



Original Paper

# Data-Driven Predictive Modeling of Lithofacies and Fe In-Situ Grade in the Assen Fe Ore Deposit of the Transvaal Supergroup (South Africa) and Implications on the Genesis of Banded Iron Formations

Glen T. Nwaila,<sup>1,4</sup> Steven E. Zhang,<sup>1,2,3</sup> Julie E. Bourdeau,<sup>1,3</sup> Elekanyani Negwangwatini,<sup>4</sup> Derek H. Rose,<sup>5</sup> Mark Burnett,<sup>6</sup> and Yousef Ghorbani <sup>7,8</sup>

Received 29 March 2022; accepted 6 July 2022  
Published online: 20 July 2022

The Assen Fe ore deposit is a banded iron formation (BIF)-hosted orebody, occurring in the Penge Formation of the Transvaal Supergroup, located 50 km northwest of Pretoria in South Africa. Most BIF-hosted Fe ore deposits have experienced post-depositional alteration including supergene enrichment of Fe and low-grade regional metamorphism. Unlike most of the known BIF-hosted Fe ore deposits, high-grade hematite (> 60% Fe) in the Assen Fe ore deposit is located along the lithological contacts with dolerite intrusions. Due to the variability in alteration levels, identifying the lithologies present within the various parts of the Assen Fe ore deposit, specifically within the weathering zone, is often challenging. To address this challenge, machine learning was applied to enable the automatic classification of rock types identified within the Assen Fe ore mine and to predict the in-situ Fe grade. This classification is based on geochemical analyses, as well as petrography and geological mapping. A total of 21 diamond core drill cores were sampled at 1 m intervals, covering all the lithofacies present at Assen mine. These were analyzed for major elements and oxides by means of X-ray fluorescence spectrometry. Numerous machine learning algorithms were trained, tested and cross-validated for automated lithofacies classification and prediction of in-situ Fe grade, namely (a) k-nearest neighbors, (b) elastic-net, (c) support vector machines (SVMs), (d) adaptive boosting, (e) random forest, (f) logistic regression, (g) Naïve Bayes, (h) artificial neural network (ANN) and (i) Gaussian process algorithms. Random forest, SVM and ANN classifiers yield high classification accuracy scores during model training, testing and cross-validation. For in-situ Fe grade prediction, the same algorithms also consistently yielded the best results. The predictability of in-situ Fe grade on a per-lithology basis, combined with the fact that CaO and SiO<sub>2</sub> were the strongest predictors of Fe concentration, support the hypothesis that the process that led to Fe enrichment in the Assen Fe ore deposit is dominated by supergene processes. Moreover, we show that predictive modeling can be

<sup>1</sup>Wits Mining Institute, University of the Witwatersrand, 1 Jan Smuts Ave., Johannesburg 2000, South Africa.

<sup>2</sup>PG Techno Wox (Pty) Limited, 39 Kiewiet Street, Randfontein 1759, South Africa.

<sup>3</sup>Geological Survey of Canada, 601 Booth Street, Ottawa, ON K1A 0E9, Canada.

<sup>4</sup>School of Geosciences, University of the Witwatersrand, 1 Jan Smuts Ave., Johannesburg 2000, South Africa.

<sup>5</sup>Department of Geology, University of Johannesburg, P.O. Box 524 Johannesburg 2006, South Africa.

<sup>6</sup>AMC Consultants, Level 7, Nicholson's House Nicholson's Walk, Maidenhead SL6 1LD, UK.

<sup>7</sup>Department of Civil, Environmental and Natural Resources Engineering, Luleå University of Technology, SE-97187 Luleå, Sweden.

<sup>8</sup>To whom correspondence should be addressed; e-mail: yousef.ghorbani@ltu.se

used to demonstrate that in this case, the main differentiator between the predictability of Fe concentration between different lithofacies lies in the strength of multivariate elemental associations between Fe and other oxides. Localized high-grade Fe ore along with lithological contacts with dolerite intrusion is indicative of intra-basinal fluid circulation from an already Fe-enriched hematite. These findings have a wider implication on lithofacies classification in weathered rocks and mobility of economic valuable elements such as Fe.

**KEY WORDS:** Assen Fe ore deposit, Banded Iron Formation, Transvaal Supergroup, Supergene enrichment, Machine learning.

## INTRODUCTION

Iron can be found in a variety of minerals, such as hematite (70% Fe), magnetite (72.4% Fe), limonite (60% Fe), siderite (48.3% Fe) and pyrite (46% Fe)<sup>1</sup> (Muwanguzi et al., 2012). However, in Fe ores, Fe percentages are lower because of impurities. Banded iron formation- (BIF) hosted high-grade (> 60% Fe) hematite ore deposits constitute the most important source of the Fe mined, both historically and currently (Hagemann et al., 2016; Smith & Beukes, 2016). In 2020, 2.4 million metric tons of Fe ore were mined worldwide, with the majority of the ore (98%) being used to produce steel (USGS, 2021). BIFs are chemical sedimentary rocks comprised of alternating layers of Fe-poor chert and Fe-rich minerals (Klein, 2005; Trendall, 2005). These alternating layers can reach several hundred meters in thickness and may extend laterally for hundreds of kilometers. Economically important BIF-hosted Fe ore deposits were largely formed during the Archean and Paleoproterozoic eons (Smith & Beukes, 2016). A general consensus for the depositional mechanisms for BIFs has not yet been reached; however, most are thought to have been formed as a response to diverse environmental changes, including but not limited to those time periods when large igneous provinces were emplaced and the great oxygenation event (Bekker et al., 2010; Pufahl & Hiatt, 2012; Hagemann et al., 2016; Dreher et al., 2021). Significant BIF-hosted Fe ore deposits are found in Western Australia's Hamersley Province (e.g., Pilbara and Yilgarn cratons), São Francisco and Amazon cratons in Brazil, Singhbhum and Bastar cratons in India and the Kaapvaal Craton in South Africa (Hagemann et al., 2016 and references therein).

BIFs can be divided into three types (i.e., Algoma, Lake Superior and Rapitan-Urucum) based

on their geotectonic settings (Hagemann et al., 2016; Smith & Beukes 2016; Smith, 2018). Subaqueous volcanic rocks found in the convergent margins of Archean and Paleoproterozoic granite-greenstone belts are stratigraphically linked or interlayered with Algoma-type BIFs (Gross, 1980, 1993). This type of deposit encompasses the Archean-aged Yilgarn (55% Fe) and Pilbara-type deposits in Western Australia (Teitler et al., 2014; Hagemann et al., 2016). The largest known BIF-hosted Fe ore deposits, Lake Superior-type BIFs, are thought to have been formed in a passive-margin environment during the Proterozoic era (Gross, 1980, 1993). Deposits of this type include the Hamersley (high-grade ores typically contain 64% Fe) in Australia and the Transvaal (ca. 61% Fe) in South Africa (Beukes and Gutzmer, 2008; Thorne et al., 2008). The Rapitan-Urucum-type BIFs are constrained to the Neoproterozoic (715–580 Ma) in glaciogenic sedimentary sequences (Halverson et al., 2011; Hagemann et al., 2016). Deposits of this type include the Urucum (54% Fe) in Brazil and the Rapitan (43% Fe) in Canada (Urban, 1992; PorterGeo, 2021). This classification scheme is important in selecting an exploration strategy as it provided the basis of metallogenic models and provides constraints on expected rock types and Fe ore distribution for training machine learning algorithms.

BIFs are already significantly enriched in Fe (averaging 15–35% Fe; Klein, 2005; Gutzmer et al., 2008; Dreher et al., 2021). In order to be economically viable, under current economic and process constraints, BIF deposits require an in-situ average grade of > 60% Fe. For this to occur, additional processes are necessary to enrich the Fe content via the removal of silica (Smith & Beukes, 2016). The Fe enrichment in BIFs is controlled by (a) structural permeability (e.g., long-lived fault systems), (b) hypogene modification due to ascending fluids (magmatic or basinal) and descending meteoric water and (c) supergene enrichment due to

<sup>1</sup> The iron percentages listed above assume no impurities.

intensive weathering and low-temperature fluid circulation (Hagemann et al., 2016 and references therein; Smith & Beukes, 2016). Primary hypogene and supergene ore-enrichment stages can be found in most Fe ore deposits, namely (a) silica leaching and development of magnetite and carbonate minerals, (b) oxidation of magnetite to hematite with the ongoing dissolution of quartz and formation of carbonate minerals and (c) further oxidation, replacement of Fe-silicates by hematite and dissolution of carbonate minerals; and (Hagemann et al., 2016 and references therein). During the stages, none-ore constituents such as Ca–Mn–Mg may be accumulated and may lead to the formation of low-grade and complex Fe ore deposits.

BIF-hosted Fe ore deposits may be enriched by both depositional and post-depositional processes. However, post-depositional processes such as alteration and weathering can obscure the geological characteristics of lithofacies within such deposits. Furthermore, rock identification is commonly performed using traditional geological techniques which often lead to incorrect lithofacies identification where rocks are highly weathered. Machine learning provides a useful method to accurately perform lithofacies discrimination, especially in altered/weathered zones. There are many desirable effects using this approach, namely: (1) a reduction on the reliance of geoscientific subject matter expertise, which may increase discrimination objectivity and/or performance; (2) a transfer of the balance of complexity and interpretability using manual geological interpretation to a balance of data abundance/coverage and performance characteristics using data-driven methods; which is more objective and useful to guide further sampling; and, (3) an ability to perform data-driven science, which can assist with scientific hypothesis testing on a per-deposit and per-dataset basis. In this study, we demonstrate these benefits using geochemical and geological data. We demonstrate that there are statistically correlated relationships between Fe grade and the major/minor (rock-forming) elements, allowing for their use in determining the Fe grade and original lithology. We show that feature importance testing can be used to examine the nature of ore-forming processes and to distinguish between supergene versus magmatic enrichment of Fe. Using Fe ore data from the Assen Fe ore deposit, 50 km northwest of Pretoria in South Africa, we demonstrate the value of the data-driven machine learning approach.

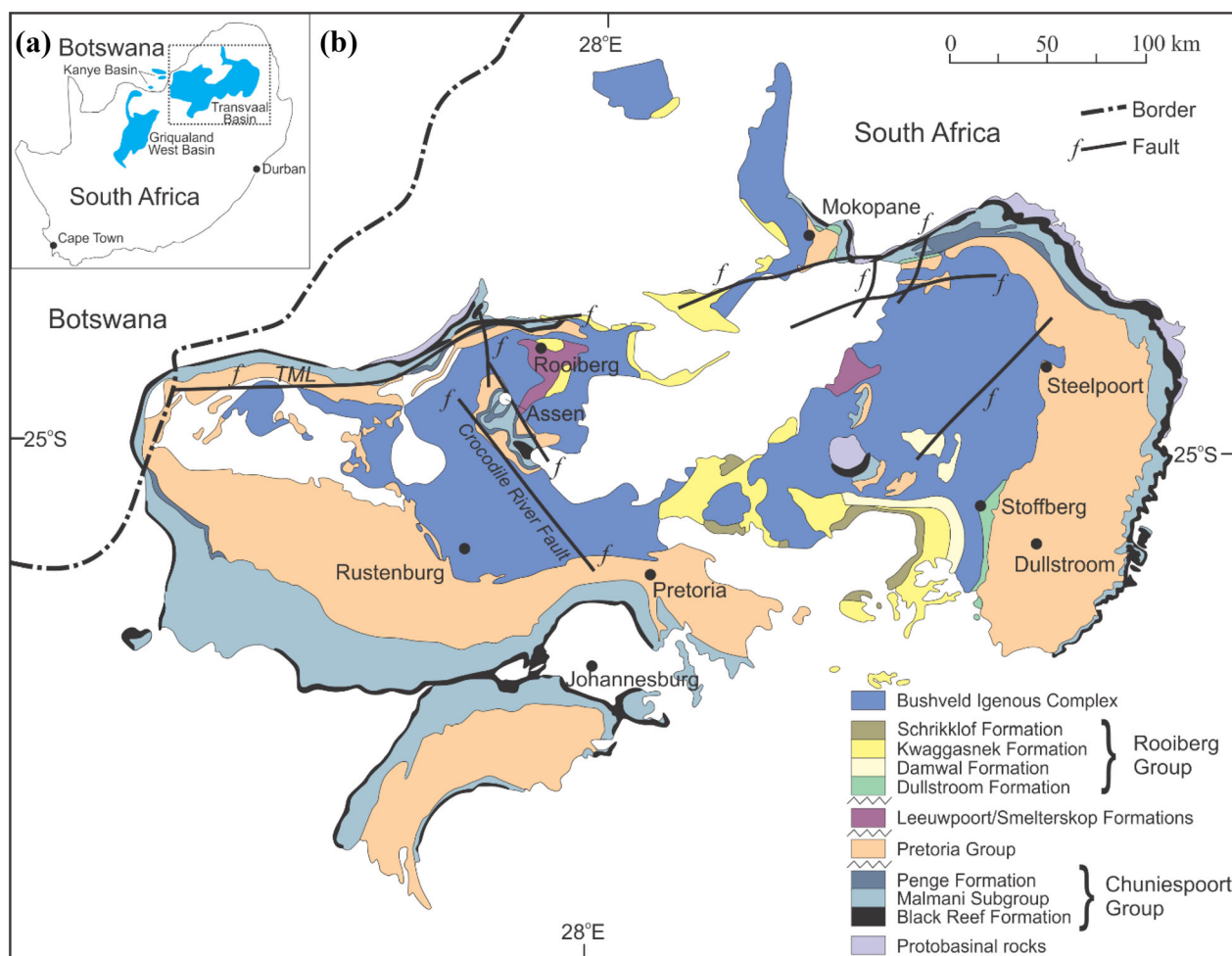
## GEOLOGICAL SETTING

### Regional Geology

On the Kaapvaal Craton in southern Africa, the 2.642–2.055 Ga Transvaal Supergroup is preserved in three basins (Eriksson et al., 1993, 1995, 2006) (Fig. 1a). The Transvaal Basin (to the east) and the Griqualand Basin (to the west) are in South Africa, whereas the Kanye Basin is in southern Botswana (Eriksson et al., 2006). The Transvaal Supergroup (Button, 1986; Eriksson et al., 1993, 2006) is made up of mudrocks, sandstones, dolomites, Fe formations and volcanic rocks that cover the Archean basement (Kaapvaal Craton), the Witwatersrand and Ventersdorp Supergroups (Fig. 1b). The Transvaal Basin was intruded by the Bushveld Igneous Complex at 2.06 Ga (Zeh et al., 2015; Mungall et al., 2016). Previous studies have demonstrated that the intrusion of the Bushveld Igneous Complex may have led to additional mineral tenor in the Transvaal Basin, such as the formation of hydrothermal Au deposits in carbonate rocks (e.g., the Elandshoogte and Pilgrims rest gold deposits; Harley & Charlesworth, 1992; Eriksson et al., 1995).

The stratigraphic sequence of the Transvaal Supergroup, in the Transvaal Basin, can be divided into four subdivisions, namely the (from bottom to top; Fig. 2): (1) Protobasinal rocks (a descriptive term rather than a formal group; approx. 2.66 Ga; Eriksson et al., 2006) characterized by mudrocks, sandstones and basaltic-to-rhyolitic volcanic rocks; (2) the auriferous Black Reef Formation (2.59 Ga; Henry & Master, 2008) characterized by mature quartz arenites with lesser conglomerates and mudrocks; (3) Chuniespoort Group rocks (2.64–2.43 Ga; Eriksson et al., 2006) which consist of dolomites, chert, limestones, shales and BIFs; and (4) Pretoria Group rocks (ca. 2.22 Ga; Burger & Coertze, 1975) comprised of mudrocks alternating with quartzose sandstone, basaltic-andesite lavas and subordinate conglomerates, diamictites and carbonate rocks.

