

# Deriving big geochemical data from high-resolution remote sensing data via machine learning: Application to a tailing storage facility in the Witwatersrand goldfields



Steven E. Zhang <sup>a,b,\*</sup>, Glen T. Nwaila <sup>b</sup>, Julie E. Bourdeau <sup>a,b</sup>, Yousef Ghorbani <sup>c,d</sup>, Emmanuel John M. Carranza <sup>e</sup>

<sup>a</sup> Natural Resources Canada, Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario, K1A 0E8, Canada

<sup>b</sup> Wits Mining Institute WMI, University of the Witwatersrand, Private Bag 3, 2050, Wits, South Africa

<sup>c</sup> Department of Civil, Environmental and Natural Resources Engineering, Luleå University of Technology, SE 97187, Luleå, Sweden

<sup>d</sup> School of Chemistry, University of Lincoln, Joseph Banks Laboratories, Green Lane, Lincoln, Lincolnshire, LN6 7DL, United Kingdom

<sup>e</sup> Department of Geology, University of the Free State, 205 Nelson Mandela Dr, Bloemfontein, 9301, South Africa

## ARTICLE INFO

### Keywords:

Remote sensing  
Big geochemical data  
Machine learning  
Tailing storage facilities  
Witwatersrand Basin  
Dry labs

## ABSTRACT

Remote sensing data is a cheap form of surficial geoscientific data, and in terms of veracity, velocity and volume, can sometimes be considered big data. Its spatial and spectral resolution continues to improve over time, and some modern satellites, such as the Copernicus Programme's Sentinel-2 remote sensing satellites, offer a spatial resolution of 10 m across many of their spectral bands. The abundance and quality of remote sensing data combined with accumulated primary geochemical data has provided an unprecedented opportunity to inferentially invert remote sensing data into geochemical data. The ability to derive geochemical data from remote sensing data would provide a form of secondary big geochemical data, which can be used for numerous downstream activities, particularly where data timeliness, volume and velocity are important. Major benefactors of secondary geochemical data would be environmental monitoring and applications of artificial intelligence and machine learning in geochemistry, which currently entirely relies on manually derived data that is primarily guided by scientific reduction. Furthermore, it permits the usage of well-established data analysis techniques from geochemistry to remote sensing that allows useable insights to be extracted beyond those typically associated with strictly remote sensing data analysis. Currently, no generally applicable and systematic method to derive chemical elemental concentrations from large-scale remote sensing data have been documented in geosciences. In this paper, we demonstrate that fusing geostatistically-augmented geochemical and remote sensing data produces an abundance of data that enables a more generalized machine learning-based geochemical data generation. We use gold grade data from a South African tailing storage facility (TSF) and data from both the Landsat-8 and Sentinel remote sensing satellites. We show that various machine learning algorithms can be used given the abundance of training data. Consequently, we are able to produce a high resolution (10 m grid size) gold concentration map of the TSF, which demonstrates the potential of our method to be used to guide extraction planning, online resource exploration, environmental monitoring and resource estimation.

## 1. Introduction

Remote sensing provides some of the cheapest and sometimes open exploration data. A diversity of remote sensing platforms includes satellite-based remote sensing and are being leveraged for many purposes (Rajesh, 2004; Song et al., 2020; Ge et al., 2020; Shirmard et al., 2022). From a trans-disciplinary perspective (a data-centric manner that

is unspecific to any particular discipline), these types of data are modern (machine-readable and of high quality) and abundant (combined with sufficient velocity to be near or at the level of big data). Such types of data are a critical input into data-centric laboratories, which can rapidly consume data and depend on the availability of timely geoscientific data (e.g., dry labs; Ghorbani et al., 2022, 2023). Other benefits of remote sensing data include their broad (spatial and temporal) coverage and

\* Corresponding author. Natural Resources Canada, Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario, K1A 0E8, Canada.  
E-mail address: [ezhan053@uottawa.ca](mailto:ezhan053@uottawa.ca) (S.E. Zhang).

their constant improvement in capability (e.g., spatial and spectral resolution). For example, the Landsat satellites through Landsat 1 to 8 have continuously and iteratively improved in their resolution and sensor capability (NASA 1997; Barsi et al., 2014). We expect that improvements in satellite-based remote sensing will continue. However, although both surficial geochemistry and remote sensing probe the surface of the Earth, their data contain highly dissimilar content. In the variable domain, spectral bands are composite responses of minerals and chemical elements (and other molecular responses, e.g., water, anthropogenic and biological responses), whereas geochemical data typically describe elemental or oxide concentrations. Remote sensing is already desirable for geological mapping and/or mineral exploration, particularly during the reconnaissance stage (Diaz-Rodriguez et al., 2021) and essential where physical access is difficult, seasonal, or where political and environmental sensitivities exist (Ciampalini et al., 2013; Ali et al., 2015).

Inverting remote sensing data into secondary geochemical data creates many benefits, the most relevant ones realized by this study are: (1) the ability to directly infer chemical information from remote sensing data; (2) the ability to supply secondary geochemical data at a spatial resolution, velocity and volume far surpassing that of primary geochemical data; and (3) a vast reduction in cost of secondary geochemical data compared to that of the primary type. These benefits are crucial to sustain the use of artificial intelligence in geosciences, because primary geoscientific data is expensive, opportunistic (Demetrides et al., 2018; Dramsch, 2020) and impractical or impossible to scale-up towards big data. Although geochemical data are important to many uses, such as prospectivity mapping (Zuo, 2020; Lawley et al., 2021), such data are generated through manual and full-contact processes (e.g., Govett, 1983; Friske and Hornbrook, 1991; Moon et al., 2006), and were engineered according to the philosophy of scientific reduction (towards increasing accuracy and precision, e.g., from spectral analysis to mass spectrometry; Cohen et al., 2010). Reductive data and machine learning is a poor match, because machine learning treats data at the system level. For example, the deep learning task to classify animal species uses images of animals in a variety of contexts and perspectives (e.g., Hou et al., 2020; Prudhivi et al., 2021). This task does not rely on expensive reductive data, such as genetic sequencing data. As geoscientific data consumers shift from in-domain to trans-disciplinary experts (Ghorbani et al., 2022, e.g., Lawley et al., 2021), data generation must also evolve accordingly (Ghorbani et al., 2023) and ideally acquire some big data characteristics (He et al., 2022). The awareness of changes in data requirements induced by the adoption of artificial intelligence in the geoscientific community is immature but growing (Chen and Lin, 2014; Karpatne et al., 2018; Zhang and Lu, 2021; He et al., 2022).

Dry tailing storage facilities (TSFs) present an ideal method development environment for the inversion of remote sensing data into geochemical data. This is because: (1) TSF chemical and mineral compositions are partly man-made; (2) geochemical surveys are often conducted to profile resource distribution and guide waste reuse; and (3) they exhibit a surficial signature and are usually remote from civic infrastructure, which favour remote sensing and disfavour manual survey methods (Ciampalini et al., 2013). There are other motivations that include: (1) tailings could become an unconventional resource (Nwaila et al., 2021a, 2021b) of critical raw materials (Rachidi et al., 2021 and references therein), particularly historic tailings because extraction methods were less selective and orebodies were generally higher in grade (Prior et al., 2012; Nwaila et al., 2019); (2) unlike conventional mineral resources, TSFs contain low resource grades and the cost-to-benefit ratio for geochemical surveys is less favourable; (3) managed TSFs require ongoing environmental monitoring and evaluation; and (4) the gridded nature of remote sensing implies that downstream geostatistical treatments may be unnecessary and, therefore, an additional source of estimation error could be eliminated. For these reasons, satellite remote sensing data is likely useful.

Analysis of remote sensing imagery includes techniques such as false colour composites, image processing, band ratios and principal component analysis (Ciampalini et al., 2013; Beiranvand Pour and Hashim, 2014). These methods leverage the differential spectral response of minerals and generally constituents of land cover across the electromagnetic spectrum. Band ratio maps that proxy areal lithology and mineralogy (Agar and Coulter, 2007; Asokan et al., 2020; Marghany, 2022) have been combined with downstream processing to identify objects, including mineral deposits (Amer et al., 2010; Noori et al., 2019). Therefore, band ratio analysis is a type of knowledge-driven remote sensing data-inversion to derive mineral composition. Knowledge-driven inversion pre-suppose the existence of local knowledge in the form of target minerals and their spectral responses (Marghany, 2022; e.g., Cao et al., 2020). Inversion of remote sensing data to geochemistry is possible, where the context is suitable (e.g., good coverage and insignificant land cover). This relies on the fact that where rock and mineral compositions are similar, spectral responses would also be similar, and elemental concentrations are therefore predictable (Marghany, 2022). Inverting remote sensing data into geochemistry using only knowledge seems intractable, as the first principles that spawned band ratio analysis cannot unambiguously relate general spectral reflections to material chemistry, because spectral response encodes information at the macroscopic scale (of photon-mineral interactions) and not the microscopic scale (of elements) (Marghany, 2022). However, the absence of a general knowledge-driven solution does not eliminate specific solutions. A different perspective is the data-driven perspective, which treats the inversion problem as an inferential modelling problem (e.g., Guo et al., 2016; Shen et al., 2019; Cheng et al., 2021; Xiao and Wan, 2021). Although the physics of spectral response are scale-invariant from the macroscopic scale upwards, the inferential approach seems to be better developed in material stream profiling, within the extraction to beneficiation stages (e.g., Xiao et al., 2019, 2021). The desire to use non-contact sensors result from the harsh nature of industrial processes, although sensors vary, e.g., near-infrared spectroscopy (Bath, 2007; Cao, 2013), mid-infrared Fourier transform spectroscopy (Griffiths et al., 2007) and Raman spectroscopy (Rudolph and Heftner, 2009).

Machine learning is a discipline concerned with machines' ability to mimic biological learning (Mitchell, 1997). Machine learning algorithms learn by modelling patterns encoded by data, which are useful for various tasks, including predictive modelling, anomaly detection, dimensionality reduction, etc. A common use of machine learning in remote sensing is classification of land use or cover, vegetation and water (Mather and Tso, 2009; Sehgal, 2012; Al-doski et al., 2013; Madhuanand et al., 2021) and to discover mineral deposits (Cracknell and Reading, 2014; Harvey and Fotopoulos, 2016; Bachri et al., 2019; Chakouri et al., 2020; Song et al., 2020; Madhuanand et al., 2021). The combination of remote sensing and machine learning can overcome some drawbacks of knowledge-driven approaches, such as the subjectivity of the interpreter, model and the (sometimes) simplicity of manually derived models (e.g., band composites), and issues that are specific to remote sensing (Beiranvand Pour et al., 2016, 2019; Lary et al., 2016; Dai et al., 2017). Inferential inversion of remote sensing data into geochemical data is rare in literature, but has been documented for a range of metallic (e.g., Fe, Cu, As, Hg, Mo, Pb and Zn) and non-metallic (e.g., P) elements (e.g., Guo et al., 2016; Shen et al., 2019; Cheng et al., 2021; Xiao et al., 2021; Xiao and Wan, 2021). All approaches (Guo et al., 2016; Shen et al., 2019; Xiao et al., 2021; Xiao and Wan, 2021) employ labelled data of high quality (e.g., abundant, captures necessary relationships and is representative of the deployment area). Although the abundance and modernity of remote sensing data suit machine learning-based solutions to the remote sensing inverse problem, the limitation is the paucity of geochemical data. The manner in which the existing approaches overcome geochemical data scarcity is their key differentiator. The use of parametrized models, particularly multilinear regression (e.g., Cheng et al., 2021) reduces the amount of

data required relative to machine learning models with more degrees of freedom. Another popular approach is to employ algorithms that are training data-efficient, such as variations of the extreme learning machine (Xiao et al., 2021; Xiao and Wan, 2021), which does not necessitate backpropagation. However, limited training data precludes the benefit of a broader range of applicable algorithms.

