 2019; Benaissi et al., 2022; Rajan et al., 2019; Adiri et al., 2020; Soydan et al., 2021). Analyzing satellite data to determine the geographical locations of hydrothermally altered rocks can help identify the main outflow zones of hydrothermal systems, potentially leading to the discovery of mineral deposits. Although Cu or any other metallic mineralization can be rarely identified directly through any existing remote sensing techniques, certain minerals such as iron oxides and clays detected through their spectral signatures in the visible and near/shortwave/thermal infrared portion of the electromagnetic spectrum can be used to indicate the presence of hydrothermal alteration zones, which are associated with Cu occurrence (Pour and Hashim, 2012; Atwizukye, 2022).

The Fast Line-of-sight Atmospheric Analysis of Hypercubes (FLAASH) technique, together with the sub-arctic summer (SAS) and Maritime aerosol models, were used to eliminate atmospheric attenuation interference on satellite data, converting it from sensor radiance to surface reflectance using PCI Geomatics code. No atmospheric adjustment was applied on Landsat-8 TIR data and hence used with the original brightness values. Following that, the pre- and post-processing results of Landsat-8 and Sentinel-2 satellite imagery were compared as shown in Fig. 6.

The VNIR (Visible Near Infrared) and SWIR (Short wave infrared) spectral regions have already been highlighted in terms of their importance and high potential for mapping alteration materials in considerable detail by several researchers (Mahboob et al., 2019; Benaissi et al., 2022; Rajan et al., 2019; Pour and Hashim, 2012; Atwizukye, 2022). In the VNIR wavelength range, the Sentinel-2 data includes nine bands, while the Landsat-8 data has four bands. However, in the SWIR wavelength range, both data sets have two bands. As a result, Sentinel-2 and Landsat-8 multispectral data have the potential to effectively map various hydrothermally altered minerals.

Several hydrothermal mineral alterations associated with Cu mineralization such as Iron-oxide alteration (Bauer et al., 2022), Argillic alteration (Nasab and Agah, 2023), Phyllic alteration (Soloviev et al., 2019) and Propylitic alteration (Yang et al., 2022a) was mapped using the band rationing, principal component analysis and Intensity- Hue-Saturation (IHS) transformation from remotely sensed satellite data.

The terrain can also influence the spatial distribution of mineralization, which can be considered as an indirect sign of the subsurface geology (Carranza, 2009; Keykhay-Hosseinpoor et al., 2020). The study area is characterized by a rugged mountainous landscape, where the spatial attributes and land cover are significantly influenced by topographic variables such as surface elevation and slope. The guidelines developed by Grunsky (1996) for mineral exploration also included the topographic elevation as an important parameter for effective mineral exploration and included in the list of must require datasets. Thus, topographic factors such as surface elevation and slope are taken into consideration when studying the spatial distribution of Cu mineralization in the study area as shown in Figs. 7 and 8 respectively. The topographic predictor maps were derived and processed from the Space Shuttle Radar Topog-

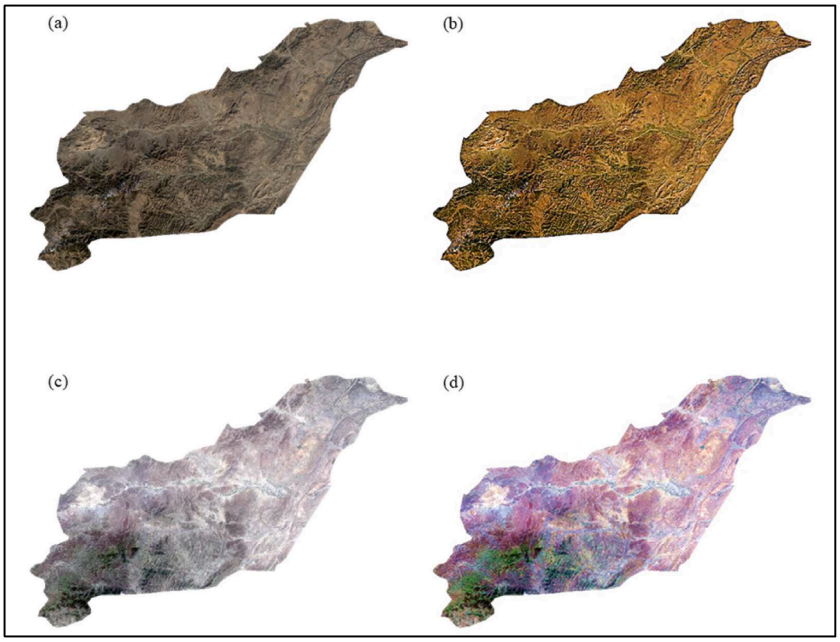

Fig. 6. Digital image processing of satellite imagery (a) raw Landsat-8 image; (b) pre-processed Landsat-8 image; (c) raw Sentinel-2 image; (d) pre-processed Sentinel-2 image. The illustration is for False Color Composite (FCC) highlighting Band 3 as Red, Band 4 as Green and Band 2 as Blue. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

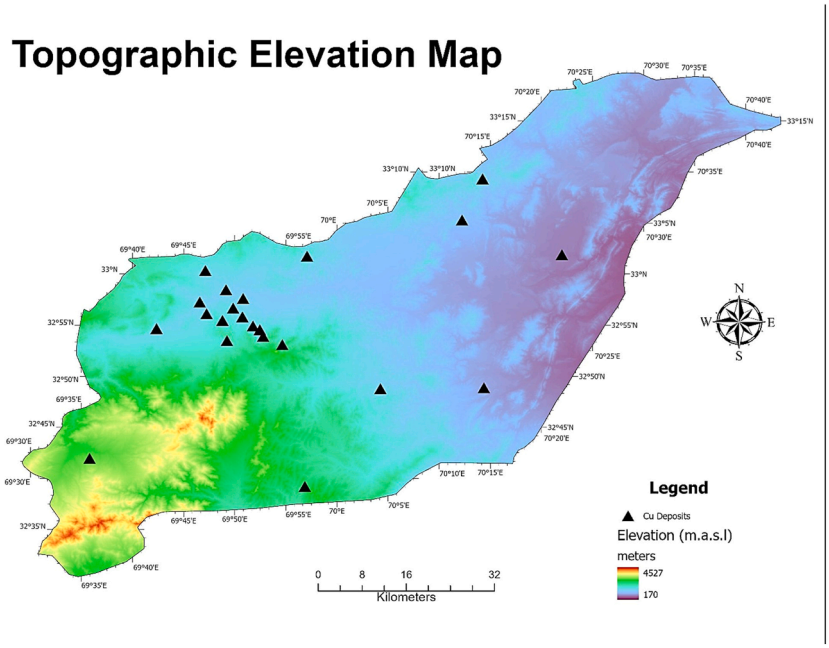

Fig. 7. The topographic elevation of the study area emphasises the predominantly difficult mountainous environment.