Of interest to this study is the Chuniespoort Group, which attains a maximal thickness of 3,500 m in the Transvaal Basin (Button, 1981). It is subdivided into a number of formations, with the Penge Iron Formation defining the stratigraphic top of the Group (Fig. 2). Underlying the Penge Iron Formation ( $\leq 640$  m thick) is the Malmani Subgroup dolostones. Quartzites and shales of the Rooihooft Formation, Pretoria Group, overly the formation.



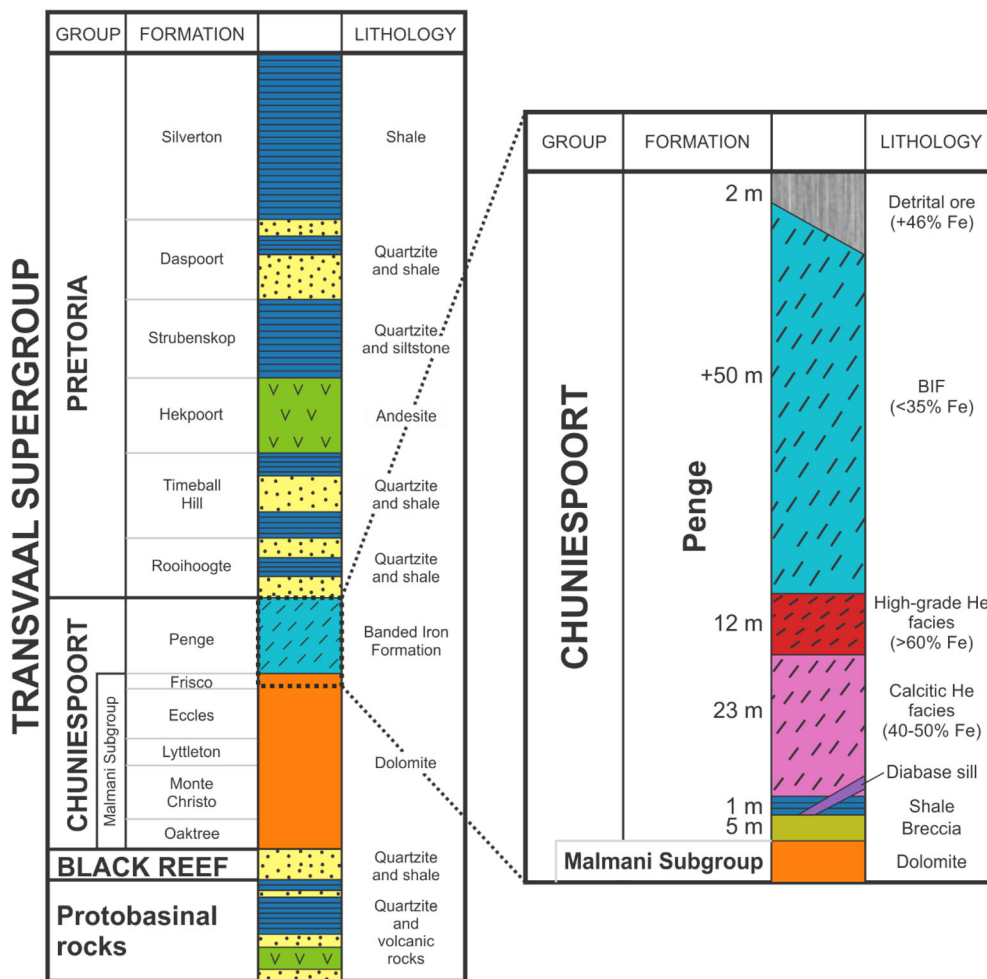
**Figure 1.** Simplified geological maps of the Transvaal Basin, including the Bushveld Igneous Complex, showing the location of the Assen Fe ore deposit. (a) Location of Transvaal Supergroup rocks and their basins within southern Africa. (b) Map of the Transvaal Basin. Figure modified and updated from Eriksson et al. (1995). Coordinates of the Assen Mine: 25° 7' 44.5692'' S, 27° 36' 23.292'' E. TML = Thabazimbi–Murchison Lineament.

The Penge Iron Formation is highly affected by regional contact metamorphism caused by the intrusion of the Bushveld Complex. Within the area of the Assen Fe mine, rocks have been folded, resulting in the formation of domes, as well as being subjected to low- to medium-grade metamorphism (410–510 °C; Hartzer, 1987, 1989, 1995). However, the ultramafic–mafic magmas of the Bushveld Complex (Rustenburg Layered Suite) are not in direct contact with the rocks of the study area, which has been confirmed using geophysical methods (Hartzer, 1989).

The Assen Fe ore deposit is located near the center of the Crocodile River dome (Fig. 1). Domes in the area were created due to the intrusion of the Bushveld Complex into the Pretoria Group sedi-

mentary rocks. The density contrast between the intrusion and overlying sedimentary rocks, combined with the heat and pressure from the intrusion, increased the ductility of the sedimentary sequences such that they buckled (Gerya et al., 2003). The Assen deposit consists of approximately 350 m thick of Fe-rich rocks, which are divided into three facies, namely: (1) calcitic hematite (23 m thick; 40–50% Fe); (2) high-grade hematite (12 m thick; > 60% Fe); and (3) BIF (> 50 m thick; < 35% Fe) facies (Fig. 2). Iron deposits are found within the high-grade hematite and BIF facies. Orebodies, defined as containing > 60% Fe and < 15% SiO<sub>2</sub>, occur as irregular, tabular bodies reaching 80 m in thickness and extending for 12 km along strike. To the east of the Assen deposit is a major ENE-striking syncline





**Figure 2.** Stratigraphic succession of the Transvaal Supergroup. The figure includes the detailed stratigraphic succession of the Penge Formation, which hosts the Assen Fe ore deposit (this study). Stratigraphic columns were modified after Hartzer (1995) and Catuneanu and Eriksson (1999).

that marks the center of the Crocodile River dome. As a result, the stratigraphic sequence at the Assen Fe mine strikes E–W and dips between 35 and 60°N. The dome is bounded by between 3,000 and 4,000 m SE–NW strike-slip faults to the east and west, with the western fault intersecting the westernmost extent of the Assen mine. Lithologies near the western fault are highly brecciated (the fault damage zone measures up to 100 m across where exposed on the surface). A number of minor folds and faults have been recorded in the mine lease area. In addition to folding and faulting, rocks of the Assen deposit have been plastically deformed, partially metamorphosed, metasomatized and recrystallized. Contact metamorphism temperatures reached 410–510 °C (Hartzer, 1987), resulting in the formation of amphibole

(tremolite), talc, calcite, crystalline quartz and dolomite.

### Geology of the Assen Fe Ore Deposit

At the Assen Fe ore mine, the lowermost stratigraphical unit exposed consists of dolomites belonging to the Malmani Subgroup (Figs. 2 and 3). The deposit itself is hosted within the Penge Iron Formation, which overlies the rocks of the Malmani Subgroup. A thin layer of shale (1–2 m thick) lies on top of the dolomites, marking the base of the Penge Iron Formation (Fig. 4). The shale layer is frequently brecciated due to karstification of the underlying dolomites; hence, from here onward, it

will be referred to as “breccia.” Overlying the shale is approximately 350 m of Fe-rich rocks which are divided into three, namely the: 1) calcitic hematite (23 m thick; 40–50% Fe), 2) hematite (12 m thick; > 60% Fe) and 3) BIF (46–152 m thick; < 35% Fe) facies (Figs. 2, 3 and 4).

The calcitic hematite facies is characterized by alternating layers of hematite and calcite (Figs. 3 and 5a, b and c). The facies is further separated into three sub-facies based on mineralogical compositions, physical characteristics and textural appearances. The first sub-facies is characterized by alternating layers of hematite and calcite, with both hematite and calcite bands measuring ~ 1.5 cm in thickness, and is spatially closer to the overlying hematite facies (Fig. 5a). The second sub-facies occurs when both the hematite and calcite bands disappear due to a gradual increase in calcite, approximately 4 m below the contact with the overlying hematite facies (Fig. 5b). The last sub-facies is characterized by hematite which is intercalated with calcite in certain areas (Fig. 5c).

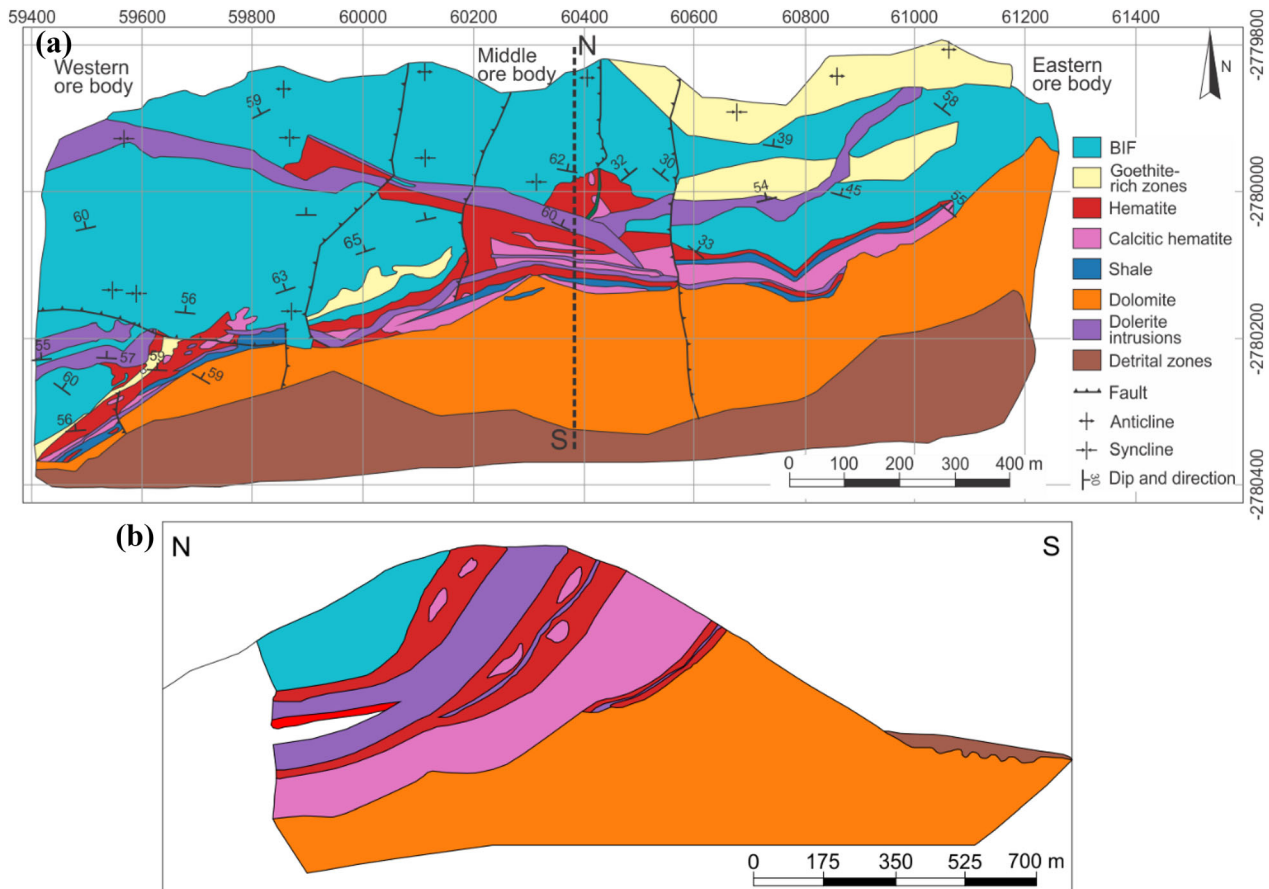
The calcitic hematite facies grade upward to high-grade laminated- or massive-hematite ores, which define the hematite facies (Figs. 2 and 3). Laminated hematite ore is relatively high grade and is identified as fine to medium grained with a reddish-gray color (Fig. 5d and e). Massive hematite ore is dominated by hematite with minor amounts of goethite and micro-quartz, resulting in a relatively unweathered appearance, as compared to all other lithologies at the Assen mine. Hematite crystals in massive hematite ore are fine grained, non-porous and tightly packed, appearing as a lustrous, compact mass (Fig. 5f).

The hematite facies is overlain by the BIF facies (Figs. 2 and 3). The BIF facies is identified by relatively continuous (some minor pinching and swelling) brownish-white chert bands alternating with grayish Fe oxide (hematite) bands (Fig. 5g and h). Bands are composed of quartz, goethite, hematite and smectite. The mineralogical assemblage of the BIF facies changes along strike, with the northern half of the west body being magnetite-dominated with minor weathered goethite, whereas hematite dominates in the central body. Furthermore, there are local goethite-rich zones (Fig. 5i). The modal mineralogy varies with depth, as observed by visible changes in the thickness of the bands. Specifically, the thickness of hematite bands decreases, whereas silica bands increase progressing up stratigraphy. The BIF facies can be divided into three sub-facies.

Sub-facies 1 BIF occupies the upper bounds of the stratigraphy and is extensively silicified, with alternating thin Fe (< 2.5 cm) bands and thick (5 cm) chert bands (Fig. 5g). Contacts between bands are sharp. Sub-facies 2 BIF is characterized by micro-banding, observed as alternating reddish-gray hematite and chert bands, each measuring between 0.5 and 6 mm in thickness. Recrystallized quartz veins and calcite veinlets (0.5–1.5 mm across) with fine-grained greenish-to-yellow amphibole at their center cut through the bands. Sub-facies 2 BIF grades into sub-facies 3 BIF. Sub-facies 3 BIF is characterized by well-preserved primary banding, composed of alternating, 10–50 mm thick, hematite and chert bands, accompanied by minor amounts of amphibole (Fig. 5h).

Dolerite intrusions (< 50 m thick) crosscut all facies at the Assen mine (Figs. 2, 3 and 4). Overall, the intrusions are oriented parallel to the primary bedding. Primary facies are intensely sheared near the contact with dolerite intrusions. The shear lamination is oriented parallel to the contact surface and is accompanied by veins of calcite, epidote–amphibole, as well as local pitting of the hematite, giving the rock a “pig Fe” appearance. The intrusions are green-colored, phaneritic and composed of chlorite, smectite, calcite, kaolinite, anatase and magnetite/hematite crystals. Subsequent hydrothermal alteration chloritized and hematized the dolerite intrusions, especially when the intrusions contact the calcitic hematite and hematite facies.

