Artificial Intelligence in Geosciences 4 (2023) 137–149

Contents lists available at ScienceDirect

Artificial Intelligence in Geosciences

journal homepage: www.keaipublishing.com/en/journals/artificial-intelligence-in-geosciences

Big geochemical data through remote sensing for dynamic mineral resource monitoring in tailing storage facilities

Steven E. Zhang [a], Glen T. Nwaila [a,*], Shenelle Agard [b,c], Julie E. Bourdeau [a], Emmanuel John M. Carranza [c], Yousef Ghorbani [d]

[a] Wits Mining Institute, University of the Witwatersrand, 1 Jan Smuts Ave., Johannesburg, 2000, South Africa
[b] Department of Earth Sciences, Uppsala University, SE 75236, Uppsala, Sweden
[c] Department of Geology, University of the Free State, 205 Nelson Mandela Dr., Bloemfontein, 9301, South Africa
[d] School of Chemistry, University of Lincoln, Joseph Banks Laboratories, Green Lane, Lincoln, Lincolnshire, LN6 7DL, United Kingdom

ARTICLE INFO

Keywords:
Big geochemical data
Mine waste valorisation
Tailings storage facilities
Sentinel-2
Remote sensing
Machine learning

ABSTRACT

Evolution in geoscientific data provides the mineral industry with new opportunities. A direction of geochemical data generation evolution is towards big data to meet the demands of data-driven usage scenarios that rely on data velocity. This direction is more significant where traditional geochemical data are not ideal, which is the case for evaluating unconventional resources, such as tailing storage facilities (TSFs), because they are not static due to sedimentation, compaction and changes associated with hydrospheric and lithospheric processes (e.g., erosion, saltation and mobility of chemical constituents). In this paper, we generate big secondary geochemical data derived from Sentinel-2 satellite-remote sensing data to showcase the benefits of big geochemical data using TSFs from the Witwatersrand Basin (South Africa). Using spatially fused remote sensing and legacy geochemical data on the Dump 20 TSF, we trained a machine learning model to predict in-situ gold grades. Subsequently, we deployed the model to the Lindum TSF, which is 3 km away, over a period of a few years (2015-2019). We were able to visualize and analyze the temporal variation in the spatial distributions of the gold grade of the Lindum TSF. Additionally, we were able to infer extraction sequencing (to the resolution of the data), acid mine drainage formation and seasonal migration. These findings suggest that dynamic mineral resource models and live geochemical monitoring (e.g., of elemental mobility and structural changes) are possible without additional physical sampling.

1. Introduction

The threshold to 'big data' varies but is generally the result of a bottleneck in the supply-to-demand system of data (e.g., if a bottleneck exists in supply, the data is small and vice versa). Traditional geochemical data including survey data is small data because of low data velocity. Big geochemical data is improbable to achieve using traditional sampling, preparation and analysis (e.g., Govett, 1983; Friske and Hornbrook, 1991; Moon et al., 2006), resulting in expensive and low-velocity data (Dramsch, 2020). Sustained deployment of artificial intelligence methods tend to occur in domains with data that is near or at big data (Chen and Lin, 2014; Zhang and Lu, 2021; He et al., 2022). A lack of big geochemical data to sustain the use of big data methods hampers the adoption of transdisciplinary methods (artificial intelligence, machine learning and data science) in geosciences (e.g., He et al., 2022; Dramsch, 2020; Zhang et al., 2023; Ghorbani et al., 2022, 2023). Data re-purposing of traditional data is a stop-gap solution, because data is always gathered for a purpose, and changing data analysis methods necessitates changes in data generation. Furthermore, traditional geochemical data (in terms of velocity but also latency) is unsuitable for a range of tasks that include dynamic resource profiling and live environmental monitoring. However, modern sensors, data and analytics solutions are feasible, like for the exploration of critical raw materials (Booysen et al., 2020; Zhang et al., 2022) or real-time control and management (Ghorbani et al., 2022, 2023). Because geochemical data is universally useful across the mineral value chain, from prospecting (Grunsky and de Caritat, 2017, 2019; Zuo, 2020; Lawley et al., 2021), exploration (Edwards and Atkinson, 1986; Hogson, 1990; Carranza, 2012), metal accounting and reconciliation (Ghorbani et al., 2020; Mery et al., 2020), to waste valorisation (Nwaila et al., 2021a, b; Blannin

* Corresponding author.
E-mail address: glen.nwaila@wits.ac.za (G.T. Nwaila).

https://doi.org/10.1016/j.aiig.2023.09.002
Received 11 February 2023; Received in revised form 23 September 2023; Accepted 24 September 2023
Available online 26 September 2023

et al., 2022, 2023), it is important to seek modern solutions towards big geochemical data.

Remote sensing data is closely related to geochemistry and exhibits big data characteristics (e.g., scalability, automation potential, velocity, coverage and coverage rate). Remote sensing data is already useful for a variety of geoscientific tasks (Booysen et al., 2020; Xiao and Wan, 2021; Ghorbani et al., 2022, 2023). Deriving secondary geochemical data from remote sensing data would imbue the secondary data with big data-characteristics of remote sensing data (Shen et al., 2019; Cheng et al., 2021; Xiao et al., 2021; Zhang et al., 2023). Consequently, remote sensors become 'soft' or 'virtual' geochemical sensors (Fortuna et al., 2006; Kadlec et al., 2009). However, this solution is not universally practical, because of complications due to overburden, variable terrain, the scale of the application and ambiguity of inversion outcomes. Additionally, remote sensing-data inversion usually occurs to mineral species and not elemental composition (Asokan et al., 2020; Marghany, 2022). Zhang et al. (2023) developed an inferential method that used geostatistical data augmentation and machine learning to invert remote sensing data into geochemical data. Prototyping of the method occurred using geochemical and remote sensing data from a tailing storage facility (TSF). A high-resolution resource model was spatially integrated with a temporally matched remote sensing dataset, to produce an integrated dataset that paired features (spectral band data) with data labels (resource grade). Furthermore, Zhang et al. (2023) demonstrated that a variety of machine learning algorithms can generate useable models.

TSFs are engineered structures that consist of confining embankments (also known as 'tailings dams') and associated linings, which are used to store mining and mineral processing wastes, and manage water associated with them. TSFs are an unconventional source of raw materials and metals (Blannin et al., 2022, 2023; Nwaila et al., 2021a, 2021b), particularly historic tailings because extraction methods of the past were less selective and multi-elemental, and orebodies were generally higher in grade (Nwaila et al., 2019). TSFs require management and monitoring to prevent unauthorised access and use, and reduce environmental risks (e.g., of acid mine drainage (AMD), air and water dispersal, and rupture of containment walls or dams) (Carlà et al., 2017, 2019; Grebby et al., 2021). However, some historic TSFs, similar to open pit mines, either have been or will be extracted for further processing. Pragmatically, TSFs are generally remote from civil infrastructure, and some are exposed surficially, which provides a remote sensing response (Ciampalini et al., 2013).

In this study, we carry the method that was developed in Zhang et al. (2023) into deployment and demonstrate model deploy-ability across space, leveraging the knowledge that Witwatersrand tailings are consistent in mineral composition. By choosing a TSF similar to that in Zhang et al. (2023), we can ensure that mineral assemblages and chemistry are reasonably similar, such that the deployed models could produce sensible results. Specifically, we generate big geochemical data of the surface of a tailing storage facility (TSF) over an extraction period of four years (2015–2019). The resulting data is big in the sense that it is much more voluminous and of far higher velocity compared with traditional geochemical data, and that traditional resource modelling and exploration was not designed to handle such data. We compare the temporal predictions of the gold grade with that of a geostatistical model (not used for predictive modelling). We showcase the effectiveness of big geochemical data for online resource assessment, environmental monitoring and other purposes by observing changes in the distribution of gold in the TSF and relating those changes with existing knowledge.

2. Background

The data for this study originates from a TSF that hosts mine waste from the Witwatersrand Basin. It is one of many that are located near Randfontein, which is about 40 km west of Johannesburg, South Africa (Fig. 1a). The name of the TSF is Lindum, because it contains tailing materials from the now-exhausted Lindum reef orebody. The location of Lindum TSF is located 4.2 km south of the Dump 20 TSF which was studied by Zhang et al. (2023) (Fig. 1b). The Witwatersrand Basin is famous for its gold endowment, having produced, to date, >30% of the world's total gold (Frimmel, 2019; Frimmel and Nwaila, 2020). A long history of gold extraction from the Witwatersrand Basin resulted in an accumulation of mine waste immediately surrounding the city, with some of the TSFs currently surrounded by informal settlements (Fig. 1b; Ngigi, 2009; Durand, 2012; Dlamini, 2014; McCarthy, 2010; IHRC, 2016; Nwaila et al., 2017). The Lindum reef-tailings are composed of primarily crushed conglomerate remnants, with minor amounts of meta-sedimentary host rocks (e.g., quartzites, shales, greywackes). Gold is hosted inside of the conglomerate remnants. The grain size-distribution of the TSF is not exactly known, but local knowledge suggests that the particles are mainly cm-sized. This implies that the TSFs containing Witwatersrand gold-mine tailings are comparable in composition, despite their individual genetic histories, because both the nature of the reefs and extraction methods were comparable. For our purpose of generating geochemical data from remote sensing data, it is relevant to understand that the Witwatersrand gold is hosted inside thin meta-conglomerate beds with spatially extensive but geologically consistent mineral assemblages and chemistries (Frimmel, 2019). Furthermore, it was empirically demonstrated that the characteristics of the conglomerate beds at the deposit scale are quantitatively predictive of the gold grade (Nwaila et al., 2020; Zhang et al., 2021c). Consequently, it can be expected that remote sensing data capturing variations in material characteristics are likely effective covariates of the gold

