Application of Machine-Learning Algorithms to the Stratigraphic Correlation of Archean Shale Units Based on Lithogeochemistry

Steven E. Zhang,[1,*] *Glen T. Nwaila,*[2] *Julie E. Bourdeau,*[1] *Hartwig E. Frimmel,*[3] *Yousef Ghorbani,*[4] *and Riham Elhabyan*[5]

1. SmartMin, 39 Kiewiet Street, Helikon Park, 1759, South Africa; 2. School of Geosciences, University of the Witwatersrand, Private Bag 3, Johannesburg, 2050, South Africa; 3. Bavarian Georesources Centre, Department of Geodynamics and Geomaterials Research, Institute of Geography and Geology, University of Würzburg, Am Hubland, D-97074 Würzburg, Germany; and Department of Geological Sciences, University of Cape Town, Rondebosch 7700, South Africa; 4. Department of Civil, Environmental and Natural Resources Engineering, Luleå University of Technology, SE-97187 Luleå, Sweden; 5. Carleton University, Ottawa, 1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada

ABSTRACT

Data-driven methods have increasingly been applied to solve geoscientific problems. Incorporation of data-driven methods with hypothesis testing can be effective to address some long-standing debates and reduce interpretation uncertainty by leveraging larger volumes of data and more objective data analytics, which leads to increased reproducibility. In this study, lithogeochemical data from regionally persistent Archean shale units were aggregated from literature, with special reference to the Kaapvaal Craton of South Africa—namely, shales from the Barberton, Witwatersrand, Pongola, and Transvaal Supergroups—and the Belingwe and Buhwa Greenstone Belts of the Zimbabwe Craton. We examine the feasibility of using machine-learning algorithms to produce a geochemical classification and demonstrate that machine learning is capable of accurately correlating stratigraphy at the formation, group, and supergroup levels. We demonstrate the ability to extract highly useful scientific findings through a data-driven approach, such as geological implications for the uniqueness of the sediment compositions of the Central Rand and West Rand Groups. We further demonstrate that when lithogeochemistry and machine-learning algorithms are used, only about 50 samples per geological unit are necessary to reach accuracy levels of around 80%–90% for our shale samples. Consequently, for many traditional tasks, such as rock identification and mapping, some expensive analyses and manual labor can be replaced by an abundance of cheaper data and machine learning. This approach could transform large-scale geological surveys by enabling more detailed mapping than currently possible, by vastly increasing the coverage rate and total coverage. In addition, the aggregation of historical data facilitates data reuse and open science. These results justify the need to bridge data- and hypothesis-driven techniques for the stratigraphic correlation and prediction of rock units, which can improve the accuracy of the inferred stratigraphic correlation and basin setting.

Online enhancements: supplementary table.

Introduction

The fundamental goal of data collection in geosciences is to understand physical and chemical processes and their complex interaction in space and time through the extraction of increasingly nuanced and multidimensional insights. Broadly speaking, along with mapping and spatial analysis, two other significant usages of geological data are hypothesis testing and categorization. Stratigraphic classification is a type of categorization that is used to distinguish geological bodies, define assemblages of sedimentary, igneous, and metamorphic rocks, and organize sequences of rock strata in the Earth's crust into some convenient classes by delineating class boundaries between many different characteristics of these rocks (Zalasiewicz et al. 2004). As

Manuscript received January 8, 2021; accepted October 21, 2021; electronically published January 20, 2022.

* Author for correspondence; email: ezhan053@uottawa.ca.

Journal of Geology, volume 129, number 6, November 2021. https://doi.org/10.1086/717847

the Earth is a complex system and it is common for multiple processes to affect the stratigraphic classification of typical rocks, multidimensional concurrent constraints are necessary to accurately classify rocks. While the exact characteristics required to classify a rock vary by inferred rock type, they typically include some combination of physical, mineralogical, and chemical characteristics that range from whole-rock geochemistry to mineralogical composition, macroscopic and microscopic textures, and so on. Classification of sedimentary rocks, especially siliciclastic sedimentary rocks, is more complicated because of the diversity and admixtures of source rocks, depositional processes, and nature of deposurface conditions. Although major advances have been made on classifying clastic sedimentary rocks such as sandstones, more work is required when dealing with fine-grained rocks such as shales, which are difficult to distinguish without detailed and sometimes costly analyses. In reductive geosciences, if the classification power of one characteristic is insufficient, additional lines of evidence would be required. However, as classification in geosciences makes use of specific discrimination criteria such as elemental or isotopic ratios, not all of the information contained within data is readily used. The extraction of multidimensional insights and constraints to classify rocks is not suitable to traditional geoscientific techniques that make use of highly specific discrimination criteria.

Classification in the broadest sense is a typical machine-learning problem, which in the context of science becomes a data-driven approach to scientific inquiry. Data-driven science is still relatively rare in geosciences; however, there is increasing adoption of machine learning in geosciences to overcome challenges of dealing with data (Karpatne et al. 2018; Chen et al. 2020; Dramsch 2020). Part of this trend is due to the suitability of data-driven techniques for the extraction of insights from abundant, mixed quantitative-qualitative, and multidimensional data. In particular, in the context of the broader interdisciplinary definition of the task of classification, stratigraphic correlation is a discipline-specific subset of classification. In this sense, there are two main tasks that are necessary to carry out stratigraphic correlation—the construction of classes and the assignment of samples into such classes. In essence, the first task is precisely the process of creating stratigraphic classes, such as lithostratigraphic units in geosciences. Therefore, given sufficient and particularly annotated data, it is possible to apply many of the tools from data science and, in particular, machine learning to create stratigraphic classes. In this case, the two main obvious applications are the construction of classes in a manner that is driven by data instead of by hypothesis and post hoc analyses of the appropriateness of existing stratigraphic classifications. The second task is essentially the assignment of samples into either data-driven or traditional classes, using some known sample characteristics. In the context of assigning rock samples into correct stratigraphic units, this task is also known as stratigraphic correlation in geosciences. For both of these tasks, shale is an interesting challenge to modern data-driven techniques, because "shale" is a diverse group of fine-grained sedimentary rocks that form by compaction of silt- and clay-sized particles. Constituent particles can originate from parental igneous, metamorphic, and even sedimentary rocks, and therefore the possibilities of characteristic overlap between stratigraphic classes are much larger than those for pure igneous or metamorphic rocks. For example, it is entirely possible that two shale samples from different stratigraphic units exhibit the same lithogeochemistry and that the differentiation between the two requires finding other distinctive features by using either another line of evidence or a statistical consideration.