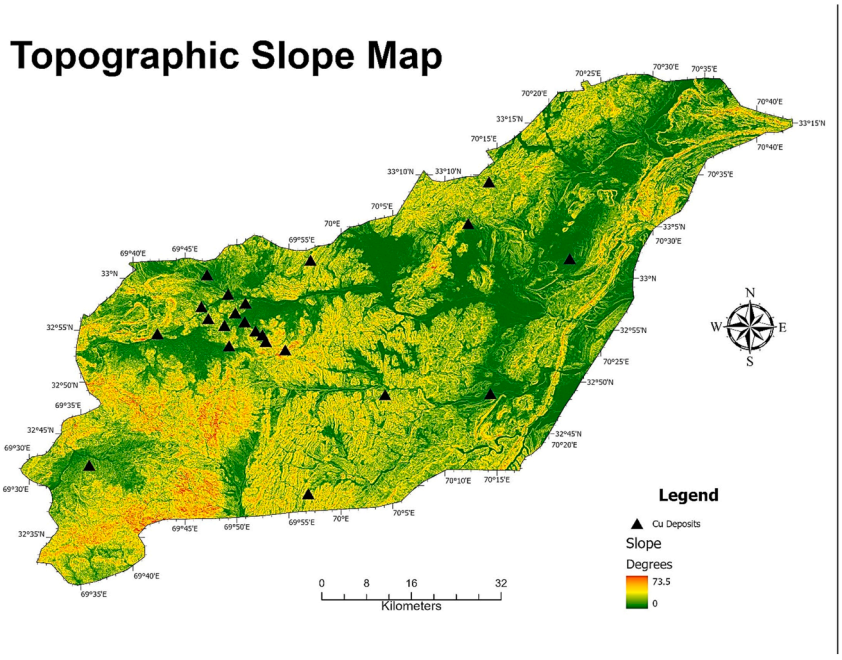

Fig. 8. The topographic slope of the study area, highlighting the western high slopes.

raphy Mission (SRTM) based 1-arc second Digital Elevation Model (DEM) with a spatial resolution of about 30 m as it has been used in several research studies.

2.3.2. Target variable

During training and testing, the occurrence of mineral deposits is represented as a binary variable, with a value of 1 if the deposit exists and 0 if it does not. The 22 known Cu deposits were used as positive samples for the presence of mineralization. The following criteria were used to select sample locations for non-deposit occurrences; first to ensure adequate training data and an equitable balance of positive and negative samples, the quantity of non-deposit and deposit samples must be equal. This will allow for an accurate assessment of the ML algorithms' prediction capability. Second, non-deposit locations should be far away from any known deposits,

since those near to known deposits may exhibit similar characteristics to those of the Cu mineralization and thus have a high chance of new deposit discoveries. Point pattern analysis was applied to assess the adequacy of the distance between deposit locations. The distances between each deposit as well as its closest adjacent deposit were computed, and the findings were statistically assessed and shown in Fig. 9. It is evident that all the deposits are within 8247 m of each other, indicating a 100% likelihood of another deposit being present within that range for any given deposit. As deposits are unlikely to be found beyond 8247 m, a buffer distance of 5200 m, with an 80% chance of finding deposits near any identified deposit, was established.

The non-deposit samples must be chosen outside of known deposit areas. Favorable regions within a 3500 m radius of intrusive rocks were identified through the distance distribution analysis conducted in the area. Several buffer intrusions appear to coincide with delineated areas of known deposit sites, indicating areas of extensive Cu mineralization for further exploration. The other buffered zones of intrusions outside of known deposits may provide favorable indicators in underexplored locations. Hence, the locations of non-deposit should not only be selected outside of any buffer zone, but also distant from known deposit areas. Non-deposit locations should be randomly dispersed, as opposed to mineral deposits, which are the outcome of rare events and non-random mineralization processes that tend to cluster spatially. As target samples, 22 non-deposit locations were chosen at random using the criteria in Fig. 10.

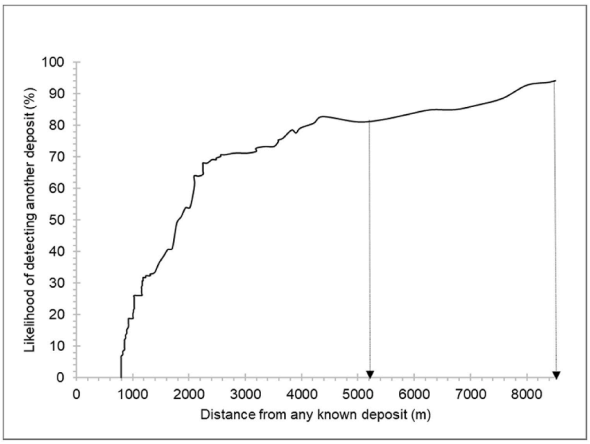

Fig. 9. Point pattern analysis indicating the likelihood of discovering another Cu deposit relative to known Cu deposits.

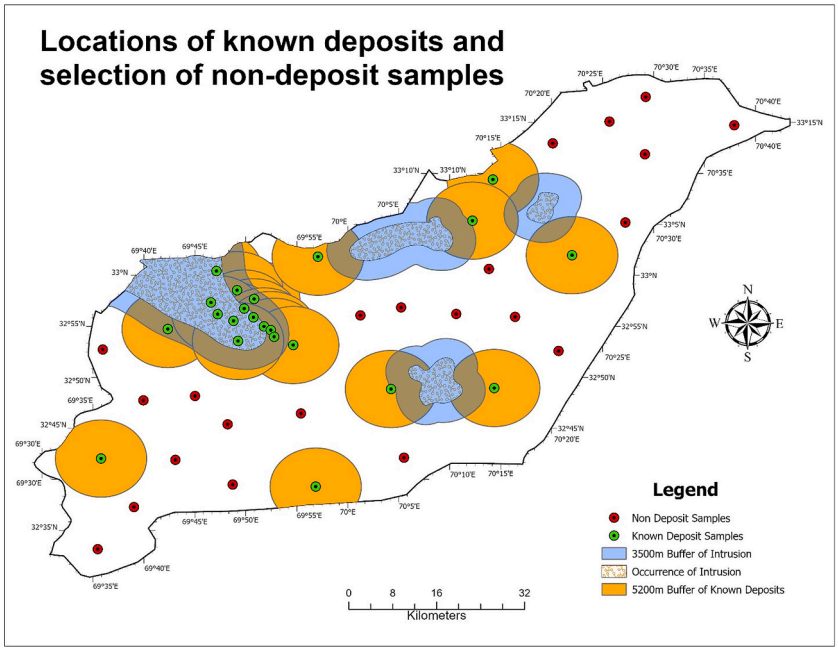

Fig. 10. Point pattern analysis indicating the likelihood of discovering another Cu deposit relative to known Cu deposits.

2.3.3. Input dataset

Maps of evidentiary features need to be converted into raster data with numerical values representing the evidence in each cell before applying prospectivity modelling. A methodology proposed by Carranza and Laborte (2015) was used to objectively select the cell size. According to the point pattern analysis, the shortest distance between two deposits is 900 m, indicating that using a cell size larger than 900 m will likely involve multiple deposits. Thus, a suitable cell size should be within the range of 900 m, as the upper limit. The minimum cell size can be determined based on the map scale of the spatial information. The best readable spatial resolution (S_R) can be calculated through equation (7).

$$S_R = MS \times 0.00025 \tag{7}$$

The MS (scale of the map) determines the minimum cell size, which is 250 m. Finally, the raster maps were generated with 40,500 pixels, by using a cell size of 400 m within the appropriate size range. Notably, the original geographic data was collected using separate surveys at a lower scale, yielding in raster images that may not provide a single sample per cell. The spatial interpolation methods of Inverse Distance Weighted were utilized in the rasterizing processes, which influenced the accuracy of feature delineation. In most Geographical Information Systems (GIS) based mineral prospectivity modelling applications, the predictor maps were combined into a single grid data instead of being used directly as an input for creating prospectivity models (Rodriguez-et al., 2015; Zuo et al., 2021). The nine predictor maps were integrated into a single integral map, along with an associated feature dataset, using the digital overlay approach in GIS. This database contains 40,500 records, each containing a nine-dimensional input vector for training. After the training processes were done, the results were assigned to a new field in the same table.

From the 40,500 cells, 44 cells having occurrence sites (deposit and non-deposit) were extracted to create a labelled dataset. While 66% of the extracted data was utilized for training the CNN, RF, and SVM models, 34% was retained as a test dataset to evaluate the accuracy of these models.