The solution to geochemical data scarcity is not unique. One solution is data augmentation (e.g., Shorten et al., 2019), which serves many functions for predictive modelling and has been used with success in other fields to overcome data scarcity (e.g., Bird et al., 2020). In our case, training data were augmented through the use of geostatistical modelling. This process transformed sparse geochemical data into a high-resolution spatial model, which when fused with the remote sensing data, created a sufficient amount of training data for a variety of machine learning algorithms. Geostatistically augmented data are in principle well-matched with remote sensing data, in the sense that both types of data capture spatially averaged responses and are gridded. Therefore, geostatistical modelling is also a particular solution to the change-of-support problem (e.g., representation of an area by point-wise samples) and enables rigorous geospatial data fusion (Gelfand et al., 2001; Gotway and Young, 2002). Solutions to overcome data scarcity can enable more widespread use of artificial intelligence in geosciences.

This study demonstrates a method using machine learning, augmented legacy geochemical data and satellite-based remote sensing data to derive high-resolution geochemical maps from remote sensing data. We use a combination of Landsat-8 and Sentinel-2 satellite data and legacy geochemical data for rapid metal content prediction from a TSF in the Witwatersrand goldfields, South Africa. This, in turn, will assist with mine waste valorisation, enable online resource assessment during extraction and continuous monitoring of spatio-temporal changes, and create geochemical maps. In addition, our approach enables the supply of geochemical data as truly big data.

## 2. Methodology

### 2.1. Remote sensing data

The remote sensing data used in this study were sourced from two satellites due to their differential temporal coverage of the TSF. The satellites are the Landsat-8 (hereon referred to as Landsat) and the Copernicus Programme's Sentinel-2 (hereon referred to as Sentinel). Extraction of the TSF pre-dated the launch of the Sentinel satellites, which created a large spatial discrepancy between the Sentinel data and the geochemical data, thus preventing spatial data fusion (integration) (Abdalla, 2016) using the Sentinel data. Instead, Landsat data were used to optimize the spatial alignment. After optimizing the spatial alignment of the geochemical and remote sensing datasets, the higher-resolution Sentinel data were used for predictive modelling. Although hyperspectral data are also available, we employed multispectral data for two reasons: (1) multispectral data have a higher spatial resolution, which is important in our case to maximize the amount of training data; and (2) hyperspectral data's number of bands implies that the amount of machine learning features are higher, and therefore data density is lower in the feature space. If we had employed hyperspectral data from similar sources as compared with the multispectral data, the reduction of spatial resolution from 10 m to 30 m implies that training data would be approximately 1/9th of the original volume. This would be greatly exacerbated by the increase in the number of bands in hyperspectral data relative to multispectral data due to the curse of dimensionality, if each band was treated as a machine learning feature (Trunk, 1979; Chandrasekaran and Jain, 1974; McLachlan, 2004). In other applications where these factors are not problematic, then hyperspectral data may be used instead of multispectral data.

For information on the Landsat-8 satellite, see USGS (2022) and for its Operational Land Imager and Thermal Infrared Sensor, see Irons et al. (2012) and Roy et al. (2014). For information on the Sentinel satellites

**Table 1**

Capabilities of landsat 8/9 operational land imager and thermal infrared sensor (USGS, 2022).

Bands, common abbreviations and descriptions	Wavelength (μm)	Resolution (m)
Band 1 - Coastal aerosol	0.43–0.45	30
Band 2 - Blue	0.43–0.45	30
Band 3 - Green	0.53–0.59	30
Band 4 - Red	0.64–0.67	30
Band 5 - Near-infrared (NIR)	0.85–0.88	30
Band 6 - Short-wave infrared 1 (SWIR 1)	1.57–1.65	30
Band 7 - Short-wave infrared 2 (SWIR 2)	2.11–2.29	30
Band 8 - Panchromatic	0.50–0.68	15
Band 9 - Cirrus	1.36–1.38	30
Band 10 - Thermal infrared 1 (TIRS 1)	10.6–11.19	100
Band 11 - Thermal infrared 2 (TIRS 2)	11.50–12.51	100

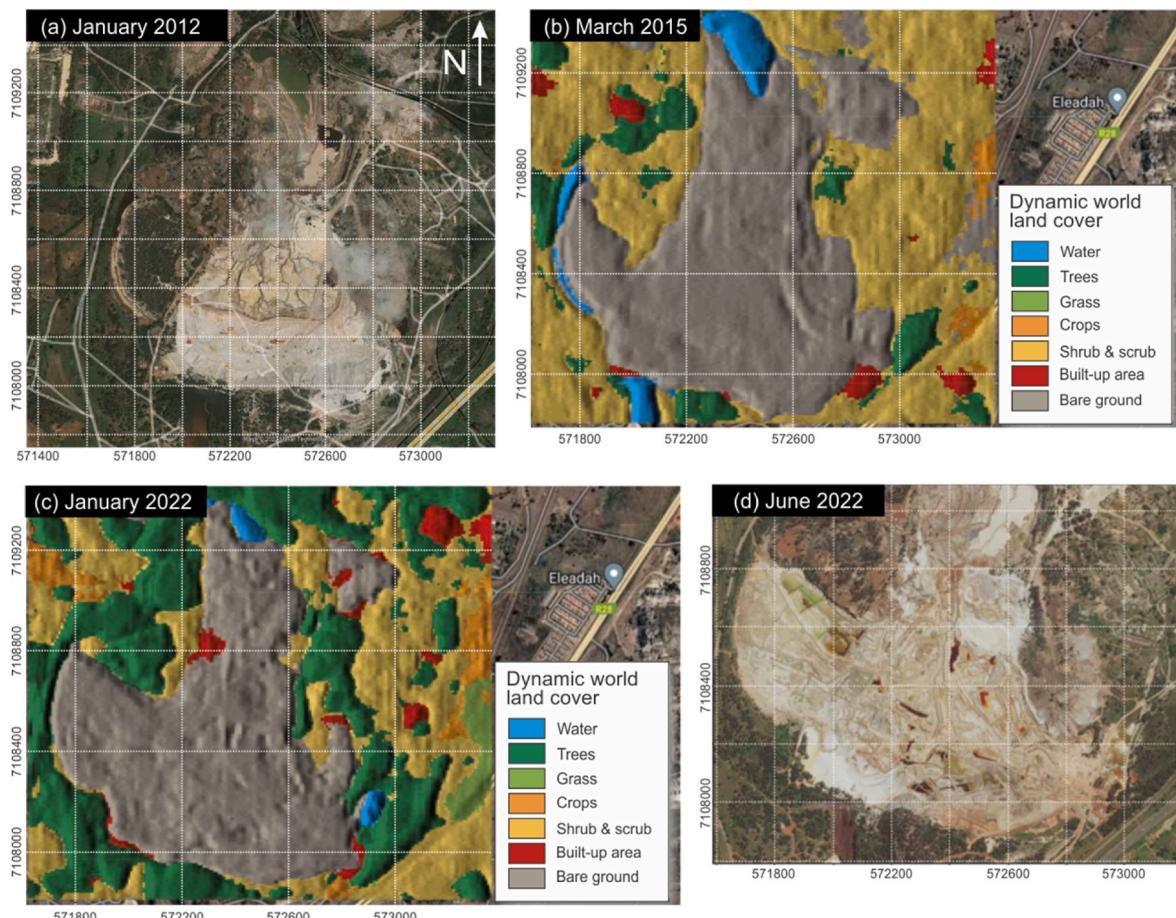
(2 A and 2 B), see Ge et al. (2018) and for its single multispectral sensor, see Vaiopoulos and Karantzalos (2016) and Park et al. (2017). Spectral and spatial capabilities of the Landsat 8/9 and Sentinel spectral instruments are summarized in Tables 1 and 2. Sentinel data surpass data from conventional geological multispectral imaging satellites, such as the Landsat and the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), in terms of spectral, spatial and temporal resolution from a visible and near-infrared (VNIR) to short-wave infrared (SWIR) range (Beiranvand Pour et al., 2016, 2019). In addition, Sentinel is the first civil Earth observation mission to be able to display three bands in the red edge. The 10-m spatial resolution within the 13 bands enables ongoing coordination with the SPOT-5 and Landsat missions (Van der Meer et al., 2014; Van der Werff and van der Meer, 2015). Radiometric resolution of both Landsat and Sentinel satellites is 12 bits. For both satellites, we make use of the top-of-atmosphere subset, which is cloud-free and has undergone atmospheric correction to produce a surface reflectance product using the European Space Agency's Sen2Cor processing algorithm's ATCOR model (ESA; Kristoffersson and Karathanassi, 2020). All images were cloud-free level 1 T (L1T, terrain-corrected) and were collected from the USGS Earth Resources Observation and Science Centre (USGS-EROS at <http://earthexplorer.usgs.gov/>).

The proof-of-concept area centres around a legacy TSF in the Witwatersrand goldfields, South Africa, with some surrounding vegetation and anthropogenic structures (Fig. 1). The Sentinel and Landsat data cover an area of 1.86 km<sup>2</sup> (458.39 acres). Both datasets used the Universal Transverse Mercator (UTM) map projection and the World Geodetic System 84 (WGS84) datum. The Landsat image used in this study was collected on the May 5, 2013 (metadata: coordinate reference system (CRS) from European Petroleum Survey Group (EPSG) 32,735, Transform: Affine [30, 0, 536,985, 0, -30, -2765,085]; ID: LANDSAT/LC08/C01/T1\_SR/LC08\_170, 078\_20,130,505; Version: 1522720687378920.0.) The Sentinel image used in this study was collected on the August 24, 2015 (metadata: coordinate reference system (CRS) from European Petroleum Survey Group (EPSG) 32,735, Transform: Affine [20, 0, 499,980, 0, -20, 7,200,040]; ID: COPERNICUS/S2/20150824T082656\_20150824T082659\_T35JNM; Version: 1,618,001,523,716,097; Data taken identifier: GS2A\_20150824T082656\_000890\_N02.04; 'SPACECRAFT\_NAME': 'Sentinel-2A'). The bands of the Landsat data employed in this study include bands 1 through 7. The panchromatic band (band 8), cirrus cloud band (band 9) and the TIRS bands were excluded as their resolutions are substantially lower than those of bands 1 to 7. Similarly, bands 1 to 12, except for 1, 9 and 10, were used for the Sentinel data.

The remote sensing data were pre-processed using the Rasterio and GEEMAP Python libraries (Wu et al., 2019; Wu, 2020) and consisted of radiometric calibration, atmospheric correction, geometric correction through an affine translation and noise removal. For mineral exploration and lithological classification, radiometric correction is important to minimize pixel error in spectral data (Rajendran and Nasir, 2014; examples in Cooley et al., 2002; Salem et al., 2016). Radiometric

**Table 2**Spectral bands of the Sentinel-2 satellites (Clerc et al., 2022; <https://sentinel.esa.int/web/sentinel/missions/sentinel-2>).