Geologically, shale is differentiated from other mudstones by the extent of diagenetic/low-grade metamorphic overprint. Shale units and formations are common in almost all sedimentary basins throughout the history of the Earth and varying geological environments. There are several ways to classify shales, although a purely lithogeochemical classification scheme does not exist, because shales are theoretically difficult to distinguish on the basis of their chemistry. On the basis of organic matter, two types of shales can be classified, namely, black, carbon-rich and gray, carbon-poor shales. Carbon-rich shales contain appreciable amounts of organic carbon (>0.5 vol%) and have been common sedimentary rocks since Neoarchean times (Condie et al. 2001; Meyer and Kump 2008; Lyons et al. 2009). In contrast, carbon-poor shales contain low amounts of organic carbon, are usually laminated, and were mostly deposited during Archean times. As shales record changes in crustal processes and may be associated with natural resources, their characterization is very important to the geosciences as well as the energy and mineral industries.

Various studies have revealed that oxidized basins can produce stratiform Cu-Co, Zn-Pb, and U deposits, whereas reduced basins can produce Au-As, P, Ba, and metalliferous black shales (Ni, Mo, Zn, Cu, U, and the platinum-group elements [PGEs]: Pt, Pd, Ir, Os, Ru, and Rh; Jiang et al. 2007; Large et al. 2015). Within the metalliferous shales,

the most sought-after type is pyritic black shales, which are commonly known to be a source of desirable metals (Ni, Mo, As, Zn, Cu, Co, U, V, Ag, Au, and the PGEs). Black shales, mainly those of Phanerozoic and, to some extent, Proterozoic age, have been studied for their potential as either sources or hosts of mineral deposits (e.g., Talvivaara, Finland [Kontinen et al. 2013]; the Alum Shale, Sweden [Leventhal 1991]; and the Cambrian Ni-Mo-rich shales in southern China [Xu et al. 2013]). The ore-concentrating processes leading to the formation of mineralized black shales include primary enrichment during deposition (e.g., syngenetic) and secondary concentration of metals through interaction with hydrothermal fluids (e.g., Vine and Tourtelot 1970; Coveney et al. 1991; Jiang et al. 2007; Lehmann et al. 2007; Large et al. 2015). Apart from black shales' association with metals, they have been explored for oil and gas, their often rich paleontological and palynological record, and their ability to record and archive changes in Earth's broader environment and hydrosphere-atmosphere evolution (Johnson et al. 2017).

In contrast to Phanerozoic black shales, most of the shales that were deposited during the Archean are carbon poor and usually not metalliferous. In some instances, Archean shales have been studied for their lithogeochemical composition in order to understand sediment provenance and crustal evolution through time and to infer tectonic settings (Wronkiewicz and Condie 1987; Feng and Kerrich 1990). More often, the mineralogical assemblage of Archean shale has been used to interpret metamorphic conditions and for reconstruction of sedimentary facies (Wronkiewicz and Condie 1987; Nwaila et al. 2017). More recently, studies have been conducted on Archean shales to assess their background concentration of siderophile elements and their links to mineralization (Nwaila and Frimmel 2019). Strides have also been made in understanding mechanisms that cause discontinuities in local sediment supply (flux; Kim et al. 2006), which in turn cause discontinuities in deposition (accumulation) of detritus and preservation. Clearly, lithogeochemistry is useful in solving problems of shale classification and stratigraphic correlation and for shale-associated resource exploration, be it for metal resources or for oil and gas. In this context, an abundance of geochemical data is highly suitable for modern data-driven approaches to classification and prediction.

Shale stratigraphic prediction, or pinpointing the origin of shale samples, traditionally has been a geoscientific classification problem that is typically addressed using a cascading series of evidence that begins with geological observations and physical characterization and proceeds to fine-scaled studies, such as microscopy and geochemical analyses. Field surveys are generally expensive but yield valuable specimens and are therefore unavoidable. Lithogeochemical analyses of shales are common practice and have the potential to be more informative and systematic for stratigraphic correlation than purely descriptive field observations, thus forming the base of chemostratigraphy. Isotopic analyses of shales are also sufficiently common; however, the exact selection of elements and isotopes depends on the hypothesis being tested and application-specific subjective and physical constraints. In addition, as isotopic analyses are much more expensive per dimension of information and are usually reserved for select samples, this type of compositional data is rare, and the relative variation of results and procedures, including quality control and assurance between analytical laboratories, is significant compared to that in the analyses of major- or trace-element concentrations. In this manner, historic data are hampered by inconsistencies in the selection of analyzed elements. The usefulness of petrographic analyses is limited because of the very fine grain size of shales. Petrographic analyses are also extremely time-consuming, and the information that they produce is highly qualitative. If it were not for the abundance and willingness of graduate students, petrography in general would be a very costly laboratory affair. However, for as long as there has been interest in shale, there have been disaggregated data that are still accumulating in various data silos. Unfortunately, detailed analyses of shale, such as chemical analyses, are relatively expensive, and no single study has ever attempted to look at a large-scale picture of the lithogeochemistry of shales and whether it could be useful to fingerprint shales for purposes of classification and genetic analysis. This need not be the case if there are sufficient data, such as through aggregation of historical data on the composition of shales. The value of historical data in this case far exceeds the original intent at the time of data gathering. A principal requirement for the adoption of data-driven analytics and modeling is the availability of high-quality data. Aggregation of historical data is one method to generate sufficient data for the application of data-driven scientific methods to geosciences.

In this study, we investigate the use of the lithogeochemistry of Archean shales by repurposing aggregated geochemical data from a variety of sources to perform stratigraphic classification, correlation, and prediction with machine learning. Our whole-rock major-element geochemical data

of Archean terranes has been aggregated from peer-reviewed and published sources, with special reference to the Kaapvaal Craton, South Africa. Although more abundant data are available for other types of rocks, the challenging case of shale classification and the variable data coverage of our data set provide an ideal opportunity to simultaneously examine the feasibility of data-driven classification in the context of geosciences and analyze the data requirements to provide guidance for those wanting to adopt data science techniques as a tool in geosciences. Therefore, the first goal of this article is to demonstrate the value of unsupervised machine learning to rapidly categorize shales using solely major-element lithogeochemistry, although the resulting classification schemes are expectedly different from geological classifications, which use multiple lines of evidence. Given that conventional (geological) classification of shales is dominant in geosciences, a second goal of this article is to demonstrate that major-element concentrations can be used to classify shale samples into traditional classes at the formation, group, and supergroup levels with excellent accuracy. We demonstrate that there are suitable machine-learning algorithms for both raw and transformed compositional data and that, on average, most of them yielded excellent results. Furthermore, we show that about 50 samples from each geological unit are sufficient to achieve a high level of accuracy in the classification of shales using whole-rock major-element concentration data. Thus, we identify data availability as the key limiting factor to the adoption of modern data science in geochemical classification problems.

