UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Geociências

ELIAS MARTINS GUERRA PRADO

AVALIAÇÃO DE MÉTODOS DE APRENDIZAGEM DE MÁQUINA PARA O MODELAMENTO DA PROSPECTIVIDADE E VETORIZAÇÃO DE MINERALIZAÇÕES DO TIPO IOCG

EVALUATION OF MACHINE LEARNING METHODS FOR PROSPECTIVITY MODELING AND VECTORING OF IOCG MINERALIZATIONS

CAMPINAS
2022

ELIAS MARTINS GUERRA PRADO

EVALUATION OF MACHINE LEARNING METHODS FOR PROSPECTIVITY MODELING AND VECTORING OF IOCG MINERALIZATIONS

AVALIAÇÃO DE MÉTODOS DE APRENDIZAGEM DE MÁQUINA PARA O MODELAMENTO DA PROSPECTIVIDADE E VETORIZAÇÃO DE MINERALIZAÇÕES DO TIPO IOCG

THESIS PRESENTED TO THE INSTITUTE OF GEOSCIENCES OF THE UNIVERSITY OF CAMPINAS TO OBTAIN THE DEGREE OF DOCTOR IN SCIENCES IN AREA OF GEOLOGY AND NATURAL RESOURCES

ORIENTADOR: PROF. DR. CARLOS ROBERTO DE SOUZA FILHO

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELO ALUNO ELIAS MARTINS GUERRA PRADO E ORIENTADA PELO PROF. DR. CARLOS ROBERTO DE SOUZA FILHO

CAMPINAS
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Geociências
Marta dos Santos - CRB 8/5892

P882e

Prado, Elias Martins Guerra, 1988-
Evaluation of machine learning methods for prospectivity modeling and vectoring of IOCG mineralizations / Elias Martins Guerra Prado. – Campinas, SP : [s.n.], 2022.

Orientador: Carlos Roberto de Souza Filho.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Geociências.

1. Minas e recursos minerais – Exploração. 2. Aprendizado de máquina. 3. Modelamento de prospectividade mineral. 4. Espectroscopia de reflectância. 5. Geologia – Estimativas. I. Souza Filho, Carlos Roberto de, 1965-. II. Universidade Estadual de Campinas. Instituto de Geociências. III. Título.

Informações Complementares

Título em outro idioma: Avaliação de métodos de aprendizagem de máquina para o modelamento da prospectividade e vetorização de mineralizações do tipo IOCG
Palavras-chave em inglês:
Mines and mineral resources - Exploration
Machine learning
Mineral prospectivity modeling
Reflectance spectroscopy
Geology - Estimates
Área de concentração: Geologia e Recursos Naturais
Titulação: Doutor em Geociências
Banca examinadora:
Carlos Roberto de Souza Filho [Orientador]
Gelvam André Hartmann
Álvaro Penteado Crósta
Aline Tavares Melo
Lena Virginia Soares Monteiro
Data de defesa: 21-12-2022
Programa de Pós-Graduação: Geociências

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-2461-8246
- Currículo Lattes do autor: http://lattes.cnpq.br/4551088996437373

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE GEOCIÊNCIAS

UNICAMP

AUTOR: Elias Martins Guerra Prado

Evaluation of machine learning methods for prospectivity modeling and vectoring of IOCG mineralizations

Avaliação de métodos de aprendizagem de máquina para o modelamento da prospectividade e vetorização de mineralizações do tipo IOCG

ORIENTADOR: PROF. DR. CARLOS ROBERTO DE SOUZA FILHO

Aprovada em: 21/12/2022

EXAMINADORES:

Prof. Dr. Carlos Roberto de Souza Filho - Presidente

Prof. Dr. Gelvam André Hartmann

Prof. Dr. Álvaro Penteado Crósta

Profa. Dra. Aline Tavares Melo

Profa. Dra. Lena Virgínia Soares Monteiro

A Ata de Defesa assinada pelos membros da Comissão Examinadora consta no processo de vida acadêmica do aluno.

Campinas, 21 de dezembro de 2022.

SÚMULA/BIOGRAFIA

Elias M. G. Prado é bacharel em Geologia (2012) pela Universidade de Brasília (UnB). Obteve mestrado na UnB em 2015, desenvolvendo um estudo sobre análise quantitativa da mineralogia de formações ferríferas por meio de dados espectrais. Em 2014, se tornou pesquisador em geociências do Serviço Geológico do Brasil (SGB), onde atuou no mapeamento geológico do estado de Rondônia por três anos. Atua desde 2019 como coordenador executivo do Centro de Geociências Aplicadas (CGA) do SGB, onde desenvolve pesquisas utilizando técnicas de inteligência artificial para criar soluções na área das geociências. Iniciou seu Ph.D. na Universidade de Campinas (Unicamp), em 2019, com o objetivo de investigar a aplicação de métodos de aprendizado de máquina na exploração de depósitos minerais do tipo IOCG.

BIOGRAPHY

Elias M. G. Prado has a bachelor’s degree in Geology (2012) from the University of Brasília (UnB). He obtained a master’s degree from UnB in 2015, developing a study on quantitative analysis of the mineralogy of iron formations through spectral data. In 2014, he became a researcher in geosciences at the Geological Survey of Brazil (SGB), where he worked on the geological mapping of the state of Rondônia for three years. He has served since 2019 as executive coordinator of the SGB's Center for Applied Geosciences (CGA), where he develops research using artificial intelligence techniques to create solutions in the geosciences. He started a Ph.D. at the University of Campinas (Unicamp) in 2019, aiming to investigate the application of machine learning methods in the exploration of IOCG-type mineral deposits.

AGRADECIMENTO

To my father, Edson, in loving memory. To my mother, Germana, and sister, Cássia, my sincere thanks for the caring, example, and full support. To my wife Flávia and son Caio, thanks for the companionship, encouragement and all the support.

I thank Prof. Dr. Carlos Roberto de Souza Filho of the University of Campinas (UNICAMP) for his dedication and guidance during this learning period. To the Prof. E. John Carranza of the University of the Free State (Bloemfontein, South Africa) for the great collaboration in this work.

I thank the Geological Survey of Brazil (SGB) for the release for training through internal selection process. I direct this acknowledgment to the head of the Applied Geosciences Center (CGA) of the SGB, Noevaldo Araujo Teixeira, who supports my work so many times

To the various geologists and colleagues in Brazil who participate directly or indirectly in the research. To João Motta, Leandro, Rodrigo, Cimara, and Christian for all the support and knowledge exchange.

Thanks to the department of Geology and Natural Resources, Institute of Geosciences, UNICAMP for providing all the institutional support and infrastructure needed in this work.

RESUMO

O desenvolvimento da exploração mineral com foco no aumento da disponibilidade dos recursos minerais e na redução do impacto ambiental é crucial para o desenvolvimento sócio-econômico sólido e sustentável da sociedade. A sociedade está atualmente passando por um rápido crescimento na extração e consumo de recursos minerais devido ao aumento da população humana, aumento dos níveis de atividade econômica e transição para novas tecnologias. Técnicas de inteligência artificial, análise de grandes volumes de dados e outras tecnologias da Industria 4.0 são soluções promissoras para contornar muitos desses problemas. Este trabalho apresenta novas abordagens para a exploração mineral que incorpora métodos de última geração focados na aplicação de algoritmos de aprendizagem de máquina para o mapeamento da prospectividade mineral e estimativa de teor de minério por meio de dados espectrais.

As abordagens são exemplificadas pela exploração de depósitos de cobre, ouro e óxido de ferro (IOCG), fonte de commodities economicamente importantes, como o cobre (Cu) e elementos de terras raras (REE), que atualmente têm uma demanda alta e crescente. Os métodos desenvolvidos neste trabalho abrangem duas etapas distintas de exploração mineral, a identificação de novas zonas mineralizadas e a mineração de jazidas conhecidas. Como estes métodos são inovadores e, ainda assim, são testes, o núcleo do projeto se concentra nos numerosos aspectos do processamento de dados, otimização de algoritmos, arquitetura de modelos e ajuste de hiperparâmetros de algoritmos de aprendizagem de máquina.

Técnicas de aprendizagem de máquina foram adaptadas para processar dados geológicos e geofísicos da província mineral de Carajás, Brasil, para modelar a prospectividade dos depósitos minerais do IOCG na região. Os resultados mostram que os modelos de prospectividade desenvolvidos utilizando algoritmos de aprendizagem de máquina têm o desempenho espacial e de classificação melhores do que os métodos tradicionais baseados em dados, como peso da evidência.

Este trabalho também mostra uma nova abordagem pela qual algoritmos de aprendizagem profunda são usados para prever as classes Cu no depósito Olimpic Dam, Austrália, por meio de dados hiperespectrais. Os resultados mostraram que a abordagem proposta pode ser usada em sistemas de automação para a identificação de zonas mineralizadas, permitindo mineração seletiva, diminuindo assim os custos e o impacto ambiental, e aumentando o desempenho das operações de mineração.

Além disso, foi proposto um novo fluxo de trabalho para automatizar a identificação de limites litológicos e de alteração utilizando dados hiperespectrais adquiridos de testemunhos. Os resultados mostraram que os aglomerados obtidos pela abordagem proposta têm uma correlação significativa com as litologias registradas e as concentrações de Cu, foram capazes de estimar corretamente os limites litológicos e de alteração, bem como identificar padrões de alteração associados ao grau do minério que não foram identificados durante a extração visual.

Os resultados desta tese indicam que as técnicas de aprendizagem de máquinas superam as técnicas tradicionais utilizadas para a modelagem de prospectividade e vetorização de mineralizações IOCG. Os métodos desenvolvidos podem ser perfeitamente adaptados e utilizados na exploração de outros tipos de depósitos minerais.

Palavras-chave: Exploração mineral; aprendizagem de máquina; mapeamento de prospectividade mineral; dados espectrais; estimativa de teor de minério.

ABSTRACT

The development of mineral exploration focused on increasing the availability of mineral resources and reducing environmental impact is crucial for a solid and sustainable socio-economic development of society. Society is currently experiencing rapid growth in the extraction and consumption of mineral resources due to an increasing human population, increasing levels of economic activity, and transition to new technologies. Artificial intelligence techniques, big data analysis, and other Industry 4.0 technologies are promising solutions to circumvent many of these problems. This work presents new approaches to mineral exploration targeting that incorporates state-of-the-art methods focused on the applicability of machine learning algorithms for mineral prospectivity mapping and estimation of ore grade by means of spectral data.

The approaches are exemplified by the exploration of Iron-Oxide Cupper Gold (IOCG) deposits, source of economically important commodities, as copper (Cu) and rare earth elements (REEs), that currently have a high and growing demand. The methods developed at this work cover two distinct stages of mineral exploration, the identification of new mineralized zones and mining of known deposits. As these methods are innovative and yet being testes, the core of the project focuses on the numerous aspects of data processing, algorithm optimization, model architecture and hyperparameters tunning of machine learning algorithms.

Machine learning techniques were adapted to process geological and geophysical data of Carajás mineral province, Brazil, to modelling the prospectivity of IOCG mineral deposits in the region. Results shows that prospectivity models developed using machine learning algorithms have a better classification and spatial performance than traditional data-driven methods, as weight of evidence. Therefore, these methods are likely to be dominant in this field in the coming years.

This work also shows a novel approach whereby deep learning algorithms are used to predict Cu grades at Olympic Dam deposit, Australia, by means of hyperspectral data. Results shown that the proposed approach has the potential to be used in automation systems for the identification of mineralized zones, allowing selective mining, thus lowering costs and environmental impact, and increasing performance of mining operations.

In addition, a novel workflow was proposed for automating the identification of lithologic and alteration boundaries using hyperspectral data acquired from drill cores. Results shown that the clusters obtained by the proposed approach has a significative correlation with the logged lithologies and the Cu concentrations, been able to correctly estimate the lithological and alteration boundaries, as well as identify alteration patterns associated with the ore grade which were not identified during visual logging.

The results of this thesis indicate that machine learning techniques overperform traditional techniques used for prospectivity modelling and vectoring of IOCG mineralizations. The developed methods can be seamlessly adapted and used in the exploration of other types of mineral deposits.

Keywords: Mineral exploration; machine Learning; mineral prospectivity mapping; spectral data; ore grade estimation.

SUMÁRIO

1 GENERAL INTRODUCTION

1.1 Industry 4.0 in the mining industry

The socioeconomic development of a nation is linked to the development of mineral exploration. Increasing the availability of raw materials from mineral resources, and reducing the environmental impact generated by the mineral industry are critical factors to promote a solid and sustainable socioeconomic development. However, the mineral industry is currently experiencing some important challenges that need to be overcome, such as the high global demand for raw material, the reduction of ore content in mines, the tight labor market, and the high standards required for mineral production, coupled with the preservation of natural resources (Dorin et al., 2014). Industry 4.0 technologies that encompass the Internet of Things (IoT), Industrial Internet of Things (IIoT), cyber-physical systems, big data, and artificial intelligence have been pointed out as the best solutions to circumvent the problems of modern industry (Sishi and Telukdarie, 2020).

Also known as the fourth industrial revolution, industry 4.0 is a strategy developed by the German government in 2013. After steam engines, electricity, and electronics, Industrie 4.0 consists of the implementation of technologies for process automation, real-time monitoring of processes, and more assertive decision making based on large volumes of data, which allow normal factories to be transformed into smart factories. The German concept is formulated in the report "Recommendations for implementing the strategic initiative Industrie 4.0 - Final report of the Industrie 4.0 Working Group" (Kagermann et al., 2013). Other nations have adopted similar concepts, such as Made in China 2025 (Wübbeke et al., 2016) promoted by the Chinese government, and Society 5.0 (Fukuda, 2020) promoted by the Japanese government.

The concepts of Industry 4.0 are also applicable to the mining industry. In fact, some mines have taken important steps in this direction, and already use the technologies involved in this concept. Gradually, the mineral industry is approaching the concept of Industry 4.0 and fully automated mines, with more technologically sophisticated ore processing facilities, as well as more precise mineral exploration campaigns driven by big data. Despite this, the minerals industry is behind other industries when it comes to employing Industry 4.0 technologies in its endeavors (Sishi and Telukdarie, 2020). One of the key technologies involved in Industry 4.0 is artificial intelligence. The use of artificial intelligence techniques such as machine learning (ML) algorithms to solve problems in the mining industry has been gaining prominence in recent years.

1.2 Overview of technologies for mineral exploration

Mineral exploration and production activities present multiple risks of being unsuccessful, which derive from geological, technological, environmental, social, political and economic uncertainties (Eggert, 2010; Singer and Kouda, 1999). During the exploration process, efficient prospecting technologies are required to minimize uncertainties about the presence of mineralized zones. Similarly, during production, the employment of technologies for modeling the ore body, and identifying undesirable contaminants for the metallurgical processing is of critical importance to mitigate risks. The most used technologies for these purposes are presented in this section.

1.2.1 Mineral prospectivity mapping

One important advance on the identification of regions of ore accumulation is the development of techniques and technologies for the integration of large quantities of spatial geoscientific data via Mineral-Prospectivity Mapping (MPM). The MPM, first proposed in the late 1980s by Bonham-Carter et al. (1988), consists of a statistical method for the recognition and parameterization of spatial patterns and features in geoscience data which are indicative of a particular style of mineralization. The recognized parameters and patterns are used to predict the likelihood of mineralization with similar characteristics occurring in another region. In this way, MPM can significantly reduce the area to be investigated by companies of mineral exploration, minimizing operation costs. The MPM is fundamental for the implementation of Industry 4.0 concepts in mineral exploration, as it allows target selection to be done in an efficient and automated way, based on data, enabling the systematization of the exploration activity. Although MPM has contributed to advances in mineral exploration data integration, existing methods for producing prospectivity maps have some important limitations that can be circumvented by modern data analysis methods that use ML algorithms to find complex patterns in the data.

1.2.2 Integration of hyperspectral and geochemical data

In addition to advances in prospectivity modeling, another important advance in mineral exploration is the use of ultraspectral data for the mineralogical characterization of mineralized zones. Ultraspectral sensors measure reflected light in many narrow, contiguous bands across the visible and near-infrared (VNIR; 350-1000 nm), shortwave infrared (SWIR; 1000-2500 nm) and thermal infrared (TIR; 6000-14500 nm) wavelengths. When interacting with minerals, light is preferably absorbed at certain wavelengths, while it is transmitted at other wavelengths. The position (wavelength) of these absorption features is determined by the crystal structure and chemical composition of minerals (Adams, 1975, 1974; Burns, 1993; Hunt, 1977). Ultraspectral sensors provide the necessary spectral resolution to resolve diagnostic absorption features of many minerals, including iron oxyhydroxide, clays, carbonates, and silicates.

The spectra absorption features can be directly correlated to the chemical composition and crystal structure of minerals (Van der Meer, 2004). Therefore, the correlation between the information obtained by ultraspectral data and geochemistry can be used to infer the concentration of certain elements from ultraspectral data. This correlation can be used to: (1) assist ore body modeling and ore grade mapping systems; (2) detect undesirable minerals (that affects milling and/or metallurgical processing or have occupational health implications); (3) classify ore and waste during loading operations at the mine face; and (4) detect and map hydrothermal alteration patterns that may be used to vector further mineralization. All these can provide inputs to assist in the control of autonomous mining equipment (Fraser et al., 2006).

Drilling during mineral exploration to identify new mineralized zones and to modeling the ore body often requires an expert to characterize the mineralized intervals through core description, and geochemical analysis to quantify the presence of the element of interest at the selected locations. Ore quantification using ultraspectral data enables non-destructive identification and quantification of mineralized zones and hydrothermal alteration patterns to be performed soon after core sampling, allowing faster and more assertive decision making at a lower cost. During ore mining, it is common that the collected material comes with a significant volume of non-mineralized or low-grade material, usually containing mineral phases undesirable for processing plants. The use of ultraspectral data for mineralogical characterization and quantification of mineralized zones, allows the ore mining to be more selective, discarding the material with contaminants, and provides the ore processing plant with important information, such as the mineralogical composition of the collected material. This technology increases the efficiency of the mine and decreases its environmental impact. Therefore, the characterization of mineralized zones using ultraspectral data is also essential for the implementation of Industry 4.0 concepts in the minerals industry.

1.3 Machine Learning: an emerging tool for mineral exploration and production

Recent advances in ML algorithms are driving profound changes in data analysis and integration in various areas of the mineral industry. However, the application of ML algorithms in mineral exploration is still restricted. Classic ML algorithms as artificial neural networks, support vector machines, decision trees and random forest (Burkov, 2019; Friedman et al., 2001) and deep learning (DL) algorithms as Convolutional Neural Networks (CNN; Lecun and Bengio, 1995) have gained a lot of attention in geoscientific journals over the last 20 years, especially in papers on MPM and classification of hyperspectral images (CHI). Despite the considerable number of publications using these algorithms, many advances still need to be made to improve the performance of these techniques and allow their application in the mineral industry.

Given the ability of ML algorithms to extract complex patterns, which may be difficult for conventional statistical methods, these algorithms are able to circumvent the limitations of classical methods in MPM. For this reason, currently, most of the methods used for produce prospectivity maps are based on ML models (Abedi et al., 2012; Brown et al., 2000; Carranza and Laborte, 2015a; Chen et al., 2014; Chen and Wu, 2017a; Leite and de Souza Filho, 2009a; Oh and Lee, 2010; Rodriguez-Galiano et al., 2014; Zuo and Carranza, 2011a). However, one of the main problems in applying ML algorithms to MPM is to deal with imbalanced dataset.

Also, ML algorithms have shown accurate results in the classification of hyperspectral images in recent years due to their ability to learn complex relationships between the spectrum obtained in each pixel and the imaged material (Gewali et al., 2018). In addition, DL algorithms as CNN can recognize patterns related to the spatial relationship between the image pixels (Castelluccio et al., 2015; Krizhevsky et al., 2012). Recently, the number of publications applying ML methods, especially DL algorithms, in CHI has increased considerably, with many journals publishing special issues in the subject (Alavi et al., 2016; Camps-valls and Bioucas-dias, 2016; Chi et al., 2015; Tuia et al., 2014). However, most of these studies have focused on classification of spectral data rather than correlating them with geochemical data.

1.4 Current limits of machine learning for mineral prospectivity mapping

ML algorithms assumes a balanced number of classes on the training dataset. However, the number of mineralized sites is naturally much larger than the number of non-mineralized sites. The huge imbalance between known mineralized sites (minority class) and known non-mineralized sites (majority class) makes it difficult for ML algorithms to learn the classification rules between majority and minority classes during training compromising the performance of the models.

Recent studies involving MPM have attempted to deal with the problem of unbalanced data by using the synthetic minority over-sampling technique (SMOTE) for synthetic oversampling of the mineralized sites (Hariharan et al., 2017a; Li et al., 2019). These studies suggest that using SMOTE can improve the performance of mineral prospectivity models trained on imbalanced data. However, none of these studies has made a systematic analysis of the influence of SMOTE on the performance of mineral prospectivity models.

1.5 Current limits of machine learning for integration of hyperspectral and geochemical data

Most algorithms used for integration of spectral and geochemical data are based on classical statistical methods. Generally, a polynomial regression is calculated to find the fit function between spectral parameters and chemical analyses for a given element (Clark and Roush, 1984; T. Cudahy et al., 2009; Haest and Cudahy, 2012; Prado, 2016). However, polynomial regression is not able to handle some of the complexities of hyperspectral data. The variability of illumination within and between data acquired under different conditions or on uneven surfaces and the presence of impure spectra, formed by mixtures of minerals, considerably decreases the performance of these methods. Furthermore, for accurate results, these methods require large spectral libraries and prior knowledge to identify and map zones of interest. Implementing ML algorithms for processing and interpreting hyperspectral data can circumvent many of these problems. Although some approaches based on ML techniques, recently, have been applied to correlate drill-core geochemical data with hyperspectral data (Acosta et al., 2020), as far as we are aware, deep learning algorithms have not been used to predict element concentration by means of hyperspectral data.

1.6 Organization of the Thesis

The organization of this thesis is as follows. This chapter introduces the main concepts of our work and presents a brief description on the technologies and methods used. Chapter 2 provides the main objectives and contributions of our work. Chapters 3 and 4 outlines a review of the current knowledge and limits in the research topic and introduces the approach adopted in the thesis. The following chapters are structured around scientific articles resulting from the thesis and submitted to or published in peer-reviewed journals. These articles describe the results obtained at each step of the approach. The article “Modeling of Cu-Au prospectivity in the Carajás mineral province (Brazil) through machine learning: Dealing with imbalanced training data” presented in chapter 5 was published in the journal Ore Geology Reviews in September 2020, and the article “Ore Grade Estimation from Hyperspectral Data Using Convolutional Neural Networks: A Case Study at The Olympic Dam IOCG Deposit, Australia” presented in chapter 6 was submitted for publication in the journal Economic Geology in June 2022. The article “Clustering of Hyperspectral Drill Core Measurements Using Deep Autoencoders and Self-Organizing Maps” presented in Chapter 7 is intended to be submitted for publication in the journal Computers & Geosciences in 2023. Chapter 8 gives a summary of the main achievements of this thesis as well as discusses on directions for future work. Appendix 1 contains the code to compute the fractal dimension of geological features (geological complexity), written in python.

2 OBJECTIVES OF THE THESIS

Intended to contribute to the implementation of industry 4.0 concepts in mineral exploration and production, the present PhD thesis aims to develop new workflows based on machine learning and deep learning algorithms for identify and quantify mineralized zones in IOCG deposits. The use of machine learning algorithms for MPM was explored and a systematic analysis on the influence of imbalanced datasets in the performance of mineral prospectivity models was conducted. The use of deep neural networks for the integration of spectral and geochemical data was explored and a new method to quantify Cu ore grade using hyperspectral reflectance data was developed. This work resulted in the following key contributions:

- A code to compute the fractal dimension of geological features (geological complexity) was written in python which is based in raster arithmetic to speed up processing time.
- A systematic analysis on the use of synthetic minority over-sampling (SMOTE) technique for mineral prospectivity mapping.
- A mineral prospectivity map for IOCG deposits in the Carajás mineral province, Brazil.
- Convolutional neural networks were applied to quantify Cu grade using hyperspectral reflectance data for the first time.
- Development of a new method for mapping Cu ore grade using hyperspectral data in the Olympic Dam IOCG deposits.
- Development of a new method based on Deep autoencoders, self-organizing maps and agglomerative clustering for detection of lithological and alteration boundaries in drill-cores using hyperspectral data.

3 STATE-OF-THE-ART OF MINERAL PROSPECTIVITY MAPPING

3.1 Introduction

MPM was first proposed in the late 1980s by geoscientists as a statistical method for the integration and interpretation of spatial patterns in geoscience data (Bonham-Carter et al., 1988). The concept was to determine the link between various geoscience datasets (ie: geology, geophysics, geochemistry) and the existence or absence of economic mineralization in an objective way.

The mineral potential mapping methods can be classified into data-driven, knowledge-driven, and hybrids of the two methods. Data-driven methods make use of the relationship between discovered mineral deposits and their surrounding map patterns to set up a mineral potential mapping model (Carranza, 2011). Knowledge-driven methods, however, are estimated based on the expert knowledge of the processes that resulted in the formation of mineral deposits in the given geological setting (Abedi et al., 2013a; Carranza, 2008).

3.2 Previous works

One of the original formulations of MPM, termed Weights of Evidence (WofE) (Agterberg et al., 1990), used posterior probability as the mapping function, which is calculated by counting the relative number of occurrences within and without a series of binary thematic layers (ie: Granite Contact). Other authors have explored using logistic regression (Harris and Pan, 1999), wherein weights are calculated for a series of geological variables using least-squares regression on the probability of mineralization and the binary presence or absence of mineralization. To address the inherent uncertainty in geoscience data, handle categorical variables, and incorporate some expert knowledge, some authors have used fuzzy logic (Porwal et al., 2003a), in which a membership function acts in the place of uncertainty to quantify the degree to which statements are true.

In the last 20 years more sophisticated algorithms have been borrowed from the ML field. Feed forward neural networks, now present in all forms of artificial intelligence, started with few works in the early 1990s and by the 2000s were being applied in simple architectures for MPM (Barnett and Williams, 2009; Brown et al., 2000; Harris and Pan, 1999; Porwal et al., 2003b; Rodriguez-Galiano et al., 2015; Singer and Kouda, 1997). Support vector machines, a maximum margin classifier which gained prominence in the 1990s as a favored algorithm with strong theoretical background, has been applied by several authors (Abedi et al., 2012; Porwal et al., 2010; Rodriguez-Galiano et al., 2015; Zuo and Carranza, 2011a).

Weights of evidence, fuzzy logic and neural networks are some of the more commonly adopted methods (Partington and Sale, 2004; Raines et al., 2010). The popularity of these methods can be attributed to ease of use, flexibility, and successful application in other fields.

3.3 Machine learning for mineral-prospectivity mapping

In the last decades, the use of ML algorithms in geosciences, for both supervised and unsupervised learning, was predominantly concentrated on classification of lands and vegetation, and mapping using remote sense images (Huang et al., 2002; Rogan et al., 2008; Wulder et al., 2004). Only recently, state of art ML methods has become more commonplace in other fields of geosciences (Barnet and Williams, 2006; Caté et al., 2017; Cracknell and Reading, 2014; Davidson, 2017; Lary et al., 2016; Poulton, 2002; Zhu et al., 2017). The increase in available data, computational power, and availability of new algorithms are creating opportunities for the application of ML algorithms to more complex problems. The development of open-source libraries dedicated to making ML more accessible in high-level programming languages has helped popularize these techniques. A good example of such libraries are scikit-learn (Pedregosa et al., 2011) and tensorflow (Abadi et al., 2015). Both are open-source modules for the Python programming language. Scikit-learn implements the most used ML algorithms and tensorflow implements deep learning algorithms.

A range of method by which ML techniques can be applied to MPM have begun to be explored in several recent studies. Carranza and Laborte (2015b) used Random Forest (RF) algorithm for predictive mapping of gold prospectivity in Baguio district, Philippines. The spatial datasets used include a geological map, map of faults/fractures, and locations of gold deposits. Their results shown that RF modeling is a much more potentially useful method compared to other methods that are currently used for data-driven mapping of mineral prospectivity, given its stability in training and its ability to yield predictive maps with high success- and prediction-rates. Similarly, Rodriguez-Galiano et al. (2015) evaluated the performance of neural networks (NN), RF, regression trees (RT) and support vector machine (SVM) for MPM of gold in Rodalquilar mining district, southeast Spain. The algorithms were trained with physical–chemical data such as a geochemical survey, gravity and magnetic survey, geological information regarding fractures and lithology and hyperspectral EO1-Hyperion image. Their results showed that decision-tree-based algorithms (RT and RF) involve a lesser difficulty in their training, additionally the greatest accuracy of classifications was achieved by RF and SVM, but the performance of RF for all the parameter combinations was better than that of the rest in terms of stability and accuracy.

More recently, a data driven approach using CNN was undertaken by Granek et al. (2016) for MPM of copper-gold porphyry mineralizations on QUEST (QUesnelia Exploration STrategy) project in central British Columbia, Canada. Authors used airborne gravity, magnetic and electromagnetic data, inductively coupled plasma mass spectrometry (ICP-MS) analysis of stream and sediment samples (providing compositional information for 35 elements), geological era, period, rock class and rock type, for training the deep learning architecture. The same dataset was used to train a SVM model. Results showed that the main advantages of CNNs over SVM in MPM is the ability to recognize anomalous structure in the data rather than simply anomalous values. Comparably, Chen and Wu (2017) used extreme learning machine (ELM) regression for mapping polymetallic prospectivity in Lalingzaohuo district, China. Their data-driven approach used geochemical analysis of stream and sediment samples, regional geological entities, contact zones, faults, magnetic data, and linear features for training the ELM algorithm. Results from this study shown that the ELM algorithm is stable and reproducible, and that the learning speed of the ELM regression is much faster than that of logistic regression, and the ELM regression algorithm slightly outperforms logistic regression in mapping polymetallic prospectivity. Additionally, Cracknell and Caritat (2017) used unsupervised ML clustering method, specifically Self-Organizing Maps (SOM), for catchment-based gold prospectivity analysis in northern Australia. The SOM was trained with geochemical analysis of stream and sediment samples, airborne gravity and magnetic data, terrain slope and surface geology.

3.4 Dealing with imbalanced datasets

The application of ML algorithms assumes a balanced number of sites with known mineral occurrences and sites where there are no occurrences of the type/style of mineralization of interest. However, mineralization, being a singular event, is rare (Carranza and Laborte, 2015a, 2015b; Cheng, 2007; Granek et al., 2016; Granek and Haber, 2015). Therefore, the number of non-mineralized sites is naturally much larger than the number of mineralized sites. The huge imbalance between known mineralized sites (minority class) and known non-mineralized sites (majority class) makes it difficult for ML algorithms to learn the classification rules between majority and minority classes during training. This leads to biased estimation of decision boundaries for the majority (non-mineralized) class and a higher misclassification rate (false negative) of the minority (mineralized) class based on the test sites (Japkowicz and Stephen, 2002a). High rates of false negatives (i.e., mineralized sites incorrectly classified as non-mineralized) in MPM can result in missed opportunities for the discovery of a new mineral deposit.

Most of the papers using ML and DL algorithms for MPM attempts to solve the imbalanced data problem in a naive way, selecting an amount of non-mineralized sites equal to the amount of known mineralized sites to train the algorithms. However, generally the number of known mineralized sites is small, a few tens of points, resulting in models trained with less than 100 samples, in some extreme cases trained with less than 20 samples (adding mineralized and non-mineralized sites). Models trained with such a small amount of data are not statistically meaningful. In addition, they are prone to overfitting because the ability to generalize is compromised by the lack of statistical representativeness of the data.

Recently, some studies have explored in depth how to address the problem of imbalanced data for both majority and minority classes. Xiong and Zuo (2017) proposed a cost-sensitive neural network to minimize the error of classification of known mineralized sites. The proposed algorithm incorporates into a neural network the cost of misclassification of known negative and positive classes. To decrease the bias of the model towards the majority class, Xiong and Zuo (2018) proposed a rare event logistic regression algorithm, which incorporate corrections in the sampling and decision threshold of the original logistic regression algorithm. Wang et al. (2020) proposed a semi-supervised random forest algorithm to deal with the small number of known mineralized sites in MPM. The proposed classifier exploits the unclassified data to improve the performance of the classification model. At the data level, the use of Synthetic Minority Over-sampling Technique (SMOTE) for synthetic oversampling of the non-mineralized class was proposed by Hariharan et al. (2017) to improve the prospectivity model produced by a random forest classifier. Li et al. (2019) recently demonstrated the usefulness of SMOTE in MPM based on random forest models. These studies suggest that using SMOTE to oversample the minority class and undersample the majority class can improve the performance of mineral prospectivity models trained on unbalanced data.

3.5 Conclusion

This section outlines the current knowledge in MPM. Recently, many authors reported success cases on the application of ML and DL algorithms for MPM. However, the imbalanced nature of MPM problems makes training the ML and DL algorithms difficult, as it usually results in biased models with high misclassification rates of the minority class (mineralized locations in MPM problems). Many publications deal with this problem in a naive way, training the models with balanced datasets that are not statistically meaningful given their small number of samples. In the last five years, some authors delved in solve this problem using better approaches. Based on these studies, SMOTE seams to be an effective approach, which showed to improve the MPM models and is easier to implement. Although this approach has been used by some authors, a systematic analysis of the influence of SMOTE on MPM has never been performed in previous studies. This analysis is developed in this thesis.

4 STATE-OF-THE-ART OF INTEGRATION OF SPECTRAL AND GEOCHEMICAL DATA

4.1 Introduction

The analysis of spectral data consists of observing and analyzing the light spectrum reflected by different materials. In geology this analysis is used for the identification of minerals and rocks. When interacting with minerals, light is preferably absorbed at certain wavelengths, while it is transmitted and/or reflected at other wavelengths. The position of the absorption features in the visible and near-infrared (VNIR) are determined by the presence of some transition metals such as Cr, Ti, Fe, etc., in the mineral structure (Adams, 1975, 1974; Burns, 1993; Hunt, 1977; Laukamp et al., 2021). In the short-wave infrared (SWIR) region, vibration processes related to H_2O, OH– and CO_3 bounds also produce characteristic absorption features in minerals (Hunt, 1977; Laukamp et al., 2021). Also, fundamental vibrational frequencies of silicates (Si-O), carbonates (CO_3^{2-}), sulfates (SO_4^{2-}), and phosphates (PO_4^{3-}) show spectral features in the thermal infrared (TIR) region (Hunt, 1976; Hunt and Salisbury, 1974, 1970; Laukamp et al., 2021; Lyon, 1965; Lyon and Burns, 1963; Riley and Hecker, 2013; Vincent et al., 1975).

The crystalline structure and chemical composition of the minerals control the position, shape, and depth of the absorption features. In this way, the spectral analysis, in addition to identifying the minerals, can be used to estimate the chemical composition and degree of crystallinity of the minerals. The macroscopic characteristics of the sample can also affect the depth and shape of the absorption features, such as grain size and surface roughness (Clark et al., 1990; Murray and Lyons, 1955; Van der Meer, 2004).

Spectral data is acquired by spectroradiometers, a tool with sensors that measure reflected light in the VNIR, SWIR and TIR wavelengths. These tools are classified according to the spectral resolution of their sensors into multispectral, hyperspectral and ultraspectral. Multispectral sensors collect information at a few (<100) discrete bands, commonly at an interval larger than 20 nm, therefore, do not produce the "spectrum" of a mineral. Hyperspectral sensors measure the reflected light in many (between 100-350) narrow, contiguous bands, commonly at an interval smaller than 10 nm, providing a continuous spectrum. At this work, we also refer to ultraspectral data as data acquired by sensors that similarly to hyperspectral ones, provide continuous spectrum, however, has a better spectral resolution and a wider spectrum coverage, collecting information in the VNIR-SWIR (350-2500 nm) and TIR (6500-14500 nm) regions at narrow intervals (6-4 nm in VNIR-SWIR and 25 nm in TIR), resulting in spectra data with more than 800 bands.

Multi- and hyperspectral data is normally acquired from airborne and satellite imaging platforms and has provided new insights into the structure and functioning of the Earth's systems (Goetz, 2009; Goetz et al., 1985; van der Meer et al., 2012). Recently, uncrewed aerial vehicles (UAVs) (also referred as drones) are also been used to acquire multi- and hyperspectral data at a high spatial resolution (cm scale) (Booysen et al., 2020; Fairley et al., 2018; Jackisch et al., 2018; Padró et al., 2019). Hyperspectral sensors mounted on field-based platforms capable to provide images with high spectral and spatial resolution (mm scale) are also available, thanks to the recent advances in scanning technology (Kurz et al., 2011, 2008; Monteiro et al., 2009; Murphy and Monteiro, 2013).

Laboratory hyperspectral scanning systems, which became available in the late- 2000s, made possible to scan and to analyze drill core samples at a high resolution and to detail spectral variations over the entire sample. Because of a fine spatial resolution (commonly 0.5 mm/pixel) of these systems, it is possible to acquire the spectral signature of the mineral network (rock texture), providing picture elements that are smaller than many of the mineral grains and veins of interest. The spatial resolution of these scanners minimizes the effects of mixtures, and the mineral maps generated from these data can provide information about the sample structure, such as foliation, bedding, and veining that can be present in the core samples. In addition, these maps provide detailed mineralogical mapping, allowing the characterization of zones of hydrothermal alteration, and the characterization of the mineral assemblage present in the ore zones (Mathieu et al., 2017).

Ultraspectral sensors, different from multi- and hyperspectral, usually are not imaging systems, and are only capable to acquire punctual measurements. These sensors can be used both in laboratory and field. Its advantage is the possibility to acquire the measurement at a short distance from the sample (a few mm when using the contact probe), enabling a high signal-to-noise ratio. In addition, as already mentioned, these sensors have high spectral resolution covering a wide range of the spectrum. Core logging systems mounted with ultraspectral sensors are also available, as the HyLogger-3^{TM} system (Schodlok et al., 2016). As hyperspectral scanners, these systems also have a high spatial resolution (< 8 mm), however, as the measurements are taken punctually, these core logging systems do not produce images.

4.2 Previous works

The concentration of certain elements is significantly correlated to the spectral response of certain rocks and minerals. The absorption intensity of a particular mineral in relation to the background (usually considered the continuum of the spectrum) is a function of the absorption coefficient and abundance of a mineral; thus, when the mineral is identified, its abundance may be determined by the absorption intensity (Clark and Roush, 1984; Hunt and Ashley, 1979). Additionally, crystal chemistry of a particular mineral determines the wavelength of its diagnostic absorption feature (Clark et al., 1990; Cudahy and Ramanaidou, 1997; Hunt and Salisbury, 1970). Whole rock geochemical analysis is strictly related to the abundance and composition of minerals, and therefore it is also related to the spectral response of rocks and minerals.

Several publications use these principles to correlate the depth and position of absorption features to geochemical data. Generally, these studies combine multiple spectral features to identify and quantify minerals. For mineral quantification, the depth of the main absorption feature of certain minerals are correlated with geochemical analysis to produce a linear or polynomial regression to estimate the concentration of elements by means of the spectral features. Some of the products developed in these studies include: estimation of wt % Fe using the depth of iron (oxyhydr-)oxides (hematite, goethite) feature (~900 nm) (Ducart et al., 2016; Haest and Cudahy, 2012; Prado et al., 2016); estimation of wt % Al_2O_3 using the depth of Al clays (kaolinite group, white micas, and Al smectites) feature (~2200 nm) (Haest and Cudahy, 2012; Prado et al., 2016; Silversides and Murphy, 2017); estimation of wt %

MgO using the depth of Talc feature (~2310 nm) (Prado et al., 2016); estimation of wt % K_2O using the depth carnallite feature (~1200 nm) (J.-T. Qiu et al., 2017).

4.3 Machine Learning for spectral data analysis

To date, many ML techniques have been used for hyperspectral images classification (Bioucas-Dias et al., 2013; Plaza et al., 2009). The ability of ML algorithms to identify complex patterns on high dimensionality data, the ability to generalize, and the computational efficiency of these algorithms, makes them suitable for hyperspectral image classification.

Remote sensing researchers have developed numerous deep learning based remote sensing data analysis methods which has produced top performances. A popular deep learning architecture for vision tasks is a convolutional neural network (CNN; Krizhevsky et al., 2012). Inspired by the mammalian visual system, these neural networks contain layers for learning low-level to high-level features. Networks with one dimensional (Hu et al., 2015), two dimensional (Romero et al., 2015), and three dimensional (Chen et al., 2016) convolutional layers have been developed for analyzing hyperspectral data in some papers. Results of these studies show that deep neural networks outperform classical spectra classification methods as spectral angle mapper, as well as classic ML algorithms as support vector machines and random forest.

Recurrent neural networks (RNN) are popular architectures for modeling sequential data. They contain feedback loops in their computation allowing the current output to be dependent on the current input and the previous input. Mou et al., (2017) proposed using RNN to model pixel spectra in a hyperspectral image as a 1-D sequences for classification. They experimented with architectures based on two kinds of recurrent units, namely, long short-term memory (LSTM; Graves and Schmidhuber, 2005) and gated recurrent unit (GRN; Mou et al., 2017). They found that the GRN worked better than the LSTM for modeling hyperspectral data and both of the recurrent networks outperformed traditional approaches and baseline CNN. Similarly, Wu and Prasad (2017) showed that a convolutional RNN (a network that has few convolutional layers followed by RNN; Zuo et al., 2015) is better choice for spectra classification than LSTM and baseline CNN.

Most of the studies using ML algorithm for spectral data analysis have focused on mineral classification rather than correlating them with geochemical data. Recently, some approaches based on ML techniques have been applied to correlate drill-core geochemical data with spectral data. Barker et al. (2021) developed a ML model for quantitative mineral mapping of drill core surfaces. Authors used micro-X-ray fluorescence analysis and TIR spectra collected by hyperspectral drillcore scanners to train a ML model to predict minerals in each image pixel. Acosta et al. (2020) proposed a new ML-based technique for the integration of drill-core geochemical and hyperspectral data. The algorithm is trained to classify the spectral data into classes defined by the geochemical analysis; Therefore, it can be used to upscale the information obtained from the geochemical assays to the entire borehole. Although these works use ML techniques to integrate geochemical and spectral data, as far as we are aware, ML algorithms have not been used to predict element concentration by means of spectral data.

4.4 Conclusion

This section outlines the current knowledge in the integration of spectral and geochemical data. Spectral data is a powerful tool in mineral exploration, as it can be used to identify and quantify the mineralogy of rocks. Many authors reported success cases on the application of ML and DL algorithms for the classification of spectral data, especially in remote sensing applications. However, these techniques based on artificial intelligence were only used recently for mineral quantification and correlation of spectral and geochemical data. Despite these recent advances, as far as we are aware, ML and DL algorithms have not been used to predict element concentration by means of spectral data. Therefore, a deep neural network model was developed in this thesis with the aim to explore the use of DL algorithms to predict ore concentration by means of spectral data.

5 MODELING OF CU-AU PROSPECTIVITY IN THE CARAJÁS MINERAL PROVINCE (BRAZIL) THROUGH MACHINE LEARNING: DEALING WITH IMBALANCED TRAINING DATA

5.1 Introduction

As the volume and complexity of geoscience data increase, the need for effective means of analysis and integration of such "big data" also increases. This difficulty of finding new mineral deposits is addressed by increasing the amount and type of suitable data for analysis and synthesis to derive exploration targets (Bergen et al., 2019; Karpatne et al., 2017; Yousefi et al., 2019). Methods for mineral prospectivity mapping (MPM) provide objective tools for the integration of large quantities of geoscientific data.

The procedure of MPM using a geographic information system (GIS) was first demonstrated in the late 1980s by Bonham-Carter et al. (1988) to facilitate data analysis and synthesis. It consists of statistical analysis for the recognition and parameterization of spatial patterns in geoscience data layers. The spatial patterns are deemed meaningful proxies of processes associated with the occurrence of a particular type/style of mineralization. They are also used to predict the existence of areas where there is a likelihood for the occurrence of the same type/style of mineralization, with an implicit assumption that predicted areas have similar characteristics as the known mineralized locations. Therefore, parts of predicted areas with no known occurrence of the same type/style of mineralization are targets for mineral exploration. In this way, MPM can significantly reduce the size of a region to be investigated, with the corresponding minimization of operational costs.

Methods of GIS-based MPM can be classified as either data- or knowledge-driven, as well as hybrids of these two end-members. Data-driven methods make use of quantified spatial relationships between discovered mineral deposits and individual layers of spatial patterns of interest in the region being investigated (Bonham-Carter, 1994; Carranza, 2011; Ford et al., 2016; Liu et al., 2014; Pan and Harris, 2000). Knowledge-driven methods, however, are based on subjective evaluations of spatial patterns of interest based on expert knowledge of processes that might have operated during the formation of mineral deposits in the geological setting of the region under analysis (Abedi et al., 2013b; An et al., 1991; Carranza, 2008; Ford et al., 2016; Pan and Harris, 2000). Data-driven methods of MPM need a significantly large number of mineralized locations to produce statistically accurate results (Bonham-Carter et al., 1988). In contrast, knowledge-driven methods can be applied in areas with few or no known mineralized locations; however, their drawbacks are the subjective bias and the cost and difficulty to have access to relevant expert knowledge. Regardless of the method used, the main challenge nowadays is to extract and parameterize the spatial patterns of interest from large volumes of suitable data that are available for MPM.

The ability of MLalgorithms to recognize complex spatial patterns circumvents the limitations of classical, statistical methods of MPM. Currently, many data-driven methods of MPM are based on ML algorithms such as artificial neural network (Brown et al., 2000; Harris and Pan, 1999; Leite and de Souza Filho, 2009b; Porwal et al., 2003b; Rodriguez-Galiano et al., 2015; Singer and Kouda, 1996; Xiong and Zuo, 2017), support vector machine (SVM) (Abedi et al., 2012; Granek et al., 2016; Porwal et al., 2010; Rodriguez-Galiano et al., 2015; Shabankareh and Hezarkhani, 2017; Zandiyyeh et al., 2016; Zhang et al., 2018; Zuo and Carranza, 2011a), random forest (Carranza and Laborte, 2016, 2015c, 2015b; Gao et al., 2016; Hariharan et al., 2017b; McKay and Harris, 2016; Person et al., 2008; Radford et al., 2018; Rodriguez-Galiano et al., 2015, 2014; Zhang et al., 2016), decision tree (Chen et al., 2014) and extreme learning machine (Chen and Wu, 2017a). The application of any of these ML algorithms assumes a balanced number of locations of known mineral occurrences and number of locations where occurrences of the type/style of mineralization of interest definitely do not exist. However, mineralization, being a singular event, is rare (Carranza and Laborte, 2015c, 2015b; Cheng, 2007; Granek et al., 2016; Granek and Haber, 2015). Therefore, the number of mineralized locations is naturally outnumbered to a large extent by non-mineralized locations. The huge imbalance between known mineralized locations (i.e., the minority and positive class) and known non-mineralized locations (i.e., the majority and negative class) makes it difficult for ML algorithms to learn the classification rules between the majority and minority classes during training (i.e., learning). This leads to biased estimation of the decision boundaries towards the majority (i.e., non-mineralized) class and a higher misclassification (false negative) rate of the minority (i.e., mineralized) class based on testing locations (Japkowicz and Stephen, 2002b; Sun et al., 2009). High false negative rate (i.e., mineralized locations incorrectly classified as non-mineralized) in MPM will result in a missed opportunity for discovery of a new mineral deposit.

Reported solutions for handling the learning difficulties of ML algorithms when trained with imbalanced data regarding the majority and minority classes were previously proposed at data and algorithm levels (e.g., Chawla et al., 2004; Sun et al., 2009). Solutions at data level consist of re-sampling methods developed to rebalance data for the majority and minority classes, namely by over-sampling or under-sampling of a relevant class and by a combination of both methods (Chawla et al., 2002, 2004; Jiang et al., 2013; Sun et al., 2009; Zhi-Hua Zhou and Xu-Ying Liu, 2006). Over-sampling methods increase the number of minority class samples, whereas under-sampling methods decrease the number of majority class samples, to balance the amount of sample data before training. At algorithm level, solutions try to improve the learning ability of ML algorithm to correctly classify the minority class (Quinlan, 1991; Zadrozny and Elkan, 2001), such as cost-sensitive learning, one-class learning and ensemble learning. Cost-sensitive learning solutions attempt to assume higher misclassification cost to samples of the minority class, reducing the bias toward the majority class on the estimation of the decision (i.e., classification) boundaries (Domingos, 1999; Drummond and Holte, 2003; Elkan, 2001; Zadrozny et al., 2003; Zhi-Hua Zhou and Xu-Ying Liu, 2006). One-class learners are ML algorithms trained only with the target class (i.e., mineralized class in MPM) by recognition-based learning. Different from discrimination learning are models that attempt to define boundaries between classes, whereas recognition-based learning attempts to define a classification boundary that surrounds the target class. The classification by one-class learning is done by measuring the similarity between a query object and the target class, and defining a threshold on the similarity value (Chen and Wu, 2017b; Japkowicz, 2001; Manevitz and Yousef, 2001; Schölkopf et al., 2001; Tax, 2001; Xiong et al., 2018; Xiong and Zuo, 2016). Ensemble learning algorithms combine multiple classifiers and aggregate their predictions in order to improve the generalization ability of the model. Most ensemble algorithms are based on Boosting (Freund, 1995) and Bagging (Breiman, 1996) methods. The Bagging method creates multiple classifiers by training a classification model several times with different training datasets. The training data is obtained by randomly sampling the original dataset with replacement. The performance gain obtained by Bagging methods is usually the result of a reduction in the variance of model predictions. Different from Bagging, where the classifiers are trained independently to each other, in Boosting the classifiers are trained in sequence in a very adaptative way. The classification models are adjusted to give more importance to observations in the dataset that were poorly handled by previous models in the sequence. In such way, the Boosting method creates a stronger learner with lower bias than the initial classifier.