Bands and descriptions	Sentinel-2A		Sentinel-2B		Spatial resolution (m)
	Central wavelength (nm)	Bandwidth (nm)	Central wavelength (nm)	Bandwidth (nm)	
Band 1 – Coastal aerosol	442.7	21	442.2	21	60
Band 2 – Blue	492.4	66	492.1	66	10
Band 3 – Green	559.8	36	559.0	36	10
Band 4 – Red	664.6	31	664.9	31	10
Band 5 – Vegetation red edge	704.1	15	703.8	16	20
Band 6 – Vegetation red edge	740.5	15	739.1	15	20
Band 7 – Vegetation red edge	782.8	20	779.7	20	20
Band 8 – NIR	832.8	106	832.9	106	10
Band 8 A – Narrow NIR	864.7	21	864.0	22	20
Band 9 – Water vapour	945.1	20	943.2	21	60
Band 10 – SWIR – Cirrus	1373.5	31	1376.9	30	60
Band 11 – SWIR	1613.7	91	1610.4	94	20
Band 12 – SWIR	2202.4	175	2185.7	185	20

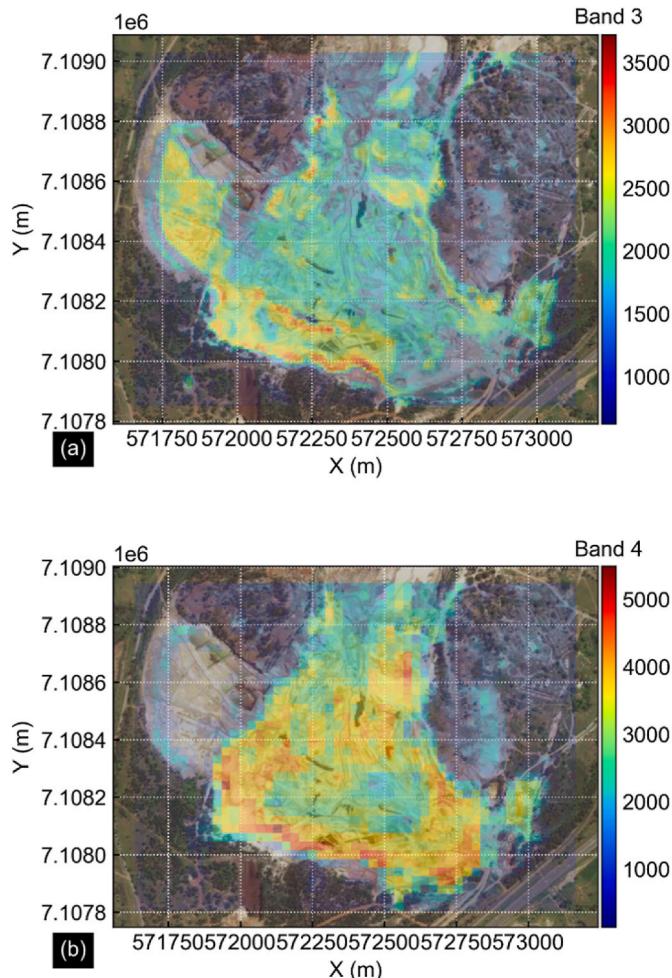
**Fig. 1.** The tailing storage facility (TSF) used in this study was imaged in (a) 2012, (b) 2015 with land cover classification overlay, (c) 2022 with land cover classification overlay, and (d) 2022. The fan-shaped feature stretches left to the right (most visible in (d) is the main storage area).

calibration optimized remote sensing images for best radiance, reflectance or brightness temperatures. For this purpose, the remote sensing data were processed using the first-principles-based Fast Line-of-Sight Atmospheric Analysis of Spectral Hypercubes (FLAASH) method, which can correct wavelengths up to 3 m in the visible, near-infrared and short-wave infrared ranges (ENVI, 2009). For our data, the tropical atmospheric and rural aerosol models were used for the FLAASH algorithm (Manakos et al., 2011). Therefore, spectrometer radiance data was converted to reflectance data. Band rescaling is a common procedure for remote sensing data, if the data would be used for band ratio analysis (e.g., Pohl and Van Genderen, 2014; Salem et al., 2016).

However, we employed band-amplitude normalization as a part of machine learning (feature rescaling). For a visualization of the remote sensing data, see Fig. 2.

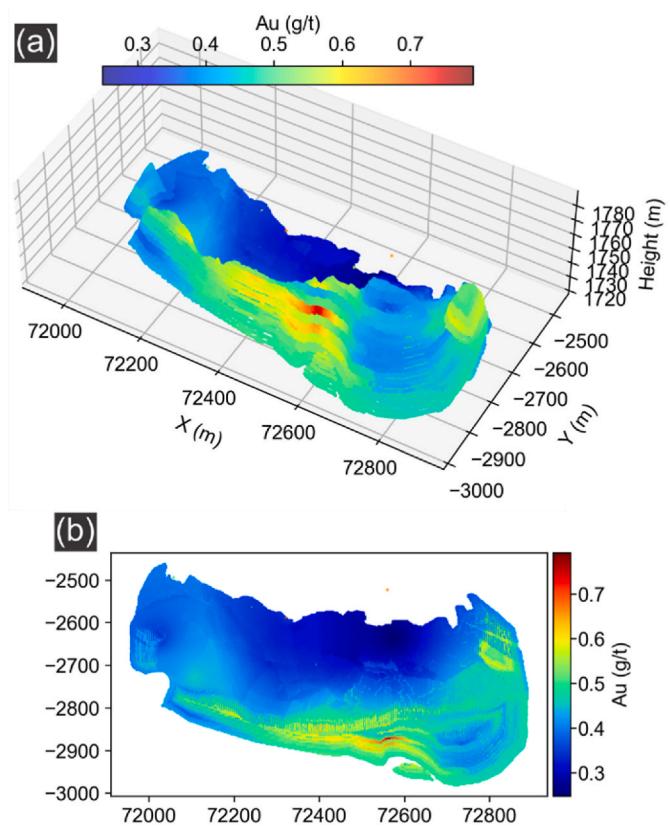
## 2.2. Geochemical data source and training data

Geochemical data containing concentrations of the target (gold [Au]) need to be integrated with remote sensing data, such that the fused data are in the form of paired features-and-label format (He et al., 2010; e.g., Aiazz et al., 1999; Bannari et al., 2016), in preparation for predictive modelling. Data fusion has been used for other similar tasks,

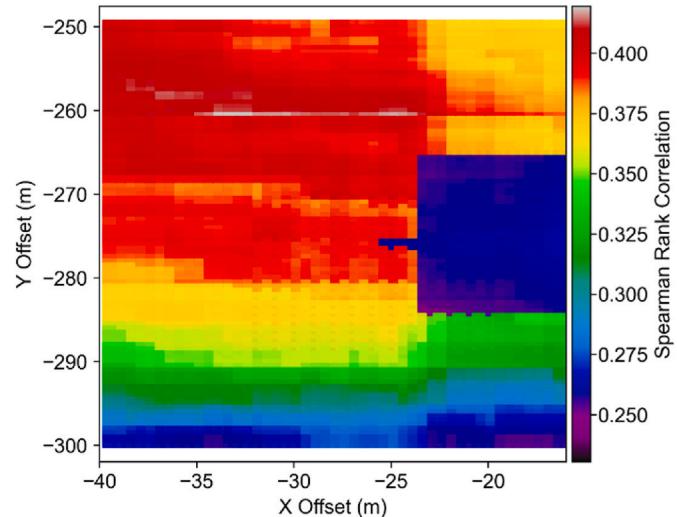


**Fig. 2.** Visualization of (a) Sentinel, and (b) Landsat remote sensing data overlaid on top of local satellite imagery. Notice the higher spatial resolution of the Sentinel data compared to the Landsat data. Additionally, the response from the north-western portion of the tailing body is mostly missing in the Landsat data in (b) due to heavy vegetation (see Fig. 1a). Not all bands are shown for brevity.

such as lithological mapping (Wang et al., 2021). Before data fusion (Abdalla, 2016), remote sensing and geochemical data exist in different datasets with different projections and a spatial offset. Sibanye-Stillwater supplied the geochemical data in two forms: (1) drill core data of Au assays; and (2) a geostatistical resource model derived from (1) and additional grab sampling. The data characterize the resource contained within selected areas of the TSF. The geostatistical model augmented the primary drill core (and grab sample, which were unavailable for this study) data by providing a high-resolution 3D point cloud that is resolution-biased towards the shallower depths. The augmented geochemical data are a variable density 3D resource model composed of a scatter cloud of 276,622 gridded points at a resolution of 0.5 m (Fig. 3a). Due to a confidentiality agreement, the coordinates were partially altered to protect the proprietary information of Sibanye-Stillwater. Additionally, the coordinates provided were in a local coordinate reference system based on the Gauss Conformal projection used by the business (Fig. 3a). Due to the dimensionality and projection mismatches, the 3D model cannot be directly integrated/fused with the 2D remote sensing data and hence, we performed dimensional reduction on the 3D model to produce a 2D model by segmenting the 3D model into discrete 3D vertical columns at a resolution of 1 m (both X and Y coordinates) and skimming the data from the top 10 cm of each column (Fig. 3b). This produced a thin surficial model,



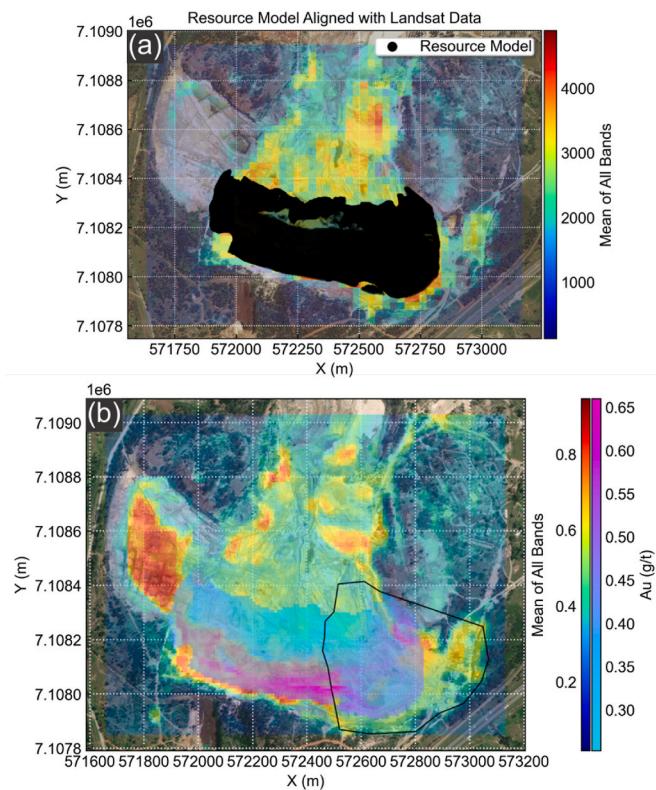
**Fig. 3.** The (a) 3D, and (b) 2D models of the Au resource within the tailing storage facility. Both X- and Y-coordinates are in a local proprietary coordinate reference system.



**Fig. 4.** Results of the grid search using the Spearman rank correlation coefficient to determine an optimal match between the 2D resource model and the remote sensing data. The Y offset is in addition to the constant  $2.89 \times 10^6$  m observed.

which is then vertically averaged to produce a 2D resource model. The 2D resource model retained 63.65% of the original 3D model data, amounting to 176,071 data points.

Conversion of the local CRS to UTM coordinates was straightforward except for a constant offset detected in the resource model relative to the remote sensing data. Qualitatively, the offset was on the order of  $2.89 \times$



**Fig. 5.** (a) Visualization of the best alignment of the 2D resource model (Fig. 3b) with the Landsat remote sensing data. (b) Data integration using the 2D resource model (Fig. 3b) and the normalized Sentinel remote sensing data (e.g., Fig. 2a). The polygon depicts the portion of the resource model that was discarded after data integration, as this area had already been extracted.

$10^6$  m in the Y direction and decametre in the X-direction. However, due to the private nature of the data, we could not verify the exact offset. Instead, we performed a grid search to maximize the match of the 2D model with the remote sensing data using the Spearman rank correlation, averaged across all bands (Fig. 4). Note that this is a heuristic approach and was necessary given the obfuscated context of the supplied geochemical data. This procedure is not a requirement of our method in general. The grid search was performed using the Landsat data because it is temporally the best match with the resource model. Although the Landsat data matched the best with the resource model, it was lower in resolution (30 m resolution) compared with the Sentinel data (10 m resolution). In addition, the unextracted north-western portion of the TSF was under heavy vegetation as captured by the Landsat data (Figs. 1a and 5a, upper left corner, notice the lack of remote sensing response over the north-western part of the TSF). The Sentinel data are a more current representation of the state of the TSF. Once the 2D resource model was optimally matched with the remote sensing data, data integration was performed using the resource model and the Sentinel data at the native resolution of the Sentinel dataset (10 m grid size). This resulted in the creation of 17,346 spatially-matched pixels containing both the band amplitudes of the Sentinel data and unweighted averages of the Au grades within each grid cell over the study area. Thereafter, because portions of the TSF had been extracted when the Sentinel satellite was launched, we removed the mismatched portion (Fig. 5b). After removing the mismatched data, the dataset contained 14,721 data points, 84.87% of the original set.