A Data-Driven Approach for Categorizing and Predicting Rock Types

Conventional classification of shale is largely based on textural characteristics, mineralogy, and trace-element ratios. For most Archean terranes, major-element concentrations in various shale units have been determined, but the trace-element data, especially in older studies, tend to be incomplete. Whole-rock major-element (technically, major- and minor-element) geochemistry should be an ideal setup for fingerprinting shales. In addition, major elements are rock-forming elements and, as such, constitute the bulk of each sample. However, much emphasis has been placed on the use of trace elements because of their resistance to weathering, their mobility and immobility characteristics, and the ability of some of them to preserve information on the original detritus. Nevertheless, major-element concentration data remain the most ubiquitous and consistent form of high-dimensional and quantitative information about shales that is available across publications and databases. In the current geochemical tradition, because the major elements exhibit a differential diagnostic power, some elements are essentially neglected for classification and correlation of stratigraphic units or discrimination purposes. As a whole, there is likely far more information contained within major-element concentration data than is currently extracted through traditional geochemical analyses. Simultaneously, many shale units experienced such insignificant alteration that their major-element concentrations still provide excellent information on the original sediments. Because of the abundance of such information, a data-driven approach is likely to be able to provide new insights and a highly reproducible and objective method to perform stratigraphic correlation and prediction. Nonetheless, whole-rock major-element data remain underutilized, compared to the more powerful trace-element and isotope analyses. There are several reasons for this apparent neglect of whole-rock major-element data, one of which is the fact that many geoscientists gather data solely for the purpose of testing specific hypotheses. While emphasizing certain trace-element and occasionally isotope ratios, their data sets typically also include major elements. The other reason is that geochemistry is generally high-dimensional, and traditional data analysis techniques are manual and use highly specific models to extract known dimensions of scientific insights. Therefore, synthesizing general information from geochemical data without a hypothesis a priori is challenging and is a less well-defined problem in the context of traditional geochemistry. On the other hand, machine-learning algorithms can easily tackle high dimensionality, given that there are sufficient data; in the case that more powerful trace-element and isotopic composition data are unavailable or inconsistent, it is possible to systematically classify rocks and predict their classification. This may be particularly useful for Archean terranes. A contributing factor is the general high cost of data in geosciences (e.g., compared to commerce) and the fact that data tend to become siloed. As a result, disaggregated data sets typically reside with various data stewards at institutions, which makes it a difficult task to access and understand them and to determine their comparability. Fortunately, the most comparable data usually concern major elements, because their detection limits and precision are typically much smaller than their absolute concentrations. Therefore, our major-element concentration data and machine-learning algorithms are highly

suitable for data-driven methods for the classification and prediction of shale through leveraging data, algorithms, and an infrastructure that provide the required computational power. There are a multitude of other benefits that are realizable in the short to long term by utilizing this approach: (1) the high cost of data generation in geosciences is mitigated by sharing of data and building a discipline culture that reuses data formally, which minimizes the need for resampling; (2) the reuse and accumulation of data lead to statistically robust interpretations, constructive debates, more reliable and predictive hypotheses, and increased reproducibility of results; (3) a large backdrop of data facilitates outlier and even deception detection for new data and also an enhanced separation of data, real novel findings, and reproducible results. In this study, we also reflect on these additional benefits of our approach and results, where appropriate.

Overview of Machine Learning

In machine-learning parlance, each chemical element for the purpose of rock classification is a feature, and specific ratios of various elements or all types of mathematical manipulations of features are essentially discipline-specific forms of feature engineering (Hastie et al. 2009; Domingos 2012). The features form a high-dimensional vector space that is the feature space (Hastie et al. 2009). Machine-learning algorithms in general do not assume the geometry or the existence of properties of the feature space. For example, tree-based methods are not at all aware of the geometry of the feature space. Algorithms that are geometry aware in the feature space makes use of notions of distance, such as distance metrics. However, the exact metric employed is not restricted to the Euclidean metric: other metrics are also used, and the vector space can be of any geometry, such as hyperbolic and spherical (e.g., Gu et al. 2019). The choice of vector space to embed the data's features depends on the data's native structure, the choice of algorithms, and, ultimately and most importantly, model performance assessment (Karpatne et al. 2018) using performance metrics (e.g., Gu et al. 2019). In practice, the transformation of data from one structure to another is a subtype of data preprocessing that overlaps with feature engineering, and, as with all data-preprocessing tasks, it should be self-evident that the purpose should be to enhance predictive modeling performance. Selection of features that will lead to better predictive or classification accuracy is often an automatable process in machine learning and is known as feature selection (Hastie et al. 2009). If the features encode characteristics of rocks, which also differ by rock type, then it is possible for the machine to identify these differences and their relationships that can be used for both classification and prediction. In this manner, chemical data are readily usable for the classification and prediction of rocks. In addition, in the machine-learning approach, class boundaries need not be parametric, and assumptions about the shape of classes in feature space (e.g., clusters) can be explicitly examined and discarded.

There are two major types of machine-learning algorithms that are suitable for rock classification: (1) supervised and (2) unsupervised learning. Semisupervised machine learning, as a hybrid of 1 and 2, is also possible. In unsupervised learning, data are unlabeled, which means that the classes are unknown a priori, and the machine attempts to deduce natural categorizations within the data to create a classification scheme, which is then used to classify new data (Hastie et al. 2009). In supervised learning, the data are labeled (e.g., the categories or rock types are known), and the algorithm's hyperparameters are tuned with training and cross-validation data sets (Hastie et al. 2009). The resulting models are then used to predict either continuous (e.g., the amount of metal in a sample) or discrete (e.g., types of rocks) labels. The results of the predictions can be assessed for accuracy through yet another data set that does not overlap the training and cross-validation data sets.