Recent studies on MPM have attempted to deal with the problem of imbalanced data for both the majority and minority classes. Xiong and Zuo (2017) proposed a cost-sensitive neural network to minimize misclassification of known mineralized locations. The proposed algorithm incorporates the cost of misclassification of known negative and positive classes in a neural network algorithm to learn the mineral prospectivity model. To address the bias toward the majority and negative class, Xiong and Zuo (2018) proposed a rare event logistic regression algorithm, which embeds sampling and decision threshold corrections into the original logistic regression algorithm. Wang et al. (2019) proposed a semi-supervised random forest algorithm to deal with the lack of known mineralized locations in MPM. The proposed classifier exploits the unlabeled data to enhance the performance of the classification model. At data level solutions, the use of SMOTE for synthetic over-sampling the non-mineralized class was proposed by Hariharan et al., (2017) to enhance prospectivity modeling by a random forest classifier. Li et al. (2019) recently further demonstrated the usefulness of SMOTE in random forest-based modelling of mineral prospectivity. These studies suggest that the use of SMOTE to over-sample the minority class and under-sample the majority class can improve the performance of mineral prospectivity models trained with unbalanced data.

In this context, here the effects of SMOTE on MPM are explored. We aim to evaluate if SMOTE can significantly improve the performance of ML algorithms in MPM. To achieve this objective, the original training dataset was modified. Using SMOTE, the mineralized locations were over-sampled and the non-mineralized locations randomly under-sampled at different ratios. Some 400 training datasets were produced with ratios of mineralized-to-non-mineralized samples ranging from 20:1 to 1:20. This strategy was used to evaluate the performance of an SVM algorithm under different ratios of mineralized-to-non-mineralized samples for modeling prospectivity of Cu-Au deposits in the Carajás Mineral Province (CMP), Brazil. The SVM algorithm was chosen because, unlike other ML algorithms, it is less sensitive to class imbalance, as class boundaries are calculated with respect to only a few data points and the class size may not affect the class boundary excessively (Japkowicz and Stephen, 2002b; Sun et al., 2009).

5.2 Study area

The CMP, located in northern Brazil, was chosen as a case study region (Fig. 1). It is the largest polymetallic mineral province in Brazil, hosting giant, enriched iron and manganese deposits and a world-class cluster of Cu-Au deposits (Grainger et al., 2008). The geology of the CMP has been mapped by the Geological Survey of Brazil (GSB/CPRM) at 1:250,000 scale. It is also covered by airborne magnetic, gamma-spectrometric and gravimetric surveys. The shape and size of the study area (Fig. 1) were defined by the airborne gravimetric survey coverage.

1. Geology and Cu-Au mineralization

The CMP is in the oldest part of the Amazonian craton, with Archean/Paleoproterozoic age (Cordani and Teixeira, 2007; Gibbs et al., 1986; Machado et al., 1991). Two tectonic blocks are distinguished in the CMP, namely the Rio Maria greenstone terrain to the south and the Itacaiúnas shear belt to the north(Santos et al., 2000; Tassinari and Macambira, 1999; Vasquez and Rosa-Costa, 2008). The basement of the northern block is composed of granulites, gneisses, and migmatites with age ranging from ~3.0 Ga to ~2.8 Ga (Althoff et al., 2000; de Oliveira et al., 2009; de Souza et al., 2001; Feio et al., 2013; Machado et al., 1991; Pidgeon et al., 2000). The Archean metavolcanic-sedimentary sequences, called Rio Novo Group and Itacaiúnas Supergroup ~2.75Ga (Martins et al., 2017), overlie the Itacaiúnas shear belt basement (DOCEGEO, 1988). These sequences are composed of iron formations, clastic sediments and both mafic and felsic volcanic rocks (DOCEGEO, 1988; Machado et al., 1991; Wirth et al., 1986). Between ~2.76 and ~2.56 Ga, these rocks were intruded by layered mafic-ultramafic complexes, gabbro dikes and sills, and by syn-tectonic alkaline granites (Barros et al., 2001; Dall'Agnol et al., 1997; Ferreira Filho et al., 2007; Galarza et al., 2007; Machado et al., 1991; Pimentel et al., 2003; Souza et al., 1996). Within-plate A-type granites and alkaline to sub-alkaline granites also intruded the metavolcanic-sedimentary sequences at ~1.88 Ga (Dall'Agnol et al., 2005, 1994; Tallarico et al., 2004).

The rocks of the Itacaiúnas Supergroup host the main Cu-Au mineralization in the CMP, which are generally associated with granites and rocks with hydrothermal alteration (Grainger et al., 2008; Xavier et al., 2012). The metal associations in these mineralization are iron oxide-Cu-Au-(Mo-Ag-U-REE) and Cu-Au-(W-Sn-Bi) (Grainger et al., 2008; Xavier et al., 2012). The former represents the largest Cu-Au deposits in the CMP, most of which are word-class (>200 Mt) deposits, such as Salobo, Cristalino, Sossego and Igarapé Bahia, Cento e Dezoito (118), Alemão, Breves, Águas Claras, Gameleira, and Estrela. The largest of them, Salobo, hosts 789 Mt @ 0.96 % Cu and 0.86 g/t Au (Craveiro et al., 2019; deMelo et al., 2017; Farias and Saueressig, 1982; Grainger et al., 2008; Huhn et al., 1999; Lindenmayer, 2003; Lindenmayer et al., 2005; Requia et al., 2003; Tallarico et al., 2017; Torresi et al., 2012; Vieira et al., 1988).

The iron oxide-Cu-Au-(Mo-Ag-U-REE) mineralization is commonly hosted by brecciated lower volcanic to volcano-sedimentary rocks and display intense Fe-K±Na metasomatism with chloritization and carbonatization. The ore minerals are mainly chalcopyrite, bornite and chalcocite. All the deposits are typically highly enriched in LREE with variable enrichments in Co, Ni, Pb, As, Mo, P, Th and U. The deposits are all located within shear zones, and the orebodies appear to be structurally controlled (Grainger et al., 2008; Haddad-Martim et al., 2017; Requia and Fontboté, 2000; Ronzê et al., 2000; Soares et al., 1999; Souza and Vieira, 2000; Tazava, 2000)

The Cu-Au mineralization with Cu-Au-(W-Sn-Bi) metallic association shares some characteristics with the iron oxide-Cu-Au group of deposits such as enrichment in As, Co, F, LREE, P, Th and U; Fe-K hydrothermal alteration with associated chloritization; and spatial association with faults and shear zones. However, this mineralization has distinct characteristics such as additional enrichment in Bi, Sn, and W; absence of Mo; abundance of quartz rather than Fe-oxides and silicates; and spatial association with Proterozoic granite stocks and/or dikes. The Cu-Au deposits in this group are smaller, generally <50 Mt in size (Grainger et al., 2008; Pollard et al., 2019).

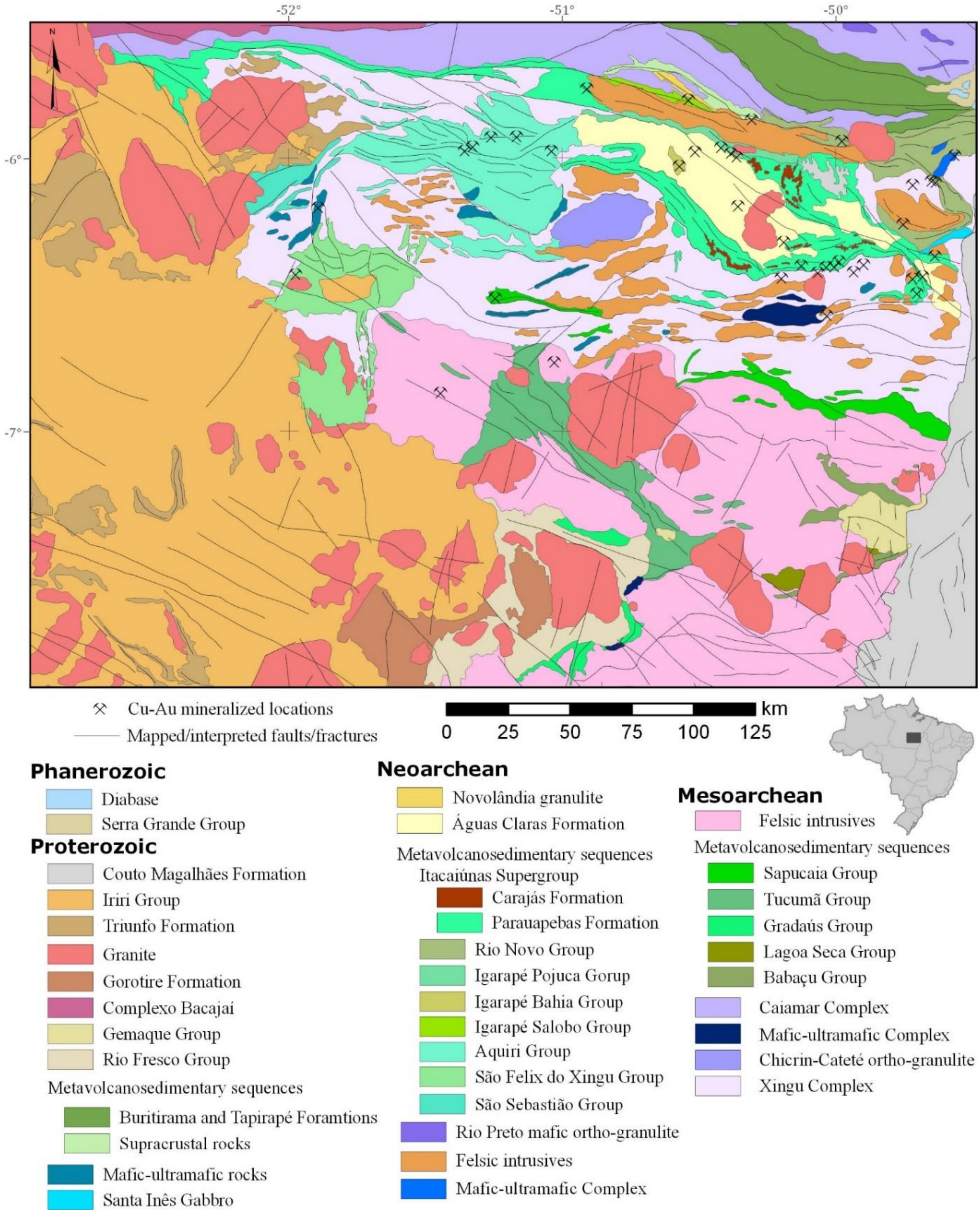


Fig. 1 Simplified geological map of the study area (modified from Vasquez and Rosa-Costa, 2008) and locations of Cu-Au deposits in the Carajás Mineral Province. (see Appendix B for mineralized location names).

5.2.1 Conceptual model for prospectivity of Cu-Au deposits

The spatial recognition criteria for prospectivity modeling of Cu-Au deposits in the CMP can be defined based on the mineral system approach to exploration targeting (Mccuaig et al., 2010). Table 1 lists the mineral system processes used to define the targeting model and their correspondent targeting elements and mappable criteria in the CMP. Based on the current knowledge of the CMP geology discussed above and the similarities of the Cu-Au mineralization with the class of iron oxide-copper-gold (IOCG) deposits, the processes considered critical for Cu-Au mineralization include (i) source of Cu and Au, (ii) active metal-carrying pathways, (iii) physical throttles that promote fluid flow and (iv) chemical scrubber that act as traps for the precipitation of metals.

Deep alkaline magma derived from melted metal-rich metasomatized subcontinental lithospheric mantle are important sources of Cu and Au for IOCG deposits (Groves et al., 2010, 2005; Groves and Vielreicher, 2001). The presence of alkaline, basic and ultrabasic lithologies and the proximity to contact between Mesoarchean, Neoarchean, and Paleoproterozoic lithologies were used here as proxies of possible metasomatized subcontinental lithospheric mantle zones. The latter targeting criterion represents trans-crustal structural zones (e.g., Motta et al., 2019), where subcontinental lithospheric mantle could have been metasomatized during previous tectonic events. These zones in the CMP are represented by the contact between Mesoarchean granulites, gneisses and migmatites of the basement and the Neoarchean meta-volcano-sedimentary sequences (Rio Novo Group and Itacaiúnas Supergroup), which are near or defined by regional shear zones (Vasquez and Rosa-Costa, 2008). The contact between Paleoproterozoic and Archean also likely represents one of these metal enriched zones, because Paleoproterozoic rocks are mainly represented by alkaline to sub-alkaline granites in the CPM, which are spatially associated with Cu-Au-(W-Sn-Bi) mineralization (Grainger et al., 2008). Most of the Cu deposits in the CMP occur near or within regional structures, such as the Cinzento and Carajás shear zones (Haddad-Martim et al., 2017; Xavier et al., 2012). Some of these regional structures likely acted as metal pathways that connected the deep metal source to the mineralized sites. Multi-scale edge algorithms, also known as worms (Fedi and Florio, 2001; Hornby et al., 1999), were applied to gravity and magnetic data to map deep-seated structures in the CMP. Suture zones between terrains of distinct ages also represent favorable regions for the presence of trans-crustal fault zones and ore-forming fluid transit. Therefore, proximity to the contact between Archean-

Paleoproterozoic and Proterozoic lithologies was also considered as a mappable targeting criterion for pathways for ore-forming fluids.

A map of faults and lithological boundaries, as a function of scale, known as geological complexity (Hodkiewicz, 2003), can be used as a proxy of strain accommodation zones, which permit significant fluid flow, as well as a proxy of important physical throttle for fluids (Ford and Blenkinsop, 2008). Therefore, geological complexity and proximity to Archean-Paleoproterozoic magmatism were used as mappable targeting criteria to define decompression zones with a high geothermal gradient; hence, zones with potential to transport large volumes of fluid to form the Cu-Au mineralization.

Fluid mixing and reaction within wall rocks are essential in the genesis of IOCG deposits (Groves et al., 2010, 2005; Groves and Vielreicher, 2001). The presence of certain alteration minerals can indicate the occurrence of these processes. In the CMP, the Cu-Au deposits generally exhibit alteration zones enriched in Fe-K and REE. Proxies for these alteration styles derived from the gamma-ray spectrometric data and the magnetic data were used for mapping zones with high concentrations of K, U, and magnetite, where Cu-Au deposition possibly occurred.

Considering the conceptual model for prospectivity of Cu-Au deposits in the CMP, only some stratigraphic units are associated with the IOCG Cu-Au mineral system (i.e., alkaline, basic and ultra-basic lithologies). However, in this work, all stratigraphic units (Fig. 1) were considered as model inputs, and the importance or weight of each stratigraphic unit in the prospectivity modeling was determined during the training of the SVM algorithm.

Table 1 Processes, targeting elements and mappable features comprising the conceptual model for prospectivity of Cu-Au deposits in the Carajás Mineral Province.

Critical Process	Source (magma, metals)	Active Pathway	Physical Throttle	Chemical Scrubber
Constituent Processes	Deep alkaline magmatic source	Trans-crustal and/or craton-scale fault zones	Decompression evidenced by brecciation zones	Fluid mixing
	Metasomatized subcontinental lithospheric mantle	Lithospheric craton margins	High geothermal gradient	Reaction with wall rocks
Targeting Elements	Alkaline magmatism associated with ultrabasic to basic rocks	Suture zones between terrains of distinct ages	Occurrence of large brecciation zone	Key alteration minerals (hematite, biotite, sericite, albite, Na amphibole, chlorite, uranium and/or REE rich minerals)
	Suture zones with multiple orogeny events		Volume of magmatic activity	Rocks with favorable chemistry (magnetite rich alteration zones)
Mappable Targeting Criteria	Geologic map (alkaline, basic and ultrabasic lithologies)	Deep structures mapped by gravity and magnetic worms	Geological Complexity map	Gamma Th/K, eU
	Contacts between Mesoarchean, Neoarchean Paleoproterozoic	Contacts between Mesoarchean, Neoarchean Paleoproterozoic	Geologic map (proximity to magmatism with the same age of mineralization)	Magnetic highs

5.3 Spatial data input

The spatial datasets used to produce the model input features include a geological map, map of faults/fractures, airborne geophysical data and locations of known Cu-Au deposits, which were provided by the GSB/CPRM. These datasets were pre-processed using GIS and Python libraries to extract the relevant proxy information. The datasets and the methodologies used to generate the input features for the prospectivity modeling are described in the following sections.

Geochemical data, such as those obtained from regional stream sediment surveys, are commonly used in MPM (e.g. Chen and Wu, 2016; Gao et al., 2016; Nykänen et al., 2008; Zuo and Carranza, 2011b; Zuo and Xiong, 2018). The geochemical data acquired at the CMP by the GSB/CPRM is rather voluminous but do not cover the study area homogenosly. In some specific sectors, the data display a good sample density, but for most of the study area the sample density is low or the data not acquired. This is largely due to the difficulty of access in the region imposed by the high density of forest and rivers. For this reason, we do not use geochemical data in this work.

Spatial pattern analysis of the known Cu-Au deposits was conducted first to calculate a suitable pixel size based on the spatial distribution of the deposits (Carranza, 2009). The analysis suggested that a 350 m pixel size is suitable for prospectivity modeling in the study area. However, because the airborne geophysical data have a smaller spatial resolution of 125 m, this was adopted as the spatial resolution for the prospectivity modeling in order not to lose information in the re-sampling of these data. Other raster inputs were re-sampled, and vector inputs were rasterized to this pixel size.

Before combining the spatial inputs into a set of feature vectors for SVM model training, a linear transformation was applied to all continuous spatial data in order to encode each evidential layer with values varying from 0 to 1, using the following normalization equation:

$$x_{norm} = x/x_{max}$$

where x_{norm} is the transformed value, x is the original data value, and x_{max} is maximum value of the original data (Burkov, 2019; Juszczak et al., 2002). Normalization of input data, also known as feature scaling, is an important step to avoid instability in the training of ML models that use gradient descent optimization techniques, as SVMs do. The weights of features with higher values will update much faster than others, causing the model to learn incorrect patterns.

1. **Geological map**

The digital 1:1,000,000 scale geological map of Pará state, Brazil (Vasquez and Rosa-Costa, 2008) was subjected to vector map operations to produce the following inputs: the presence of a specific stratigraphic unit (for all stratigraphic units in the work area; Fig. 1), proximity to contact between Mesoarchean, Neoarchean and Paleoproterozoic units, and geological contacts (all contacts). Prior to processing, the geologic map was simplified; that is, empty spaces corresponding to river paths were filled with the underlying stratigraphic units while Quaternary stratigraphic units were replaced with the underlying bedrock (Fig. 1).

The geologic map is categorical data, but SVM requires all input variables to be numerical. Therefore, one-hot encoding was adopted to transform it into a numerical input. In this method, each category is mapped to a vector that contains 1 and 0, denoting the presence or absence of a feature, respectively. Thus, each stratigraphic unit of the geologic map was individually rasterized into a binary raster, which has values 1 inside the stratigraphic unit and 0 outside it. Consequently, the presence of a stratigraphic unit was represented by 38 input features in the final model, each representing an independent stratigraphic unit.

The contacts between Mesoarchean and Proterozoic stratigraphic units, Neoarchean and Proterozoic stratigraphic units, as well as the contacts between Paleoproterozoic and Proterozoic stratigraphic units were extracted from the geologic map. Then, they were rasterized and a map of Euclidean distance from these contacts was generated as another input feature for the prospectivity modeling (Fig. 2). This feature was used as a proxy of possible metal sources and fluid pathways (Table 1).

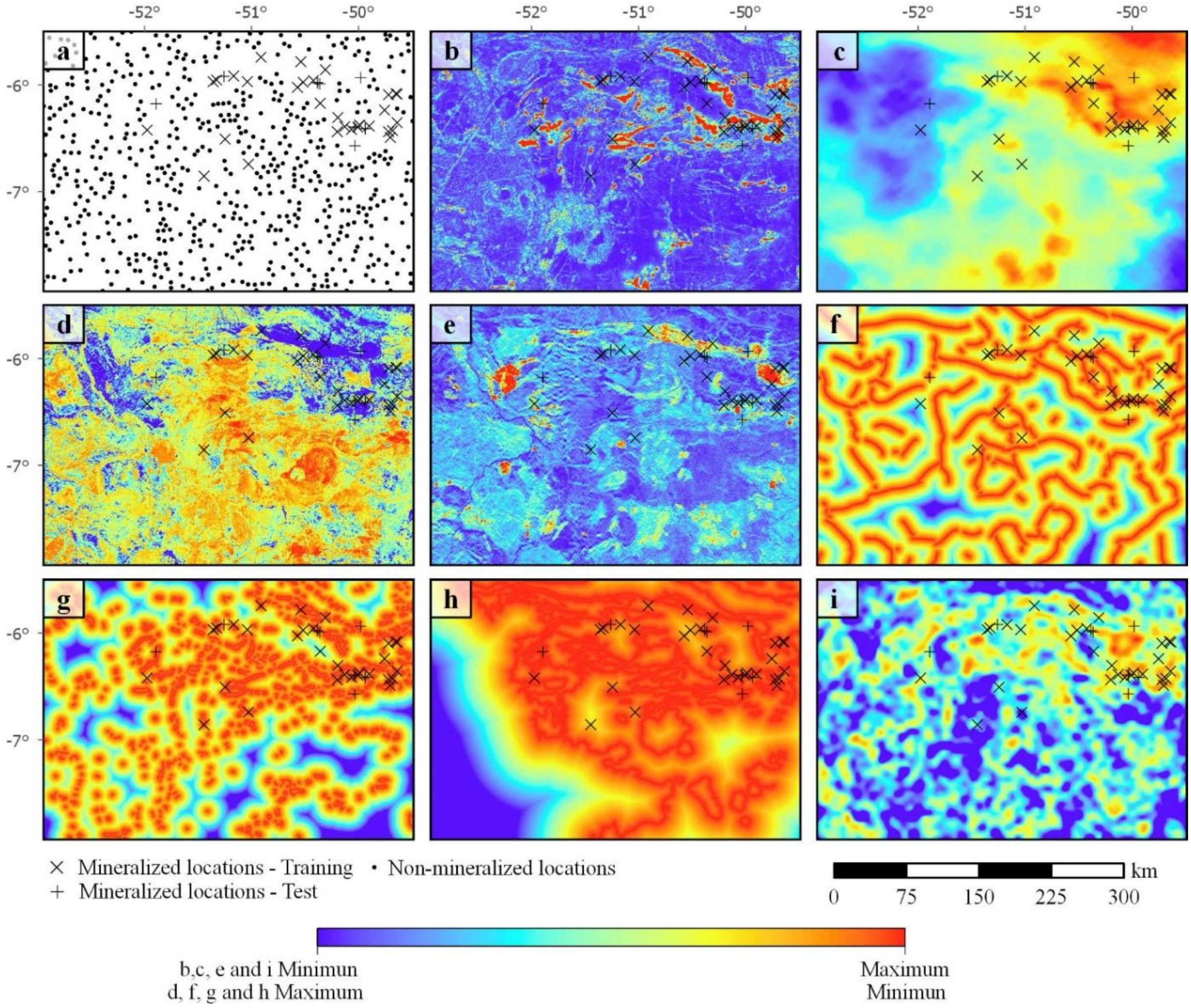


Fig. 2 Spatial data input used in SVM modeling. (a) Target variables, showing deposits (training and testing) and non-deposits locations. (b) Total gradient of anomalous magnetic field map derived from airborne magnetic data. (c) Simple Bouguer anomaly map derived from airborne gravity data. (d) eTh/K ratio map derived from airborne gamma-ray spectrometric data. (e) eU concentration map derived from airborne gamma-ray spectrometric data. (f) Gravity worms derived from 10km upward continuation of Bouguer data. (g) Magnetic worms derived from 5km upward continuation of magnetic data. (h) Proximity to Mesoarchean/Proterozoic, Neoarchean/Proterozoic and Paleoproterozoic/Proterozoic stratigraphic units contacts. (i) Geological complexity map.

5.3.1 Geological Complexity

Fractal analysis of faults and lithological boundaries (Ford and Blenkinsop, 2008; Hodkiewicz, 2003) was performed to produce a map of geological complexity of the CMP (Fig. 2) and used as a proxy of physical throttle (Table 1). The map of faults/fractures was produced by interpretation of shaded-relief images derived from the 30 m resolution Shuttle Radar Topographic Mission digital elevation model. This map indicates both shallow structures and deep structures that reach the surface. The map of faults/fractures and the map of lithological contacts of interest, extracted from the geological map (Vasquez and Rosa-Costa, 2008), were used to represent these features in the fractal analysis.

Geological complexity was then quantified by calculating the fractal dimension of the spatial distribution of faults/fractures and lithological contacts in the CMP using the box-counting method (Hirata, 1989; Mandelbrot, 1983). This method consists of superimposing grids with square cells of size d, and counting the number of cells, N_d, containing any of the analyzed features (faults/fractures and/or lithological contacts). This counting process is repeated n times, with grids of cell size $d_n = \frac{d}{2^{n-1}}$. To calculate the fractal dimension, D, the slope of a line on a log-log plot of N_{d_n} vs d_n is measured, such that:

$$N_{d_n} \propto {d_n}^{-D}$$

where D is a value between 1 and 2 for a two-dimensional map and it is directly proportional to geological complexity.

To calculate geological complexity in this study, a grid of points p regularly spaced at 5 km was defined in the whole of the CMP, and fractal dimension D was measured for each point. It was considered appropriate to adopt 5 km grid spacing for the level of detail according to the 1:1,000,000 scale CMP geological map, and based on the methodologies of previous publications on geological complexity using the box-counting method (Gillespie et al., 1993; Hodkiewicz, 2003; Walsh and Watterson, 1993). The maximum cell size d adopted to compute D at each point was 5 km, equal to the spacing between the points, resulting in four cells centered at p when $n = 1$. The box-counting process was repeated five times at each point, counting N_{d_n} for cell grids with size d_n equal to 5, 2.5, 1.25, 0.625 and 0.3125 km, respectively. Finally, the grid with computed D was interpolated to generate the geological complexity surface raster using a two-dimensional minimum curvature spline function.

The algorithm used in this study to calculate the geological complexity was written as part of this work and it is based on the methodology outlined in Ford and Blenkinsop (2008) and Hodkiewicz (2003). The code is available online (Prado, 2020).

5.3.2 Airborne geophysics

The airborne geophysical dataset consists of magnetic, gamma-ray spectrometric and gravity data provided by the GSB/CPRM. The aeromagnetic and gamma-ray spectrometric data are a compilation of four surveys: Rio Maria, Oeste de Carajás, Tucuruí and Anapu-Tuerê (GSB/CPRM - Geological Survey of Brazil, 2015a, 2015b, 2010, 2004), acquired between 2004 and 2015. These surveys had N-S flight lines 500 m apart, E-W control lines 10 km apart and a mean sensor clearance of 100 m above surface. The data were interpolated into a grid using the bi-directional line gridding method with a 125 m cell size.

The aeromagnetic data are available as total magnetic intensity values. To center magnetic anomaly values above their magnetic source, the total gradient (or analytic signal amplitude; Roest et al., 1992) of the anomalous magnetic field was calculated. The total gradient was used to identify magnetic highs as a proxy of chemical scrubber (Fig. 2b; Table 1).

Regions of K and/or U enrichment were mapped from the gamma-ray spectrometric data. Given the high geochemical mobility of K, the K concentrations were normalized to Th, which is less mobile than K and U. The ratio eTh/K was used to map K enriched zones (Shives et al., 2000) (Fig. 2d). U enriched zones were mapped directly with eU concentrations (Fig. 2e). Both eTh/K and eU maps were used as proxies of chemical scrubber (Table 1).

Airborne gravity data comprise measurements from the aerogravimetric Carajás survey (GSB/CPRM - Geological Survey of Brazil, 2015c), with N-S flight lines 3 km apart, E-W control lines 12 km apart and a mean sensor clearance of 900 m above surface. The Bouguer anomaly values provided by the survey were interpolated into a grid using the bi-directional line gridding method with a 600 m cell size.

The interpretation and modeling of this Bouguer gravity anomaly map carried by Motta et al. (2019) and Oliveira (2018) suggests that the main regional NNE-striking gravity high (Fig. 2c) have a strong correlation with the oldest part of the Amazonian Craton, which corresponds to the Rio Maria, Carajás and Bacajá tectonic domains, according to the tectonic subdivision proposed in Vasquez and Rosa-Costa (2008). Almost all known Cu-Au mineralization are located in this region. The regional gravity lows at NW and SE of the gravity high (Fig. 2c) are correlated with younger rocks of the Iriri-Xingu domain and the Araguaia belt, respectively, according to the same tectonic subdivision. The Iriri-Xingu domain is composed by a volcano-plutonic association of Orosirian age; the Araguaia belt is a Neoproterozoic fold belt composed mainly of metasedimentary rocks (Cordani and Teixeira, 2007; Fonseca et al., 2004; Vasquez and Rosa-Costa, 2008). The linear gravity highs to the NE of the Bouguer anomaly map, oriented mainly in E-W and WNW-ESE direction (Fig. 2c), are correlated with outcrops of Itacaíunas Supergroup metavolcano-sedimentary rocks, composed by the Parauapebas and Carajás Formation. Most of the known Cu-Au mineralization is related or hosted by the Parauapebas Formation. The spatial resolution of the aerogravimetric survey allows only regional-scale interpretations, therefore gravimetric variations at deposit scale are not identifiable.

An algorithm for detecting multi-scale edges, also known as worms (Fedi and Florio, 2001; Hornby et al., 1999), was used to automatically extract the positions of gradients in the Bouguer gravity data and the total gradient of magnetic data. These gradients represent the contacts or edges between bodies of contrasting density or magnetic susceptibility and can provide additional information about the location of deep-seated structures. The worms (gradient strings) were derived from the 10 km upward continuation of the Bouguer gravity data, and from the 5 km upward continuation of the total gradient of magnetic data. The upward continuation level was empirically determined based on the analysis of the proximity of Cu-Au deposits to worms generated at different levels. The levels chosen were those where the largest number of deposits were close to the worms. Structures mapped with worms have depth estimated as half the level of the upward continuation (Hornby et al., 1999; Jacobsen, 1987); thus, the computed worms represent structures that may persist roughly 5 km below surface. After this processing, the worms were rasterized with 125 m cell size, and Euclidean distance from such features was calculated and used as a proxy of possible fluid pathways (Figs. 2f and 2g; Table 1).

5.3.3 Target variable

SVM algorithms are binary classifiers; therefore, they require that a training dataset comprises data of two distinct classes. As MPM attempts to distinguish prospective zones from non-prospective ones, mineralized and non-mineralized locations were used in the analysis. Mineralized locations were labeled as 1 and non-mineralized locations as -1. The locations of discovered Cu-Au deposits used for the modeling were obtained from a compilation of GSB/CPRM datasets and data provided in Haddad-Martim et al. (2017) and Motta et al. (2019). The data include locations of 38 known Cu-Au deposits. To measure the performance of the model and ensure that the model can deal with new data, 30 (or ~80%) of the mineralized locations were randomly selected for training the model and the remaining 8 (or ~20%) were used as an independent set for testing the model. (Figs. 1 and 2). According to previous publications, the proportion of 80%/20% for training/testing is suitable for the development of ML models (Shahin et al., 2004; Swingler, 1996).

The non-mineralized locations were randomly selected, with the restriction of being at a distance greater than 12.5 km from known deposits and a minimum distance of 5 km between them. Point pattern analysis of the deposit locations was performed using the pair correlation function (Boots and Getis Arthur, 1988; Diggle, 1983) $g(r)$ to determine the first restriction. The pair correlation function calculates the probability of finding a deposit at a given distance from another deposit. According to the point pattern analysis, the known Cu-Au deposits in the CMP are arranged in clusters that have two main grouping distances (Fig. 3). At deposit scale, deposits are clustered within 2.7 km from each other, and at district scale, deposits are clustered within 8.5 km from each other. The analysis also suggested that at distances greater than 12.5 km from a known deposit, the probability of finding another deposit is considerably reduced. Hence, the non-mineralized locations were sampled from areas beyond 12.5 km of every known Cu-Au deposit location. The second restriction was imposed to force non-mineralized locations to sample a greater diversity of input features, avoiding clustering of non-mineralized locations at regions with similar patterns. Given these restrictions, 600 non-mineralized locations were randomly generated for use in training the prospectivity model (Fig. 2a).

The mineral system proposed in section 2.2 states that Cu-Au mineralization in the CMP is related to certain lithostratigraphic units. Here, these units were not used to restrict non-mineralized sites. Such restriction could create a bias in the model. As such, the lithostratigraphic units with mineralization in would be overweighted by the model, hiding the potential of areas that have characteristics similar to those found in the mineralized zones, but are located in other stratigraphic units.

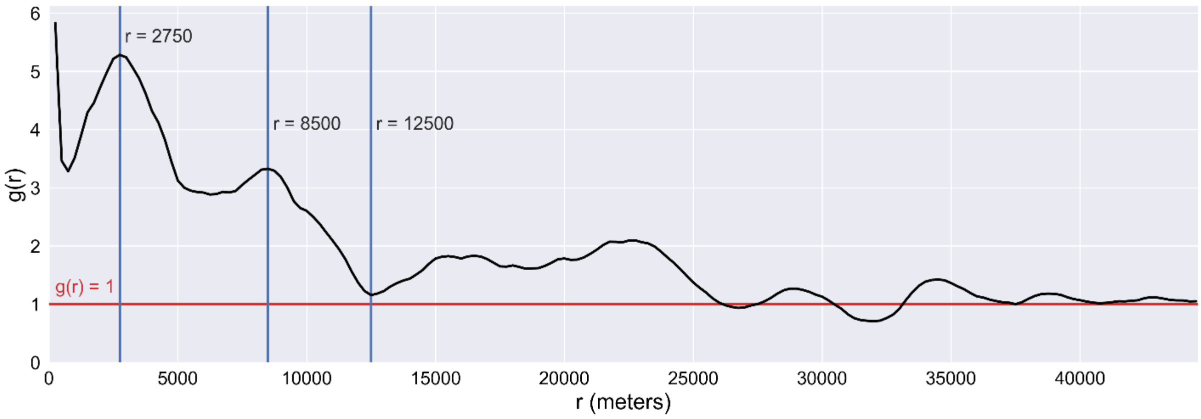


Fig. 3. Pair correlation function curve of CMP Cu-Au deposits. Values of g(r) > 1 indicate that the deposits are more clustered than expected under completely spatial randomness, with g(r) < 1 indicating that the deposits are more dispersed than expected under completely spatial randomness.

5.4 Methodology

1. Support vector machines classification algorithm

Proposed in the late 1960s, SVMs are learning algorithms based on maximum margin strategy (Vapnik, 1963, 2000). When firstly proposed, SVMs were for linear two-class classification. That is, if data are linearly separable, SVMs will try to find the equation of the optimal separating hyperplane that best separates the classes, where the distance from the separating hyperplane to the closest data point (margin) is minimal. The SVM algorithm considers only data points at the margin for learning the optimal hyperplane; these data points are called support vectors.

Mathematically, a training dataset is given by $(x_1, y_1), \ldots, (x_l, y_l), x \in R^n, y \in \{+1, -1\}$, where x_l is a vector of n input-features with label y_l being 1 (e.g., mineralized class) or -1 (e.g., non-mineralized). The hyperplane function has the equation $\boldsymbol{w}^{\mathsf{T}}x_l + b$, where $\boldsymbol{w}$ is a weight vector that is normal to the hyperplane, and b is the bias that moves the hyperplane away from the origin. This function is defined so that, for all data points of class $y_l = -1$, $\boldsymbol{w}^{\mathsf{T}}x_l + b \leq -1$, and for data points of class $y_l = 1$, $\boldsymbol{w}^{\mathsf{T}}x_l + b \geq 1$, both of which are the same as $y_l(\boldsymbol{w}^{\mathsf{T}}x_l + b) \geq 1$. The region between the hyperplanes $\boldsymbol{w}^{\mathsf{T}}x_l + b \leq -1$ and $\boldsymbol{w}^{\mathsf{T}}x_l + b \geq 1$ should have no data points and is called the "margin". The function $\boldsymbol{w}^{\mathsf{T}}x_l + b = 0$ defines the dividing plane. The class of a vector x_l is defined by it sign; that is, $\widehat{y_l} = siqn(\boldsymbol{w}^{\mathsf{T}}x_l + b)$, where $\widehat{y_l}$ is the predicted class of a data point (Vapnik, 1963).

Considering two support vectors that lie on opposite margins (x_+ and x_-) and are as close as possible to one another, the vector that connects these two vectors will be perpendicular to the hyperplane defining the margins $\boldsymbol{w}^{\mathsf{T}}x_+ + b = 1$ and $\boldsymbol{w}^{\mathsf{T}}x_- + b = -1$. Subtracting the equations of the two margins generates $\boldsymbol{w}^{\mathsf{T}}(x_+ - x_-) = 2$. Because the vectors $\boldsymbol{w}$ and $x_+ - x_-$ are perpendicular to the hyperplane, and are parallel to each other, the equation can be written as $||\boldsymbol{w}|| * ||x_+ - x_-|| = 2$, where $||\boldsymbol{w}||$ is the length of vector $\boldsymbol{w}$. Simplifying this equation, by dividing both sides by $||\boldsymbol{w}||$, gives $||x_+ - x_-|| = 2/||\boldsymbol{w}||$, which is the distance between the hyperplanes. In order to maximize the margin, the SVM algorithm tries to find the minimum $||\boldsymbol{w}||$ that maintains the relationship $y_l(\boldsymbol{w}^{\mathsf{T}}x_l + b) \geq 1$ for all vectors in the training dataset. To solve the problem, Vapnik (1963) proposed to transform $||\boldsymbol{w}||$ into the equivalent function $\frac{1}{2}||\boldsymbol{w}||^2$ in order to use quadratic programming optimization to get the solution. The saddle point of the Lagrange function corresponds to the solution of the optimization problem, and the Lagrange multipliers α are determined by maximizing the output of

$$w(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{ij=1}^{l} \alpha_i\alpha_j y_i y_j (x_i x_j)$$

while satisfying the constraints that all $\alpha_i \geq 0$ and $\sum_{i=1}^{l} \alpha_i y_i = 0$. The optimal multipliers α are composed mostly of zeros. The vectors x_i that have non-zero α_i are those that fall on the margin, satisfying $y_i(\boldsymbol{w}^{\mathsf{T}}x_i + b) = 1$, and are the only vectors that contribute to the calculation of $\boldsymbol{w}$. These are the support vectors for the model.

To deal with non-linearly separable datasets, a kernel function $\boldsymbol{k}(x_i x_j)$ is used to capture a non-linear hyperplane, and the multipliers α are determined by maximizing the kernel function (Vapnik, 2000):

$$w(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{ij=1}^{l} \alpha_i\alpha_j y_i y_j \boldsymbol{k}(x_i, x_j)$$

The objective of this kernel function is to project the data into a transformed space, with higher dimension, where a linear dividing hyperplane can be found. The kernel function for a linear SVM is $k(x_i, x_j) = x_i^{\mathsf{T}} x_j$. The most used non-linear kernels are the polynomial $\mathrm{k}\left(x_i, x_j\right) = \left(\gamma x_i^\intercal x_j + \mathrm{C}\right)^{\mathrm{d}}$, the radial basis function $k\left(x_i, x_j\right) = exp\left(-\gamma||x_i - x_j||^2\right)$, and the sigmoid $k\left(x_i, x_j\right) = tanh\left(\gamma x_i^\intercal x_j \ + \ C\right)$, where γ, C and d are hyperparameters known, respectively, as kernel coefficient, kernel independent term, and degree of polynomial kernel.

After training, and then computing the optimal Lagrange multipliers α that maximize $w(\alpha)$, the SVM algorithm computes the class of a new data point by utilizing the output of the function

$$\hat{y} = sign\left(\sum_{i=1}^{l} \alpha_i y_i k\left(x_{i,} x\right) + b\right)$$

where $\hat{y}$ is the predicted class for vector x.

5.4.1 Synthetic minority over-sampling technique (SMOTE)

There are ways to enhance the performance of a SVM classifier when trained with an imbalanced number of control data points. Chawla et al. (2002) proposed SMOTE to over-sample the minority class by generating “synthetic” training samples or points. The technique to generate the synthetic training points involves taking the difference between a feature vector (sample) and its nearest neighbor, multiplying the difference by a random number between 0 and 1, and adding the result to the feature vector under consideration. This technique generates a random feature vector along the line segment between the sample and its nearest neighbor. The minority class is over-sampled by taking each minority class data point and calculating the random synthetic sample that lies between any or all the *k* minority class nearest neighbors. The number of nearest neighbors used is proportional to the amount of over-sampling needed (Chawla et al., 2002). The synthetic samples modify the decision boundaries for the minority class space by spreading it further into the majority class space, creating more general regions for the minority class and preventing majority class over-fitting. The effect is a classifier that generalizes the training data and improves the classification performance (Batista et al., 2004; Chawla et al., 2002; He and Garcia, 2009).

5.4.2 Sampling of training data

The original training dataset (30 mineralized locations and 600 non-mineralized locations) was interactively re-sampled using SMOTE. A Python package called imbalanced-learn (Lemaître et al., 2017), which implements the SMOTE algorithm, was used to produce the synthetic samples. The code is available online (https://github.com/Eliasmgprado/GeologicalComplexity_SMOTE; Prado, 2020). The class of mineralized locations was over-sampled at a rate N varying from 100% to 2000% at increments of 100 to generate synthetic samples. The class of non-mineralized locations was randomly under-sampled at a rate N varying from 100% to 5% at steps of 5. The total number of samples generated at each re-sampling interaction was calculated as $\frac{N}{100} * D$, where N is either over-sampling or under-sampling rate and D is the corresponding total number of mineralized or non-mineralized locations in the original training dataset. That is, the interaction with 300% over-sampling of the mineralized class and 100% under-sampling of the non-mineralized class, produced a training dataset with 90 mineralized samples (30*300/100; 30 original locations and 60 synthetic samples) and 600 non-mineralized samples (600*100/100). In this way, 400 training datasets with different ratios of mineralized to non-mineralized samples were obtained, one per over-sampling/under-sampling rate pair, including the original dataset, which corresponds to the 100%/100% pair.

It is emphasized here that the increase in the number of samples of mineralized class by over-sampling the known mineralized locations does not represent an increase in the number of mineralized locations in the work area. The over-sampling only increases the number of training vectors of the mineralized class, generating synthetic vectors, in feature or vector space but not in geographic space. That is, these vectors do not have geographical coordinates, so they do not represent locations on the map, but the samples (both original and synthetic) exist only in the vector space of the model. Assuming that the 30 known mineralized locations used for training have similar characteristics in the input vector space, SMOTE increases the importance of features related to these locations. To assess the changes in model performance produced by re-sampling the original dataset, an SVM model was trained for each one of the 400 training datasets, and then the performance for each model was compared. Details on the training methodology adopted and the performance metrics used are provided in the next session.

5.4.3 Performance Metrics

When dealing with imbalanced training data, care should be taken in the selection of the performance metric used for assessing the classification performance and guiding the classifier modeling. Some metrics are not suitable, like accuracy, given by the ratio of number of correct predictions (true positives + true negatives) to the total number of predictions made (total number of samples classified). The issue is that the minority (or positive) class has very little impact on the accuracy as compared to that of the majority (or negative) class. The performance metric selection needs to consider which class has the higher misclassification cost (the weight the model will give for misclassifying a determined class) according to the model objectives. In MPM, although both mineralized and non-mineralized classes have a high cost of being misclassified, misclassifying a mineralized area as non-mineralized (i.e., false negative) has a higher cost than misclassifying a non-mineralized area as mineralized (i.e., false positive), as the focus of MPM is to make the ML model learn to identify patterns associated with mineralized areas. Consequently, the chosen performance metric should evaluate the prospectivity model according to its ability to identify patterns associated with mineralized areas. The model's ability to identify non-mineralized areas is of lesser importance.

Therefore, in order to evaluate the prospectivity model performance appropriately, the F1 score (Van Rijsbergen, 1979) was used here to assess the classification performance and guide the selection of model hyperparameters. The metric is obtained by calculating the harmonic mean between recall and precision. Recall, also known as true positive rate, is defined as the percentage of samples from the positive (or mineralized) class that were correctly classified as positive; that is, the ratio of true positives to the total number of positive class samples (i.e., true positives + false negatives):

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Precision is defined as the percentage of samples classified by the model as belonging to the positive class, which were correctly classified as positive. In other words, it is the ratio of true positives to the total number of samples classified as positive (i.e., true positives + false positives):

$$Precission = \frac{true\ positives}{true\ positives + false\ positives}$$

Mathematically, the F1 score can be expressed as:

$$\text{F1 Score} = 2 \cdot \frac{precision \cdot recall}{\text{precision} + \text{recall}}$$

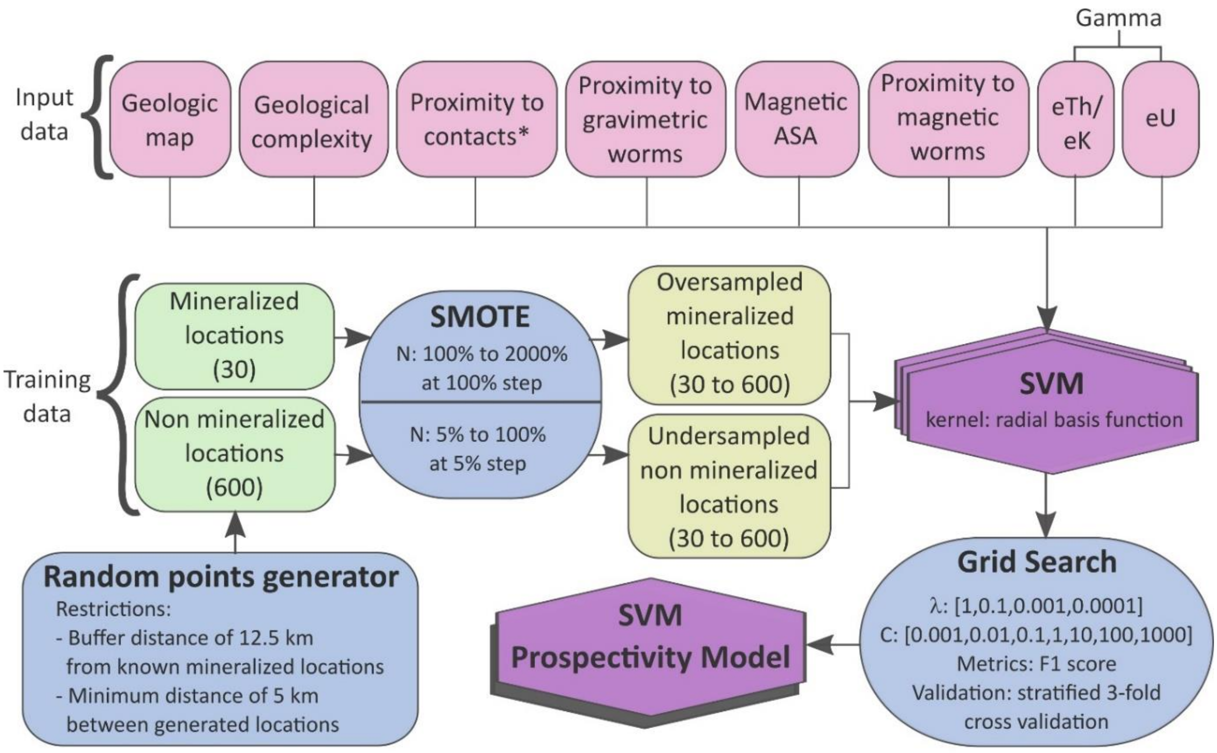
The F1 score is the harmonic mean of the precision and recall. It considers only the performance of the positive class to evaluate a model. It evaluates the precision of the model in the classification of positive samples; that is, how many samples were correctly classified as positive (or mineralized), as well as how robust the classification is, measuring the capability of the model not to classify a negative sample (non-mineralized) as positive (or mineralized).

It is important to note that these metrics assume that the classes assigned to the training and validation data are correct. Therefore, if a non-mineralized location used for training is, in fact, a mineralized location, the number of false positives used for the metric calculation will be wrong, distorting the precision value, and consequently the F1 score. In this work, the non-mineralized locations were randomly selected, assuming some restrictions (see section 3.4) that do not prevent the occurrence of mineralization, only decreasing their probability. So, the absolute value of the F1 score for some models may be underestimated. However, for performance comparison purposes between the models, this distortion can be disregarded, since it affects all models in the same proportion.

5.4.4 Application of SVM

The Scikit-learn (Pedregosa et al., 2011) python library, which implements ML algorithms, including SVM, was used for the ML model generation and prediction. The hyperparameters needed to be set for the SVM implementation are the penalty of the error term C, the kernel function k, and the kernel coefficient λ. The radial basis function was used as the kernel function k for training the models. The kernel coefficient λ determines the width of the radial basis function. The penalty of the error term C determines the influence of the misclassification on the estimation of hyperplane. To determine the optimal hyperparameter C and λ for each model an exhaustive search was performed using the grid search algorithm available in the Scikit-learn library.

The grid search algorithm was configured to obtain the best parameter based on the F1 score of a stratified 3-fold cross validation. The k-fold cross-validation method (Plutowski et al., 1994) consists of subdividing the training dataset into k subsets (folds) of approximately equal size. In the stratified variation of the k-fold cross-validation, k subsets are made by preserving the original percentage of samples for each class. After splitting, the ML model is trained k times, each using the samples in k-1 subsets for training and the samples in the remaining subset for model validation. Also, the algorithm was configured to search for the following values of C: 0.001, 0.01, 0.1, 1, 10, 100, 1000, and for the following values of λ: 1, 0.1, 0.001, 0.0001. A flowchart of the model construction process is shown in Fig. 4.