Metamorphism, deformation, supergene and weathering processes affected all lithologies at the Assen mine. Geological mapping reveals that there are two main anticlines and three synclines within the deposit. These anticlines and synclines folded and shaped the deposit topography in an E–W orientation (Fig. 3a). Specifically, bands in orebodies have an amplitude between 30 and 85 m and a wavelength between 245 and 400 m. Lithologies dip between 35° and 60° to the north (Fig. 3b). Specifically for the BIF facies, metamorphism and deformation resulted in the formation of high-grade hematite and calcite-rich hematite areas, which are mined. Dolerite intrusions into the BIF facies further remobilized chert to anticlines. Mined areas are both folded and faulted (on a macro- and micro-scale) with shearing along strike, which resulted in localized thickening of bands (Fockema, 1948). Both micro-folds and micro-faults, which often crosscut each other, appear to have served as conduits for fluid flow. Weathering of the deposit resulted in a



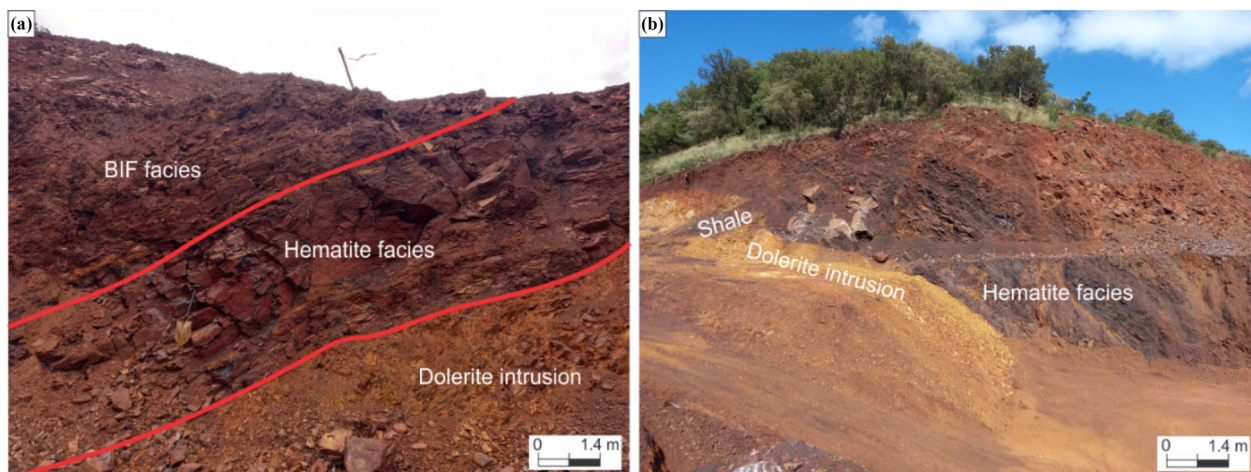
**Figure 3.** Geological map of the Assen Fe ore deposit showing the distribution of lithofacies and geological structures. Both synclines and anticlines modified the dip of the orebody. (a) Overview of the geology of the Assen deposit. (b) Stratigraphic cross section of the Assen deposit.

sporadically developed talus deposit of hematite fragments and boulders, which is currently mined as detrital Fe ore. The talus deposit is developed on the southern slopes of the hill, often obscuring underlying outcrops (Fig. 4).

Iron orebodies are chiefly found within the calcitic hematite, hematite and BIF facies, as well as at contacts between facies and dolerite intrusions (Fig. 3). Orebodies, defined as containing  $> 60\%$  Fe and  $< 15\%$   $\text{SiO}_2$ , occur as irregular, tabular bodies reaching 80 m in thickness and extending for 12 km along strike. The Fe ore is typically finely-banded (5–30 mm) composed of alternating layers of hematite (massive and/or laminated) and white calcite, with practically no chert (Fig. 5). The three most common Fe ore minerals present in the study area are magnetite, hematite and goethite. Magnetite occurs as fine-to-coarse-grained euhedral crystals.

Both hematite and brown–yellow goethite are products of magnetite oxidation, commonly found in altered zones or near the surface of the deposit. Gangue minerals include crystalline quartz (from metamorphosed chert), as well as carbonate (siderite, ankerite, calcite and dolomite), sulfide (pyrite) and other oxide minerals (pyrolusite). Three high Fe grade ore zones can be defined based on their relationship with mapped structural and lithofacies features:

1. Ore zone 1 is located on the hanging and foot-wall of dolerite intrusions,
2. Ore zone 2 is parallel and above the mapped Fe-rich shale layer and is comparatively enriched in silica compared to ore zone 1, and
3. Ore zone 3 appears to be structurally controlled and is associated with SW–NE normal faults.



**Figure 4.** Field photographs of the Assen orebody showing the stratigraphic succession and pronounced weathering. **(a)** Highly foliated BIF facies, underlain by the hematite facies (goethite-rich; 2.5 m thick), intruded by a dolerite sill (at the base). **(b)** N–S cross section of the Assen stratigraphy in the middle orebody. A thin shale layer (1 m) separates dolomite (to the left, buried under the talus) from the hematite facies. Lithologies dip 60° toward the north. A highly weathered dolerite sill (yellow color) intruded below the thin shale layer. Reddish areas are goethite-dominated, whereas dark-brown areas are goethite–hematite-dominated.

## METHODOLOGY

### Sample Acquisition and Data Generation

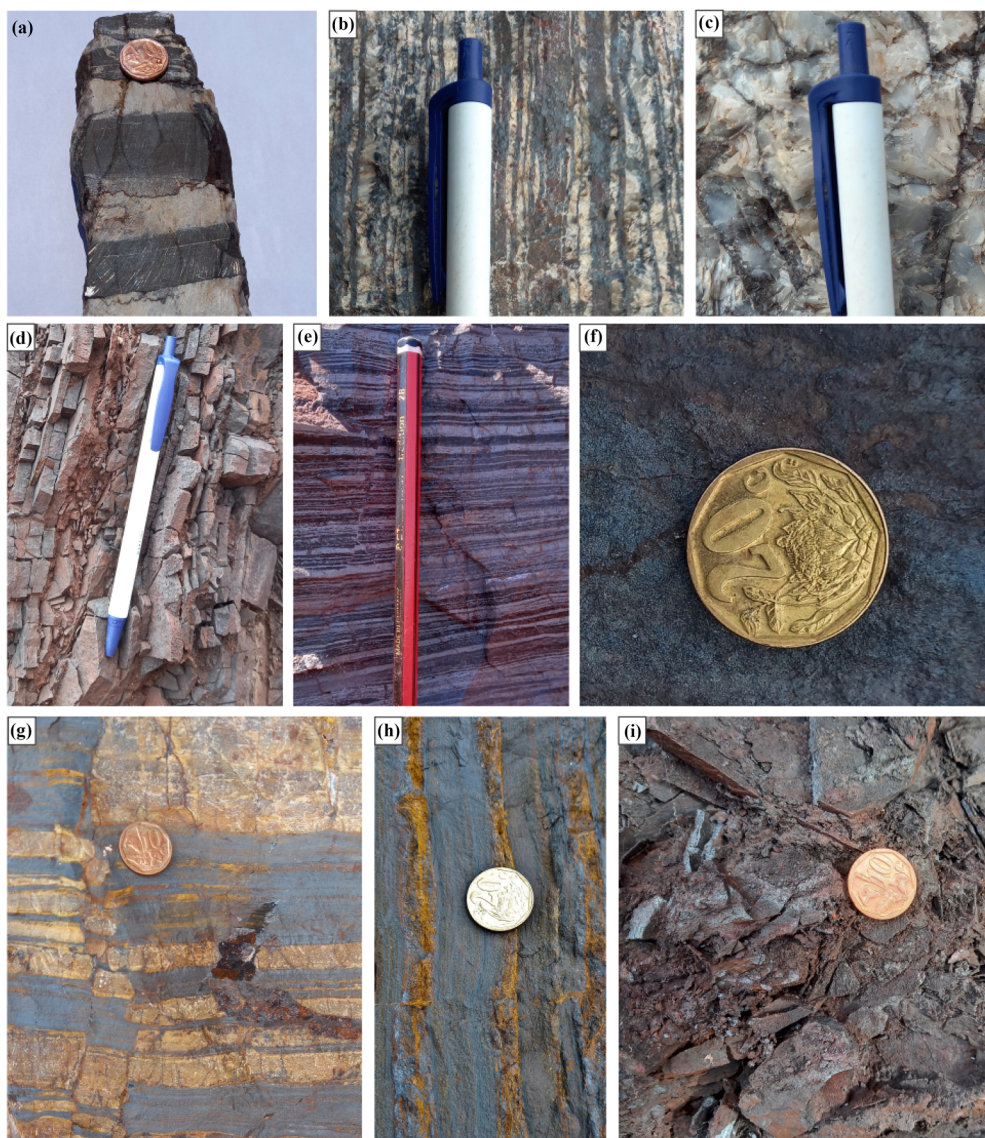
Geological and structural mapping was carried out at the Assen Fe ore mine, which is located on the farms Assen 140 JQ, Assen 161 JQ, Buffelspoort 149 JQ, Doornkloof 141 JQ, Swarthoek 10 JQ in the Brits district (25° 7' 44.5692" S, 27° 36' 23.2920" E) of the North West Province, South Africa (Fig. 1b). Field geological mapping was conducted to map lithofacies geological structures and to identify areas that would require additional drilling due to structural complexity/transported cover. The Assen mine has been mining the detrital Fe ore and stratabound hematite ore located at the base of the Penge Iron Formation since 2013. The terms west, middle and east ore bodies are used by the Assen mine staff, based on their geographical positions along the ridge, although the nature of the boundary between the middle and east bodies has not yet been determined.

Geological field mapping, covering an area of 1.3 km<sup>2</sup>, measured and recorded lithological contacts and geological structures (faults, dykes and sills). Coordinates of stations and outcrops were determined using a Garmin GPS Map64s as well as a Trimble GPS. A total of 24 bulk rock samples (four BIF, six dolerite, four laminated hematite, six massive hematite and four calcitic hematite) were taken for petrographic and mineralogy determination

analyses. Assessment of mineralization, weathering and alteration in different rock types and their field relationship to key features were also mapped.

We repurposed a database that captures a total of 21 diamond drill cores from the Assen mine (Fig. 6). The database was originally used to guide mine operations. For that purpose, to ensure sufficient areal and lithological coverage, the drill cores were drilled in the western, middle and eastern orebodies. The drill cores were logged, orientated and vertically split into axial halves and sampled at one-meter intervals. One-half of the cores were retained in the core tray and were returned to the core yard. The remaining half core was used for analytical and petrographic analyses. Lithochemical analyses (supplementary file) were conducted using X-ray fluorescence (XRF) spectrometry analysis at the South African Bureau of Standards lab (SABS) in Pretoria (South Africa). The samples were analyzed for the following major elements and oxides: SiO<sub>2</sub>, Al<sub>2</sub>O<sub>3</sub>, FeO (total), TiO<sub>2</sub>, CaO, MgO, K<sub>2</sub>O, MnO and P<sub>2</sub>O<sub>5</sub>, all expressed in weight percent (wt%). An ARL Advant XP Sequential XRF (Thermo Fisher) instrument was used to analyze the bulk chemical composition of the samples (Nwaila et al., 2017). The analytical error for most elements is less than 0.10%, with the exception of MgO (0.50%). After heating previously dried particles to 950 °C, the loss on ignition (LOI) was calculated. Total Fe content was verified using volumetric titration (Beyeme-Zogo, 2009). Total sulfur (TS) was analyzed using



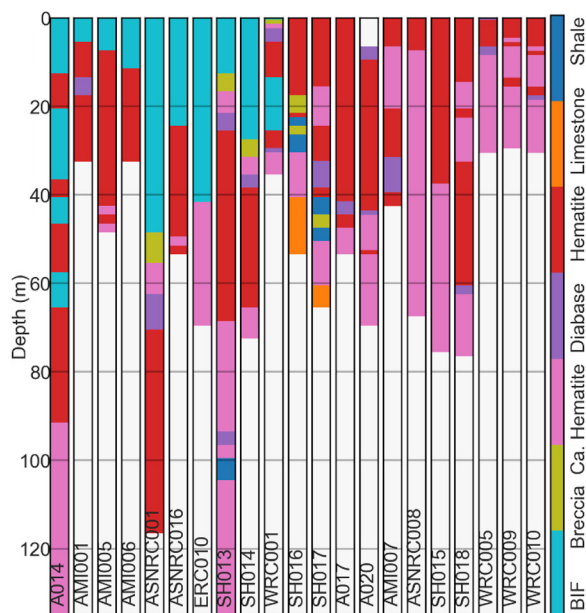


**Figure 5.** Photographs of lithofacies and structures observed in the Assen Fe ore deposit. The calcitic hematite facies is presented in (a), (b) and (c); the hematite facies in (d), (e) and (f); and the BIF facies in (g), (h) and (i). (a) Thick alternating band of hematite (black) and calcite (white). (b) Thinner alternating bands of hematite and calcite. (c) Locally, banding is replaced by massive calcite with inclusions of hematite. (d) Laminated hematite. The hematite bands measure 0.5 mm in thickness and are separated by a thin layer of goethite and chert. (e) Micro-banded laminated hematite. (f) Massive hematite composed of lustrous, bluish-gray, fine-grained hematite. (g) Alternating bands of chert (brown) and hematite (gray), defining BIF sub-facies 1. The bands have been fractured and displaced (left of coin). (h) Micro-bands of hematite (gray) and chert (brown), defining BIF sub-facies 2. The micro-banded BIF sub-facies have a higher Fe and lower SiO<sub>2</sub> content than sub-facies 1. (i) Goethite-rich zone within the BIF facies.

LECO and found to be mostly below the instrument lower detection limit of 0.01%. Despite the original range of oxides analyzed within our dataset as it was obtained, there are frequent missing values of FeO and MgO (about 20% and 15%, respectively). Note that MgO and Na<sub>2</sub>O are largely below the detection

limit at the Assen Fe ore deposit. These missing data entries are commonly uncaptured (even if analyzed) in the mine's geochemical database, as such data were deemed unimportant for the data's original purpose. However, the Fe grade in the form of elemental Fe (wt%) is complete, and it can be shown



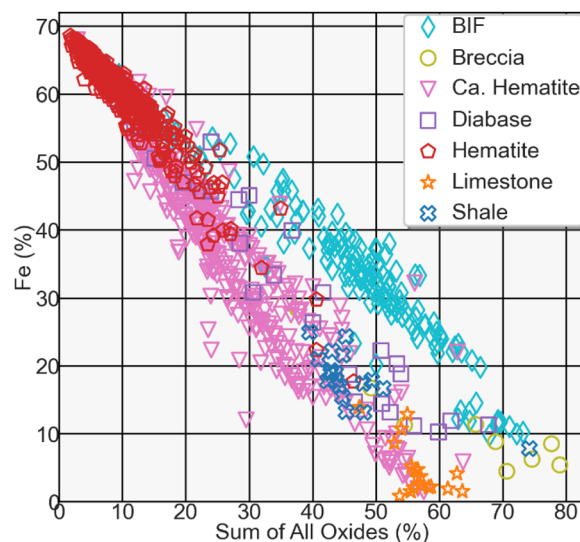


**Figure 6.** Borehole drill cores section profiles are color-coded based on lithofacies. In total, 21 borehole drill cores were used in this study.

that the Fe oxidation state varies by lithology and that distinct trends can be seen (Fig. 7). For our purpose of machine learning-based predictive modeling of lithofacies and understanding the relationships between the Fe grade and the other oxides, the lithochemical information captured by the data were sufficient.

### Multivariate Geochemical Analysis

To examine the geochemical data aside from predictive modeling, we used principal component analysis (PCA), which is a multivariate analysis tool that can be applied to geochemical datasets (Barnett, 2017). Essentially PCA is a re-coordination method that generates new coordinates of the chemical components using linear combinations of existing chemical coordinates such that they best capture the variability of the data cloud. The principal components (PCs) successively decrease in their explanatory power. Scatter plots of PCs can be used to visualize the chemical characteristics of various types of rocks. Although PCA can be a machine learning algorithm if used in an appropriate data-driven context (e.g., dimensionality reduction), in this study, PCA was used strictly as a multivariate geochemical analysis tool to understand multidimensional



**Figure 7.** Fe grade (%) versus the sum of all oxides. Multiple oxidation states exist, and on average, the banded Fe formation (BIF) facies features the lowest oxidation state and the calcitic hematite (Ca. Hematite) facies, limestone and shale are the most common oxidized. The hematite facies tends toward the more oxidized trend but is generally between the two. Diabase is distributed erratically but mainly in the more oxidized trend. Breccia is also irregularly distributed and is mostly in the less oxidized trend.

elemental associations. Other methods such as self-organizing maps (Kohonen, 1982, 2001) have also been used with success (e.g., Iwashita et al., 2011; Cracknell & de Caritat, 2017). However, in this study, the purpose of multivariate geochemical analysis was mainly to provide an in-discipline basis for the understanding of machine learning results and, therefore, PCA is an appropriate choice that is highly explainable and intuitive.