Clustering is a type of unsupervised machine learning and refers to grouping a set of data points into subsets or clusters (Hastie et al. 2009). Each cluster is a class exhibited by the data that is, in some sense, different from other classes. In geosciences, clustering can be applied to a variety of tasks, such as facies classification, mapping, stratigraphic ranking, and domaining of mineral resources. The goal of clustering algorithms is to leverage computational power and data with a mathematically or algorithmically formalized notion of similarity, such as a distance metric in the feature space, to create classes such that they best separate the available data. In other words, the data-driven approach to classification is similar to methods used to create geological classifications, such as a shale formation. In general, clustering algorithms can be divided along two major axes: (1) centroid/parametric or density/nonparametric and (2) flat or hierarchical. Centroid-based algorithms, such as the *K*-means algorithm, seek to find the centroid of clusters and effectively partition the feature space by volume. Density-based algorithms delineate clusters by variations in data density

in feature space. Flat clustering algorithms examine all features simultaneously, whereas hierarchical clustering algorithms build clusters that are linked by various features hierarchically in a treelike structure that can be represented by a dendrogram. There are two approaches to hierarchical clustering, which can be represented by agglomerative and divisive algorithms. An agglomerative approach begins with each observation in a distinct (singleton) cluster and successively merges clusters until a stopping criterion is satisfied. A divisive method begins with all patterns in a single cluster and performs splitting until a stopping criterion is met. In this article, as the density of data in feature space is relatively sparse, we utilize both the *K*-means and the agglomerative clustering algorithms.

The *K*-means clustering algorithm, also called the Lloyd or Lloyd-Forgy algorithm (Lloyd [1957] 1982; Forgy 1965), constructs K nearly equal-variance clusters C, using N samples X by minimizing the intracluster sum of squares of the Euclidean distances between each member x_i and the centroid or mean μ_j of each cluster j; for example, $\sum_{i=0}^{N}\min_{\mu_j \in C}\|x_i - \mu_j\|^2$ (an objective function). The *K*-means algorithm proceeds in three stages: (1) either random or guided assignment of initial μ_j^t per cluster at $t = 0$; (2) association of each x_i with a μ_j^t at each iteration t, such that the objective function is minimized; for example, $k_j^t = x_i : \|x_i - \mu_j^t\|^2 \leq \|x_i - \mu_k^t\|^2 \; \forall \, k, \; 1 \leq k \leq N$; (3) assignment of each μ_j^t to be the centroid of the jth cluster; for example, $\mu_j^{t+1} = (1/|k_j^t|)\sum_{[x_i \in k_j^t]} x_i$. Convergence is achieved when the centroids become effectively stationary under repetitions of stages 2 and 3. This algorithm makes convex and isotropic assumptions about the clusters because its objective function employs the Euclidean metric. Since this algorithm effectively partitions the feature space regardless of data density, it is suitable for low-data-density applications. The number of clusters is a free hyperparameter in this algorithm. This algorithm is aware of the geometry of the feature space, and by the choice of the Euclidean distance metric, the geometry should ideally be Euclidean.

The agglomerative clustering algorithm builds a hierarchy of clusters by successively merging existing clusters (which can be singletons), by minimizing an objective function and a linkage criterion, the former in a manner similar to the *K*-means algorithm by also employing a metric (Rokach 2005). However, unlike the *K*-means algorithm, agglomerative clustering can make use of any metric, including non-Euclidean metrics. The linkage criteria quantify the dissimilarity of sets of data points as a function of their distances, which is evaluated by the employed metric. A common linkage criterion is Ward's method (Ward 1963), which minimizes the total intracluster variance over all clusters. Algorithmically, during each iteration, pairs of clusters that lead to the smallest increase in total intracluster variance after merging are merged. This process is repeated until only a single cluster remains. The results of this process can be visualized as a tree of cluster members at each step, which is called a dendrogram. The tree can be cut at any depth to produce a desired number of clusters. Either the number of clusters or the distance threshold are free parameters in this algorithm. This has been a popular algorithm in the discipline of crystallography since the development of X-ray powder diffraction (XRD) to process mineral crystallographic phases by clustering the resulting data. In limited applications, hierarchical clustering, in combination with principal component analysis (PCA), has been used to combine XRD mineralogical analyses and X-ray fluorescence chemical analyses for robust clustering of geological samples. To our knowledge, no study has been done in the past to attempt classification of rock units, particularly Archean shales, using clustering algorithms. This algorithm is not necessarily aware of the feature space geometry, because of its flexibility in the choice of metrics.

There are various methods to determine the appropriate number of clusters and the effectiveness of algorithms. In this article, we use the average-silhouette method, which computes the average silhouette scores for a range of the number of clusters (Kaufman and Rousseeuw 1990). This method uses the intra- and intercluster distance of all data points and computes a number between −1 and 1 for a clustering configuration, such that values near 1 indicate that samples are far away from other clusters and therefore that the intercluster distance is maximized. The highest silhouette score indicates the most appropriate choice for the number of clusters, and other high scores represent alternative configurations. Silhouette scores are generally higher for convex clusters.

Supervised classification algorithms automate the deduction of relationships between the data's features and the class labels (Russell and Norvig 2010). The trained model is then used to predict labels for a population in an inductive manner. There are many supervised classification algorithms, including but not limited to those used in this article: support vector machine, logistic regression, naïve Bayes, decision trees, *k*-nearest neighbors, random forest, AdaBoost, neural networks, and Gaussian processes. The choice of algorithm is

based on many factors, such as computation time; data density including feature space density; bias-variance trade-off; function complexity; feature space dimensionality; input and prediction noise; and feature interactions. In many cases, a trial-and-error approach is used to select the final prediction algorithm through cross validation. The prediction of a label in classification is discontinuous, meaning that each sample definitively belongs to one class. Prediction error can be decomposed into three main components: bias, variance, and noise. The bias of the model is the tendency of the model to default to some class label. The variance of the model gauges the relative change in the model output, given a relative change in the model's input. The noise is the unavoidable component of the prediction error that is neither bias nor variance. The total prediction error is the root-mean-square sum of the three sources of errors. Various algorithms generally exhibit different behaviors along these sources of errors, and in some cases, it might be desirable to trade some bias for a correspondingly larger reduction in overall variance for a particular algorithm in a given scenario.

The k-nearest-neighbors algorithm (KNN; Fix and Hodges [1951] 1989; Cover and Hart 1967) uses k training samples (a hyperparameter of the model) to locally construct a consensus in feature space that is used to estimate unknown targets that fall within the locality in feature space (Witten and Frank 2005; Kotsiantis 2007). An optimal value for k can be determined by cross validation. Excessively large values of k may lead to model overfitting (Hastie et al. 2009), which increases the prediction variance. The distance metric employed in the KNN algorithm is not necessarily Euclidean; however, where it is employed, the algorithm assumes a Euclidean geometry in the feature space. Other metrics, such as the Minkowski metric, are also available.