### 2.3. Machine learning-based predictive modelling

Machine learning algorithms are used to model relationships between the remote sensing response (features) and geochemical

**Table 3**

Parameter grid for employed machine learning algorithms.

Algorithm	Parameter Grid
kNN	$k = \{2 \text{ to } 16 \text{ in intervals of } 1\}$
SVM	$C = \{10, 100, 200, 400, 600, 800, 1000\}$ , $\epsilon = \{0.00001, 0.0001, 0.001, 0.01, 0.1, 0.5, 1.0\}$ , kernel = {linear, radial basis function (RBF)}
Random forest	Ensemble size = {1000, 1500}; maximum depth = {9, 11, 13, 15, unlimited}, maximum number of features = {2, 3, 4, 5, 6, 7, 8, unlimited}, minimum number of samples for a split = {2, 3, 4, 5, 6}, minimum number of samples for a leaf = {1, 2, 3, 4, 5, 6}
AdaBoost	Number of classifiers = {100, 200, 300}, base algorithm = decision tree with the same parameter grid as the random forest algorithm
ANN	$\alpha = \{0.0001, 0.001, 0.01, 0.1, 1.0\}$ , activation = {identity, logistic, tanh, relu}, learning rate = {constant, inverse scaling, adaptive}, maximum number of iterations = {1500, 2500}, hidden layer sizes = {100, 150, 200}

concentration (data label). There are many suitable supervised algorithms available that range in algorithmic complexity and rationale. Simpler algorithms construct models or produce results that are more humanly interpretable. Complex algorithms are less explainable by design but, under some circumstances, may produce higher-performing models. We intentionally explored several characteristic shallow learning algorithms for scientific purposes, where predictive modelling performance is not a singular essential outcome, but also method explainability and replicability. The algorithms that were explored in this study are: k-nearest neighbours (kNN), support vector machine (SVM), random forest, adaptive boosting of decision trees (AdaBoost), and artificial neural networks (ANN). These algorithms represent a diversity of potentially useful approaches that include a simple non-parametric approach (kNN), a parametric approach (SVM), ensemble and boosted approaches (random forest and AdaBoost, respectively), and a simple neural network-based approach (ANN). Details of these algorithms and their hyperparameters are fully described in Zhang et al. (2021, 2022). In our case, machine learning algorithms were used to build models that capture relationships between spectral response and Au grade. The spectral response data for each band was rescaled using feature rescaling, such that the rescaled data spans a numerical range from 0 to 1 (Fig. 5b). Algorithm and model selection, and performance assessments were conducted using 10-fold cross-validation. We employed the coefficient of determination (CoD or  $R^2$ ) metric for model tuning and selection and, additionally, employed the median absolute error (MedAE) and the mean absolute percentage error (MAPE) to profile cross-validation results. The training data were the entire dataset resulting from spatial data integration of the resource model and the Sentinel data (where Au grades exist). Aside from cross-validation using the training data, we also deployed the best models to the remainder of the TSF, where there is no known resource information. The parameter grid for the algorithms is given in Table 3. The data pre-processing, machine learning and visualization were implemented in Python and particularly for machine learning, using the Scikit-Learn library (Buitinck et al., 2013).

## 3. Results

After the training data generation, the relationship between various bands and the Au grade was complex and generally nonlinear (Fig. 6). As seen in Fig. 6, there is a tendency towards increasing Au grade with increasing band amplitudes. However, this relationship is not monotonic, because the response may be reversed at high band amplitudes. For example, there is a negative correlation between band amplitudes and gold grade at high band amplitudes (between 2215 and 3375 band amplitude and 0.4–0.6 g/t in Fig. 6c). In particular, there seem to be two discernible clusters of data, with one cluster occurring at relatively lower band amplitudes and the other higher (Fig. 6). Based on this exploratory analysis of the integrated/fused dataset, we expected that

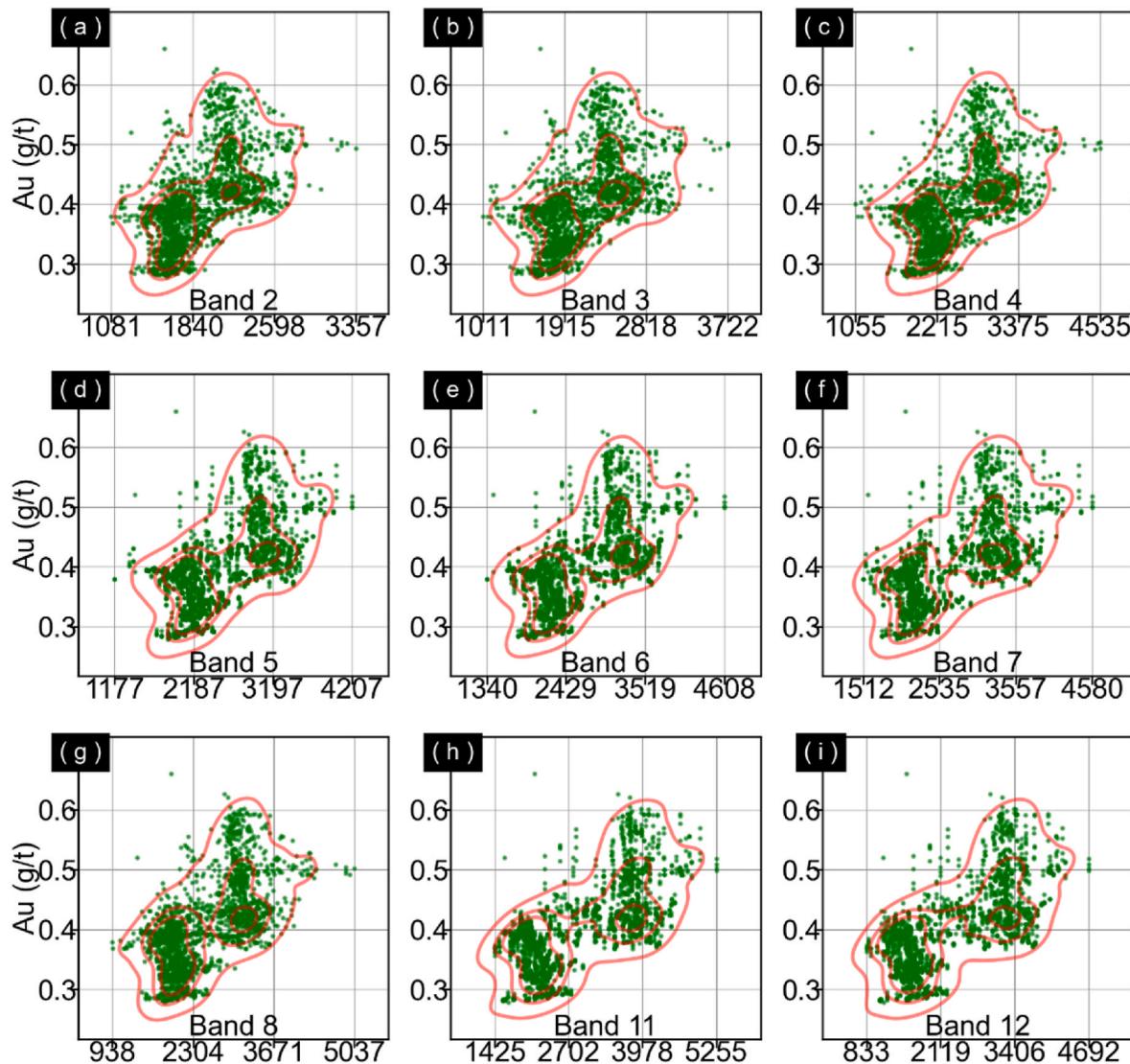


Fig. 6. Relationship of the Au grade ( $\text{g/t}$ ) with band amplitudes 1 to 12 (a to i, respectively) for the integrated dataset.

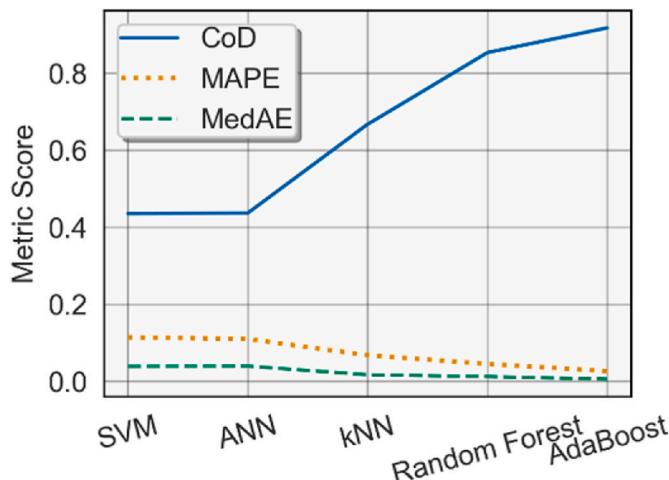


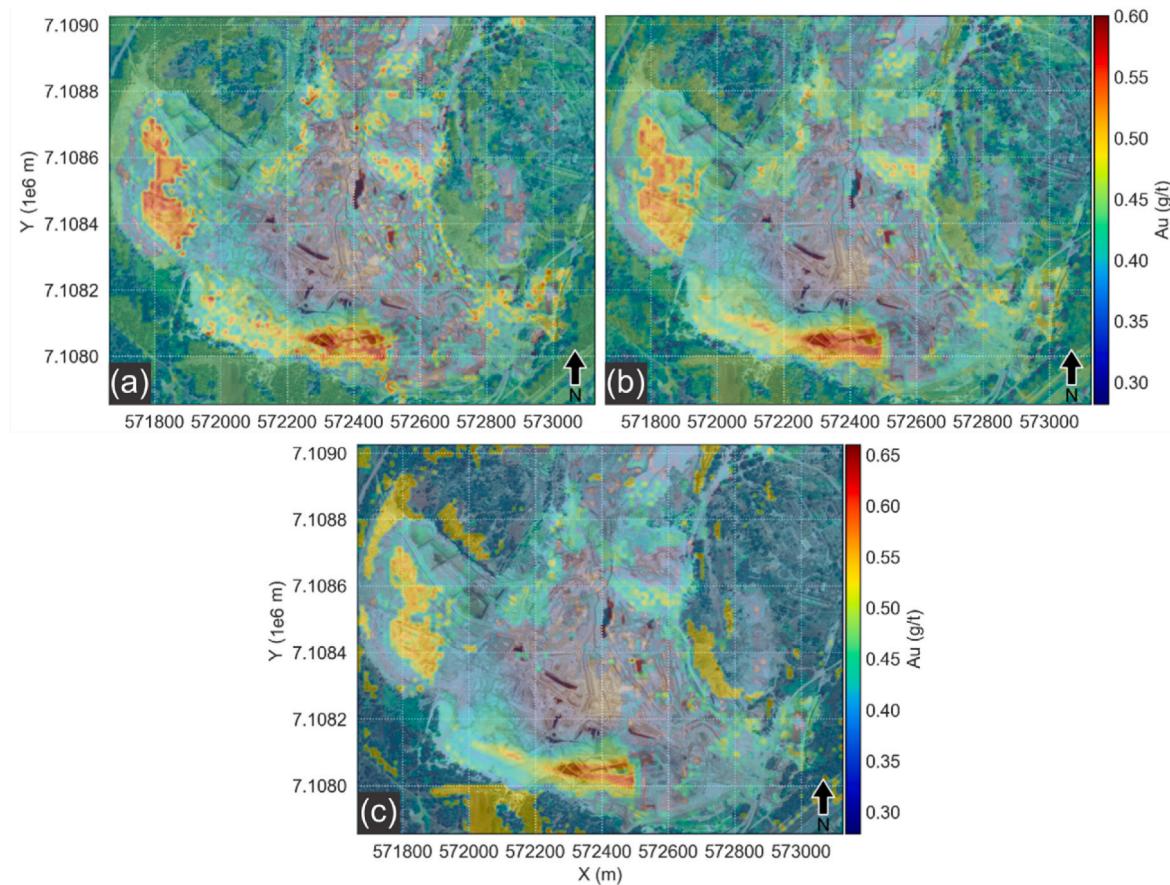
Fig. 7. Comparison of various metrics for several machine learning algorithms for 100 randomized cross-validations.