### Machine Learning-based Predictive Modeling

The lithochemistry of samples (rock-forming elements) were used as machine learning features in this study. Lithochemistry provides a wealth of information about surface and internal geological processes such as elemental distribution, lithological variability, hydrothermal overprinting and tectonic setting. (Nwaila et al., 2017). Lithochemistry in this study served two purposes—to predict the type of rock (lithofacies) in the Assen Fe ore deposit and to predict the Fe grade (%). For the Fe grade prediction, the primary intent is to generate insights into the relationships between Fe and

other chemicals, which facilitates scientific hypothesis testing.

Lithochemical data were used as machine learning features. Feature engineering is a technique for manipulating features in order to improve a measurable algorithmic outcome (Hastie et al., 2009; Domingos, 2012). Feature engineering, such as log-ratio-transformations, can change the shape of the feature space (Hastie et al., 2009), removing the finite range restriction of compositional data (Aitchison, 1982). Many machine learning techniques do not require certain feature space geometries or distributions of feature variables (e.g., normality). As a result, range restriction is not always an issue (especially with feature rescaling). Where there is a discrepancy between the assumed feature space geometry and the geometry of the data embedding, spatial distortions could be introduced, affecting the algorithms' performance. Geometry (and its associated properties, such as transformations and metrics) can be chosen heuristically based on the data structure and historical practice, but it must be validated through algorithm performance profiling. Gu et al. (2019) illustrate how manipulating the geometry of feature space can improve algorithm performance. Compositional data exhibits Aitchison geometry (Aitchison, 1982), which is different from the Euclidean geometry assumed by many statistical techniques (Aitchison, 1982; Grunsky et al., 2014; Harris et al., 2015; Chen et al., 2018; Grunsky & de Caritat, 2019). The traditional practice is to employ log-ratio transformations on the data to minimize the impacts of the Aitchison geometry on statistical techniques, such as multivariate data analysis. Machine learning algorithms that are unaware of the feature space geometry (e.g., they do not use spatial distance metrics or assume linearity properties of the space), it is unclear whether any data transformation would be heuristically useful. Spatially aware algorithms may be benefitted from a more Euclidean-like geometry in the feature space. The centered log-ratio (clr) transformation is a popular choice in traditional multivariate compositional data analysis (Aitchison, 1982; Grunsky et al., 2014; Harris et al., 2015; Chen et al., 2018; Grunsky & de Caritat, 2019), which takes a logarithm of the ratio of all sample compositions by their geometric mean. In previous studies (e.g., Zhang et al., 2021, 2022), clr-transformed and raw data were empirically demonstrated to produce a similar performance for classification and regression tasks (over a range of algorithms) that are similar to those used in this study. However,

since our data were frequently missing FeO and MgO (and other elements to lesser extents, such as S), it is unclear whether the clr transformation would be effective if Fe were substituted for FeO. To explore the feasibility of this type of data pre-processing for our tasks and our data, we also employed the clr transformation for one workflow, with the other parallel workflow being identical in all other regards with the exception of the use of raw data. The clr-transformed and raw datasets in both workflows were subjected to further feature rescaling to ensure that all features can contribute equally to predictive modeling.

Predictive modeling with geochemical data can be done using both supervised and unsupervised machine learning methods. Data are unlabeled in unsupervised learning, which means class labels (such as concentrations or rock types) are unavailable, and the machine attempts to deduce categorizations within the data to develop a classification scheme, which is then used to categorize fresh data (Hastie et al., 2009; Zhang et al., 2021, 2022). The data were labeled (for example, an elemental concentration is recorded) and utilized to tune the algorithm's hyperparameters using training and cross-validation datasets in supervised learning (Hastie et al., 2009). The tuned models are used to regress continuous (e.g., an element's concentration) or discrete (e.g., rock kinds) labels (Zhang et al., 2021, 2022). Data that do not overlap with the training and cross-validation datasets can be used to evaluate model performance. Classification algorithms are appropriate for predicting rock type, while regression techniques are appropriate for predicting Fe contents.

The extraction of links between features and class labels is automated using supervised regression and classification algorithms (Russell & Norvig, 2010). Indicatively, trained models are used to predict labels for a population. Performance characteristics, computational complexity, data density, including feature space density, bias-variance trade-off, model complexity, feature space dimensionality, input and prediction noise, and feature interactions are all factors to consider when choosing an algorithm. In many circumstances, cross-validation is utilized to identify an optimal method using an experimental technique. The root-mean-square sum of bias, variance, and noise can be used to model prediction error in the bias-variance decomposition. The model's inclination to default to a particular class label is referred to as bias. Variance describes

how a model's output changes in response to changes in its input. The irreducible component of the prediction error is noise. Given that algorithms often exhibit diverse bias-variance behaviors, the selection of a suitable algorithm for a particular job is sometimes meant to intentionally raise bias for a correspondingly bigger reduction in variance, which promotes model generalizability (Zhang et al., 2021, 2022).

We used several performance metrics for regression tasks, including the coefficient of determination (CoD or  $R^2$ ) and the median absolute error (MedAE) for performance evaluation. MedAE is the median value of the absolute value of the prediction residual (predicted minus actual), and it has the same units as the predicted quantity. MedAE is more robust to outliers and is a measure of the absolute amount of deviation expected in a typical prediction. To cross-compare prediction quality across lithofacies with different chemical concentrations and to facilitate performance interpretation, we also quantified the prediction errors using the relative error metric (RE, absolute value of the predicted minus the actual, divided by the actual). This metric is intended to facilitate cross-comparisons, as it is non-dimensionalized, and its value can be understood as the fraction of the predicted result that is likely to be incorrect. For classification tasks, we used confusion matrices, accuracy and the F1 score (Fawcett, 2006). An F1 score can be explained as a harmonic mean of precision and recall, where the best score is 1, and the worst score is 0.

([https://scikit-learn.org/stable/modules/generate\\_d/sklearn.metrics.f1\\_score.html](https://scikit-learn.org/stable/modules/generate_d/sklearn.metrics.f1_score.html)).

In this study, we employed the following algorithms:  $k$ -nearest neighbors (Tikhonov, 1943; kNN, Fix & Hodges, 1951; Cover & Hart, 1967; Elastic-Net (for regression), Santosa & William, 1986; Tibshirani, 1996; Witten & Frank, 2005; Zou & Hastie, 2005 Gaussian Process, Rasmussen & Williams, 2006; Kotsiantis, 2007); support vector machines (Vapnik, 1998; Hsu & Lin, 2002; Karatzoglou et al., 2006); tree-based algorithms such as random forest and adaptive boosting or AdaBoost (Ho, 1995; Breiman, 1996a, 1996b; Freund & Schapire, 1997; Breiman, 2001a; Kotsiantis, 2014; Sagi & Rokach, 2018); logistic regression (for classification), Cramer, 2002); naïve Bayes (for classification, Rennie et al., 2003; Hastie et al., 2009); and artificial neural network (ANN, Curry, 1944; Rosenblatt, 1961; Rumelhart et al., 1986; Hastie et al., 2009; Lemaré-

chal, 2012). For details of these algorithms aside from Gaussian Process, their functionality and parameters, as well as for two application examples that are similar to those of this study, see Zhang et al. (2021, 2022). For Gaussian Process, details can be found in Rasmussen and Williams (2006).

Except for one drill core used to simulate deployment testing, we used the same lithochemical dataset for all predictive modeling tasks. For this task, it is important that the numerical range of the training data are at least as broad as those would be present in the deployment. This is particularly important for decision tree-based methods, such as random forest, where due to the nature of averaging an ensemble of trees and of leaves within each tree, the results do not extrapolate outside of trained numerical domains. As such, this limits the prediction to within the range covered by the training data and therefore extrapolation is not well suited to such algorithms. Where this capability is required, other methods may be more useful, such as Elastic-Net, with the usual caveats regarding the uncertainty of extrapolation. Cross-validation is commonly used for algorithm selection and tuning in supervised machine learning algorithms (an out-of-sample testing technique). The dataset was divided into several non-overlapping sets for cross-validation, the largest of which was the training dataset, and it was used to train the models. The remaining validation dataset was then used to profile the models' performance and adjust model hyperparameters. This process reduces issues such as excessive model variance and selection bias. The data partitioning scheme chosen was determined by data properties and computation complexity requirements (Zhang et al., 1999; An et al., 2007). In order to determine the best algorithms for both regression and classification tasks, we used a grid search (Table 1) combined with fivefold cross-validation. Subsequently, we used an exhaustive fivefold cross-validation over the entire training and testing datasets (not including the reserved deployment testing dataset) to profile model performance, and in the case of lithofacies classification, we also used the reserved drill core data to perform deployment testing.

Feature importance testing is a process that examines the predictive power of various features, which in this case are the chemical components. Permutation of covariates can be used to determine the importance of features via the random forest algorithm (Breiman, 2001a, b; Strobl et al., 2008; Altmann et al., 2010; Gregorutti et al., 2015; Zhu et al., 2015;

Datta et al., 2016; Gregorutti et al., 2017). This process serves several purposes depending on the context. For predictive modeling, feature importance testing often permits the removal of irrelevant or less useful features. This reduces the feature space dimensionality, which results in increases in data density and, therefore, performance. However, for our purpose, feature importance testing served strictly as a tool to diagnose the strength of chemical associations between Fe and other chemical components, in particular, taking advantage of random forest's ability to leverage non-linear feature interactions. This is an approach that is different from standard multivariate analysis, such as PCA and correlation matrices. The key differences lie in the use of highly flexible models in feature importance testing to describe the high-dimensional multivariate elemental relationships, and the metrics that are used to evaluate the strengths of the relationships are indicative of predictive power. Compared to PCA or correlation matrices, both of which assume linearity of relationships and convexity of the data cloud, this approach is more powerful and general. However, the models generated by predictive modeling may be difficult to interpret (e.g., random forest models or artificial neural network models) compared to simpler parametric and often linear models (e.g., PCA coordinates). Nevertheless, this process allowed us to gain insight into the feasibility of various formation hypotheses. For instance, any formation hypothesis that requires a particular relationship between a feature and Fe would be more probable for the Assen Fe ore deposit if the feature is demonstrably important. On the contrary, if important features or their interactions are completely irrelevant for a hypothesis, then the hypothesis itself may be irrelevant, incomplete and/or inadequate. However, as predictive modeling is inferential, results and important studies are not direct proof or falsification of any particular hypothesis, similar to traditional multivariate analyses, but rather serve as data-driven evidence. In our application, we used the random forest classification and regression algorithms with optimized hyperparameters to perform feature importance testing.

## RESULTS AND DISCUSSION

### Core Logging and Lithochemical Characterization

The Assen Fe ore deposit showed a variable chemical profile in all the studied drill cores. Sedi-

mentary provenance resulted in the initial concentration of Fe in the BIF facies, with values ranging between 15 and 30% Fe. Secondary supergene enrichment led to the formation of the deposit by upgrading in-situ Fe concentration up to 60% Fe. In lithochemical profiles and field outcrops, we observed localized alteration due to dolerite intrusions, specifically near or at their contact with other lithologies. Where the dolerite occurs in proximity to Fe-rich horizons such as the BIF, hematite and calcitic hematite facies, the rocks are enriched in Fe. The Fe content was measured as being significantly higher (58 and 64% Fe) when compared to surrounding rocks (42% Fe) (Fig. 8). Although dolerite has been subjected to both severe weathering and erosion, preserved dolerite outcrops help in the determination of its broader impact on Fe mineralization. In Figure 8, the lithochemical profile shows relatively high SiO<sub>2</sub>, Al<sub>2</sub>O<sub>3</sub>, TiO<sub>2</sub>, CaO and P<sub>2</sub>O<sub>5</sub> contents and low Fe concentration in the BIF facies. This systematically reverses to high Fe concentration and low SiO<sub>2</sub>, Al<sub>2</sub>O<sub>3</sub>, TiO<sub>2</sub>, CaO and P<sub>2</sub>O<sub>5</sub> content in the hematite facies.

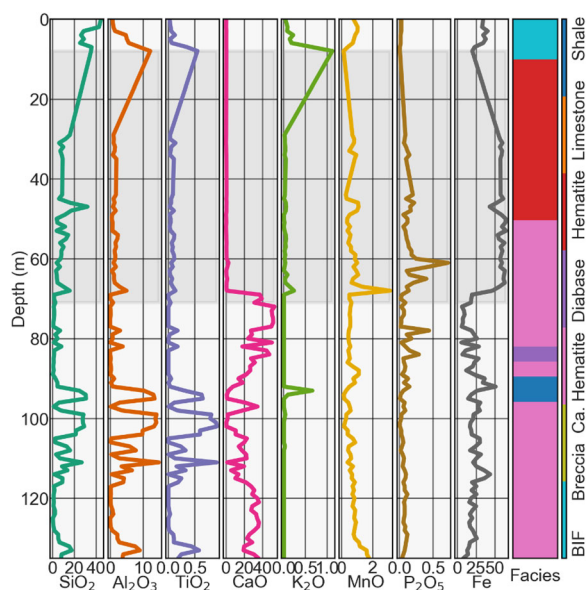
Relationships between major oxides per lithofacies are multidimensional and mostly nonlinear (Fig. 9). Lithofacies of the Assen deposit demonstrate a highly variable geochemical composition, e.g., there was a positive correlation between SiO<sub>2</sub> and Fe in the BIF facies and a negative correlation of the same elements in the calcic hematite facies due to secondary replacement of certain cations by Ca<sup>2+</sup>. A nonlinear correlation was observed between Al<sub>2</sub>O<sub>3</sub> and CaO. This demonstrates that multiple and potentially complex relationships occur between the different major oxides present in the lithofacies.