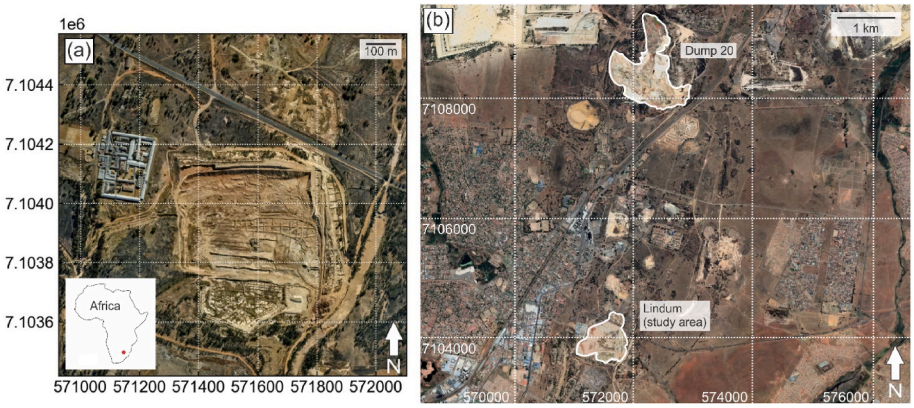

Fig. 1. Location of the studied tailings storage facility located near Randfontein (South Arica). The coordinates provided are in UTM. (a) Close-up of the Lindum tailings facility. Included in the figure is the location of the facility within Africa. (b) The location of Lindum tailings facility is located 4.2 km south of the Dump 20 tailings facility. The satellite image was acquired on 2021-09-25.

grade.

According to high-resolution satellite imagery, the TSF consists of a main portion south of a public roadway and a smaller portion to the north-eastern portion of the map (Fig. 1a). The areal extent of the tailing body is approximately 3.6×10^5 m^2. The TSF is tabular in shape, with a maximal vertical height of about 20 m, as measured in 2014. According to the mine, the TSF in this study was first assessed for its resource potential in 2014, with extraction starting in 2015 and lasting until 2020. At the time of this writing, there is a small amount of remaining material at the TSF. Additionally, drilling was conducted during at least two campaigns in the TSF, with the most recent results used to produce a single static resource model (a geostatistical model), which was used to guide mine planning and the extraction process. Using the supplied borehole data, we recreated the geostatistical model for comparison purposes, but we did not rely on any geochemical data from the studied TSF for training purposes. The erosion of TSFs creates a constant dispersal of fine dust, which implies that there is a contamination of surrounding areas that could give rise to spectral signatures associated with the tailing material (Schonfeld et al., 2014). Hence, we constrain our analysis to the area inside the TSF.

3. Data and methods

3.1. Sentinel-2 dataset

The remote sensing data used in this study is sourced from the constellation of Sentinel-2 Earth observation satellites, which are part of a mission that continuously gathers optical images with a high spatial resolution (10–60 m) over land and coastal areas (Ge et al., 2018). Details of the satellite capabilities were given in Zhang et al. (2023). Here, we summarize the key details. The Sentinel-2 satellites each contains a multispectral imager (MSI) that captures four bands at 10 m, six bands at 20 m, and three atmospheric correction bands at 60 m spatial resolution (Vaiopoulos and Karantzalos, 2016; Park et al., 2017). The atmospheric correction bands include: band 1 for aerosol scattering and cloud detection; band 9 mainly for aerosol retrieval; and band 10 for cloud detection (uncalibrated, see Ge et al., 2020). There are ten visible and near-infrared (VNIR) bands and three short-wave infrared (SWIR) bands make up the image (Clerc and Team, 2022; https://sentinel.esa.int/web/sentinel/missions/sentinel-2). Sentinel-2's MSI is superior to that of the Landsat and the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), in terms of spectral, spatial and temporal resolution from the VNIR to SWIR range (Beiranvand Pour et al., 2016, 2019). The radiometric resolution is 12 bits (per-pixel brightness intensity range of 0–4095). This study uses cloud-free level 1T (L1T, terrain-corrected) images from the top-of-atmosphere (TOA) subset, which and has automatically undergone atmospheric correction to produce a surface reflectance product using the European Space Agency's Sen2Cor processing algorithm's ATCOR model (ESA; Kristollari and Karathanassi, 2020). The images were sourced through the USGS Earth Resources Observation and Science Centre (USGS-EROS at http://earthexplorer.usgs.gov/).

Sentinel-2 images of the deployment area that contains a legacy TSF cover an area of 1.12 km^2 (277.63 Acres). All suitable images from the August 24, 2015 to the October 4, 2019 were collected, which resulted in a total of 17 useable images. Images that were unable to meet the criteria of being cloud-free L1T were discarded. The metadata information for the first image of the series is: "*coordinate reference system (CRS) from European Petroleum Survey Group (EPSG) 32735, Transform: Affine [20, 0, 499980, 0, -20, 7200040]; ID: COPERNICUS/S2/20150824T082656_20150824T082659_T35JNM; Version: 1618001523716097; Data taken identifier: GS2A_20150824T082656_000890_N02.04; 'SPACECRAFT_NAME': 'Sentinel-2A'*". Bands 1 to 12, except for the atmospheric correction bands (1, 9 and 10), were used for the data. In total, the images captured the deployment area on: 2015-08-24; 2015-10-03; 2016-02-07; 2016-04-30; 2016-07-06; 2016-10-07; 2017-03-16; 2017-04-02; 2017-07-06; 2017-10-04; 2018-03-03; 2018-04-07; 2018-10-07; 2019-01-22; 2019-04-27; 2019-07-01; and 2019-10-04.

Data pre-processing is purposely kept maximally identical to that employed in Zhang et al. (2023), which used the Rasterio and GEEMAP Python libraries (Wu et al., 2019; Wu, 2020). The only change from Zhang et al. (2023) in our data pre-processing methodology for machine learning-based predictive modelling is that we employed a clipping process to remove excessively high-intensity pixels. Here, we provide a brief summary of the steps that were executed: (1) affine translation to project the satellite images into the Universal Transverse Mercator (UTM) projection and the World Geodetic System 84 (WGS84) datum; (2) radiometric correction/calibration using the FLAASH method (ENVI, 2009); and, (3) clipping the pixels to the 99.8th percentile and rescaling the bands using a MinMax-based re-scaler (re-scaling all data to the range of the data per band). The only image that contained pixel intensities substantially above the theoretical maximum of 4095 was the image taken on 2017-04-02, which contained an anomalous pixel with an intensity of 11,250. Clipping all images to the 99.8th percentile heuristically standardised the dynamic range of all images and made the entire series of images more comparable in terms of exposure. For more detailed descriptions, see Zhang et al. (2023). All images taken capture the same area, which is primarily centred around the TSF (Fig. 1a). Outside of the TSF, the images show some vegetation coverage, anthropogenic structures associated with the TSF and local roads (Fig. 2).

3.2. Training data

The training data were generated using the methodology from Zhang et al. (2023) and contains a spatially fused geochemical and remote sensing dataset over a different TSF in the vicinity of the studied TSF in South Africa (Fig. 1b). To ensure that the training data are maximally comparable to the deployment data, we adopted the same remote sensing data pipeline as that of the deployment area. Silbanye-Stillwater supplied the geochemical data in the form of gold assays contained in a high-resolution 3D resource model. The geostatistical modelling essentially served to both: (1) augment the data to achieve an abundance of training data; and (2) address the change-of-support problem of point samples representing areal averages. Zhang et al. (2023) performed dimensionality reduction on the 3D model by skimming the top surface of the model, which was then spatially fused (integrated) with remote sensing data from a Sentinel-2 satellite. The resulting dataset contained spatially-matched data points that contained both the band amplitudes of the remote sensing data and unweighted averages of the gold grades within each grid cell. Thereafter, because portions of the TSF had been extracted when the Sentinel-2 satellite was launched, the mis-matched portion was removed (Fig. 3). For detailed background and methodology on the generation and validation of the training data, see Zhang et al. (2023). The training dataset contains a total of 14,721 records.