2.3.4. Model training

Choosing the right criteria to train a model is a critical aspect of predictive modelling, necessary to ensure reliable predictions. Nonetheless, it can be challenging to determine the best configuration for achieving the desired accuracy, as there is no defined method for determining the appropriate parameters for every given case. This study uses a 10-fold cross-validation technique to assess the predictions made by different combinations of parameters. While empirical terms are helpful, finding the ideal arrangement requires a very subjective trial-and-error approach (Sun et al., 2020). The input training dataset is divided into ten equal-sized parts, one of which is selected as the validation dataset and the other nine are selected as the training dataset. This step was repeated ten times to ensure that all subsets were only used once as the validation data set (Raschka, 2018) and to balance bias and variance. The mean squared error was used to assess cross-validation results, which can be expressed as equation (8).

$$MSE = \frac{1}{D_N} \sum_{i=1}^{D_N} \left(\widehat{P}_i - P_i \right)^2 \tag{8}$$

For a validation dataset containing D_N data points, $\widehat{P}_i$ represents the predicted class value (deposit as 1, non-deposit as 0) and P_i is the actual class value of each target data. Given the consideration that 10-fold cross-validation typically requires a larger dataset, alternative cross-validation techniques such as Monte Carlo cross-validation and 5-fold cross-validation could also be utilized. Monte Carlo cross-validation involves multiple random splits of the data into training and validation sets, providing flexibility and robustness for smaller datasets. Similarly, 5-fold cross-validation, which divides the data into five parts, reduces the computational burden while still offering reliable performance estimates. Furthermore, 10-fold cross-validation offers a good balance between computational efficiency and model evaluation rigor, rendering it a suitable choice for the dataset and objectives of this study. The optimal configuration was established by identifying the model with the least Mean Squared Error (MSE). Table 3 lists the parameters of each ML model, their characteristics, as well as the recommended range of parameter values indicated by different authors (Sothe et al., 2020; Imamverdiyev and Sukhostat, 2019; Ghorbanzadeh et al., 2019; Adem, 2022) for CNN; (Sothe et al., 2020; Tuba et al., 2017; Xu et al., 2019b; Imran et al., 2022) for SVM and (Yin and Li, 2022; Carranza and Laborte, 2015; Agrawal et al., 2022) for RF. The hyperparameters for each model were rigorously optimized. For SVM, we explored multiple kernel functions and optimized the gamma and C parameters using grid search cross-validation. For CNN, we started with a basic architecture and optimized the number of layers, fil-

Table 3
Training hyper-parameters for machine learning models.

Model	Parameters	Description	Reference Range
RF	Number of trees	Number of trees in a random forest	10–500
	Number of features	Number of features utilized to construct each tree	1–8
	Maximum depth	Maximum number of iterations to split	2–20
	Minimum leaf size	Minimum sample size in each leaf node	1–20
SVM	Gamma	RBF width parameter determining the affecting range of each support vector	0.1–1
	Cost	Penalty for incorrect classification	0.1–50
CNN	Number of feature maps	Convolutional filters utilized to create new feature maps	8–64
	Number of neurons	The total neurons present in the connected layer	8–64

ter sizes, and dropout rates. RF parameters, including the number of trees and maximum depth, were also fine-tuned to ensure model robustness.

Information Gain (IG) was utilized to evaluate the impact of input features on the trained model (Yang et al., 2022b). The formula in equation (9) can compute the value of IG for a feature E_f associated with output class C (deposit or non-deposit).

$$IG\ \left(C, E_f\right) = H\left(C\right) - H\left(C\middle|E_f\right) \tag{9}$$

Where $H(C)$ denote the entropy value of C, and $H\ (C|E_f)$ be the entropy value of C when evidentiary features E_f have been assigned values.

2.3.5. Model evaluation

To evaluate the ML model's efficacy thoroughly, the research employed a confusion matrix, predictive accuracy metrics, a ROC curve, and a success-rate curve. The confusion matrix can effectively summarize the predictions of the model. It shows four outcomes of classification, i.e., TP (true positive; deposit sample identified correctly as a deposit), FN (false negative; deposit sample misclassified as a non-deposit), FP (false positive; non-deposit sample incorrectly classified as a deposit) and TN (true negative; non-deposit sample accurately identified as such) (Sun et al., 2019). The confusion matrix was used to measure the predictive performance of all the models using a range of statistical indices. These indices were formulated as given in equation (10) (Maria et al., 2016; Beguería, 2006).

$$\begin{aligned}
&Sensitivity = \frac{TP}{TP + FN}\\
&Specificity = \frac{TN}{TN + FP}\\
&Positive\ predictive\ value = \frac{TP}{TP + FP}\\
&Negative\ predictive\ value = \frac{TN}{TN + FN}\\
&Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
\end{aligned} \tag{10}$$

The effectiveness of ML models can be assessed through ROC and success-rate curves. A ROC curve depicts how a binary classification system performs when the discrimination threshold is changed. It illustrates how well the system can distinguish between two classes. The curve is produced by plotting the sensitivity (TPR) on the y-axis and the false positive rate (FPR, 1-specificity) on the x-axis with various thresholds. The threshold establishes the criterion for distinguishing the predictive outcomes. Cells with values higher or less than the limit were labelled as “deposit” or “non-deposit” respectively. A set of data points (TPR, FPR) were collected by varying the threshold in order to generate the ROC curve. In general, the closer the ROC curve is to the upper left corner, the better the model's performance will be. The Area Under the Curve (AUC) was used to quantify the performance of various predictive models. This metric ranges from 0 to 1, with a score of 1 indicating a perfect accuracy, i.e., a sensitivity of 1 and a specificity of 0 whereas a value of 0.5 would signify a completely random model. The success rate curve is derived by comparing the percentage of deposits in the target area with the total number of deposits at different thresholds.

3. Results and discussion

3.1. Satellite derived hydrothermal alterations

The hydrothermal alterations associated with the potential identification of Cu mineralization contain critical diagnostic information at the RGB, NIR and SWIR regions of the electromagnetic spectrum. The spectral bands of Landsat-8 and Sentinel-2 data demonstrate a strong capability in detecting hydrothermal alterations. The results of Iron-oxide alteration, Argillic alteration, Phyllic alteration and Propylitic alteration based on Sentiel-2 are shown in Fig. 11. Iron-oxide alteration is a key geological process in developing and exploitation of copper deposits. Iron-oxide alteration is distinguished by the presence of iron oxides such as hematite and magnetite and are formed by the reaction of hot fluids with iron-bearing minerals in rocks. Other minerals, such as feldspar, may also be replaced by iron oxides as a result of the process. Because the fluids that induce the alteration frequently contain large amounts of copper, there is a strong relationship between iron-oxide alteration and copper mineralization (Khaleghi et al., 2020). As these fluids interact with the rocks, copper minerals such as chalcopyrite, bornite, and chalcocite are deposited. These copper minerals are frequently discovered in the altered rocks that surround the iron-oxide alteration zone, which is highlighted in tone of purple colors of Map of Iron-Oxide alteration in Fig. 11. Argillic alteration is a significant geological process that contributes to the development of copper deposits and serves as a crucial indicator for geologists in their search of discovering new copper deposits (Beygi et al., 2021). Argillic alteration is a geological process that develops when hydrothermal fluids come into contact with rocks that comprise minerals such as feldspar, mica, and others. The minerals present in the rocks undergo alteration due to the previously mentioned interaction, leading to the formation of a type of clay mineral known as kaolinite. This phenomenon has the potential to spread across a vast area and possibly extend to considerable depths below the surface. Fig. 11 depicts the argillic alteration, which is visually distinguished through chromatic variations ranging from cyan to pink shades. Phyllic alteration is another hydrothermal alteration process that is frequently related with the formation of porphyry copper deposits. This process transforms primary minerals into a suite of minerals that commonly includes quartz, sericite, and pyrite. The existence of phyllic alteration zones is often an indication of the possible oc-