A PCA biplot shows clusters of samples based on their similarity and is a method to optimally visualize multidimensional data variability. Our data can be depicted using this method (Fig. 10). In PCA space, the variance along PC1 is 46% of the total, and that along PC2 is 22%, and therefore, these two PCs capture a majority of the variability described by our dataset. PC1 is associated negatively with CaO and positively associated with K<sub>2</sub>O, SiO<sub>2</sub>, Al<sub>2</sub>O<sub>3</sub> and TiO<sub>2</sub>. PC2 is associated positively mainly with Fe and associated negatively with CaO. P<sub>2</sub>O<sub>5</sub>, MnO and Fe are correlated positively. The separation of lithofacies by Fe is somewhat clear in the PC1, and PC2 coordinates, similar to that observed in the binary scatter plots (Fig. 9). However, the separation of hematite and calcitic hematite facies is quite pronounced along with CaO, with minor



**Table 1.** Model parameters used in the grid search

Algorithm	Parameter grid
kNN	$k = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$
SVM	$C = \{10, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 750, 1000\}$ , $\epsilon = \{0.00001, 0.0001, 0.001, 0.01, 0.1, 0.5, 1.0\}$ , kernel = {linear, RBF}
Elastic-Net	$\rho = \{0.1, 0.5, 0.7, 0.9, 0.95, 0.99, 1.0\}$
Gaussian process	$\alpha = \{1e-06, 1e-05, 0.0001, 0.001, 0.01, 0.1, 1.0, 10.0\}$ , kernel = {RBF(L) for L from $10^{-3}$ to $10^2$ in logarithmic intervals, DotProduct( $\sigma_0$ ) for $\sigma_0$ from $10^{-3}$ to $10^2$ in logarithmic intervals}
Random forest	Ensemble size = 500, 1000, 1500; maximum depth = {7, 6, 5, 4, 3, 2, 1, unlimited}, maximum number of features = {1, 2, 3, 4, 5, 6}, minimum number of samples for a split = {2, 3, 4}, minimum number of samples for a leaf = {1, 2, 3, 4, 5}, criterion (for classification) = {Gini, entropy}
AdaBoost	Learning rate = 1, number of classifiers = {250, 500, 1000}, base algorithm = decision tree with the same parameter grid as the random forest algorithm
ANN	$\alpha = \{0.001, 0.01, 0.1, 1.0\}$ , activation = {identity, logistic, tanh, relu}, learning rate = {constant, inverse scaling, adaptive}, max iterations = {500, 1000, 1500, 2000}, number of hidden layer neurons = {10, 20, 50, 100}
Naïve Bayes	$\alpha = \{10^{-1}, 10^{-2}, \dots, 10^{-11}\}$
Logistic regression	$C = \{0.1, 0.2, \dots, 19.9\}$ , penalty = {L <sub>1</sub> , L <sub>2</sub> , mix of L <sub>1</sub> and L <sub>2</sub> , none}, class weight = {balanced, none}



**Figure 8.** Lithochemical profile of major elements distribution in the Assen Fe ore deposit. The approximate location of the main ore zone is shaded. The calcitic hematite facies is abbreviated as Ca. Hematite.

overlap (Fig. 10). Furthermore, transitioning from calcitic hematite to hematite and finally to BIF is associated with a change in chemical association from  $P_2O_5$  to MnO to Fe (Fig. 10).

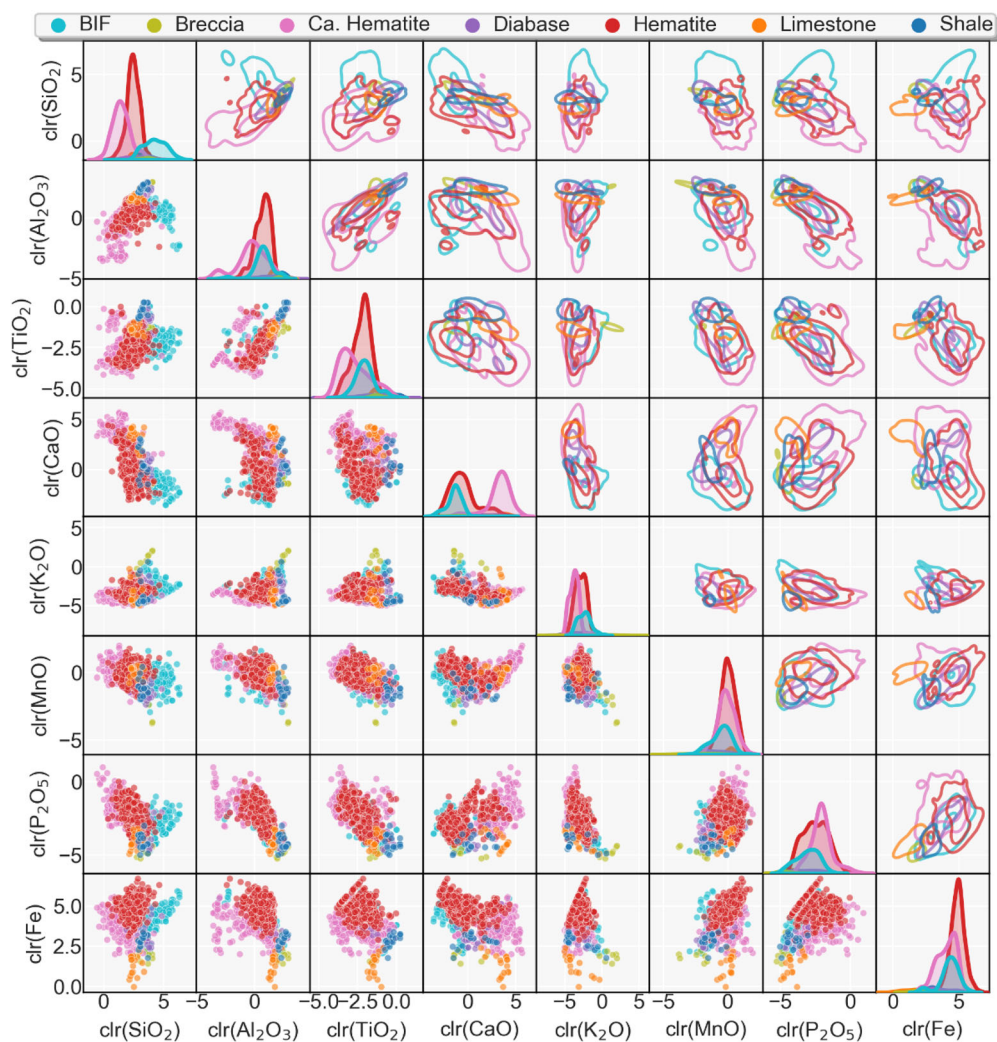
While the hematite facies did not show a dominant direction of variability (the cluster appears relatively isotropic and convex), the calcitic hematite facies showed a clear dominant direction of variability that is very well aligned with  $P_2O_5$ ,  $TiO_2$

and  $Al_2O_3$ . The overlap between non-hematite lithofacies and other lithofacies is generally visible, except for a great extent of overlap between the BIF and hematite facies, and diabase and other lithofacies, although the BIF facies tends to contain more  $K_2O$ ,  $SiO_2$  and less Fe. Diabase appeared to be uncharacterizable as distinct lithofacies as it overlapped greatly with the calcitic hematite and hematite facies and to a lesser extent, the BIF facies (Fig. 10). This degree of overlap may present challenges in the use of geochemistry to distinguish diabase and potentially the BIF facies from other lithofacies, especially given the amount of data for diabase.

### Machine Learning-based Facies Classification

Facies classification was performed using the major oxides and Fe. The results averaged over 100 runs suggest that, in general, facies can be predicted reliably using both raw and clr-transformed data. However, raw data generally produced better results for most algorithms (Fig. 11). The F1 score is noticeably higher for all algorithms except naïve Bayes, and the standard deviation of scores is comparable for a few algorithms (AdaBoost, logistic regression and SVM) but is substantially improved for the ANN algorithm using raw data. The accuracy metric resulted in qualitatively identical observations (not shown). The remainder of the results is presented for the workflow that employed raw data,





**Figure 9.** Scatter matrix of different oxides in clr-coordinates based on lithofacies. The upper diagonal is a kernel density estimation of the data. The calcitic hematite facies is abbreviated as Ca. Hematite.

as that workflow exhibited generally higher performance.

For all algorithms with the exception of naïve Bayes and Gaussian process, the F1 score (the accuracy score was slightly higher but similar in behavior) was over 0.9 in model training and close to 0.9 during model testing (Fig. 12). The confusion matrices for the highest performing algorithms (random forest and SVM) are similar (Fig. 13). Diabase seems to be difficult to predict and, to a lesser extent, breccia (Fig. 13). Due to the extent of overlap between BIF and hematite, there was a significant proportion of predictions of BIF that were incorrectly classified as hematite (Fig. 14).

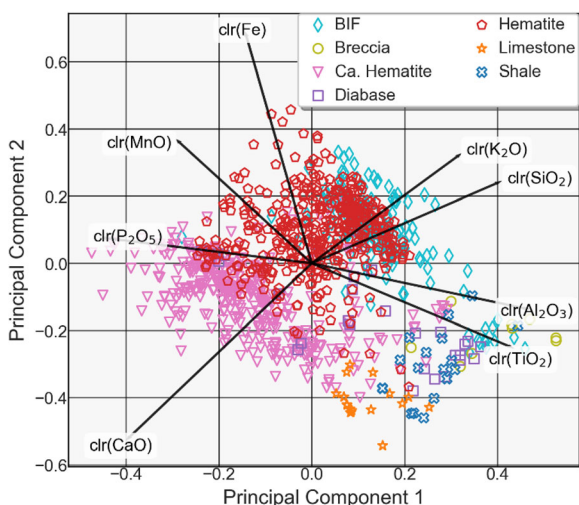
To further test the generalizability and accuracy of our predictive models, we deployed the trained models on a simulated deployment drill hole. The dataset for the deployment drill hole was not used for any previous training or testing (but otherwise is of a similar origin and structure) and it contained the same lithofacies as those covered by the training and testing datasets. Furthermore, the data for the deployment drill hole for all features and targets fell within the ranges covered by the training data and therefore, no extrapolation was required. The ranking of classifier performance is shown in Table 2. The best performing algorithm was AdaBoost, which predicted the facies perfectly. With the

exception of naïve Bayes, all other classifiers performed within expectation based on testing results. It is worth noting that as per field observations, some lithofacies' boundaries were not discrete but transition gradually into one another. Prediction performance was limited by the extent of chemical overlap between some lithofacies. However, the performance observed during the training and testing phase of predictive modeling was likely more indicative of the deployment performance than any single deployment borehole application because the

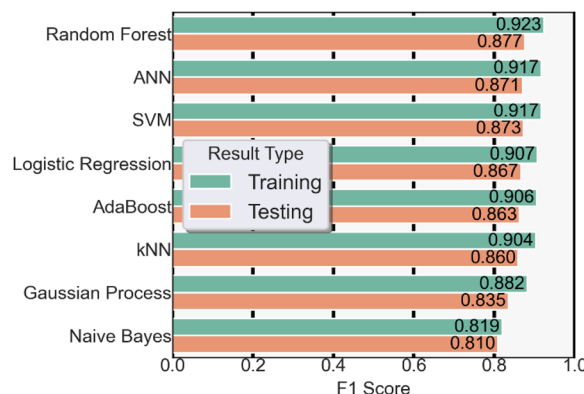
data in the former case were more abundant and the cross-validation testing was more statistically robust.

**Prediction of In-situ Fe grade**

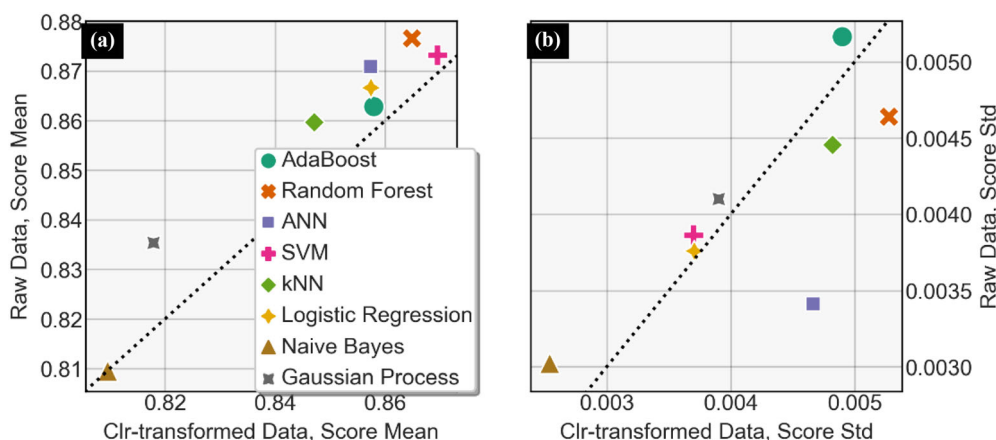
The prediction performance of Fe grade, as averaged over 100 runs was consistently better using raw data for most of the algorithms across various metrics (Figs. 15 and 16). With the exception of SVM, Gaussian process and ANN, whose performance as measured by the CoD metric seemed to slightly improve with the use of clr-transformed data, the performance differentials were heavily in favor of the workflow using raw data. Subsequent results presented are those obtained using the raw



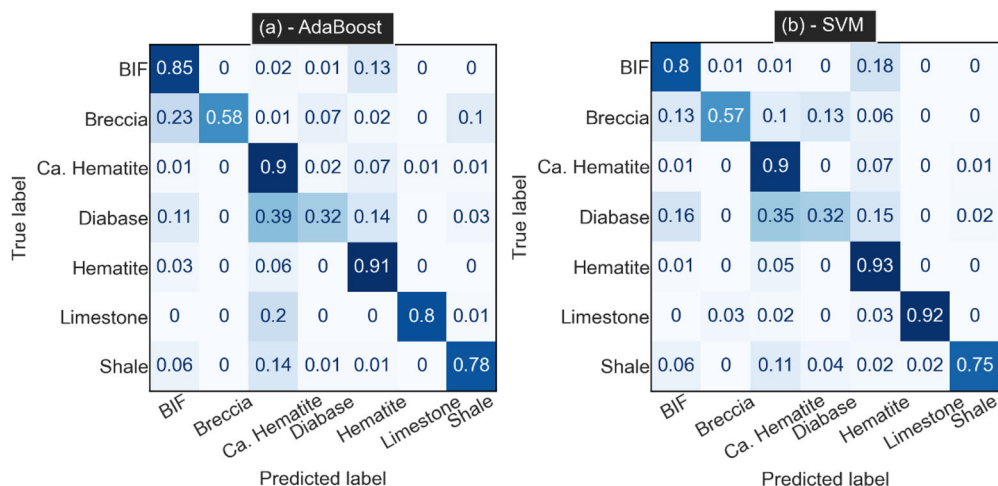
**Figure 10.** Principal component biplot of Fe and oxides. Component loadings are shown as additional lines—each line for each chemical component points toward increasing values of that variable. The calcitic hematite facies is abbreviated as Ca. Hematite.



**Figure 12.** F1 scores for various algorithms used for facies classification. Training and testing scores are both shown.



**Figure 11.** F1 scores of various algorithms using both raw and clr-transformed data. The mean and standard deviation (std) are shown in (a) and (b), respectively. The dotted line is the 1:1 line. Data points on the line indicate no change in algorithm behavior as measured through the F1 score metric.



**Figure 13.** Normalized confusion matrices for the (a) AdaBoost and (b) support vector machine (SVM) algorithms used for facies classification.

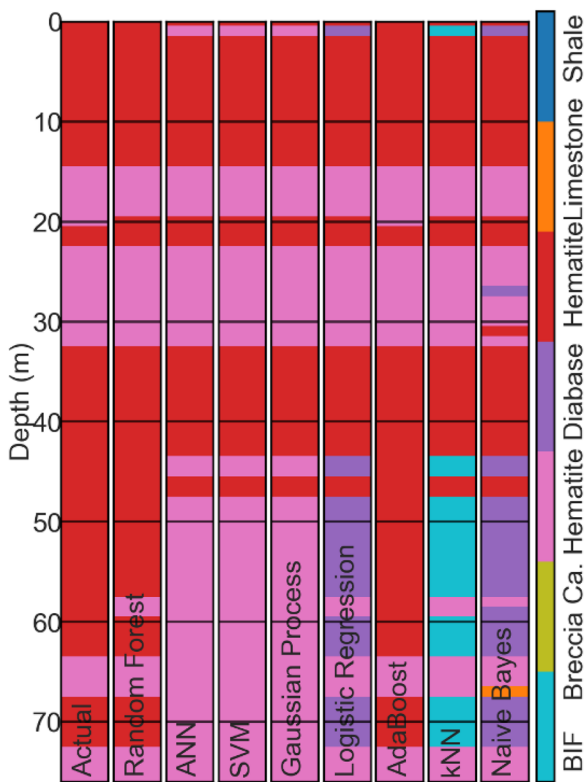
**Table 2.** Model performance based on optimum hyperparameters for lithofacies classification

Algorithm	Parameters	Training accuracy	Training F1 score	Testing accuracy	Testing F1 score	Borehole test accuracy	Borehole F1 score
SVM	$C = 300$ , $\gamma = 0.5$ , kernel = RBF	0.915	0.917	0.878	0.873	0.929	0.929
Gaussian process	kernel = RBF(0.268), multi-class method = one versus one	0.887	0.882	0.853	0.835	0.929	0.929
Random forest	Criterion = entropy, max depth = 7, max features = 5, min samples leaf = 2, min samples split = 4, number of estimators = 1000	0.925	0.923	0.882	0.877	0.946	0.947
ANN	Activation = relu, $\alpha = 0.001$ , learning rate = adaptive, max iteration = 1000, hidden layer size = 100	0.915	0.917	0.877	0.871	0.929	0.929
AdaBoost	Number of estimators = 1000, max features = 5, min samples = 2, min samples split = 4, criterion = Gini, min samples split = 4	0.906	0.906	0.866	0.863	1.000	1.000
kNN	$k = 7$	0.906	0.904	0.866	0.860	0.929	0.937
Logistic regression	$C = 0.1$ , class weight = none, penalty = none	0.906	0.907	0.871	0.867	0.929	0.937
Naive Bayes	$\alpha = 0.001$	0.811	0.819	0.807	0.810	0.875	0.907

data because the performance was generally better across a variety of metrics (MedAE exhibited the same qualitative behavior as the CoD metric, not shown).