3.3. Geostatistical resource model

To understand the feasibility of our predicted resource concentration maps using remote sensing data and machine learning, we opt to perform a qualitative and quantitative comparison between the predicted maps and a geostatistical model built using all data supplied by the mine (borehole data only, no grab samples). Because grab-sample data were unavailable for the construction of the geostatistical model, the surface of the geostatistical model is substantially lower in resolution and in representativeness of the actual TSF condition as compared to the geostatistical model in Zhang et al. (2023). The geostatistical model is tabular with a maximal depth of about 20 m and a physical extent that approximates the boundaries of the TSF. The grid spacing is purposely kept the same as the grid spacing of the remote sensing images (10 m), while the vertical grid size is 5 m. The geochemical survey was conducted in 2014 and has not been updated since that time. This implies

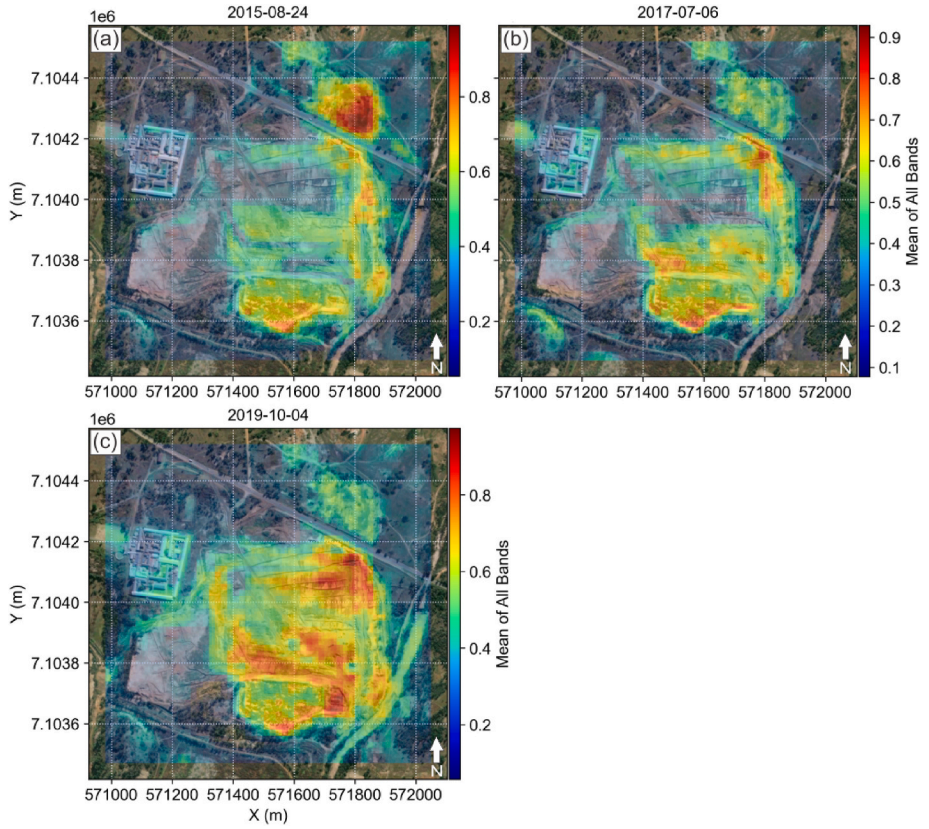

Fig. 2. Mean of all bands (re-scaled) for the remote sensing images taken on the (a) 24th of August 2015, (b) 6th of July 2017, and (c) 4th of October 2019. The coordinates provided are in UTM.

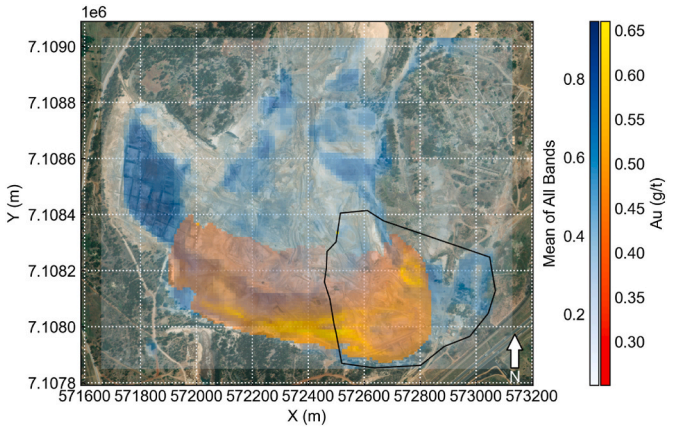

Fig. 3. Spatial data fusion of the resource model and the Sentinel-2 data following Zhang et al. (2023). The polygon shows the mismatched portion due to extraction of the Dump 20 TSF (Fig. 1b), which was removed from the training dataset. The coordinates provided are in UTM.

that the model does not account for subsequent changes to the TSF due to internal (e.g., consolidation, slumping and drainage) and external processes (e.g., extraction and other anthropogenic disturbances). Therefore, to facilitate the fairest comparison between the predicted maps and the geostatistical model, we chose to use the remote sensing image that was temporally best matched with the creation of the geostatistical model (which was just before the onset of extraction). This image was captured by a Sentinel-2 satellite on 2015-08-24, which corresponds to the first image of our series of snapshots.

As the geostatistical model occurs in 3D, we performed two separate dimensional reduction methods on the 3D model to produce two 2D models for comparison purposes: (1) a surficial model that skims only the blocks of the 3D model at the highest elevations (Fig. 4a); and (2) a vertically averaged model that uniformly averages blocks column-wise (Fig. 4b). The model produced through the process (1) leads to a surficial resource map intended to be used for a static comparison of the prediction results of the satellite image taken on 2015-08-24. The model produced through process (2) is intended for comparisons with the subsequent images, as ongoing extraction would have revealed deeper layers that may be more resemblant of the vertical average. Unfortunately, matching the topography exactly with the depths of the 3D model is not possible because: (1) the digital elevation model in this area was not available at the beginning of the extraction process and its resolution is insufficient to allow us to reconstruct actual topography; and (2) slumping and internal processes were significant during the imaging period from 2015 to 2019. In relative comparison, the surficial model is generally lower in grade, especially in the interior of the TSF. The vertically averaged model generally features higher grades, which is particularly prominent near the south edge and the south-eastern portion of the TSF (Fig. 4). This implies that the gold grade generally increases with depth. The heterogeneity of the gold distribution vertically and horizontally (e.g., enrichment at the south edge) are caused by several mechanisms: chemical reactions within the TSF with rainwater, which produced metal-laden water (e.g., AMD), down-slope transport of water and fine fragments that included gold and gold-bearing particles; and a lack of surficial grab samples in the construction of the geostatistical model.

3.4. Machine learning methodology

Inferential data modelling can be performed using machine learning approaches (Mitchell, 1997). Aspects of geoscientific inquiry that are inferential in nature can be formulated into machine learning tasks,

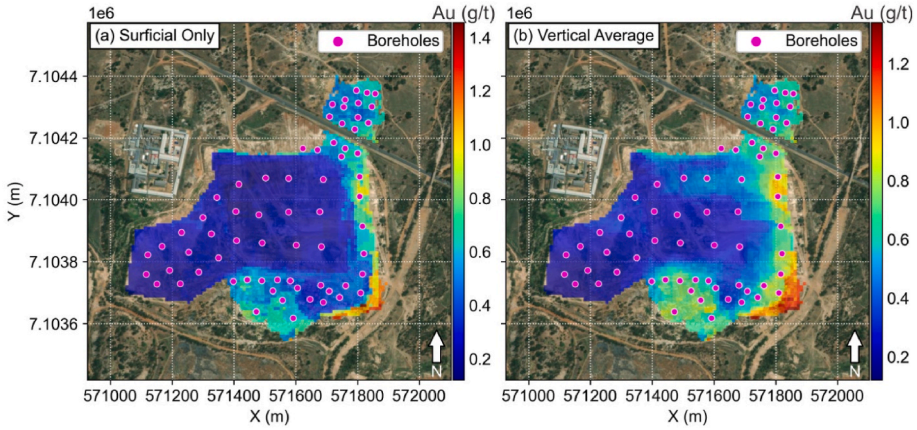

Fig. 4. (a) Surficial 2D geostatistical resource model (block model) of the studied TSF, constructed from borehole data and overlaid on top of the satellite-born imagery of the TSF. (b) Surficial 2D geostatistical resource model (block model) that is produced through vertical averaging of the 3D model of the TSF and overlaid on top of satellite-born imagery of the TSF. The coordinates provided in the figure are in UTM.

which has been used in remote sensing for a variety of purposes (Cracknell and Reading, 2014; Harvey and Fotopoulos, 2016; Bachri et al., 2019; Chakouri et al., 2020; Mather and Tso, 2009; Sehgal, 2012; Al-doski et al., 2013; Madhuanand et al., 2021). A method development study (Zhang et al., 2023) specifically focused on a machine learning-based remote sensing-data inversion method, by modelling relationships between the remote sensing response and geochemical data. A key limiting factor is the availability of high quality and abundant geochemical data, that could be fused with remote sensing data for model training. Zhang et al. (2023) addressed the data abundance issue through geostatistical data augmentation, which transformed sparse survey data into a high-resolution resource model. Additionally, because remote sensing captures an areal response, whereas geochemical samples are essentially point objects, the use of geostatistical data augmentation also solved the change-of-support problem (Gelfand et al., 2001; Gotway and Young, 2002). However, some obvious limitations were identified as well, which include essentially all limitations associated with remote sensing data, such as its mainly two-dimensional nature and occurrences of overburden (Zhang et al., 2023). For the deployment data, remote sensing data from Sentinel-2 was used from bands 1 to 12, except for bands 1, 9 and 10 (Zhang et al., 2023).