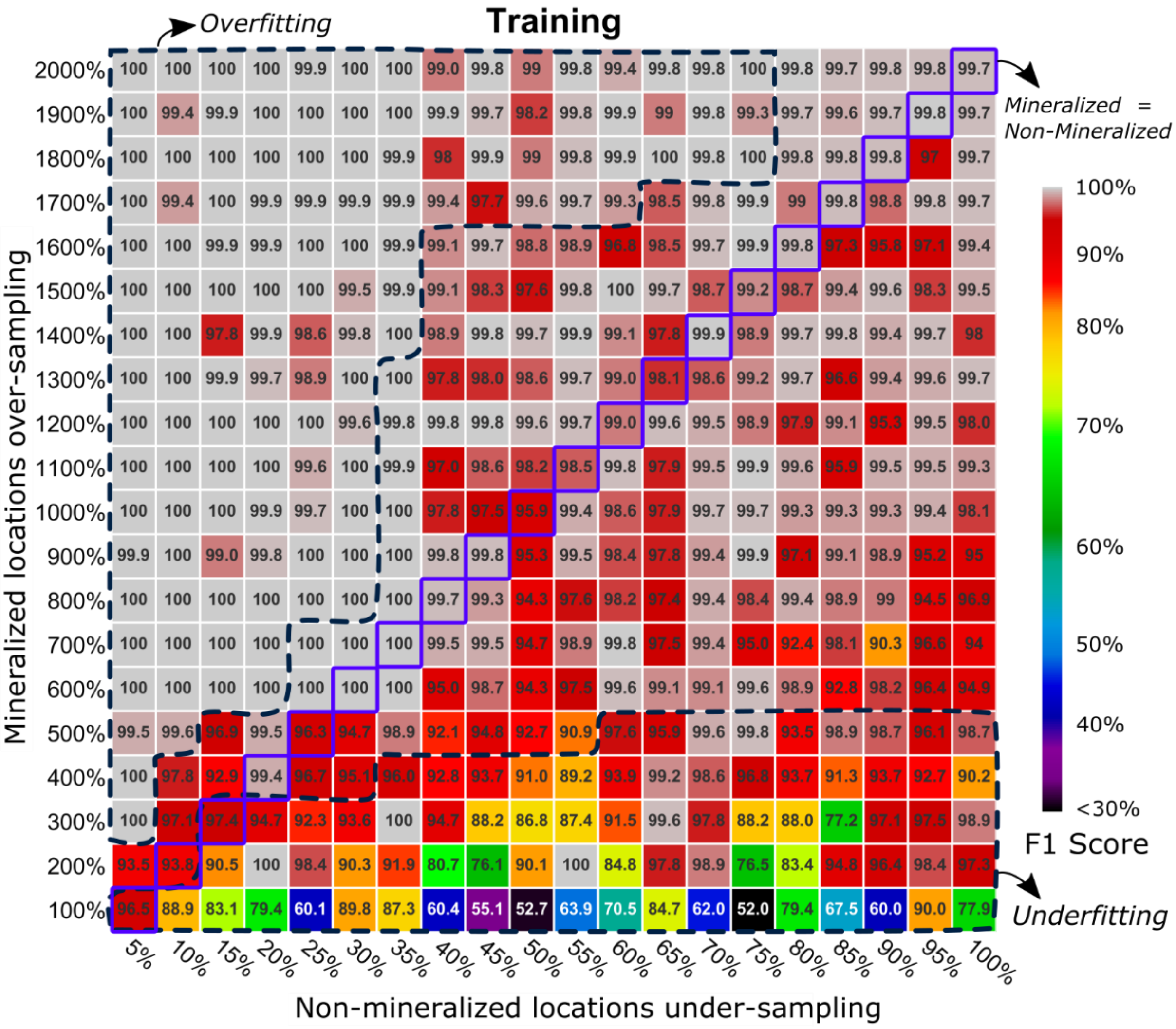

Gamma
Input data
Geologic map
Geological complexity
Proximity to contacts*
Proximity to gravimetric worms
Magnetic ASA
Proximity to magnetic worms
eTh/eK
eU
Training data
Mineralized locations (30)
Non mineralized locations (600)
SMOTE
N: 100% to 2000% at 100% step
N: 5% to 100% at 5% step
Oversampled mineralized locations (30 to 600)
Undersampled non mineralized locations (30 to 600)
SVM
kernel: radial basis function
Random points generator
Restrictions:
- Buffer distance of 12.5 km from known mineralized locations
- Minimum distance of 5 km between generated locations
SVM Prospectivity Model
Grid Search
λ: [1,0.1,0.001,0.0001]
C: [0.001,0.01,0.1,1,10,100,1000]
Metrics: F1 score
Validation: stratified 3-fold cross validation

*Proximity to contacts between Mesoarchean and Proterozoic stratigraphic units, Neoarchean and Proterozoic stratigraphic units, as well as contacts between Paleoproterozoic and Proterozoic stratigraphic units.

Fig. 4. Flowchart of the model construction process.

5.5 Results and discussion

By following the methodology described above, a SVM model was trained using each one of the 400 training datasets. After training, the F1 scores against training and testing sets were computed for individual models for measuring their performance. The values obtained are shown in Figs. 5 and 6.

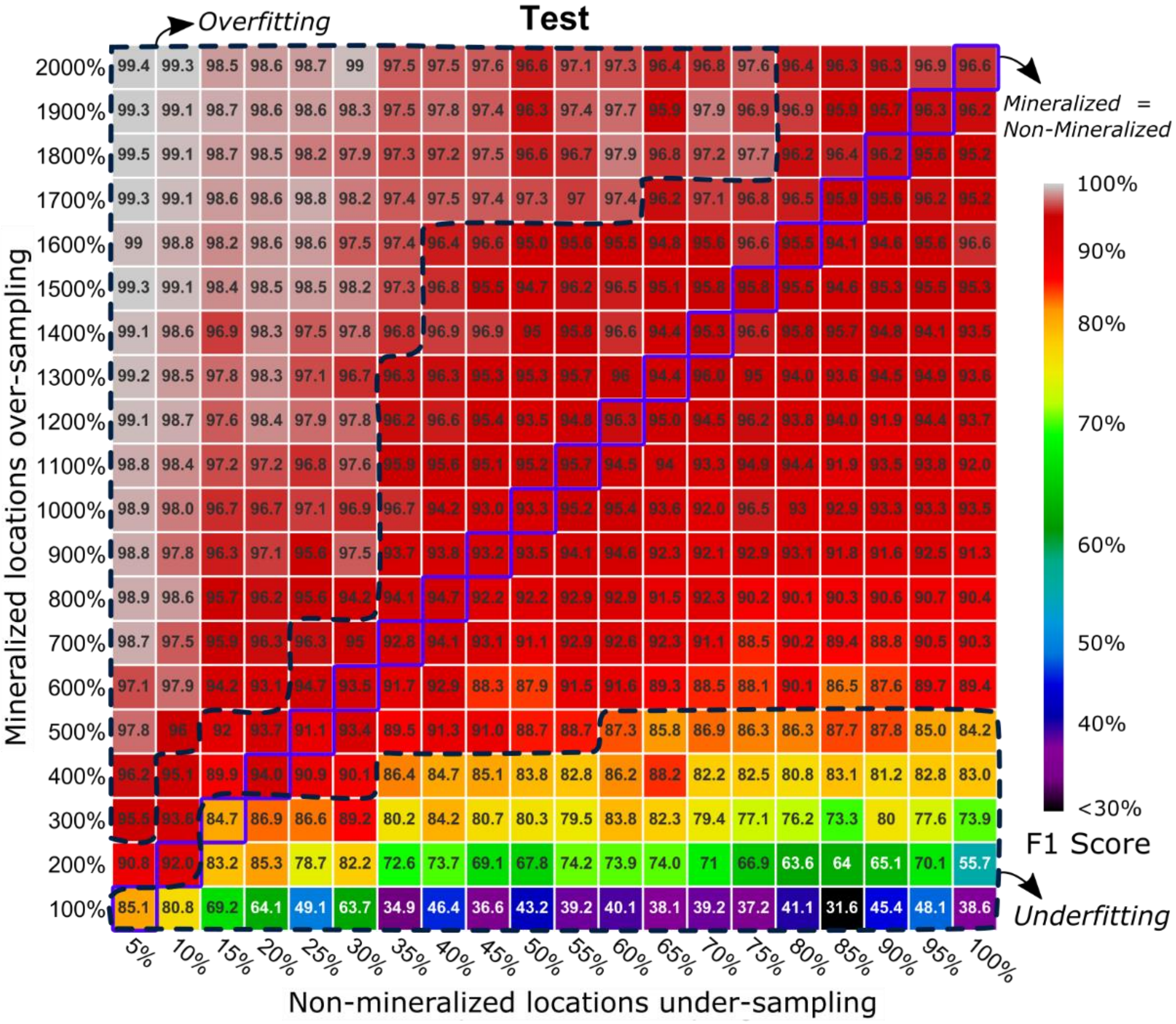


	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
2000%	100	100	100	100	99.9	100	100	99.0	99.8	99	99.8	99.4	99.8	99.8	100	99.8	99.7	99.8	99.8	99.7
1900%	100	99.4	99.9	100	100	100	100	99.9	99.7	98.2	99.8	99.9	99	99.8	99.3	99.7	99.6	99.7	99.8	99.7
1800%	100	100	100	100	100	100	99.9	98	99.9	99	99.8	99.9	100	99.8	100	99.8	99.8	99.8	97	99.7
1700%	100	99.4	100	99.9	99.9	99.9	99.9	99.4	97.7	99.6	99.7	99.3	98.5	99.8	99.9	99	99.8	98.8	99.8	99.7
1600%	100	100	99.9	99.9	100	100	99.9	99.1	99.7	98.8	98.9	96.8	98.5	99.7	99.9	99.8	97.3	95.8	97.1	99.4
1500%	100	100	100	100	100	99.5	99.9	99.1	98.3	97.6	99.8	100	99.7	98.7	99.2	98.7	99.4	99.6	98.3	99.5
1400%	100	100	97.8	99.9	98.6	99.8	100	98.9	99.8	99.7	99.9	99.1	97.8	99.9	98.9	99.7	99.8	99.4	99.7	98
1300%	100	100	99.9	99.7	98.9	100	100	97.8	98.0	98.6	99.7	99.0	98.1	98.6	99.2	99.7	96.6	99.4	99.6	99.7
1200%	100	100	100	100	100	99.6	99.8	99.8	99.8	99.6	99.7	99.0	99.6	99.5	98.9	97.9	99.1	95.3	99.5	98.0
1100%	100	100	100	100	99.6	100	99.9	97.0	98.6	98.2	98.5	99.8	97.9	99.5	99.9	99.6	95.9	99.5	99.5	99.3
1000%	100	100	100	99.9	99.7	100	100	97.8	97.5	95.9	99.4	98.6	97.9	99.7	99.7	99.3	99.3	99.3	99.4	98.1
900%	99.9	100	99.0	99.8	100	100	100	99.8	99.8	95.3	99.5	98.4	97.8	99.4	99.9	97.1	99.1	98.9	95.2	95
800%	100	100	100	100	100	100	100	99.7	99.3	94.3	97.6	98.2	97.4	99.4	98.4	99.4	98.9	99	94.5	96.9
700%	100	100	100	100	100	100	100	99.5	99.5	94.7	98.9	99.8	97.5	99.4	95.0	92.4	98.1	90.3	96.6	94
600%	100	100	100	100	100	100	100	95.0	98.7	94.3	97.5	99.6	99.1	99.1	99.6	98.9	92.8	98.2	96.4	94.9
500%	99.5	99.6	96.9	99.5	96.3	94.7	98.9	92.1	94.8	92.7	90.9	97.6	95.9	99.6	99.8	93.5	98.9	98.7	96.1	98.7
400%	100	97.8	92.9	99.4	96.7	95.1	96.0	92.8	93.7	91.0	89.2	93.9	99.2	98.6	96.8	93.7	91.3	93.7	92.7	90.2
300%	100	97.1	97.4	94.7	92.3	93.6	100	94.7	88.2	86.8	87.4	91.5	99.6	97.8	88.2	88.0	77.2	97.1	97.5	98.9
200%	93.5	93.8	90.5	100	98.4	90.3	91.9	80.7	76.1	90.1	100	84.8	97.8	98.9	76.5	83.4	94.8	96.4	98.4	97.3
100%	96.5	88.9	83.1	79.4	60.1	89.8	87.3	60.4	55.1	52.7	63.9	70.5	84.7	62.0	52.0	79.4	67.5	60.0	90.0	77.9

Fig. 5. Training F1 scores of SVM models. The scores are arranged by over-sampling/under-sampling rates used to generate the training data points. Models with a tendency to over-fit and under-fit are highlighted inside dashed contours. Models trained with the same number of mineralized and non-mineralized samples are highlighted in purple.

The models with the lowest training and testing F1 scores are those trained with low mineralized over-sampling rates and high non-mineralized under-sampling rates (bottom right of Figs. 5 and 6). The models with the highest F1 scores are those trained with high mineralized over-sampling rates and low non-mineralized under-sampling rates (top left of Figs. 5 and 6). The SVM model trained with 2000% over-sampling rate of mineralized class (600 deposits, of which 570 were generated synthetically) and 5% under-sampling rate of non-mineralized class (30 non-deposits) achieved the highest F1 score (99.4%) against testing. Although most models trained with over-sampling rates above 500% (150 mineralized samples) and under-sampling rates below 40% (240 non-mineralized samples) present F1 scores equal to 100 on the training set, the testing F1 scores for these models vary widely from 93 to 99. This behavior indicates that models trained with imbalanced training datasets, which have high mineralized to non-mineralized ratios, tend to over-fit. In contrast, models trained with over-sampling rates below 500% (150 mineralized samples) exhibit the lowest training and testing F1 scores. These models have some testing F1 scores below 50, indicating that models trained with imbalanced training datasets, which have low mineralized to non-mineralized ratios, tend to under-fit. The SVM models trained with balanced training datasets, i.e., trained with the same number of mineralized and non-mineralized samples (highlighted diagonal of Figs. 5 and 6), exhibit F1 scores varying from 85.1 to 96.6. The F1 scores for these models consistently increase as the over-sampling rate of mineralized class increases and as the under-sampling rate of non-mineralized class decreases; that is, as the total number of training samples increases (from bottom left to top right of Figs. 5 and 6). This behavior suggests that models trained with balanced training datasets tend to be more stable, being less prone to over-fitting or under-fitting.

The variations in F1 scores caused by changes in sampling of the training data behaved logically. Models trained with a much larger number of mineralized samples compared to the number of non-mineralized samples will tend to learn only patterns associated with the former samples used for training. Therefore, such models struggle to generalize and classify new locations as non-mineralized. In contrast, models trained with a much larger number of non-mineralized samples compared to the number of mineralized samples will tend to learn only patterns associated with former samples used for training, making it difficult for such models to identify new locations as mineralized. However, models trained with balanced training datasets do not suffer from these problems generated by the asymmetry in the number of samples of each class, tending to generalize better without reducing performance.



	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
2000%	99.4	99.3	98.5	98.6	98.7	99	97.5	97.5	97.6	96.6	97.1	97.3	96.4	96.8	97.6	96.4	96.3	96.3	96.9	96.6
1900%	99.3	99.1	98.7	98.6	98.6	98.3	97.5	97.8	97.4	96.3	97.4	97.7	95.9	97.9	96.9	96.9	95.9	95.7	96.3	96.2
1800%	99.5	99.1	98.7	98.5	98.2	97.9	97.3	97.2	97.5	96.6	96.7	97.9	96.8	97.2	97.7	96.2	96.4	96.2	95.6	95.2
1700%	99.3	99.1	98.6	98.6	98.8	98.2	97.4	97.5	97.4	97.3	97	97.4	96.2	97.1	96.8	96.5	95.9	95.6	96.2	95.2
1600%	99	98.8	98.2	98.6	98.6	97.5	97.4	96.4	96.6	95.0	95.6	95.5	94.8	95.6	96.6	95.5	94.1	94.6	95.6	96.6
1500%	99.3	99.1	98.4	98.5	98.5	98.2	97.3	96.8	95.5	94.7	96.2	96.5	95.1	95.8	95.8	95.5	94.6	95.3	95.5	95.3
1400%	99.1	98.6	96.9	98.3	97.5	97.8	96.8	96.9	96.9	95	95.8	96.6	94.4	95.3	96.6	95.8	95.7	94.8	94.1	93.5
1300%	99.2	98.5	97.8	98.3	97.1	96.7	96.3	96.3	95.3	95.3	95.7	96	94.4	96.0	95	94.0	93.6	94.5	94.9	93.6
1200%	99.1	98.7	97.6	98.4	97.9	97.8	96.2	96.6	95.4	93.5	94.8	96.3	95.0	94.5	96.2	93.8	94.0	91.9	94.4	93.7
1100%	98.8	98.4	97.2	97.2	96.8	97.6	95.9	95.6	95.1	95.2	95.7	94.5	94	93.3	94.9	94.4	91.9	93.5	93.8	92.0
1000%	98.9	98.0	96.7	96.7	97.1	96.9	96.7	94.2	93.0	93.3	95.2	95.4	93.6	92.0	96.5	93	92.9	93.3	93.3	93.5
900%	98.8	97.8	96.3	97.1	95.6	97.5	93.7	93.8	93.2	93.5	94.1	94.6	92.3	92.1	92.9	93.1	91.8	91.6	92.5	91.3
800%	98.9	98.6	95.7	96.2	95.6	94.2	94.1	94.7	92.2	92.2	92.9	92.9	91.5	92.3	90.2	90.1	90.3	90.6	90.7	90.4
700%	98.7	97.5	95.9	96.3	96.3	95	92.8	94.1	93.1	91.1	92.9	92.6	92.3	91.1	88.5	90.2	89.4	88.8	90.5	90.3
600%	97.1	97.9	94.2	93.1	94.7	93.5	91.7	92.9	88.3	87.9	91.5	91.6	89.3	88.5	88.1	90.1	86.5	87.6	89.7	89.4
500%	97.8	96	92	93.7	91.1	93.4	89.5	91.3	91.0	88.7	88.7	87.3	85.8	86.9	86.3	86.3	87.7	87.8	85.0	84.2
400%	96.2	95.1	89.9	94.0	90.9	90.1	86.4	84.7	85.1	83.8	82.8	86.2	88.2	82.2	82.5	80.8	83.1	81.2	82.8	83.0
300%	95.5	93.6	84.7	86.9	86.6	89.2	80.2	84.2	80.7	80.3	79.5	83.8	82.3	79.4	77.1	76.2	73.3	80	77.6	73.9
200%	90.8	92.0	83.2	85.3	78.7	82.2	72.6	73.7	69.1	67.8	74.2	73.9	74.0	71	66.9	63.6	64	65.1	70.1	55.7
100%	85.1	80.8	69.2	64.1	49.1	63.7	34.9	46.4	36.6	43.2	39.2	40.1	38.1	39.2	37.2	41.1	31.6	45.4	48.1	38.6

Fig. 6 Testing F1 scores of SVM models. The scores are arranged by over-sampling/under-sampling rates used to generate the training data points. Models with a tendency to over-fit and under-fit are highlighted inside dashed contours. Models trained with the same number of mineralized and non-mineralized samples are highlighted in purple.

Although the F1 score assessments of the SVM models provide some indication of their classification performances, they do not provide information about the spatial context of the prospectivity models. In order to examine the spatial efficiency of classification by the SVM models, the success-rate curve for each prospectivity model was calculated by plotting the cumulative prospective (i.e., predicted mineralized) area against the cumulative number of training mineralized locations delineated in predicted mineralized zones, as described in Agterberg and Bonham-Carter (2005).

The change in success-rate curves due to changes in over-sampling/under-sampling rates are shown in Figure 7. The wide differences among success-rate curves obtained illustrate the sensitivity of prospectivity modeling to the number of samples and distribution of classes in the training set. Models trained with variable mineralized over-sampling rates and a fixed non-mineralized under-sampling rate of 5% (i.e., trained with only 30 non-deposits samples), exhibit similar success-rate curves, keeping the spatial efficiency of the models practically unchanged even when trained with a larger number of mineralized samples (Fig. 7a). Most of these models obtained the maximum F1 score of 100 on the training set, and testing F1 scores above 99. However, the success-rate curves show that they are less efficient than the models trained with a larger number of non-mineralized samples, reinforcing the over-fitting of these models discussed above.

In contrast, models trained with fixed mineralized over-sampling rates of 100% (i.e., without over-sampling mineralized samples) clearly show a variation in the success-rate curves (Fig. 7b). Models trained with a higher number of non-mineralized samples (under-sampling rates near 100%) exhibit better spatial efficiency, correctly classifying 80% of training mineralized locations in less than 4% of the study area. However, models trained with a small number of non-mineralized samples (under-sampling rates near 5%) were less efficient, needing approximately 17% of the study area to correctly classify 80% of the mineralized training locations. These models, trained without over-sampling the mineralized class, were the ones that obtained the lowest F1 scores on training and testing sets, with most of the models having F1 scores below 50. For this reason, although the models trained with a larger number of non-mineralized samples have better spatial efficiency, the F1 scores are very low, indicating under-fitting as discussed above.

Success-rate curves for the prospectivity models derived by using nearly balanced training datasets (Fig. 7c) show that models trained with a larger number of mineralized and non-mineralized samples are more spatially efficient than models trained with a smaller number of mineralized samples. As seen in Figure 6, the testing F1 scores of models derived by using nearly balanced training datasets also increase with increasing number of mineralized samples. This means that models trained with nearly the same number of mineralized and non-mineralized samples increase their performance, both in terms of classification and the spatial efficiency of the classification, as the number of synthetically generated mineralized samples in the training dataset increases.

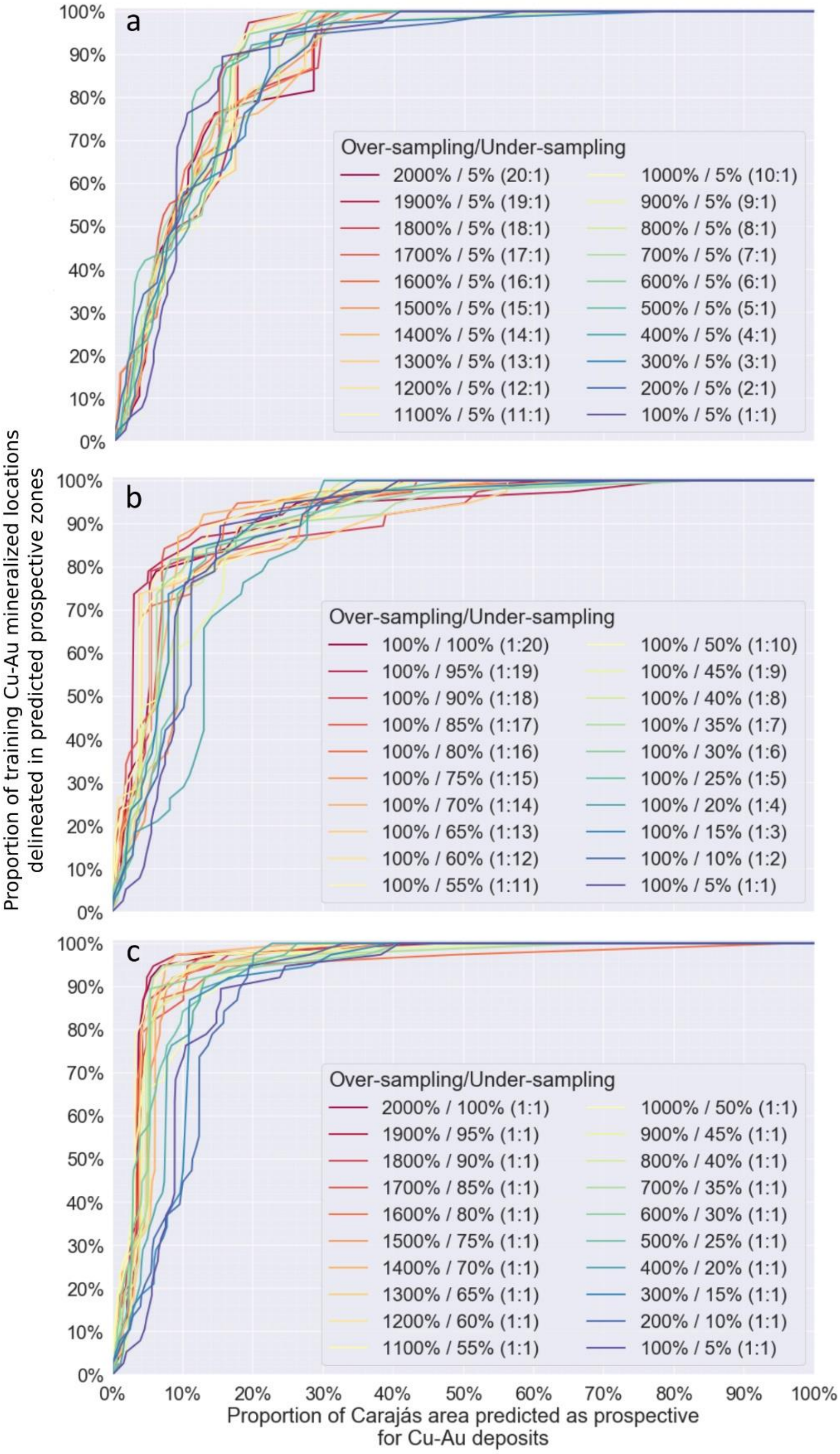


Fig. 7. Success-rate curves of SVM-based prospectivity models derived using training data (a) with fixed under-sampling rate of 5% for non-mineralized samples and variable over-sampling rates for mineralized samples ranging from 100-2000%, (b) with fixed over-sampling rate of 100% for mineralized samples and variable under-sampling rates for non-mineralized samples ranging from 5%-100%, and (c) with

same number of mineralized and non-mineralized samples. The rate between mineralized and non-mineralized locations for each model is shown in the legend in brackets.

Based on the success-rate curves, the best prospectivity map is the one obtained from the model trained with 2000% over-sampling of the mineralized class (600 mineralized samples, of which 570 were synthetically generated by SMOTE) and 100% under-sampling of the non-mineralized class (600 non-mineralized samples). This prospectivity model is not the one with the highest F1 score, showing F1 scores of 99.7 and 96.6 against the training and testing sets, respectively. However, the highest spatial efficiency obtained by this model together with its high F1 score against the testing set (the highest F1 score among the models derived using balanced training datasets), indicate that this is the model with the best performance. The use of a balanced training dataset avoids model over-fitting, which occurs with models trained with non-mineralized samples under-sampling rates near 5%. Such condition also increases the performance of the model for identifying true negatives (non-mineralized areas), significantly reducing the prospective area mapped by the model to classify the training set correctly.

The model trained without mineralized class over-sampling (100% sampling; 30 samples) and without non-mineralized class under-sampling (100% sampling; 600 samples), that is, without using SMOTE, is one of the models that obtained the worst F1 scores (i.e., 77.9 and 38.6 against the training and testing sets, respectively). The model trained without over-sampling of the mineralized class (100% sampling, 30 samples) and with 5% under-sampling of the non-mineralized class (30 samples) is also one of the models that obtained the worst F1 scores (i.e., 96.5 and 85.1 against the training and testing sets, respectively). These results show that instead of training models for MPM without over-sampling the mineralized class (i.e., without using SMOTE), it is better to produce a balanced training dataset (i.e., to use the same or nearly the same number of samples for the mineralized and non-mineralized class). It is clear from the results that using SMOTE to over-sample the mineralized class significantly increases the F1 scores of the models.

Again, it is important to emphasize that the 570 samples of the mineralized class generated synthetically for the training of the best performance model do not represent geographical locations, but vectors of the mineralized class. Therefore, as

600 mineralized locations within an area being explored are geologically improbable, the synthetic samples exist only in the vector space of the model and cannot be interpreted in this way. Synthetic vectors represent only a mathematical tool to improve the performance of the model.

The success-rate curves measure the spatial efficiency of the prospectivity model to identify locations of known deposits; i.e., the ones used for training. To quantify the spatial efficiency of a model for finding undiscovered prospects, the mineralized samples reserved for testing were used as independent data to create a prediction-rate curve for every model. The prediction-rate curve is computed similarly as the success-rate curve, as described in Agterberg and Bonham-Carter (2005). The difference is that the prediction-rate curve computes the cumulative prospective area against the cumulative number of testing mineralized samples delineated in predicted prospective zones, instead of training mineralized samples.

The success- and prediction-rate curves for the model with the best performance (based on F1 scores) are shown in Figure 8a. The curves illustrate that the model trained with 2000%/100% over-sampling/under-sampling rates, respectively, is the best predictive model for Cu-Au prospectivity in the CMP because with only 7% of the study area being prospective, it correctly classifies 100% (30) of the training mineralized locations and almost 80% (6) of the testing mineralized locations. The prospectivity map derived from this model is shown in Figure 9. This map can be used to guide further exploration of Cu-Au prospects in the CMP. According to the interpretation of the prospectivity map, six zones in the study area should be prioritized for Cu-Au exploration (polygons with dashed outlines in Figure 9). Although these zones are near known deposits and exhibit high likelihood values in the prospectivity map, no Cu-Au mineralization have been found inside them so far.

The modeled prospective map can be an essential tool for the exploration of Cu-Au prospects; However, it is important to understand the uncertainties associated with it. There are uncertainties associated with training data, such as those related to the detection limits of geophysical instruments and data processing procedures, as well as qualitative uncertainties related to expert interpretation and sampling bias that occur in geological mapping. Another source of uncertainty is related to training data labels, especially those associated with non-mineralized sites.

Besides, there are uncertainties related to the proposed mineral system, such as how well the targeting criteria reflect the targeting elements; how well the targeting elements reflect the constituent processes; and whether these processes are critical to the genesis of the mineral deposit.

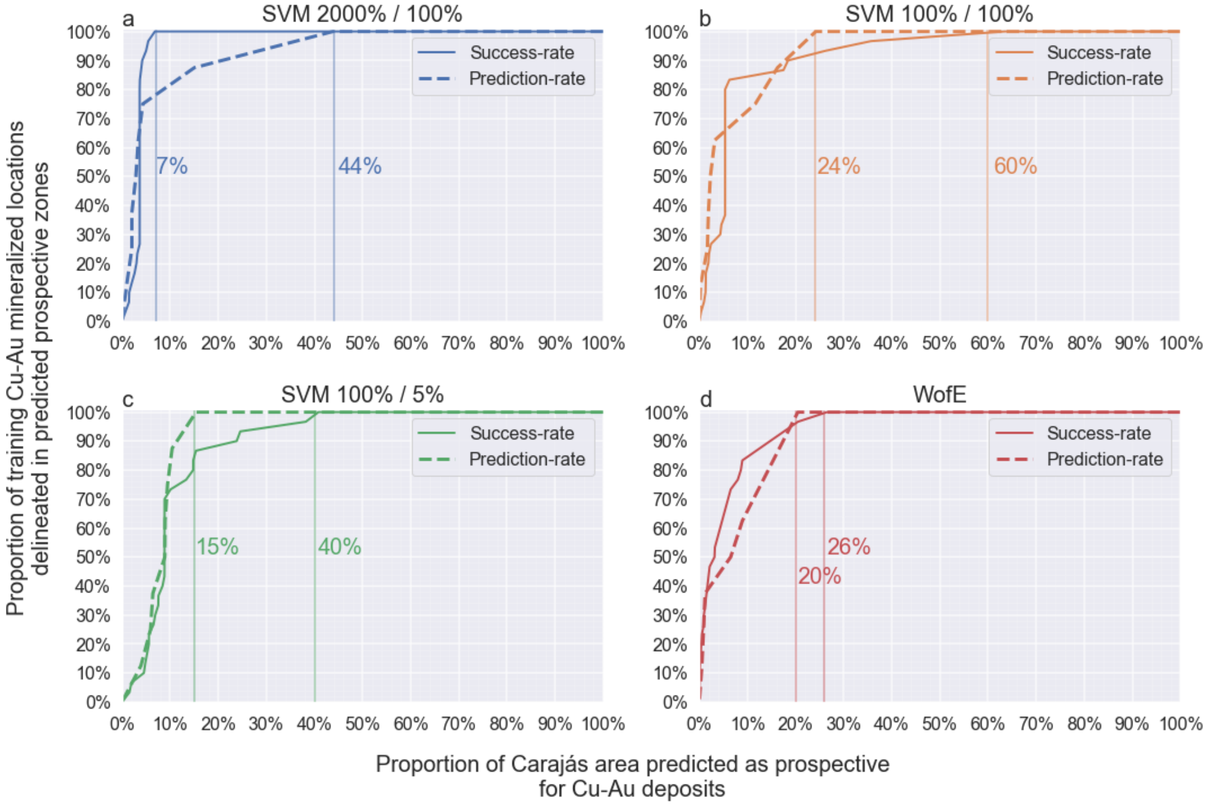


Fig. 8. Success-rate and prediction-rate curves of CMP Cu-Au prospectivity models. (a) SVM model trained with 2000% over-sampling rate of mineralized class and 100% under-sampling rate of non-mineralized class (balanced class distribution). (b) SVM model trained without over-sampling of mineralized class (100% over-sampling rate) and without under-sampling non-mineralized class (100% under-sampling rate). (c) SVM model trained without over-sampling of mineralized class (100% over-sampling rate) and 5% under-sampling rate of non-mineralized class (balanced class distribution). (d) weight of evidence (WofE) model trained with the same input features of the SVM models and the original training mineralized locations. Vertical lines highlight which proportion of Carajás area predicted as prospective for Cu-Au deposits covers 100% of training/test (success-rate/prediction-rate) Cu-Au mineralized locations.

To evaluate further the added value of using SMOTE, the success- and prediction-rate curves of the prospectivity models trained without using SMOTE are shown in Fig. 8, as well as for a weight of evidence (WofE) model (Bonham-Carter, 1994) trained with the same input features using in SVM modeling and the original mineralized locations. The SVM model without SMOTE (i.e., trained without over-sampling the mineralized class and without under-sampling the non-mineralized class (100%/100% over/under-sampling rate; Fig. 8b) needed 60% of the study area as prospective to correctly classify 100% of the training mineralized locations, and about 25% of the study area as prospective to correctly classify 36 mineralized locations (the number of mineralized locations correctly classified as prospective using 7% of the study area by the model trained with 2000%/100% over/under-sampling rate). The SVM model trained with a balanced training data and without using SMOTE, represented by the model trained with 100%/5% over/under-sampling rate (Fig. 8c), correctly classified 100% of the training mineralized locations within 40% of the study area, and needed about 22% of the study area to correctly classify 36 mineralized locations. The WofE model (Fig. 8d; Supplementary Material) outperformed the SVM models trained without SMOTE, correctly classifying 100% of the training mineralized locations with 26% of the study area, and 36 mineralized locations with about 20% of the study area. These results show that not using SMOTE to balance the training data for MPM significantly reduces the spatial efficiency of SVM modeling for the classification of training mineralized locations, as can be deduced from Fig. 8. However, the prediction-rate curves show that the model with the best spatial efficiency for the classification of testing mineralized locations is the SVM model trained with 100%/5% over/under-sampling rate, that is, without SMOTE and with a balanced class distribution, classifying 100% of the mineralized test sites within 15% of the study area (Fig. 8c). Despite this, the model with the best spatial efficiency to classify both training and testing mineralized location is the SVM model trained with 2000%/100% over/under-sampling rate, because it is the model that correctly classified the largest number of mineralized locations in only 7% of the study area.

To qualitatively determine the most influential features to the prospective potential of the CMP, the variance-based sensitivity analysis approach introduced by Sobol (2001) was carried on the best performing SVM model (trained with 2000%/100% over/under-sampling rate). The Sobol sensitivity analysis was performed using the SALib package (Herman and Usher, 2017). The method aims to quantify the relative influence of input features to the model output. The influence of each input is given by the sensitivity indices. The total-order sensitivity indices measure the contribution to the output variance caused by a model input, including both the effects of each input individually and the effects of interactions between the inputs. The second-order sensitivity indices measure the contribution to the output variance caused by the interaction of two model inputs.

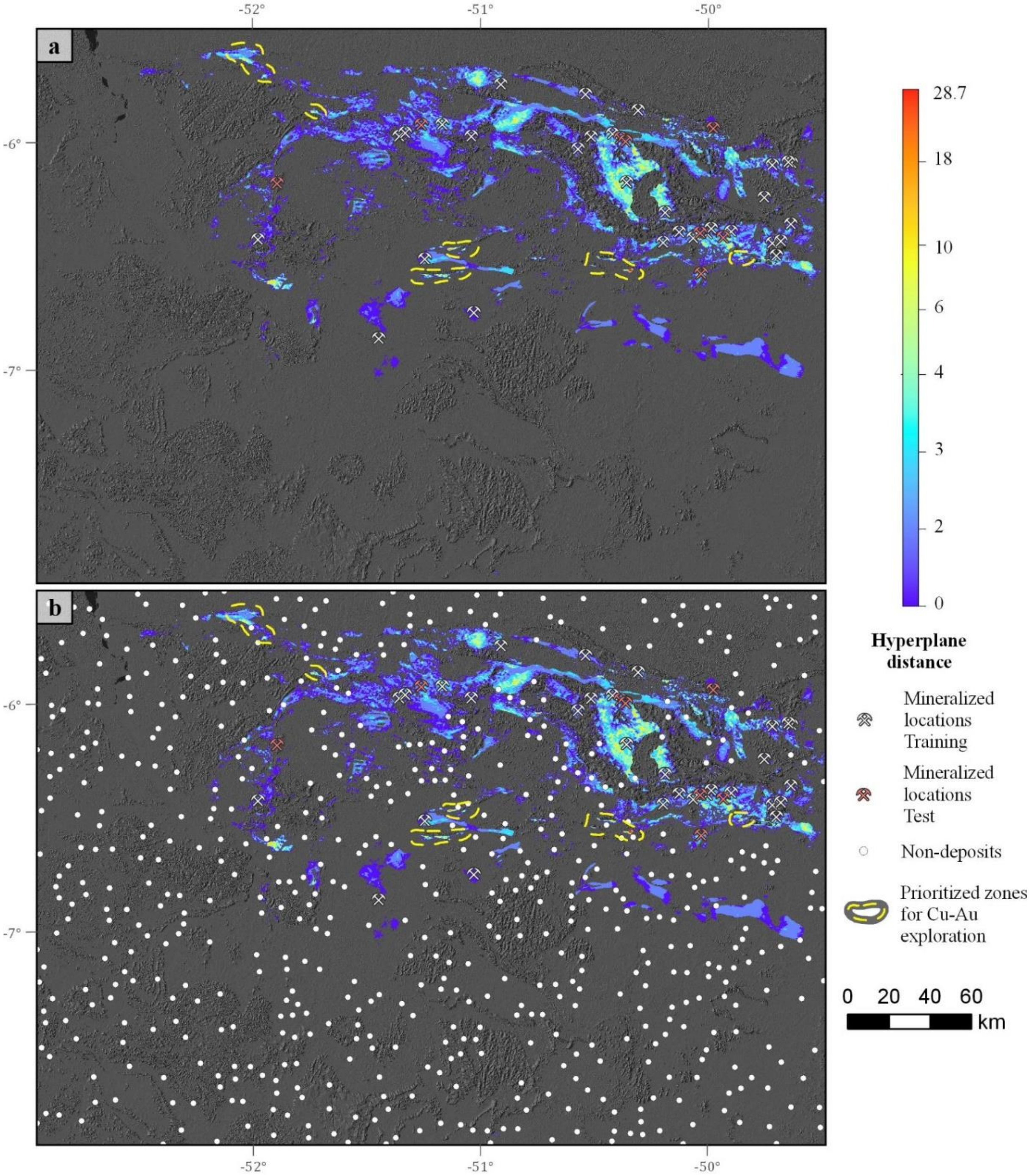

a
b
-52°
-51°
-50°
-6°
-7°
28.7
18
10
6
4
3
2
0
Hyperplane distance
Mineralized locations Training
Mineralized locations Test
Non-deposits
Prioritized zones for Cu-Au exploration
0 20 40 60
km

Fig. 9. (a) Predictive map for Cu-Au prospectivity obtained by a SVM model trained with 2000% over-sampling rate and 100% under-sampling rate of mineralized and non-mineralized samples, respectively. The map is colored according to the hyperplane distance (distance from the feature vector to the surface computed by the SVM model that defines the boundaries between classes), which is directly proportional to the predicted likelihood of finding a Cu-Au deposit. Dashed polygons in red highlight zones that could be prioritized for Cu-Au exploration according to the interpretation of the prospectivity map. (b) The same predictive map with the target variables used to train and test the model plotted.

The Fig. 10 shows the total and second-order sensitivity indices for the ten most critical input features for the SVM model according to the total-order sensitivity indices. The most relevant input is the magnetic worms, followed in order of relevance by the Xingu Complex, the fractal dimension, the proximity to contacts between Mesoarchean, Neoarchean and Paleoproterozoic, and the gamma-ray eU (Fig. 10a). Therefore, these inputs are the main targeting criteria for the SVM model. The second-order sensitivity indices indicate that the most relevant input pair is gravimetric worms and fractal dimension, followed by pairs Parauapebas Formation and gamma-ray eU, and fractal dimension and Xingu Complex (Fig. 10b). These results suggest that the relationship between deep structures outlined by the gravimetric/magnetic worms and their connectivity to shallow structures mapped with the fractal dimension is the strongest control on the modelled prospectivity of Cu-Au mineralization. Also, the second-order sensitivity indices indicate that the uranium enrichment outlined by the gamma-ray eU is more relevant for modeling the prospectivity of mineralization related with the Parauapebas Formation, and that the fractal dimension, which outline the connectivity of the structures, is more relevant for modelling the prospectivity of mineralization related with the Xingu Complex.

The relationship between the mappable targeting criteria, according to the sensitivity analysis, with their respective targeting elements and constituent processes proposed in section 2.2 (Table 1), suggests that the main processes related to the Cu-Au mineral system at CMP are the migration of metal-enriched fluids in trans-crustal and/or craton-scale structural zones located between Meso/Neoarchean-Paleoproterozoic contacts. These structures are outlined by the magnetic/gravimetric worms and by the proximity to contacts between Mesoarchean, Neoarchean and Paleoproterozoic, respectively. This relationship also suggests that the deposition of the metals is commonly associated with brecciated zones and uranium enriched alteration zones, outlined by the fractal dimension map and by gamma-ray eU, respectively. The uranium enriched zones are more relevant for Cu-Au mineralization related to the Parauapebas Formation, and the structural controls are more critical for mineralization related to other lithostratigraphic units.

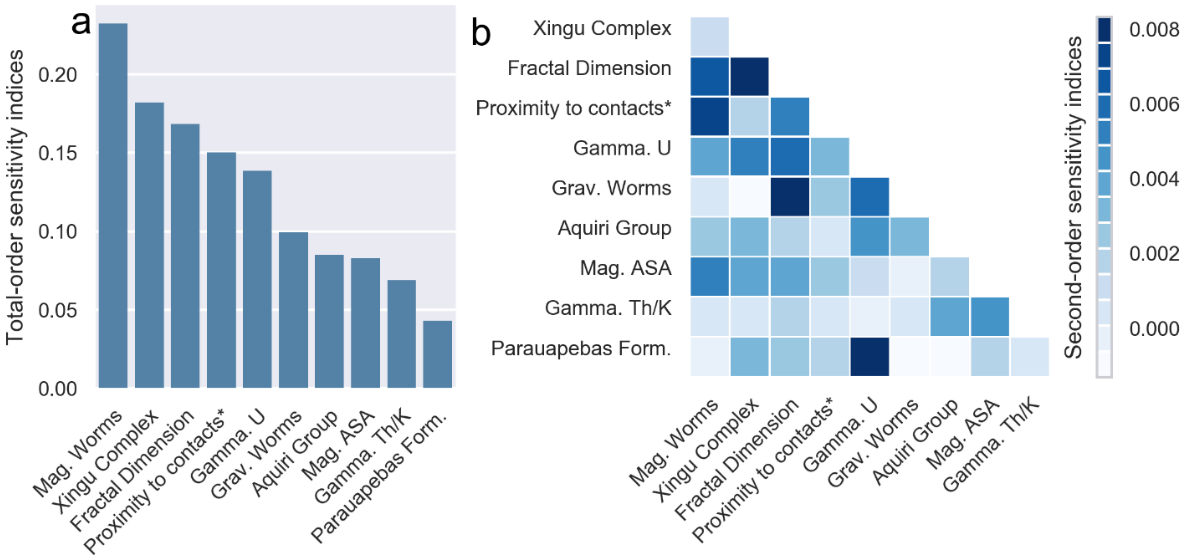


Fig. 10. (a) Total-order sensitivity indices and (b) second-order sensitivity indices for the ten most critical input features for the SVM model trained with 2000% over-sampling rate and 100% under-sampling rate of mineralized and non-mineralized samples, respectively. *Proximity to contacts between Mesoarchean and Proterozoic stratigraphic units, Neoarchean and Proterozoic stratigraphic units, as well as contacts between Paleoproterozoic and Proterozoic stratigraphic units.

5.6 Conclusion

Over the years, ML has become a primary tool for data-driven mineral prospectivity mapping (MPM). One of the major challenges in the use of ML methods is the need for large volumes of labeled data (i.e., locations of known deposit occurrences), which are difficult and expensive to obtain in most cases, as well as in mineral exploration problems where commonly few deposits are known. The small number of deposit locations and the large number of non-deposit locations inherent in the nature of mineralized regions is an obstacle to the efficient performance of a ML method in data-driven MPM. That is, the commonly used ML methods have difficulty in learning when trained with imbalanced training datasets. However, as shown in this study, the synthetic generation of mineralized samples using the SMOTE algorithm can significantly increase the classification performance and the spatial efficiency of data-driven MPM through ML. However, care should be taken when increasing the number of mineralized samples and without changing the number of non-mineralized samples. According to the results, prospectivity models derived using balanced training datasets are more stable on training and have better spatial efficiency. Therefore, the performance of prospectivity models derived by a ML method can be increased using

SMOTE only if the model is trained with the same or nearly the same number of mineralized and non-mineralized samples. Nevertheless, this result needs further verification by testing this approach using other ML methods, and in other regions with different numbers of known deposits. Also, other sampling techniques need to be trialed and compared with SMOTE.

The advantage of using spatial sampling techniques to deal with the imbalanced training data problem in MPM is that they are easy to implement and does not require in-depth knowledge of ML architectures, which is necessary when implementing algorithm level solutions. The use of these techniques can lower the risk in mineral exploration and should allow exploration geologists to make spatially-informed decisions.

5.7 Acknowledgments

The authors would like to thank the Geological Survey of Brazil (GSB/CPRM) for free access to geological and geophysical datasets. The Geological Survey of Finland (GTK) is acknowledged for providing access to the new ArcSDM toolbox. We are grateful to OGR editor and reviewers for providing helpful and detailed comments that significantly improved the final version of the manuscript. This research was supported by the University of Campinas (UNICAMP), the Geological Survey of Brazil and the National Council for Scientific and Technological Development (CNPq) through research project Proc. Nr. 401316/2014-9. C.R.S.F. (Proc. Nr. 309712/2017-3) and E.J.C. (Proc. Nr. 401316/2014-9) acknowledge CNPq for research grants.

6 ORE GRADE ESTIMATION FROM HYPERSPECTRAL DATA USING CONVOLUTIONAL NEURAL NETWORKS: A CASE STUDY AT THE OLYMPIC DAM IOCG DEPOSIT, AUSTRALIA

6.1 Introduction

The acquisition of information on the spatial distribution of chemical properties of subsurface rocks is required during exploration and mining of mineral resources. The concentration of elements in the subsurface outlines the volume with commercially viable ore, being essential to verify the economic viability of a deposit, as well as to design ore mining and to guide mine operation. To acquire this information, drilling machines are used to collect subsurface rock samples, known as drill-cores. Commonly, a few equally spaced small sections along the drill-cores are selected and sampled for geochemical analysis. X-ray fluorescence, inductively coupled plasma mass spectrometry (ICP-MS) and inductively coupled plasma optical emission spectrometry (ICP-OES) are the analytical techniques typically employed in most cases. This strategy has been used widely by mining companies over the past decades; however, in most cases the high analytical cost involved does not allow for adequate sampling (large sampling intervals; ⌧ 1 m) of the drill-cores leading to low spatial resolution data. Usually, the geochemical analysis needs to be combined with drill-core descriptions (lithological boundaries and units) performed by geologists and other analytical techniques to increase the precision of the modeled ore body.

The spatial distribution of minerals in the subsurface is also an important information to be acquired during exploration and mining operations. Recognition and spatialization of minerals in drill-cores aid the delineation of an ore body and the identification of ore accumulations. Optical microscopy and X-ray diffraction are examples of techniques typically employed to obtain information about the ore mineralogy of drill-cores. In addition to these techniques, the use of hyperspectral imaging systems to collect mineralogy information of drill-cores has grown in recent years (Acosta et al., 2019; Calvin and Pace, 2016; Kruse, 1996; Taylor, 2000). Hyperspectral sensors are noninvasive and nondestructive tools capable of scanning drill-cores in a short time. These sensors acquire images in hundreds of continuous spectral bands, capturing detailed spectral information for each image pixel. Certain minerals absorb electromagnetic radiation at specific wavelengths as a result of the fundamental electronic and vibrational processes of their molecular bonds (Clark, 1999; Van der Meer, 2004). Therefore, the analysis of spectral absorption features can be used for mineral identification, allowing hyperspectral sensors to collect detailed information (centimetric to millimetric spatial resolution) on the spatial distribution of different minerals, assisting the identification of ore zones and associated gangue.

Several studies have shown that concentrations of certain elements are significantly correlated to the spectral responses of certain rocks and minerals (Clark and Roush, 1984; T. J. Cudahy et al., 2009; Dalm et al., 2017, 2014; Ducart et al., 2016; Haest and Cudahy, 2012; Prado et al., 2016; J. T. Qiu et al., 2017; Silversides and Murphy, 2017). Therefore, hyperspectral data can potentially be used to predict ore grade distribution across drill-holes at centimetric to millimetric scale. However, unlike conventional images with only few bands, hyperspectral images have hundreds of bands, resulting in high volume and high complexity data. For this reason, the analysis of hyperspectral data is rarely performed visually; rather, algorithms are developed to extract the desired information from the images, such as mineral abundance, composition, and crystallinity. Most of the algorithms used to correlate these spectral parameters with geochemical data are based on classical statistical methods such as linear and polynomial regression algorithms, which are not able to handle some of the complexities of hyperspectral data. Furthermore, to obtain accurate results, these methods require the constant interaction of specialists to interpret the spectra and fine-tune the acquisition of spectral parameters.