The regression results for in-situ prediction of Fe (in %) are shown in Figure 17, and the performance summary is given in Table 3. The SVM algorithm performed better relative to the others across all metrics (Fig. 17a, Table 3). For all algo-

rithms, the prediction errors were heteroscedastic and in a manner that seemed non-intuitive, such that the prediction error was the lowest at the highest concentrations, which were strongly dominated by the hematite facies (Fig. 17). The results of the ANN algorithm were the least heteroscedastic. This implies that the chemical relationships were more predictive of the Fe grade in the hematite facies relative to those in other lithofacies.

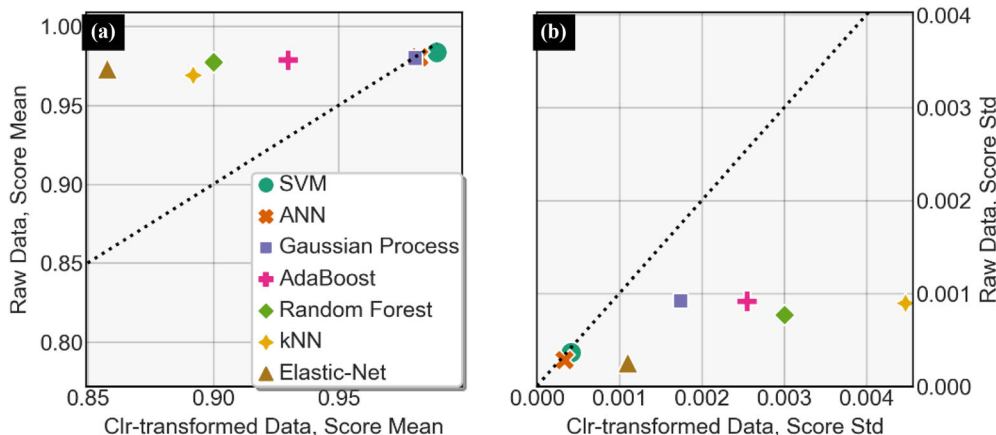


**Figure 14.** Automated lithofacies classification in a borehole using major elements geochemical data and machine learning algorithms. The actual is shown first to the left, and following to the right are the predicted facies per algorithm. Abbreviations: kNN = k-nearest neighbors; SVM = support vector machine; and ANN = artificial neural network.

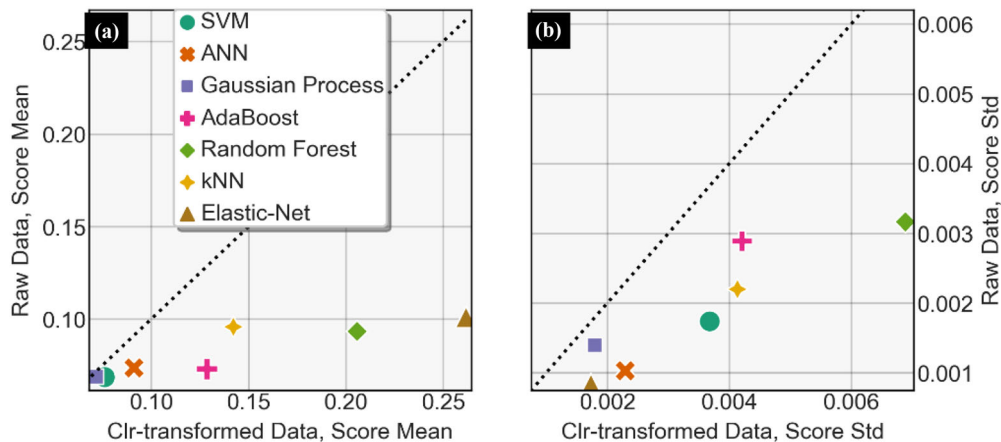
### Feature Importance and the Effect of Facies Distribution on Performance

The results of feature importance testing for classification and regression shows that for both tasks, the importance ranking is largely similar (Fig. 18a and b). For the purpose of Fe grade prediction, the two most important features for the random forest algorithm were CaO and SiO<sub>2</sub>, and to a lesser extent, Al<sub>2</sub>O<sub>3</sub> (Fig. 18b). This was an interesting observation, which indicated that the most relevant relationships between Fe and other oxides in high-dimensional chemical space were those with CaO and SiO<sub>2</sub>. For lithofacies classification, the most important features were CaO, SiO<sub>2</sub> and Fe (Fig. 18a). It might seem unintuitive that Fe was a less relevant feature for lithofacies classification compared with CaO. However, the overlap of Fe with various lithofacies in our dataset was indeed significant, and the separation was the best along with CaO (Figs. 9 and 10). This observation, along with the significant overlap of Fe across lithofacies, implied that the changes in Fe concentration across the lithofacies can be gradual, and Fe enrichment is unlikely to be stratabound for some lithofacies.

In addition to feature importance testing, the predictive performance is also an indicator of the strength of high-dimensional relationships between various oxides and Fe. From the prediction versus actual graphs (Fig. 17), we observed that the various



**Figure 15.** Performance comparison of the prediction of Fe grades using the CoD metric. The score means (a) and standard deviations (std; b) of various algorithms are shown using the raw data versus the clr-transformed data. The dotted line is the 1:1 line.

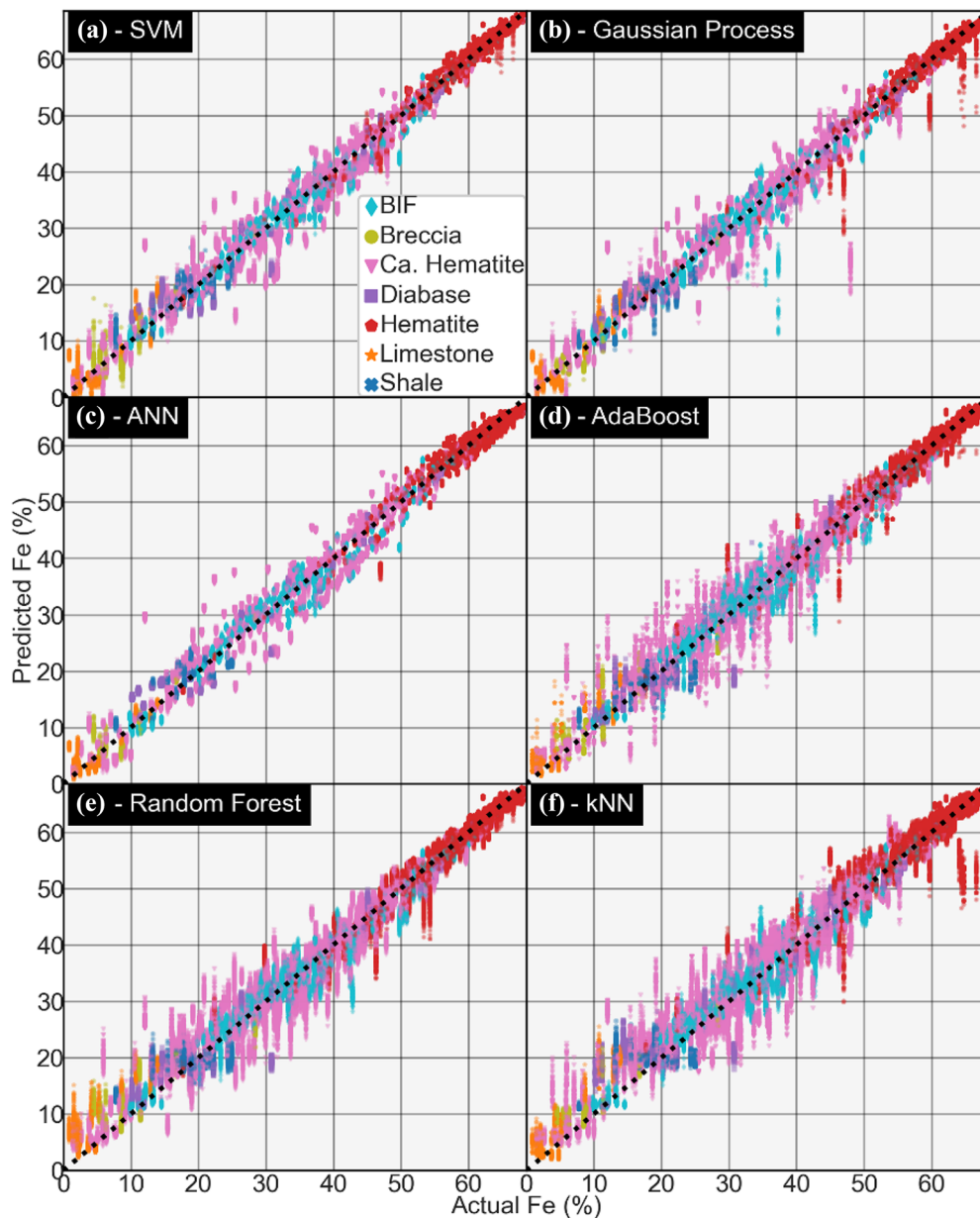


**Figure 16.** Performance comparison using the relative error metric. The score means (a) and standard deviations (std; b) of various algorithms are shown using the raw data versus the clr-transformed data. The dotted line is the 1:1 line.

lithologies were predicted within a range of accuracy. The histogram and density estimation of prediction residuals for all lithologies demonstrate that the hematite facies was predicted the most accurately, followed distantly by the BIF and calcitic hematite facies (Fig. 19). This pattern was observed consistently across all algorithms employed and, therefore, was not an artifact of the choice of algorithm. Model performance was related to the predictive power of various features, which may vary between lithofacies, but it was also affected by the number of data points. The latter effect seemed to be a threshold-type of behavior, which after about 100 data points, the prediction performance reached a plateau (Fig. 20a). The implications of the amount of training data per facies (or class label) extended beyond performance for those particular facies. Significant class imbalance (e.g., a differential on the order of a magnitude or more for some facies compared to others in the dataset) can result in performance reductions as the algorithms will tend to default to the most commonly seen class labels. This was observed for the more sparsely sampled facies in the dataset in the form of false negatives, which fell strongly within facies that were the most abundant in the dataset (e.g., shale and limestone in Fig. 13a tended to default to calcitic hematite when predicted incorrectly). The effects of class imbalance were generally more problematic for some algorithms, such as SVM. However, there are methods to overcome this limitation where such algorithms must be used and effects of class imbalance are undesirable or intolerable (e.g., Batuwita and Palade, 2013).

The effects of class imbalance in the case of scientific data were unfortunately unavoidable, as sampling occurred mainly using scientific methods (for hypothesis testing and therefore sampling was biased toward samples that best tested the hypothesis) and coverage of facies in geoscientific data depended mainly on exposure. As such, for scenarios beyond method development and into the deployment stage, this is a significant factor that should be addressed through some combination of algorithm and data selection. However, the exact compromise on prediction performance depends on the context and requires a per-application consideration. The prediction performance contained two anomalous behaviors, namely the further reduction in RE for the hematite facies relative to the BIF and calcitic hematite facies, and the excessively high prediction error for limestone. It is clear that the prediction error for all lithologies was strongly and nonlinearly related to their mean Fe grade (Fig. 20b). The effect of the sample population size on the prediction performance can be understood by examining the relationship between the mean Fe grade and the number of samples per lithology (Fig. 20c). Several lithologies (limestone, diabase, shale and breccia) had similar population sizes in the dataset, yet their prediction performances were substantially different (Fig. 20b). Therefore, their prediction performance was more strongly influenced by their mean Fe grades rather than their class populations in the dataset. Similarly, the change in the number of data points across diabase, the BIF, calcitic hematite and hematite facies was relatively





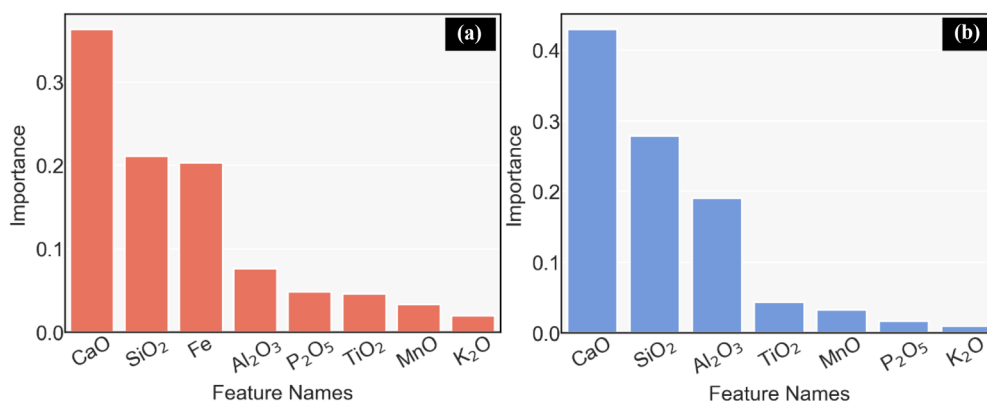
**Figure 17.** In-situ Fe grade (%) prediction results using multiple machine learning algorithms show actual versus predicted values (Elastic-Net not shown). Abbreviations: kNN = k-nearest neighbors; SVM = support vector machine; and ANN = artificial neural network.

large, yet their prediction performance did not change in a corresponding manner. As such, we concluded that the main effect controlling the prediction performance of the Fe grade was the mean Fe concentration. Based on our feature importance testing results, this implies that the relationships among CaO, SiO<sub>2</sub> and Al<sub>2</sub>O<sub>3</sub> were stronger throughout the deposit, especially where the Fe grade was high. It seemed probable that the mech-

anism(s) of Fe enrichment in this deposit impacted the multi-elemental associations the least in limestone and the most in the hematite facies. To further support this idea using predictive modeling-performance analysis, we would require a significantly larger dataset, such that all of the lithofacies have reached the performance plateau and in a manner that avoids observable class imbalance (the latter of which may not be possible for this type of data).