The entire methodology used in this study was developed in Zhang et al. (2023). Zhang et al. (2023) explored a variety of supervised learning algorithms and determined that for the task of inferential inversion of remote sensing data, and specific to the training dataset that was used, the best algorithms among the explored ones were the: k-nearest neighbours (kNN); random forest (RF); and adaptive boosting of decision trees (AdaBoost or AB) algorithms. In this deployment study, we adopt the same algorithms and workflow. The kNN algorithm (Cover and Hart, 1967; Fix and Hodges, 1951) is a non-parametric method that averages the data label of a number of training samples (hyperparameter k) that are closest to the target's features to produce an inference (Kotsiantis et al., 2007; Witten and Frank, 2005). RF is an ensemble method that constructs a forest of decision trees in parallel and averages their output. This mitigates the noise sensitivity of decision trees, provided that tree outputs are decorrelated through, e.g., feature bootstrapping (Ho, 1995). The hyperparameters of the RF algorithm includes: maximum number of features per tree; the number of trees; and the minimum number of samples per split, in addition to the tree depth parameter that is inherited from decision trees. Similarly, AB is also an ensemble method of decision trees but uses adaptive boosting, which is conceptually trees constructed in series (Freund and Schapire, 1995). It uses a weighted sum of the output of decision trees as the inference. Weight adjustment occurs through adaptation, whereby the weights of successor trees depend on the error of predecessor trees. The number of trees is the main model hyperparameter, in addition to those inherited from decision trees. Further details of these algorithms, as well as their hyperparameters, are fully described in Zhang et al. (2021a, b and references therein). The choice of algorithms for scientific purposes should also consider algorithm complexity in addition to performance, to address model and result explainability (Zhang et al., 2023). In this deployment study, we do not focus on model explainability, although it would be feasible to conduct SHAP analysis (SHapley Additive exPlanations, Lundberg and Lee, 2017; Rodríguez-Pérez and Bajorath 2020). Model selection was performed through 10-fold cross-validation using the coefficient of determination (CoD or R^2) metric, while the performance profiling was conducted over an average of 100 random runs using the CoD, median absolute error (MedAE) and the mean absolute percentage error (MAPE, which is usually given in fractions despite the name). The MAPE metric is the absolute value of the true minus the predicted value, divided by the true value. The MedAE is the median of all absolute differences between the true and the predicted values. The MedAE is robust to outliers. The algorithm parameter grid is given in Table 2, and the top 3 models and their performance metric scores are given in Table 3 (for model parameters and other details, see Zhang et al., 2023).

In this study, deployment of the method in Zhang et al. (2023) occurs by our deploying the best kNN, RF and AB machine learning models to a different TSF. This permitted us to generate temporal snapshots and animations of resource distribution in the TSF. However, in this study, we added a data pre-processing step, which is the clipping of over-saturated pixel values due to discrepancies between images as part of a series of snapshots. This change in the data pre-processing led to a small but consistent improvement (of a few percent maximum in terms of net change) in predictive modelling performance across all chosen algorithms in Zhang et al. (2023). A performance comparison shows that there is improvement across the CoD and MAPE metrics, but a loss of performance as evaluated by the MedAE metric, see Table 3. The training dataset otherwise remains identical to that in Zhang et al. (2023).

Table 2
Parameter grid for employed machine learning algorithms.

Algorithm	Parameter Grid
kNN	k = {2 to 16 in intervals of 1}
RF	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}
AB	Number of classifiers = {100, 200, 300}, base algorithm = decision tree with the same parameter grid as the random forest algorithm

Table 3
Best prediction results compared across several performance metrics and a comparison with previously published results.

Algorithm	Zhang et al. (2023)			This Study		
	CoD	MAPE	MedAE	CoD	MAPE	MedAE
kNN	0.667	0.068	0.017	0.700	0.014	0.059
RF	0.853	0.045	0.012	0.884	0.011	0.040
AB	0.917	0.027	0.006	0.946	0.005	0.022

4. Deployment results and analysis

4.1. Static deployment and geostatistical results

Deployment of the trained models to the TSF in this study is straightforward. For each algorithm that was found to work well for this task (kNN, RF and AB), the best model is deployed to the sequence of processed satellite-born remote sensing images. To validate the prediction results, we compare our machine learning-based results with our surficial geostatistical model. Because the geostatistical model was created using borehole data from 2014 in anticipation of waste reuse, the most comparable image snapshot for this comparison was taken on 2015-08-24. Subsequent images capture the TSF in a dynamic state during various stages of an extraction process. To facilitate this comparison, we spatially trimmed the predicted maps at the grid level (without re-gridding or re-interpolating to minimize errors) to best match the physical extent of the geostatistical model. The comparisons are provided for the best models using the kNN, RF and AB algorithms (Fig. 5). At a large scale and qualitatively across all results, there exists a high-grade cluster of blocks (or pixels) at the north-eastern tip of the TSF, which is internally graded (Fig. 5a, b and c). This is corroborated by the geostatistical model (Fig. 5d). The internal contrast within the TSF is remarkable across all results, with the lowest grades occurring in a west-east trend towards the south-eastern end of the TSF. This depletion pattern is the result of the drainage of the TSF due to its topography. Drainage in TSFs is known to re-distribute various elements internally and leads to chemical stratification (Hansen, 2018; Nwaila et al., 2021a, b). The high-grade region near the southeastern end of the TSF in the geostatistical model can be attributed to this internal reaction-transport process. At the southern tip of the TSF, all predicted maps demonstrate a gold enrichment, which is corroborated by a similar observation in the geostatistical model. The existence of a weak west-east enrichment trend near the north of the TSF below the public roadway is not visually observable in the surficial geostatistical model (Fig. 4a). However, this trend is visible in the vertically averaged geostatistical model (Fig. 4b). The south-western lobe of the TSF exhibits a similar level of gold enrichment that is not seen in either the surficial or vertically averaged geostatistical models. In comparison with the geostatistical models, the predicted gold concentration maps are systematically less smooth with more detail internal to the TSF. The drainage pathway is completely unobvious in the geostatistical model.

Qualitatively, the maximum grade of all the predicted maps is substantially lower (by a factor of over 2) than the geostatistical models (either the surficial or vertically averaged model). However, this is almost entirely due to the presence of a high-grade region near the south-eastern portion of the TSF in the geostatistical model. However, although it is clear from the predicted maps that this enrichment is most likely the result of internal reaction-transport due to weathering and subsequent drainage, it is actually not sampled by any of the boreholes (Fig. 4). Hence, as far as any geostatistical model is concerned, this region is extrapolated; as such, we have less confidence in its actual gold concentration. A quantitative comparison of key statistical moments is given in Table 4. The total grade is computed as the sum total of the area of each grid cell, multiplied by the grade of that cell, which has dimensions of $g/t \cdot m^2$. The total surficial grade is substantially different between the geostatistical model and the remote-sensing derived models

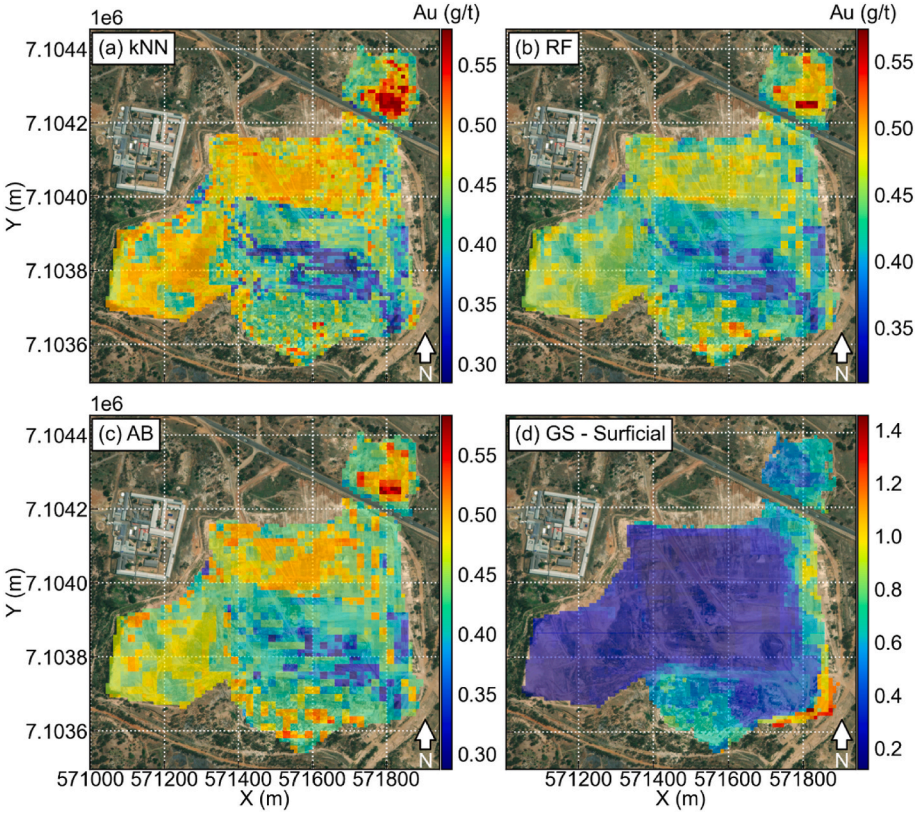

Fig. 5. Predictions of the gold resource map for the image taken on 2015-08-24 using (a) kNN, (b) RF, and (c) AB algorithms, as compared to the geostatistical surficial (GS) model in (d). The coordinates provided in the figure are in UTM.

Table 4
Comparison of predicted resource maps between several methods. STD is the abbreviation for standard deviation.