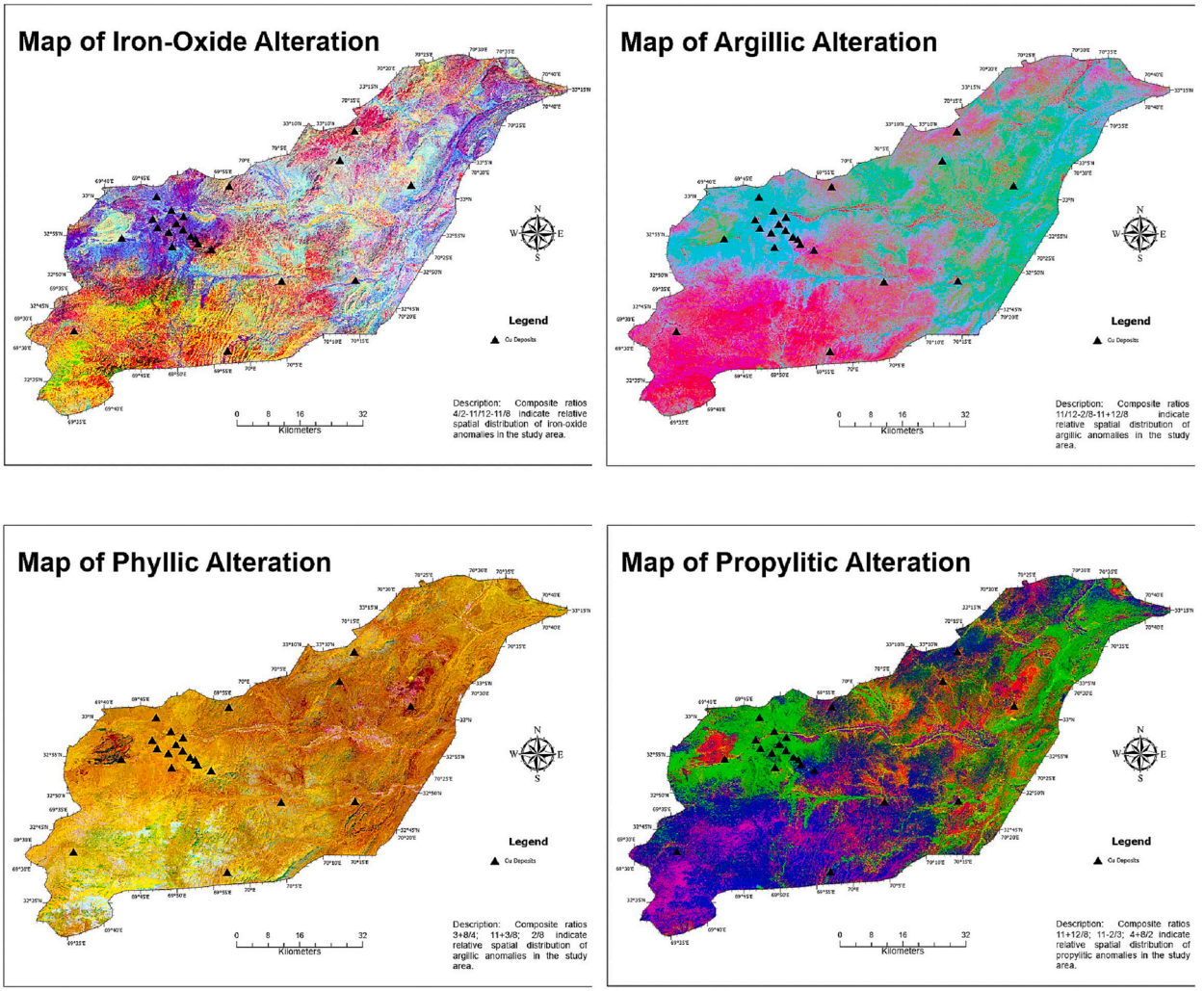

Fig. 11. Hydrothermal maps of the study area used to map hydrothermal alternation zones associated with Cu mineralization based on Sentinel-2 satellite data.

currence of porphyry copper mineralization. The phyllic alteration is highlighted in Fig. 11 as light orange to yellowish in color. Another alteration of rocks produced by hydrothermal fluids is referred to as propylitic alteration. These fluids contain a high concentration of water, carbon dioxide, and other minerals that react with the host rock to generate new minerals. Propylitic anomalies are zones of propylitic alteration seen in rocks around copper deposits (Parcutela et al., 2022). These anomalies are distinguished by the presence of minerals indicative of hydrothermal alteration, such as chlorite, epidote, and actinolite. The existence of these minerals indicates that the rocks were changed by hydrothermal fluids migrating from the copper deposit (Parcutela et al., 2022). The propylitic alteration is highlighted in Fig. 11 because it is greenish-gray in color and has a fine-grained texture on the ground.

3.2. Sensitivity analysis

Parameters configured for model training can have a major effect on the reliability and generalizability of ML techniques, which significantly influences the predictions accuracy. The 10-fold cross-validation results revealed significant variations in the misclassification rates of the three models when different parameter settings are used which can be seen through overall accuracy results as shown in Table 4.

The different parameters yield different levels of classification precision when training individual ML models. However, RF and CNN models show similar variation patterns of MSE in response to changes in the specific parameter, i.e., no significant change was observed in the MSE of both models. It does not indicate that the MSE values were comparable. Generally, these two models demonstrate more accurate and stable performance with overall accuracy values greater than 0.8 and are less sensitive to parameter variations. On the other hand, SVM models produce inadequate results with an average overall accuracy of 0.73. The complexity of a model can be determined by the number of ML parameters associated with its architecture, such as the number of trees in a Random Forest or the number of feature maps and neurons in a Convolutional Neural Network. Many research studies (Shirmard et al., 2022; Jung and Choi, 2021; Agrawal et al., 2022; Sun et al., 2019; Caruana et al., 2015) have established that more precise predictions are

Table 4
Analysis of 10-fold cross-validation showing Overall Accuracy results.

Model	Overall Accuracy			
	Minimum	Maximum	Mean	Standard deviation
CNN	0.79	0.89	0.86	±0.03
SVM	0.70	0.79	0.73	±0.02
RF	0.79	0.85	0.82	±0.02

often produced by complex ML models. This study investigated the influence of architecture-related parameters on the MSE results of prospectivity modelling. The result revealed that increasing the complexities of the model did not increase the efficacy of prospectivity modelling in the area. As the parameters increase, the statistical indices used to quantify classification error, such as the lowest and average MSE, have not decreased significantly. Moreover, the use of a greater number of trees, neurons, and feature maps increases computation power. Consequently, it is suggested that the modelling processes discussed here do not necessarily require complex ML model architectures. These findings are also consistent with those from earlier research (Rodriguez-et al., 2015; Sun et al., 2019) which showed that complex models did not necessarily lead to improved accuracy of predictions. The limited size of training datasets, typically consisting of only a few deposit and non-deposit places, may explain the achieved accuracy in these results. Simple architectures and an appropriate amount of training can help avoid over-fitting errors that could be caused by using more complex models or extensive training.

3.3. Selection of prospectivity map

The optimized models will provide a probability score for each cell, assigned as a floating-point value ranging from 0 to 1, indicating the likelihood of a mineral presence. ML algorithms classify cells with probability values greater than 0.5 as areas with potential mineralization, while those below are marked as barren and without prospecting likelihood.

The classification accuracies of the three ML models are represented by the confusion matrices in the training and test datasets. Table 5 displays the confusion matrices for the three ML models, which are used to quantitatively measure the accuracy of binary classification through various statistical indices listed in Table 6.