**Table 3.** Model testing performance based on optimum hyperparameters for several metrics for Fe grade prediction

Algorithm	CoD	MedAE (Fe%)	RE	Parameters
SVM	0.983	1.049	0.069	$C=200$ , $\gamma=1.0$ , kernel = RBF
Gaussian process	0.980	1.209	0.069	$\alpha=0.1$ , kernel = RBF(0.00268)
AdaBoost	0.979	1.072	0.073	Number of estimators = 250, max features = 4, max depth = none, min samples leaf = 1, min samples split = 2,
Random forest	0.973	1.259	0.099	Max depth = none, max features = 4, min samples leaf = 1, min samples split = 2, number of estimators = 1500
kNN	0.970	1.426	0.086	$k=5$
ANN	0.980	1.242	0.078	Activation = tanh, $\alpha=0.0001$ , learning rate = invscaling, max iteration = 2000, hidden layer size = 100
Elastic-Net	0.973	1.664	0.101	$\rho=0.90$

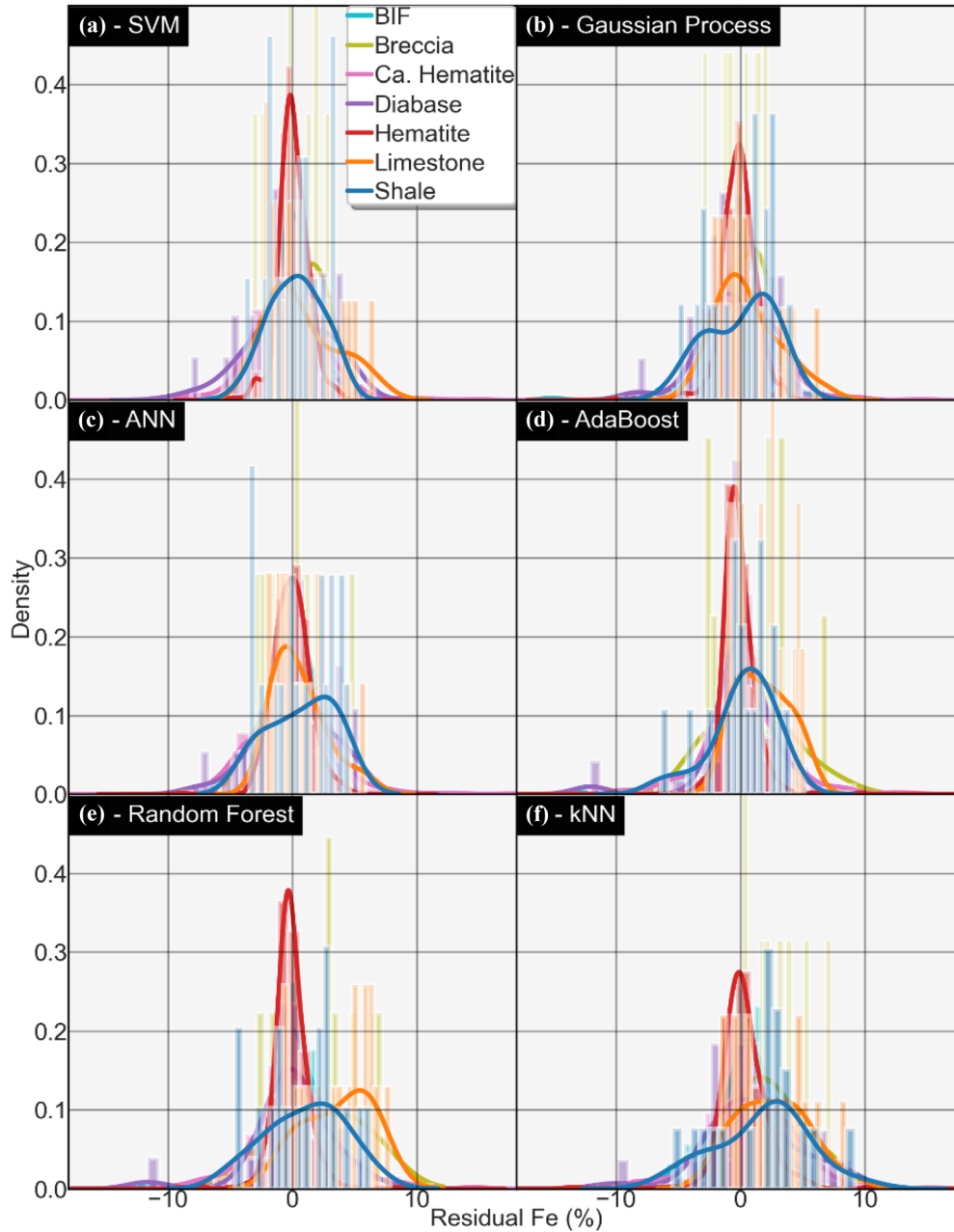
**Figure 18.** Ranking of feature importance for (a) lithofacies classification and (b) Fe concentration prediction.

Such a dataset would enable the use of stratified sampling and performance profiling for a range of lithofacies to profile prediction performance changes as a function of the fraction of the dataset used.

### Implications of Data-driven Approaches in the Characterization of Fe Ore Deposits

For lithofacies classification, CaO and SiO<sub>2</sub> were both highly important, more so than Fe. This implies that the mechanism of Fe enrichment at the Assen Fe ore deposit is plausibly highly related to CaO and SiO<sub>2</sub>. It seems almost certain that the main Fe concentration mechanism could not have possibly been hypogene in this regard, as systematic, cross-stratigraphic coupling of CaO, SiO<sub>2</sub> and Fe would be improbable as a result of any primary hypogene processes. In addition, the poor lithofacies classifi-

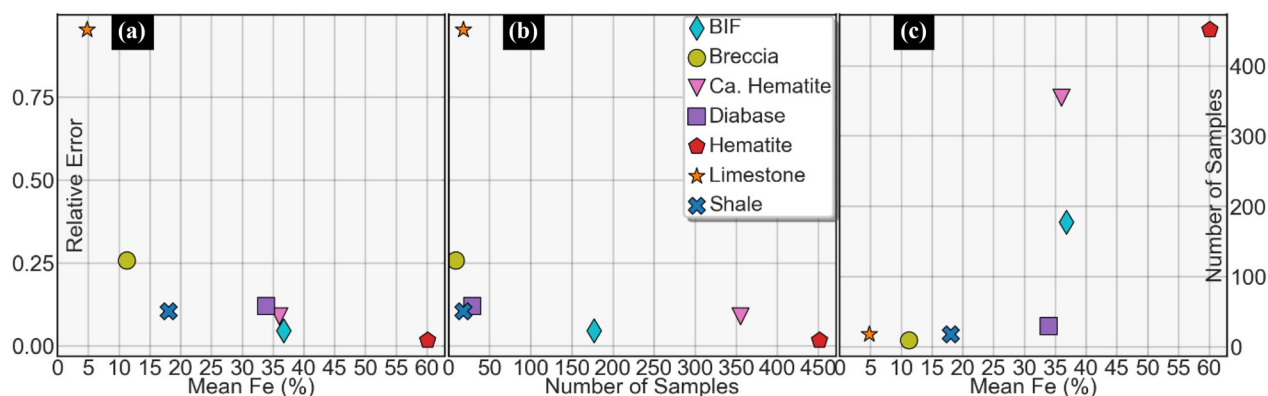
cation performance of diabase and, to a lesser extent, of breccia seems to imply that both were chemically not very contrasting with the other lithofacies, particularly given the amount of data available for those lithofacies. Therefore, it would seem improbable that Fe across the lithofacies would have come from diabase intrusions, which is a possible hypogene source of Fe because, in that context, Fe would have been a stronger predictor of lithofacies. Furthermore, the relationships among P<sub>2</sub>O<sub>5</sub>, Fe and MnO were positive, whereas, between CaO and Fe, the relationship seemed to be a generally negative one (Fig. 9). This was also confirmed from the PCA biplots (Fig. 10). Assuming that Fe, MnO and P<sub>2</sub>O<sub>5</sub> were mainly co-genetic, it would imply that the enrichment of Fe was mainly supergene and of a diagenetic nature (Smith & Beukes, 2016). Certainly, subsequent processes (e.g., secondary or tertiary) did not degrade (or perhaps



**Figure 19.** Histogram of prediction residuals (predicted minus actual) averaged over 100 runs for various algorithms. Data points are color-coded by lithofacies. Abbreviations: kNN = k-nearest neighbors; SVM = support vector machine; and ANN = artificial neural network.

strengthen) the primary multi-elemental relationships because Fe can be predicted with high CoD scores, which implies that a vast majority of the variability of the Fe grade was captured by the different models. From the PCA (Fig. 10) and the scatter matrices (Fig. 9), CaO seemed to be a key differentiator between calcitic hematite and hema-

tite facies. In addition, there seemed to be some of the hematite facies that bear a calcitic hematite facies signature along PC1 and PC2 (Fig. 10). Given that the relationship between CaO and Fe was a negative correlation and that CaO was highly predictive of Fe concentrations, it would mean that if the Fe enrichment processes occurred in a supergene



**Figure 20.** (a) Relative error versus the mean Fe grade. (b) Relative error versus the number of samples. (c) The number of samples versus the mean Fe grade. The results were obtained over 100 runs using a parameter-optimized support vector machine model.

setting, then Fe mainly replaced calcium selectively or it was added instead of mainly calcium to the deposit. However, this requires further investigation to ascertain. If this were true, then it implies that there existed a very close genetic relationship between the calcitic hematite and hematite facies, and their variation in the deposit captured environmental changes in the deposition condition. In addition, Fe being a moderate predictor of lithofacies implies that, under the supergene Fe-enrichment hypothesis, the enrichment occurred in a manner that, in some cases, was a gradual onset process to any particular strata. This would be consistent with a supergene and spatially diffuse type of enrichment process.

Although predictive modeling is most often used for predictions that could guide decision-making (e.g., exploration planning), its origin as a formalization of data modeling (e.g., manual regression modeling) implies that it can also be used to test rapidly various lines of scientific hypotheses. Traditional and often discipline-specific forms of data modeling, when used for hypothesis testing, attempt to determine how well the data fit a particular hypothesis. With machine learning-based predictive modeling, the quality of the fit depends on the algorithm used, the algorithm tuning and the data. In this case, the quality of predictive models can be analyzed, and the predictive power of various features can be examined in the context of competing scientific hypotheses. The manner in which predictive modeling can aid hypothesis testing is similar to that with traditional data modeling, with the usual caveats, e.g., of data bias toward a particular lithofacies. In our application, the hypotheses that cor-

rectly predicted certain relationships were more likely to be correct than the ones that did not make these predictions. For the Assen Fe ore deposit, any hypothesis that attempted to explain the enrichment of Fe must address adequately the observation that calcium and silicon were intricately related to the occurrence of Fe and that this relationship seemed to hold across the stratigraphy to within data availability and bias. There is a possibility of yet unobserved and additional correlation that may be demonstrable with additional data or data from other similar deposits. This is a typical caveat of data-driven insight generation, which is not an issue for our purposes, as we did not generalize our observations beyond the Assen deposit and we do not claim universally the validity of any particular class of scientific hypotheses.

Sedimentary facies reflect the particular physical, chemical and biological conditions that rocks experienced during sedimentation and post-depositional alteration. There is a consensus that BIF-hosted Fe ore deposits are upgraded via supergene and hydrothermal leaching of silica and re-precipitation of Fe under oxidizing conditions (Smith & Beukes, 2016) associated with varying degrees of deep weathering (Hagemann et al., 2006) that affected the hypogene enrichment process. In the case of the Assen Fe ore deposit, the predictive qualities of the rock-forming elements were reflective of the nature of the formation of the deposit, in the sense that the Fe enrichment processes are wholly dependent on the proto-rock composition (i.e., BIF and its associated physical properties). The results of our data-driven analysis were strongly supportive of

a supergene enrichment process followed by tertiary dolerite alteration, whereby most of the Fe enrichment occurred in a supergene setting. The enrichment process enhanced specifically the relationships among Fe, CaO and SiO<sub>2</sub> in the high-Fe lithology—the hematite facies, relative to all other lithologies, to the extent that these relationships are highly predictive of the Fe grade in the hematite facies. In comparison, the effect of the Fe-enrichment processes was weaker in other lithologies and particularly limestone, which is common in the marine environment. The dolerite intrusions in the deposit did not seem to contribute materially to the Fe balance; however, they provided a reducing agent during supergene enrichment processes, especially along with their lithological contacts. Permeability due to early-stage deformation, the doming relayed specifically to the intrusion of the Bushveld Igneous Complex and, later, the Karoo Supergroup-age dykes and sills (Drakensberg Group, 183–180 Ma; SACS, 1980) may have been such conduits. This may have resulted in the depletion of the BIF of non-Fe metals along with the intrusive/BIF contacts, with the mobilized Fe being re-deposited in the hematite facies. We deduced that the spatial relationships between the dolerite and the high-grade hematite ores suggest that the dolerite intrusions influenced the deposition of the hematite locally as reflected by the mineral assemblage proximal and distal to the dolerite intrusions. This demonstrates the role of dolerite intrusions in controlling and remobilizing Fe ore mineralization locally on the precipitation and deposition of orebodies. The deformed BIF facies provided the channel ways for Fe-rich fluids to flow down the stratigraphy during the process of ore enrichment. This may well have caused the observation that Fe was not the main distinguishing factor of rock types within the stratigraphy of the Assen deposit. The ability to predict Fe content using associated, major element, whole-rock geochemistry analysis demonstrates that the primary Fe ore was concentrated by a similar process and that the multielement associations remained intact regardless of weathering. Phosphorus and sulfur are considered to be contaminants in Fe ores (Muwanguzi et al., 2012). The Assen Fe ore deposit has low phosphorus and sulfur content, which facilitate it to be viable economically because key contaminants are known to lower the solidification temperature, increase fluidity and render the metal very fluid indirectly during the production of a low melting constituent in Fe and steel making (Muwanguzi et al., 2012). How-

ever, enrichment of Fe is associated with enrichment of P<sub>2</sub>O<sub>5</sub>; hence, complete avoidance of phosphorus may be in general impossible if the underlying mechanisms of P<sub>2</sub>O<sub>5</sub> and Fe enrichment are generalizable throughout similar deposits. The ability to use machine learning algorithms to identify the primary lithology using lithochemochemistry indicates that different rock types are generally distinguishable chemically, with the exceptions of some of the altered lithofacies boundaries. In addition to the ability to predict the Fe grade, an approach similar to ours could be developed readily to predict the phosphorous and sulfur content in material samples to aid downstream processing by providing advanced material characterization knowledge.

## CONCLUSIONS

Machine learning-based lithofacies classification and prediction of in-situ Fe grade at the Assen Fe ore mine were both highly accurate. For lithofacies classification, a tuned AdaBoost model was able to predict a borehole that was unseen during predictive modeling with perfect accuracy. For in-situ Fe grade prediction, despite the oxidation state within the sampled lithologies being highly variable on a per-lithology basis, the performance as measured through the CoD metric was consistently over 97% for all algorithms explored, and with a MedAE of no worse than 1.664 Fe%. These results suggest that it is both practical and effective to use machine learning-based methods to perform lithofacies classification and Fe grade prediction. Aside from the obvious use of predictive modeling for inferential purposes, feature importance testing showed that machine learning-based methods can also assist with hypothesis testing, by identifying the predictive powers of various features (chemicals in this study) and understanding the relationships between prediction performance and key descriptive statistics of the dataset (number of samples and Fe grade in this paper). We demonstrated that the prediction performance of lithofacies and Fe grade using chemistry was controlled strongly by the use of CaO, SiO<sub>2</sub> and Fe as features for the classification of lithofacies, while CaO, SiO<sub>2</sub> and Al<sub>2</sub>O<sub>3</sub> were strong predictors of in-situ Fe grade across the stratigraphy. Furthermore, we showed that the Fe enrichment across the lithofacies controlled strongly the in-situ Fe grade prediction performance, to an extent that cannot be accounted for by data abundance on a per lithofacies



basis. Based on our data and for the Assen Fe ore deposit, our results are supportive of enrichment of Fe through mainly a supergene process, likely of a diagenetic nature. Dolerite intrusions seemed to have played an insignificant role in the direct enrichment of Fe ore at the Assen Fe mine.