Resource model type	Number of samples	Total surficial resource (g/t·m^2)	Surficial grade STD (g/t)	Surficial grade mean (g/t)	Surficial grade median (g/t)
Geostatistical - surficial	3819	132,267.4	0.22	0.35	0.23
Geostatistical - vertical average	3819	166,529.4	0.23	0.44	0.38
kNN	4048	177,798.9	0.05	0.44	0.44
RF	4048	175,367.4	0.04	0.43	0.43
AB	4048	176,607.6	0.05	0.44	0.43

using any algorithm. Even accounting for the difference in the number of cells, the relative difference of the total surficial grade of the average of the machine learning resource models is about 25.81% higher than the surficial geostatistical model. However, this difference narrows to just 0.08% if the average of the machine learning-based predictions is compared with the vertically averaged geostatistical model. The standard deviations of the surficial grade are the most contrasting between the geostatistical models and the machine learning-based predictions. This is also clear from a histogram analysis (Fig. 6). It is obvious that the geostatistical models feature gold grades that are more asymmetrically distributed with a pronounced shift of the mean towards lower gold grades, whereas the machine learning-based predictions are more normal with a higher mean. In terms of the resource model mean and median values, the vertically averaged geostatistical model is the most comparable with those predicted by the machine learning algorithms (Table 4). In general, we would expect that the geostatistical model is of higher confidence since it is produced with direct sampling of the TSF, whereas the machine learning-based predictions are made using models that are trained on a different (although materially similar) TSF. However, we make an exception to the level of confidence of the geostatistical model at the high-end of the gold concentration since its estimation occurs outside of the borehole coverage and is a clear result of reaction-transport due to drainage within the TSF. In our case, as some time has elapsed since the sampling for the geostatistical model has taken place, we are unable to eliminate the possibility of extraction and natural physical processes that could have changed the surficial grade distribution.

4.2. *Dynamic model deployment – online mineral resource modelling*

Deployment of the trained models can also occur to the TSF over time, from 2015 to 2019 (Figs. 7 and 8). This produces temporal snapshots of the surficial resource model dynamically, through the extraction period. The results from the kNN, RF and AB models are qualitatively

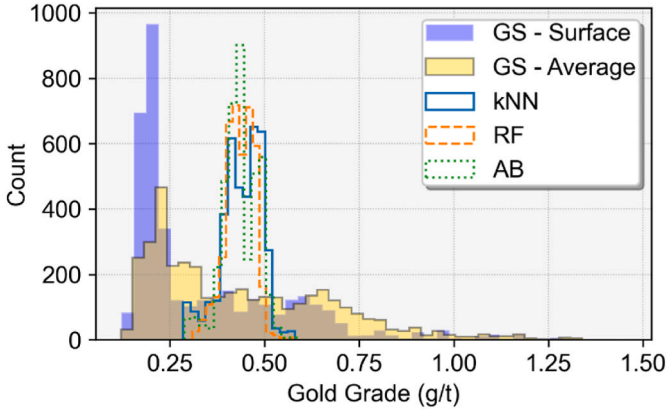

Fig. 6. Histograms of various predictions of the surface gold grade of the TSF and comparison with the geostatistical (GS) resource models.

similar. Therefore, we chose to present the results of the best AB model and provide similar results for the RF model as a comparison. In terms of the temporal coverage, we chose to illustrate the predicted TSF surficial resource grades using one remote sensing image per year, with a temporal interval that is as close to 1 year as possible, except for the inclusion of the image from 2017-04-02 (Figs. 7c and 8c). Our choice to include this image is because this image is the most dissimilar with the rest of the images, as it contained the highest anomalous pixel intensities. The predicted resource grade map of this image is compared with the predictions of the closest related (temporally) maps. This provides qualitative verification that the data pre-processing was effective in remediating the outlier nature of this image. Without the clipping of pixel values, the resource model predicted using this image is substantially lower in grade in general. For all resource models, they (Figs. 7 and 8) demonstrate a change in the surface of the TSF every year. The significant qualitative changes include the gradual disappearance of the drainage channel, which is no longer discernible by 2017-07-06. The results are similar between the predictions of all algorithms, although the RF algorithm yielded the smoothest image (e.g., Figs. 7 and 8).

A major change during the imaging period is the gradual disappearance of the drainage channel. Remnants of the drainage channel are visible in all predictions by July of 2017 (Figs. 7d and 8d). This implies that the extraction had substantially altered the topography of the TSF, such that surficial drainage patterns were affected. However, the timespan of this activity would have occurred substantially before July of 2017, as the cumulative changes of surficial gold grade that follows changes in TSF drainage patterns is not immediate. This process could occur on the order of months to years, depending on a competition of further extraction-induced structural changes and natural processes and is accelerated during the wet seasons. In addition, the highest gold grades near the northeastern portion of the TSF have been extracted the earliest, which is consistent with typical extraction sequencing. As the extraction progressed into 2018 and 2019, a major occurrence is the presence of high gold grades in isolated spots near the southern tip of the TSF, which is likely indicative of extraction-induced exposure of consolidated material through internal reaction and transport due to drainage in the years prior to extraction. The enrichment of the deeper depths of the TSF, especially to the south-eastern edge of the TSF (but also along the southern edge), can be expected from the geostatistical models, which exhibited a similar trend (Fig. 4). Another visible change within the TSF is the gradual disappearance of moderate gold-grade areas to the northern portion of the main body of the TSF. Although this appeared to have been in progress from 2015 to 2017, the most prominent visual changes occurred in 2018–2019. By 2019, the moderate enrichment of gold to the northern portion of the main body seems to have disappeared entirely. This is consistent with the known time of onset of the second extraction phase, targeting less enriched areas. Another approach that demonstrates the big geochemical data nature of our method is the visualisation of changes in daily snapshots in the form of a movie. For this purpose, we linearly interpolated between all 17 images, using a standard frame interval of 3 days to create a smoothly varying animation of the surficial changes of the TSF in gold grade. The images are then rendered into animated movies for each of the algorithms considered (kNN, RF and AB; see Supplementary data). The

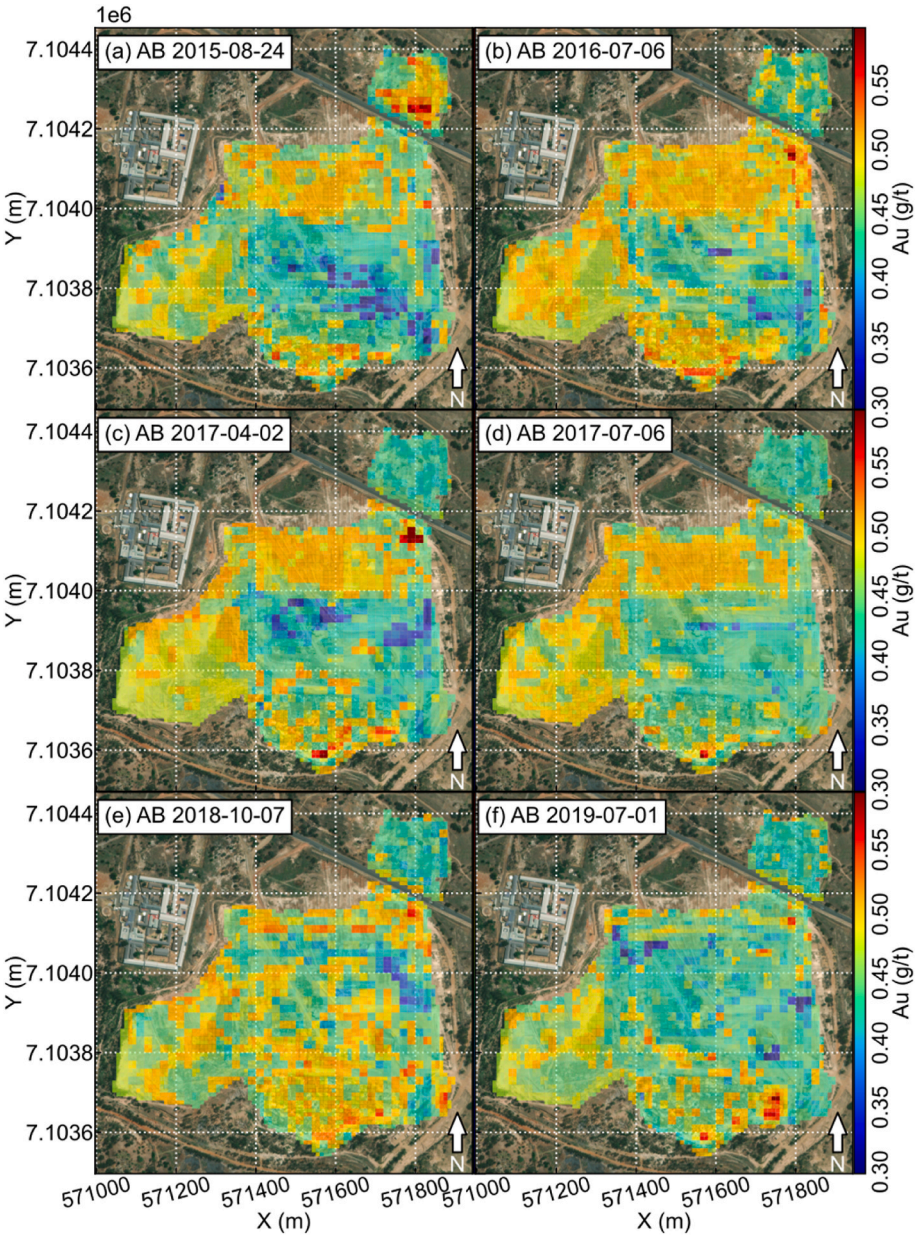

Fig. 7. Temporal snapshots of the surficial resource model of the TSF as predicted using the AdaBoost (AB) algorithm. The snapshots have been rendered into a movie for each of the algorithms used (kNN, RF and AB) and can be viewed in Supplementary Information. The coordinates provided are in UTM.

animated movies are intended for qualitative purposes since the remote sensing images were unevenly spaced temporally due to discarding of images that do not meet cloud-free L1T criteria. This implies that the frames of the animations are variably representative of the state of the TSF. Statistical comparisons between the dynamic geochemical state of the TSF as predicted by various models is similar (RF shown in Fig. 9). In general, there is a fluctuation of the overall surface grade both in terms of the mean and median. Both the mean and median surficial grades as predicted through time are close to the mean but not the median grade in the vertically-averaged geostatistical model (Fig. 9). This can be expected given that the gold grades of the geochemical models are heavily non-normally distributed, whereas the predicted grades are more normally distributed (Fig. 6). The surficial geostatistical model is a poor match of any of the predicted snapshots through time (Fig. 9).