**Table 4**  
Best prediction results and model hyperparameters.

ALGORITHM	COD	MAPE	MEDAE	PARAMETERS
KNN	0.667	0.068	0.017	$k = 3$
RANDOM FOREST	0.853	0.045	0.012	Maximum depth = unlimited, maximum features = 7, minimum number of samples for a leaf = 1, minimum number of samples for a split = 5, number of estimators = 1500
ADABOOST	0.917	0.027	0.006	Maximum depth = unlimited, maximum number of features = 4, minimum number of samples for a leaf = 1, minimum number of samples for a split = 4, Ensemble size = 300

there would be non-trivial spatial relationships between the band amplitudes and Au grades. In particular, while we expected Au grades to correlate with local band amplitudes in general, the more detailed variations would be nonlinear and potentially counterintuitive (e.g., higher band amplitudes do not always lead to higher Au grades).

The algorithm selection demonstrates that the best algorithms for this task were the kNN, random forest and AdaBoost algorithms. Block-



**Fig. 8.** Predicted map of gold distribution using the (a) kNN, (b) random forest, and (c) AdaBoost algorithms.

based cross-validation results obtained over an average of 100 runs further supported this finding (Fig. 7). The three best algorithms and their performance are summarized in Table 4. SVM and ANN were unable to produce satisfactory results. The cross-validation results show that the median error as measured by the MedAE for the AdaBoost algorithm was generally 0.006 g/t, and the MAPE was 0.027 or 2.7%. The errors were roughly doubled for the random forest in comparison and about three times for the kNN algorithm. The prediction maps showed excellent visible concordance with the resource model within the training region for the three best algorithms (kNN in Fig. 8a; random forest in b; and AdaBoost in c). In particular, the separations between two east-to-west linear stretches of high Au grades were observable in both the random forest and AdaBoost prediction maps (Fig. 8b and c, respectively). This delineation is less evident for the kNN prediction map (Fig. 8a). It is worth noting that the physical location of the two east-west stretches of high Au grades does not overlap with the two east-west stretches of peak band amplitudes (see Fig. 5). It is clear that the high Au grade areas were not simply where the remote sensing data also exhibit high amplitudes, as evidenced by the east-west gradation in Au grades at the southern tip of the TSF, which was unobserved in the remote sensing data (e.g., Figs. 3 and 5). This behaviour was qualitatively anticipated through the exploratory data analysis of the integrated data (Fig. 6). The portion of the TSF that was already extracted (Fig. 5b), within the polygon to the lower right of the figure, was predicted to be negligibly low in Au grades by all deployed models (Fig. 8), and hence served inadvertently as a deployment test. This behaviour, combined with the cross-validations scores and the observation of two clearly separable Au-rich stretches in the TSF, increases our confidence in the deployability of our models. Prediction outside of the TSF clearly shows that although the training data did not capture any portion outside of the TSF, the algorithms that performed well could often correctly assign

these areas with low Au grades. Although we could have presented only results pertaining to the interior of the TSF, we wanted to understand algorithmic behaviour over terrain that was vastly dissimilar to those in the training data. This is important because manual verification of the nature of physical locations (whether they are comparable to terrain in the training data) cannot always occur, especially in a remote setting. Thus, removing all terrain unseen in any training is not always practical or even desirable, given the use of remote sensing over inaccessible locations.

## 4. Discussion

### 4.1. Deriving geochemistry from remote sensing data

The prediction performance demonstrated in this case study using the TSF was excellent for a proof-of-concept study (Table 4). Typical metrics such as the CoD, MAPE and MedAE indicate that both average and median prediction errors were low (on the order of 0.01 g/t MedAE and less than 10% MAPE; Fig. 7). Some key features are observable in the random forest and AdaBoost results, such as the separation of the two east-west trends in the concentration of Au at the southern tip of the TSF (Fig. 8b compared with c, respectively). However, prediction outside of the TSF was more contrasting between the random forest and the AdaBoost algorithms. Slight enrichments in gold (possibly unrealistic and physically unverifiable in this study) outside of the TSF are more selective but generally higher in concentration in the AdaBoost results (Fig. 8c) relative to the other algorithms. In the case of Witwatersrand TSFs (which are open/uncovered), substantial surficial erosion and dispersal of dust (e.g., Oguntoke et al., 2013; Kneen et al., 2015) can create enrichment signatures outside of TSFs. Nonetheless, prediction of terrain that was unsampled in training data is usually poor practice, and

significantly so where the numerical ranges exceed that covered by the training data (and hence extrapolation is expected to occur). This is more important for algorithms that extrapolate poorly, such as the random forest algorithm. Although in the case of remote sensing data, it is not apparent in general that unsampled terrain would exhibit responses that are numerically outside of the training data. An implication of this observation combined with the prediction results is that where the physical location is inaccessible (and coincidentally where remote sensing is more useful), verification of the representativeness of the training data to the deployment area would be a challenge.

The benefits of remote sensing data relative to traditional geochemical data include: coverage, spatial resolution and data velocity. However, inversion of remote sensing data to specific minerals and elements is a challenge. Our proposed inferential method using machine learning leverages a key recognition, that data augmentation is an effective method to overcome data scarcity for suitable spatial data. Once the data scarcity issue has been addressed, many machine learning algorithms become feasible for workflow development. In our case, we explored several suitable shallow learning-type of machine learning algorithms. However, this need not be the case, and where data is even more abundant and explainability is unimportant or secondary (e.g., for engineering purposes as opposed to scientific ones), deep learning could be applied, provided that a cost-benefit analysis supports its use. For our application, it is difficult to realize benefits of even higher performance, given the high CoDs that were observed through cross-validation.

Remote sensing data will likely remain the most affordable, and one of the most accessible types of geoscientific data for exploration and environmental monitoring. Similarly, traditional (primary) geochemical data are unlikely to be displaced by inferred geochemical data. However, between the highly accurate traditional geochemical data and the big data characteristics of remote sensing data, there remains a gap of voluminous and high velocity secondary (inferred) geochemical data. Inferential inversion of remote sensing data to geochemical data can clearly fill this gap and therefore build a bridge between geochemistry and remote sensing. Bipartite benefits include the ability to leverage geochemical data analysis techniques using remote sensing data (e.g., resource assessment or prospectivity mapping), and bringing truly (and in our case, free) big data into geochemistry. Big geochemical data are key to the sustained development and deployment of artificial intelligence methods that are specific to geochemistry. In addition, re-purposing traditional geochemical data can revitalize legacy geochemical data, and extract additional and sometimes irreplaceable value (e.g., in cases where physical access is no longer possible). The generation of geochemical data from remote sensing data benefits geochemical exploration and monitoring by providing high-resolution, timely (including historical availability) and orders of magnitude cheaper geochemical knowledge relative to current practices.

The availability of high-quality training data is a pre-requisite for machine learning, which is identical to the requirement of a-priori terrain knowledge for knowledge-driven inversion techniques. For example, for band ratio-based mineral maps (e.g., iron oxides or clays), the user must know that target minerals exist in an area. Otherwise, the results are unlikely to be acceptable. The use of geostatistical interpolation to create a high-resolution geochemical model was key to creating a sufficient amount of training data. In the context of transdisciplinary approaches, this method can be seen as a discipline-specific form of data augmentation. Using geostatistics for data augmentation is a novel approach to overcome data scarcity. Because the cost to generate traditional geochemical data is substantial, data augmentation could be an enabler to broader adoption of artificial intelligence in geochemistry. Re-purposing legacy data also requires good data stewardship, such that data (especially expensive and often irreplaceable data, e.g., of exhausted mineral deposits) is managed appropriately to facilitate their re-purposing. Where primary data are unavailable, have become irrelevant (e.g., not representative of current physical condition), or are of poor quality (e.g., insufficient or inconsistent data), additional primary

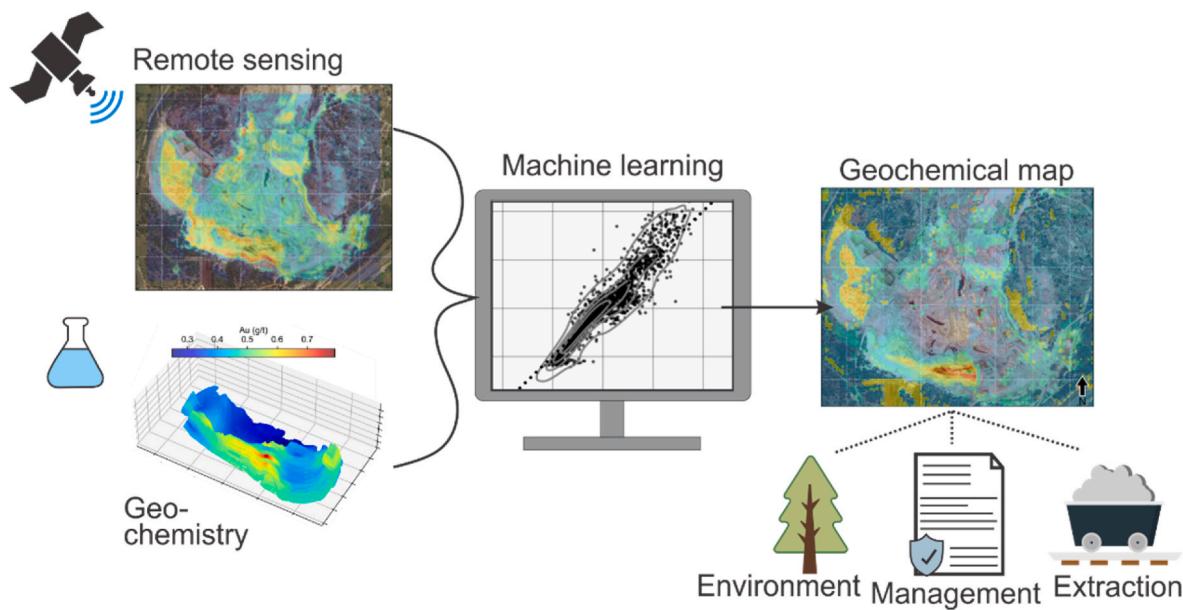
geochemical data would be required before using our method.

An obvious disadvantage of remote sensing data is that they are essentially two-dimensional. Spectral response originates from a shallow surface layer relative to the depths of typical TSFs and, presumably, mineral deposits (e.g., Marghanay, 2022). This is problematic where the geochemical distribution has substantial vertical heterogeneity and a volumetric representation is required. However, the same is true for surficial geochemistry, where borehole geochemistry would be the appropriate choice. Another obvious disadvantage of remote sensing is that it may be completely useless in areas with substantial overburden, such as vegetation, ice or water. Where these issues are substantial and insurmountable, surficial remote sensing-based inferential geochemical data is not possible.