The support vector machine (SVM) algorithm (Vapnik 1998) is similar to other Euclidean-metric regression algorithms and is typically used to define nonlinear decision boundaries or models in high-dimensional variable space (Hsu and Lin 2002; Karatzoglou et al. 2006). The SVM algorithm maximizes the Euclidean distance between training samples that are closest to the decision or regression hyperplane, which are known as support vectors. This is mathematically formulated into an objective function, or loss function, that measures the amount of margin, which is the sum of the Euclidean distances between the support vectors and the hyperplane. A hyperparameter C defines a penalty for misclassifying support vectors. Increasing C promotes an increasingly more complex hyperplane that eventually becomes prone to overfitting (which increases model prediction variance and decreases bias). Another parameter in SVM is ε, which specifies a boundary region around the hyperplane such that no penalty in the loss function is associated with points predicted within this region of the actual value. As SVM uses only support vectors, it can automatically ignore some outliers. By choice of the Euclidean metric, this algorithm assumes Euclidean geometry for the feature space.

Decision trees are flowchart-like tree structures that partition the trees recursively. Internal nodes represent features, branches represent decision rules, and each leaf represents an outcome. This algorithm learns to partition the data on the basis of feature values. The flowchart-like structure is easy to interpret and visualize. The decision rule to split a node is based on a metric to maximize some notion of difference between the resulting leaves. There are numerous metrics to measure leaf difference. For example, the Gini impurity measures how often a random node from the tree would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the branch's subset. It is a form of an information entropy measure. For N class labels, the probability of selecting a data point with label i, $1 \le i \le N$, is $p(i)$. Therefore, the Gini impurity (G) of a set of training data can be given by $G = \sum_{i=1}^{N} p(i) \cdot [1 - p(i)]$. For a binary classification problem, the worst Gini impurity at a split is 0.5 (randomly assigned leaves) and the best is 1.0 (perfectly split leaves). The objective for picking a particular split is then to maximize the difference of the Gini impurity of the current split configuration against a completely random split. Similarly, a different criterion to split nodes is based on entropy of information, which is measured similarly to the Gini impurity. The entropy criterion is given by $E = \sum_{i=1}^{N} p(i) \cdot \log(p_i)$. The depth of the tree is a hyperparameter. Decision trees are weak classifiers, in the sense that they classify above chance but not substantially. It is possible to convert a weak classifier into a strong one by using various statistical approaches, such as ensemble methods (Freund and Schapire 1995; Ho 1995; Breiman 1996; Kotsiantis 2014), provided that each individual weak classifier is better than a random guess. Random forests are a type of bagged decision tree that mitigates the noise sensitivity of individual trees by constructing an ensemble of trees and averaging the output, as long as the trees are not correlated (Ho 1995). The removal of correlation between individual trees is via sampling of random subsets of features (e.g., bootstrap sampling)

to build individual trees. This leads to better model performance than decision trees in general, because the model variance is reduced but without introducing additional bias. The maximum number of features per tree, the number of trees, and the minimum number of samples per split are model hyperparameters, in addition to the tree depth parameter that is inherited from decision trees. AdaBoost can use decision trees as a base weak classifier and, in this form, is a boosted decision tree that uses adaptive boosting (Freund and Schapire 1995). In this method, the output of weak classifiers is combined into a weighted sum that represents the final output. Adaptation occurs by modifying subsequent weak classifiers in favor of those instances misclassified by previous classifiers (an attempt to perform prediction correction). The rate of adaptation and the number of trees are model hyperparameters. The tree-based methods generally do not make assumptions about the geometry of the feature space.

Logistic regression is a type of binary regression that estimates the logarithm of the odds of a label, given a set of features (Cramer 2004). The probability of the sample belonging to a particular class is derived from the logarithm of the odds, using the logistic function. As the model is logarithmic in feature space, changing any of the features multiplicatively affects the odds of the outcome at a constant rate, which is unique for each feature. Nonbinary logistic regression (multinomial) is a generalization of the binary form to find the probability of the data point belonging to multiple classes, and a scheme such as a rank order can be used to identify the most probable label. For an observation i and its N-dimensional features X_i, the probability of the data point belonging to class $1 \leq l < L$ can be written as $f(l, i) = A_l \cdot X_i$, where A_l is vector of regression parameters. The probability of a data point y belonging to a particular class q, $(1 \leq q < L)$ is then $P(y = q) = e^{A_q \cdot X_i}/(1 + \sum_{j=1}^{N-1} e^{A_j \cdot X_i})$. The regression parameters are usually jointly estimated by the regularized maximum likelihood criterion that uses weights to increase model bias in an attempt to reduce model variance. The regularization strength parameter C is a model hyperparameter. Similarly to SVMs, smaller values of C lead to stronger regularization (less complex decision planes in feature space). This algorithm is geometry aware in the feature space.

Naïve Bayes is a class of classification algorithms based on conditional probability that assigns labels using features and assumes that the features are independent and contribute to the classification independently (Hastie et al. 2009). Conditional probability calculates the posterior probability of an outcome, given the prior probability, times the likelihood, divided by the evidence. Given a feature vector X_i, the probability of an outcome l within a total of L outcomes is $P(l|X_i) = P(l)P(X_i|l)/P(X_i)$. A classifier built on this probability model also requires a decision criterion, which is usually the selection of the most probable label. The assumed probability distribution of the features is a model hyperparameter, although, aside from the Gaussian distribution, other distributions are not commonly used. Naïve Bayes does not assume any feature-space geometry.