The CNN model has overall accuracy greater than the RF and SVM models while predicting positive and negative samples. Notably, SVM produces poorer predictions in both training and testing, particularly in the testing process that misclassifies most of the deposits. The CNN model was found to be the most sensitive by correctly identifying 87% of deposit locations, followed by the RF model (85%). The CNN model shows a specificity of 85%, which indicates that 85% of cells without mineralization being correctly identified as non-deposits. Similarly, the RF and SVM models also achieved reasonable specificities of 75% and 71%, respectively. The CNN and RF models successfully predicted Cu occurrences in the majority of the predicted cells (>80%), demonstrating a high positive predictive value. The CNN and RF models had the highest negative predictive rate (85%), indicating that 85% of predicted non-deposit cells are true non-deposit samples followed by SVM (60%). In terms of overall accuracy, the CNN model achieved the best value of 80%, indicating that 80% of all samples were correctly classified, and the same pattern was also followed by RF (80%).

The ROC curve was used to assess the predictive capability of high-probability zones. The discriminative thresholds were changed from high to low probability values for evaluation purposes. The ROC curve with a y-value of 1 and an area under the curve (AUC) of 1 means the probability of occurrence samples is greater than that of non-occurrence samples. Any deviation from this line, whether due to a non-occurrence cell with a high probability or an occurrence cell with a low probability, reduces the AUC value.

The RF model has the ROC curve closest to the upper left corner of the graph, indicating the highest predictive capability with AUC values greater than 0.95. The AUC values of the CNN model are similarly good, while the SVM model shows the least accurate performance in terms of ROC curves as shown in Fig. 12.

Table 5
Confusion matrix of machine learning algorithms in test procedures.

	RF		CNN		SVM	
	Actual Deposit	Actual Non-Deposit	Actual Deposit	Actual Non-Deposit	Actual Deposit	Actual Non-Deposit
Predicted deposit	6	2	7	1	3	5
Predicted non-deposit	1	6	1	6	4	3

Table 6
Predictive efficiency of machine learning models.

	RF	CNN	SVM
Sensitivity	85%	87%	75%
Specificity	75%	85%	71%
Positive predictive value	75%	85%	38%
Negative predictive value	85%	85%	60%
Accuracy	80%	80%	60%

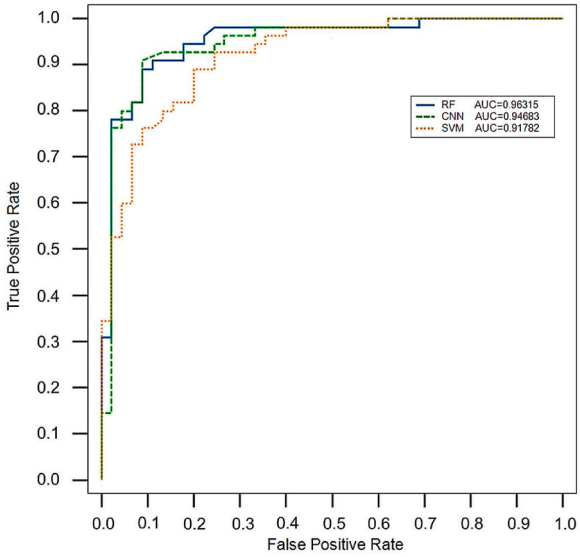

Fig. 12. The ROC curve and AUC analysis of RF, CNN and SVM models.

The majority of ML models can accurately predict >80% of known events under the initial classification parameters. Nonetheless, exploring all these "prospective areas" is too costly for "real world" mineral exploration, thus making target regions with high confidence and minimal area the ideal option. Hence it is vital for such applications to evaluate the demarcation of these zones as well as the effectiveness of the models. The success rate curve analysis highlights the performance of models in making predictions, with a steeper slope indicating a higher level of efficiency. Identifying the cut-off limits indicated by the three regression lines, which signify high, moderate and low probability for finding new Cu deposits, can be used to carry out prospectivity zoning. This enables delineated areas to capture more known mineral occurrences. Areas with a success rate curve slope more than five are considered high potential and are suitable for exploration. Areas with a slope equal to or less than the natural distributed density (for example, 1% of the area containing 1% of mineral occurrences) are not considered prospective. The success-rate curves show that the RF model has the highest slope, making it the most efficient in predicting outcomes. In the RF model, 8.99% of the study area accounted for 63.46% of the known deposits, compared to 5.98% for the CNN model containing 57.67% of the deposits and 3.14% for the SVM model containing 53.85%. The RF model identified 18 known deposits (81.81%) within 14.79% of the study area, including very high and high potential zones. CNN and SVM generated similar results, with 16 and 14 deposits respectively, but covered a larger area (26.18%).

Based on the threshold values for the predictive models, the area was classified into potential zones for discovering new deposits as shown in Fig. 13.

The CNN model was found to be the best to correctly identify the Cu deposits in the study area, however it is not the best predictor. One of the possible explanations of this can be associated with the availability of limited field dataset. The RF model, on the other hand, is both a good identifier and the best predictor of Cu deposits, making it the optimal choice for exploration targeting since predictive accuracy is of great importance for further mineral exploration decisions. The areas with very high potential of new Cu deposits were also in the close vicinity of the geological lineaments and mostly present in the suitable hosting geological rocks. The predicted zones are also aligned with the hydrothermal alterations as extracted by processing of satellite remote sensing data. The ongoing mining in the region also falls under the very high potential zones towards the North-West side as predicted under this research study (Fig. 14). Hence the predictive zones highlighted by the RF model were used as the final prospectivity map. The RF model showed the most impressive results in comparison to other ML based Mineral Prospectivity Mapping or Modelling (MPM) applications. Several research studies (Shirmard et al., 2022; Jung and Choi, 2021; Rodriguez-et al., 2015; Woodhead and Landry, 2021) has also reported that the RF model has outperformed other ML models in the application of machine learning for minerals exploration and mapping. This can be due to the random sampling conditions employed by the RF training, which provides a high-level of diversity with limited data and prevents over-fitting.

Although the hydrothermal alteration maps were produced from satellite data, field-based investigations for the verification of host rocks are extremely important and should be explored for future research in the region. To further enhance the identified potential Cu mineralization prospects, more detailed geological information, particularly data on the geochemistry of rocks and soil, needs to be collected and analyzed.

4. Conclusions

In this study, a data science approach was applied based on ML algorithms including CNN, RF, and SVM. A set of ML models, including RF, SVM and CNN were trained based on nine predictor maps using different parameters. During the training and testing process, both the CNN and RF models demonstrated high accuracy and consistency despite any changes to their parameters. According to the sensitivity analysis, model design with complex structure does not necessarily improve the prediction accuracy of MPM in