A last concern with inferred geochemical data is that they do not feature quality assurance and quality control (QA/QC) that is associated with analytical geochemistry (Knott et al., 1993; Geboy and Engle, 2011; Piercy, 2014). Quality of inferred geochemical data is a completely new area of research and, for example, standards may need to be met for rigorous use through physical calibration patches. Ground truthing using portable spectrometer-based measurements alone are insufficient due to the change-of-support problem. In a feed-forward sense, multiple factors control the quality of inferred geochemical data, which include: (1) the representativity of the training data of the target site; (2) the quality of training data and if employed, data augmentation (including the quality of the primary geochemical data); and (3) the performance of predictive model(s). We used our knowledge of the internal structure of the TSF to ensure that (1) is met. In other cases, traditional remote sensing-analysis techniques and physical examination may be necessary to ensure that (1) is met. Factor (2) depends on the context of the training data and where traditional data are used. QA/QC documentation of primary data and some knowledge of its generation method and representativeness of the area would be useful to ascertain its quality. It is important to highlight that data augmentation is important to supply a necessary abundance of training data (which are limited by the abundance of geochemical data). This is a canonical requirement of many artificial intelligence-based predictive modelling methods (He et al., 2022 and references therein). It is possible to anticipate the data abundance requirement. For a linear model, at least 2 data points are required in 2 dimensions. Increasing the number of dimensions (e.g., for multilinear modelling), the number of data points required increases exponentially if data density is to be preserved, such that in 10 dimensions, the amount of data required is  $2^9 = 512$  data points. Nonlinear and non-parametric models require even more data. Geostatistical data augmentation is particularly useful for point-wise geochemical samples, because it effectively handles the change-of-support problem (e.g., Gelfand et al., 2001; Gotway and Young, 2002) and therefore generates data that is fuse-able with remote sensing data. However, geostatistical modelling introduces interpolation errors and hence predictive modelling performance is only one indicator of overall deployment performance. This compromise is similar to the use of interpolated data layers in prospectivity mapping and its effects on nonlinear predictive modelling are not yet understood. Factor (3) pertains to the explanatory power of the model, with everything else being equal, as some models are better at some tasks (He et al., 2022). In a deployment situation, the performance would generally be less than that observed in cross-validation. However, exact ground truth of actual geochemical concentrations is difficult to establish without a change of support, which introduces an additional and unavoidable source of error.

#### 4.2. Potential application - rapid *in-situ* assessment of tailings for mine waste valorisation and monitoring

Mine waste valorisation enables mine waste reuse. Regardless of valorisation frameworks (e.g., Nwaila et al., 2021a, b), cost is a fundamental constraint in the exploration of mine waste, because they are less



**Fig. 9.** Schematic diagram of the method demonstrated in this study. Remote sensing and geochemical datasets were integrated together. Machine learning algorithms were then used to model the fused data and to create a geochemical map. The predictively generated geochemical data can be used for various downstream processes such as environmental monitoring, as well as management, exploration and extraction of resources.

resource dense than orebodies. This constrains the coverage and resolution of contact-based sampling. Efficacious and cost-effective reconnaissance methods can promote waste valorisation. Inferred geochemical data could be used to profile desirable and deleterious (e.g., refractory) elements to facilitate extraction planning or monitoring. Online monitoring is a particularly suitable use of inferred geochemical data, because almost all TSFs known to the authors employ surficial extraction methods, similar to open-pit mines. This means that planar resource models derived using remote sensing data can be dynamically updated. To the best of our knowledge, there are no known examples of mine waste assessment using remote sensing-data inversion. However, a similar application has been documented on reclaimed soil to detect Cu (and other heavy metal) contamination levels (Shen et al., 2019). Furthermore, a different application specifically expresses the desire to perform online assessment of saline and alkaline levels over land (Xiao and Wan, 2021). An obvious advantage to online assessment is that perturbances of a TSF (e.g., due to extraction, slumping or consolidation) could be taken into account for extraction sequencing, material stockpiling and metal reconciliation. Drone-based remote sensing could be substituted for satellite-based remote sensing to increase feedback rate or resolution.

For environmental monitoring, the ability to derive elemental concentrations from remote sensing data, and especially how such concentrations evolve over time due to intrinsic and extrinsic processes (e.g., erosion and weathering) could potentially increase the accuracy of local knowledge through more precise and temporal chemical information about the area. The possibility of real-time monitoring was also recognised by Xiao and Wan (2021). The velocity of satellite-based remote sensing data enables the ability to derive historical snapshots and animations of geochemical concentration maps. In this case, our method permits remote and timely geochemical knowledge without necessitating further assays (aside from initial assays and perhaps ongoing sampling to construct a training dataset). Reasonably, this implies that costs associated with environmental management of TSFs or otherwise (that are suitable to be imaged through remote sensing) could be potentially reduced, and management action can occur with lower latency. Aside from operational cost reductions, this can improve the human and environmental safety of managed TSFs.

#### 4.3. Potential application - surficial mineral deposit detection and prospectivity mapping

Downstream uses of inferred geochemical data could include prospectivity mapping (especially where geochemical data is untimely or unavailable, e.g., Cheng et al., 2021; Mohamed et al., 2021) and material stream profiling (where spectral sensors are used for in-situ material analyses, e.g., Xiao et al., 2021). Where the environment is adequately exposed and remote sensing data is available, it is, in principle, possible to extend our method to create multi-elemental data. Presently, this does not appear to be a popular use of remote sensing data and sparse examples are available (e.g., derivation of elemental concentrations of As, Cu, Hg, Mo, Pb and Zn in Cheng et al., 2021). The ability to rapidly generate geochemical data for mapping is critical for time-sensitive exploration activities, such as those associated with critical raw materials (e.g., battery minerals and the overall green energy transition, e.g., Warner, 2019). There is an explicit and large-scale desire to re-invent exploration to vastly reduce time-to-discovery, especially for critical materials (Government of Canada, 2022). Hence, traditional geochemical exploration may be unsuitable. Deriving advanced and probable information rapidly is advantageous, which is the driver behind the proliferation of geophysical exploration.

Nevertheless, inferred data can be substituted for primary data for prospectivity and exploration. For example, inferred data could be subjected to geochemical anomaly detection (e.g., Grunsky and de Caritat, 2019; Zhang et al., 2021, 2022) for further exploration. An exploration pipeline could be created to automate inferential data generation and anomaly detection. This could lead to the establishment of an extensive training database and broadly deployable models (including deep learning models). Such a pipeline may become essential in future exploration campaigns as the quest for minerals becomes more urgent (critical minerals) and the tools of data generation to data analysis rapidly evolve. Another strong argument in favour of using remote sensing-derived geochemical data is their superior spatial resolution relative to traditional geochemical surveys (e.g., Friske and Hornbrook, 1991; Demetrides et al., 2018). The increased resolution could permit the discovery of fainter or more complex geochemical signatures, which is important for the discovery of, for example, deeper mineral deposits.

## 5. Conclusions

In this study, we demonstrated a general method based on machine learning to invert remote sensing data into geochemical data (Fig. 9). Our method attempts to provide a form of secondary geochemical data that integrates the most desirable aspects of remote sensing data (e.g., big data characteristics) and the elemental-level of knowledge of geochemical data. To derive a dataset that is suitable for predictive modelling using machine learning, we fused the remote sensing and geostatistically-augmented geochemical datasets. We expect our method to be useable in similar settings to profile areas where training data exist. In general, we expect that predictions would be reliable if the deployment area is compositionally similar to that covered in training data. Inferred geochemical data could be analysed using tools that are standard to geochemistry. Hence, inferential geochemistry connects remote sensing and geochemistry, and can provide high spatial-resolution geochemical data for downstream uses, such as resource quantification, historic resource analysis, animated or online extraction/operation monitoring, environmental monitoring and prospectivity mapping. We envision that geochemical exploration and monitoring could be rendered to be much cheaper and targeted by using multi-platform, remote sensing-based inferential geochemistry to guide the generation of primary geochemical data. This would benefit the entire geoscientific and natural resource communities.

## Funding

Funding for this research was provided by the Department of Science and Innovation (DSI)-National Research Foundation (NRF) Thuthuka Grant (Grant UID: 121,973) and DSI-NRF CIMERA. Yousef Ghorbani acknowledges financial support from the Centre for Advanced Mining and Metallurgy (Camm), a strategic research environment established at the Luleå University of Technology funded by the Swedish government. We also thank Sibanye-Stillwater Ltd. For their funding through the Wits Mining Institute (WMI).

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The authors would like to thank Sibanye-Stillwater Ltd. For access to their geochemical data. We also thank several anonymous reviewers and Prof. Kunfeng Qiu (University of Beijing) for providing insightful comments, which have greatly improved this paper.

## Supplementary data

The remote sensing data is publicly available. Geochemical data is the property of Sibanye-Stillwater Ltd.