Given the ability of machine learning (ML) algorithms to extract complex patterns, which may be difficult for conventional statistical methods, these algorithms are able to circumvent the limitations of classical methods in hyperspectral data analysis. ML algorithms have yielded satisfactory results in the analysis of drill-core hyperspectral data in recent years (Acosta et al., 2019; Khodadadzadeh and Gloaguen, 2019; Koirala et al., 2019; Schneider et al., 2014; Tusa et al., 2019; Wang et al., 2016) due to their ability to learn complex relationships between the spectrum obtained in each pixel and the imaged material (Gewali et al., 2018). However, most of these studies have focused on mineral classification of spectral data rather than correlating them with geochemical data.

Recent advances in ML algorithms are driving profound changes in data analysis and integration in various aspects of the mineral industry. Recent technological advances have promoted a significant increase in the processing power of computers, made possible by the development of graphic processing unit (GPU) in recent times. GPUs enable the training of larger neural networks with more complex architectures, leading to the development of ML algorithms based on deep architectures, called deep learning (DL) algorithms. These algorithms overcome many of the limitations present in traditional ML techniques, which resulted in major success stories in a wide range of commercial applications, such as computer vision, speech recognition, and natural language translation (Lecun et al., 2015). Deep architectures can learn exclusively from data, removing the need for manual extraction of features in the input data. This ability makes DL algorithms suitable for solving problems with complex data structure, data acquired by different sensors, or big data. Usually, manual extraction of features in these problems is difficult and requires the knowledge of an expert to be performed adequately.

Although ML models are to overcome the inadequacy of classical statistical methods in the analysis of hyperspectral data, these algorithms depend on the preliminary extraction of features and are not able to analyze the spatial relations in the input data. The advances made by the DL architectures in the data analysis can contribute significantly to the improvement of the methods currently used. One of the recent DL algorithms that deserve prominence is known as convolutional neural networks (CNN; Lecun and Bengio, 1995). These networks can recognize patterns related to the spatial location of data, that is, considering that the training data are images (data with strong spatial correlation), unlike other architectures that analyze the pixels of an image individually, CNNs are able to identify patterns related to the neighboring pixels in an image (Castelluccio et al., 2015; Krizhevsky et al., 2012). Thus, CNNs incorporate in their analysis the spatial structure of hyperspectral data, besides being able to learn to extract automatically the best features for classification of data, even in complex databases. However, one of the major challenges in the training of deep architectures is the need for large volumes of labeled data, which are difficult and expensive to obtain in most cases, as in mineral exploration problems. Although some approaches based on ML techniques, recently, have been applied to correlate drill-core geochemical data with hyperspectral data (Acosta et al., 2020), as far as we are aware, DL algorithms have not been used to predict element concentration by means of hyperspectral data.

Based on the motivations described above and the need to develop an understanding of the capabilities of CNNs for the integration of geochemical and hyperspectral data, this work investigates the applicability of CNN architectures to predict copper grade in drill-cores from the Olympic Dam deposit (ODD) by means of hyperspectral data. Applicability, in this context, is constrained by the fact that DL architectures can learn exclusively from data, removing the need for manual extraction of features in the input data, and that CNN can recognize complex patterns, including those related to the spatial location of absorption features in the hyperspectral data.

6.2 Study area

The ODD in South Australia was chosen as a case study region. It is one of the largest Cu–U–Au–Ag deposits in the world (Kathy Ehrig et al., 2012), with a total resource (open pit + underground) of 11.680 Gt @ 0.70 % Cu, 0.31 g/t Au, 1.3 g/t Ag, 0.23 kg/tonne U_3O_8 (BHP Annual Report, 2022). The Australian government provides a large amount of information about the ODD deposit in a free access database, which contains lithological description, as well as geochemical and hyperspectral data of several boreholes (most of them intercepted the ore zone). Furthermore, the ODD has a well-known geology and mineralization style with a well-described hydrothermal alteration zonation.

1. Geology and Cu mineralization

The ODD is the largest deposit of the Olympic Cu–Au Province, located on the eastern margin of the Archean to Mesoproterozoic Gawler Craton. The mineralization is spatially and temporally associated with the formation of the bimodal Gawler Silicic Large Igneous Province (SLIP) at ~1.6 Ga (Allen et al., 2008). This province comprises felsic and mafic volcanic rocks assigned to the Gawler Range Volcanics and granitoid intrusions of the Hiltaba Suite. The deposit is located within a hematite-rich hydrothermal breccia complex, known as Olympic Dam Breccia Complex

(ODBC; Reeve, 1990), hosted by the Roxby Downs Granite (RDG), part of the Hiltaba Suite (Creaser, 1989). Felsic and mafic–ultramafic volcanic rocks assigned to the Gawler Range Volcanic and clastic sedimentary rocks are also described in the ODBC (Kathy Ehrig et al., 2012). Neoproterozoic to Cambrian sedimentary rocks of the Stuart Shelf overlies the Olympic Dam Breccia Complex and the Roxby Downs Granite (Dmitrijeva et al., 2019; Kathy Ehrig et al., 2012).

The RDG is a coarse to medium-grained, quartz-poor syenogranite with A-type affinities (Creaser, 1989; Reeve, 1990). It is composed mainly of alkali feldspar (~45%), quartz (~30%), and sodic plagioclase (~20%), with minor biotite and amphibole, and accessory to trace amounts of igneous magnetite, sphene, apatite, zircon, allanite, fluorite, ankerite, synchysite and uranothorite (Kathy Ehrig et al., 2012). The ODBC exhibit a well-defined hydrothermal alteration zonation as documented by Ehrig et al. (2012). The outer margins of the ODBC are defined by the complete replacement of igneous biotite from RDG to chlorite (biotite 'out' in Figs. 1 and 2). The brecciated zones inside the ODBC are divided into granite-rich breccias (5% Fe outline in Figs. 1 and 2) and hematite-rich breccias (20% Fe outline in Figs. 1 and 2). The rocks between the granite-rich breccias and the ODBC outer margins exhibit hematite alteration of igneous magnetite and sericite alteration of igneous plagioclase and orthoclase. The hematite and sericite alterations increase towards the geographic center of the deposit.

The granite-rich breccias consist of fractured and veined granites with the presence of hematite as matrix, clasts, or veins. Within these breccias, the igneous plagioclase is completely replaced by sericite and the sericite alteration of orthoclase intensifies (Fig. 2). Pyrite is the predominant sulfide with minor chalcopyrite. Siderite often occurs as fragments or veins. The contact between granite-rich breccias and hematite-rich breccias defined by the 20% Fe outline (Fig. 1) is marked by the increase in abundance of hematite, with hematite becoming the most abundant component in the breccia. The orthoclase is completely altered to sericite inside hematite-rich breccias, and sericite is altered to hematite toward deposit center. These breccias are the main host rocks of Cu mineralization, containing economic concentrations of Cu–(Fe)–sulphides as chalcopyrite, bornite, and chalcocite, with fluorite, barite and minor siderite typically associated. As the hematite alteration intensifies towards the deposit center the abundance of quartz, hematite and barite increases, and the abundance of aluminosilicate and sulfide minerals decreases, defining the hematite–quartz–barite hydrothermal assemblage. The Cu mineralization does not occur within this alteration assemblage.

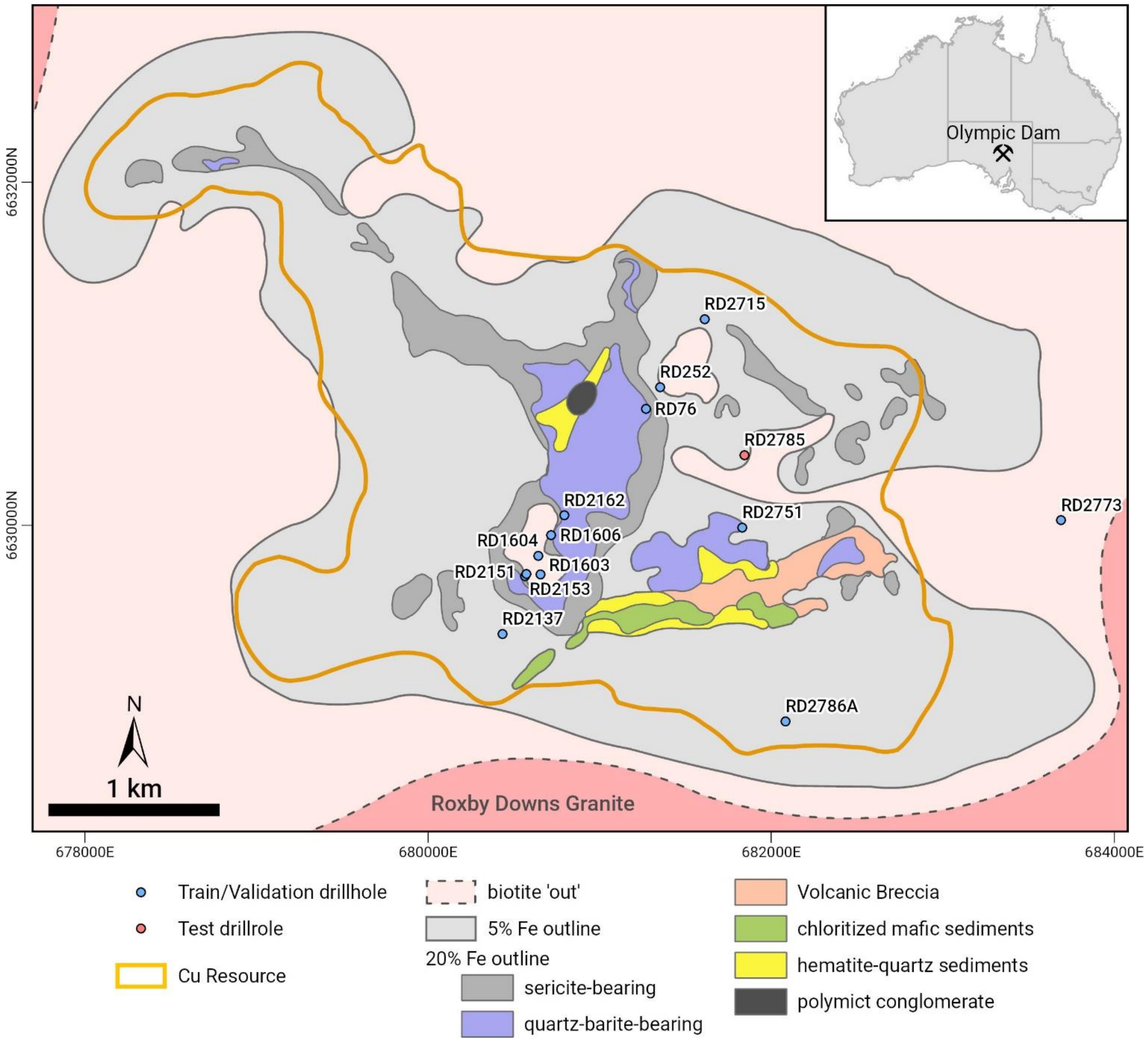


Fig. 1 Simplified geologic map of Olympic Dam deposit (modified from Dmitrijeva et al., 2019) projected at 350 m below surface showing collar locations of drillholes selected for training, validation and test of the CNN models. Coordinate system: WGS 1984 UTM Zone 53S.

Three hydrothermal alteration assemblages within the granite- and hematite-rich breccias were defined by Ehrig et al. (2012) according to the Fe and S oxidation states, namely reduced Fe oxide (magnetite + apatite + siderite + chlorite + quartz), oxidized Fe oxide (hematite + sericite + fluorite), and hematite–quartz–barite assemblages (Fig. 2). The main sulfide within the reduced Fe oxide assemblage is pyrite with less amounts of chalcopyrite. Chalcopyrite, bornite and chalcocite are the predominant sulfides within the oxidized Fe assemblage, with less amounts of pyrite.

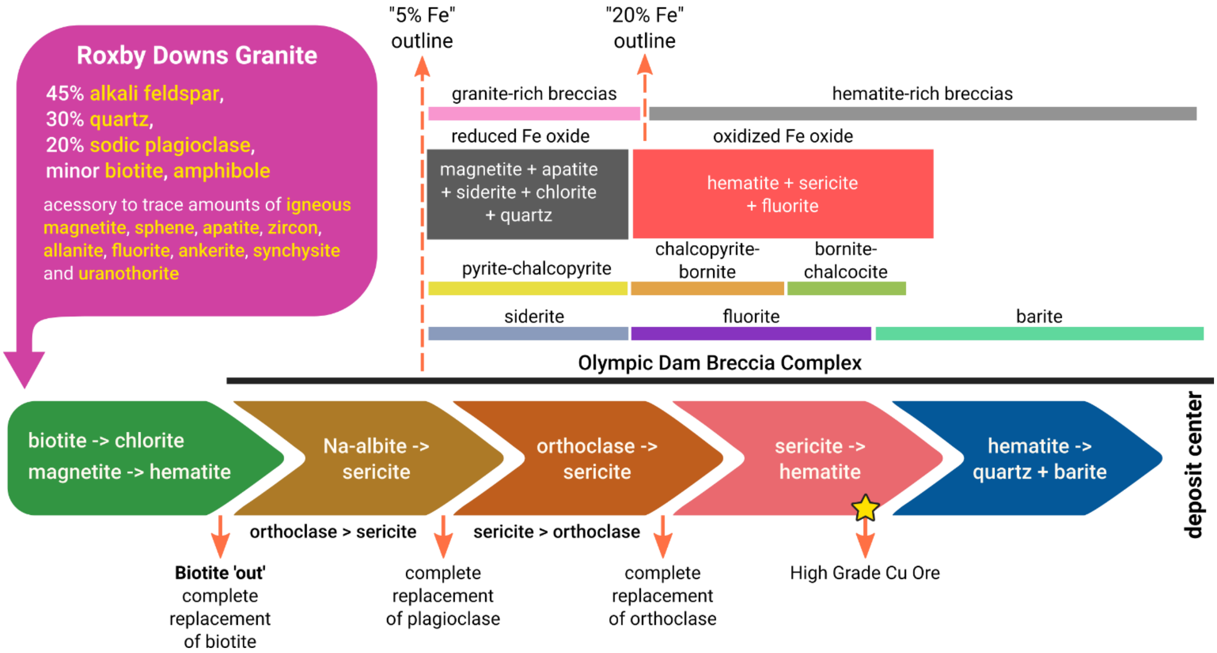


Figure 2 Schematic diagram of hydrothermal alteration and mineral zonation of the Olympic Dam deposit (as described in Ehrig et al., 2012).

6.3 Methods

1. Hyperspectral data and geochemical analysis

The datasets used to train the model include point spectra and geochemical data collected from 14 drill cores (with average length of 1275 m) from the ODD. The spectral data were provided by the AuScope National Virtual Core Library (NVCL), and the geochemical data by the South Australian Resources Information Gateway (SARIG). Both spectral and geochemical datasets can be downloaded free of charge at AuScope Discovery Portal (http://portal.auscope.org/).

The drill cores were profiled with the CSIRO Hylogger-3: a visible, near infrared (VNIR) to shortwave (SWIR) and thermal (TIR) infrared reflectance spectrometer system for drill core logging. The spectral data were collected continuously throughout the drill cores at a sampling interval of 8 mm. The field of view of the spectrometers was 10 mm across the track (across the drill hole extension) and 18 mm along the track (along the drill hole extension) (Schodlok et al., 2016). The VNIR–SWIR data were acquired for a wavelength range of 380–2500 nm (4 nm bandwidth; 531 bands) and the TIR data for a wavelength range of 6000–14500 nm (25 nm bandwidth; 341 bands).

The cores were also sampled to obtain geochemical elemental analyses for Ag, Al, As, Au, Ba, CO2, Ca, Ce, Co, Cu, Fe, K, La, Mg, Mn, Mo, Na, Ni, P, Pb, S, Si, Ti, U3O8, Zn, Zr, Cr, Sr, V, Y, Sb, and Sc. The concentrations of most of these elements were determined by inductively coupled plasma mass spectrometry (ICP-MS) and inductively coupled plasma optical emission spectrometry (ICP-OES) techniques, except for all the Au and CO2 analyses, and most of the U3O8 analysis, which were determined by fire assay, combustion, and X-ray fluorescence (XRF) techniques, respectively.

The sample spacing for geochemical analysis was variable, but most samples were obtained at a spacing of 1 m (~85% of the samples) and 5 m (~11% of the samples), totaling 8868 samples with geochemical analysis. The average reflectance spectrum was calculated for the spectra that encompass the interval where a sample was collected for geochemical analysis (typically 1–5m long; average of 200 spectra). The obtained reflectance spectra and respective Cu data were used to train deep neural networks regression models.

6.3.1 Dataset pre-processing

6.3.1.1 Dataset verification and correction

The hyperspectral data were provided as normalized reflectance spectra, consisting of 1,814,569 spectra. Some of these spectra (4,931; 0.27% of total) have anomalous reflectance values (i.e., > 1), and were discarded because it is expected that reflectance values fall in the range [0,1]. In addition, some spectra (30,882; 1.7% of total) have reflectance values lower than 0, which are also anomalous; negative values close to zero (<0.01) were replaced with 0 to keep the reflectance values in the [0,1] range. After this procedure, the average of the spectra, which encompass each sampled interval for geochemical analysis, was calculated, reducing the hyperspectral data to 8868 spectra, one for each Cu analysis.

6.3.1.2 Continuum removal

Normally, during spectral analysis, the continuum is removed (Clark and Roush, 1984) to isolate absorption features of the spectrum from other effects, such as level changes and slopes generated by other materials. These effects can cause shifts of the local minimum in the spectra and can significantly reduce absorption depths. When the continuum is removed, the minima show more stable positions. In addition, the effects of illumination geometry, as well as the effects of contaminants and grain size are reduced because the continuum removal normalizes the spectra (Clark et al., 2003).

The continuum removal algorithm from the pysptools python's package (Therien, 2018) was used. This algorithm computes the convex hull of the signal and removes it by hull quotient. Example results of continuum removal are shown in Fig. 3.

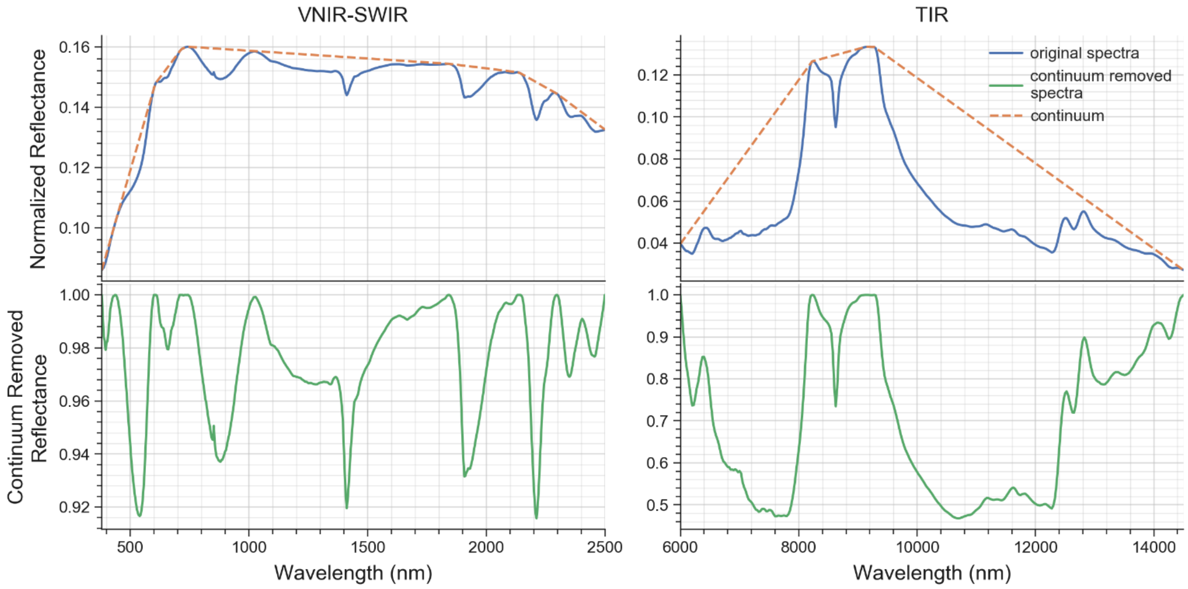


Fig. 3 Examples of continuum removal. Top panels: original spectra as solid lines with respective convex hull continuum as dashed lines. Bottom panels: continuum removed spectra. Left panels: VNIR–SWIR data. Right panels: TIR data.

6.3.1.3 Spectrogram

Spectrograms are typically used to represent signals, such as audio data, in 2D space (e.g., time–frequency for audio; wavelength–frequency for reflectance spectra) raster, where the magnitude of a signal with a given frequency at a given position (time for audio; wavelength for reflectance spectra) is represented by the value of a pixel. The spectrograms were computed using SciPy library (Virtanen et al., 2020) for the Python programming language. The algorithm decomposes the signal into overlapping segments of equal length, and then applies the short-time fast Fourier transformation to each segment (Griffin and Lim, 1984; Oppenheim, 1999). To generate the spectrograms, we used the Hann window (Blackman and Tukey, 1958), a segment length of 20, with 10 observations of overlap, and a sampling frequency of 0.25 (1/4) for the VNIR–SWIR spectra and 0.04 (1/25) for the TIR spectra. After computation, the spectrograms values were transformed to logarithmic scale in order to normalize the distribution of pixel values.

The spectral data were transformed to spectrograms, generating a raster representation of the spectrum, which is more appropriate for CNN models based on 2D convolutions to process. The original hyperspectral data samples consist in vectors of length 872 (VNIR–SWIR + TIR available bands), which were transformed to matrices of 11 x 85 (frequency x wavelength) after transformation. Examples of the original data and the resulting spectrograms are shown in Fig. 4. The drops in reflectance can be identified in the spectrogram as large (yellow) values.

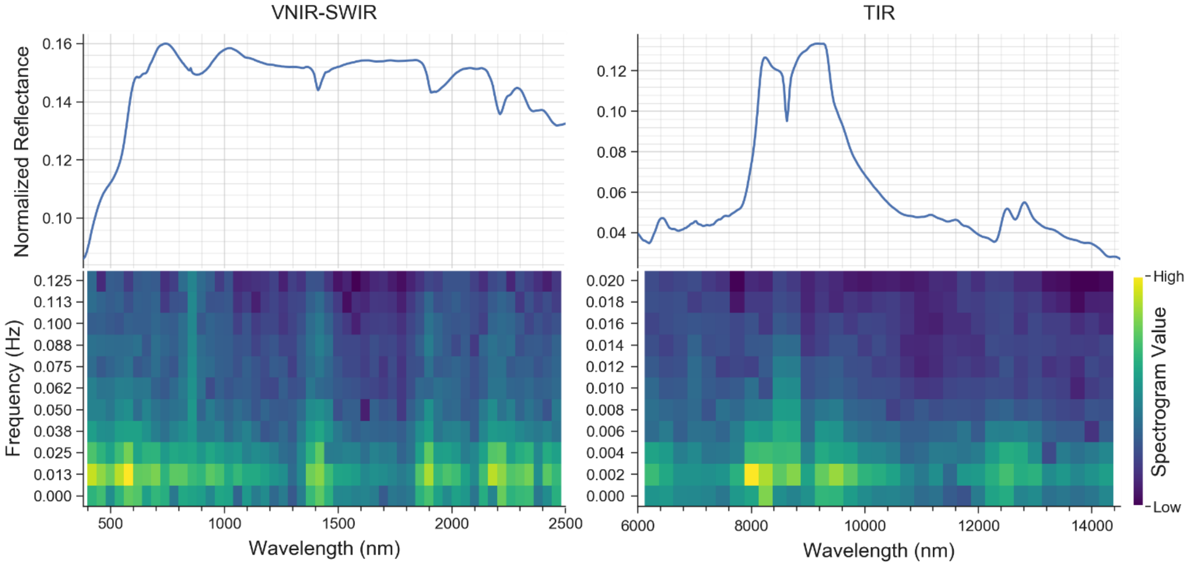


Fig. 4 Example of spectral data encoded as spectrograms. Top panels: original spectral data. Bottom panels: spectrograms with amplitude (color) in log scale. Left panels: VNIR–SWIR data. Right panels: TIR data.

6.3.1.4 Standardization

Before model training, the mean of each feature (reflectance value of each band for spectra data, or magnitude of each pixel for spectrogram data) was subtracted from the feature values and divided by the standard deviation of the feature, thus:

$$z = (x - \mu)/\sigma$$

where x is the feature value, μ and σ are the mean and standard deviation of the feature, respectively, in the training samples, and z is the standardized feature value (Grus, 2015; Pal and Sudeep, 2016). This procedure centralizes the data distribution at 0 mean and sets the standard deviation to 1. Standardization of input data is an important step to avoid instability in the training of neural networks models, as it scales the features into unit variance facilitating the updating of weights during gradient descent. Dataset standardization avoids the weighting of features with higher values, which update much faster than those with lower values and cause a model to learn incorrect patterns.

6.3.2 Dataset analysis

To better understand the relationship between the Cu data and the spectra data, the samples were grouped according to their Cu % so that the non-mineralized samples (statistically low values of Cu) were grouped together (bin 1) and the mineralized samples were subdivided into three groups, lower grade ore (bin 2), moderate grade ore (bin3), high grade ore (bin 4), resulting in the creation of four bins. The number of bins and the ranges per bin were defined by box plot statistics. The minimum and the median define the first bin (0.00–0.30% Cu), the median and the third quartile define the second bin (0.30–1.01% Cu), the third quartile and the upper whisker define the third bin (1.01–2.44% Cu), and finally the upper whisker and the maximum define the fourth bin (2.44–8.4% Cu). The upper whisker was calculated as 1.5*(Q3 - Q1), where Q3 and Q1 are the third and first quartiles, respectively. The number of samples per bin is shown in Fig. 5.

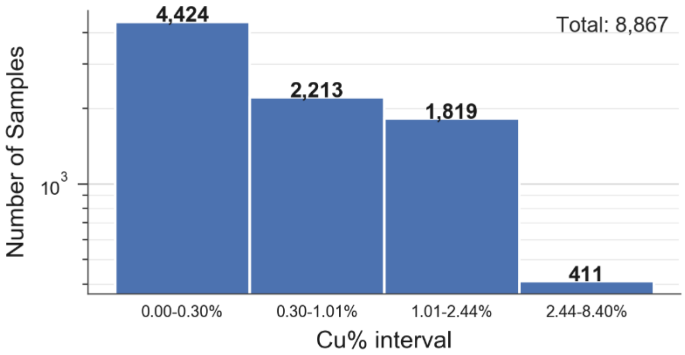


Figure 5 Number of samples in each Cu % bin.

The averages of spectra obtained per geochemical sample are shown in Figs. 6 and 7, grouped by their Cu % as defined above. The decrease in reflectance was the most prominent pattern associated with the increase in Cu %, in both VNIR–SWIR and TIR spectra (Fig. 6). The samples of the first bin (0.0–0.3% Cu) had VNIR–SWIR reflectance values of 0–0.86, and most samples had a reflectance value of 0.3. The reflectance values decreased as one moves to bins with higher Cu %. In bins 3 and 4, the maximum VNIR–SWIR reflectance values for most samples were less than 0.20. The same pattern was true with TIR spectra, which had reflectance values of 0–0.54 in the first bin, decreasing progressively in maximum reflectance value on reaching bin 4, which had a maximum reflectance of 0.36.

By analyzing the continuum removed average spectra of the VNIR–SWIR (Fig. 7), it was observed that a broad absorption feature in the range 650–1600 nm, centered at 1250 nm, often occurred in the samples of the first bin, and it gradually disappeared in samples with higher Cu %. This absorption feature is characteristic of reduced iron (hydro-)oxide as magnetite, and its depth is proportional to the Fe2+ content (Hunt and Salisbury, 1970). Likewise, it was observed that some absorption features decreased their depth and/or the number of samples with increase in Cu %, as in a double absorption feature in the range 2250–2320 nm, centered at 2270 nm and 1310 nm, which is characteristic of chlorite (Hunt and Salisbury, 1970), and an absorption feature in the range 2150–2250 nm, centered at 2200 nm, which is characteristic of white micas as muscovite, illite, paragonite and phengite (Hunt and Salisbury, 1970). It was also noted that the absorption feature in the range 750–1050 nm, centered at 900 nm, which is characteristic of hematite and goethite (Curtiss, 1985; Morris et al., 1985), was usually deeper for samples with higher Cu %.

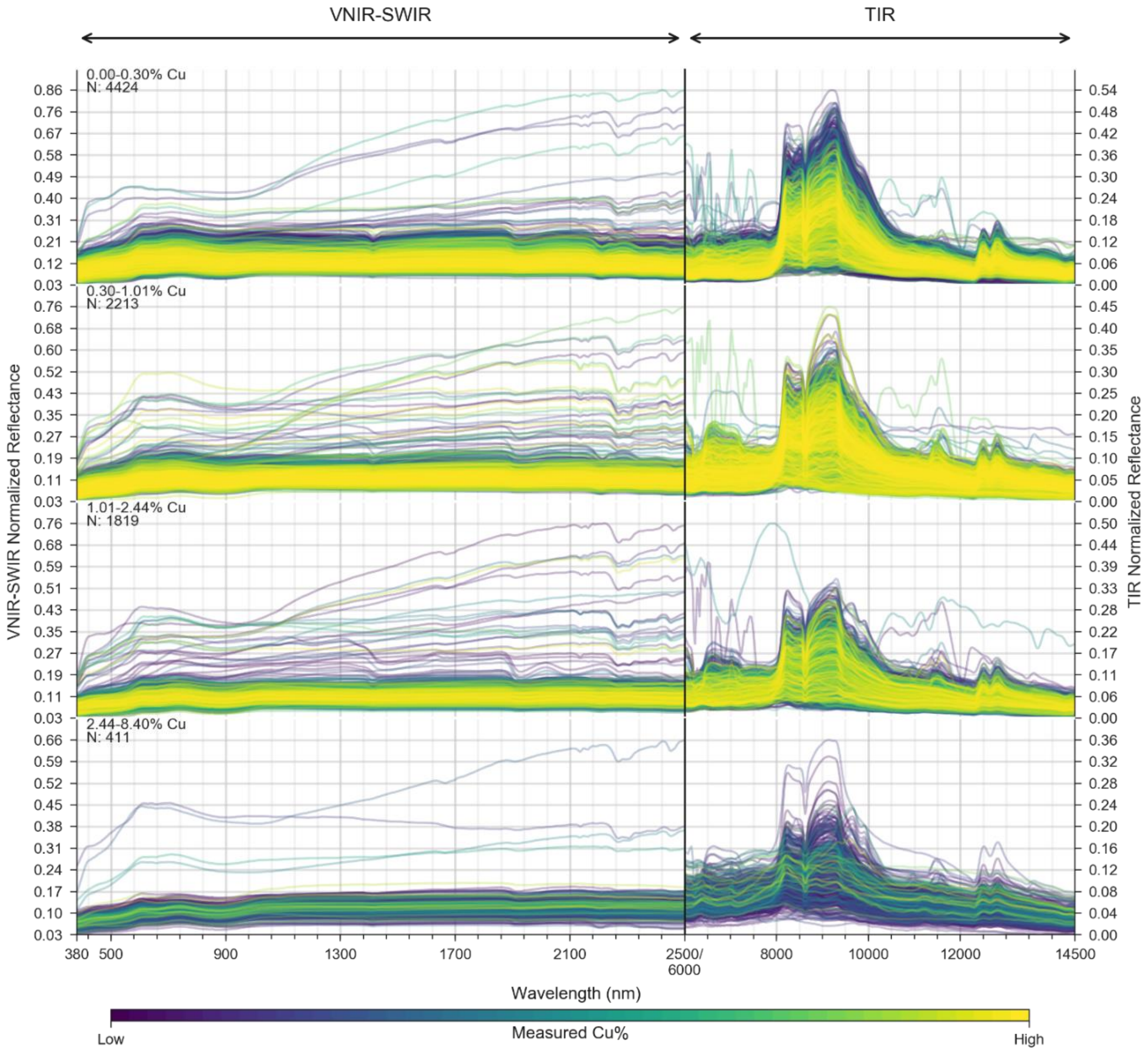


Fig. 6 Average spectra for each geochemical sample grouped by Cu %. N = number of samples.

The continuum removed TIR spectra (Fig. 7) also showed some absorption features that were more notable in samples with lower Cu %, such as the absorption feature in the range 6400–8200 nm, which presented a flat pattern in the range 7200–7800 nm only up to bin 3. The absorption feature in the range 9300–12500 nm changed its symmetry as the Cu % increased, by decreasing the depth of absorption at 12250 nm. In addition, the absorption feature in the range 12800–14500 nm gradually became less notable for samples with higher Cu %.

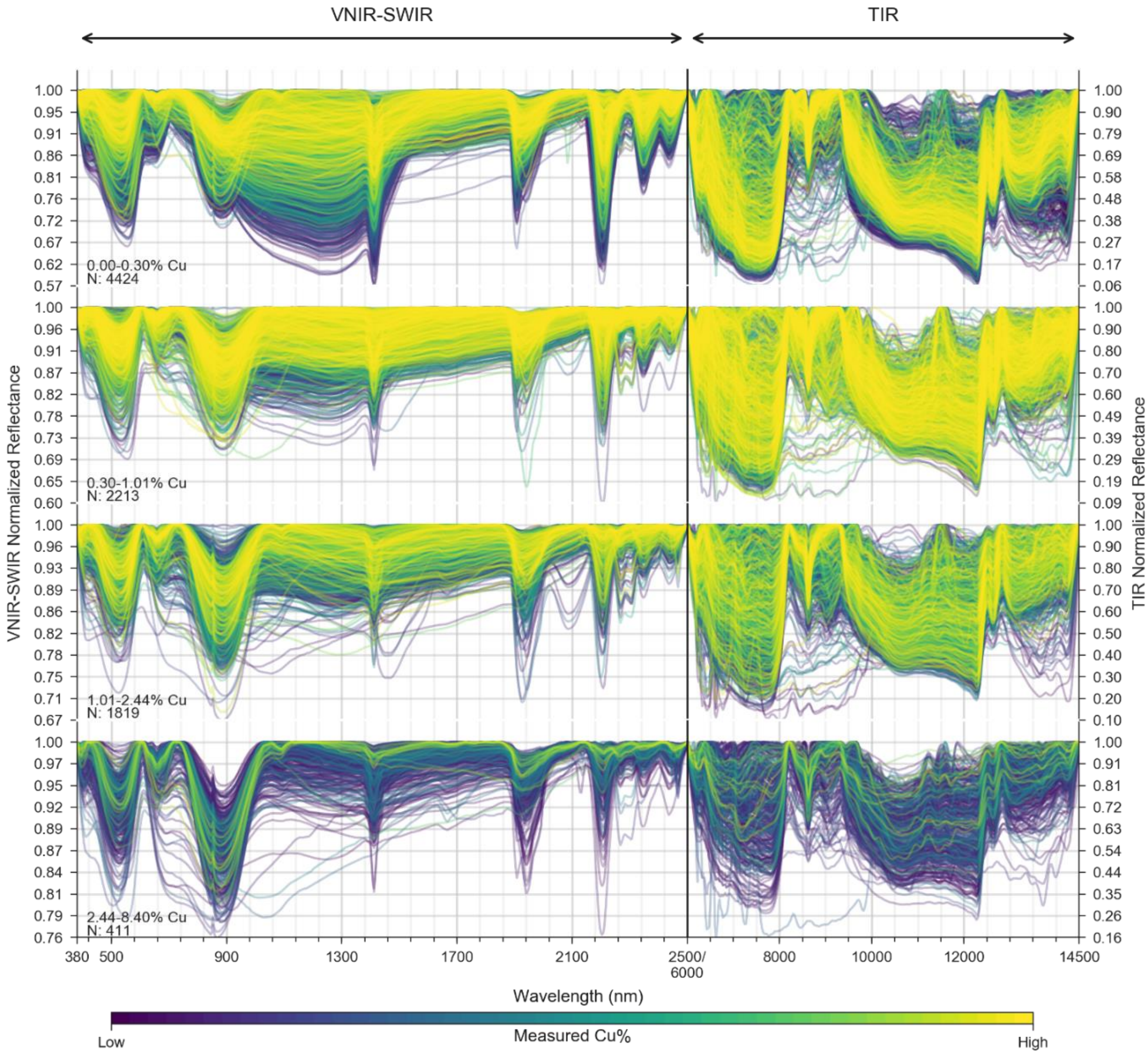


Fig. 7 Continuum removed average spectra for each geochemical sample grouped by Cu %. N = number of samples.

The original hyperspectral data acquired from AuScope were provided as TSG™ software (CSIRO Earth Science and Resource Engineering - CESRE - Division, Sydney, Australia) file format. In these files, in addition to the spectral curves, the automatic mineral interpretation obtained by "The Spectral Assistant™" (TSA) module was provided, for both VNIR–SWIR and TIR spectra. The percentage of samples identified for each mineral is shown in Figure 8, either for all data or data grouped by Cu %. When we analyzed the relationship between the identified minerals and the Cu %, it was observed that some minerals were identified in a similar percentage of samples in all bins, and others were identified in a large or smaller number of samples as the Cu % increases. Quartz, muscovite, phengite, microcline, siderite, calcite, hematite, and montmorillonite were the minerals mostly identified in the samples (all data in Fig. 8). The most identified minerals in samples with higher Cu % (bins 3 and 4) were quartz, muscovite, phengite, calcite, siderite, and hematite, with quartz being the most frequently identified mineral (identified in 65% of the samples of bin 4). The percentage of samples with calcite, siderite, hematite, goethite, gypsum, biotite, kaolinite-WX and dolomite increased as the Cu % increased. However, for some of these minerals (e.g., siderite, biotite, and dolomite), the percentage of samples decreased in bin 4. Moreover, the percentage of samples with quartz, muscovite, phengite, albite, orthoclase, and microcline decreased as the Cu % increased.

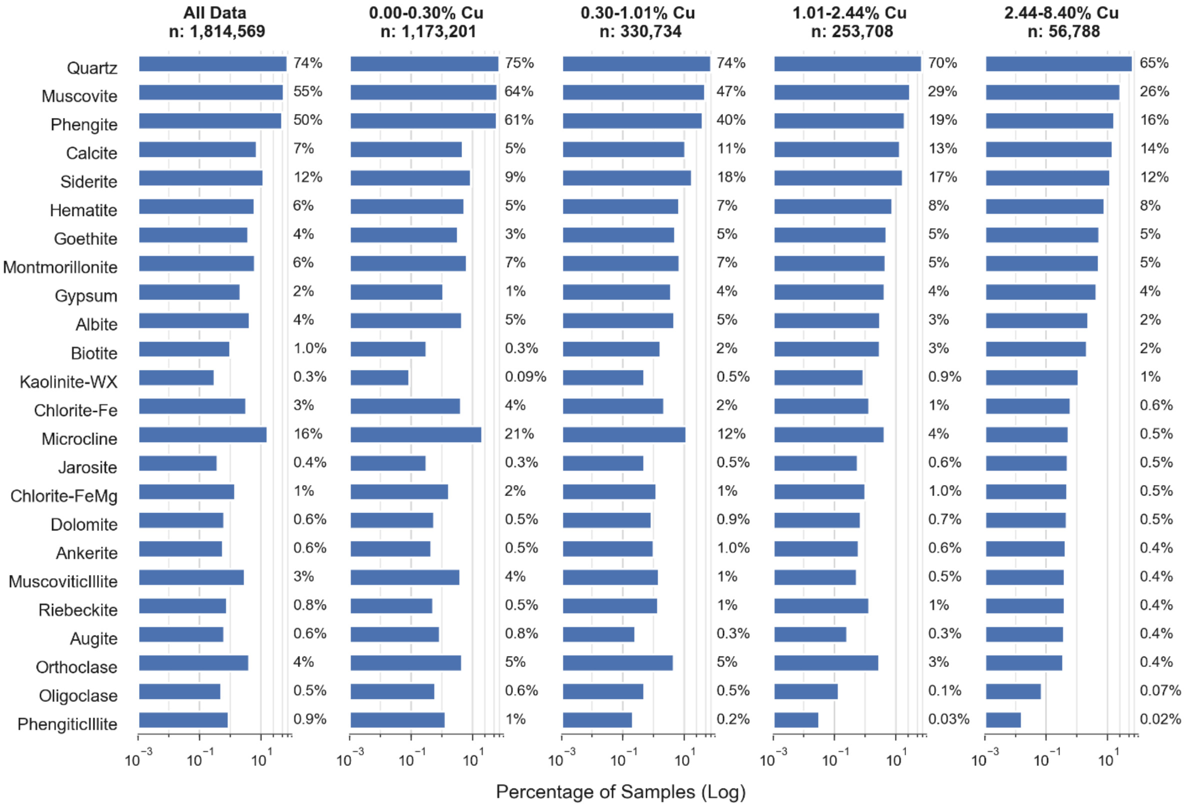
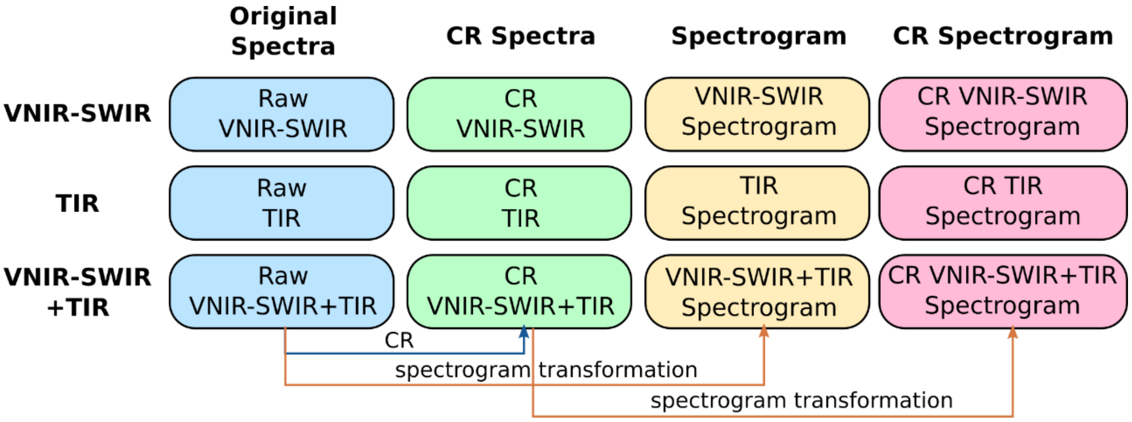


Fig. 8 Percentage of samples in which a given mineral has been identified, for each mineral identified by the TSA module of the TSG software (only minerals identified in more than 0.5% of the total samples in all Cu bins are shown), computed for all the spectral data and for data in each Cu % bin.

The variation in mineralogy of the samples obtained from the hyperspectral data is the same variation observed/described by Ehrig et al. (2012) in the hydrothermal alteration zones of the ODD, where feldspar (microcline, orthoclase) and plagioclase (albite) are altered to sericite (muscovite, phengite) as the alteration intensifies. The relation between the increase of Cu % and the decrease of feldspar and plagioclase identified in the samples indicates that the intervals sampled for geochemical analysis intercepted hydrothermal zones with progressively increasing alteration intensity as the Cu % increases. Samples with Cu % higher than 1.01% (bins 3 and 4) probably mostly intercepted altered rocks, in which almost all feldspars and plagioclase are altered to sericite, being composed mainly of quartz and sericite. This correlation observed for variations in spectral behavior and Cu content of the samples reaffirms the spatial distribution and relationship among alteration minerals and mineralized zones observed by Ehrig et al. (2012) and Tappert et al. (2013), and supports the use of hyperspectral data to estimate Cu content.

6.3.3 Model data input

Twelve different input datasets were generated from the original hyperspectral data (Fig. 9). First, the original data were divided into three datasets, namely VNIR–SWIR data, TIR data, and VNIR–SWIR + TIR data (blue color in Fig. 9). These three datasets consist of the raw spectra, without preprocessing. Another set of three datasets were generated by removing the continuum from the raw spectra datasets (green color in Fig. 9). The other six datasets were generated from these, transforming the spectra of the three raw datasets and the three datasets with continuum removal into spectrograms (yellow and red colors in Fig. 9). All the datasets were standardized before model training (see Section 2.5.). Data not transformed into spectrograms consist of 1D vectors, and transformed data consist of 2D vectors. Two CNN architectures based on convolutional neural networks, one for 1D data and one for 2D data, were developed and trained with the generated input datasets. Then, the performances of models obtained per input dataset were compared to investigate which spectrum region, VNIR–SWIR, TIR or both, is the best for model training, as well as to investigate which pre-processing is more adequate, continuum removal, transformation to spectrogram, both, or none.



VNIR = visible and near-infrared; SWIR = short-wave infrared; TIR = thermal infrared; CR = Continuum Removed

Fig. 9 Model input datasets.

6.3.4 CNN Model

Developing a CNN model is highly interactive. The number and type of layers used, kernel and filters size at each layer, dropout rate, regularization constant, and learning rate are some of the hyperparameters that need to be defined to design the model. In this section, we introduce the 1D and the 2D CNN models and present the networks with the best performing combination of hyperparameters.

6.3.4.1 Introduction to CNN

Convolutional neural networks (LeCun et al., 1989) are deep learning algorithms specialized in processing data with known grid-like topology, as image data. These networks have been tremendously successful in practical applications. CNNs employ a mathematical operation called convolution, which is a specialized kind of linear operation (Goodfellow et al., 2016). CNNs are composed of three basic layers, known as convolutional layers, nonlinearity layers, and pooling layers. These layers, when connected, establish a network architecture designed to take advantage of the spatial structure of the input data (Zhang and Goh, 2016).

The convolutional layers consist in a set of n rectangular matrices, also called filter, where n is the number of filters. Each filter is slid over the entire input data to produce an output, or feature map. At each iteration, the algorithm computes the inner product between a patch of the input data x^i and a filter k_j^i, where i is the data patch index and j is the filter index. This operation can be expressed as a single matrix multiplication of the form $K_j * X$ where K_j is the large Toeplitz matrix of convolution operations and X is a matrix of vectorized inputs concatenated together. The result of this operation is then passed as input to a nonlinearity layer, where it is combined with an activation function $f(\cdot)$ and a bias, b_j, thus:

$$Z_j = f(K_j * X + \ b_j)$$

where Z_j is the feature map produced by the convolution of filter j. A commonly chosen activation function is the rectified linear unit (ReLU) function $f(x) = \max{(0, x)}$. The output of the convolutional layer is a matrix of n feature maps concatenated together.

The pooling layer function reduces the size of the input, replacing the values of the input matrix with a summary statistic of the nearby values. The max pooling

(Zhou and Chellappa, 1988) is a commonly chosen pooling operation; it calculates the maximum value *P* within a rectangular neighborhood S of each feature map j, thus:

$$P_j^S = \max_{i \in S} Z_j$$

The last component in any CNN is a final layer that takes as input the feature maps derived from all previous layers and it outputs a predicted value. To solve for the numerous parameters in the network, some objective function must be defined that penalizes poor performance during training. In general, the objective function has the form:

$$\phi = \|f(\theta; X) - y\|_p$$

where y represents the true labels, $f(\cdot)$ is some function of the input data X (the images), which the network is trying to learn, θ represents the parameters of the function (in this case this would include all the convolution kernels and biases of the CNN), and $\|\cdot\|_p$ is some p-norm of the misfit. The parameters are then optimized iteratively using the back-propagation algorithm (LeCun et al., 1989), which in most cases is some form of (stochastic) gradient descent. As the network is optimized, the convolution kernels are "learned" to provide the best set of feature maps from which to minimize the error expressed by the objective function.

6.3.4.2 Network architecture

Two CNN architectures were developed, one composed of 1D convolutional and max-pooling layers, and the other composed of 2D convolutional and max-pooling layers. The architecture of both models was based on the image recognition VGG model (Simonyan and Zisserman, 2014), which is one of the state-of-art CNN models for image classification.

The input of the 1D CNN architecture is a fixed size 1xN vector, built from normalized reflectance or continuum removed reflectance data, where N is the number of spectral bands in the input data (531 for VNIR–SWIR data, 341 for TIR data, and 872 for VNIR–SWIR + TIR data). The only preprocessing applied to these data besides continuum removal was standardization, computed on the training set (Section 2.5.). The input 1D vector was passed through a stack of convolutional layers. The first two layers had 1x1 and 1x3 kernel sizes, respectively, and 64 filters each, which were followed by a stack of three convolution blocks. Each block was composed of two convolutional layers, with 1x1 and 1x3 kernel sizes, respectively, followed by a dropout regularization. The filter sizes of the convolutional layer inside the convolutional block were the same, starting with 128 filters in the first block and doubling at each block, going to 256 and 512, in the second and third blocks, respectively. In all these convolutional layers, the convolution stride was fixed to 1 and the spatial padding was set to preserve the spatial resolution after convolution. After the convolutional layers, spatial pooling was carried out by a max-pooling layer with a 1x2 window size and stride of 2, which follow the convolution block. Then, the output was flattened (transformed into a 1D vector) and passed through two fully-connected layers, with 32 channels each. The final layer was a fully-connected layer with 1 channel (prediction value) and a linear activation function (Fig. 10a). All convolutional and fully-connected layers were followed by a rectification (ReLU) non-linearity (Krizhevsky et al., 2012), with the exception of the last fully connected layer.