## ACKNOWLEDGMENTS

We acknowledge critical comments received from the AMC Consultants team, particularly, Alex Virisheff and Chris Arnold. We would also like to thank Prof. John Carranza for editorial handling and two anonymous reviewers, whose comments helped to significantly improve this manuscript.

## FUNDING

Open access funding provided by Lulea University of Technology. Funding for this research was provided by the Department of Science and Innovation (DSI)-National Research Foundation (NRF) Thuthuka Grant (Grant UID: 121973).

## DECLARATIONS

**Conflict of 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.

## OPEN ACCESS

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To

view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## SUPPLEMENTARY INFORMATION

The online version contains supplementary material available at <https://doi.org/10.1007/s11053-022-10105-z>.

## REFERENCES

- Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), 139–160.
- Altmann, A., Tolosi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10), 1340–1347.
- An, S., Liu, W., & Venkatesh, S. (2007). Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. *Pattern Recognition*, 40(8), 2154–2162.
- Barnett, R. M. (2017). Principal component analysis. In J. L. Deutsch (Ed.), *Geostatistics lessons*. Retrieved January 5, 2022, from <http://geostatisticslessons.com/lessons/principalcomponentanalysis.html>.
- Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector machines. In H. He & Y. Ma (Eds.), *Imbalanced learning: Foundations, algorithms and applications*. Wiley Online Library. <https://doi.org/10.1002/9781118646106.ch5>.
- Bekker, A., Slack, J. F., Planavsky, N., Krapez, B., Hofmann, A., Konhauser, K. O., & Rouxel, O. J. (2010). Iron formation: The sedimentary product of a complex interplay among mantle, tectonic, oceanic, and biospheric processes. *Economic Geology*, 105, 467–508.
- Beyeme-Zogo, J.-C. (2009). *Beneficiation potential of low-grade iron from a discard lumpy stockpile and fines tailings dam at Beeshoek Mine, Northern Cape Province, South Africa*. MSc Thesis, University of Johannesburg, South Africa. pp.151.
- Beukes, N. J., & Gutzmer, J. (2008). Origin and paleoenvironmental significance of major iron formations at the Archean-Paleoproterozoic boundary. In S. Hagemann, C. A. Rosière, J. Gutzmer, & N. J. Beukes (Eds.), *Banded Iron Formation related high-grade iron ore* (pp. 5–47). Reviews in Economic Geology. <https://doi.org/10.5382/Rev.15.01>.
- Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- Breiman, L. (1996b). Stacked regressions. *Machine Learning*, 24(1), 49–64.
- Breiman, L. (2001a). Random forests. *Machine Learning*, 45, 5–32.
- Breiman, L. (2001b). Statistical modelling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231.
- Burger, A. J., & Coertze, F. J. (1975). Age determinations - April 1972 to March 1974. *Annals of the Geological Survey of South Africa*, 10, 135–141.
- Button, A. (1981). The cratonic environment, The Pongola Supergroup. In D. R. Hunter (Ed.), *Precambrian of the Southern Hemisphere* (pp. 501–510). Elsevier.

- Button, A. (1986). The Transvaal sub-basin of the transvaal sequence. In C. R. Annhaesser & S. Maske (Eds.), *Mineral deposits of Southern Africa* (pp. 811–817). Geological Society of South Africa.
- Catuneanu, O., & Eriksson, P. G. (1999). The sequence stratigraphic concept and the Precambrian rock record: An example from the 2.7–2.1 Ga Transvaal Supergroup. *Kaapvaal craton. Precambrian Research*, 97(3–4), 215–251.
- Chen, S., Hattori, K., & Grunsky, E. C. (2018). Identification of sandstones above blind uranium deposits using multivariate statistical assessment of compositional data, Athabasca Basin, Canada. *Journal of Geochemical Exploration*, 188, 229–239.
- Cover, T., & Hart, P. (1967). Nearest neighbour pattern classification. *IEEE Transactions on Information Theory*, 13, 21–27.
- Cracknell, M. J., & de Caritat, P. (2017). Catchment-based gold prospectivity analysis combining geochemical, geophysical and geological data across northern Australia. *Geochemistry: Exploration, Environment, Analysis*, 17(3), 204–216.
- Cramer, J. S. (2002). The origins of logistic regression. *Tinbergen Institute*, working paper no. 2002-119/4, pp. 16. Doi: <https://doi.org/10.2139/ssrn.360300>.
- Curry, H. B. (1944). The method of steepest descent for non-linear minimisation problems. *Quarterly Applied Mathematics*, 2, 258–261.
- Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 598–617. Doi: <https://doi.org/10.1109/SP.2016.42>.
- Dreher, C. L., Schad, M., Robbins, L. S., Konhauser, K. O., Kappler, A., & Joshi, P. (2021). Microbial processes during deposition and diagenesis of Banded Iron Formations. *PalZ*, 95, 593–610.
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87.
- Eriksson, P. G., Schweitzer, J. K., Bosch, P. J. A., Schreiber, U. M., Van Deventer, J. L., & Hatton, C. J. (1993). The transvaal sequence: An overview. *Journal of African Earth Sciences*, 16(1–2), 25–51.
- Eriksson, P. G., Hattingh, P. J., & Altermann, W. (1995). An overview of the geology of the transvaal sequence and Bushveld complex, South Africa. *Mineralium Deposita*, 30(2), 98–111.
- Eriksson, P. G., Altermann, W., & Hartzer, F. J. (2006). The transvaal supergroup and its precursors. In M. R. Johnson, C. R. Anhaeuser, & R. J. Thomas (Eds.), *The geology of South Africa* (pp. 237–260). Geological Society of South Africa and Council for Geoscience.
- Fawcett, T. (2006). Introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861–874.
- Fix, E., & Hodges, J. L. (1951). An important contribution to nonparametric discriminant analysis and density estimation. *International Statistical Institute*, 57, 233–238.
- Fockema, R. A. P. (1948). The geology south of the confluence of the Pienaars and Crocodile rivers. *Transactions of the Geological Society of South Africa*, 51(1), 63–75.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalisation of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139.
- Gerya, T. V., Uken, R., Reinhardt, J., Watkeys, M. K., Maresch, W. V., & Clarke, B. M. (2003). Cold fingers in hot magma: Numerical modeling of country-rock diapirs in the Bushveld complex, South Africa. *Geology*, 31, 753–756.
- Gregorutti, B., Michel, B., & Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90, 15–35.
- Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27(3), 659–678.
- Gross, G. A. (1980). A classification of iron formations based on depositional environments. *Canadian Mineralogist*, 18, 215–222.
- Gross, G. A. (1993). Industrial and genetic models for iron ore in iron-formations. In R. V. Kirkham, W. D. Sinclair, R. I. Thorpe, & J. M. Duke (Eds.), *Mineral deposit modelling (Special Paper 40)* (pp. 151–170). Publisher Geological Association of Canada.
- Grunsky, E. C., Mueller, U. A., & Corrigan, D. (2014). A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: Application for predictive geological mapping. *Journal of Geochemical Exploration*, 141, 15–41.
- Grunsky, E. C., & de Caritat, P. (2019). State-of-the-art analysis of geochemical data for mineral exploration. *Geochemistry: Exploration, Environment, Analysis*, 20, 217–232.
- Gu, A., Sala, F., Gunel, B., & Ré, C. (2019). Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*.
- Gutzmer, J., Chisonga, B. C., Beukes, N. J., & Mukhopadhyay, J. (2008). The geochemistry of banded iron formation-hosted high-grade hematite-martite iron ores. In S. Hagemann, C. Rosière, J. Gutzmer, & N. J. Beukes (Eds.), *Banded Iron Formation related high-grade iron ore* (pp. 157–183). Publisher Reviews in Economic Geology. <https://doi.org/10.5382/Rev.15.06>.
- Hagemann, S. G., Rosière, C. A., Lobato, L., Baars, F., Zucchetti, M., & Figueiredo e Silva, R. C. (2006). Controversy in genetic models for Proterozoic high-grade, banded iron formation (BIF)-related iron deposits – unifying or discrete model(s)? *Applied Earth Science IMM Transactions Section B*, 115, 147–151.
- Hagemann, S. G., Angerer, T., Duuring, P., Rosière, C. A., Figueiredo e Silva, R., Lobato, L., Hensler, A. S., & Walde, D. H. G. (2016). BIF-hosted iron mineral system: A review. *Ore Geology Reviews*, 76, 317–359.
- Halverson, G. P., Poitrasson, F., Hoffman, P. F., Nédélec, A., Montel, J.-M., & Kirby, J. (2011). Fe isotope and trace element geochemistry of the Neoproterozoic syn-glacial Rapitan iron formation. *Earth and Planetary Science Letters*, 309(1–2), 100–112.
- Harley, M., & Charlesworth, E. G. (1992). Thrust-controlled gold mineralisation at the Elandshoogte Mine, Sabie-Pilgrim's Rest goldfield, South Africa. *Mineralium Deposita*, 27, 122–128.
- Harris, J. R., Grunsky, E., Behnia, P., & Corrigan, D. (2015). Data-and knowledge-driven mineral prospectivity maps for Canada's North. *Ore Geology Reviews*, 71, 788–803.
- Hartzer, F. J. (1987). *Die Geologie van die Krokodil rivier fragment, Transvaal*. MSc thesis, Rand Afrikaans University, Johannesburg, South Africa.
- Hartzer, F. J. (1989). Stratigraphy, structure, and tectonic evolution of the Crocodile river fragment. *South Africa Journal of Geology*, 92(2), 110–124.
- Hartzer, F. J. (1995). Transvaal supergroup inliers: Geology, tectonic development and relationship with the Bushveld complex, South Africa. *Journal of African Earth Sciences*, 21(4), 521–547.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). Springer.
- Henry, G., & Master, S. (2008). *Black reef project*. Council for Scientific and Industrial Research (CSIR) and University of the Witwatersrand.
- Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and*

- Recognition (pp. 278–282). Montréal, Canada. Doi: <https://doi.org/10.1109/ICDAR.1995.598994>.
- Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13, 415–425.
- Iwashita, F., Friedel, M. J., de Souza Filho, C. R., & Fraser, S. J. (2011). Hillslope chemical weathering across Paraná, Brazil: A data mining-GIS hybrid approach. *Geomorphology*, 132(3–4), 167–175.
- Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector machines in R. *Journal of Statistical Software*, 15(9), 1–28.
- Klein, C. (2005). Some Precambrian banded iron-formations (BIFs) from around the world: Their age, geologic setting, mineralogy, metamorphism, geochemistry, and origin. *American Mineralogist*, 90, 1473–1499.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59–69.
- Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Springer.
- Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. In I. G. Maglogiannis (Ed.), *Emerging artificial intelligence applications in computer engineering* (Vol. 160, pp. 3–24). IOS Press.
- Kotsiantis, S. B. (2014). Bagging and boosting variants for handling classifications problems: A survey. *The Knowledge Engineering Review*, 29, 78–100.
- Lemaréchal, C. (2012). Cauchy and the gradient method. *Doc Math Extra*, 251(254), 10.
- Mungall, J. E., Kamo, S. L., & McQuade, S. (2016). U-Pb-geochronology documents out-of-sequence emplacement of ultramafic layers in the Bushveld Igneous complex of South Africa. *Nature Communications*, 7, 1–13.
- Muwanguzi, A. J. B., Karasev, A. V., Byaruhanga, J. K., & Jönsson, P. G. (2012). Characterisation of chemical composition and microstructure of natural iron ore from Muko deposits. *International Scholarly Research Network*, 2012, 174803.
- Nwaila, G., Frimmel, H. E., & Minter, W. E. L. (2017). Provenance and geochemical variations in shales of the Mesoproterozoic Witwatersrand supergroup. *The Journal of Geology*, 125, 399–422.
- PorterGeo (2021). Raptian iron formation—crest iron deposit, Yukon Territory, Canada. Porter GeoConsultancy Ltd. Retrieved October 3, 2021, from <http://www.portergeo.com.au/database/mineinfo.asp?mineid=mn104>.
- Pufahl, P. K., & Hiatt, E. E. (2012). Oxygenation of the earth's atmosphere–ocean system: A review of physical and chemical sedimentologic responses. *Marine and Petroleum Geology*, 32(1), 1–20.
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.
- Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of Naive Bayes text classifiers. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)* (pp. 616–623). Washington DC.
- Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books. [https://doi.org/10.1007/978-3-642-70911-1\\_20](https://doi.org/10.1007/978-3-642-70911-1_20).
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition I: Foundation*. MIT Press.
- Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice-Hall.
- Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249.
- Santosa, F., & William, W. S. (1986). Linear inversion of band-limited reflection seismograms. *Journal on Scientific and Statistical Computing*, 7, 1307–1330.
- Smith, A. J., & Beukes, N. J. (2016). Palaeoproterozoic Banded Iron formation hosted high-grade hematite iron ore deposits of the Transvaal Supergroup, South Africa. *Episodes*, 39(2), 269–284.
- Smith, A. J. B. (2018). The iron formations of Southern Africa. In S. Siegesmund, M. A. S. Basei, P. Oyhantçal, & S. Oriolo (Eds.), *Geology of Southwest Gondwana, regional Geology reviews* (pp. 469–491). Springer International Publishing.
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1), 307.
- Teitler, Y., Duuring, P., & Hagemann, S. G. (2014). Styles and controls of BIF-hosted iron ore in Archean terranes, comparison between the Pilbara and Yilgarn Cratons. In *Australian Earth Sciences Convention (AESC)*. Newcastle.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.
- Tikhonov, A. N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39, 195–198.
- Thorne, W., Hagemann, S., Webb, A., & Clout, J. (2008). Banded iron formation-related iron ore deposits of the Hamersley Province, Western Australia. In S. Hagemann, C. Rosière, J. Gutzmer, & N. J. Beukes (Eds.), *Banded iron formation related high-grade iron ore* (pp. 197–221). Reviews in Economic Geology. <https://doi.org/10.5382/Rev.15.08>.
- Trendall, A. (2005). Banded iron formations. In R. C. Selley, L. Robin, R. M. Cocks, & I. R. Plimer (Eds.), *Encyclopedia of geology*. Elsevier. <https://doi.org/10.1016/B0-12-369396-9/00440-8>.
- Urban, H., Stribny, B., & Lippolt, H. J. (1992). Iron and manganese deposits of the Urucum District, Mato Grosso do Sul, Brazil. *Economic Geology*, 87, 1375–1392.
- USGS (2021). Iron ore statistics and information. National minerals information center. Retrieved October 3, 2021, from <https://www.usgs.gov/centers/nmic/iron-ore-statistics-and-information>.
- Vapnik, V. (1998). *Statistical learning theory*. Springer.
- Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). Morgan Kaufman.
- Zeh, A., Ovcharova, M., Wilson, A. H., & Schaltegger, U. (2015). The Bushveld Complex was emplaced and cooled in less than one million years - results of zirconology, and geotectonic implications. *Earth and Planetary Science Letters*, 418, 103–114.
- Zhang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research*, 116(1), 16–32.
- 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. *Artificial Intelligence in Geosciences*, 2, 60–75.
- Zhang, S. E., Nwaila, G. T., Bourdeau, J. E., Frimmel, H. E., Ghorbani, Y., & Elhabyan, R. (2022). Application of machine-learning algorithms to the stratigraphic correlation of Archean shale units based on litho-geochemistry. *Journal of Geology*, 129(6), 647–672.
- Zhu, R., Zeng, D., & Kosorok, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512), 1770–1784.
- Zou, H., & Hastie, T. (2005). Regularisation and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodology)*, 67(2), 301–320.