## References

- Abdalla, R., 2016. Geospatial Data Integration. In: Introduction to Geospatial Information and Communication Technology (GeoICT). Springer, Cham. [https://doi.org/10.1007/978-3-319-33603-9\\_6](https://doi.org/10.1007/978-3-319-33603-9_6).
- Agar, B., Coulter, D., 2007. Remote sensing for mineral exploration – a decade perspective 1997–2007. In: Milkereit, B. (Ed.), Fifth Decennial International Conference on Mineral Exploration, pp. 109–136.
- Aiazz, B., Alparone, L., Barducci, A., Baronti, S., Pippi, I., 1999. Multispectral fusion of multisensor image data by the generalized Laplacian pyramid. IEEE Int. Geosci. Remote Sens. Symp. <https://doi.org/10.1109/IGARSS.1999.774572>.
- Al-doski, J., Mansor, S.B., Zulhaiddi, H., Shafri, M., 2013. Image classification in remote sensing. Environ. Earth Sci. 3 (10), 141–148.
- Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., Notarnicola, C., 2015. Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Rem. Sens. 7, 16398–16421. <https://doi.org/10.3390/rs71215841>.
- Amer, R., Kusky, T., Ghulam, A., 2010. Lithological mapping in the central eastern desert of Egypt using ASTER data. J. Afr. Earth Sci. 56, 75–82. <https://doi.org/10.1016/j.jafrearsci.2009.06.004>.
- Asokan, A., Anitha, J., Ciobanu, M., Gabor, A., Naaji, A., Hemanth, D.J., 2020. Image processing techniques for analysis of satellite images for historical maps classification – an overview. Appl. Sci. 10, 4207. <https://doi.org/10.3390/app10124207>.
- Bachri, I., Hakdaoui, M., Raji, M., Teodoro, A.C., Benbouziane, A., 2019. Machine learning algorithms for automatic lithological mapping using remote sensing data: a case study from Souk Arbaa Sahel, Sidi Ifni Inlier, Western Anti-Atlas, Morocco. ISPRS Int. J. Geo-Inf. 8 (6), 248. <https://doi.org/10.3390/ijgi8060248>.
- Bannari, A., El-Battay, A., Saquaque, A., Miri, A., 2016. PALSAR-FBS L-HH mode and Landsat-TM data fusion for geological mapping. Adv. Rem. Sens. 5 (4), 246–268. <https://doi.org/10.4236/ars.2016.54020>.
- Barsi, J.A., Lee, K., Kvaran, G., Markham, B.L., Pedelty, J.A., 2014. The spectral response of the landsat-8 operational land imager. Rem. Sens. 6 (10), 10232–10251. <https://doi.org/10.3390/rs61010232>.
- Bath, A., 2007. Infrared spectroscopy of proteins. Biochim. Biophys. Acta Bioenerg. 1767 (91), 1073–1101. <https://doi.org/10.1016/j.bbabi.2007.06.004>.
- Beiranvand Pour, A., Hashim, M., 2014. ASTER, ALI and Hyperion sensors data for lithological mapping and ore minerals exploration. SpringerPlus 3, 130. <https://doi.org/10.1186/2193-1801-3-130>.
- Beiranvand Pour, A., Hashim, M., Hong, J.K., Park, Y., 2019. Lithological and alteration mineral mapping in poorly exposed lithologies using Landsat-8 and ASTER satellite data: north-eastern Graham Land, Antarctic Peninsula. Ore Geol. Rev. 108, 112–133. <https://doi.org/10.1016/j.oregeorev.2017.07.018>.
- Beiranvand Pour, A., Hashim, M., Makoundi, C., Zaw, K., 2016. Structural mapping of the Bentong-Raub Suture Zone using PALSAR remote sensing data, Peninsular Malaysia: implications for sediment-hosted/orogenic gold mineral systems exploration. Resour. Geol. 66 (4), 368–385. <https://doi.org/10.1111/rge.12105>.
- Bird, J.J., Faria, D.R., Premeida, C., Ekart, A., Ayrosa, P.P.S., 2020. Overcoming data scarcity in speaker identification: dataset augmentation with synthetic MFCCs via character-level RNN. IEEE Int. Conf. Autonom. Robot. Syst. Compet. 146–151. <https://doi.org/10.1109/ICARSC49921.2020.9096166>.
- Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., 2013. API Design for Machine Learning Software: Experiences from the Scikit-Learn Project. European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases. <https://doi.org/10.48550/arXiv.1309.0238>.
- Cao, N., 2013. Calibration Optimization and Efficiency in Near Infrared Spectroscopy. Ph.D. thesis, Iowa State University, p. 183.
- Cao, J.S., Deng, Z.Y., Li, W., Hu, Y.D., 2020. Remote sensing inversion and spatial variation of land surface temperature over mining areas of Jixi, Heilongjiang, China. PeerJ 8, 10257. <https://doi.org/10.7717/peerj.10257>.
- Chakouri, M., El Harti, A., Lhissou, R., El Hachimi, J., Jelloul, A., 2020. Geological and mineralogical mapping in Moroccan central Jebilet using multispectral and hyperspectral satellite data and machine learning. Int. J. Adv. Trends Comput. Sci. Eng. 9, 5772–5783. <https://doi.org/10.30534/ijatcse/2020/234942020>.
- Chandrasekaran, B., Jain, A.K., 1974. Quantization complexity and independent measurements. IEEE Trans. Comput. 23 (8), 102–106. <https://doi.org/10.1109/T-C.1974.223789>.
- Chen, X.-W., Lin, X., 2014. Big data deep learning: challenges and perspectives. IEEE Access 2, 514–525. <https://doi.org/10.1109/ACCESS.2014.2325029>.
- Cheng, G., Huang, H., Li, H., Deng, X., Khan, R., Soh Tamehe, L., Atta, A., Lang, X., Guo, X., 2021. Quantitative remote sensing of metallic elements for the Qishitan gold polymetallic mining area, NW China. Rem. Sens. 13, 2519. <https://doi.org/10.3390/rs13132519>.
- Ciampalini, A., Garfagnoli, F., Del Ventisette, C., Moretti, S., 2013. Potential use of remote sensing techniques for exploration of iron deposits in Western Sahara and southwest of Algeria. Nat. Resour. Res. 22, 179–190. <https://doi.org/10.1007/s11053-013-9209-5>.
- Clerc, S., MPC Team, 2022. Sentinel-2 Data Quality Report, Report Issue 55. ESA-CS, France. [https://sentinel.esa.int/documents/247904/685211/Sentinel-2\\_L1C\\_Data\\_Quality\\_Report](https://sentinel.esa.int/documents/247904/685211/Sentinel-2_L1C_Data_Quality_Report). (Accessed 27 June 2022).
- Cohen, D.R., Kelley, D.L., Anand, R., Coker, W.B., 2010. Major advances in exploration geochemistry: 1998–2007. Geochem. Explor. Environ. Anal. 10, 3–16. <https://doi.org/10.1144/1467-7873/09-215>.
- Cooley, T., Anderson, G.P., Felde, G.W., Hoke, M.L., Ratkowski, A.J., Chetwynd, J.H., Gardner, J.A., Adler-Golden, S.M., Matthew, M.W., Berk, A., Bernstein, L.S., Acharya, P.K., Miller, D., Lewis, P., 2002. FLAASH, a MODTRAN4-Based Atmospheric Correction Algorithm, its Application and Validation. In: Proceedings of the Geoscience and Remote Sensing Symposium. IEEE International, pp. 1414–1418. <https://doi.org/10.1109/IGARSS.2002.1026134>.
- 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>.
- Dai, J., Wang, D., Dai, H., Liu, L., Wu, Y., 2017. Geological mapping and ore prospecting study using remote sensing technology in Jiajika area of western Sichuan Province. Chin. Geol. 44 (2), 389–398.
- Demetrides, A., Smith, D.B., Wang, X., 2018. General concepts of geochemical mapping at global, regional, and local scales for mineral exploration and environmental

- purposes. *Geochim. Bras.* 32 (2), 136–179. <https://doi.org/10.21715/GB2358-2812.2018322136>.
- Diaz-Rodriguez, J., Dietmar Muller, R., Chandra, R., 2021. Predicting the emplacement of cordilleran porphyry copper systems using a spatio-temporal machine learning model. *Ore Geol. Rev.* 137, 104300 <https://doi.org/10.1016/j.oregeorev.2021.104300>.
- Dramsch, J.S., 2020. 70 years of machine learning in geoscience in review. *Adv. Geophys.* 61, 1–55.
- ENVI, 2009. Atmospheric correction module: QUAC and FLAASH user's guide. [http://www.exelisvis.com/portals/0/pdfs/envi/Flaash\\_Module.pdf2009](http://www.exelisvis.com/portals/0/pdfs/envi/Flaash_Module.pdf2009). (Accessed 31 March 2022).
- Friske, P.W.B., Hornbrook, E.H.W., 1991. Canada's national geochemical reconnaissance programme. In: transactions of the Institution of Mining and Metallurgy, Section B. *B. Appl. Earth Sci.* 100, B47–B56.
- Ge, W., Cheng, Q., Jing, L., Wang, F., Zhao, M., Ding, H., 2020. Assessment of the capability of Sentinel-2 imagery for iron-bearing minerals mapping: a case study in the cuprite area, Nevada. *Rem. Sens.* 12, 3028. <https://doi.org/10.3390/rs12183028>.
- Ge, W., Cheng, Q., Tang, Y., Jing, L., Gao, C., 2018. Lithological classification using Sentinel-2A data in the Shibanjing ophiolite complex in inner Mongolia, China. *Rem. Sens.* 10 (4), 638. <https://doi.org/10.3390/rs10040638>.
- Geboy, N.J., Engle, M.A., 2011. Quality Assurance and Quality Control for Geochemical Data: a Primer for the Research Scientist. U.S. Geological Survey, Open File 2011-1187, p. 33.
- Gelfand, A.E., Zhu, L., Carlin, B.P., 2001. On the change of support problem for spatio-temporal data. *Biostat* 2 (1), 31–45. <https://doi.org/10.1093/biostatistics/2.1.31>.
- Ghorbani, Y., Zhang, S.E., Nwaila, G., Bourdeau, J.E., 2022. Framework components for data-centric dry laboratories in the minerals industry: a path to science-and-technology-led innovation. *Extr. Ind. Soc.* 10, 101089 <https://doi.org/10.1016/j.eis.2022.101089>.
- Ghorbani, Y., Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., Safari, M., Hoseinne, S.H., Nwaila, P., Ruuska, J., 2023. Dry laboratories – mapping the required instrumentation and infrastructure for online monitoring, analysis, and characterization in the mineral industry. *Miner. Eng.* 191, 107971 <https://doi.org/10.1016/j.mineng.2022.107971>.
- Gotway, C.A., Young, L.J., 2002. Combining incompatible spatial data. *J. Am. Stat. Assoc.* 97 (458), 632–648. <https://doi.org/10.1198/016214502760047140>.
- Govett, G.J.S., 1983. Geochemistry in the exploration sequence. In: Govett, G.J.S. (Ed.), *Handbook of Exploration Geochemistry*, vol. 3. Elsevier. <https://doi.org/10.1016/B978-0-444-42021-3.3.0006-9>.
- Government of Canada, 2022. The Canadian Critical Minerals Strategy. Government of Canada. <https://www.canada.ca/en/campaign/critical-minerals-in-canada/canadian-critical-minerals-strategy.html>. (Accessed 30 January 2023).
- Griffiths, P.R., De Haseth, J.A., Winefordner, J.D., 2007. *Fourier Transform Infrared Spectrometry*, second ed. Wiley.
- Grunsky, E.C., de Caritat, P., 2019. State-of-the-art analysis of geochemical data for mineral exploration. *Geochem. Explor. Environ. Anal.* 20, 217–232. <https://doi.org/10.1144/geochem2019-031>.
- Guo, P., Su, Y., Cha, Z., Lin, Q., Luo, W., Lin, Z., 2016. Prediction of leaf phosphorus contents for rubber seedlings based on hyperspectral sensitive bands and back propagation artificial neural network. *Trans. Chin. Soc. Agric. Eng.* 32 (1), 177–183.
- Harvey, A., Fotopoulos, G., 2016. Geological mapping using machine learning algorithms. *Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci.* 41-B8, 423–430. <https://doi.org/10.5194/isprarchives-XLI-B8-423-2016>.
- He, H., Yang, X., Li, Y., Li, Y., Wang, L., 2010. Multi-source data fusion technique and its application in geological and mineral survey. *J. Earth Sci. Environ.* 32 (1), 44–47.
- He, Y., Zhou, Y., Wen, T., Zhang, S., Huang, F., Zou, X., Ma, X., Zhu, Y., 2022. A review of machine learning in geochemistry and cosmochemistry: method improvements and applications. *Appl. Geochim.* 140, 105273 <https://doi.org/10.1016/j.apgeochem.2022.105273>.
- Hou, J., He, Y., Yang, H., Connor, T., Gao, J., Wang, Y., Zeng, Y., Zhang, J., Huang, J., Zheng, B., Zhou, S., 2020. Identification of animal individuals using deep learning: a case study of giant panda. *Biol. Conserv.* 242, 108414 <https://doi.org/10.1016/j.biocon.2020.108414>.
- Irons, J.R., Dwyer, J.L., Barsi, J.A., 2012. The next Landsat satellite: the Landsat data continuity mission. *Remote Sens. Env.* 122, 11–21. <https://doi.org/10.1016/j.rse.2011.08.026>.
- Karpnate, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V., 2018. Machine learning for the geosciences: challenges and opportunities. *IEEE Trans. Knowl. Data Eng.* 31 (8), 1544–1554.
- Kneen, M.A., Ojelede, M.E., Annegarn, H., 2015. Housing and population sprawl near tailings storage facilities in the Witwatersrand: 1952 to current. *South Afr. J. Sci.* 111, 1–9.
- Knott, J.M., Glysson, G.D., Malo, B.A., Schroder, L.J., 1993. Quality assurance plan for the collection and processing of sediment data by the U.S. Geological Survey, water resources division. U.S. Geological Survey, Open File 92-499, 22. <https://doi.org/10.3133/ofr92499>.
- Kristollari, V., Karathanassi, V., 2020. Fine-tuning self-organizing maps for Sentinel-2 Imagery: separating clouds from bright surfaces. *Rem. Sens.* 12 (12), 1923. <https://doi.org/10.3390/rs12121923>.
- Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. *Geosci. Front.* 7 (1), 3–10. <https://doi.org/10.1016/j.gsf.2015.07.003>.
- Lawley, C.J.M., Tschirhart, V., Smith, J.W., Pehrsson, S.J., Schetselaar, E.M., Schaeffer, A.J., Houle, M.G., Eglington, B.M., 2021. Prospectivity modelling of Canadian magmatic Ni ( $\pm$  Cu  $\pm$  Co  $\pm$  PGE) sulphide mineral systems. *Ore Geol. Rev.* 132, 103985 <https://doi.org/10.1016/j.oregeorev.2021.103985>.
- Madhuandan, L., Sadavarte, P., Visschedijk, A.J.H., Denier Van Der Gon, H.A.C., Aben, I., Osei, F.B., 2021. Deep convolutional neural networks for surface coal mines determination from Sentinel-2 images. *Eur. J. Remote Sens.* 54 (1), 296–309. <https://doi.org/10.1080/22797254.2021.1920341>.
- Manakos, I., Manevski, K., Kalaitzidis, C., Edler, D., 2011. Comparison between Atmospheric Correction Modules on the Basis of Worldview-2 Imagery and in Situ Spectroradiometric Measurements. In: 7<sup>th</sup> EARSeL SIG Imaging Spectroscopy workshop, Edinburgh, pp. 1–14.
- Margany, M., 2022. *Remote Sensing and Image Processing in Mineralogy*. CRC Press. <https://doi.org/10.1201/9781003033776>.
- Mather, P.M., Tsai, B., 2009. *Classification Methods for Remotely Sensed Data*, second ed. Taylor & Francis.
- McLachlan, G.J., 2004. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience.
- Mitchell, T.M., 1997. *Machine Learning*. McGraw Hill.
- Mohamed, M.T.A., Al-Naimi, L.S., Mgbeojedo, T.I., Agoha, C.C., 2021. Geological mapping and mineral prospectivity using remote sensing and GIS in parts of Hamissana, Northeast Sudan. *J. Pet. Explor. Prod.* 11, 1123–1138. <https://doi.org/10.1007/s13202-021-01115-3>.
- Moon, C.J., Whateley, M.K.G., Evans, A.M., 2006. *Introduction to Mineral Exploration*, second ed. Blackwell Publishing.
- NASA, 1997. NASA Facts: Landsat 7. National Aeronautics and Space Administration, pp. 1–4. FS-1997(03)-007-GSFC.
- Noori, L., Beiranvand Pour, A., Askari, G., Taghipour, N., Pradhan, B., Lee, C.W., Honarmand, M., 2019. Comparison of different algorithms to map hydrothermal alteration zones using ASTER remote sensing data for polymetallic vein-type ore exploration: toroud-Chahshirin magmatic belt (TCMB), North Iran. *Rem. Sens.* 11, 495. <https://doi.org/10.3390/rs11050495>.
- Nwaila, G.T., Ghorbani, Y., Becker, M., Frimmel, H.E., Petersen, J., Zhang, S.E., 2019. Geometallurgical approach for implications of ore blending on cyanide leaching and adsorption behaviour on Witwatersrand Gold ores, South Africa. *Nat. Resour. Res.* 29, 1007–1030. <https://doi.org/10.1007/s11053-019-09522-4>.
- Nwaila, G.T., Ghorbani, Y., Zhang, S.E., Frimmel, H.E., Tolmay, L.C., Rose, D.H., Nwaila, P.C., Bourdeau, J.E., 2021a. Valorization of mine waste – Part I: characteristics of, and sampling methodology for, consolidated mineralized tailings by using Witwatersrand gold mines (South Africa) as an example. *J. Environ. Manag.* 295, 113013 <https://doi.org/10.1016/j.jenvman.2021.113013>.
- Nwaila, G.T., Ghorbani, Y., Zhang, S.E., Frimmel, H.E., Tolmay, L.C., Rose, D.H., Nwaila, P.C., Bourdeau, J.E., Frimmel, H.E., 2021b. Valorization of mine waste – Part II: resource evaluation for consolidated and mineralized mine waste using the Central African Copperbelt as an example. *J. Environ. Manag.* 299, 113553 <https://doi.org/10.1016/j.jenvman.2021.113553>.
- Ogunkoto, O., Ojelede, M.E., Annegarn, H.J., 2013. Frequency of mine dust episodes and the influence of meteorological parameters on the Witwatersrand area, South Africa. *Int. J. Atmos. Sci.* 128463 <https://doi.org/10.1155/2013/128463>, 2013.
- Park, H., Choi, J., Park, N., Choi, S., 2017. Sharpening the VNIR and SWIR bands of Sentinel-2A imagery through modified selected and synthesized band schemes. *Rem. Sens.* 9 (10), 1080. <https://doi.org/10.3390/rs9101080>.
- Piercey, S.J., 2014. Modern analytical facilities 2. A review of quality assurance and quality control QA/QC) procedures for lithogeochemical data. *Geosci. Can.* 41, 75–88. <https://doi.org/10.12789/geocanj.2014.41.035>.
- Pohl, C., Van Genderen, J., 2014. Remote sensing image fusion: an update in the context of Digital Earth. *Int. J. Digit. Earth* 7 (2), 158–172. <https://doi.org/10.1080/17538947.2013.869266>.
- Prior, T., Giurco, D., Mudd, G., Mason, L., Behrisch, J., 2012. Resource depletion, peak minerals and the implications for sustainable resource management. *Global Environ. Change* 22 (3), 577–587. <https://doi.org/10.1016/j.gloenvcha.2011.08.009>.
- Prudhivi, L., Narayana, M., Subrahmanyam, C., Krishna, M.G., 2021. Animal species image classification. *Mater. Today*. <https://doi.org/10.1016/j.matpr.2021.02.771>.
- Rachidi, N.R., Nwaila, G.T., Zhang, S.E., Bourdeau, J.E., Ghorbani, Y., 2021. Assessing cobalt supply sustainability through production forecasting and implications for green energy policies. *Resour. Pol.* 74, 102423 <https://doi.org/10.1016/j.resourpol.2021.102423>.
- Rajendran, S., Nasir, S., 2014. Hydrothermal altered serpentized zone and a study of Ni magnesioferrite–magnetite–awaruite occurrences in Wadi Hib, Northern Oman Mountain discrimination through ASTER mapping. *Ore Geol. Rev.* 62, 211–226. <https://doi.org/10.1016/j.oregeorev.2014.03.016>.
- Rajesh, H., 2004. Application of remote sensing and GIS in mineral resource mapping-An overview. *J. Mineral. Petrol. Sci.* 99, 83–103. <https://doi.org/10.2465/jmps.99.83>.
- Roy, D.P., Wulder, M.A., Loveland, T.R., Woodcock, C.E., Allen, R.G., Anderson, M.C., Helder, D., Irons, J.R., Johnson, D.M., Kennedy, R., Scambos, T.A., Schaaf, C.B., et al., 2014. Landsat-8: science and product vision for terrestrial global change research. *Remote Sens. Env.* 145, 154–172. <https://doi.org/10.1016/j.rse.2014.02.001>.
- Rudolph, W., Hefta, G., 2009. Quantitative analysis in alkaline aluminate solutions by Raman spectroscopy. *Anal. Methods* 1 (2), 132–138. <https://doi.org/10.1039/B9AY00064J>.
- Salem, S.M., El Sharkawi, M., El-Alfy, Z., Soliman, N.M., Ahmed, S.E., 2016. Exploration of gold occurrences in alteration zones at Dungash district, Southeastern Desert of Egypt using ASTER data and geochemical analyses. *J. Afr. Earth Sci.* 117, 389–400.
- Sehgal, S., 2012. Remotely Sensed Landsat image classification using neural network approaches. *Int. J. Eng. Res. Afr.* 2 (5), 43–46.