The 2D CNN architecture took as input a fixed size 11xN image, derived by the transformation of the normalized reflectance or continuum removed reflectance in a spectrogram, where N is the number pixels in the wavelength (spectral band) dimension of the spectrogram (52 for VNIR–SWIR data, 33 for TIR data, and 85 for VNIR–SWIR + TIR data), which varies according to the number of input data spectral bands. As in the 1D CNN architecture, before training, the input data were standardized against the training data (Section 2.5.). The image was passed through a stack of two convolutional blocks, each block with two convolutional layers, with 1x1 and 3x3 kernel size, respectively. The filter sizes of both convolutional layers inside the block were equal, being 32 in the first block and 64 on the second. The stride of all convolutional layer was set to 1 and the padding was set to preserve the spatial resolution. A max-pooling layer with a 2x2 window size and stride of 2 was applied to the output of the last convolutional layer. The result of the pooling operation was flattened and passed through two fully-connected layers, with 16 channels each. The final layer was a fully-connected layer with 1 channel (prediction value) and a linear activation function (Fig. 10b). All convolutional and fully-connected layers were followed by a rectification (ReLU) non-linearity, with the exception of the last fully connected layer.

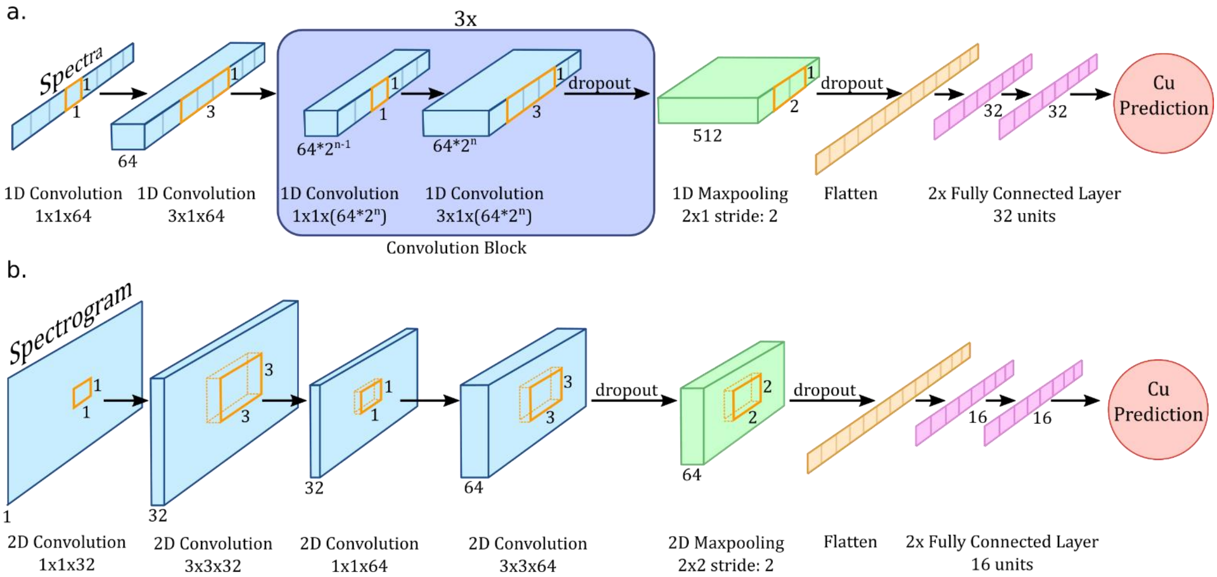


Fig. 10 CNN architectures: (a) 1D and (b) 2D.

6.3.4.3 Training the network

Before training, the datasets were divided into a training set, a validation set and a test set. The validation set was used to evaluate the model performance during training for parameter selection, and the test set was used as an independent set for measuring model performance after parameter selection. The samples of the drill-core RD2785 (Fig. 1), which had 1,479 samples with Cu data (~16% of the total number of samples), were selected and used as test set. From the remaining samples, 20% (1,478 samples) were randomly selected and used as validation set, and the other 80% (5,911 samples) were used as training set.

The training of both networks was conducted using *Adam* optimizer (Kingma and Ba, 2014) with mini-batch gradient descent and root mean squared error (RMSE) as the loss function. The batch size was set to 128. Dropout regularization was used in both networks (Srivastava et al., 2014). In the 1D CNN, the last convolutional layer of the convolution block and the max-pooling layer were followed by dropout layers with dropout ratio set to 0.5 (Fig. 10a). Similarly, in the 2D CNN, the last convolutional layer and the max-pooling layer were followed by dropout layers with dropout ratio set to 0.1 (Fig. 10b). It is important to note that, despite the similarity, the 1D CNN had four dropout layers, one for each of the three convolution blocks, and one for the max-pooling layer, while the 2D CNN had only two dropout layers. In both networks, the convolutional layer regularized with dropout also had weight decay regularization (Ng, 2004), with L_2 penalty multiplier set to 0.0003. The learning rate was initially set to 10^{-2} in the 1D CNN, and to 7.5x10^{-4} in the 2D CNN. In both networks, these initial values were set to decrease by a factor of 10 when the validation RMSE stopped improving (i.e., the validation RMSE did not drop 10^{-4} units during 100 epochs). Early stopping was used on both architectures, and the training was set to stop when the validation RMSE did not drop 10^{-4} units during 120 epochs. The weights in both networks were initialized using the random uniform initialization procedure of Glorot and Bengio (2010), also known as Xavier uniform initialization. The CNN models were implemented in Python programming language using the packages Keras (Chollet and others, 2015) and TensorFlow (Martín et al., 2015).

6.3.5 Synthetic minority over-sampling technique for regression with gaussian noise (SMOGN)

When analyzing the distribution of samples by the Cu % (Fig. 5), it was observed that the number of samples with low Cu % (< 1.98%; 5,452 samples) was much larger than the number of samples with high Cu % (>= 1.98%; 459 samples), indicating that the dataset had imbalanced domains (~1:12 ratio of samples in minority/majority classes). Machine learning algorithms, including neural networks, commonly have problems when trained with imbalanced datasets, both for classification and regression task (Branco et al., 2017; He and Garcia, 2009; López et al., 2013; Prado et al., 2020; Torgo et al., 2013). Usually, models trained with imbalanced dataset present a tendency to output values associated to the majority class (target values with many samples); that is, a bias towards the majority class. Therefore, these models commonly are not good for predicting samples of the minority class (target values with few samples).

To address the problem of imbalanced domains in the training dataset and to enhance the performance of the CNN model, the Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) proposed by Branco et al. (2017) was used. This technique combines random under-sampling of the majority class with SmoterR (Torgo et al., 2013) and introduction of Gaussian Noise (Branco et al., 2016) to over-sample the minority class. As suggested by the authors of the SMOGN algorithm, the relevance function proposed by Ribeiro (2011) was used to select the minority class (rare/extreme target values) and majority class (normal target values) samples.

The relevance function maps the target variable domain to the range [0,1], where 1 represents the maximum relevance. The method proposed by Ribeiro (2011) automatically estimates the relevance function from the target variable boxplot statistics. The median and the outlier threshold ($OT = Q3 + O_{coef}\ x\ IQR$, where $Q3$ is the third quartile, O_{coef} is the outlier coefficient (commonly set to 1.5), and IQR is the inter-quartile range) were used to compute the relevance function with the piecewise cubic Hermite interpolation method (Dougherty et al., 1989; Ribeiro, 2011), which maps to 0 the target values below the median, to 1 the target values above the OT, and to the range (0,1) the target values between the median and the OT. A threshold t_R on the relevance values must be defined by the user, whereby values of the target variable with relevance below t_R are assigned to the majority class, and values of the target variable with relevance above t_R are assigned to the minority class.

The SMOGN algorithm takes the samples assigned to the majority class and applies a random under-sampling procedure. Samples assigned to the minority class are over-sampled using either SmoteR or the introduction of Gaussian Noise strategy to generate new cases. The distances in the dataset vectorial space between the minority class samples and its nearest neighbors are used to determine which over-sampling strategy the SMOGN algorithm will apply. When the distance between the samples is considered “safe” (short enough) to perform interpolation through SmoteR, the SmoteR strategy is used, otherwise the SMOGN algorithm generates new cases by introducing Gaussian Noise to the samples.

The Python implementation of SOMGN algorithm, provided in https://github.com/nickkunz/smogn, was used for resampling the original dataset. This implementation provides two strategies to obtain the under-/over-sampling rates used by the SMOGN algorithm, called “balance” and “extreme”. The “extreme” strategy, used in this work, first calculates a scale factor by the formula $S = 2b/\sum\frac{b^2}{N_{class}}$, where $b = n/N_{class}$, n is the total number of training samples (5,911) and N_{class} is the number of classes obtained by the relevance function (2 in this work, corresponding to the minority and majority classes). Then calculates the resampling rates by the formula $\%_{resampling} = S_{class}/N_{class}$, where $S_{class} = b^2/S\ x\ N_{class}$. This strategy generates a larger number of synthetic samples, resulting in a resampled dataset with a larger number of samples than the original training dataset. In addition, two other parameters must be set by the user in this SMOGN implementation, namely the boxplot outlier coefficient O_{coef} and the relevance threshold t_R, which were set to 1.5 and 0.8, respectively. Both the O_{coef} and t_R were selected based on performance (RMSE) on validation set by doing an exhaustive search over these hyperparameters values. During this procedure, all possible combinations for the values of $O_{coef} = [1, 1.2, 1.5, 2, 3]$ and $t_R = [0.5, 0.6, 0.7, 0.8, 0.9, 1]$ were used to generate different training datasets by the SMOGN algorithm. Then, a model was trained for each dataset. The RMSE obtained on validation data by each model was used to select the best combination of hyperparameters. With these hyperparameters the Cu % threshold obtained by the relevance function was 1.98 %, therefore, samples with Cu % less than this threshold belong to the majority class (being undersampled), and samples with Cu % larger than or equal this threshold belong to the minority class (being oversampled). After applying the SMOGN algorithm to the training set, the number of samples of the majority class changed from 5,452 to 4,955 samples, and the number of samples of the minority class changed from 459 to 5,414 samples, balancing the training dataset to approximately 1:1 ratio of samples in the minority and majority classes.

6.4 Results and discussion

1. Model performance

As mentioned, 12 regression models were developed, one per training data (Fig. 9); six models for training the 1D CNN architecture (Fig. 10a) with the 1D data (reflectance curve) and six models for training the 2D CNN architecture (Fig. 10b) with the 2D data (spectrograms). To measure the performance of each model, the RMSE against the training, validation and test sets were computed for individual models. The values obtained are show in Table 1. The 2D CNN model trained with the Raw VNIR–SWIR+TIR Spectrogram dataset obtained the best performance, with RMSE of 0.065 on the training set, 0.495 on the validation set, and 0.494 on the test set. The SMOGN algorithm was then used to resample the Raw VNIR–SWIR+TIR spectrogram dataset (best performing dataset) to increase the performance of the CNN model. The 2D CNN model trained with the resampled dataset obtained a RMSE of 0.051, 0.487, and 0.482 on the training, validation, and test datasets, respectively, performing slightly better than the model trained without SMOGN.

Table 1 Absolute RMSE obtained per training dataset.

	RMSE		
Dataset	Train	Validation	Test
2D Raw VNIR–SWIR+TIR Spectrogram with SMOGN	0.051	0.487	0.482
2D Raw VNIR–SWIR+TIR Spectrogram	0.065	0.495	0.494
2D CR VNIR–SWIR+TIR Spectrogram	0.092	0.511	0.537
2D Raw TIR Spectrogram	0.090	0.524	0.572
2D CR TIR Spectrogram	0.088	0.528	0.613
2D Raw VNIR–SWIR Spectrogram	0.074	0.608	0.492
2D CR VNIR–SWIR Spectrogram	0.085	0.610	0.516
1D CR VNIR–SWIR+TIR	0.670	0.705	0.480
1D CR TIR	0.734	0.751	0.567
1D Raw VNIR–SWIR	0.777	0.798	0.550
1D Raw VNIR–SWIR+TIR	0.886	0.896	0.744
1D CR VNIR–SWIR	0.886	0.896	0.743
1D Raw TIR	0.886	0.896	0.740

The 1D CNN models trained both with raw and continuum removed reflectance spectra were the worst performing models, obtaining higher RMSE than those obtained by the 2D CNN models, indicating that the CNN architectures perform better with the 2D spectrogram data. Models trained with all available spectral bands (VNIR–SWIR+TIR) performed better than models trained with only part of the spectral bands (VNIR–SWIR or TIR), as expected, reinforcing the importance of using the VNIR-SWIR spectral bands together with the TIR bands for a more accurate quantitative analysis of the spectral data. The best performance of 1D models was obtained with the continuum removed VNIR–SWIR+TIR data, and the second-best performance with the continuum removed TIR data, showing that the continuum removal played an important role in reducing the RMSE of 1D models. However, by analyzing the RMSE of 2D models it can be seen that models trained with continuum removed datasets performed worse than their correspondent pair; that is, the model trained with the same spectral bands but without continuum removal. The improvement in the RMSE of 2D CNN models when using the dataset without continuum removal is probably because the 2D CNNs can extract important information from raw spectra, as albedo, which is lost when continuum removal is applied.

To better visualize the results, the predictions of Cu % obtained by the best performing model were compared to the measured values along the drill-cores for training, validation, and test samples. Fig. 11 shows the measured and predicted Cu % of the training and validation samples along the interval 450–750 m in all drill-cores used for training. The values of predicted Cu % were remarkably close to the measured values. The less accurate predictions were those obtained for samples with high measured Cu %, especially for the validation samples (not used for training), and commonly are related to the underestimation of Cu. However, despite this, the model was very precise in distinguishing qualitatively and quantitatively the high-grade in the low-grade zones represented in the training drill-cores. The measured and predicted values of Cu % of the test samples are shown in Fig. 12. The predictions obtained for the test samples simulate the behavior of the model in a production environment, where hyperspectral data were acquired for a drill-core that had no geochemical analysis and the model was used to predict the Cu % by means of hyperspectral data. The predictions for the test samples were far from the measured values compared to those obtained for the training and validation samples, which is expected as the test samples were not used for training the model, and the training dataset do not have any samples from the test drill-core. As for the training/validation samples, the less accurate predictions were those obtained for samples with high measured Cu %, and commonly were related to the underestimation of Cu %. Predictions made in the high-grade zones, between the intervals 490–510 m and 650–700 m (Fig. 12), were quite distant from the measured values. In the 650–700 m interval, for example, the measured values were mostly in the 2–4% range, but the predictions were close to 1%, underestimating the Cu % by 1–3%. Despite this, the model prediction can be used to qualitatively distinguish between high-grade and low-grade zones, because the predicted values for the high-grade zones (close to 1%) were higher compared to those predicted for the low-grade zones, which are close to 0%.

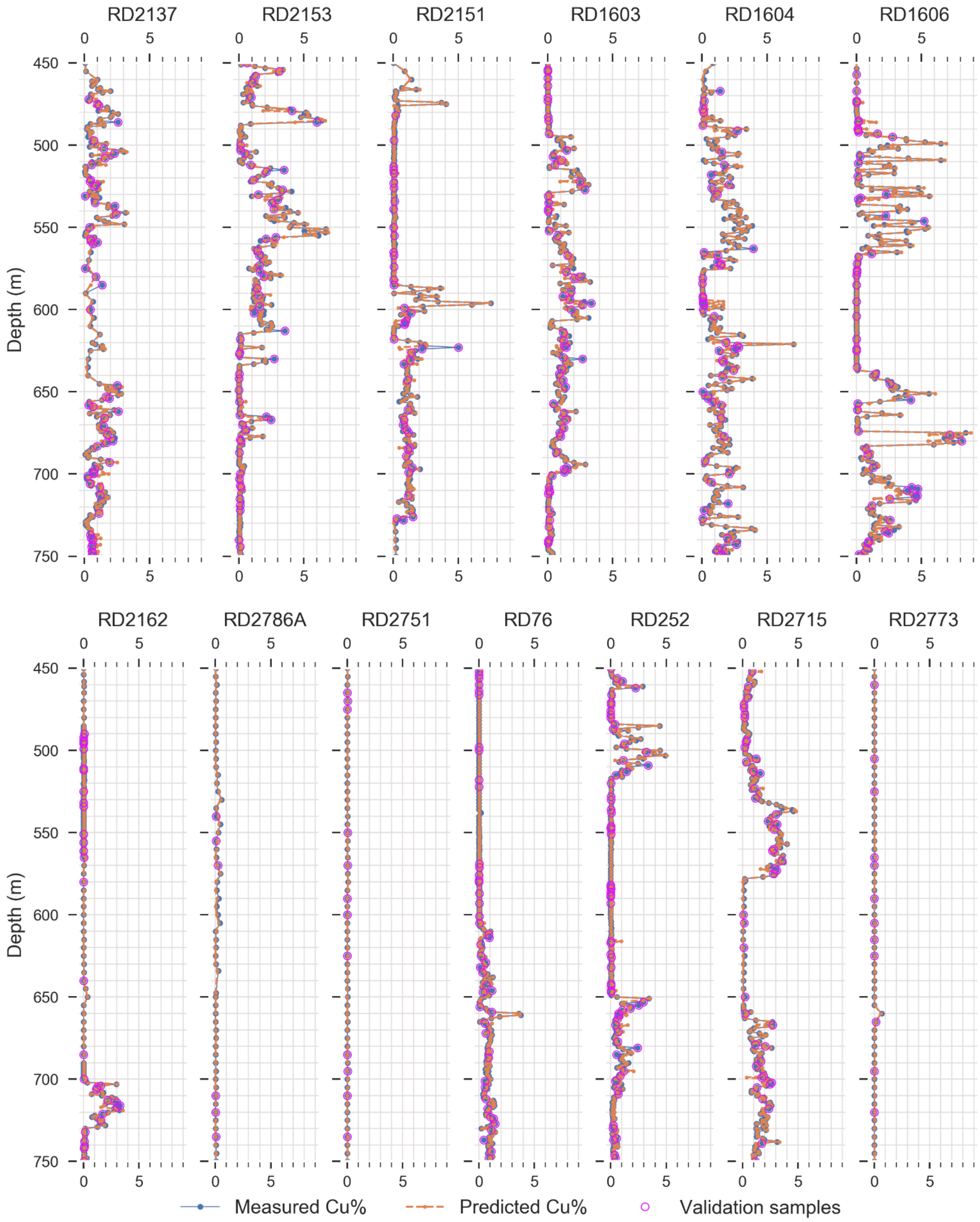


Fig. 11 Measured and predicted Cu % of training and validation samples for the interval between 450–750m of all drill-cores used for training. Predictions are obtained from the best performing model (2D CNN trained with raw VNIR–SWIR+ TIR spectrogram data).

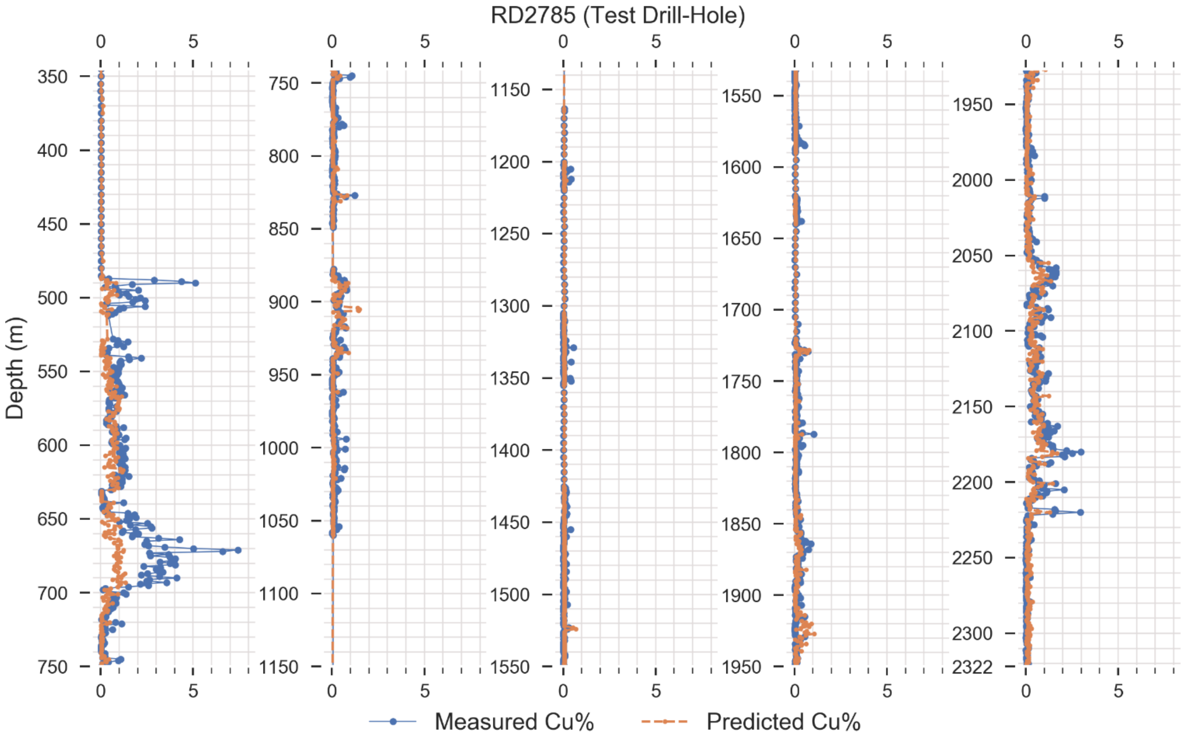


Figure 12 Measured and predicted Cu % of the drill-core reserved for test. Predictions were obtained by the best performing model (2D CNN trained with raw VNIR–SWIR+ TIR spectrogram data).

As can be seen from the results, the model shows a clear bias towards the lower Cu values, tending to underestimate the higher Cu values. This behavior suggests that, although the SMOGN technique improved the performance of the model, balancing the data by creating synthetic samples was not effective in reducing bias towards the minority class. Recent work (Yang et al., 2021) shows that this approach has some intrinsic drawbacks, as it does not take the distance between targets into account, it makes a simplified division of the data into rare and frequent, and the synthetic samples generated from high dimensional data are generally not meaningful to the model. Therefore, other solutions that can deal with imbalanced regression problems need to be tested in future work, such as the one proposed by Yang et al. (2021).

The model can be improved further in a production environment by collecting a few samples of the drill-core of interest, send them for geochemical analysis and use these data to update the model weights. To simulate this procedure, we randomly selected 10% of the test samples (148 samples) and used them to update the weights of the best performing model (2D CNN trained with raw VNIR–SWIR+TIR dataset). This experiment was repeated 100 times (each time with a different randomly picked sample set) and the RMSE obtained by the updated model on the test samples was recorded for each iteration. The mean RMSE on the test data of the updated models was of 0.394, with standard deviation of 0.04. The predictions for the test samples after the update of the model weights obtained in one of these experiments are shown in Fig. 13, as well as the measured values. The model had significantly enhanced the performance in the test samples after this procedure, especially in the high-grade zones, as in the 650–700 m interval. Therefore, we recommend this procedure when applying the model in a production framework, which not only eliminates the needs for collecting geochemical analysis from the drill-cores but also considerably reduces the number of samples to be analyzed, thus reducing costs.

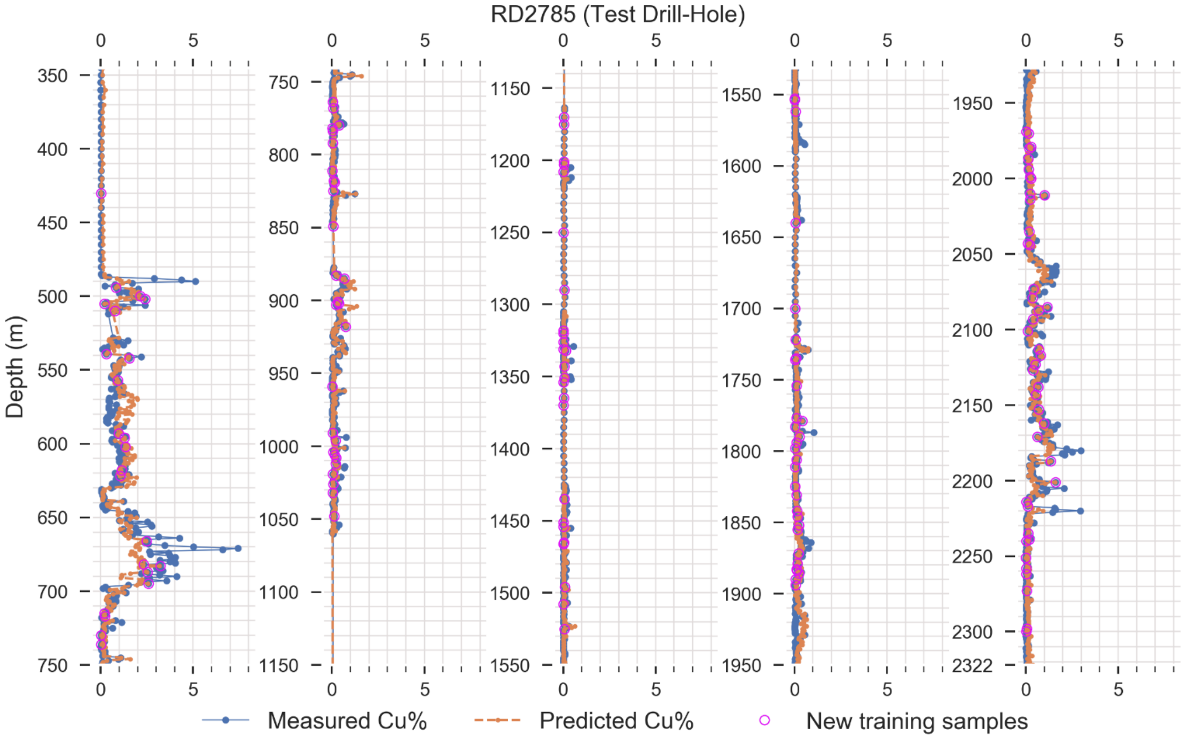


Fig. 13 Measured and predicted Cu % of the drill-core reserved for test obtained by the best performing model after updating the weights with test samples. The samples used to update the model weights are highlighted in pink circles.

Given the high spatial resolution (0.8 mm) of hyperspectral data, the trained model can be used to predict Cu % at a higher spatial resolution than the one used in geochemical sampling. To qualitatively evaluate the model performance in predicting the Cu % at a more detailed spatial resolution, a dataset was built by computing the average spectrum every 20 cm of test drill-core. Then, the best performing model with weights updated (explained above) was used to predict the Cu % of this dataset (Fig. 14). The predicted values were, in general, nearby the measured values, with the high-grade zones being clearly distinguished from the low-grade zones. As for the other datasets, the model tends to underestimate the Cu % in the high-grade zones, although that the predicted values in these zones were sufficiently higher enough to identify these zones.

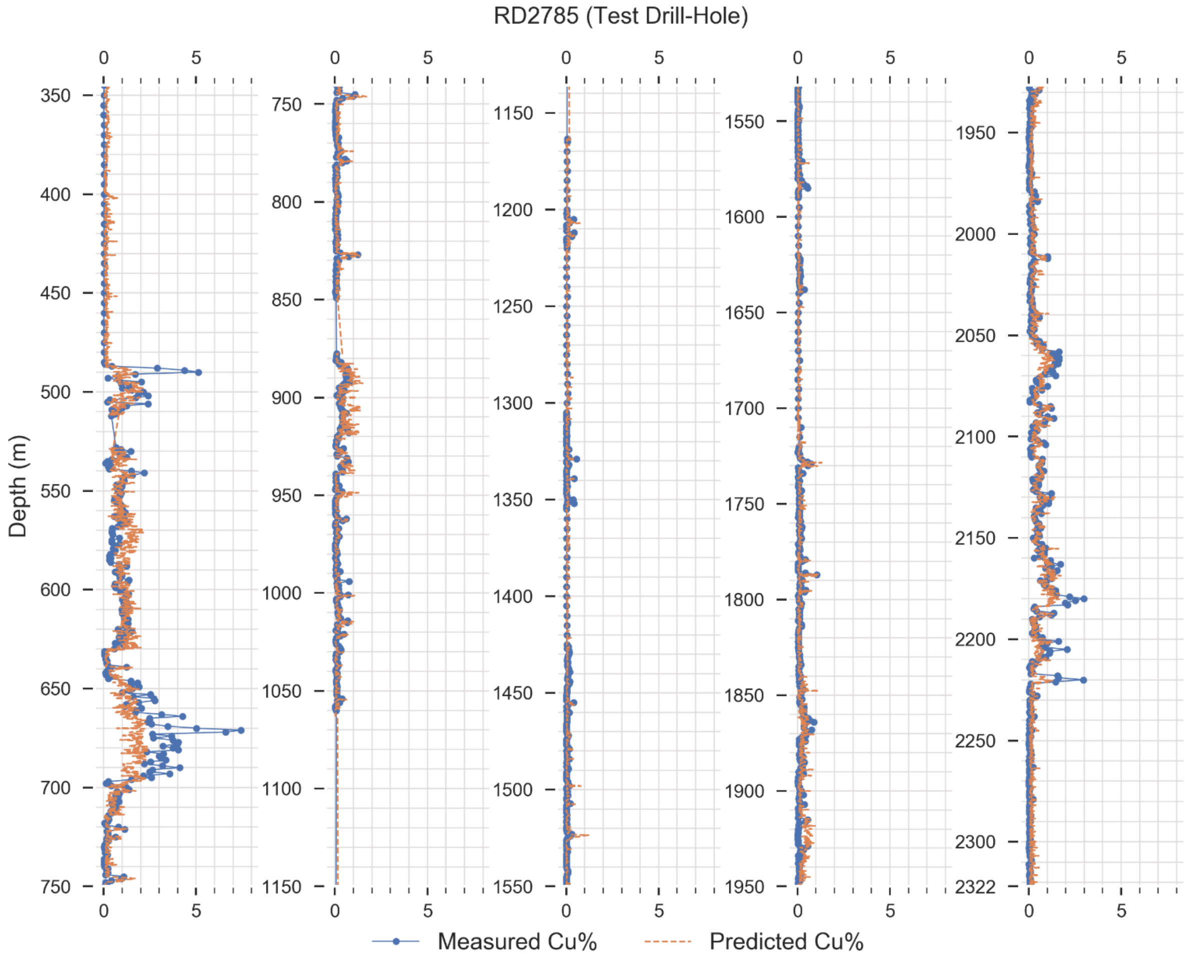


Fig. 14 Predicted Cu % at 20 cm spatial resolution of the drill-core reserved for testing, plotted with measured Cu %. Predictions were obtained by the best performing model (2D CNN trained with raw VNIR-SWIR+ TIR spectrogram data) after updating the weights with test samples.

6.4.1 Shapley Additive Explanations (SHAP)

The Shapley Additive Explanations (SHAP) approach to explain model predictions introduced by Lundberg and Lee, (2017) was carried on the best performing model (2D CNN trained with raw VNIR–SWIR+TIR dataset). The SHAP is based on the coalitional game theory. It assumes that each input feature of a model is a member of a coalition ('player') in a game where the model prediction is the coalition output or payout. The Shapley value method (Shapley, 1953) was used to fairly allocate the payout (model prediction) among the members of the coalition (features). The contributions of the input features to the model output for each sample are indicated by the Shapley values, also known as SHAP values, returned by the SHAP algorithm. A SHAP value is proportional to the contribution of a feature to the change in the expected value of a model's output (the mean value of a model output over training samples); a positive SHAP value corresponds to an increase in model output and a negative value to a decrease in model output.

To compute SHAP values in this study, we used the Deep SHAP explainer to compute the SHAP values, which is a model explainer function made for deep neural networks provided in the SHAP python library (version 0.40.0; Lundberg and Lee, 2017). To identify the features that drive the predictions of the CNN model, the SHAP values were computed for each feature of all test samples, using as background 1000 samples randomly selected from the training set. After that, we also calculated the maximum absolute SHAP value (MAXSHAP) for each column of each sample using the following equation:

$$MAXSHAP_j = \begin{cases} \max\limits_{1 \le i \le m} x_{ij} & \max\limits_{1 \le i \le m} \left|x_{ij}\right| \ge \min\limits_{1 \le i \le m} \left|x_{ij}\right| \\ \min\limits_{1 \le i \le m} x_{ij} & \max\limits_{1 \le i \le m} \left|x_{ij}\right| < \min\limits_{1 \le i \le m} \left|x_{ij}\right| \end{cases}$$

where $MAXSHAP_j$ is the maximum SHAP value for the column j of the input sample, x_{ij} is the SHAP value for the feature at row i and column j, and m is the total number of rows (11 in this study). As the input data of the CNN model were spectrograms, the columns of the input data represent a wavelength range. Therefore, the MAXSHAP values can be used as a proxy for the importance of these wavelength ranges (or spectral bands) to the model output (i.e., Cu %).

Figure 15a shows the mean of the absolute MAXSHAP values for all the test samples obtained for the best performing model. The importance of each spectral band used by the model for the prediction of Cu % is proportional to the height of the bars. The four spectral bands most critical for increasing the predicted Cu % were 14250 (14125–14375), 620 (600–640), 12500 (12375–12625), and 11500 (11375–11625), and the four spectral bands most critical for reducing the predicted Cu % were 12750 (12625–12875), 900 (880–920), 9500 (9475–9625), 860 (840–880). These spectral bands must be related to absorption features of minerals that differentiate the high grade from lower grade zones, i.e., the presence and/or shape and/or intensities of these features influences the Cu content. It is worth mentioning that these wavelengths correspond to the center of the most sensitive pixels of the spectrogram, but each pixel comprises a range of wavelengths, and the pixel widths for the VNIR–SWIR and TIR spectrograms are different, being 40 and 250 nm, respectively.

To better understand the relationship between the MAXSHAP values and the spectrogram values (feature values), we calculated the sum of each column of the spectrogram and used these values as the feature values related to the MAXSHAP values obtained for a sample. Figure 15b shows the MAXSHAP and the feature values obtained per test sample. The bars indicate the MAXSHAP values obtained for all test samples at a certain wavelength (column) of the spectrograms and are colored according to the variation of the feature values in relation to the MAXSHAP values. Bars that have higher features values (spectrogram values) associated with higher MAXSHAP values suggest that the presence of spectra features at these wavelengths are associated with higher Cu % compared to the mean Cu % of the background samples (0.76 %). In contrast, bars that have lower feature values associated with higher MAXSHAP values, suggests that the absence of spectra features at these wavelengths are associated with high Cu % higher than the mean Cu % of the background samples.

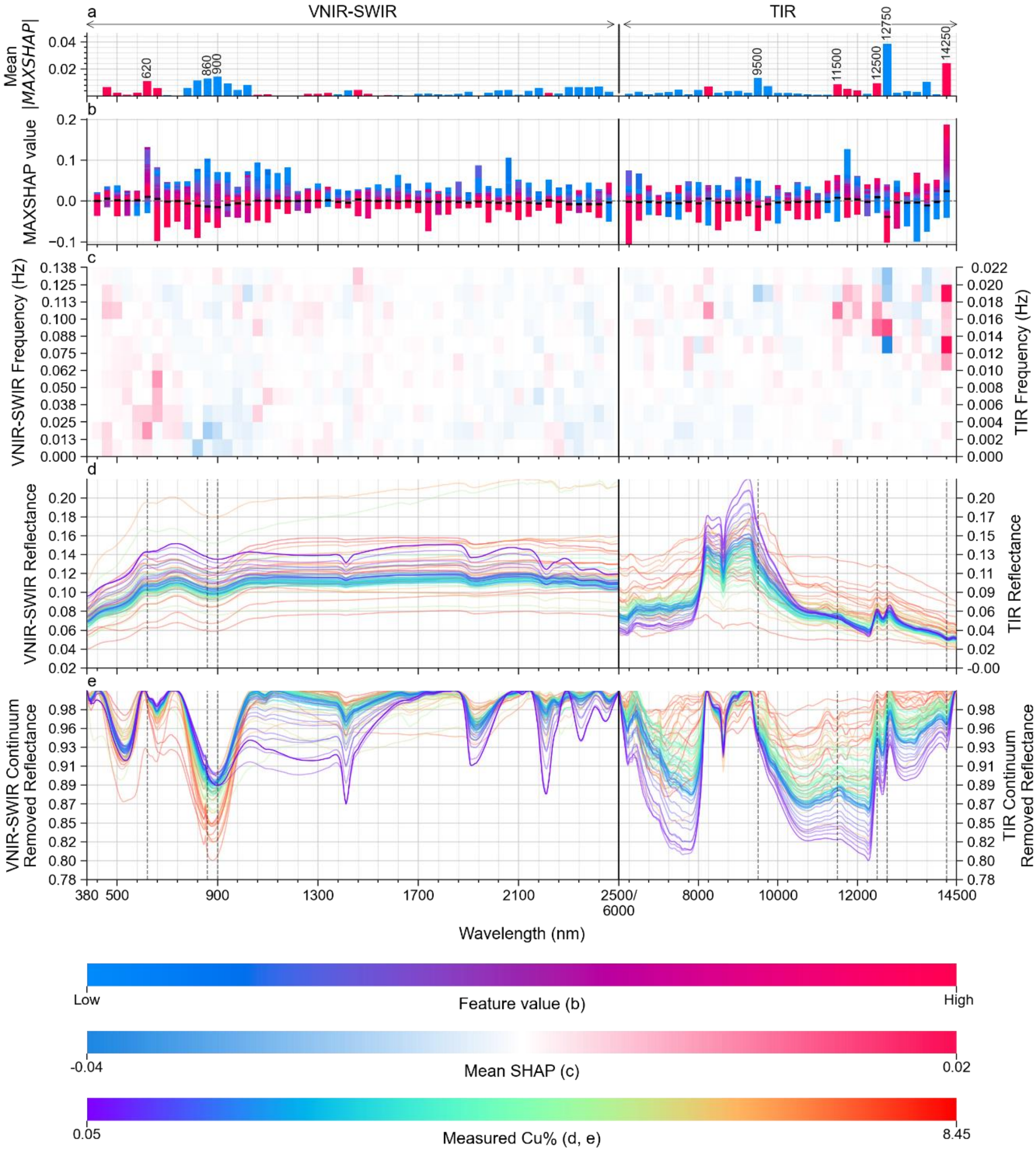


Fig. 15 SHAP analysis of the best performing model. (a) Mean MAXSHAP value per wavelength. Red bars indicate positive values and blue negative. (b) MAXSHAP values for all test samples at each wavelength. Bars are colored according to the feature values of each sample, where blueish colors are associated with low values and reddish colors with high values. The mean MAXSHAP value per wavelength is shown as a black horizontal solid line. (c) Mean SHAP value per input feature (spectrogram pixel). The pixels are colored according to their values, which if next to zero have lighter colors, with 0 corresponding to the white color, blueish colors are assigned to negative values and reddish to positive values. (d) Mean reflectance spectra obtained for samples with similar Cu % (i.e., difference of Cu % less than 0.1), covering the entire range of Cu %, colored by the measured Cu % (e) The same reflectance spectra with continuum removal. The position of the eight most significant wavelengths to the model output are shown in (a) and are highlighted as dashed lines in (d) and (e).

Figure 15c shows the mean SHAP value per spectrogram pixel (input feature) of the test samples. Pixels with higher absolute SHAP values (blueish/reddish pixels) has a greater impact on the model output. Therefore, spectra features located in the wavelength and frequency range of these pixels are the main drivers of the predictions made by the model.

The spectra of all the (training, validation, and test set) samples used in this work were grouped according to their Cu %, and the mean spectra of each group was calculated. The minimum and maximum Cu % of each group i were defined as $\min{(Cu\%)_i} = (i-1)0.1$ and $\max(\mathrm{Cu}\ \%)_i = i0.1$, respectively, for i varying from 1 to 85, Figure 15d shows these spectra colored by the mean Cu % of the group. Figure 15e shows the same spectra with the removal of continuum.

Considering the hyperspectral dataset used in this work, the 900, 860, and 620 nm wavelengths correspond to the absorption features of hematite/goethite. The first two wavelengths coincide with the main absorption feature of hematite/goethite between 850 and 1000 nm (Curtiss, 1985; Morris et al., 1985). Both of them had negative mean MAXSHAP values (Fig. 15a) and the feature values were inversely proportional to the MAXSHAP values (Fig. 15b). As can be seen in Figure 15c, the spectrogram pixels with higher absolute mean SHAP value at these wavelengths (blue pixels) were in the last two rows of the spectrogram, corresponding to spectral features with frequency lower than 0.025 Hz (> 40 nm wide). Therefore, the presence of wider absorption features at these wavelengths has a high impact on reducing the predicted Cu %. As can be seen in Figure 15e, the spectra of samples with higher Cu % tend to have the position of the 900 nm absorption feature at shorter wavelengths (~880nm), resembling hematite spectra. In contrast, spectra associated with lower Cu % have this feature positioned at longer wavelengths (~900nm), as in goethite spectra. Beyond that, the spectra associated with low Cu % present a wide absorption feature between 800 and 1700 nm, which is commonly associated with the presence of Fe^{2+} and magnetite (Hunt and Salisbury, 1970). Both the position of the 900 nm absorption feature and the presence of the wide absorption feature between 800 and 1700 nm, can be associated with changes in the spectrogram pixels indicated by the SHAP values (pixels with frequency lower than 0.025 Hz between 800 and 920 nm). Therefore, this indicates that rocks with high abundances of goethite and magnetite are associated with lower Cu % and the Cu mineralization is related to the predominance of hematite as the iron (hydro-)oxide phase. As reported in previous studies, hematite is strongly associated with ore zones in the ODD (see Section 2.1), its abundance increases toward the deposit center as magnetite decreases (K Ehrig et al., 2012; Mauger et al., 2016; Reynolds, 2001).

The 620 nm wavelength had a positive mean MAXSHAP value (Fig. 15a) and most of the feature values were proportional to the MAXSHAP values (Fig. 15b). Hematite also had an absorption feature near 620 nm (feature between 600 and 740 nm, centered at ~660 nm), which can be used to distinguish pure hematite spectra from spectra with mixtures of hematite and other minerals that have absorption features at the VNIR region, as goethite, limonite and siderite (Cudahy and Ramanaidou, 1997; Curtiss, 1985; Morris et al., 1985). This absorption feature is more prominent in pure hematite spectra, and thus it is indirectly related to hematite abundance and Cu content.

In TIR the 9500, 12500 and 12750 nm wavelengths must be associated with quartz, which has diagnostic spectral features at these regions (Spitzer and Kleinman, 1961). A broad absorption feature between 9500 and 12250 nm is typical of quartz spectra. The 9500 nm wavelength with a negative mean MAXSHAP value (Fig. 15a), and the feature values were inversely proportional to the MAXSHAP values (Fig. 15b). Therefore, the presence of the quartz absorption feature between 9500 and 12250 nm can be related with changes in the spectrogram pixels at 9500 nm, where samples with higher quartz abundance have higher pixel values at this wavelength and lower Cu %. Near 12500 and 12750 nm, quartz has two characteristic peaks. The high impact of these two wavelengths to the model output must also be related to the abundance of quartz in the samples. As can be seen in Figures 15d and 15e, samples with low Cu % have deeper absorption features between 9500 and 12250 nm, with the onset (near 9500 nm) better defined. Also, the peaks near 12500 and 12750 nm are more prominent in samples with low Cu %. These observations are consistent with the mineral assemblage of the Cu mineralization at ODD, which is composed predominantly of hematite (> 95%; Reynolds, 2001), thus having a low abundance of quartz. It is important to note that, although the Cu mineralization at ODD is associated with a hematite-quartz breccia, quartz is more abundant in low Cu rocks.

Carbonates have diagnostic spectral features near 11500 and 14250 nm. Siderite, which is associated with the Cu mineralization at ODD (K Ehrig et al., 2012; Reynolds, 2001), has a characteristic peak near 11500 nm and a less prominent trough-peak feature near 14250 nm (Green and Schodlok, 2016; Lane and Christensen, 1997). Both these wavelengths had a positive mean MAXSHAP value (Fig. 15a), and the feature values are proportional to the MAXSHAP values (Fig. 15b). As can be seen in Figure 15e, the spectra associated with higher Cu % had a peak near 11500 nm and the trough near 14250 nm was more symmetric than in the low Cu % spectra. Therefore, the spectral features of siderite must be related to changes in the spectrogram pixels at these wavelengths, where samples with siderite features had high pixel values and high Cu %. Although siderite is also associated with rocks with low Cu content as reported in previous works (K Ehrig et al., 2012; Reynolds, 2001), the minor amount of quartz in high Cu samples and, thus, the weaker quartz spectra features contribute to the enhancement of the siderite spectra features.

6.5 Conclusions

This study examined the use of DL algorithms for the estimation of Cu content based on drill-cores hyperspectral data. An open-source dataset of the ODD obtained from the AuScope database (http://portal.auscope.org/) was used for training two CNN models. The dataset was composed of Cu concentrations and hyperspectral data in VNIR–SWIR–TIR collected by the HyLogger system. Some data preprocessing strategies were tested, such as continuum removal and spectrogram transformation. In addition, we investigated which spectral range (VNIR–SWIR, TIR, or VNIR-SWIR-TIR) is best suited for addressing this problem. Results showed that the best performing models were those trained with the VNIR–SWIR–TIR spectrograms, generated from the reflectance spectra without continuum removal. The proposed approach can estimate Cu % in drill-cores scanned by the HyLogger system and it allows the upscaling of the Cu % information obtained by geochemical assays to the spatial resolution of the hyperspectral data (centimetric scale).

Results show that the proposed approach can be used effectively to estimate Cu concentrations along drill-cores at a centimetric spatial resolution. Commonly, the methods used in the literature to correlate geochemical and hyperspectral data need prior extraction of spectral features, which is usually a time-consuming task performed only by specialists (T. J. Cudahy et al., 2009; Dalm et al., 2017; Ducart et al., 2016; Haest and Cudahy, 2012; Prado et al., 2016; J. T. Qiu et al., 2017; Silversides and Murphy, 2017). In contrast, the only data pre-treatment needed for the application of the proposed method is to transform the reflectance spectra into spectrogram representation, which is done automatically by the algorithm. Moreover, the experimental results show that the best performing model provides an accurate tool to distinguish between low grade and high grade ore zones, which can be used to assist and support decisions during mining and mineral exploration. By mapping the ore grade in detail, the method (a) supports the selection of zones of interest where more detailed analyses are appropriate, (b) reduces the number of samples needed to characterize and identify the ore zones, and (c) assists in the estimation of the volume with commercially viable ore. In addition, the developed model can be used in autonomous sensor-based sorting systems to discriminate ore and waste. Nevertheless, these findings need further verification by testing the approach using other DL methods, and in other regions with different datasets.

Neural networks are changing many industries, increasing their performance, and reducing operational costs, making their activities more sustainable. This paper shows how neural networks has the potential to change the way we model drill-core hyperspectral data, conduct drill-core geochemistry surveys, and distinguish ore and waste during mining operation.

7 CLUSTERING OF HYPERSPECTRAL DRILL CORE MEASUREMENTS USING DEEP AUTOENCODERS AND SELF-ORGANIZING MAPS

7.1 Introduction

During exploration and mining of mineral resources, the identification of lithological and alteration boundaries is critical to managing operations. This information is necessary for the understanding of the mineral system and can be used from planning to the control of various process. For example, it can be used to estimate the continuity of a mineralized zone during exploration, or to control mill feed and speed during mining. Acquiring this information is usually done visually by geologists and mineralogists, supported by laboratory tests and sensor measurements.

Hyperspectral drill core sensing systems are noninvasive and nondestructive tools capable to characterize in detail (centimetric to millimetric spatial resolution) the mineral assemblages along drill-cores, assisting the identification of lithological and alteration boundaries. The use of these sensors to support the description of drill-cores is becoming more frequent due to their cost-effectiveness (Calvin and Pace, 2016; Kruse, 1996; Kruse et al., 2012; Littlefield et al., 2012; Tappert et al., 2013; Taylor, 2000). However, the high spatial and spectral resolution of these systems led to datasets comprised of millions of reflectance spectra, each one with hundreds of features (spectral bands). Due to the large volume and complexity of these datasets, the development of workflows to automate the extraction of the desired information from the reflectance spectra, such as the lithological and alteration boundaries, is of extreme importance.

The methods used for mineral classification of drill holes by means of hyperspectral data can vary from algorithms that utilize libraries of spectral signatures, to those that make use of specific absorption features for a given mineral target. The formers usually are based on supervised classification algorithms which measure similarity between acquired spectra and reference spectra. These algorithms were improved by researchers from traditional algorithms such Euclidean distance (Gower, 1985), spectral angle mapper (Kruse et al., 1993), and spectral information divergence (Chang, 1999), to algorithms based on machine learning, such as support vector machines (Melgani and Bruzzone, 2004), random forest (Acosta et al., 2020), artificial neural networks (Adep et al., 2017; Paya et al., 1997), and convolutional neural networks (Liu et al., 2021). The limitation of these algorithms is that the results are highly dependent on the quality of the spectra library used. That is, the spectral libraries must reflect the diversity of minerals on the analysed samples, as well as the variability of mineral mixtures on them, to make the model perform well (Davis et al., 1978). The collection of such libraries is difficult, expensive, and time-consuming (Li et al., 2009). Methods that make use of specific absorption features to identify a given mineral are based on the acquisition of spectra features parameters, which has as downside the need for constant interaction with specialists to interpret the spectra and fine-tune the acquisition.

Unsupervised classification algorithms, known as clustering algorithms, such as k-means clustering (Hartigan and Wong, 1979; Steinhaus and others, 1956) and agglomerative clustering (Dubes and Jain, 1976; Jain et al., 1999), also has been used for mineral classification of drill-cores (Guo et al., 2013; Ren et al., 2019; Rodger et al., 2021). These methods group the data such that similar spectra are within one cluster, and dissimilar ones are assigned to different clusters. The main advantage of these methods over supervised methods is that they do not require labeled data for training, such as spectral libraries. However, these models are sensitive to high-dimensional data, and in general fail to find meaningful clusters in such datasets (Assent, 2012; Beyer et al., 1999). Beyond that, the computational cost of these algorithms increases fast with the increase of the dataset. Hierarchical clustering algorithms for example, such as agglomerative clustering, in general require memory space of the order of $O(N^2)$, in which N is the number of records in the dataset (Xu and Wunsch, 2005). A common approach to deal with that is to reduce the dimensionality of the dataset before clustering. Algorithms like the Principal Component Analysis (PCA) (Pearson, 1901), diffusion maps (du Plessis et al., 2009) or Minimum Noise Fraction (MNF) (Green et al., 1988) have been traditionally applied to reduce data dimensionality.