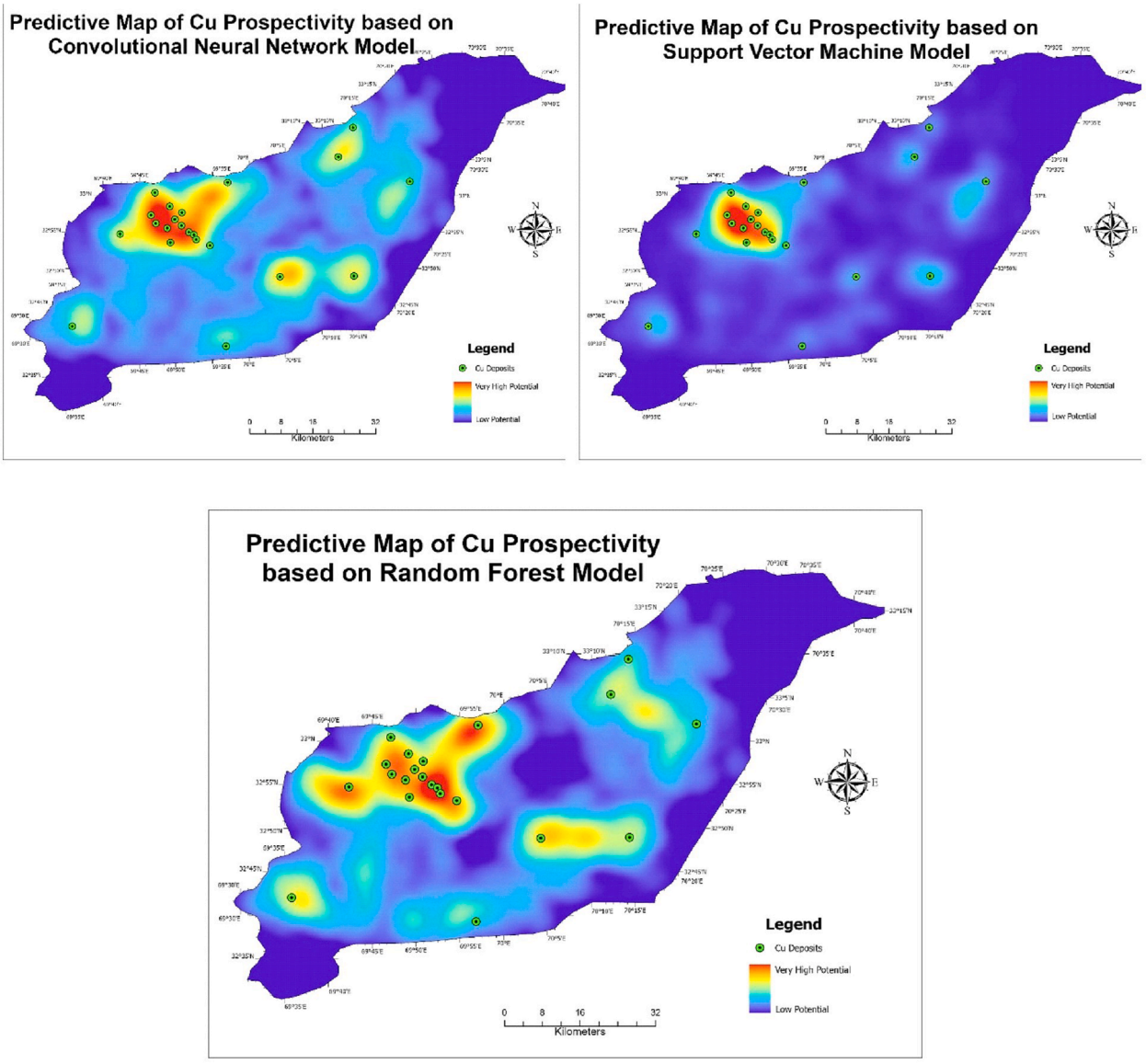

Fig. 13. Predictive maps of CNN, SVM, and RF models highlighting low to very high potential areas delineated based on threshold values.

the study area. In terms of classification accuracy, the CNN model achieves the highest level of accuracy, with 92.38% of labelled samples correctly identified, while RF and SVM trail closely behind. The success rate curves showed that the RF model can detect a greater number of Cu deposits in smaller, well-defined regions, resulting in the most promising predictive accuracy. Given the model evaluation measurements obtained, the RF model stands out as the best ML model for developing prospectivity maps for further research in the study area. The study reveals that satellite-derived hydrothermal alteration maps, proximity to intrusive contacts, and the number and density of geological fault intersections are the most influential criteria for Cu prospectivity predictions. The findings suggest that Cu hydrothermal alterations in the host rocks along with ML based models represent an underutilized exploration technique, providing important information on the origin of Cu mineralization.

The recent discovery of a known Cu deposit through field mineral exploration in the study area validates the predictive efficiency of the ML models applied in this study. The data science approach proposed and evaluated in this study for mineral prospectivity prediction has huge significance, not only for delineating of known Cu deposits, but also for exploring other deposits related to the remote sensing satellite derived hydrothermal mineral system on a larger scale.

The study provides significant insights into mineral prospectivity mapping, but several factors may influence the model outcomes and their generalizability. The focus on the North Waziristan region limits the applicability of findings to other geological settings. The choice of algorithms (RF, SVM, CNN) was effective but constrained by computational limits. Furthermore, the spatial resolution of the remote sensing data may have limited the detection of small-scale geological features, resulting in incomplete predictions. Considering these factors, future studies should incorporate more comprehensive field data such as LiDAR and hyperspectral imaging, ex-

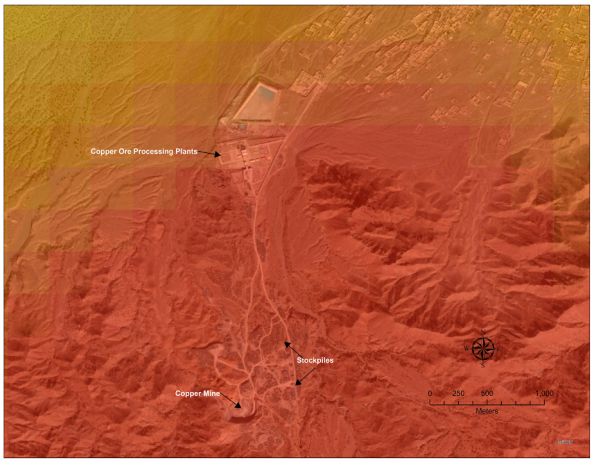

Fig. 14. Current copper mining in the study area lies within the very high potential mineralization zone, as predicted by the RF model.

plore other advanced modelling techniques, and consider temporal variability and higher resolution data to improve the accuracy and generalizability of predictive models in mineral prospectivity mapping.

Declaration of generative AI and AI-assisted technologies

The authors did not use AI and AI-assisted technologies in the writing of this manuscript.

Funding

This research received no external funding.

CRediT authorship contribution statement

Muhammad Ahsan Mahboob: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Turgay Celik:** Writing – review & editing, Visualization, Supervision. **Bekir Genc:** Writing – review & editing, Visualization, Supervision.

Declaration of competing interest

We, the authors of this research paper, declare that we have no financial or non-financial conflicts of interest related to the research presented in this manuscript. Our work is solely driven by the pursuit of scientific inquiry and the advancement of knowledge in our respective fields.

Acknowledgments

The work presented here is part of a PhD research work in the School of Mining Engineering at the University of the Witwatersrand. We would like to thank Professor Sarfraz Ali (CEO of GeoSync, Pakistan) for providing the field data and results verification used in this study.

Data availability

Data will be made available on request.

References

Abedi, M., Norouzi, G.-H., Bahroudi, A., 2012. Support vector machine for multi-classification of mineral prospectivity areas. Comput. Geosci. 46, 272–283.
Adem, K., 2022. Impact of activation functions and number of layers on detection of exudates using circular Hough transform and convolutional neural networks. Expert Syst. Appl. 203, 117583.
Adiri, Z., Lhissou, R., El Harti, A., Jellouli, A., Chakouri, M., 2020. Recent advances in the use of public domain satellite imagery for mineral exploration: a review of Landsat-8 and Sentinel-2 applications. Ore Geol. Rev. 117, 103332.
Agrawal, N., Govil, H., Chatterjee, S., Mishra, G., Mukherjee, S., 2022. Evaluation of machine learning techniques with AVIRIS-NG dataset in the identification and mapping of minerals. Adv. Space Res.
n a heterogeneous reservoir: a comparative study. Comput. Geosci. 36, 2010, 1494–1503.
Atwizukye, T., 2022. Using Satellite Based Remote Sensing in Copper Ore Exploration, A Case Study of Kasese District. Makerere University.
Bauer, T.E., Lynch, E.P., Sarlus, Z., Drejing-Carroll, D., Martinsson, O., Metzger, N., Wanhainen, C., 2022. Structural controls on iron oxide copper-gold mineralization and related alteration in a paleoproterozoic supracrustal belt: insights from the nautanen deformation zone and surroundings, northern Sweden. Econ. Geol. 117, 327–359.
Beguería, S., 2006. Validation and evaluation of predictive models in hazard assessment and risk management. Nat. Hazards 37, 315–329.
Benaissi, L., Tarek, A., Tobi, A., Ibouh, H., Zaid, K., Elamari, K., Hibti, M., 2022. Geological mapping and mining prospecting in the Aouli inlier (Eastern Meseta,