5. Discussion

5.1. Transferability of trained machine learning models to new environments

In the context of this study, a key factor contributing to the ability to deploy existing trained machine learning models from their original TSF to the current TSF is that we have a-priori knowledge of similarity of the composition of both TSFs. This knowledge was not obtainable remotely, because remote sensing data are unable to directly capture elemental or mineral distributions. A data-only comparison is ambiguous and insufficient to ensure model transferability due to the physics of (low energy) remote sensing. Ground reflections originate from photon-mineral interactions primarily. Consequently, ground reflections contain an ambiguous combination of both compositional and microstructural signatures (e.g., mineral structure; Marghany, 2022). Hence, even where

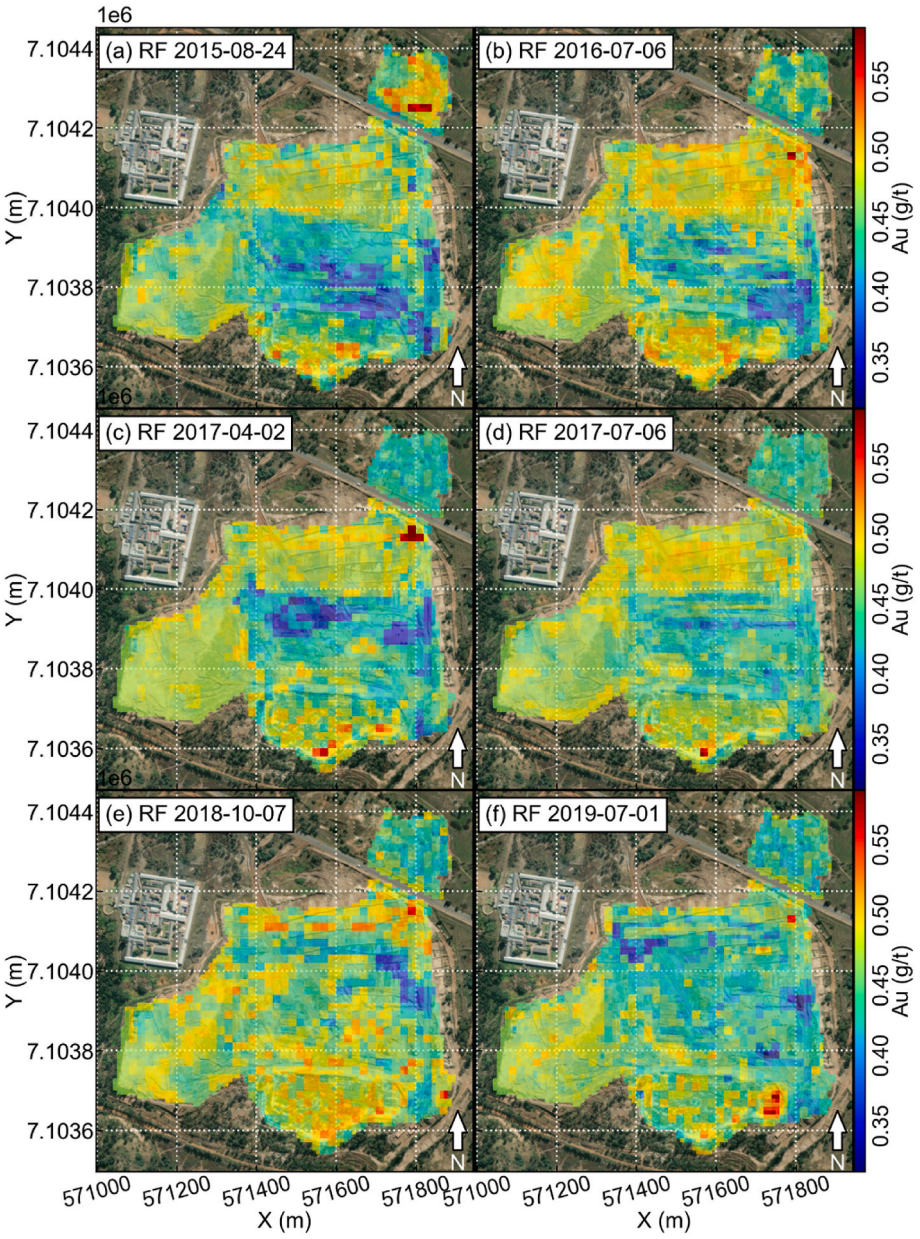

Fig. 8. Temporal snapshots of the surficial resource model of the TSF as predicted using the random forest (RF) algorithm. The snapshots have been rendered into a movie for each of the algorithms used (kNN, RF and AB) and can be viewed in Supplementary Information. The coordinates provided are in UTM.

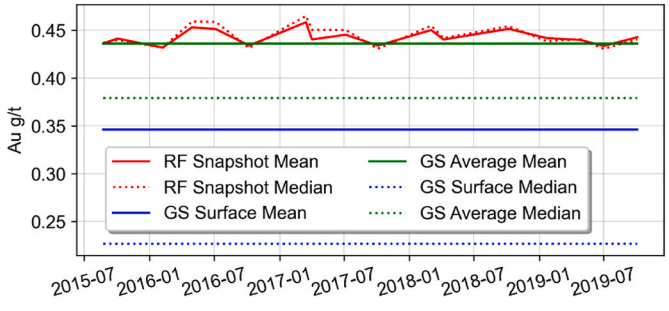

Fig. 9. Statistics of the snapshot predictions of the surface gold grade of the TSF using the random forest (RF) algorithm and comparison with the geostatistical (GS) resource models.

a deployment area shows spectral signatures that are similar to a training area, it is not possible to conclude that the material composition is similar without additional knowledge. In this case, deployment of models could be met with misleading or unrealistic results. As was noticed by Zhang et al. (2023), although trained models could be theoretically deployed over any surface, the results require further validation. The general case of geostatistical transferred learning is not yet solved and, cross-validation in the training area can seriously over-estimate actual model performance due to spatial correlation of covariates (and other effects, Hoffimann et al., 2021). In Zhang et al. (2023), the authors purposely deployed the trained models outside of their TSF and observed that although predicted resource grades were expectedly low, they were not zero. Incorporating training data outside of the TSF can suppress the models' response to non-relevant spectral signatures (e.g., of buildings and trees). However, this cannot be performed universally because the types of landcover that does not correspond to one or a few types of the composition of interest can arise from

an infinite combination of terrains. Hence, the feature space spanned by non-target land cover is infinitely complex in general. If this approach were to be taken, care would also have to be taken such that the spectral response of non-interesting compositions does not significantly overlap with that of the interesting compositions. Open tailings such as the Dump 20 and Lindum TSFs are affected by material dispersal mechanisms, such as surficial erosion, wind, rain and transport losses (e.g., Tutu et al., 2008; Ojelede et al., 2014). Unfortunately, this means that remote sensing-image pixels near such TSFs are likely contaminated by spectral signatures of minerals from the TSF. It would seem that some knowledge of the nature of remote sites, where there is uncertainty regarding their composition, is important to the spatial transferability of trained models.

5.2. Spatio-temporal changes of gold grade distribution in the tailings storage facilities

The most important contribution of this study lies in the demonstration of streams of geochemical data. Certainly, the derived geochemical data that was used to analyze temporal changes in the TSF are produced at a spatial and temporal resolution far surpassing that of traditional geochemical surveys. In the case of monitoring TSFs, our method is capable of both identifying natural processes (e.g., drainage pattern) and monitoring anthropogenic activity (extraction). The response rate, or latency of monitoring depends on the imaging frequency, which in this case is controlled by Sentinel-2 satellite's capability. The data are considered 'big' in the sense of a stream of constant images that cover the area of interest. This type of data is essentially modernised non-contact sensor data, which is yet to be popularised in the domains of geochemistry, geology, resource estimation and mine operations. However, clearly, the big data approach to geochemistry adds tremendous value to every discipline where geochemistry is currently significant in practice. For resource estimation, the resource model could be dynamically generated and modified, at the pace of sensor output. Should even higher rates of output, or alternatively, higher spatial and/or spectral accuracy be required, it is an engineering matter to switch from satellite-born remote sensing to drone-based remote sensing (hybridised approaches are also possible). In this case, the frequency of snapshots could become as high as desired and certainly once per day is feasible (Booysen et al., 2020). In addition, with drone-based remote sensing, spatial resolution at the cm-scale would be possible. Even in the case of satellite-born remote sensing data, the idea of weekly or monthly updates to a resource model at a resolution of decametres is unheard of within the traditional practices of geochemical sampling and analysis. To be able to monitor extraction, where surface exposure exists (TSFs or e.g., open pit mines), remote sensing-derived geochemical data greatly facilitates the adoption of big data and modern sensors into traditional workflows. If such workflows were utilised, a company could dynamically prioritise extraction sequencing to maximise their business outcome. This would undoubtedly create enhanced competitiveness within the industry and especially in the case of waste valorisation, where resource grades are lower and therefore profit margins are thinner and business continuity is more influenced by external variables, such as market conditions. Where the companies are more competitive, more waste could be re-used, enhancing primary resource sustainability. In the case of the Witwatersrand gold TSFs, because they contain substantial radionuclides (uranium), heavy and semi-metals and are near human settlements (Schonfeld et al., 2014), the environmental and health argument for their immediate and complete re-use and land remediation is strong. This is in addition to their relatively high mineral resource grades, given that their accumulation from mining began over 100 years ago (Tutu et al., 2008; Frimmel, 2019).