Recently, works in computer sciences have shown that deep neural networks such as Deep Autoencoders (DAEs) (Kingma and Welling, 2013; Lecun et al., 2015) and variational autoencoders, have been used in combination Self-Organizing Map (SOM) (Kohonen, 2012) to substantially increase the performance of clustering methods (Forest et al., 2019; Manduchi et al., 2019; Tao et al., 2018). Both, autoencoders and SOM can be considered dimensionality reduction techniques that are based on neural networks, which has as advantage the ability of neural networks to identify complex patterns on high dimensionality data. Although there are some studies in the literature that uses DAE and SOM in combination with clustering algorithms for hyperspectral data classification, most of them are for remote sensing applications. Goncalves et al., (2008), proposed an unsupervised method for classifying remotely sensed images using SOM and agglomerative clustering. Some authors proposed the use of DAE for clustering hyperspectral images (Gao et al., 2021; Zhang et al., 2022). However, we are not aware of any work that has evaluated the use of both DAE and SOM for clustering drill hole hyperspectral data.

In this study we propose a novel workflow for clustering high-dimensional drill hole hyperspectral data, which uses both DAE and SOM for dimensionality reduction of the dataset before clustering. Our model uses the DAE network to reduce the dimensionality of the hyperspectral dataset in the feature space, encoding the input spectra to a lower number of features. Then a Maximum Data Variance Power Transform (MDVPT) is applied to the encoded features to enhance their variance. The dimensionality of the transformed encoded features is then reduced in the samples space using the SOM algorithm. The SOM results are finally clustered using the agglomerative clustering algorithm. In addition, a voting strategy is used to aggregate the clusters along the drill holes, smoothing the results. Hyperspectral data collected from drill holes of the Prominent Hill IOCG deposit, South Australia, were used to evaluate the proposed workflow. The substantial reduction in dimensionality, and therefore in the size of the dataset, provided by the proposed workflow allows the clustering of large volume hyperspectral datasets (with high spatial and spectral resolution) to be performed effectively in permissive times. The obtained clusters map regions in the drill holes with same mineralogical composition, and thus can be used to identify lithological and alteration boundaries. Furthermore, the distance between the clusters centers and labeled spectra from the JPL spectra library, were used to predict the predominant minerals in each cluster.

7.2 Study Area

The dataset used in this study were collected from drill cores of the Prominent Hill iron oxide copper-gold (IOCG) deposit in central South Australia. The publicly available information provided by the Australian government about these drill cores contains hyperspectral and geochemical data, accompanying downhole lithology log. The existence of these three types of information in the drill cores provide an ideal case study as the clustering results can be compared with lithology log and ore grade (geochemical data), providing a way to measure the performance of the clustering workflow, and highlight lithologies and alteration zones that were not discriminated during visual drill core logging.

7.2.1 Geology and Cu-Au mineralization

The Prominent Hill deposit is one of the numerous IOCG deposits (e.g. Olympic Dam, Carrapateena) and prospects (e.g. Oak Dam East, Emmie Bluff) hosted by the Olympic IOCG province, located in the eastern and northern Archean to Mesoproterozoic Gawler Craton of South Australia (Fig. 1). The deposit is hosted by metasedimentary rocks of the Wallaroo Group, deposited between ~1760 and ~1730 Ma in the Mountain Woods domain, filling a rifting-related extensional system (Chalmers, 2007; Conor, 1995; Cowley et al., 2003; Freeman and Tomkinson, 2010). As other deposits in the province, the Cu-Au mineralization and hematite-rich alteration is temporally and spatially associated with a major tectonothermal and magmatic event between 1.60-1.58 Ga which formed the bimodal Gawler Silicic Large Igneous Province (SLIP), composed of felsic and mafic volcanic rocks of the Gawler Range Volcanics and granitoid intrusions of the Hitalba Suite (Allen et al., 2008; Belperio et al., 2007; Schlegel and Heinrich, 2015). Mesozoic to paleozoic sedimentary rocks overlies the mineralized Paleo- to Mesoproterozoic rocks of the Mountain Woods Domain, concealing the basement rocks by basin cover sediments having up to 400 m thickness (Belperio et al., 2007; Drexel and Preiss, 1995).

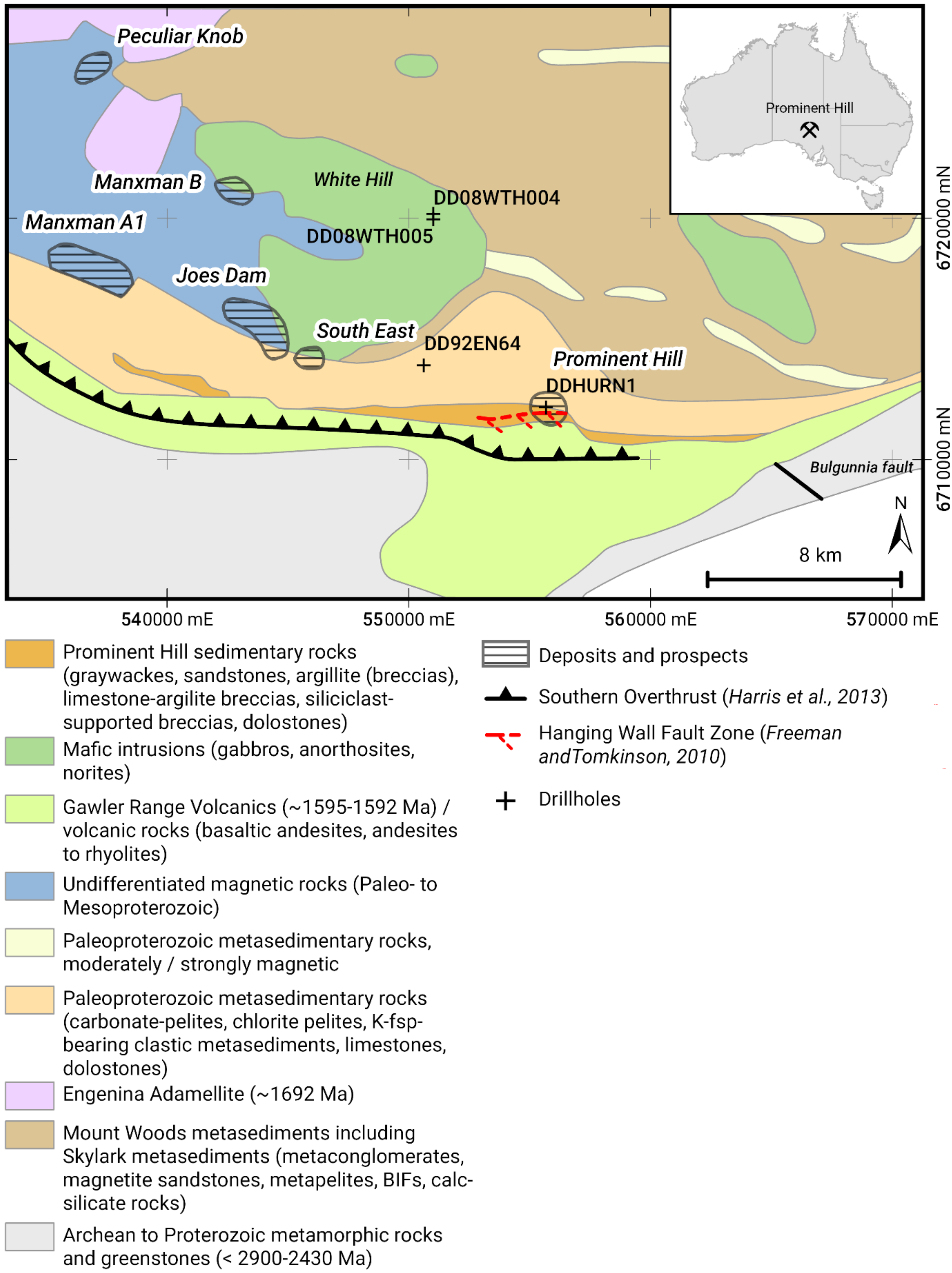

Peculiar Knob
Manxman B
Manxman A1
White Hill
DD08WTH004
DD08WTH005
Joes Dam
South East
DD92EN64
Prominent Hill
DDHURN1
Bulgunnia fault
N
8 km
Prominent Hill
6720000 mN
6710000 mN
540000 mE
550000 mE
560000 mE
570000 mE
Prominent Hill sedimentary rocks (graywackes, sandstones, argillite (breccias), limestone-argilite breccias, siliciclast-supported breccias, dolostones)
Mafic intrusions (gabbros, anorthosites, norites)
Gawler Range Volcanics (~1595-1592 Ma) / volcanic rocks (basaltic andesites, andesites to rhyolites)
Undifferentiated magnetic rocks (Paleo- to Mesoproterozoic)
Paleoproterozoic metasedimentary rocks, moderately / strongly magnetic
Paleoproterozoic metasedimentary rocks (carbonate-pelites, chlorite pelites, K-fsp-bearing clastic metasediments, limestones, dolostones)
Engenina Adamellite (~1692 Ma)
Mount Woods metasediments including Skylark metasediments (metaconglomerates, magnetite sandstones, metapelites, BIFs, calc-silicate rocks)
Archean to Proterozoic metamorphic rocks and greenstones (< 2900-2430 Ma)
Deposits and prospects
Southern Overthrust (Harris et al., 2013)
Hanging Wall Fault Zone (Freeman andTomkinson, 2010)
Drillholes

Fig. 1 Interpreted subsurface geology of the southern Mount Woods domain, modified from (Schlegel and Heinrich, 2015)

Fig 1 Interpreted subsurface geology of the southern Mount Woods domain, modified from (Schlegel and Heinrich, 2015). Basement rocks at the Mountain Woods Domain include mafic, intermediate and felsic volcanics and intrusives, and metasedimentary rocks, with metamorphism ranging from greenschist to granulite facies (Belperio et al., 2007; Belperio and Freeman, 2004; Forbes et al., 2012, 2011; Skirrow et al., 2002). Argillaceous, calcareous and siliciclastic metasediments, and banded iron formations from the Skylark metasedimentary rocks are the dominant rocks in the northern and central parts of the Mount Woods domain (Betts et al., 2003; Chalmers, 2007; Freeman and Tomkinson, 2010). These metasedimentary rocks are metamorphosed and intruded by the synorogenic Engenina Adamelite at 1692 Ma, during the Kimban oregeny (1730-1690 Ma) (Betts et al., 2003; Daly et al., 1998; Fanning, 1997; Freeman and Tomkinson, 2010). Gabbroic rocks of the White Hill mafic igneous complex (~1585 Ma), contemporaneous to the Hiltaba Suite, also intrude these metasedimentary rocks to the north of the Mountain Woods Domain (Chalmers, 2007; Freeman and Tomkinson, 2010). At north and northeast of Prominent Hill, metasedimentary rocks from the Wallaroo Group (Paleoproterozoic metasedimentary rocks in Fig. 1), metamorphosed to amphibolite facies, are separated from unmetamorphosed sedimentary rocks (Prominent Hill sedimentary rocks in Fig. 1) of potentially 1590 to 1580 Ma depositional age by a reverse fault (hanging wall fault zone) (Belperio et al., 2007; Freeman and Tomkinson, 2010). Basaltic to andesitic volcanic rocks of the Gawler Range Volcanics occur south of Prominent Hill (Carter et al., 2003; Harris et al., 2013).

The hanging wall fault zone is represented by a E-W-trending, variable steeply N-dipping chlorite breccia (Schlegel and Heinrich, 2015). North of the fault zone, in the hanging wall, the metasedimentary rocks are intruded by undeformed granitoids (~1585 Ma) (Belperio et al., 2007). In the footwall, at south of the fault zone, the sedimentary rocks are structurally overlaying mafic to intermediate composition lower Gawler Range Volcanics (Belperio et al., 2007). The Cu and Au mineralization at Prominent Hill are hosted by these sedimentary and volcanic rocks in the footwall and are characterized by the overprinting of earlier hydrothermal hematite replacement zones. The sedimentary host rocks consist of interbedded and brecciated, argillaceous, calcareous, and siliciclastic sediments. Breccias are strata bound and exhibit a variable brecciation intensity, which resulted in the formation of both clast- and matrix-supported breccia sheets. The hematite breccias, hosting the Cu mineralization, are predominately altered calcareous sedimentary rocks, formed by the hematite replacement of calcareous and siliciclastic breccia and rock components. Beyond hematite, the hydrothermal alteration includes carbonates (i.e., siderite, ankerite, and dolomite), sericite (likely phengite), chlorite, fluorite, barite, quartz, fluorapatite, REE minerals (including monazite), uraninite, coffinite, and Cu-(Fe) sulfides (Schlegel and Heinrich, 2015).

The Prominent Hill hematite breccia body is represented by two types of hematite breccia. The hematite-quartz breccia result of hematite and quartz replacement, and the hematite-aluminosilicate breccia, result of hematite, chlorite, and sericite replacement (Schlegel and Heinrich, 2015). Most of the Au mineralization is within the margins of the hematite-quartz alteration zone, and within the transition to the surrounding hematite- aluminosilicate breccia (Belperio and Freeman, 2004; Schlegel and Heinrich, 2015). The hematite-quartz-altered rocks and breccias display only week Cu mineralization. The hematite-chlorite-sericite alteration is associated with Cu mineralization in the hematite-aluminosilicate breccia matrix. Cu-(Fe) sulfide minerals including chalcocite, digenite, bornite, idaite, and chalcopyrite are abundant in the breccia matrix. Locally, chalcopyrite-pyrite mineralization is spatially associated with intense hematite-fluorite-barite alteration. Typically, the high-grade Cu mineralization is confined to areas of the breccia matrix that contain fine- to coarse-grained crystalline hematite and visible sericite (Schlegel and Heinrich, 2015).

7.3 Methods

7.3.1 Data

The datasets used in this study include point reflectance spectra, geochemical data, and lithology logs, collected from 4 drill cores (with average length of ~500 m) from the Prominent Hill deposit and surrounding areas. The spectral data were provided by the Australian National Virtual Core Library, and the geochemical data and lithology logs by the South Australian Resources Information Gateway (SARIG). All this data is publicly available and can be downloaded through AuScope's Discovery Portal (http://portal.auscope.org/).

The spectral data were collected by the CSIRO Hylogger-3 system: an automated system for drill core logging that combines a X-Y translation table and a reflectance spectrometer with sensor at visible, near infrared (VNIR), shortwave infrared (SWIR), and thermal (TIR) infrared spectral region. The spectral data were collected continuously throughout the drill cores at a sampling interval of 8 mm. The field of view of the spectrometers was 10 mm across the track (across the drill hole extension) and 18 mm along the track (along the drill hole extension) (Schodlok et al., 2016). The VNIR–SWIR data were acquired for a wavelength range of 380–2500 nm (4 nm bandwidth; 531 bands) and the TIR data for a wavelength range of 6000–14500 nm (25 nm bandwidth; 341 bands). These reflectance spectra were used as inputs to the proposed clustering workflow.

From the 4 drill cores used in this study, two are located north of the Prominent Hill deposit (DD08WTH004 and DD08WTH005), intercepting rocks from the White Hill mafic igneous complex, one is located west of the deposit (DD92EN64), intercepting metasedimentary rocks of the Wallaroo Group, and the last one is located at the Prominent Hill deposit (DDHURN1), intercepting the mineralized breccias (Fig. 1).

7.3.2 Hyperspectral Data Clustering Workflow

The proposed workflow is carried out using a two-level approach (Fig. 2). The dataset is first passed through a dimensionality reduction stage where a deep autoencoder network and a SOM were trained to reduce the dimensionality of the dataset both in the feature and sample space, respectively, and then, the transformed spectra were partitioned into clusters using the agglomerative clustering method and a segmentation strategy. Reducing the dimensionality of the dataset before clustering is a well-known approach, due to the considerable decrease in computational load, making it possible to cluster large data sets in a reasonable time. Beyond that, dimensionality reduction methods generally increase the signal to noise ratio of the dataset, decreasing the importance of noisy features and/or samples to the clustering result (Affeldt et al., 2020; Banijamali and Ghodsi, 2017; Goncalves et al., 2008; Tian et al., 2014; Vesanto and Alhoniemi, 2000).

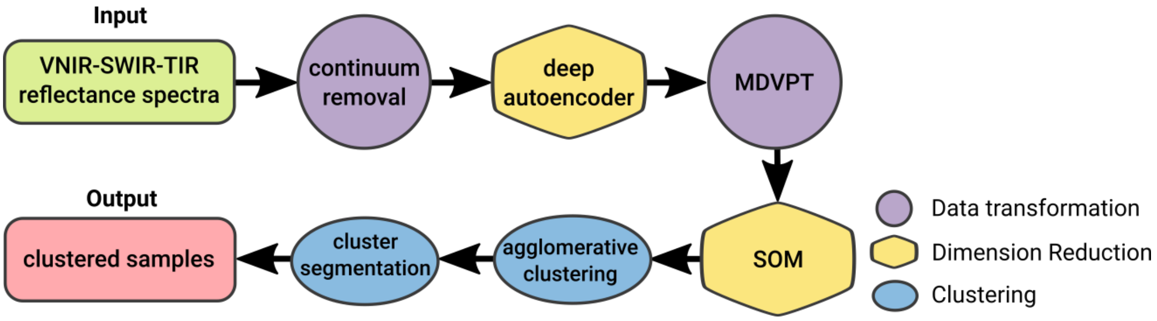


Fig. 2 Proposed hyperspectral data clustering workflow.

In the dimensionality reduction stage, before training the deep autoencoder, the continuum was removed from the reflectance spectra. The continuum removal decreases the effect of less prominent features in the learning process, which normally has lower signal to noise ratios. Furthermore, this step ensures that the autoencoder input data is scaled between the interval 0↔1. Data scaling is a common approach adopted when training neural networks, as deep autoencoders, to avoid instability on training. The deep autoencoder network was then trained with the continuum removed spectra. After trained, the network is used to encode the continuum removed reflectance spectra from 872 features (VNIR-SWIR and TIR spectral bands) to 16 features. The contrast of the encoded data is then enhanced by applying the maximum data variance power transform function before training the SOM. This transformation increases the standard deviation of the encoded features, assisting in training the SOM. After this, the data is used to train the SOM, which reduces the data into 2500 vectors with 16 features each.

The distance between the vectors returned by the SOM were then used to cluster the data using the agglomerative clustering method. A segmentation strategy is then employed to aggregate the clusters along the drillholes, smoothing the clustering results.

7.3.3 Dataset Verification and Correction

The hyperspectral data were provided as normalized reflectance spectra, consisting of 220,160 spectra. It is expected that reflectance values fall in the range [0,1], therefore, spectra with reflectance values greater than 1 were discarded (797; 0.13% of total), and spectra with reflectance values lower than 0 were clamped (19,929; 9% of total), i.e. negative values close to zero (<0.01) were replaced with 0 to keep the reflectance values in the range [0, 1].

7.3.4 Continuum removal

Normally, during spectral analysis, the continuum is removed (Clark and Roush, 1984) to isolate absorption features of the spectrum from other effects, such as level changes and slopes generated by other materials. These effects can cause shifts of the local minimum in the spectra and can significantly reduce absorption depths. When the continuum is removed, the minima show more stable positions. In addition, the effects of illumination geometry, as well as the effects of contaminants and grain size are reduced because the continuum removal normalizes the spectra (Clark et al., 2003).

The continuum removal algorithm from the pysptools python's package (Therien, 2018) was used. This algorithm computes the convex hull of the signal and removes it by hull quotient. Example results of continuum removal are shown in Figure 3.

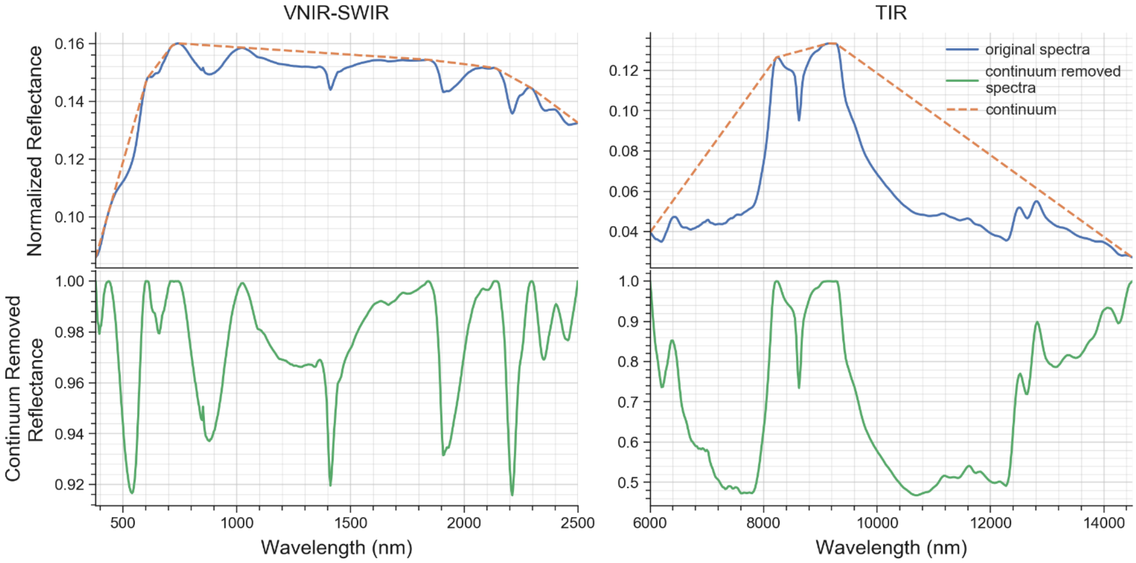


Fig. 3 Examples of continuum removal. Top panels: original spectra as solid lines with respective convex hull continuum as dashed lines. Bottom panels: continuum removed spectra. Left panels: VNIR–SWIR data. Right panels: TIR data.

7.3.5 Deep Autoencoder Network

An autoencoder network is a type of feedforward neural network designed to reproduce the input data. The network is trained using the same data as input and target, and thus the size of the output layer is always the same as the input layer. Autoencoders can be used for dimensionality reduction when the number of neurons in the hidden layers is smaller than the input/output size, resulting in a bottleneck structure which can be visualized as containing two parts an encoder and a decoder. The encoder, which can be represented by the function $e = f(x)$ where x is the input and e is the encoded input, is used to reduce the size of the input, decreasing the number of neurons in each layer. On the other hand, the decoder produces a reconstruction $r = g(e)$ of the encoder output, increasing the number of neurons in each layer until it reaches the size of the input data. After trained the network learns to encode the input to a lower dimension and decode it back returning an approximate copy of the input. This way, the model is forced to prioritize which aspects of the input should be copied, leading it to capture the most salient features of the training data (Lecun et al., 2015).

Autoencoders may be trained with all the same techniques as feedforward neural networks, typically minibatch gradient descent following gradients computed by back-propagation. The network is trained by minimizing the loss function

$$L(x, g(f(x)))$$

where L is a loss function penalizing $g(f(x))$ for being dissimilar from x, such as the mean squared error.

Autoencoders architectures can have a single hidden layer or multiple hidden layers. When these networks have more than one hidden layer, they are called deep autoencoders. Using deep networks offers many advantages, as reducing the computational cost of representing some functions and decreasing the amount of training data needed to learn some function. Moreover, experimental results show that deep autoencoders yield much better compression than corresponding shallow or linear autoencoders (Hinton and Salakhutdinov, 2006). Deep autoencoders with nonlinear encoder functions f and nonlinear decoder functions g can learn more powerful nonlinear generalization of the training data than Principal Component Analisis (PCA) (Lecun et al., 2015). For this reason, we choose deep autoencoders for dimensionality reduction of the hyperspectral data.

7.3.5.1 Network architecture

The developed deep autoencoder consists of a stack of fully connected layers followed by rectification non-linearity (ReLU) (Krizhevsky et al., 2012) or linear activation functions. Developing deep neural networks is highly interactive, as the number of layers and neurons on each layer, as well as the activation functions, regularization constant, and learning rate are some of the hyperparameters that need to be defined to design the model. In this section, we present the network with the best performing combination of hyperparameters.

The encoder network receives as input a fixed size 1x872 vector, build from the continuum removed hyperspectral data, where 872 is the number of spectral bands of the concatenated VNIR-SWIR + TIR spectra. The input vector is than passed through a stack of four fully connected layers having 256, 128, 64, 32 neurons respectively, all of them followed by a ReLU activation function. The output of the encoder is a fully connected layer with 16 neurons and a linear activation function.

The output of the encoder network is provided as input to the decoder network. The encoded input in than passed through a stack of four fully connected layers having 32, 64, 128, and 256 neurons respectively, all of them followed by a ReLU activation function. The output of the decoder is a fully connected layer with 872 neurons and a linear activation function.

7.3.5.2 Training the network

The training was conducted using Adam optimizer (Kingma and Ba, 2014) with mini-batch gradient descent and root mean squared error (RMSE) as the loss function. The batch size was set to 256. The fully connected layers had weight decay regularization (Ng, 2004), with L2 penalty multiplier set to 0.00001. The learning rate was initially set to 10-3. This initial value was set to decrease by a factor of 10 when the validation RMSE stopped improving (i.e., the validation RMSE did not drop 10-4 units during 10 epochs). Early stopping was also used, and the training was set to stop when the validation RMSE did not drop 10-5 units during 10 epochs. The weights in both networks were initialized using the random uniform initialization procedure of Glorot and Bengio (2010) also known as Xavier uniform initialization. The model was implemented in Python programming language using the PyTorch (Paszke et al., 2019) package.

7.3.6 Self-organizing Maps

Self-organizing maps (SOM) (Kohonen, 2012) is a type of artificial neural network, which is designed to map the input samples to vectors with the same topological structure of the input space, where the number of vectors is less than the number of input samples, and each vector has the same number of features as the input data. Therefore, SOM can also be considered a dimensionality reduction method, but unlike PCA and autoencoders which are used to reduce the dimensionality in feature space, SOM reduce the dimensionality in sample space and is thus also considered a clustering method. As each vector returned by the SOM can be considered a cluster.

SOM are composed of two layers, an input layer, and an output layer. Different from others artificial neural networks, SOM are not trained using backpropagation and stochastic gradient descend, this network uses competitive learning to update its weights. During training the network modify the weights of the output vectors to reposition them to match the distribution and structure of the original input data using measures of vector similarity. After training, these vectors are known as “best matching units” (BMUs). The BMUs are then projected into a two-dimensional grid and the Euclidean distance between the BMUs are computed to produce the “unified distance matrix” (U-Matrix).

Training the network: After encoding the continuum removed spectra, and apply the MDVP transformation, the transformed encoded spectra were used as input for the SOM model. To train the network we used a Python implementation of SOM named quicksom (Mallet et al., 2021). After some experimentation, a map size of 50 x 50 was chosen for this exploratory study. The weights were randomly initialized. A batch size of 256 was used and the network was trained for a total of 30 epochs.

7.3.7 Agglomerative Clustering

Clustering methods aim to partition a data set into a set of clusters $Q_i, i = 1, \cdots, N$. The main two method used are hierarchical and partitive methods. Partitive clustering algorithms produce one partition of the data set into N clusters, generally by optimizing an objective function. The number of clusters is usually predefined, but it can also be part of the objective function. Hierarchical clustering algorithms produce a nested series of partitions, which can be generated using two approaches, agglomerative or divisive. In the agglomerative method the algorithm is initialized with each sample being a distinct cluster, and successively merges clusters together until a stopping criterion is satisfied. In the other hand, a divisive approach begins with all samples in a single cluster and performs splitting until a stopping criterion is met. A commonly used stopping criterion is the number of clusters N, i.e., when the number of clusters N is reached the algorithm stops merging or splitting the clusters (Jain et al., 1999).

Partitive methods are considered better than hierarchical ones, because they do not depend on previously found clusters. However, partitive methods assumes that the clusters have a predefined shape. One of the most used partitive methods, k-means (Steinhaus and others, 1956), tries to find spherical clusters. This assumption is generally not in line with the intended segmentation of the data. Among the hierarchical methods, agglomerative clustering techniques are more understandable and more commonly used than divisive methods. Considering this, after some experimentation we found that the agglomerative clustering technique was the most suitable for clustering the SOM results in this work.

In agglomerative clustering, a distance function D is used to determine the pair of distinct clusters $\left(C_i, C_j\right)$ that will be merged to form a new cluster C_k, i.e. $C_k = C_i \cup C_j$, so that:

$$D\left(C_i, C_j\right) = \min_{\substack{1 \le m,l \le N \\ m \ne l}} D(C_m, C_l)$$

after this the algorithm calculate the new distance, $D(C_k, C_l)$, between the new cluster C_k and the others that are left. These steps are repeated until a stopping criterion is satisfied or all the samples are in a single cluster.
In this work we used the Euclidean distance as distance metric and the average linkage (Sokal, 1958) as the distance function D, which defines the new distance $D(C_k, C_l)$, as the average distance between samples in C_k and C_l, that is:

$$D(C_k, C_l) = \sum_{ij} \frac{D(C_k[i], C_l[j])}{|C_k||C_l|}$$

where $x[i]$ represent a sample i in cluster x, and $|x|$ represents the number of samples in cluster x.

In this work, a fixed number of clusters K was defined as the stopping criterion for the agglomerative clustering algorithm. The number of clusters K was obtained by applying a peak detection function to the U-Matrix values computed by the SOM multiplied by -1. Peaks are local maxima separated by at least a minimum distance d, in pixels. Therefore, the obtained peaks represent BMUs or U-Matrix pixels, where the distance between the samples and the BMU are locally the minimum ones. The *peak_local_max* function provided in the scikit-image library for Python (van der Walt et al., 2014) was used as the peak detection function, and the number of clusters K was set to the number of peaks obtained by using a minimum distance d of 2 pixels.

7.3.8 Cluster Segmentation

Generally, conventional clustering methods, which do not consider the spatial correlation of samples, as the one adopted in this work, are affected by salt-and-pepper noise (Zhai et al., 2021), mainly due to small variations in illumination intensity and the roughness of the analyzed surface, and thus cannot accurately model the spatial neighborhood of samples along the drillholes. Although the clustering results are consistent when viewed in feature space, the high frequency of cluster changes along the drillholes makes interpretation difficult and is usually not consistent with the expected geological variations.

The clusters produced by the agglomerative clustering algorithm were aggregated according to their location along the drillholes, enhancing the spatial homogeneity of the clustering result. For this we segmented each drillhole H at a regular interval d (~2 m), creating a unidimensional vector of depths $P_H = [p_1, p_2, \cdots, p_z]$, where p_1 is the minimum depth for drillhole H, p_z is the maximum depth for drillhole H, and $p_{i+1} - p_i = d$. Then, we applied the voting strategy:

$$V_k[p_i] = S_k[p_i]$$

where $V_k[p_i]$ is the voting for cluster k at depth p_i, and $S_k[p_i]$ is the number of samples assigned to cluster k between the depths p_s and p_e, defined as:

$$p_s = \begin{cases} p_i, & for\ i = 1 \\ p_i - d, & for\ i > 1 \end{cases}$$

$$p_e = \begin{cases} p_i + d, & for\ i < z \\ p_i, & for\ i = z \end{cases}$$

The wining cluster k' is then assigned to each depth p_i, according to the equation:

$$k'[p_i] = \max_{1 \leq k \leq N} S_k[p_i]$$

where N is the number of clusters. Finally, the wining clusters $k'[p_i]$ are assigned to the neighborhood samples of p_i, defined as the samples between the depths p_m and p_l, according to the equations:

$$p_m = \begin{cases} p_i, & for\ i = 1 \\ p_i - \frac{d}{2}, & for\ i > 1 \end{cases}$$

$$p_l = \begin{cases} p_i + \frac{d}{2}, & for\ i < z \\ p_i, & for\ i = z \end{cases}$$

7.4 Results and Discussion

Firstly, the clustering performance of the proposed workflows to the four drill cores was compared with the performance of simpler clustering strategies. The clustering performance was evaluated by comparing the results with the provided down hole lithology. Then, the results obtained by the proposed workflow were detailed. Lastly, the mineral assemblages of each cluster were estimated and compared with the alteration assemblages described at Prominent Hill deposit.

7.4.1 Comparison with other clustering strategies

The proposed workflow is characterized by a twostep dimensionality reduction stage, where the dataset passes through stacked DAE and SOM networks. The clustering performance of this strategy was compared with the performance of simpler strategies, which have a single dimensionality reduction step or directly cluster the dataset without reducing the dimensionality, been them: (1) encoding the continuum removed spectra with a DAE and then clustering; (2) training a SOM with the continuum removed spectra and then clustering; (3) directly clustering the continuum removed spectra. In all the experiments agglomerative clustering was used as the clustering method. The DAE and SOM networks are all trained with the same hyperparameters. The results are shown in Figure 4.

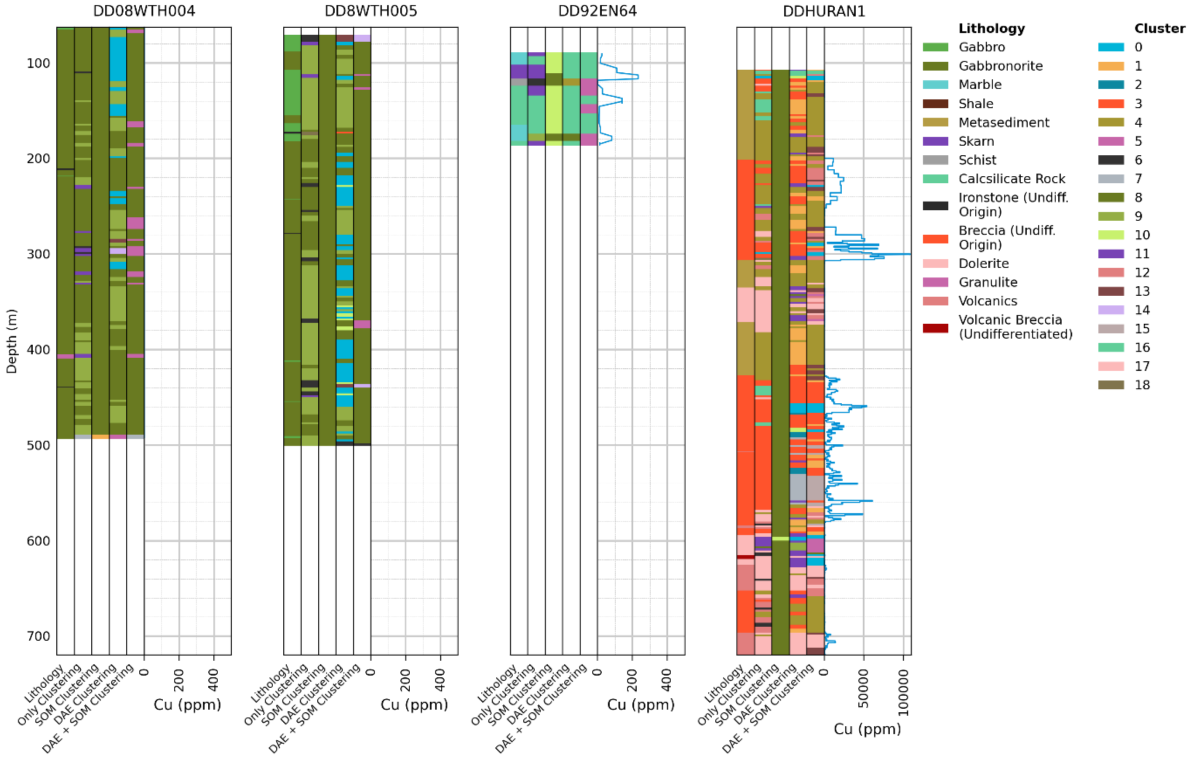


Fig. 4 Clustering results plotted with the logged lithology and Cu analysis for all the four drill cores.

All the clustering strategies, with exception of the second one (SOM clustering), were able to identify most of the lithologic boundaries. However, it can be observed that our proposed workflow (DAE + SOM clustering) results in smoother cluster distribution along the drill cores then the other strategies. The second strategy, which uses a SOM for dimensionality reduction before clustering, had by far the worst clustering performance. The clustering algorithm for this strategy assigned almost all the samples to cluster 8, resulting in a lack of correlation between the cluster distribution and the lithologic descriptions. These results indicate that the SOM network of this strategy failed to fit the data set adequately. The large number of features in the dataset makes training the SOM network difficult, leading to an under-fitting of the network. For this reason, we do not consider this strategy in the analysis bellow.

According to the logged lithologies, drill holes DD08WTH004 and DD08WTH005 are mainly composed of two lithologies, gabbronorite and gabbro, intercalated with thin layers of volcanic rocks and ironstones. As can be seen in Figure 4, clustering strategies other than the proposed workflow assigned six or more cluster labels along this drill holes, resulting in a more complex distribution of the labels, making it difficult to relate them with the described lithologies. The proposed workflow assigned just four cluster labels along this drill holes (same number of different lithologies in the logs), with most of the samples been assigned to one of them (cluster 8). When compared to the others, this cluster distribution is closer to the lithological descriptions provided, facilitating the correlation with the lithologies. Cluster 8 for examples, can be easily correlated to the gabbronorite lithology, which is extensively described along these drill holes. It is worth mentions that, in all the clustering strategies, the predominant cluster labels assigned to the samples in these drill holes (dark and light green in Fig. 4) were not assigned to samples in other drill holes. According to the lithological descriptions, these clusters represent the gabbro and gabbronorites, which were also not described in the other drill holes. This shows that all the clustering strategies were able to effectively differentiate these samples from the others.

Drill hole DD92EN64 intercepts calcsilicate rocks, marbles, skarns, and schists, according to the logs (Fig. 4). The performance between the different clustering strategies for this drill hole was similar. In the logs most of the samples of this drill hole were described as calcsilicate rocks. In the same way, all the clustering strategies assigned most of the samples of this drill hole to the same cluster (cluster 16). The lithological boundaries of this drill hole can be identified in most of the clustering results. Despite the clustering performance of the proposed workflow is not clearly better then the others in this drill hole, when comparing the distribution of cluster with the Cu analysis, the clusters obtained by the proposed workflow showed a better correlation with the Cu grade.

Lastly, the logs of drill hole DDHURAN1, which intercepts the mineralization at Prominent Hill deposit, are composed mainly of metasediments and undifferentiated breccias, with minor intercalations of dolerite, volcanic rocks and volcanic breccias (Fig. 4). Most clustering strategies correctly identified the lithological boundaries logged in this drill hole. However, in general the number of clusters are higher than the number of logged lithologies, resulting in a more complex distribution of classes. Despite that, some clusters obtained by the DAE + SOM clustering strategy showed a high correlation with the logged lithologies, as cluster 4 and the sedimentary rocks, cluster 3 and the undifferentiated breccias, and clusters 12 and 17 and the dolerite. Between 200 and 300 m approximately, mineralized undifferentiated breccias were logged. In this interval, none of the samples were assigned to cluster 3 by the proposed method. Cluster 3, is the cluster assigned to most of the samples of the mineralized interval bellow, also logged as undifferentiated breccia. Beyond that, the cluster distribution provided by the proposed workflow showed a better correlation with the Cu grades in this interval than the logged lithologies, where the presence of some clusters was correlated with high Cu grades, as clusters 0 and 12. Mineralized undifferentiated breccias were also logged between 420 and 600 m approximately. In this interval, the cluster distribution provided by the proposed workflow also showed a good correlation with the Cu grades. Furthermore, it can be noted that some clusters which occur in these mineralized intervals are strongly correlated with high Cu grades, as cluster 0 (~300m and ~460 m).

7.4.2 Clustering results of the proposed workflow

In this section, the clustering results obtained by the proposed workflow are evaluated. The U-matrix computed by the SOM network (Fig. 5a) shows the distance of nodes from their neighbors. Larger values (pixels in yellowish colors in Fig. 5a) imply larger spectral differences. The given U-matrix shows that the spectra of samples near to the nodes of the upper right are significantly separated from the rest. Most of this samples were assigned to cluster 8 (Fig. 5b), which has a high correlation with the gabbronorite logged in drill holes DD08WTH004 and DD08WTH005. Mafic igneous rocks have a distinct mineralogical assembly from sedimentary rocks and felsic igneous rocks. Therefore, it is expected that the spectra of these samples are considerably distinct from the others. It is also possible to observe other well-defined boundaries in the U-matrix (yellowish pixels Fig. 5a), showing that spectra of samples from drill holes DD92EN64 and DDHURAN1 present a relevant distinction. Most of these boundaries were preserved after clustering, as can be seen in Figure 5b, indicating good clustering performance.

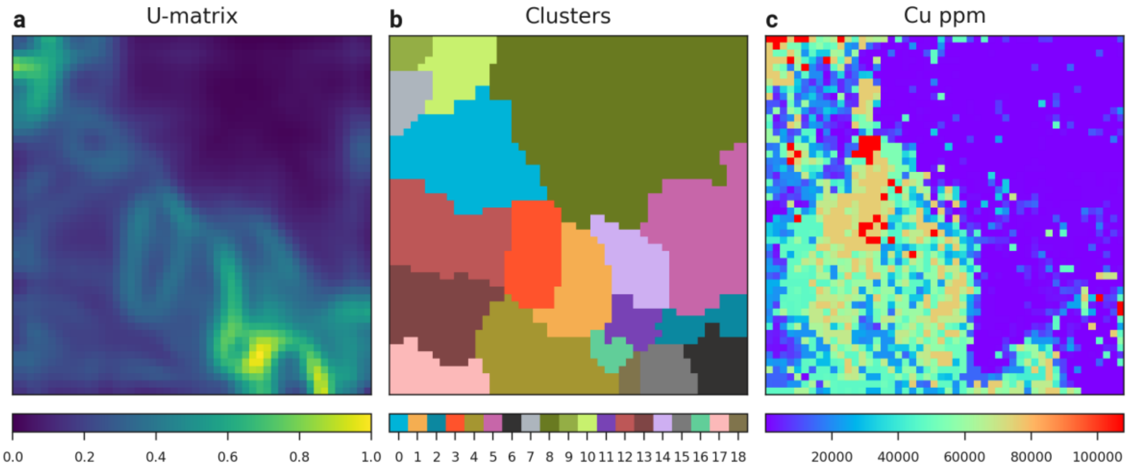


Fig. 5 SOM network results. (a) U-matrix. (b) Clusters assigned to each node of the U-matrix. (c) Maximum Cu analysis of samples assigned to each node of the U-matrix.

The maximum Cu analysis of the sample of each node of the U-matrix are plotted in Figure 5c. As can be seen, some regions exhibit a grouping of high Cu values, indicating that some spectra are strongly correlated with high Cu grades. Furthermore, the clusters were able to distinguish between non-mineralized and mineralized samples, as well as differentiating the higher-grade samples. Most of the non-mineralized samples were assigned to clusters 5, 6, 8, 11, 14, and 16, and most of the higher-grade samples were assigned to clusters 0, 1, 3, 9, 10, and 12.

To visualize the clusters distributions across all the samples, the first two principal components (PC1 and PC2) of the SOM network input dataset (continuum removed spectra encoded with DAE and transformed with MDVPT method) were computed using Principal Component Analysis (PCA) and used to visualize the multidimensional dataset in 2D (Fig. 6). The clusters have varied shapes and their distribution follows the major trends formed by the clouds of samples with similar lithologies (Fig. 6a and 6b). When the Cu analyses of the samples are plotted on the same chart (Fig. 6c), it is possible to notice that the mineralized samples are grouped in three well-defined point clouds. The first is located at the bottom of the chart, with an almost horizontal trend, the second is located a little above and to the right of the first, with a more accentuated positive trend, and the third is in the upper left corner of the chart, with a slight vertical trend. Furthermore, some high-grade samples were grouped in the lower left corner of the chart. The recorded lithologies and the predicted clusters show a clear correlation with these groups. Samples logged as metasediments, volcanics and volcanic breccias are distributed mainly along the first

Cu point-cloud. The high-grade samples in the lower left corner were logged as undifferentiated breccias, as well as most of the samples along the second Cu point-cloud. Samples in the third point-cloud were mainly logged as volcanics, volcanic breccias, and dolerite. Also, it is important to notice that almost all the samples logged as gabbronorite and gabbros are associated with the low Cu value point cloud in the upper right of the chart. The clusters obtained by the proposed workflow, in addition to preserving these relationships, were able to distinguish the high content samples, assigned to cluster 0, from the undifferentiated breccia, generally assigned to cluster 3.

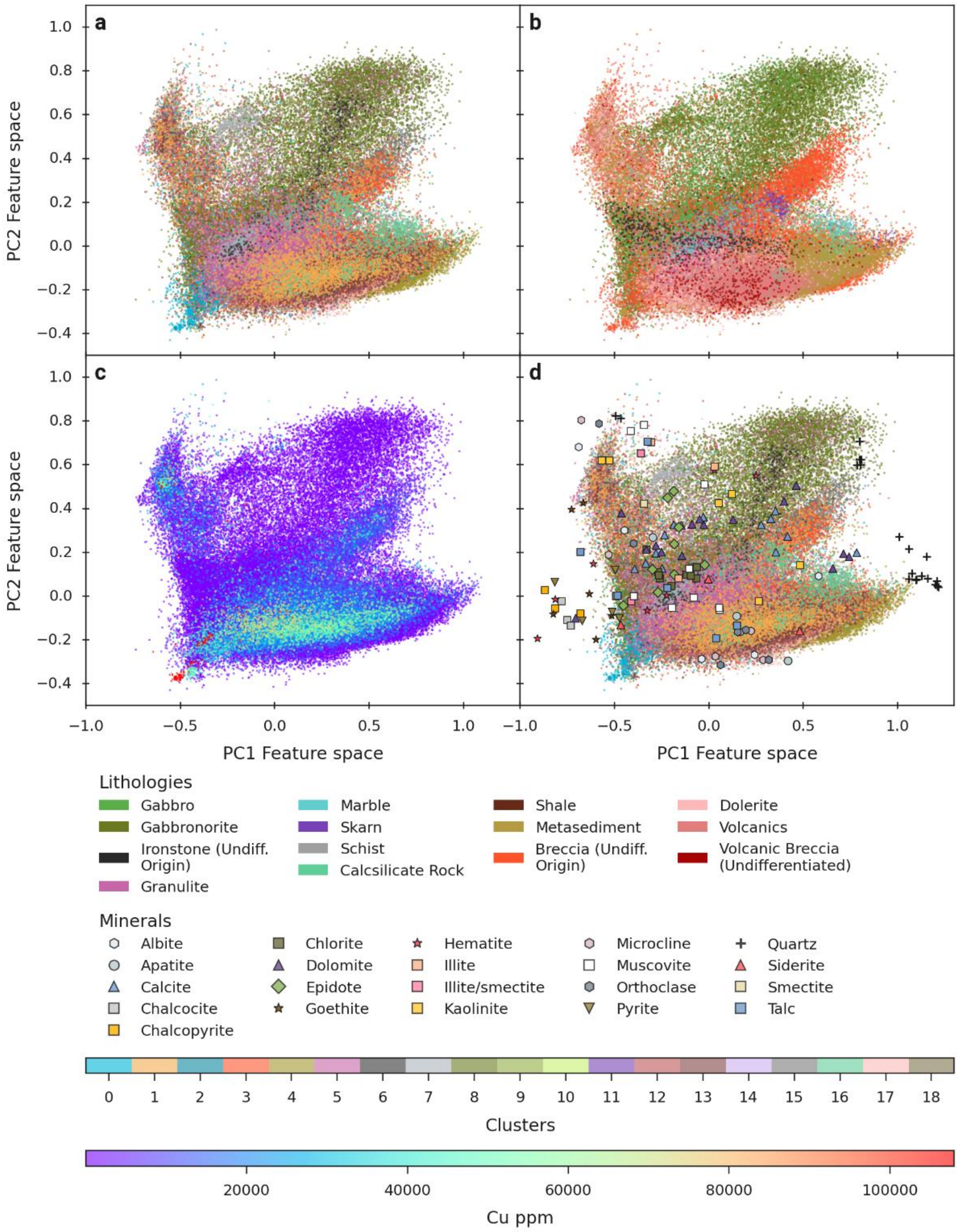


Fig. 6 Plots of the first two principal components of the dataset. (a) Samples colored by the clusters predicted with the proposed workflow. (b) Samples colored by the logged lithologies. (c) Samples colored by Cu analysis. (d) Same as (a) but with the selected minerals from the ASTER version 2 spectral library plotted together.

7.4.3 Clusters mineral assemblage

To better understand the relationship between the clusters and the lithologic and alteration boundaries, the minerals associated with the average spectrum of each cluster were estimated. For this, we used the spectra from the ASTER version 2 spectral library (Baldridge et al., 2009), which provides spectra of minerals collected in the same spectral region as the spectra used for the clustering

(VNIR-SWIR + TIR). From this library we selected 142 spectra of minerals (Fig. 6d) which could be associated with the logged lithologies and the alteration zones according to previous works. The Euclidean distance between these spectra and the average spectra of each cluster were then calculated. The minerals with the smallest distances for each cluster are shown in Figure 7.