Morocco) based on remote sensing and geographic information systems (GIS). China Geology 5, 614–625.
Beygi, S., Talovina, I.V., Tadayon, M., Pour, A.B., 2021. Alteration and structural features mapping in Kacho-Mesqal zone, Central Iran using ASTER remote sensing data for porphyry copper exploration. International Journal of Image and Data Fusion 12, 155–175.
Blandine, K.T.A., Jules, T.K., Martial, F.E., Ludovic, A.M., Julios, E.A., Robinson, S.B., Ousmanou, S., Maurice, K., 2023. Geological mapping and structural interpretation of the Dschang-Santchou-escarpment (West, Cameroon), using Landsat 8 OLI/TIRS sensors/SRTM and field observations. Geol. J. 58, 1111–1130.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Carranza, E.J.M., 2008. Geochemical Anomaly and Mineral Prospectivity Mapping in GIS. Elsevier.
Carranza, E.J.M., 2009. Controls on mineral deposit occurrence inferred from analysis of their spatial pattern and spatial association with geological features. Ore Geol. Rev. 35, 383–400.
Carranza, E.J.M., Laborte, A.G., 2015. Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: application of Random Forests algorithm. Ore Geol. Rev. 71, 777–787.
Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N., 2015. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1721–1730.
Chen, Q., Xia, J., Zhao, Z., Zhou, J., Zhu, R., Zhang, R., Zhao, X., Chao, J., Zhang, X., Zhang, G., 2022. Interpretation of hydrothermal alteration and structural framework of the Huize Pb–Zn deposit, SW China, using Sentinel-2, ASTER, and Gaofen-5 satellite data: implications for Pb–Zn exploration. Ore Geol. Rev. 105154.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.
da Silva, G.F., Silva, A.M., Toledo, C.L.B., Junior, F.C., Klein, E.L., 2022. Predicting mineralization and targeting exploration criteria based on machine-learning in the Serra de Jacobina quartz-pebble-metaconglomerate Au-(U) deposits, São Francisco Craton, Brazil. J. S. Am. Earth Sci. 116, 103815.
Fu, Y., Cheng, Q., Jing, L., Ye, B., Fu, H., 2023. Mineral prospectivity mapping of porphyry copper deposits based on remote sensing imagery and geochemical data in the duolong ore district, tibet. Rem. Sens. 15, 439.
Geranian, H., Tabatabaei, S.H., Asadi, H.H., Carranza, E.J.M., 2016. Application of discriminant analysis and support vector machine in mapping gold potential areas for further drilling in the Sari-Gunay gold deposit, NW Iran. Nat. Resour. Res. 25, 145–159.
Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S.R., Tiede, D., Aryal, J., 2019. Evaluation of different machine learning methods and deep-learning convolutional neural networks for landslide detection. Rem. Sens. 11, 196.
Grunsky, E., 1996. Spatial data integration for mineral exploration, resource assessment and environmental studies-A guidebook. By International Atomic Energy Agency. NONRENEWABLE RESOURCES 5, 78–79.
Imamverdiyev, Y., Sukhostat, L., 2019. Lithological facies classification using deep convolutional neural network. J. Petrol. Sci. Eng. 174, 216–228.
Imran, M., Ahmad, S., Sattar, A., Tariq, A., 2022. Mapping sequences and mineral deposits in poorly exposed lithologies of inaccessible regions in Azad Jammu and Kashmir using SVM with ASTER satellite data. Arabian J. Geosci. 15, 538.
Jung, D., Choi, Y., 2021. Systematic review of machine learning applications in mining: exploration, exploitation, and reclamation. Minerals 11, 148.
Keykhay-Hosseinpoor, M., Kohsary, A.-H., Hossein-Morshedy, A., Porwal, A., 2020. A machine learning-based approach to exploration targeting of porphyry Cu-Au deposits in the Dehsalm district, eastern Iran. Ore Geol. Rev. 116, 103234.
Khaleghi, M., Ranjbar, H., Abedini, A., Calagari, A.A., 2020. Synergetic use of the Sentinel-2, ASTER, and Landsat-8 data for hydrothermal alteration and iron oxide minerals mapping in a mine scale. Acta Geodyn. Geromater 17, 311–329.
Khan, E., 2000. Flotation of copper minerals from north waziristan copper ore, on pilot-scale. Quarterly Sci. Vision 6, 10–20.
Khan, S.R., Jan, M.Q., Khan, T., Khan, M.A., 2007. Petrology of the dykes from the waziristan ophiolite, NW Pakistan. J. Asian Earth Sci. 29, 369–377.
Khan, I.U., Khan, A., Ullah, A., 2022. Causes and factors responsible for Operation Zarb-e-Azb: perspective of internally displaced persons of North Waziristan, Pakistan. Liberal Arts and Social Sciences International Journal (LASSIJ) 6, 181–200.
Köhler, M., Hanelli, D., Schaefer, S., Barth, A., Knobloch, A., Hielscher, P., Cardoso-Fernandes, J., Lima, A., Teodoro, A.C., 2021. Lithium potential mapping using artificial neural networks: a case study from central Portugal. Minerals 11, 1046.
Kong, Y., Chen, G., Liu, B., Xie, M., Yu, Z., Li, C., Wu, Y., Gao, Y., Zha, S., Zhang, H., 2022. 3D mineral prospectivity mapping of zaozigou gold deposit, west qinling, China: machine learning-based mineral prediction. Minerals 12, 1361.
Krupski, J., Graniszewski, W., Iwanowski, M., 2021. Data transformation schemes for cnn-based network traffic analysis: a survey. Electronics. 10, 2042.
Lim, E.-P., Foo, S., Khoo, C., Chen, H., Fox, E., Shalini, U., Thanos, C., 2002. Digital libraries: people, knowledge, and technology. In: 5th International Conference on Asian Digital Libraries, ICADL 2002, Singapore, December 11-14, 2002, Proceedings, vol. 2555. Springer Science & Business Media.
Mahboob, M., Genc, B., Celik, T., Ali, S., Atif, I., 2019. Mapping Hydrothermal Minerals Using Remotely Sensed Reflectance Spectroscopy Data from Landsat, vol. 119. Journal of the Southern African Institute of Mining and Metallurgy, pp. 279–289.
Mahboob, M., Celik, T., Genc, B., 2022. Review of Machine Learning-Based Mineral Resource Estimation, vol. 122. Journal of the Southern African Institute of Mining and Metallurgy, pp. 655–664.
Malkani, M.S., Mahmood, Z., Alyani, M.I., Siraj, M., 2017. Mineral resources of khyber Pakhtunkhwa and FATA, Pakistan. Geological Survey of Pakistan, Information Release 996, 1–61.
Maria Navin, J., Pankaja, R., 2016. Performance analysis of text classification algorithms using confusion matrix. Int. J. Eng. Tech. Res. IJETR 6, 75–78.
Mehsud, S., 2012. Combating militancy in bajaur and North-waziristan agency in federally administered tribal areas (FATA) of Pakistan: a comparative analysis. Tigah. J. Peacebuilding Dev. . FATA Research Centre, Islamabad.
Namdeo, A., Singh, D., 2021. Challenges in evolutionary algorithm to find optimal parameters of SVM: a review. Mater. Today: Proc.
Nasab, M.H., Agah, A., 2023. Mapping hydrothermal alteration zones associated with copper mineralization using ASTER data: a case study from the mirjaveh area, southeast Iran. Transactions: Basics 36, 720.
Parcutela, N., Dimalanta, C., Armada, L., Austria, R., Gabo-Ratio, J., Yumul, G., 2022. Band processing of Landsat 8-OLI multi-spectral images as a tool for delineating alteration zones associated with porphyry prospects: a case from Suyoc, Benguet, Philippines. In: Proceedings of the IOP Conference Series: Earth and Environmental Science. 012022.
Pour, A.B., Hashim, M., 2012. The application of ASTER remote sensing data to porphyry copper and epithermal gold deposits. Ore Geol. Rev. 44, 1–9.
Rajan Girija, R., Mayappan, S., 2019. Mapping of mineral resources and lithological units: a review of remote sensing techniques. International Journal of Image and Data Fusion 10, 79–106.
Rajesh, H., 2004. Application of remote sensing and GIS in mineral resource mapping-An overview. J. Mineral. Petrol. Sci. 99, 83–103.
Raschka, S., 2018. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv preprint arXiv:1811.12808*.
Robinson, R., 2018. Available online: https://mlnotebook.github.io/post/CNN1/. 21 March.
Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818.
Santos, D., Cardoso-Fernandes, J., Lima, A., Müller, A., Brönner, M., Teodoro, A.C., 2022. Spectral analysis to improve inputs to random forest and other boosted ensemble tree-based algorithms for detecting NYF pegmatites in Tysfjord, Norway. Rem. Sens. 14, 3532.
Sekandari, M., Masoumi, I., Beiranvand Pour, A., M Muslim, A., Rahmani, O., Hashim, M., Zoheir, B., Pradhan, B., Misra, A., Aminpour, S.M., 2020. Application of landsat-8, sentinel-2, ASTER and WorldView-3 spectral imagery for exploration of carbonate-hosted Pb-Zn deposits in the central Iranian terrane (CIT). Rem. Sens. 12, 1239.
Shimoda, H., Kimura, T.1, 2017. 09 Japanese space program. Comprehensive Remote Sensing 246.
Shirmard, H., Farahbakhsh, E., Müller, R.D., Chandra, R., 2022. A review of machine learning in processing remote sensing data for mineral exploration. Rem. Sens. Environ. 268, 112750.
Smirnoff, A., Boisvert, E., Paradis, S.J., 2008. Support vector machine for 3D modelling from sparse geological information of various origins. Comput. Geosci. 34, 127–143.
Soloviev, S.G., Kryazhev, S.G., Dvurechenskaya, S.S., Vasyukov, V.E., Shumilin, D.A., Voskresensky, K.I., 2019. The superlarge Malmyzh porphyry Cu-Au deposit, Sikhote-Alin, eastern Russia: igneous geochemistry, hydrothermal alteration, mineralization, and fluid inclusion characteristics. Ore Geol. Rev. 113, 103112.