- Shen, Q., Xia, K., Zhang, S., Kong, C., Hu, Q., Yang, S., 2019. Hyperspectral indirect inversion of heavy-metal copper in reclaimed soil of iron ore area. *Spectrochim. Acta Mol. Biomol. Spectrosc.* 222, 117191 <https://doi.org/10.1016/j.saa.2019.117191>.
- Shirmard, H., Farahbakhsh, E., Müller, R.D., Chandra, R., 2022. A review of machine learning in processing remote sensing data for mineral exploration. *Remote Sens. Env.* 268, 112750 <https://doi.org/10.1016/j.rse.2021.112750>.
- Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. *J. Big Data* 6, 60. <https://doi.org/10.1186/s40537-019-0197-0>.
- Song, K., Wang, E., Yao, Y., Fu, J., Hao, D., You, X., 2020. Spectral alteration zonation based on close-range hyperspec-320 m imaging spectroscopy: a case study in the Gongchang high-grade iron ore deposit, Liaoning Province, NE China. *Appl. Sci.* 10, 8369. <https://doi.org/10.3390/app10238369>.
- Trunk, G.V., 1979. A problem of dimensionality: a simple example. *IEEE Trans. Pattern Anal. Mach. Intell.* 3, 306–307. <https://doi.org/10.1109/TPAMI.1979.4766926>.
- USGS, 2022. Landsat 8. <https://www.usgs.gov/landsat-missions/landsat-8>. (Accessed 1 June 2022).
- Vaiopoulos, A., Karantzalos, K., 2016. Pansharpening on the narrow VNIR and SWIR spectral bands of Sentinel-2. *ISPRS* 41, 723–730.
- Van der Meer, F.D., van der Werff, H.M.A., van Ruitenbeek, F.J.A., 2014. Potential of ESA's Sentinel-2 for geological applications. *Remote Sens. Env.* 148, 124–133. <https://doi.org/10.1016/j.rse.2014.03.022>.
- Van der Werff, H.M.A., van der Meer, F.D., 2015. Sentinel-2 for mapping iron absorption feature parameters. *Rem. Sens.* 7 (10), 12635–12653. <https://doi.org/10.3390/rs71012635>.
- Wang, Z., Zuo, R., Jing, L., 2021. Fusion of geochemical and remote-sensing data for lithological mapping using Random Forest metric learning. *Math. Geosci.* 53, 1125–1145. <https://doi.org/10.1007/s11004-020-09897-8>.
- Warner, J.T., 2019. Overview and Comparison of Different Lithium-Ion Chemistries. Lithium-Ion Battery Chemistries. Elsevier. <https://doi.org/10.1016/B978-0-12-814778-8.00004-1>.
- Wu, Q., 2020. Geemap: a Python package for interactive mapping with Google Earth engine. *J. Open Source Softw.* 5 (51), 2305. <https://doi.org/10.21105/joss.02305>.
- Wu, Q., Lane, C.R., Li, X., Zhao, K., Zhou, Y., Clinton, N., DeVries, B., Golden, H.E., Lang, M.W., 2019. Integrating LiDAR data and multi-temporal aerial imagery to map wetland inundation dynamics using Google Earth Engine. *Remote Sens. Env.* 228, 1–13. <https://doi.org/10.1016/j.rse.2019.04.015>.
- Xiao, D., Le, B.T., Ha, T.T.L., 2021. Iron ore identification method using reflectance spectrometer and a deep neural network framework. *Spectrochim. Acta Mol. Biomol. Spectrosc.* 248, 119168 <https://doi.org/10.1016/j.saa.2020.119168>.
- Xiao, D., Liu, C., Le, B.T., 2019. Detection method of TFe content of iron ore based on visible-infrared spectroscopy and IPSO-TELIM neural network. *Infrared Phys. Technol.* 97, 341–348. <https://doi.org/10.1016/j.infrared.2019.01.005>.
- Xiao, D., Wan, L., 2021. Remote sensing inversion of saline and alkaline land based on an improved seagull optimization algorithm and the two-hidden-layer extreme learning machine. *Nat. Resour. Res.* 30, 3795–3818. <https://doi.org/10.1007/s11053-021-09876-8>.
- Zhang, C., Lu, Y., 2021. Study on artificial intelligence: the state of the art and future prospects. *J. Ind. Inf. Integr.* 23, 100224 <https://doi.org/10.1016/j.jii.2021.100224>.
- Zhang, S.E., Bourdeau, J.E., Nwaila, G.T., Corrigan, D., 2022. Towards a fully data-driven prospectivity mapping methodology: a case study of the Southeastern Churchill Province, Québec and Labrador. *Artif. Intell. Geosci.* 2, 128–147. <https://doi.org/10.1016/j.aiig.2022.02.002>.
- Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., Ashwal, L.D., 2021. Machine learning-based prediction of trace element concentrations using data from the Karoo large igneous province and its application in prospectivity mapping. *Artif. Intell. Geosci.* 2, 60–75. <https://doi.org/10.1016/j.aiig.2021.11.002>.
- Zuo, R., 2020. Geodata science-based mineral prospectivity mapping: a review. *Nat. Resour. Res.* 29, 3415–3424. <https://doi.org/10.1007/s11053-020-09700-9>.