5.3. Implication of the proposed method in geosciences

The ability to generate big geochemical data may be of high relevance to the exploration community. This benefit mainly includes the ability to generate data at a pace and a resolution that is far higher than current standards of practice, which spans years from survey planning to data availability at the regional level. Where there had been geochemical surveys, legacy data, especially surficial geochemical data is of tremendous value because it is essentially training data to derive secondary geochemical data. However, as with all big data types of approaches, the increase in data abundance can be expected to offset some losses due to the precision of the data generation method using predictive modelling compared with analytical instrumentation. This may be a worthwhile balance in favour of activities such as exploration (and hence mapping) because such activities are exploratory in nature and makes use of a system-level of data (e.g., coherent spatial patterns rather than a single concentration outlier). For this activity, the availability of geochemical data at a fixed grid spacing with a resolution on the order of decametres implies that much finer signatures could be recognised. In fact, the resolution of remote sensing-derived geochemical data may be sufficiently high to essentially render regional scale surveys into greenfield surveys, where they were once brownfield due to existing discoveries and resolution limitations of traditional primary geochemical data. This can obviously facilitate the discovery of more mineral occurrences that may lead to the additional discovery of deposits. An additional benefit of the big data aspect of geochemical data is that they can be averaged through time and space to increase the precision of the data. Such an approach is very popular in the astrophysics/astronomy community, for example, for the purpose of planetary and space exploration. This is predicated on the fact that natural geological landscapes evolve very slowly relative to the pace of remote sensing-based data generation. Hence, averaging through time can be used to improve the resolving power of the data. This aspect would be entirely impossible without the big data aspect of geochemical data.

The case for adopting big geochemical data is even stronger in the modern usage of geoscientific data through data analytics. Transdisciplinary techniques that primarily use methods from artificial intelligence, machine learning and data science require copious amounts of data that cannot be sustained through legacy data alone. In fact, the sparsity of big data in geosciences is probably the most important obstacle to adopting modern data talents (Karpatne et al., 2018; He et al., 2022). Skilled scientists trained in artificial intelligence for example, already face a barrier of entry into geosciences that primarily consists of discipline-specific knowledge and context (He et al., 2022). The added issue of sparsity of a sustainable data supply means that even the keenest geodata scientists would have to switch between isolated projects as data are quickly consumed. An implication of this is that geodata scientists may be forced to adopt projects from multiple sub-disciplines of geosciences to maximise success (as data-driven activities are exploratory and high risk). This type of multi-disciplinary agility is rare because each sub-domain of geosciences requires its own knowledge to understand its data, pre-processing methods (e.g., denoising, transformations and subjective interpretations) and scientific concepts. In addition, data-driven and hypothesis-driven science are philosophically incompatible in terms of their requirements on data. In other domains where transdisciplinary techniques have found success (e.g., behavioural analytics and recommender engines in online commerce), models are typically trained and deployed on identical data streams. Deployment of models is a relatively rare event in traditional geosciences because of data characteristics that include a low data velocity (data streams are rare). As it pertains to geochemistry, big geochemical data are not a concept that could be arrived at from within traditional analytical laboratories and within the discipline. As such, it represents an evolutionary leap and not an incremental development. Integrating remote sensing, legacy geochemical data, geostatistical data augmentation and machine learning proved to be an effective bridge to

bring the benefits of high-data-velocity from remote sensing into geochemistry. Alternatively, using the tools of geochemistry, such as resource estimation, it is possible to continuously analyse sustainable sources of remote sensing data, to derive stronger proxies to natural processes in the form of elemental concentrations.

6. Conclusion

This study deployed a machine learning-based inferential inversion method to produce snapshots of big geochemical data. The pre-requisites to our approach include the availability of a high-quality training dataset that covers a terrain with a similar composition to an area of deployment. In the sense that remote sensing data are considered big data, the derived geochemical data would also be big data. The ability to generate big geochemical data brings many benefits to both geoscience and engineering communities, in the form of: (1) cheap secondary geochemical data that realises the additional value in legacy geochemical data; (2) a sustainable supply of geochemical data to sustain data-driven applications; (3) the ability to dynamically model and monitor locations of interest; (4) the ability to adopt data analysis approaches (e.g., resource estimation) that were previously foreign to the remote sensing community; and (5) a low-cost method to plan surficial mining processes that support resource sustainability. We think the ability to provide a sustainable supply of geochemical data to downstream applications is probably our study's most significant and timely implication. By our observation, there is a pervasive desire from many scientific and engineering domains to adopt data-driven methods, but the need for sustainable data generation and the system-nature of data-driven methods (as opposed to the typical needs of scientific reduction) has been overshadowed by the more prominent rise of glamorous applicative examples of artificial intelligence, machine learning and data science. However, a sustainable supply of geoscientific data is an absolute necessity to sustainably attract geodata scientists, whose stability and career progression within the field of geosciences is the ultimate determinant of the long-term success of data-driven methods.

Funding

This study was supported by a Department of Science and Innovation (DSI)-National Research Foundation (NRF) Thuthuka Grant (Grant UID: 121973) and DSI-NRF CIMERA.

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 thank Sibanye-Stillwater for providing geochemical data used to train various machine learning algorithms. We also thank Prof. Kunfeng Qiu (China University of Geosciences, Beijing) for highly constructive comments which have greatly improved this manuscript. We thank Dr. Hua Wang for editorial handling.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aiig.2023.09.002.

References

Al-doski, J., Mansor, S.B., Zulhaidi, H., Shafri, M., 2013. Image classification in remote sensing. Environ. Earth Sci. 3 (10), 141–148.

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.

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.

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.

Blannin, R., Frenzel, M., Tolosana-Delgado, R., Gutzmer, J., 2022. Towards a sampling protocol for the resource assessment of critical raw materials in tailings storage facilities. J. Geochem. Explor. 236, 106974 https://doi.org/10.1016/j.gexplo.2022.106974.

Blannin, R., Frenzel, M., Tolosana-Delgado, R., Buttner, P., Gutzmer, J., 2023. 3D geostatistical modelling of a tailings storage facility: resource potential and environmental implications. Ore Geol. Rev. 105337 https://doi.org/10.1016/j.oregeorev.2023.105337.

Booysen, R., Jackisch, R., Lorenz, S., Zimmermann, R., Kirsch, M., Mex, P.A.M., Gloaguen, R., 2020. Detection of REEs with lightweight UAV-based hyperspectral imaging. Sci. Rep. 10, 17450 https://doi.org/10.1038/s41598-020-74422-0.

Carlà, T., Farina, P., Intrieri, E., Botsiala, K., Casagli, N., 2017. On the monitoring and early-warning of brittle slope failures in hard rock masses: examples from an open-pit mine. J. Eng. Geol. 228, 71–81. https://doi.org/10.1038/s41598-019-50792-y.

Carlà, T., Intrieri, E., Raspini, F., Bardi, F., Farina, P., Ferretti, A., Colombo, D., Novali, F., Casagli, N., 2019. Perspectives on the prediction of catastrophic slope failures from satellite InSAR. Sci. Rep. 9 (1), 1–9. https://doi.org/10.1038/s41598-019-50792-y.

Chakouri, M., El Harti, A., Lhissou, R., El Hachimi, J., Jellouli, 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.

Carranza, E.J.M., 2012. Primary geochemical characteristics of mineral deposits: implications for exploration. Ore Geol. Rev. 45, 1–4. https://doi.org/10.1016/j.oregeorev.2012.02.002.

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., Team, M.P.C., 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. (Accessed 27 June 2022).

Cover, T., Hart, P., 1967. Nearest neighbour pattern classification. IEEE Trans. Inf. Theor. 13, 21–27.

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.

Dlamini, T.C., 2014. Radionuclides and toxic elements transfer the Prince Dump to the surrounding vegetation. In: Roodepoort South Africa: Potential Radiological and Toxicological Impact on Humans. University of North West, South Africa. M.Sc. thesis.

Dramsch, J.S., 2020. 70 years of machine learning in geoscience in review. Adv. Geophys. 61, 1–55. https://doi.org/10.1016/bs.agph.2020.08.002.

Durand, J.F., 2012. The impact of gold mining on the Witwatersrand on the rivers and karst system of Gauteng and North West Province, South Africa. J. Afr. Earth Sci. 68, 24–43.

Edwards, R., Atkinson, K., 1986. Ore Deposit Geology, first ed. Springer.

ENVI, 2009. Atmospheric Correction Module: QUAC and FLAASH User's Guide. http://www.exelisvis.com/portals/0/pdfs/envi/Flaash_Module.pdf2009. (Accessed 31 March 2022).

Fix, E., Hodges, J.L., 1951. An important contribution to nonparametric discriminant analysis and density estimation. Int. Stat. Rev. 57 (3), 233–238. https://doi.org/10.2307/1403796.

Fortuna, L., Graziani, S., Rizzo, A., Xibilia, M.G., 2006. Soft Sensors for Monitoring and Control of Industrial Processes: Advances in Industrial Control. Springer.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalisation of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

Frimmel, H.E., 2019. The Witwatersrand Basin and its gold deposits. In: Kröner, A., Hofmann, Al (Eds.), The Archaean Geology of the Kaapvaal Craton, Southern Africa. Springer, Cham, pp. 255–275. https://doi.org/10.1007/978-3-319-78652-0_10.