Sothe, C., De Almeida, C., Schimalski, M., La Rosa, L., Castro, J., Feitosa, R., Dalponte, M., Lima, C., Liesenberg, V., Miyoshi, G., 2020. Comparative performance of convolutional neural network, weighted and conventional support vector machine and random forest for classifying tree species using hyperspectral and photogrammetric data. GIScience Remote Sens. 57, 369–394.
Soydan, H., Koz, A., Düzgün, H.Ş., 2021. Secondary iron mineral detection via hyperspectral unmixing analysis with sentinel-2 imagery. Int. J. Appl. Earth Obs. Geoinf. 101, 102343.
Spychała-Kij, M., 2020. Civil-Military Cooperation in Post Conflict Development: A Case of North Waziristan, vol. 3. NUST Journal of International Peace and Stability, pp. 102–107.
Sun, T., Chen, F., Zhong, L., Liu, W., Wang, Y., 2019. GIS-based mineral prospectivity mapping using machine learning methods: a case study from Tongling ore district, eastern China. Ore Geol. Rev. 109, 26–49.
Sun, T., Li, H., Wu, K., Chen, F., Zhu, Z., Hu, Z., 2020. Data-driven predictive modelling of mineral prospectivity using machine learning and deep learning methods: a case study from southern Jiangxi Province, China. Minerals 10, 102.
Tompolidi, A.-M., Sykioti, O., Koutroumbas, K., Parcharidis, I., 2020. Spectral unmixing for mapping a hydrothermal field in a volcanic environment applied on ASTER, landsat-8/OLI, and sentinel-2 MSI satellite multispectral data: the Nisyros (Greece) case study. Rem. Sens. 12, 4180.
Tuba, E., Ribic, I., Capor-Hrosik, R., Tuba, M., 2017. Support vector machine optimized by elephant herding algorithm for erythemato-squamous diseases detection. Procedia Comput. Sci. 122, 916–923.
Turlapaty, A.C., Anantharaj, V.G., Younan, N.H., 2010. A pattern recognition based approach to consistency analysis of geophysical datasets. Comput. Geosci. 36, 464–476.
Van der Werff, H., Van der Meer, F., 2016. Sentinel-2A MSI and Landsat 8 OLI provide data continuity for geological remote sensing. Rem. Sens. 8, 883.
van Leeuwen, W.J., 2009. Visible, Near-IR, and Shortwave IR Spectral Characteristics of Terrestrial Surfaces. pp. 33–50.
Wang, X., Yuan, P., Mao, Z., You, M., 2016. Molten steel temperature prediction model based on bootstrap feature subsets ensemble regression trees. Knowl. Base Syst. 101, 48–59.
Wijaya, D.R., Afianti, F., Arifianto, A., Rahmawati, D., Kodogiannis, V.S., 2022. Ensemble machine learning approach for electronic nose signal processing. Sensing and Bio-Sensing Research 36, 100495.
Woodhead, J., Landry, M., 2021. Harnessing the power of artificial intelligence and machine learning in mineral exploration—opportunities and cautionary notes. SEG Discovery 19–31.
Xiong, Y., Zuo, R., 2020. Recognizing multivariate geochemical anomalies for mineral exploration by combining deep learning and one-class support vector machine. Comput. Geosci. 140, 104484.
Xu, S., Zhao, Q., Yin, K., Zhang, F., Liu, D., Yang, G., 2019a. Combining random forest and support vector machines for object-based rural-land-cover classification using high spatial resolution imagery. J. Appl. Remote Sens. 13. 014521-014521.
Xu, S., Zhao, Q., Yin, K., Zhang, F., Liu, D., Yang, G.J.J.o.A.R.S., 2019b. Combining Random Forest and Support Vector Machines for Object-Based Rural-Land-Cover Classification Using High Spatial Resolution Imagery, vol. 13. 014521.
Yang, H.-h., Wang, Q., Li, Y.-b., Lin, B., Song, Y., Wang, Y.-y., He, W., Li, H.-w., Li, S., Li, J.-l., 2022a. Geology and mineralization of the Tiegelongnan supergiant porphyry-epithermal Cu (Au, Ag) deposit (10 Mt) in western Tibet, China: a review. China Geology 5, 136–159.
Yang, N., Zhang, Z., Yang, J., Hong, Z., 2022b. Mineral prospectivity prediction by integration of convolutional autoencoder network and random forest. Nat. Resour. Res. 31, 1103–1119.
Yin, J., Li, N., 2022. Ensemble learning models with a Bayesian optimization algorithm for mineral prospectivity mapping. Ore Geol. Rev. 145, 104916.
Zuo, R., 2020. Geodata science-based mineral prospectivity mapping: a review. Nat. Resour. Res. 29, 3415–3424.
Zuo, R., Kreuzer, O.P., Wang, J., Xiong, Y., Zhang, Z., Wang, Z., 2021. Uncertainties in GIS-based mineral prospectivity mapping: key types, potential impacts and possible solutions. Nat. Resour. Res. 30, 3059–3079.