Multilayer perceptron classifier (MLP) is a class of feedforward artificial neural network, which is a collection of connected nodes (artificial neurons) that loosely resembles biological brains (Hastie et al. 2009). Connections between the neurons transmit real numbers to other neurons, and the output of each neuron is a nonlinear function of the sum of its inputs (similar to the activation potential in biological neurons). The connections and the neuron outputs are typically weighted, and the weights are adjusted through experience. Neurons activate according to some function, which may exhibit a threshold or may be linear (Hastie et al. 2009). Neurons are usually connected layer-wise, and each layer performs a different transformation on their inputs. It is possible for signals to recurrently travel the same network multiple times in other artificial neural network designs, although the feedforward designs are single pass. Artificial neural networks are universal function approximators and are extremely useful algorithms in data-rich applications such as image classification and natural language processing. To date and in an increasing number of applications, artificial neural networks are capable of surpassing human capability in a number of tasks (e.g., He et al. 2015; Lundervold and Lundervold 2019). An MLP contains a minimum of three layers of neurons—an input stage, a hidden layer, and an output layer—and because of its simplicity, it is a trivial example of an artificial neural network. Input nodes are linearly activated, while the subsequent layers are nonlinear. The supervised learning technique uses an objective function and backpropagation for model training. The objective function is any metric that evaluates the desirability of the output (e.g., its similarity to a known label). Backpropagation computes the gradient of the objective function with respect to the weights of the network for each training example by using the chain rule, iterating over each layer at a time. It allows the weights to be updated following a gradient-descent approach to

minimize the objective function (Curry 1944; Rosenblatt 1961; Rumelhart et al. 1986; Lemaréchal 2012). An MLP is capable of distinguishing data that are not linearly separable (Cybenko 1989). There are a number of hyperparameters, including the following: activation, which is the type of mathematical function used to activate the hidden and final layers and could include the identity function ($f(x) = x$); the logistic sigmoid function ($f(x) = 1/(1 + e^{-x})$); the hyperbolic tangent function ($f(x) = \tanh(x)$); the rectified linear unit function (relu; $f(x) = \max(0, x)$); the L^2-norm-based regularization parameter α, which can be tuned to balance the model bias and variance; and the learning-rate parameter, which can be constant, decreased over each time step using a power function (invscaling), and adaptive, which keeps the learning rate at the initial constant rate until the loss function ceases to decrease, at which point, the learning rate is decreased fivefold. An MLP is spatially aware in the feature space; however, the broader class of neural networks in general can make use of any type of vector space geometry (e.g., Gu et al. 2019).

Gaussian processes are a generic supervised learning method, suitable for regression and probabilistic classification problems. Essentially, Gaussian processes interpolate the observations, using various different kernels, such that the overall distribution is a joint distribution of many random variables (Rasmussen and Williams 2006). The interpolated multidimensional surface is used to measure similarity between points, which also allows unknown values to be predicted (Rasmussen and Williams 2006). The prediction output is continuous, which can be squashed through a link function to derive a probabilistic classification instead. The link function is the logistic function and provides a binary classification. Binary classifiers such as the Gaussian process classifier can be extended to support multiclass classification by treating the problem as a one-versus-rest or one-versus-one problem. In the former case, the binary classifier is used to discriminate one class from the remainder, and in the latter case, the binary classifier is fitted for all pairs of possible classes. The prediction results are combined to allow for multiclass predictions. The choice of the multiclass classification approach is a model hyperparameter. The kernel is also a model hyperparameter and can be chosen from a variety of functions, including highly flexible, nonlinear ones such as the radial basis functions (RBFs). The RBF kernel is given by $k(x_i, x_j) = e^{-|x_i, x_j|_2^2/2l^2}$, where $|x_i, x_j|_2^2$ denotes the L^2-distance of a sample pair (x_i, x_j). The parameter l is the length scale of the RBF; l and a multiplier of the kernel are the hyperparameters of this algorithm. Gaussian processes are spatially aware in the feature space.

Model selection and tuning for supervised machine-learning algorithms are usually accomplished through cross validation, which is an out-of-sample testing technique. In cross validation, the data set is split into several nonoverlapping sets, the larger of which is the training data set that is used to train the models. Then the remainder validation data set is used to profile the prediction performance of the models, and the model hyperparameters are adjusted. Subsequently, the models are retrained and revalidated to optimize the hyperparameters. Issues such as excessive model variance and selection bias are minimized through this process.

Data and Methods

Database and Geological Setting. The database refers to a total of 433 shale samples that were collected from different units of the Kaapvaal Craton (South Africa)—the Barberton, Witwatersrand, Pongola, and Transvaal Supergroups—as well as from the Zimbabwe Craton (Zimbabwe)—the Belingwe and Buhwa Greenstone Belts—a total of six supergroup-level classes. They cover various lithostratigraphic groups: the Fig Tree, Moodies, Ngezi, West Rand, Central Rand, Moozaan, Nsuze, and Pretoria Groups and the Black Reef Group and an unknown group within the Buhwa Greenstone Belt (a total of 10 group-level classes). At the formation level, there are a total of 17 classes. See table 1 for lists of all classes and the number of samples per class, and see the supplementary table (available online) for the raw data. Each class was sampled multiple times at several locations. The major-element oxides analyzed for in each sample are SiO_2, Al_2O_3, Fe_2O_3, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, and MnO, as well as loss on ignition; Fe_2O_3 refers to total Fe.

Machine-Learning Workflow. For unsupervised algorithms, we employ a two-stage workflow that includes (1) data preprocessing and cleaning and (2) algorithm hyperparameter selection. For supervised algorithms, we employ a three-stage workflow that includes (1) data preprocessing and cleaning, (2) algorithm hyperparameter selection, and (3) algorithm performance assessment. The data-processing and cleaning stage is necessary to ensure that the data set is suitable for machine-learning algorithms. As the data come from various

Table 1. Stratigraphic Classification Level and Number of Samples

Formations		Groups		Supergroups	
Name	Samples	Name	Samples	Name	Samples
Sheba	37	West Rand	133	Witwatersrand	282
Clutha	40	Central Rand	149	Barberton	77
Zeederbergs*	3	Fig Tree	37	Belingwe Greenstone Belt	15
Cheshire	12	Pretoria	30	Buhwa Greenstone Belt	20
Orange Grove*	1	Moodies	40	Transvaal	30
Parktown*	4	Ngezi	15	Pongola*	5
Promise*	3	Unknown (Buhwa)	20		
Parktown-Brixton*	4				
Palmietfontein	12				
Roodepoort	109				
Booysens	144				
Kimberley*	5				
Black Reef*	4				
Silverton	30				
Unknown (Buhwa)	20				

Note. Features with an asterisk are not used for predictions at their levels because of the sparsity of samples.