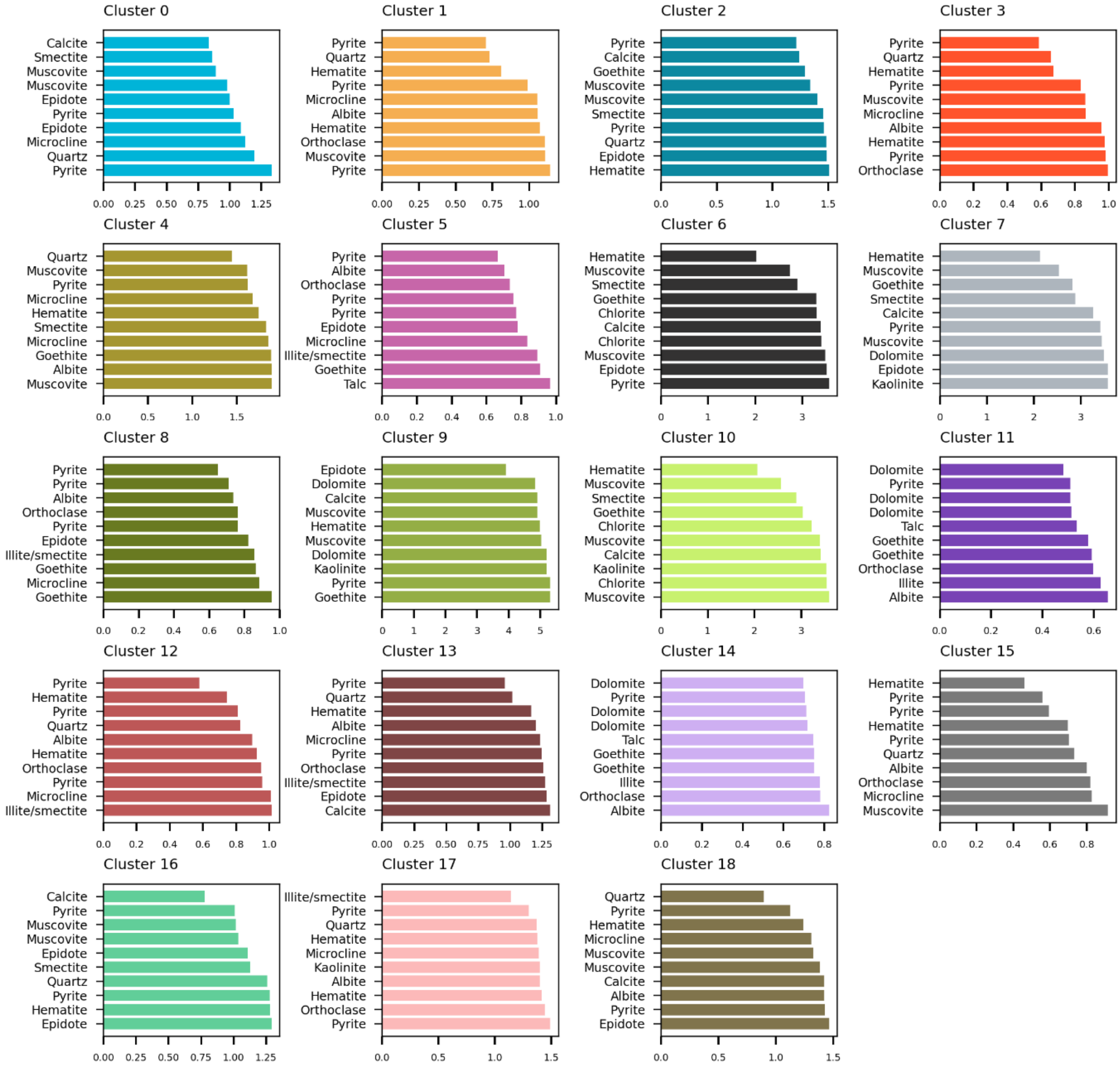


Fig. 7 Euclidean distance between spectra of minerals from the ASTER version 2 spectral library (Baldridge et al., 2009) and the average spectra of each cluster obtained by the proposed workflow. Only the ten nearest minerals of each cluster are shown.

Hematite, pyrite, and quartz are the minerals closest to the cluster associated with the mineralized zones, such as clusters 1, 3, 12, 13 and 15. Cluster 0, which is associated with the high-grade samples, is closer to calcite, smectite and muscovite spectra. These results agree with the alteration assemblies recorded in previous work at Prominent Hill (see Section 7.2.1), where Cu mineralization is associated with Fe and Cu sulfides, and intense iron oxide alteration, composed predominantly of hematite and quartz. Furthermore, the Cu mineralization at Prominent Hill is also strongly associated with pervasive replacement of calcareous lithologies by hematite, producing skarnlike assemblages.

Clusters 0, 1, 3 and 15 occur predominantly in the second mineralized interval of drill hole DDHURAN1, between 420 and 600 m, approximately (Fig. 4). As all samples from this interval were described as undifferentiated breccias, these clusters represent alteration boundaries rather than lithologic boundaries. According to the results, the main difference between these clusters is the abundance of hematite, feldspars, phyllosilicates, and carbonates. The alteration zones rich in carbonates are associated with cluster 0. Cluster 1 is associated with alteration zones where hematite, quartz and feldspars predominate. Cluster 3 is closer to muscovite than the others, being associate with an alteration assembly richer in sericite. Lastly, cluster 15 seem to be associated with an assemblage richer in hematite than the others.

7.5 Conclusions

In this work, an unsupervised method of extracting lithological and alteration boundaries from hyperspectral drill core data that exploit the properties of DAE and SOM together with hierarchical clustering methods was presented. The key point of the proposed method is to significatively reduce the dimensionality of the dataset before performing the cluster analysis, without losing important information. This approach reduces the complexity of data clustering, allowing it to be applied to large datasets in a reasonable time.

The proposed workflow has advantages that make it a promising alternative to automate the identification of lithologic and alteration boundaries of drill cores. These include: (1) the method does not require labeled data to be applied; (2) the method produces smoother cluster distribution along the drill cores, strongly correlated to the logged lithologies; (3) the clusters were able to differentiate mineralized and non-mineralized samples, as well as distinguish high-grade zones.

The cluster distribution obtained by the proposed workflow provide a more detailed allocation of lithologies and alteration boundaries than the visual drill core logging and can be used to further refine and spatially locate potential downhole changes. Therefore, the workflow is an important tool to support the decision process during mining and exploration, where potential regions of change that may not be readily apparent in visual inspections can be find or confirmed. For example, potentially significant alteration zones in the intervals logged as mineralized breccias were identified. In this way, the proposed workflow highlights composition differences between visually similar rocks, providing significant information about the distribution of lithologies, alteration zones, and even ore grades of the drill cores.

The clustering results obtained in a semi-automated manner by the proposed workflow enables the fast extraction of valuable information from large hyperspectral datasets, assisting in the definition of large-scale lithologic boundaries and characterization of small-scale variations associated with alteration mineral zonation. The latter is a powerful tool that can be used by the mining industry to guide operation and find new deposits. This workflow can be applied to other deposit types since most of the alteration minerals associated with mineral deposits are normally recognizable with hyperspectral sensors. In future works the use of recurrent autoencoders based on long short-term memory networks need to be explored for extracting spatial-spectral information during clustering, as well as other clustering strategies need to be evaluated.

8 GENERAL CONCLUSION

This thesis aimed to solve challenges required to implement machine learning techniques in mineral exploration targeting. This works describes the development and benefits of using machine learning for mineral prospectivity mapping. It has been shown what are the drawbacks of dealing with unbalanced training data sets, and how to solve them. This was achieved by demonstrating how imbalanced dataset affects model performance, and how the creation of synthetic samples can overcome this issue, resulting in more assertive prospectivity models. In addition, this work describes the development and benefits of a novel machine learning approach to estimate ore grade by means of ultraspectral data. The proposed approach can be used in autonomous systems to not only improve mining operations but also decrease environmental impacts.

The current state of the art in the application of machine learning algorithms for mineral prospectivity mapping and integration of spectral and geochemical data are presented in Chapters 3 and 4.

Chapter 5 explores the use of machine learning algorithms for mineral prospectivity mapping of IOCG deposits in the Carajás mineral province, Brazil. Prospectivity maps are important tools for mineral exploration targeting, as it contributes to the identification of new mineralized locations. It was demonstrated that the creation of synthetic mineralized samples allowed to effectively enhance the performance of prospectivity models based on machine learning techniques.

In the other study, described in Chapter 6, a novel approach based on deep neural networks was developed for the prediction of Cu grade by means of spectral data in the Olympic Dam IOCG deposit, Australia. The development of techniques capable to estimate ore grade of rocks using non-invasive and non-destructive methods are crucial for the implementation of autonomous systems in the mining industry. Ore grade estimation using ultraspectral cameras enables automatic ore detection and selective mining. Using state of the art deep learning techniques it was demostrated that ultraspectral data combined with geochemical data allow the development of a regression model which is able to predict the Cu grade along drillholes with good precision (+/- ~0.4 %).

Chapter 7 shows a novel approach based on deep autoencoders, self-organizing maps and agglomerative clustering for unsupervised classification of hyperspectral data collected from drill-cores. The proposed method significatively reduces the dimensionality of the dataset before performing the cluster analysis. This approach makes computing clusters faster, allowing it to be applied to large datasets in a reasonable time. The technique assists in the identification of lithological and alteration boundaries, improving the lithological description of the cores.

The results of this thesis indicate that machine learning techniques overperform traditional techniques used for mineral exploration targeting. The developed methods can be seamlessly used in the exploration of other deposits as well. Furthermore, techniques which can be used to integrate Industry 4.0 technologies into the mining industry, such as those proposed in this work, corroborate to a solid and sustainable development of mineral exploration.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., 2015. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv Prepr. arXiv1603.04467.

Abedi, M., Norouzi, G.H., Fathianpour, N., 2013. Fuzzy outranking approach: A knowledge-driven method for mineral prospectivity mapping. Int. J. Appl. Earth Obs. Geoinf. 21, 556–567. https://doi.org/10.1016/j.jag.2012.07.012

Acosta, I.C.C., Khodadadzadeh, M., Tolosana-Delgado, R., Gloaguen, R., 2020. Drill-Core Hyperspectral and Geochemical Data Integration in a Superpixel-Based Machine Learning Framework. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 4214–4228. https://doi.org/10.1109/JSTARS.2020.3011221

Acosta, I.C.C., Khodadadzadeh, M., Tusa, L., Ghamisi, P., Gloaguen, R., 2019. A machine learning framework for drill-core mineral mapping using hyperspectral and high-resolution mineralogical data fusion. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12, 4829–4842.

Acosta, I.C.C., Khodadadzadeh, M., Tusa, L., Ghamisi, P., Gloaguen, R., 2020. A Machine Learning Framework for Drill-Core Mineral Mapping Using Hyperspectral and High-Resolution Mineralogical Data Fusion. IEEE J Sel Top Appl Earth Obs Remote Sens 12, 4829–4842. https://doi.org/10.1109/JSTARS.2019.2924292

Adams, J.B., 1974. Visible and near-infrared diffuse reflectance spectra of pyroxenes as applied to remote sensing of solid objects in the solar system. J. Geophys. Res. 79, 4829–4836. https://doi.org/10.1029/JB079i032p04829

Adams, J.B., 1975. Interpretation of Visible and Near-infrared Diffuse Reflectance Spectra of Pyroxenes and Other Rock-forming Minerals, in: KARR, C. (Ed.), Infrared and Raman Spectroscopy of Lunar and Terrestrial Minerals. Academic Press, pp. 91–116. https://doi.org/https://doi.org/10.1016/B978-0-12-399950-4.50009-4

Adep, R.N., shetty, A., Ramesh, H., 2017. EXhype: A tool for mineral classification using hyperspectral data. ISPRS Journal of Photogrammetry and Remote Sensing 124, 106–118. https://doi.org/10.1016/j.isprsjprs.2016.12.012

Affeldt, S., Labiod, L., Nadif, M., 2020. Spectral clustering via ensemble deep autoencoder learning (SC-EDAE). Pattern Recognit 108, 107522.

https://doi.org/https://doi.org/10.1016/j.patcog.2020.107522

Agterberg, F., Bonham-Carter, G., Wright, D., 1990. Statistical pattern integration for mineral exploration. Comput. Appl. Resour. Estim. Predict. Assement Met. Pet. 1–21.

Agterberg, F.P., Bonham-Carter, G.F., 2005. Measuring the performance of mineral-potential maps. Nat. Resour. Res. 14, 1–17.

Alavi, A.H., Gandomi, A.H., Lary, D.J., 2016. Progress of machine learning in geosciences: Preface. Geosci. Front. 7, 1–2. https://doi.org/https://doi.org/10.1016/j.gsf.2015.10.006

Allen, S.R., McPhie, J., Ferris, G., Simpson, C., 2008. Evolution and architecture of a large felsic Igneous Province in western Laurentia: The 1.6 Ga Gawler Range Volcanics, South Australia. J. Volcanol. Geotherm. Res. 172, 132–147. https://doi.org/https://doi.org/10.1016/j.jvolgeores.2005.09.027

Althoff, F., Barbey, P., Boullier, A.-M., 2000. 2.8–3.0 Ga plutonism and deformation in the SE Amazonian craton: the Archaean granitoids of Marajoara (Carajás Mineral Province, Brazil). Precambrian Res. 104, 187–206.

An, P., Moon, W.M., Rencz, A., 1991. Application of fuzzy set theory for integration of geological, geophysical and remote sensing data. Can. J. Explor. Geophys. 27, 1–11.

Assent, I., 2012. Clustering high dimensional data. WIREs Data Mining and Knowledge Discovery 2, 340–350. https://doi.org/https://doi.org/10.1002/widm.1062

Baldridge, A.M., Hook, S.J., Grove, C.I., Rivera, G., 2009. The ASTER spectral library version 2.0. Remote Sens Environ 113, 711–715. https://doi.org/https://doi.org/10.1016/j.rse.2008.11.007

Banijamali, E., Ghodsi, A., 2017. Fast spectral clustering using autoencoders and landmarks, in: International Conference Image Analysis and Recognition. pp. 380–388.

Barker, R.D., Barker, S.L.L., Cracknell, M.J., Stock, E.D., Holmes, G., 2021. Quantitative Mineral Mapping of Drill Core Surfaces II: Long-Wave Infrared Mineral Characterization Using µXRF and Machine Learning. Econ. Geol. 116, 821–836. https://doi.org/10.5382/econgeo.4804

Barnet, C.T., Williams, P.M., 2006. Mineral Exploration Using Modern Data Mining

Techniques. Wealth Creat. Miner. Ind. Integr. Sci. Business, Educ. 295–310.

Barnett, C.T., Williams, P.M., 2009. Using Geochemistry and Neural Networks to Map Geology under Glacial Cover Using Geochemistry and Neural Networks to map Geology under Glacial Cover.

Barros, C.E.M., Sardinha, A.S., Barbosa, J.P.O., Krimski, R., Macambira, M.J.B., 2001. Pb–Pb and U–Pb zircon ages of Archean syntectonic granites of the Carajás metallogenic province, northern Brazil, in: Proceedings of the South American Symposium on Isotopic Geology, Expanded Abstracts. pp. 94–97.

Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. 6, 20–29.

Belperio, A., Flint, R., Freeman, H., 2007. Prominent Hill: A Hematite-Dominated, Iron Oxide Copper-Gold System. Economic Geology 102, 1499–1510. https://doi.org/10.2113/gsecongeo.102.8.1499

Belperio, A., Freeman, H., 2004. Common geological characteristics of prominent hill and olympic dam - Implications for iron oxide copper-gold exploration models. AusIMM Bulletin 67–75.

Bergen, K.J., Johnson, P.A., De Hoop, M. V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science (80-.). https://doi.org/10.1126/science.aau0323

Betts, P.G., Valenta, R.K., Finlay, J., 2003. Evolution of the Mount Woods Inlier, northern Gawler Craton, Southern Australia: an integrated structural and aeromagnetic analysis. Tectonophysics 366, 83–111. https://doi.org/https://doi.org/10.1016/S0040-1951(03)00062-3

Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When Is ``Nearest Neighbor'' Meaningful?, in: Beeri Catriel and Buneman, P. (Ed.), Database Theory — ICDT'99. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 217–235.

Bioucas-Dias, J.M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N., Chanussot, J., 2013. Hyperspectral Remote Sensing Data Analysis and Future Challenges. IEEE Geosci. Remote Sens. Mag. 1, 6–36. https://doi.org/10.1109/MGRS.2013.2244672

Blackman, R.B., Tukey, J.W., 1958. The measurement of power spectra from the point of view of communications engineering—Part I. Bell Syst. Tech. J. 37, 185–282.

Bonham-Carter, G.F., 1994. Geographic information systems for geoscientists-modeling with GIS. Comput. methods Geosci. 13, 398.

Bonham-Carter, G.F., Agterberg, F.P., Wright, D.F., 1988. Integration of Geological Data Sets for Gold Exploration in Nova Scotia. Photogramm. Engeneering Remote Sens. 54, 1585–1592.

Boots, B.N., Getis, A., 1988. Point pattern analysis. Sage, Beverly Hills.

Booysen, R., Jackisch, R., Lorenz, S., Zimmermann, R., Kirsch, M., Nex, P.A.M., Gloaguen, R., 2020. Detection of REEs with lightweight UAV-based hyperspectral imaging. Sci. Rep. 10, 1–12.

Branco, P., Ribeiro, R.P., Torgo, L., 2016. UBL: an R package for utility-based learning. arXiv Prepr. arXiv1604.08079.

Branco, P., Torgo, L., Ribeiro, R.P., 2017. SMOGN: a Pre-processing Approach for Imbalanced Regression, in: LIDTA@PKDD/ECML.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1007/BF00058655

Brown, W.M., Gedeon, T.D., Groves, D.I., Barnes, R.G., 2000. Artificial neural networks: a new method for mineral prospectivity mapping. Aust. J. earth Sci. 47, 757–770.

Burkov, A., 2019. The Hundred-Page Machine Learning Book. Andriy Burkov.

Burns, R.G., 1993. Mineralogical Applications of Crystal Field Theory, 2nd ed, Cambridge Topics in Mineral Physics and Chemistry. Cambridge University Press. https://doi.org/10.1017/CBO9780511524899

Calvin, W.M., Pace, E.L., 2016. Mapping alteration in geothermal drill core using a field portable spectroradiometer. Geothermics 61, 12–23.

Camps-valls, B.G., Bioucas-dias, J., 2016. A Special Issue on Advances in Machine Learning for Remote Sensing.

Carranza, E.J.M., 2008. Geochemical Anomaly and Mineral Prospectivity Mapping in GIS, Handbook of exploration and environmental geochemistry; 11. Elsevier. https://doi.org/10.1016/S1874-2734(09)70001-4

Carranza, E.J.M., 2009. Objective selection of suitable unit cell size in data-driven modeling of mineral prospectivity. Comput. Geosci. 35, 2032–2046. https://doi.org/10.1016/j.cageo.2009.02.008

Carranza, E.J.M., 2011. Geocomputation of mineral exploration targets. Comput. Geosci. 37, 1907–1916. https://doi.org/10.1016/j.cageo.2011.11.009

Carranza, E.J.M., Laborte, A.G., 2015a. Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: Application of Random Forests algorithm. Ore Geol. Rev. 71, 777–787. https://doi.org/10.1016/j.oregeorev.2014.08.010

Carranza, E.J.M., Laborte, A.G., 2015b. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). Comput. Geosci. 74, 60–70. https://doi.org/10.1016/j.cageo.2014.10.004

Carranza, E.J.M., Laborte, A.G., 2016. Data-Driven Predictive Modeling of Mineral Prospectivity Using Random Forests: A Case Study in Catanduanes Island (Philippines). Nat. Resour. Res. 25, 35–50. https://doi.org/10.1007/s11053-015-9268-x

Carter, D., Belperio, T., Freeman, H., 2003. The discovery of the Promi-nent Hill copper-gold deposit, South Australia. NewGenGold—case histories of discovery 15.

Castelluccio, M., Poggi, G., Sansone, C., Verdoliva, L., 2015. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. arXiv Prepr. arXiv1508.00092 1–11.

Caté, A., Perozzi, L., Gloaguen, E., Blouin, M., 2017. Machine Learning as a tool for geologists. Lead. Edge 36, 215–219. https://doi.org/10.1190/tle36030064.1

Chalmers, N.C., 2007. Mount Woods domain: Proterozoic metasediments and intrusives.

Chang, C.-I., 1999. Spectral information divergence for hyperspectral image analysis, in: IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No.99CH36293). pp. 509–511 vol.1. https://doi.org/10.1109/IGARSS.1999.773549

Chawla, N. V, Japkowicz, N., Kotcz, A., 2004. Editorial: Special Issue on Learning from Imbalanced Data Sets. SIGKDD Explor. Newsl. 6, 1–6. https://doi.org/10.1145/1007730.1007733

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

https://doi.org/10.1613/jair.953

Chen, C., He, B., Zeng, Z., 2014. A method for mineral prospectivity mapping integrating C4.5 decision tree, weights-of-evidence and m-branch smoothing techniques: A case study in the eastern Kunlun Mountains, China. Earth Sci. Informatics 7, 13–24. https://doi.org/10.1007/s12145-013-0128-0

Chen, Y., Jiang, H., Li, C., Jia, X., Member, S., 2016. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. IEE Trans. Geosci. Remote Sens. 54, 6232–6251. https://doi.org/10.1109/TGRS.2016.2584107

Chen, Y., Wu, W., 2016. A prospecting cost-benefit strategy for mineral potential mapping based on ROC curve analysis. Ore Geol. Rev. 74, 26–38. https://doi.org/10.1016/j.oregeorev.2015.11.011

Chen, Y., Wu, W., 2017a. Mapping mineral prospectivity using an extreme learning machine regression. Ore Geol. Rev. 80, 200–213. https://doi.org/10.1016/j.oregeorev.2016.06.033

Chen, Y., Wu, W., 2017b. Application of one-class support vector machine to quickly identify multivariate anomalies from geochemical exploration data. Geochemistry Explor. Environ. Anal. 17, 231–238. https://doi.org/10.1144/geochem2016-024

Cheng, Q., 2007. Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol. Rev. 32, 314–324. https://doi.org/10.1016/j.oregeorev.2006.10.002

Cheng, Q., 2007. Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol. Rev. 32, 314–324. https://doi.org/10.1016/j.oregeorev.2006.10.002

Chi, M., Plaza, A.J., Benediktsson, J.A., Zhang, B., Huang, B., 2015. Foreword to the Special Issue on Big Data in Remote Sensing. IEEE J. Sel. Top. Appl. EARTH Obs. Remote Sens. 8, 4607–4609. https://doi.org/10.1109/TGRS.2007.909915

Chollet, F., others, 2015. Keras.

Clark, R.N., 1999. Spectroscopy of rocks and minerals, and principles of spectroscopy, Remote sensing for the earth sciences: Manual of remote sensing. https://doi.org/10.1111/j.1945-5100.2004.tb00079.x

Clark, R.N., King, T.V. V, Klejwa, M., Swayze, G.A., Vergo, N., 1990. High spectral resolution reflectance spectroscopy of minerals. J. Geophys. Res. Solid Earth 95, 12653–12680. https://doi.org/https://doi.org/10.1029/JB095iB08p12653

Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. J. Geophys. Res. Solid Earth 89, 6329–6340.

Clark, R.N., Swayze, G.A., Livo, K.E., Kokaly, R.F., Sutley, S.J., Dalton, J.B., McDougal, R.R., Gent, C.A., 2003. Imaging spectroscopy: Earth and planetary remote sensing with the USGS Tetracorder and expert systems. J Geophys Res Planets 108. https://doi.org/https://doi.org/10.1029/2002JE001847

Conor, C., 1995. Moonta-Wallaroo region: An interpretation of the geology of the Maitland and Wallaroo 1:100 000 sheet areas: Adelaide, South Australia.

Cordani, U.G., Teixeira, W., 2007. Proterozoic accretionary belts in the Amazonian Craton. Geol. Soc. Am. Mem. 200, 297–320.

Cowley, W.M., Conor, C.H.H., Zang, W.L., 2003. New and revised Pro-terozoic stratigraphic units on the northern Yorke Peninsula: MESA Jour-nal, v. 29, p. 46–58. MESA Journal 29, 46–58.

Cracknell, M.J., Caritat, P. De, 2017. Catchment-based gold prospectivity analysis combining geochemical , geophysical and geological data across northern Australia 17, 204–216. https://doi.org/10.1144/geochem2016-012

Cracknell, M.J., Reading, A.M., 2014. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. Comput. Geosci. 63, 22–33. https://doi.org/10.1016/j.cageo.2013.10.008

Craveiro, G.S., Xavier, R.P., Villas, R.N.N., 2019. The Cristalino IOCG deposit: an example of multi-stage events of hydrothermal alteration and copper mineralization. Brazilian J. Geol. 49.

Creaser, R.A., 1989. The geology and petrology of Middle Proterozoic felsic magmatism of the Stuart Shelf. South Aust. [Ph. D. thesis] Canberra, Aust. Aust. Natl. Univ.

Crowley, J.K., Vergo, N., 1988. Near-Infrared Reflectance Spectra of Mixtures of Kaolin-Group Minerals: Use in Clay Mineral Studies. Clays Clay Miner. 36, 310–316. https://doi.org/10.1346/CCMN.1988.0360404

Cudahy, T., Jones, M., Thomas, M., Cocks, P., Agustin, F., Caccetta, M., Hewson, R., Verrall, M., Rodger, A., 2009. Drill core logging of plagioclase feldspar composition and other minerals associated with Archean gold mineralization at Kambalda, Western Australia, using bidirectional thermal infrared reflectance system. Rev. Econ. Geol. 16, 223–235.

Cudahy, T., Ramanaidou, E.R., 1992. Relationships between spectral properties and ferric oxides CSIRO/AMIRA Project P243 Wembley Australia [R]. CSIRO Div. Explor. Geosci. Rep. 244R 68.

Cudahy, T.J., Ramanaidou, E.R., 1997. Measurement of the hematite:goethite ratio using field visible and near-infrared reflectance spectrometry in channel iron deposits, Western Australia. Aust. J. Earth Sci. 44, 411–420. https://doi.org/10.1080/08120099708728322

Curtiss, B., 1985. Evaluation of the Physical Properties of Naturally Occurring Iron (III) Oxyhydroxides on Rock Surfaces in Arid and Semi-arid Regions Using Visible and Near Infrared Reflectance Spectroscopy. University of Washington.

Dall'Agnol, R., Lafon, J.-M., Macambira, M.J.B., 1994. Proterozoic anorogenic magmatism in the Central Amazonian Province, amazonian araton: Geochronological, petrological and geochemical aspects. Mineral. Petrol. 50, 113–138. https://doi.org/10.1007/BF01160143

Dall'Agnol, R., Souza, Z.S., Althoff, F.J., Barros, C.E.M., Leite, A.A.S., Jorge-João, X.S., 1997. General aspects of the granitogenesis of the Carajás metallogenetic province, in: Proceedings of the International Symposium on Granites and Associated Mineralizations, Salvador, Excursion Guide. pp. 135–161.

Dall'Agnol, R., Teixeira, N.P., Rämö, O.T., Moura, C.A. V, Macambira, M.J.B., de Oliveira, D.C., 2005. Petrogenesis of the Paleoproterozoic rapakivi A-type granites of the Archean Carajás metallogenic province, Brazil. Lithos 80, 101–129.

Dalm, M., Buxton, M.W.N., van Ruitenbeek, F.J.A., 2017. Discriminating ore and waste in a porphyry copper deposit using short-wavelength infrared (SWIR) hyperspectral imagery. Miner. Eng. 105, 10–18. https://doi.org/https://doi.org/10.1016/j.mineng.2016.12.013

Dalm, M., Buxton, M.W.N., van Ruitenbeek, F.J.A., Voncken, J.H.L., 2014. Application of near-infrared spectroscopy to sensor based sorting of a porphyry copper ore.

Miner. Eng. 58, 7–16. https://doi.org/https://doi.org/10.1016/j.mineng.2013.12.016

Daly, S.J., Fanning, C.M., Fairclough, M.C., 1998. Tectonic evolution and exploration potential of the Gawler craton, South Australia. Geological Society of Australia Abstracts 49, 104.

Davidson, M., 2017. Introduction to this special section: Data analytics and machine learning. Lead. Edge 36, 206–206. https://doi.org/10.1190/tle36030206.1

Davis, S. ~M., Landgrebe, D. ~A., Phillips, T. ~L., Swain, P. ~H., Hoffer, R. ~M., Lindenlaub, J. ~C., Silva, L. ~F., 1978. Remote sensing: The quantitative approach.

de Oliveira, M.A., Dall'Agnol, R., Althoff, F.J., da Silva Leite, A.A., 2009. Mesoarchean sanukitoid rocks of the Rio Maria granite-greenstone terrane, Amazonian Craton, Brazil. J. South Am. Earth Sci. 27, 146–160.

de Souza, Z.S., Potrel, A., Lafon, J.-M., Althoff, F.J., Pimentel, M.M., Dall'Agnol, R., de Oliveira, C.G., 2001. Nd, Pb and Sr isotopes in the Identidade Belt, an Archaean greenstone belt of the Rio Maria region (Carajás Province, Brazil): implications for the Archaean geodynamic evolution of the Amazonian Craton. Precambrian Res. 109, 293–315.

deMelo, G.H.C., Monteiro, L.V.S., Xavier, R.P., Moreto, C.P.N., Santiago, E.S.B., Dufrane, S.A., Aires, B., Santos, A.F.F., 2017. Temporal evolution of the giant Salobo IOCG deposit, Carajás Province (Brazil): constraints from paragenesis of hydrothermal alteration and U-Pb geochronology. Miner. Depos. 52, 709–732. https://doi.org/10.1007/s00126-016-0693-5

Diggle, P.J., others, 1983. Statistical analysis of spatial point patterns. Academic press.

Dmitrijeva, M., Ehrig, K.J., Ciobanu, C.L., Cook, N.J., Verdugo-Ihl, M.R., Metcalfe, A. V, 2019. Defining IOCG signatures through compositional data analysis: A case study of lithogeochemical zoning from the Olympic Dam deposit, South Australia. Ore Geol. Rev. 105, 86–101. https://doi.org/https://doi.org/10.1016/j.oregeorev.2018.12.013

DOCEGEO, 1988. Revisão litoestratigráfica da Província Mineral de Carajás, in: Proceedings of the Brazilian Congress of Geology, Vol. 1. pp. 11–56.

Domingos, P.M., 1999. MetaCost: A General Method for Making Classifiers Cost-

Sensitive, in: KDD.

Dorin, I., Diaconescu, C., Topor, D.I., others, 2014. The role of mining in national economies. Int. J. Acad. Res. accounting, Financ. Manag. Sci. 4, 155–160.

Dougherty, R.L., Edelman, A.S., Hyman, J.M., 1989. Nonnegativity-, monotonicity-, or convexity-preserving cubic and quintic Hermite interpolation. Math. Comput. 52, 471–494.

Drexel, J.F., Preiss, W.V., 1995. The Phanerozoic, in: The Geology of South Australia. South Australia Geological Survey Bulletin 54 (347 pp.).

Drummond, C., Holte, R.C., 2003. Erratum: Structure, Properties, and Dynamics of Oxygen Vacancies in Amorphous [Formula presented] [Phys. Rev. Lett. 89, 285505 (2002)]. Phys. Rev. Lett. 91. https://doi.org/10.1103/PhysRevLett.91.039901

du Plessis, L., Damelin, S., Sears, M., 2009. Reducing the dimensionality of hyperspectral data using diffusion maps, in: 2009 IEEE International Geoscience and Remote Sensing Symposium. pp. IV-885-IV–888. https://doi.org/10.1109/IGARSS.2009.5417519

Dubes, R., Jain, A.K., 1976. Clustering techniques: the user’s dilemma. Pattern Recognit 8, 247–260.

Ducart, D.F., Silva, A.M., Toledo, C.L.B., Assis, L.M. de, 2016. Mapping iron oxides with Landsat-8/OLI and EO-1/Hyperion imagery from the Serra Norte iron deposits in the Carajás Mineral Province, Brazil. Brazilian J. Geol. 46, 331–349.

Eggert, R.G., 2010. Mineral exploration and development: risk and reward, in: International Conference on Mining. Http://Www. Un. Org. Kh/Undp/Images/Stories/Special-Pages/Mining-Conference-2010/Docs/Mineral\% 20Exploration\% 20and\% 20Development\% 20by\% 20Roderick\% 20Eg Gert_Eng. Pdf.

Ehrig, Kathy, McPhie, J., Kamenetsky, V.S., 2012. Geology and mineralogical zonation of the Olympic Dam iron oxide Cu-U-Au-Ag deposit, South Australia.

Elkan, C., 2001. The foundations of cost-sensitive learning. Int. Jt. Conf. Artif. Intell. 973–978.

Fairley, I., Mendzil, A., Togneri, M., Reeve, D.E., 2018. The use of unmanned aerial systems to map intertidal sediment. Remote Sens. 10, 1918.

Fanning, C.M., 1997. Geochronological synthesis of southern Australia Part 2, in: The

Gawler Craton: South Australia. Primary Industries and Resources Open File Envelope 8918.

Farias, N.F., Saueressig, R., 1982. Jazida de cobre Salobo 3A. Simp. Geol. da Amaz. 61–73.

Fedi, M., Florio, G., 2001. Detection of potential fields source boundaries by enhanced horizontal derivative method. Geophys. Prospect. 49, 40–58.

Feio, G.R.L., Dall'Agnol, R., Dantas, E.L., Macambira, M.J.B., Santos, J.O.S., Althoff, F.J., Soares, J.E.B., 2013. Archean granitoid magmatism in the Canaã dos Carajás area: implications for crustal evolution of the Carajás province, Amazonian craton, Brazil. Precambrian Res. 227, 157–185.

Ferreira Filho, C.F., Cançado, F., Correa, C., Macambira, E.M.B., Siepierski, L., Junqueira-Brod, T.C., 2007. Mineralizações estratiformes de EGP-Ni associadas a complexos acamadados em Carajás: os exemplos de Luanga e Serra da Onça. Contrib. à Geol. da Amaz. SBG-Núcleo Norte, Belém 1–14.

Fonseca, M., Oliveira, C., Evangelista, H., 2004. The Araguaia Belt, Brazil: Part Of A Neoproterozoic Continental-Scale Strike-Slip Fault System. J. Virtual Explor. 17. https://doi.org/10.3809/jvirtex.2004.00107

Forbes, C.J., Giles, D., Hand, M., Betts, P.G., Suzuki, K., Chalmers, N., Dutch, R., 2011. Using P–T paths to interpret the tectonothermal setting of prograde metamorphism: An example from the northeastern Gawler Craton, South Australia. Precambrian Res 185, 65–85. https://doi.org/https://doi.org/10.1016/j.precamres.2010.12.002

Forbes, C.J., Giles, D., Jourdan, F., Sato, K., Omori, S., Bunch, M., 2012. Cooling and exhumation history of the northeastern Gawler Craton, South Australia. Precambrian Res 200–203, 209–238. https://doi.org/https://doi.org/10.1016/j.precamres.2011.11.003

Ford, A., Blenkinsop, T.G., 2008. Evaluating geological complexity and complexity gradients as controls on copper mineralisation, Mt Isa Inlier. Aust. J. Earth Sci. 55, 13–23. https://doi.org/10.1080/08120090701581364

Ford, A., Miller, J.M., Mol, A.G., 2016. A comparative analysis of weights of evidence, evidential belief functions, and fuzzy logic for mineral potential mapping using incomplete data at the scale of investigation. Nat. Resour. Res. 25, 19–33.

Forest, F., Lebbah, M., Azzag, H., Lacaille, J., 2019. Deep Architectures for Joint Clustering and Visualization with Self-organizing Maps, in: U. Leong Hou and Lauw, H.W. (Ed.), Trends and Applications in Knowledge Discovery and Data Mining. Springer International Publishing, Cham, pp. 105–116.

Fraser, S.J., Whitbourn, L., Yang, K., Ramanaidou, E., Connor, P., Poropat, G., Soole, P., Mason, P., Coward, D., Phillips, R., 2006. Mineralogical face-mapping using hyperspectral scanning for mine mapping and control. 6th Int. Min. Geol. Conf. Rising to Chall. 227–232.

Freeman, H., Tomkinson, M., 2010. Geological setting of iron oxide related mineralisation in the southern Mount Woods domain, South Austra-lia, in Porter, T.M., ed., Hydrothermal iron oxide copper-gold and related deposits: A global perspective. Adelaide, PGC Publishing 3, 171–190.

Freund, Y., 1995. Boosting a Weak Learning Algorithm by Majority. Inf. Comput. 121, 256–285. https://doi.org/https://doi.org/10.1006/inco.1995.1136

Friedman, J., Hastie, T., Tibshirani, R., others, 2001. The elements of statistical learning. Springer series in statistics New York.

Fukuda, K., 2020. Science, technology and innovation ecosystem transformation toward society 5.0. Int. J. Prod. Econ. 220, 107460.

Galarza, M.A., Macambira, M.J.B., Villas, R.N., 2007. Dating and isotopic characteristics (Pb and S) of the Fe oxide-Cu-Au-U-REE Igarapé Bahia ore deposit, Carajás mineral province, Pará state, Brazil. J. South Am. Earth Sci. 25, 377–397. https://doi.org/10.1016/j.jsames.2007.07.006

Gao, A.F., Rasmussen, B., Kulits, P., Scheller, E.L., Greenberger, R., Ehlmann, B.L., 2021. Generalized Unsupervised Clustering of Hyperspectral Images of Geological Targets in the Near Infrared, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4294–4303.

Gao, Y., Zhang, Z., Xiong, Y., Zuo, R., 2016. Mapping mineral prospectivity for Cu polymetallic mineralization in southwest Fujian Province, China. Ore Geol. Rev. 75, 16–28. https://doi.org/10.1016/J.OREGEOREV.2015.12.005

Gewali, U.B., Monteiro, S.T., Saber, E., 2018. Machine learning based hyperspectral image analysis: A survey.

Gibbs, A.K., Wirth, K.R., Hirata, W.K., Olszewski, W.J., 1986. Age and composition of the

Grão Pará Group volcanics, Serra dos Carajás. Rev. Bras. Geociências 16, 201–211.

Gillespie, P.A., Howard, C.B., Walsh, J.J., Watterson, J., 1993. Measurement and characterisation of spatial distributions of fractures. Tectonophysics 226, 113–141. https://doi.org/https://doi.org/10.1016/0040-1951(93)90114-Y

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 249–256.

Goetz, A.F.H., 2009. Three decades of hyperspectral remote sensing of the Earth: a personal view. Remote Sens. Environ. 113, S5–S16. https://doi.org/10.1016/j.rse.2007.12.014

Goetz, A.F.H., Vane, G., Solomon, J.E., Rock, B.N., 1985. Imaging Spectrometry for Earth Remote Sensing. Science (80-.). 228, 1147–1153. https://doi.org/10.1126/science.228.4704.1147

Goncalves, M.L., Netto, M.L.A., Costa, J.A.F., Zullo Junior, J., 2008. An unsupervised method of classifying remotely sensed images using Kohonen self-organizing maps and agglomerative hierarchical clustering methods. Int J Remote Sens 29, 3171–3207.

Goodfellow, I.J., Bengio, Y., Courville, A., 2016. Deep Learning. https://doi.org/10.1038/nmeth.3707

Gower, J.C., 1985. Properties of Euclidean and non-Euclidean distancematrices, in: Linear Algebra Appl. pp. 81–97.

Grainger, C.J., Groves, D., Tallarico, F.H.B., Fletcher, I.R., 2008. Metallogenesis of the Carajas Mineral Province, Southern Amazon Craton, Brazil : Varying styles of Archean through Paleoproterozoic to Neoproterozoic base- and precious-metal mineralisation. Ore Geol. Rev. 33, 451–489. https://doi.org/10.1016/j.oregeorev.2006.10.010

Granek, J., Facility, G.I., Columbia, B., 2016. Advanced Geoscience Targeting via Focused Machine Learning Applied to the QUEST Project Dataset , British Columbia 117–126. https://doi.org/10.14288/1.0340340

Granek, J., Haber, E., 2015. Data mining for real mining: A robust algorithm for prospectivity mapping with uncertainties 145–153.

https://doi.org/10.1137/1.9781611974010.17

Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18, 602–610. https://doi.org/https://doi.org/10.1016/j.neunet.2005.06.042

Green, A.A., Berman, M., Switzer, P., Craig, M.D., 1988. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. IEEE Transactions on geoscience and remote sensing 26, 65–74.

Griffin, D., Lim, J., 1984. Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust. 32, 236–243.

Groves, D.I., Bierlein, F.P., Meinert, L.D., Hitzman, M.W., 2010. Iron oxide copper-gold (IOCG) deposits through earth histoiy: Implications for origin, lithospheric setting, and distinction from other epigenetic iron oxide deposits. Econ. Geol. 105, 641–654. https://doi.org/10.2113/gsecongeo.105.3.641

Groves, D.I., Vielreicher, N.M., 2001. The Phalabowra (Palabora) carbonatite-hosted magnetite-copper sulfide deposit, South Africa: An end-member of the iron-oxide copper-gold-rare earth element deposit group? Miner. Depos. 36, 189–194. https://doi.org/10.1007/s001260050298

Groves, D.I., Vielreicher, R.M., Goldfarb, R.J., Condie, K.C., 2005. Controls on the heterogeneous distribution of mineral deposits through time. Geol. Soc. London, Spec. Publ. 248, 71–101.

Grus, J., 2015. Data Science from Scratch. 1005 Gravenstein Highway North.

GSB/CPRM - Geological Survey of Brazil, 2004. Projeto Aerogeofísico Anapu-Tuerê: relatório final do levantamento e processamento dos dados magnetométricos e gamaespectrométricos. Programa Geologia do Brasil - PGB.

GSB/CPRM - Geological Survey of Brazil, 2010. Projeto Aerogeofísico Tucuruí: relatório final do levantamento e processamento dos dados magnetométricos e gamaespectrométricos. Programa Geologia do Brasil - PGB.

GSB/CPRM - Geological Survey of Brazil, 2015a. Projeto Aerogeofísico Oeste de Carajás: relatório final do levantamento e processamento dos dados magnetométricos e gamaespectrométricos. Programa Geologia do Brasil - PGB.

GSB/CPRM - Geological Survey of Brazil, 2015b. Projeto Aerogeofísico Rio Maria: relatório final do levantamento e processamento dos dados magnetométricos e gamaespectrométricos. Programa Geologia do Brasil - PGB.

GSB/CPRM - Geological Survey of Brazil, 2015c. Levantamento Aerogravimétrico Carajás: relatório final do levantamento e processamento dos dados magnetométricos e gravimétricos. Programa Geologia do Brasil - PGB.

Guo, Y., Gao, J., Li, F., 2013. Spatial subspace clustering for hyperspectral data segmentation, in: Conference of The Society of Digital Information and Wireless Communications (SDIWC). p. 3.

Haddad-Martim, P.M., Souza Filho, C.R.D., Carranza, E.J.M., 2017. Spatial analysis of mineral deposit distribution: A review of methods and implications for structural controls on iron oxide-copper-gold mineralization in Carajas, Brazil. Ore Geol. Rev. 81, 230–244. https://doi.org/10.1016/j.oregeorev.2016.09.038

Haest, M., Cudahy, T., 2012. Quantitative mineralogy from infrared spectroscopic data. I. Validation of mineral abundance and composition scripts at the rocklea channel iron deposit in Western. Econ. Geol. 107, 209–228.

Hariharan, S., Tirodkar, S., Porwal, A., Bhattacharya, A., Joly, A., 2017. Random Forest-Based Prospectivity Modelling of Greenfield Terrains Using Sparse Deposit Data: An Example from the Tanami Region, Western Australia. Nat. Resour. Res. 26, 489–507. https://doi.org/10.1007/s11053-017-9335-6

Harris, D., Pan, G., 1999. Mineral Favorability Mapping : A Comparison of Artificial Neural Networks , Logistic Regression , and Discriminant Analysis 8, 93–109.

Harris, T.M., Murphy, F.C., Funk, C.W., Betts, P.G., 2013. Mt Woods 2D Seismic Reflection Survey, Gawler Craton, South Australia: An Integrated Minerals Exploration Case Study. ASEG Extended Abstracts 2013, 1–4. https://doi.org/10.1071/ASEG2013ab247

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm. J R Stat Soc Ser C Appl Stat 28, 100–108. https://doi.org/10.2307/2346830

He, H., Garcia, E.A., 2009. Learning from Imbalanced Data IEEE Transactions on Knowledge and Data Engineering v. 21 n. 9.

Herman, J., Usher, W., 2017. SALib: an open-source Python library for sensitivity analysis. J. Open Source Softw. 2, 97.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science (1979) 313, 504–507.

Hirata, T., 1989. Fractal dimension of fault systems in Japan: fractal structure in rock fracture geometry at various scales, in: Fractals in Geophysics. Springer, pp. 157–170.

Hodkiewicz, P., 2003. The Interplay Between Physical and Chemical Processes in the Formation of World-Class Orogenic Gold Deposits in the Eastern Goldfields Province , Western Australia.

Hornby, P., Boschetti, F., Horowitz, F.G., 1999. Analysis of potential field data in the wavelet domain. Geophys. J. Int. 137, 175–196.

Hu, W., Huang, Y., Wei, L., Zhang, F., Li, H., 2015. Deep convolutional neural networks for hyperspectral image classification. J. Sensors 2015. https://doi.org/10.1155/2015/258619

Huang, C., Davis, L.S., Townshend, J.R.G., 2002. An assessment of support vector machines for land cover classii cation. int. j. Remote Sens. 23, 725–749. https://doi.org/10.1080/01431160110040323

Huhn, S.R.B., Souza, C.I. de J., Albuquerque, M.C. de, Leal, E.D., Brustolin, V., 1999. Descoberta do depósito Cu (Au) Cristalino: geologia e mineralização associada-Região da Serra do Rabo-Carajás-PA. SBG/Núcleo Norte, Simpósio Geol. da Amaz. 6, 140–143.

Hunt, G.R., 1976. Mid-infrared spectral behavior of metamorphic rocks. Air Force Cambridge Research Laboratories, Air Force Systems Command, United~….

Hunt, G.R., 1977. Spectral Signatures of Particulate Minerals in the Visible and Near Infrared. GEOPHYSICS 42, 501–513. https://doi.org/10.1190/1.1440721

Hunt, G.R., Ashley, R.P., 1979. Spectra of altered rocks in the visible and near infrared. Econ. Geol. 74, 1613–1629. https://doi.org/10.2113/gsecongeo.74.7.1613

Hunt, G.R., Ashley, R.P., 1979. Spectra of altered rocks in the visible and near infrared. Econ. Geol. 74, 1613–1629. https://doi.org/10.2113/gsecongeo.74.7.1613

Hunt, G.R., Salisbury, J.W., 1970. Visible and near-infrared spectra of minerals and rocks : I silicate minerals. Mod. Geol. 1, 283–300.

Hunt, G.R., Salisbury, J.W., 1974. Mid-Infrared Spectral Behavior of Igneous Rocks.

Jackisch, R., Lorenz, S., Zimmermann, R., Möckel, R., Gloaguen, R., 2018. Drone-borne hyperspectral monitoring of acid mine drainage: An example from the Sokolov lignite district. Remote Sens. 10, 385.

Jacobsen, B.H., 1987. A case for upward continuation as a standard separation filter for potential-field maps. Geophysics 52, 1138–1148.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data Clustering: A Review. ACM Comput. Surv. 31, 264–323. https://doi.org/10.1145/331499.331504

Japkowicz, N., 2001. Supervised versus unsupervised binary-learning by feedforward neural networks. Mach. Learn. 42, 97–122.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: A systematic study. Intell. data Anal. 6, 429–449.

Jiang, L., Li, C., Cai, Z., Zhang, H., 2013. Sampled bayesian network classifiers for class-imbalance and cost-sensitive learning. Proc. - Int. Conf. Tools with Artif. Intell. ICTAI 512–517. https://doi.org/10.1109/ICTAI.2013.82

Juszczak, P., Tax, D., Duin, R.P.W., 2002. Feature scaling in support vector data description, in: Proc. ASCI. pp. 95–102.

Kagermann, H., Wahlster, W., Helbig, J., 2013. Recommendation for implementing the strategic initiative INDUSTRIE 4.0—Securing the Future of German Manufacturing Industry, Final report of the Industrie 4.0 Working Group. Acatech—National Acad. Sci. Eng. Forschungsunion Munchen, Ger.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V., 2017. Machine Learning for the Geosciences: Challenges and Opportunities 1–12.

Khodadadzadeh, M., Gloaguen, R., 2019. Upscaling High-Resolution Mineralogical Analyses to Estimate Mineral Abundances in Drill Core Hyperspectral Data, in: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 1845–1848. https://doi.org/10.1109/IGARSS.2019.8898441

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kingma, D.P., Welling, M., 2013. Auto-Encoding Variational Bayes. https://doi.org/10.48550/ARXIV.1312.6114

Kohonen, T., 2012. Self-organizing maps. Springer Science & Business Media.

Koirala, B., Khodadadzadeh, M., Contreras, C., Zahiri, Z., Gloaguen, R., Scheunders, P., 2019. A supervised method for nonlinear hyperspectral unmixing. Remote Sens. 11, 2458.

Krizhevsky, A., Sutskever, I., Geoffrey E., H., 2012. ImageNet Classification with Deep

Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 25 1–9. https://doi.org/10.1109/5.726791

Kruse, F.A., 1996. Identification and mapping of minerals in drill core using hyperspectral image analysis of infrared reflectance spectra. Int J Remote Sens 17, 1623–1632.

Kruse, F.A., L. Bedell, R., Taranik, J. v, Peppin, W.A., Weatherbee, O., Calvin, W.M., 2012. Mapping alteration minerals at prospect, outcrop and drill core scales using imaging spectrometry. Int J Remote Sens 33, 1780–1798. https://doi.org/10.1080/01431161.2011.600350

Kruse, F.A., Lefkoff, A.B., Boardman, J.W., Heidebrecht, K.B., Shapiro, A.T., Barloon, P.J., Goetz, A.F.H., 1993. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. Remote Sens Environ 44, 145–163. https://doi.org/10.1016/0034-4257(93)90013-N

Kurz, T.H., Buckley, S.J., Howell, J.A., Schneider, D., 2008. Geological Outcrop Modelling and Interpretation Using Ground Based Hyperspectral and Laser Scanning Data Fusion. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 37, B5.