Frimmel, H.E., Nwaila, G.T., 2020. Geologic evidence of syngenetic gold in the Witwatersrand goldfields, South Africa. In: Sillitoe, R.H., Goldfarb, R.J., Robert, F., Simmons, S.F. (Eds.), Geology of the World's Major Gold Deposits and Provinces 23. Society of Economic Geologists, Littleton, Colorado, Special Publication, pp. 645–668. https://doi.org/10.5382/SP.23.31.

Friske, P.W.B., Hornbrook, E.H.W., 1991. Canada's national geochemical reconnaissance programme. Trans. Inst. Min. Metall. Sec. 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.
Gelfand, A.E., Zhu, L., Carlin, B.P., 2001. On the change of support problem for spatio-temporal data. Biostatistics 2 (1), 31–45. https://doi.org/10.1093/biostatistics/2.1.31.
Ghorbani, Y., Nwaila, G.T., Chirisa, M., 2020. Systematic framework toward a highly reliable approach in metal accounting. Miner. Process. Extr. Metall. Rev. 43 (5), 664–678. https://doi.org/10.1080/08827508.2020.1784164.
Ghorbani, Y., Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., 2022. Framework components for data-centric dry laboratories in the minerals industry: a path of science-and-technology-led innovation. Extr. Ind. Soc. 10, 101089 https://doi.org/10.1016/j.exis.2022.101089.
Ghorbani, Y., Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., Safari, M., Hoseinie, 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. Rock geochemistry in mineral exploration. In: Govett, G.J.S. (Ed.), Handbook of Exploration Geochemistry, vol. 3.
Grebby, S., Sowter, A., Gluyas, J., Toll, D., Gee, D., Athab, A., Girindran, R., 2021. Advanced analysis of satellite data reveals ground deformation precursors to the Brumadinho Tailings Dam collapse. Commun. Earth Environ. 2, 2. https://doi.org/10.1038/s43247-020-00079-2.
Grunsky, E.C., de Caritat, P., 2017. Advances in the use of geochemical data for mineral exploration. In: Tschirhart, V., Thomas, M.D. (Eds.), Proceedings of Exploration 17: Sixth Decennial International Conference on Mineral Exploration, pp. 441–456.
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.
Hansen, R.N., 2018. Inter-comparison geochemical modelling approached and implications for environmental risk assessments: a Witwatersrand gold tailings source term characterization study. J. Appl. Geochem. 95, 71–84. https://doi.org/10.1016/j.apgeochem.2018.05.017.
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/isprsarchives-XLI-B8-423-2016.
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. J. Appl. Geochem. 105273 https://doi.org/10.1016/j.apgeochem.2022.105273.
Ho, T.K., 1995. Random decision forests. In: IEEE Proceedings of the 3rd International Conference on Document Analysis and Recognition, vol. 1. Montréal, Canada, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994.
Hoffimann, J., Zortea, M., de Carvalho, B., Zadrozny, B., 2021. Geostatistical learning: challenges and opportunities. Front. Appl. Math. Stat. 7, 689393 https://doi.org/10.3389/fams.2021.689393.
Hogson, C.J., 1990. Uses (and abuses) of ore deposit models in mineral exploration. Geosci. Can. 17 (2).
IHRC (Harvard Law School International Human Rights Clinic), 2016. Clinic Highlights Human Rights Costs of South African Gold Mining. Harvard Law Today. https://hls.harvard.edu/today/clinic-highlights-human-rights-costs-south-african-gold-mining/. (Accessed 26 August 2022).
Kadlec, P., Gabrys, B., Strandt, S., 2009. Data-driven soft sensors in the process industry comput. Chem. Eng. 33 (4), 795–814. https://doi.org/10.1016/j.compchemeng.2008.12.012.
Karpatne, 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. https://doi.org/10.1109/TKDE.2018.2861006.
Kotsiantis, S.B., Zaharakis, I., Pintelas, P., 2007. Supervised machine learning: a review of classification techniques. In: Maglogiannis, I., Karpouzis, K., Wallace, M., Soldatos, J. (Eds.), Emerging Artificial Intelligence Applications in Computer Engineering, vol. 160. IOS Press, pp. 3–24, 1.
Kristollari, V., Karathanassi, V., 2020. Fine-tuning self-organising maps for Sentinel-2 Imagery: separating clouds from bright surfaces. Rem. Sens. 12 (12), 1923. https://doi.org/10.3390/rs12121923.
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 (± Cu ± Co ± PGE) sulphide mineral systems. Ore Geol. Rev. 132, 103985 https://doi.org/10.1016/j.oregeorev.2021.103985.
Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Madhuanand, 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.
Marghany, M., 2022. Remote Sensing and Image Processing in Mineralogy. CRC Press. https://doi.org/10.1201/9781003033776.
Mather, P.M., Tso, B., 2009. Classification Methods for Remotely Sensed Data, second ed. Taylor & Francis.
McCarthy, T.S., 2010. The Decant of Acid Mine Water in the Gauteng City Region Analysis, Prognosis and Solutions. Provocations Series, Gauteng City Region Observatory. Universities of the Witwatersrand and Johannesburg, South Africa.
Mery, N., Marcotte, D., Dutaut, R., 2020. Constrained kriging: an alternative to predict global recoverable resources. Nat. Resour. Res. 29, 2275–2289. https://doi.org/10.1007/s11053-019-09601-6.
Mitchell, T.M., 1997. Machine Learning. McGraw Hill.
Moon, C.J., Whateley, M.K.G., Evans, A.M., 2006. Introduction to Mineral Exploration. Blackwell Publishing.
Ngigi, S.M., 2009. Establishing Effective, Appropriate and Applicable Technologies in Treating Contaminated Surface Water as Part of a Rehabilitation Strategy for the Princess Dump in Roodepoort. M.Sc. thesis. University of the Witwatersrand, South Africa. West of Johannesburg.
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., 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.
Nwaila, G.T., Zhang, S.E., Frimmel, H.E., Manzi, M.S., Dohm, C., Durrheim, R.J., Burnett, M., Tolmay, L., 2020. Local and target exploration of conglomerate-hosted gold deposits using machine learning algorithms: a case study of the Witwatersrand gold ores, South Africa. Nat. Resour. Res. 29, 35–159. https://doi.org/10.1007/s11053-019-09498-1.
Nwaila, P.C., Nwaila, G.T., Broadhurst, J.L., 2017. Long-term pollution and impacts on local communities from defunct gold tailings deposits: a case study of Davidsonville, South Africa. In: Proceeding of the 14th SGA Biennial Meeting, vol. 4, pp. 1453–1456, 20-23 August 2017, Québec City, Canada.
Ojelede, M.E., Annegarn, H.J., Kneen, M.A., 2014. Evaluation of aeolian emissions from gold mine tailings on the Witwatersrand. Aeolian Res 3, 477–486. https://doi.org/10.1016/j.aeolia.2011.03.010.
Park, H., Choi, J., Park, N., Choi, S., 2017. Sharpening the VNIR and SWIR bands of Sentinel-2A imagery through modified selected and synthesised band schemes. Rem. Sens. 9 (10), 1080. https://doi.org/10.3390/rs9101080.
Rodríguez-Pérez, R., Bajorath, J., 2020. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. J. Med. Chem. 63 (16), 8761–8777. https://doi.org/10.1021/acs.jmedchem.9b01101.
Schonfeld, S.J., Winde, F., Albrecht, C., Kielkowski, D., Liefferink, M., Patel, M., Sewram, V., Stoch, L., Whitaker, C., Schüz, J., 2014. Health effects in populations living around the uraniferous gold mine tailings in South Africa: gaps and opportunities for research. Cancer Epidemiol 38 (5), 628–632. https://doi.org/10.1016/j.canep.2014.06.003.
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.
Tutu, H., McCarthy, T.S., Cukrowska, E., 2008. The chemical characteristics of acid mine drainage with particular reference to sources, distribution and remediation: the Witwatersrand Basin, South Africa as a case study. Appl. Geochem. 23, 3666–3684. https://doi.org/10.1016/j.apgeochem.2008.09.002.
Vaiopoulos, A., Karantzalos, K., 2016. Pansharpening on the narrow VNIR and SWIR spectral bands of Sentinel-2. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. 41, 723–730.
Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, second ed. Morgan Kaufman, San Francisco.
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. Environ. 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., 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., 2021a. 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., Bourdeau, J.E., Nwaila, G.T., Ghorbani, Y., 2022. Advanced geochemical exploration knowledge using machine learning: prediction of unknown elemental concentrations and operational prioritization of re-analysis campaigns. Artif. Intell. Geosci. 3, 86–100.

Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., Ashwal, L.D., 2021b. 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.

Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., Ghorbani, Y., Carranza, E.J.M., 2023. Towards big geochemical data from high-resolution remote sensing data via machine learning: application to a tailings storage facility in the Witwatersrand goldfields. Artif. Intell. Geosci. https://doi.org/10.1016/j.aiig.2023.01.005 (in press).

Zhang, S.E., Nwaila, G.T., Tolmay, L., Frimmel, H.E., Bourdeau, J.E., 2021c. Integration of machine learning algorithms with Gompertz Curves and Kriging to estimate resources in gold deposits. Nat. Resour. Res. 30, 39–56. https://doi.org/10.1007/s11053-020-09750-z.

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.