sources, they reflect internal inconsistencies, which must be cleaned before the next stage. For example, the number of samples from a given formation might be too low, or analyses for certain chemical components might be at or below the analytical detection limit. Any formation, group, or supergroup that contains fewer than 10 samples was excluded from all analyses unless otherwise indicated for special purposes. For example, the Orange Grove Formation has only a single sample. However, sparse formation data are retained for group-level predictions, and the same applies for sparse group data at supergroup-level predictions. This results in three separate data sets for the formations, groups, and supergroups that are mostly overlapping. As the various data used here originally had not been intended to be used for machine learning, the coverage of different geological units is highly variable, and it is impossible to avoid bias, aside from removing sparse-data units. Bias in predictive modeling refers to the tendency for the machine to default predictions to certain labels with more data points. Although our data set is not optimal for machine learning, it is typical of the state of geochemical data coverage in much of geosciences, by our experience. It is also highly desirable to be able to predict geological units using variable data coverage, as this is an inevitable and frequently encountered situation in geosciences. All samples available for this study are summarized in table 1. The remainder of the data consists of class labels (supergroups, groups, and formations) and geochemical information on the samples, which are used as machine-learning features. Chemical analyses that were below the detection limit (which were very rare) were imputed with the KNN imputation algorithm. The imputation is verified to be acceptable if the closure of the data is unaffected, since the imputation was carried out over the rock-forming elements. In this manner, all imputed data points (25 in total) were satisfactory, with maximal deviations of about 1%. Since the data are compositional, the simplex geometry (Aitchison 1982) is the native vector space geometry. The use of compositional data usually occurs outside of its native vector space, through a choice of log ratio transformations on the data, such that the resulting vector space is Euclidean (Aitchison et al. 2000). For machine-learning algorithms, if the algorithms do not explicitly require a Euclidean distance metric or other properties of the Euclidean geometry (e.g., linear transformations), then embedding the data in its native vector space is unproblematic. We choose to both transform and use the raw compositional data for comparison purposes. Transformation for the major-element data uses centered log ratio (CLR) transformation (Aitchison 1982), which allows the use of the Euclidean distance metric, with additional properties that include scale invariance, perturbation invariance, permutation invariance, and subcompositional dominance (Aitchison et al. 2000). The CLR transformation seems to have been used with excellent success when followed by traditional multivariate geochemical data analysis routines, such as PCA (Grunsky et al. 2014; Harris et al. 2015; Chen et al. 2018; Grunsky and de Caritat 2019). In the CLR transformation, all compositions $x_i \ldots x_D$ are divided by the geometric mean of the vector, that is,

$$g(x) = \sqrt[D]{x_i \ldots x_D},$$

and, subsequently, a logarithm is taken of the ratios, that is, for sample x_j, $\mathrm{CLR}(x_j) = [\ln(x_{i,j}/g(x_j)); \ldots; \ln(x_{D,j}/g(x_j))]$. Some algorithms are sensitive to unequal feature scales, and therefore all features were rescaled to span an equal range between 0 and 1, which ensures that they influence the machine-learning algorithms equally. Note that this is not a requirement for non–distance-based algorithms, such as most tree-based methods (e.g., random forest, decision trees, and AdaBoost). For unsupervised machine-learning algorithms, the algorithm and hyperparameter selection process consists of a feedback-driven method that increments the number of clusters, determines the silhouette scores for all the algorithms, and repeats until a maximum number of clusters has been reached. This is a user-selectable parameter. The range of silhouette scores is then examined, and if a global maximum is found, it is adopted as the number of clusters for a strictly chemostratigraphic classification.

For algorithm selection and hyperparameter tuning in supervised machine-learning algorithms, we use an automated approach that consists of a randomized tenfold cross validation combined with a grid search. In this scheme, the entire data set is divided into two sets randomly for (1) training and cross validation and (2) testing. The leave-one-out cross-validation method sequentially predicts all samples during cross validation, with the remainder of the data for training. This method uses data more efficiently, compared to any variety of the K-fold cross validation, which divides the data into K equal-sized partitions, as more samples are allocated for training. However, it is computationally inefficient for large data sets. A majority of the data (90%) was used for training and cross validation. The remainder was reserved for model testing. The parameters used for each of the algorithms are listed in table 2.

The grid search for each algorithm uses the cross-validation score of each run to choose the most optimal combination of parameters on a per-algorithm basis. Subsequently, we selected a few of the most accurate algorithms for prediction of the testing data. In this case, accuracy was measured by the number of samples predicted correctly, divided by the sample size. To test the algorithm, we retrained the selected algorithm, using the training data set and the best hyperparameters, then tested the model, using the reserved test data set. The accuracy was reported, and the prediction performance was visualized using results derived from the test data set.

Results

Unsupervised Stratigraphic Classification of Shales. Silhouette scores were used to analyze clustering results of the K-means and agglomerative algorithms for 2–30 clusters using the CLR-transformed data (fig. 1). Both algorithms were unable to produce global maxima in the silhouette score using raw data. It is possible to employ dimensional reduction techniques to trade some loss of information for a massive increase in data density in feature space by projecting the 11 chemical dimensions into fewer transformed dimensions. It is also

Table 2. Algorithms and Their Hyperparameters Used in the Grid Search

Algorithm	Parameter grid
k-nearest-neighbors	Number of neighbors = {1, 2, 4, 6}
Support vector machine	C = {10, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 750, 1000}; ε = {.00001, .0001, .001, .01, .1, .5, 1.0}
Decision tree	Maximum depth = {6, 4, 2, 1, 0}
Random forests	Ensemble size = 1000; maximum depth = {7, 6, 5, 4, 3, 2, 1, unlimited}; maximum no. of features = {1, 2, 3, 4, 5, 6}; minimum no. of samples for a split = {2, 3, 4}; minimum no. of samples for a leaf = {1, 2, 3, 4, 5}; split criterion = {Gini impurity, entropy}
AdaBoost	Learning rate = 1; no. of classifiers = 100; base classifier = decision tree with maximum depth = {6, 4, 2, 1, 0}
Logistic regression	C = {.1, 1, 5, 10}
Naïve Bayes	None
Neuronet (multilayer perceptron classifier)	α = {.0001, .001, .01, .1, 1.0}; activation = {identity, logistic, tanh, relu}; learning rate = {constant, inverse scaling, adaptive}
Gaussian process	RBF kernel multipliers = continuous range between .6 and 1.0; kernel length scales = {.8, 1.0, 1.2}; method for predicting multiple classes = {one vs. rest, one vs. one}