Kurz, T.H., Buckley, S.J., Howell, J.A., Schneider, D., 2011. Integration of panoramic hyperspectral imaging with terrestrial lidar data. Photogramm. Rec. 26, 212–228. https://doi.org/10.1111/j.1477-9730.2011.00632.x

Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. Geosci. Front. 7, 3–10. https://doi.org/https://doi.org/10.1016/j.gsf.2015.07.003

Laukamp, C., Rodger, A., LeGras, M., Lampinen, H., Lau, I.C., Pejcic, B., Stromberg, J., Francis, N., Ramanaidou, E., 2021. Mineral Physicochemistry Underlying Feature-Based Extraction of Mineral Abundance and Composition from Shortwave, Mid and Thermal Infrared Reflectance Spectra. Minerals 11. https://doi.org/10.3390/min11040347

Lecun, Y., Bengio, Y., 1995. Convolutional Networks for Images, Speech, and Time-Series, in: M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks. MIT Press.

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. https://doi.org/10.1038/nature14539

LeCun, Y., others, 1989. Generalization and network design strategies, in: R. Pfeifer, Z. Schreter, F. Fogelman, L. Steels" (Eds.), Connectionism in Perspective. Elsevier, pp. 143–155.

Leite, E.P., de Souza Filho, C.R., 2009. Probabilistic neural networks applied to mineral potential mapping for platinum group elements in the Serra Leste region, Carajás Mineral Province, Brazil. Comput. Geosci. 35, 675–687. https://doi.org/10.1016/j.cageo.2008.05.003

Lemaître, G., Nogueira, F., Aridas, C.K., 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. J. Mach. Learn. Res. 18, 1–5.

Li, J., Bioucas-Dias, J.M., Plaza, A., 2009. Semi-supervised hyperspectral image classification based on a Markov random field and sparse multinomial logistic regression, in: IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. III-817-III–820. https://doi.org/10.1109/IGARSS.2009.5417892

Li, T., Xia, Q., Zhao, M., Gui, Z., Leng, S., 2019. Prospectivity Mapping for Tungsten Polymetallic Mineral Resources, Nanling Metallogenic Belt, South China: Use of Random Forest Algorithm from a Perspective of Data Imbalance. Nat. Resour. Res. https://doi.org/10.1007/s11053-019-09564-8

Lindenmayer, Z.G., 2003. Depósito de Cu–Au do Salobo, Serra dos Carajás: Uma revisão. Caracter. e Model. depósitos minerais. Ed. Unisinos, São Leopoldo 69–98.

Lindenmayer, Z.G., Fleck, A., Gomes, C.H., Santos, A.B.S., Caron, R., Paula, F. de C., Laux, J.H., Pimentel, M.M., Sardinha, A.S., Marini, O.J., others, 2005. Caracterização geológica do alvo Estrela (Cu-Au), Serra dos Carajás, Pará, in: Caracterização de Depósitos Minerais Em Distritos Mineiros Da Amazônia. DNPM, CT-Mineral/FINEP, ADIMB, p. 74p.

Littlefield, E., Calvin, W., Stelling, P., Kent, T., 2012. Reflectance spectroscopy as a drill core logging technique: An example using core from the Akutan geothermal exploration project. Geothermal Resour. Council Trans. 36, 1283–1291.

Liu, H., Wu, K., Xu, H., Xu, Y., 2021. Lithology Classification Using TASI Thermal Infrared Hyperspectral Data with Convolutional Neural Networks. Remote Sens (Basel) 13. https://doi.org/10.3390/rs13163117

Liu, Y., Cheng, Q., Xia, Q., Wang, X., 2014. Mineral potential mapping for tungsten polymetallic deposits in the Nanling metallogenic belt, South China. J. Earth Sci. 25, 689–700.

López, V., Fernández, A., Garc\'\ia, S., Palade, V., Herrera, F., 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inf. Sci. (Ny). 250, 113–141.

Lyon, R.J.P., 1965. Analysis of rocks by spectral infrared emission (8 to 25 microns). Econ. Geol. 60, 715–736.

Lyon, R.J.P., Burns, E.A., 1963. Analysis of rocks and minerals by reflected infrared radiation. Econ. Geol. 58, 274–284.

Machado, N., Lindenmayer, Z., Krogh, T.E., Lindenmayer, D., 1991. U-Pb geochronology of Archean magmatism and basement reactivation in the Carajás area, Amazon shield, Brazil. Precambrian Res. 49, 329–354. https://doi.org/10.1016/0301-9268(91)90040-H

Mallet, V., Nilges, M., Bouvier, G., 2021. quicksom: Self-Organizing Maps on GPUs for clustering of molecular dynamics trajectories. Bioinformatics 37, 2064–2065.

Mandelbrot, B.B., 1983. The fractal geometry of nature. WH freeman New York.

Manduchi, L., Hüser, M., Vogt, J., Rätsch, G., Fortuin, V., 2019. DPSOM: Deep Probabilistic Clustering with Self-Organizing Maps. https://doi.org/10.48550/ARXIV.1910.01590

Manevitz, L.M., Yousef, M., 2001. One-class SVMs for document classification. J. Mach. Learn. Res. 2, 139–154.

Martins, P.L.G., Toledo, C.L.B., Silva, A.M., Chemale, F., Santos, J.O.S., Assis, L.M., 2017. Neoarchean magmatism in the southeastern Amazonian Craton, Brazil: Petrography, geochemistry and tectonic significance of basalts from the Carajás Basin. Precambrian Res. 302, 340–357. https://doi.org/10.1016/j.precamres.2017.10.013

Mathieu, M., Roy, R., Launeau, P., Cathelineau, M., Quirt, D., 2017. Alteration mapping on drill cores using a HySpex SWIR-320m hyperspectral camera: Application to the exploration of an unconformity-related uranium deposit (Saskatchewan, Canada). J. Geochemical Explor. 172, 71–88. https://doi.org/10.1016/j.gexplo.2016.09.008

Mauger, A.J., Ehrig, K., Kontonikas-Charos, A., Ciobanu, C.L., Cook, N.J., Kamenetsky, V.S., 2016. Alteration at the Olympic Dam IOCG–U deposit: insights into distal to proximal feldspar and phyllosilicate chemistry from infrared reflectance spectroscopy. Aust. J. Earth Sci. 63, 959–972. https://doi.org/10.1080/08120099.2016.1264474

Mccuaig, T.C., Beresford, S., Hronsky, J., 2010. Translating the mineral systems approach into an effective exploration targeting system. Ore Geol. Rev. 38, 128–138. https://doi.org/10.1016/j.oregeorev.2010.05.008

McKay, G., Harris, J.R., 2016. Comparison of the Data-Driven Random Forests Model and a Knowledge-Driven Method for Mineral Prospectivity Mapping: A Case Study for Gold Deposits Around the Huritz Group and Nueltin Suite, Nunavut, Canada. Nat. Resour. Res. 25, 125–143. https://doi.org/10.1007/s11053-015-9274-z

Melgani, F., Bruzzone, L., 2004. Classification of hyperspectral remote sensing images with support vector machines. IEEE Transactions on Geoscience and Remote Sensing 42, 1778–1790. https://doi.org/10.1109/TGRS.2004.831865

Monteiro, S.T., Murphy, R.J., Ramos, F., Nieto, J., 2009. Applying boosting for hyperspectral classification of ore-bearing rocks. Mach. Learn. Signal Process. XIX - Proc. 2009 IEEE Signal Process. Soc. Work. MLSP 2009 1–6. https://doi.org/10.1109/MLSP.2009.5306219

Morris, R. V, Lauer Jr., H. V, Lawson, C.A., Gibson Jr., E.K., Nace, G.A., Stewart, C., 1985. Spectral and other physicochemical properties of submicron powders of hematite (α-Fe2O3), maghemite (γ-Fe2O3), magnetite (Fe3O4), goethite (α-FeOOH), and lepidocrocite (γ-FeOOH). J. Geophys. Res. Solid Earth 90, 3126–3144. https://doi.org/https://doi.org/10.1029/JB090iB04p03126

Motta, J.G., Souza Filho, C.R. d., Carranza, E.J.M., Braitenberg, C., 2019. Archean crust and metallogenic zones in the Amazonian Craton sensed by satellite gravity data. Sci. Rep. 9, 1–10. https://doi.org/10.1038/s41598-019-39171-9

Mou, L., Ghamisi, P., Zhu, X. ~X., 2017. Deep Recurrent Neural Networks for Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 55, 3639–3655. https://doi.org/10.1109/TGRS.2016.2636241

Murphy, R.J., Monteiro, S.T., 2013. Mapping the distribution of ferric iron minerals on a vertical mine face using derivative analysis of hyperspectral imagery (430-

970nm). ISPRS J. Photogramm. Remote Sens. 75, 29–39. https://doi.org/10.1016/j.isprsjprs.2012.09.014

Murray, H.H., Lyons, S.C., 1955. Correlation of Paper-Coating Quality with Degree of Crystal Perfection of Kaolinite. Clays Clay Miner. 4, 31–40. https://doi.org/10.1346/CCMN.1955.0040105

Ng, A.Y., 2004. Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance, in: Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04. Association for Computing Machinery, New York, NY, USA, p. 78. https://doi.org/10.1145/1015330.1015435

Nykänen, V., Groves, D.I., Ojala, V.J., Gardoll, S.J., 2008. Combined conceptual/empirical prospectivity mapping for orogenic gold in the northern Fennoscandian Shield, Finland. Aust. J. Earth Sci. 55, 39–59. https://doi.org/10.1080/08120090701581380

Oh, H.-J., Lee, S., 2010. Application of Artificial Neural Network for Gold–Silver Deposits Potential Mapping: A Case Study of Korea. Nat. Resour. Res. 19, 103–124. https://doi.org/10.1007/s11053-010-9112-2

Oliveira, R.G. de, 2018. Insights on the framework of the Carajás Province, Amazonian Craton, Brazil, and on the three-dimensional shape of the Carajás Basin, based on gravity data. J. Geol. Surv. Brazil 1, 101–112. https://doi.org/https://doi.org/10.29396/jgsb.2018.v1.n3.1

Oppenheim, A. V, 1999. Discrete-time signal processing. Pearson Education India.

Padró, J.-C., Carabassa, V., Balagué, J., Brotons, L., Alcañiz, J.M., Pons, X., 2019. Monitoring opencast mine restorations using Unmanned Aerial System (UAS) imagery. Sci. Total Environ. 657, 1602–1614. https://doi.org/https://doi.org/10.1016/j.scitotenv.2018.12.156

Pal, K.K., Sudeep, K.S., 2016. Preprocessing for image classification by convolutional neural networks, in: 2016 IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT). pp. 1778–1781. https://doi.org/10.1109/RTEICT.2016.7808140

Pan, G., Harris, D.P., 2000. Information synthesis for mineral exploration: Oxford Univ. Press. New York.

Partington, G., Sale, M., 2004. Prospectivity mapping using GIS with publicly available earth science data-a new targeting tool being successfully used for exploration in New Zealand. Pacrim 2004 Congr. Vol. Adelaide 19–22.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp. 8024–8035.

Paya, B.A., Esat, I.I., Badi, M.N.M., 1997. ARTIFICIAL NEURAL NETWORK BASED FAULT DIAGNOSTICS OF ROTATING MACHINERY USING WAVELET TRANSFORMS AS A PREPROCESSOR. Mech Syst Signal Process 11, 751–765. https://doi.org/https://doi.org/10.1006/mssp.1997.0090

Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science 2, 559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Person, M., Banerjee, A., Hofstra, A., Sweetkind, D., Gao, Y., 2008. Hydrologic models of modern and fossil geothermal systems in the Great Basin: Genetic implications for epithermal Au-Ag and Carlin-type gold deposits. Geosphere 4, 888. https://doi.org/10.1130/GES00150.1

Pidgeon, R.T., MacAmbira, M.J.B., Lafon, J.M., 2000. Th-U-Pb isotopic systems and internal structures of complex zircons from an enderbite from the Pium Complex, Carajas Province, Brazil: Evidence for the ages of granulite facies metamorphism and the protolith of the enderbite. Chem. Geol. 166, 159–171. https://doi.org/10.1016/S0009-2541(99)00190-4

Pimentel, M.M., Lindenmayer, Z.G., Laux, J.H., Armstrong, R., de Araújo, J.C., 2003. Geochronology and ND isotope geochemistry of the Gameleira Cu-Au deposit, Serra dos Carajás, Brazil: 1.8-1.7 Ga hydrothermal alteration and mineralization.

J. South Am. Earth Sci. 15, 803–813. https://doi.org/10.1016/S0895-9811(02)00127-X

Plaza, A., Benediktsson, J.A., Boardman, J.W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., Tilton, J.C., Trianni, G., 2009. Recent advances in techniques for hyperspectral image processing. Remote Sens. Environ. 113, S110–S122. https://doi.org/10.1016/J.RSE.2007.07.028

Pollard, P.J., Taylor, R.G., Peters, L., Matos, F., Freitas, C., Saboia, L., Huhn, S., 2019. 40Ar-39Ar dating of Archean iron oxide Cu-Au and Paleoproterozoic granite-related Cu-Au deposits in the Carajás Mineral Province, Brazil: implications for genetic models. Miner. Depos. 54, 329–346. https://doi.org/10.1007/s00126-018-0809-1

Porwal, A., Carranza, E.J.M., Hale, M., 2003. Artificial Neural Networks for Mineral-Potential Mapping: A Case Study from Aravalli Province, Western India. Nat. Resour. Res. 12, 155–171. https://doi.org/10.1023/A:1025171803637

Porwal, A., Carranza, E.J.M., Hale, M., 2003a. Knowledge-driven and data-driven fuzzy models for predictive mineral potential mapping. Nat. Resour. Res. 12, 1–25. https://doi.org/10.1023/A:1022693220894

Porwal, A., Carranza, E.J.M., Hale, M., 2003b. Artificial Neural Networks for Mineral-Potential Mapping: A Case Study from Aravalli Province, Western India. Nat. Resour. Res. 12, 155–171. https://doi.org/10.1023/A:1025171803637

Porwal, A., Yu, L., Gessner, K., 2010. SVM-based base-metal prospectivity modeling of the Aravalli Orogen, northwestern India. EGU Gen. Assem. … 12, 15171.

Poulton, M.M., 2002. Y2K Review Article: Neural networks as an intelligence amplification tool: A review of applications. Geophysics 67, 979. https://doi.org/10.1190/1.1484539

Prado, E.M.G., 2016. Espectrorradiometria de reflectância aplicada à análise quantitativa da mineralogia do depósito N4WS, província mineral de Carajás, Pará, Brasil ; Reflectance spectroradiometry applied to a quantitative analysis of the mineralogy of the N4WS deposit,.

Prado, E.M.G., 2020. https://github.com/Eliasmgprado/GeologicalComplexity_SMOTE.

https://doi.org/10.5281/zenodo.3659186

Prado, E.M.G., de Souza Filho, C.R., Carranza, E.J.M., Motta, J.G., 2020. Modeling of Cu-Au prospectivity in the Carajás mineral province (Brazil) through machine learning: Dealing with imbalanced training data. Ore Geol. Rev. 124, 103611. https://doi.org/https://doi.org/10.1016/j.oregeorev.2020.103611

Prado, E.M.G., Silva, A.M., Ducart, D.F., Toledo, C.L.B., de Assis, L.M., 2016. Reflectance spectroradiometry applied to a semi-quantitative analysis of the mineralogy of the N4ws deposit, Carajás Mineral Province, Pará, Brazil. Ore Geol. Rev. 78, 101–119. https://doi.org/https://doi.org/10.1016/j.oregeorev.2016.03.007

Qiu, J.T., Zhang, C., Xu, Q.J., Yao, J.L., 2017. Mapping of carnallite along with semi-quantitative estimation of potassium content of drill cores using hyperspectral imagery. Remote Sens. Lett. 8, 859–868. https://doi.org/10.1080/2150704X.2017.1333651

Quinlan, J.R., 1991. Improved estimates for the accuracy of small disjuncts. Mach. Learn. 6, 93–98.

Radford, D.D.G., Cracknell, M.J., Roach, M.J., Cumming, G. V., 2018. Geological Mapping in Western Tasmania Using Radar and Random Forests. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 3075–3087. https://doi.org/10.1109/JSTARS.2018.2855207

Raines, G.L., Sawatzky, D.L., Bonham-Carter, G.F., 2010. New fuzzy logic tools in ArcGIS 10. ArcUser Spring, 8–13.

Reeve, J.S., 1990. Olympic Dam copper-uranium-gold-silver deposit. Geol. Miner. Depos. Aust. Papua New Guinea 1009–1035.

Ren, Z., Sun, L., Zhai, Q., Liu, X., 2019. Mineral Mapping with Hyperspectral Image Based on an Improved K-Means Clustering Algorithm, in: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. pp. 2989–2992. https://doi.org/10.1109/IGARSS.2019.8899113

Requia, K., Fontboté, L., 2000. The Salobo iron oxide copper-gold deposit, Carajás, northern Brazil. Hydrothermal Iron-Oxide Copper-Gold Relat. Depos. A Glob. Perspect. Adelaide, Aust. Miner. Found. 225–236.

Requia, K., Stein, H., Fontboté, L., Chiaradia, M., 2003. Re--Os and Pb--Pb geochronology of the Archean Salobo iron oxide copper--gold deposit, Carajás mineral province, northern Brazil. Miner. Depos. 38, 727–738.

Ribeiro, R., 2011. Utility-based regression. Ph. D. Diss.

Riley, D.N., Hecker, C.A., 2013. Mineral Mapping with Airborne Hyperspectral Thermal Infrared Remote Sensing at Cuprite, Nevada, USA, in: Kuenzer, C., Dech, S. (Eds.), Thermal Infrared Remote Sensing: Sensors, Methods, Applications. Springer Netherlands, Dordrecht, pp. 495–514. https://doi.org/10.1007/978-94-007-6639-6_24

Rodger, A., Fabris, A., Laukamp, C., 2021. Feature Extraction and Clustering of Hyperspectral Drill Core Measurements to Assess Potential Lithological and Alteration Boundaries. Minerals 11. https://doi.org/10.3390/min11020136

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. https://doi.org/10.1016/j.oregeorev.2015.01.001

Rodriguez-Galiano, V.F., Chica-Olmo, M., Chica-Rivas, M., 2014. Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain. Int. J. Geogr. Inf. Sci. 28, 1336–1354. https://doi.org/10.1080/13658816.2014.885527

Roest, W.R., Verhoef, J., Pilkington, M., 1992. Magnetic interpretation using the 3-D analytic signal. GEOPHYSICS 57, 116–125. https://doi.org/10.1190/1.1443174

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., Roberts, D., 2008. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. Remote Sens. Environ. 112, 2272–2283. https://doi.org/https://doi.org/10.1016/j.rse.2007.10.004

Romero, A., Gatta, C., Camps-valls, G., Member, S., 2015. Sensing Image Classification. IEEE Trans. Geosci. Remote Sens. 1 Unsupervised 54, 1–14. https://doi.org/10.1109/TGRS.2015.2478379.

Ronzê, P.C., Soares, A.D. V, Santos, M.G.S. dos, Barreira, C.F., 2000. Alemão copper-gold (U-REE) deposit, Carajás, Brazil. Hydrothermal iron oxide copper-gold Relat. Depos. a Glob. Perspect. Aust. Miner. Found. Adelaide 191–202.

Santos, J.O.S., Hartmann, L.A., Gaudette, H.E., Groves, D.I., Mcnaughton, N.J., Fletcher, I.R., 2000. A new understanding of the provinces of the Amazon Craton based on integration of field mapping and U-Pb and Sm-Nd geochronology. Gondwana Res. 3, 453–488.

Schlegel, T.U., Heinrich, C.A., 2015. Lithology and Hydrothermal Alteration Control the Distribution of Copper Grade in the Prominent Hill Iron Oxide-Copper-Gold Deposit (Gawler Craton, South Australia)*. Economic Geology 110, 1953–1994. https://doi.org/10.2113/econgeo.110.8.1953

Schneider, S., Murphy, R.J., Melkumyan, A., 2014. Evaluating the performance of a new classifier – the GP-OAD: A comparison with existing methods for classifying rock type and mineralogy from hyperspectral imagery. ISPRS J. Photogramm. Remote Sens. 98, 145–156. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2014.09.016

Schodlok, M.C., Whitbourn, L., Huntington, J., Mason, P., Green, A., Berman, M., Coward, D., Connor, P., Wright, W., Jolivet, M., Martinez, R., 2016. HyLogger-3, a visible to shortwave and thermal infrared reflectance spectrometer system for drill core logging: functional description. Australian Journal of Earth Sciences 63, 929–940. https://doi.org/10.1080/08120099.2016.1231133

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13, 1443–1471.

Shabankareh, M., Hezarkhani, A., 2017. Application of support vector machines for copper potential mapping in Kerman region, Iran. J. African Earth Sci. 128, 116–126. https://doi.org/10.1016/J.JAFREARSCI.2016.11.032

Shahin, M.A., Maier, H.R., Jaksa, M.B., 2004. Data division for developing neural networks applied to geotechnical engineering. J. Comput. Civ. Eng. 18, 105–114. https://doi.org/10.1061/(ASCE)0887-3801(2004)18:2(105)

Shives, R.B.K., Charbonneau, B.W., Ford, K.L., 2000. The detection of potassic alteration by gamma-ray spectrometry - Recognition of alteration related to mineralization. Geophysics 65(6).

Silversides, K.L., Murphy, R.J., 2017. Identification of marker shale horizons in banded iron formation: linking measurements of downhole natural gamma-ray with measurements from reflectance spectrometry of rock cores. Near Surf. Geophys. 15, 141–153. https://doi.org/https://doi.org/10.3997/1873-0604.2016046

Šimon, I., McMahon, H.O., 1953. Study of the Structure of Quartz, Cristobalite, and Vitreous Silica by Reflection in Infrared. J. Chem. Phys. 21, 23–30. https://doi.org/10.1063/1.1698615

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition 1–14. https://doi.org/10.1016/j.infsof.2008.09.005

Singer, D.A., Kouda, R., 1996. Application of a feedforward neural network in the search for Kuroko deposits in the Hokuroku district, Japan. Math. Geol. 28, 1017–1023. https://doi.org/10.1007/BF02068587

Singer, D.A., Kouda, R., 1999. Examining Risk in Mineral Exploration. Nat. Resour. Res. 8, 111–122. https://doi.org/10.1023/A:1021838618750

Singer, D.A.D.A., Kouda, R., 1997. Use of a neural network to integrate geoscience information in the classification of mineral deposits and occurrences. Proc. Explor. 97 Fourth Decenn. Int. Conf. Miner. Explor. 97, 127–134.

Sishi, M., Telukdarie, A., 2020. Implementation of Industry 4.0 technologies in the mining industry-a case study. Int. J. Min. Miner. Eng. 11, 1–22.

Skirrow, R., Bastrakov, E., Davidson, G., Raymond, O., Heithersay, P., 2002. The geological framework, distribution and controls of Fe-oxide Cu–Au mineralisation in the Gawler Craton, South Australia. Part II. Alteration and mineralisation, in: Hydrothermal Iron Oxide Copper-Gold & Related Deposits: A Global Perspective. pp. 33–47.

Soares, A.D. V, Ronzê, P.C., Santos, M.G.C., Leal, E.D., Barreira, C.F., 1999. Geologia e mineralizações do depósito Cu-Au Alemão, Província Mineral de Carajás, PA. SBG, 6o Simpósio Geol. da Amaz. Manaus, AM. Resumos Expand. 144–147.

Sobol, I.M., 2001. Global sensitivity indices for rather complex mathematical models can be efficiently computed by Monte Carlo (or quasi-Monte Carlo) methods. Math Comput Simul 55, 271–280.

Sokal, R.R., 1958. A statistical method for evaluating systematic relationships. Univ. Kansas, Sci. Bull. 38, 1409–1438.

Souza, L.H., Vieira, E.A., 2000. Salobo 3 Alpha deposit: geology and mineralization. Hydrothermal iron oxide copper--gold Relat. Depos. a Glob. Perspect. Aust. Miner. Found. Adelaide 213–224.

Souza, S.R.B., Macambira, M.J.B., Sheller, T., 1996. Novos dados geocronológicos para os granitos deformados do Rio Itacaiúnas (Serra dos Carajás, PA, Brazil): implicações estratigráficas, in: Annals of the Amazon Geologic Symposium, Expanded Abstracts. pp. 380–383.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. 15, 1929–1958. https://doi.org/10.1214/12-AOS1000

Stanley, C.R., 2006. Numerical transformation of geochemical data: 1. Maximizing geochemical contrast to facilitate information extraction and improve data presentation. Geochemistry: Exploration, Environment, Analysis 6, 69–78.

Steinhaus, H., others, 1956. Sur la division des corps matériels en parties. Bull. Acad. Polon. Sci 1, 801.

Sun, Y., Wong, A.K.C., Kamel, M.S., 2009. Classification of Imbalanced Data: a Review. Int. J. Pattern Recognit. Artif. Intell. 23, 687–719. https://doi.org/10.1142/s0218001409007326

Swingler, K., 1996. Applying neural networks: a practical guide. Morgan Kaufmann.

Tallarico, F.H.B., de Oliveira, C.G., Figueiredo, B.R., 2017. The Igarapé Bahia Cu-Au mineralization, Carajás Province. Rev. Bras. Geociências 30, 230–233.

Tallarico, F.H.B., McNaughton, N.J., Groves, D.I., Fletcher, I.R., Figueiredo, B.R., Carvalho, J.B., Rego, J.L., Nunes, A.R., 2004. Geological and SHRIMP II U-Pb constraints on the age and origin of the Breves Cu-Au-(W-Bi-Sn) deposit, Carajás, Brazil. Miner. Depos. 39, 68–86. https://doi.org/10.1007/s00126-003-0383-y

Tao, Y., Li, Y., Lin, X., 2018. A Deep Clustering Algorithm Based on Self-organizing Map Neural Network, in: Huang De-Shuang and Gromiha, M.M. and H.K. and H.A. (Ed.), Intelligent Computing Methodologies. Springer International Publishing, Cham, pp. 182–192.

Tappert, M.C., Rivard, B., Giles, D., Tappert, R., Mauger, A., 2013. The mineral chemistry, near-infrared, and mid-infrared reflectance spectroscopy of phengite from the Olympic Dam IOCG deposit, South Australia. Ore Geol Rev 53, 26–38. https://doi.org/https://doi.org/10.1016/j.oregeorev.2012.12.006

Tassinari, C.C.G., Macambira, M.J.B., 1999. Geochronological provinces of the Amazonia craton. Episodes 22, no. 3, 174–182.

Tax, D., 2001. One-class classification. PhD thesis, Delft Univ. Technol.

Taylor, G.R., 2000. Mineral and lithology mapping of drill core pulps using visible and infrared spectrometry. Natural Resources Research 9, 257–268.

Tazava, E., 2000. The Igarapé Bahia Au-Cu-(REE-U) deposit, Carajás Mineral Province, Northern Brazil. Hydrothermal iron oxide copper-gold Relat. Depos. A Glob. Perspect. 1, 203–212.

Therien, C., 2018. Pysptools [WWW Document]. URL https://pysptools.sourceforge.io/index.html

Tian, F., Gao, B., Cui, Q., Chen, E., Liu, T.-Y., 2014. Learning deep representations for graph clustering, in: Proceedings of the AAAI Conference on Artificial Intelligence.

Torgo, L., Ribeiro, R.P., Pfahringer, B., Branco, P., 2013. SMOTE for Regression, in: Correia, L., Reis, L.P., Cascalho, J. (Eds.), Progress in Artificial Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 378–389.

Torresi, I., Xavier, R.P., Bortholoto, D.F.A., Monteiro, L.V.S., 2012. Hydrothermal alteration, fluid inclusions and stable isotope systematics of the Alvo 118 iron oxide--copper--gold deposit, Carajás Mineral Province (Brazil): Implications for ore genesis. Miner. Depos. 47, 299–323. https://doi.org/10.1007/s00126-011-0373-4

Tuia, D., Merenyi, E., Jia, X., Grana-Romay, M., 2014. Foreword to the Special Issue on Machine Learning for Remote Sensing Data Processing. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7, 1007–1011. https://doi.org/10.1109/JSTARS.2014.2311915

Tusa, L., Andreani, L., Khodadadzadeh, M., Contreras, C., Ivascanu, P., Gloaguen, R., Gutzmer, J., 2019. Mineral Mapping and Vein Detection in Hyperspectral Drill-Core Scans: Application to Porphyry-Type Mineralization. Minerals 9. https://doi.org/10.3390/min9020122

Van der Meer, F., 2004. Analysis of spectral absorption features in hyperspectral imagery. Int. J. Appl. Earth Obs. Geoinf. 5, 55–68. https://doi.org/10.1016/J.JAG.2003.09.001

van der Meer, F.D., van der Werff, H.M.A., van Ruitenbeek, F.J.A., Hecker, C.A., Bakker, W.H., Noomen, M.F., van der Meijde, M., Carranza, E.J.M., de Smeth, J.B., Woldai, T., 2012. Multi- and hyperspectral geologic remote sensing: A review. Int. J. Appl. Earth Obs. Geoinf. 14, 112–128.

https://doi.org/https://doi.org/10.1016/j.jag.2011.08.002

van der Walt, S., Schönberger, J., Nunez-Iglesias, J., Boulogne, F., Warner, J., Yager, N., Gouillart, E., Yu, T., the scikit-image contributors, 2014. scikit-image: Image processing in python. PeerJ 2:e453 . https://doi.org/https://doi.org/10.7717/peerj.453

Van Rijsbergen, C.J., 1979. Information Retrieval, 2nd ed. Butterworth-Heinemann.

Vapnik, V., 1963. Pattern recognition using generalized portrait method. Autom. Remote Control 24, 774–780.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4757-3264-1

Vasquez, M.L., Rosa-Costa, L.T., 2008. Geologia e recursos minerais do estado do Pará: texto explicativo. Escala 1:1.000.000. Programa Geol. do Bras. - PGB.

Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. IEEE Trans Neural Netw 11, 586–600.

Vieira, E.A.P., Saueressig, R., Siqueira, J.B., Silva, E.R.P., Rego, J.L., Castro, F.D.C., 1988. Caracterização geológica da jazida polimetálica do Salobo 3 A-Reavaliação, in: SBG, Congresso Brasileiro Geologia. pp. 97–111.

Vincent, R.K., Rowan, L.C., Gillespie, R.E., Knapp, C., 1975. Thermal-infrared spectra and chemical analyses of twenty-six igneous rock samples. Remote Sens. Environ. 4, 199–209.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. Pietro, Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D. V, Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.-L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T., Jones, T.R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., Vázquez-Baeza, Y., Contributors, S. 1. ., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Walsh, J.J., Watterson, J., 1993. Fractal analysis of fracture patterns using the standard box-counting technique: valid and invalid methodologies. J. Struct. Geol. 15, 1509–1512. https://doi.org/https://doi.org/10.1016/0191-8141(93)90010-8

Wang, D., Lagerstrom, R., Sun, C., Laukamp, C., Quigley, M., Whitbourn, L., Mason, P., Connor, P., Fisher, L., 2016. Automated vein detection for drill core analysis by fusion of hyperspectral and visible image data, in: 2016 23rd International Conference on Mechatronics and Machine Vision in Practice (M2VIP). pp. 1–6.

Wang, J., Zuo, R., Xiong, Y., 2020. Mapping Mineral Prospectivity via Semi-supervised Random Forest. Nat. Resour. Res. 29, 189–202. https://doi.org/10.1007/s11053-019-09510-8

Wang, J., Zuo, R., Xiong, Y., 2020. Mapping Mineral Prospectivity via Semi-supervised Random Forest. Nat. Resour. Res. 29, 189–202. https://doi.org/10.1007/s11053-019-09510-8

Wirth, K.R., Gibbs, A.K., Olszewski Junior, W.J., 1986. U-Pb ages of zircons from the Grao-Para Group and Serra dos Carajas Granite, Para, Brazil. Rev. Bras. Geociencias 16, 195–200.

Wu, H., Prasad, S., 2017. Convolutional recurrent neural networks for hyperspectral data classification. C 9. https://doi.org/10.3390/rs9030298

Wübbeke, J., Meissner, M., Zenglein, M.J., Ives, J., Conrad, B., 2016. Made in china 2025. Mercat. Inst. China Stud. Pap. China 2, 74.

Wulder, M.A., Kurz, W.A., Gillis, M., 2004. National level forest monitoring and modeling in Canada. Prog. Plann. 61, 365–381. https://doi.org/10.1016/S0305-9006(03)00069-2

Xavier, R.P., Soares Monteiro, L.V., Moreto, C.P.N., Pestilho, A.L.S., Coelho de Melo, G.H., Delinardo da Silva, M.A., Aires, B., Ribeiro, C., Freitas e Silva, F.H., 2012. The Iron Oxide Copper-Gold Systems of the Carajás Mineral Province, Brazil. SEG Spec. Publ. 16, 433–454.

Xiong, Y., Zuo, R., 2016. Recognition of geochemical anomalies using a deep autoencoder network. Comput. Geosci. 86, 75–82. https://doi.org/10.1016/j.cageo.2015.10.006

Xiong, Y., Zuo, R., 2017. Effects of misclassification costs on mapping mineral prospectivity. Ore Geol. Rev. 82, 1–9. https://doi.org/10.1016/j.oregeorev.2016.11.014

Xiong, Y., Zuo, R., 2018. GIS-based rare events logistic regression for mineral prospectivity mapping. Comput. Geosci. 111, 18–25. https://doi.org/10.1016/j.cageo.2017.10.005

Xiong, Y., Zuo, R., Carranza, E.J.M., 2018. Mapping mineral prospectivity through big data analytics and a deep learning algorithm. Ore Geol. Rev. 102, 811–817. https://doi.org/10.1016/j.oregeorev.2018.10.006

Xu, R., Wunsch, D., 2005. Survey of clustering algorithms. IEEE Trans Neural Netw 16, 645–678. https://doi.org/10.1109/TNN.2005.845141

Yang, Y., Zha, K., Chen, Y.-C., Wang, H., Katabi, D., 2021. Delving into Deep Imbalanced Regression. arXiv Prepr. arXiv2102.09554.

Yousefi, M., Kreuzer, O.P., Nykänen, V., Hronsky, J.M.A., 2019. Exploration information systems – A proposal for the future use of GIS in mineral exploration targeting. Ore Geol. Rev. 111, 103005. https://doi.org/https://doi.org/10.1016/j.oregeorev.2019.103005

Zadrozny, B., Elkan, C., 2001. Learning and Making Decisions when Costs and Probabilities Are Both Unknown, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’01. ACM, New York, NY, USA, pp. 204–213. https://doi.org/10.1145/502512.502540

Zadrozny, B., Langford, J., Abe, N., 2003. Cost-sensitive learning by cost-proportionate example weighting, in: Null. p. 435.

Zandiyyeh, F., Shayestefar, M.R., Ranjbar, H., Saadat, S., 2016. Prospectivity mapping of Iron Oxide-Copper-Gold (IOCG) deposits using support vector machine method in Feyzaabad area (east of Iran). J. Himal. Earth Sci. 49, 50–62.

Zhai, H., Zhang, H., Li, P., Zhang, L., 2021. Hyperspectral image clustering: Current achievements and future lines. IEEE Geosci Remote Sens Mag 9, 35–67.

Zhang, N., Zhou, K., Li, D., 2018. Back-propagation neural network and support vector machines for gold mineral prospectivity mapping in the Hatu region, Xinjiang, China. Earth Sci. Informatics 11, 553–566. https://doi.org/10.1007/s12145-018-0346-6

Zhang, W., Goh, A.T.C., 2016. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. Geosci. Front. 7, 45–52. https://doi.org/https://doi.org/10.1016/j.gsf.2014.10.003

Zhang, Y., Wang, Y., Chen, X., Jiang, X., Zhou, Y., 2022. Spectral-Spatial Feature Extraction with Dual Graph Autoencoder for Hyperspectral Image Clustering. IEEE Transactions on Circuits and Systems for Video Technology 1. https://doi.org/10.1109/TCSVT.2022.3196679

Zhang, Z., Zuo, R., Xiong, Y., 2016. A comparative study of fuzzy weights of evidence and random forests for mapping mineral prospectivity for skarn-type Fe deposits in the southwestern Fujian metallogenic belt, China. Sci. China Earth Sci. 59, 556–572. https://doi.org/10.1007/s11430-015-5178-3

Zhi-Hua Zhou, Xu-Ying Liu, 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans. Knowl. Data Eng. 18, 63–77. https://doi.org/10.1109/tkde.2006.17

Zhou, Y.T., Chellappa, R., 1988. Computation of optical flow using a neural network. IEEE 1988 Int. Conf. Neural Networks 71–78 vol.2.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a review. https://doi.org/10.1109/MGRS.2017.2762307

Zuo, R., Carranza, E.J.M., 2011. Support vector machine: A tool for mapping mineral prospectivity. Comput. Geosci. 37, 1967–1975. https://doi.org/10.1016/j.cageo.2010.09.014

Zuo, R., Xiong, Y., 2018. Big Data Analytics of Identifying Geochemical Anomalies Supported by Machine Learning Methods. Nat. Resour. Res. 27, 5–13. https://doi.org/10.1007/s11053-017-9357-0

Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B., Chen, Y., 2015. Convolutional

Recurrent Neural Networks: Learning Spatial Dependencies for Image Representation, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 18–26.

APÊNDICE A - GEOLOGICAL COMPLEXITY ALGORITHM

The algorithm used for calculating geological complexity in this study is given below. It was written in Python 3 by Elias Martins Guerra Prado. The algorithm is based on the program developed by Stephan Gardoll, published in Hodkiewicz (2003), to compute fractal dimension, which is based on the box-counting methodology of Hirata (1989). The advantage of this implementation compared to others is that the box-counting calculus is done using matrices (rasters) instead of spatial vectors, considerably reducing the time to compute the fractal dimension for each pixel.

The following Python libraries were used:

- Arcpy 1.4.1 (ArcGIS Pro 2.2.4)
- Geopandas 0.3.0
- Numpy 1.15.4
- Scipy 1.1.0

A brief description of the input parameters is given below:

1. Polygon shapefile with work area boundaries. This polygon defines the limits of the geological complexity map.

2. Grid spacing. Distance between the points of the fractal dimension grid. The location of each point of the grid defines the center of a fractal box. In this study, there is a 5 km spacing between grid points.

3. Polyline shapefile with geological contacts and/or faults/fractures. These linear features are sampled by the box-counting method to measure the fractal dimension. In this study, both geological contacts and faults/fractures were used.

4. Initial box size around each grid point. The initial pixel size used to rasterize the input line features. In this study, a 5 km initial box size was used.

5. Number of box count levels calculated. The number of times the initial pixel size will be halved. In this study, a five-level box count was used, resulting in box sizes of 5 km, 2.5 km, 1.25 km and 0.625 km

The program output is a raster produced by the interpolation of the fractal dimension for each grid point. The interpolation is done using a two-dimensional minimum curvature spline technique (the resulting surface passes precisely through the input points).

To compute the geological complexity the algorithm uses the following procedure:

1. First, the work area shapefile (input #1) is transformed into a raster with a pixel size equal to the grid spacing (input #2).

2. The work area raster is then converted to a point shapefile, where each point is placed in the center of each pixel. This shapefile is used as the grid for computing the fractal dimensions.

3. Then, the input line features (input parameter #3) is transformed into a binary raster, where pixels with value 1 indicate the presence of a feature. One raster is created for each box count level (input #5). The pixel size of the first raster is equal to the initial box size parameter (input #4). The pixel size of the other raster files is calculated by halving the initial pixel size N-1 times, where N is the number of box count levels (input #5).

4. The fractal dimension for each point of the grid is then determined by computing the sum of the pixels inside a square window, with side equal to two times the initial box size (input #4), centered at the grid point.

5. To determine the fractal dimension at each pixel, the slope of the line on a log-log plot of box size (pixel size) and box count result (sum of pixels for each pixel size). The value of the fractal dimension is then assigned to the corresponding grid point.

6. Finally, the fractal dimension grid is interpolated using a spline function.

```
# import Python libraries
import arcpy
from arcpy import env
from arcpy.sa import *
import geopandas as gpd
import numpy as np
from scipy import stats

#configure arcpy environment
env.overwriteOutput = True
env.extent="Default"

#input parameters
#1 SHAPE WITH WORK AREA:
work_area = 'path/to/shape'

#2 GRID SPACING SIZE (M)
grid_size = 5000

#3 FEATURES TO BE ANALYSED
input_shape = 'path/to/shape'

#4 Initial Fractal Size
init_size = 5000

#5 GET NUMBER OF RECURSIONS
Dimlimit = 5

#6 Output Fractal_dimension Raster
out_folder = 'path/to/output/folder'
output_fd_raster = out_folder + 'output_name.tif'

#7 temp files
out_work_area_raster = out_folder + 'work_area_raster.tif'  # work area raster
out_work_grid = out_folder + 'FD_grid.shp' # fractal dimension grid

# Function to calculate the slope of the fractal dimension line
def calc_Fractal_Dimension(x,y):
        slope, intercept, r_value, p_value, std_err = stats.linregress(np.log(x), np.log(y))
```

```
    fit_line = slope*np.asarray(np.log(x)) + intercept
    if math.isnan(slope):
        slope = 0
        return -1*slope

#Create grid
work_area_raster = arcpy.PolygonToRaster_conversion(
    work_area,
    "FID",
    out_work_area_raster,
    "MAXIMUM_AREA",
    "",
    grid_size)

work_grid = arcpy.RasterToPoint_conversion(
    work_area_raster,
    out_work_grid)

# Get work area limits
x_limits = []
y_limits = []

with arcpy.da.SearchCursor(work_area, ['SHAPE@']) as cursor:
    for row in cursor:
        array1 = row[0].getPart()
        for vertice in range(row[0].pointCount):
            pnt = array1.getObject(0).getObject(vertice)
            x_limits.append(pnt.X)
            y_limits.append(pnt.Y)

x_min = min(x_limits)
x_max = max(x_limits)
y_min = min(y_limits)
y_max = max(y_limits)
x_size = x_max - x_min
y_size = y_max - y_min

# Get grid limits
work_grid_gdf = gpd.read_file(out_work_grid)
```

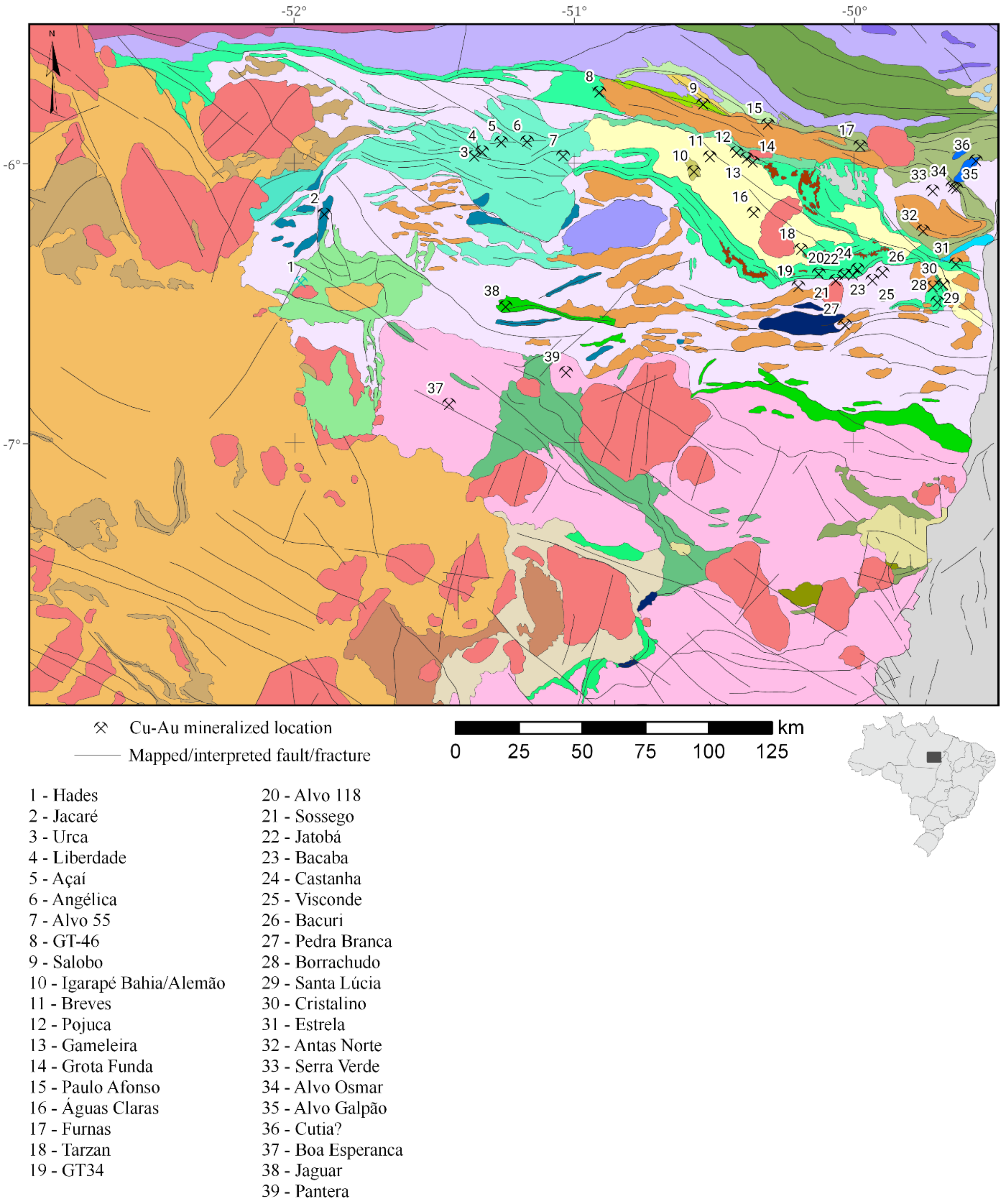
```
grid_max_x = work_grid_gdf.geometry.x.max()
grid_min_x = work_grid_gdf.geometry.x.min()
grid_max_y = work_grid_gdf.geometry.y.max()
grid_min_y = work_grid_gdf.geometry.y.min()

box_xsize = init_size
box_ysize = init_size

#Add temporary field to the input features shape to generate a binary raster
arcpy.CalculateField_management(
        input_shape,
        "temp",
        1)
#Create raster for each dimension
for dimension in range(Dimlimit):
        np_arr = np.zeros((math.ceil(((grid_max_y + init_size)-(grid_min_y - init_size)) / grid_size),
                  math.ceil(((grid_max_x + init_size) - (grid_min_x - init_size)) / grid_size)))

outConstRaster = arcpy.NumPyArrayToRaster(
        np_arr,
        arcpy.Point(grid_min_x - init_size, grid_min_y - init_size),
        grid_size,
        grid_size,
        0)

env.snapRaster = outConstRaster
env.extent = outConstRaster
out_raster = out_folder + 'input_rasterized_{}.tif'.format(dimension + 1)

arcpy.PolylineToRaster_conversion(
        Input_shape,
        "temp",
        out_raster,
        "",
        "",
        init_size*2/math.pow(2, dimension + 1))

#Calculate fractal dimension
grid_vals_ = []
```

```
for gi, row in enumerate(arcpy.da.SearchCursor(work_grid, ["SHAPE@XY"])):
        x, y = row[0]
        final_box_coord = []
        point_counts = []

        box_coord = [[[x - box_xsize, y - box_ysize],
                [x + box_xsize, y - box_ysize],
                [x + box_xsize, y + box_ysize],
                [x - box_xsize, y + box_ysize]]]

        coord_arr = [arcpy.Point(*coords) for coords in box_coord[0]]
        box_poly = arcpy.Polygon(arcpy.Array(coord_arr))

        for dimension in range(Dimlimit):
                out_raster = out_folder + 'input_rasterized_{}.tif'.format(dimension + 1)
                g_array = arcpy.RasterToNumPyArray(
                        out_raster,
                        arcpy.Point(*box_coord[0][0]),
                        math.pow(2, dimension + 1),
                        math.pow(2, dimension + 1),
                        0)
                point_counts.append(g_array.sum())
        x_sizes = [box_xsize / math.pow(2, x) for x in range(Dimlimit)]
        F_dimension = print_Fractal_Dimension_curve(x_sizes, point_counts)
        grid_vals_.append(F_dimension)

#Assign fractal dimension values to grid shape file
work_grid_gdf = gpd.read_file(out_work_grid)
work_grid_gdf['FD'] = grid_vals_
work_grid_gdf.to_file(out_work_grid)

#Interpolate fractal dimension grid and save as raster
arcpy.CheckOutExtension("spatial") # check for spatial analyst extension
out_fd_cell_size = 125 # output geological complexity raster cell size


#Spline function parameters
sline_type = "REGULARIZED" #
spline_weights = 0.1
```

```
n_points = 12
outSpline = Spline(out_work_grid, "FD", out_fd_cell_size, sline_type, spline_weights, n_points)
outSpline.save(output_fd_raster)
```

APÊNDICE B – NAMES OF MINERALIZED LOCATIONS AT CARAJÁS MINERAL PROVINCE

Names of mineralized locations used for training the prospectivity model at Chapter 5.



1 - Hades
2 - Jacaré
3 - Urca
4 - Liberdade
5 - Açaí
6 - Angélica
7 - Alvo 55
8 - GT-46
9 - Salobo
10 - Igarapé Bahia/Alemão
11 - Breves
12 - Pojuca
13 - Gameleira
14 - Grota Funda
15 - Paulo Afonso
16 - Águas Claras
17 - Furnas
18 - Tarzan
19 - GT34
20 - Alvo 118
21 - Sossego
22 - Jatobá
23 - Bacaba
24 - Castanha
25 - Visconde
26 - Bacuri
27 - Pedra Branca
28 - Borrachudo
29 - Santa Lúcia
30 - Cristalino
31 - Estrela
32 - Antas Norte
33 - Serra Verde
34 - Alvo Osmar
35 - Alvo Galpão
36 - Cutia?
37 - Boa Esperanca
38 - Jaguar
39 - Pantera