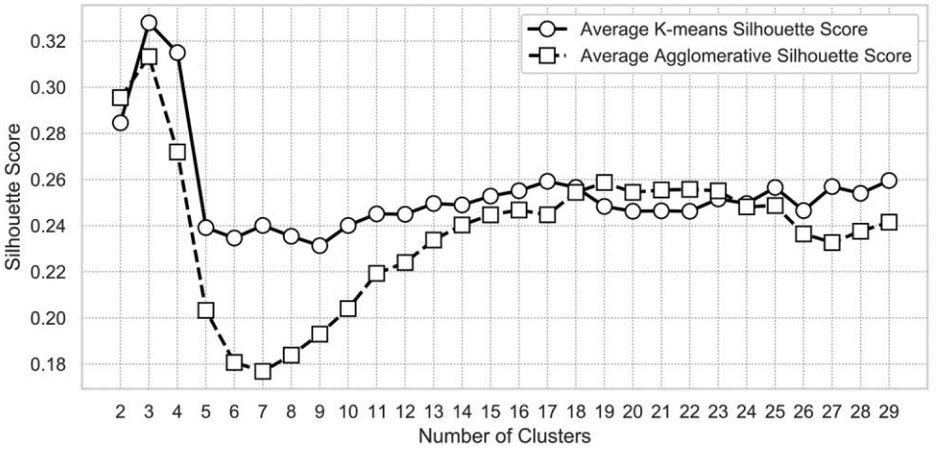

Figure 1. Silhouette score versus number of clusters. The results are averaged using the formation, group, and supergroup features.

possible to increase cluster convexity through dimensional reduction; however, this is not guaranteed. In our case, we use PCA, which is a type of dimensional reduction technique that defines an orthogonal coordinate system up to a specified number of components that best capture the variability of the data; PCA is a common traditional choice to process high-dimensional geochemical data (e.g., Grunsky et al. 2014; Gazley et al. 2015; Harris et al. 2015; Chen et al. 2018; Grunsky and de Caritat 2019). The number of principal components is determined by balancing the density of data against the explanatory power of the resulting components. Clusters can become more convex through PCA-based dimensional reduction (e.g., a quadratic loss of vertices for cubic clusters with a linear reduction in the number of dimensions), because dimensional reduction generally reduces the geometric complexity of the original data cloud. At five components, the transformed data are capable of producing global peaks in the silhouette score sweep. These components yield a total explained variance of 90%, 90%, and 89% at the formation, group, and supergroup levels, respectively. A caveat here is that for shales, the interpretation of component loadings may be complicated by complex chemical stoichiometry. However, this is not a problem for our data-driven application, as we do not require any manual interpretation. To measure silhouette scores that are meaningful for the clusters' chemical separation, while the clustering is performed on PCA-transformed features, the scoring is performed on the original features. The results of clustering on PCA-transformed features are qualitatively similar to those for non–PCA-transformed features (fig. 2; compare with fig. 1). The optimal choice, on average, appears to be either three or four clusters, using either the *K*-means or the agglomerative algorithm. At three clusters using the *K*-means algorithm, the intercluster chemical separation is visibly substantial and highly multidimensional (fig. 3). The generally lower silhouette score of the agglomerative clustering results, compared to the *K*-means clustering results, indicates that agglomerative clustering is less able to create effective classifications using our data. In the context of shale classification, the choices of the number of clusters could be understood as different classes of samples by chemostratigraphy in a manner similar to, but not identical with, that of lithostratigraphic units.

It is interesting to understand the degree of conformity between a chemostratigraphic classification and the geologically derived lithostratigraphic classification (formations, groups, and supergroups). To determine the level of conformity, the formation, group, and supergroup labels of the samples are compared with the unsupervised classification labels, using the adjusted Rand index or score. This scoring scheme compares two categorizations of a set of objects, such that if the objects are consistently considered distinct by class by up to a permutation of classes, then the score reaches its maximum of 1. Randomly classified objects produce scores close to 0. Therefore, a perfect conformity of the chemostratigraphy to the lithostratigraphy classes should yield a score of 1 (up to a permutation of labels). Results for the adjusted Rand score (figs. 4, 5) indicate that the best match between the unsupervised clustering algorithms, regardless of the algorithm or the use of PCA, occurs at the formation and group levels. The lithogeochemical

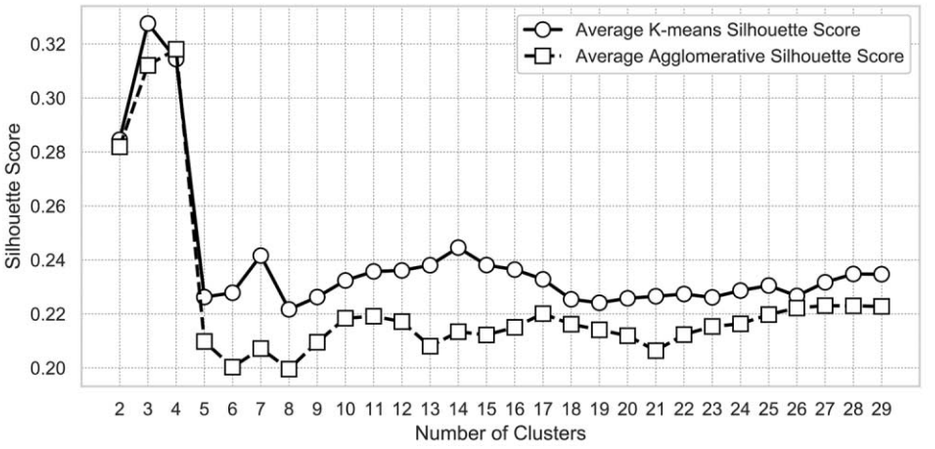

Figure 2. Silhouette score versus number of clusters. The first five principal components were used for the clustering algorithm. The results are averaged using the formation, group, and supergroup features.

overlap at the supergroup level seems to be substantially greater than that at the group and formation levels. Therefore, because the unsupervised clustering attempts to divide samples chemically, the level of conformity is low where the chemical overlap is large. In general, clustering results at the formation level exhibit the highest adjusted Rand score, which indicates that chemical separation at the formation level is usually the highest, and if lithogeochemistry was the sole basis for stratigraphic classification of shales, then the classification should occur at the formation level. The peaks in the adjusted Rand score of both algorithms occur at eight or more clusters (figs. 4, 5), more than the number of formations after data preprocessing within the data set. Non–PCA-transformed features seem to be better at generating clusters that conform best to geological classifications, and neither algorithm seems to be substantially better than the other. However, since the score is not very close to 1 at any number of clusters, the chemical composition of the chemostratigraphic classes does not fully conform to the chemical compositions of the shales at any classification level. This is not a surprise, because lithostratigraphic units of shales are usually defined using many lines of evidence that far exceed chemical properties.

Supervised Stratigraphic Classification of Shales. For supervised machine-learning algorithms, there are three sets of labels: formation, group, and supergroup, and two sets of features: CLR-transformed and raw data. Algorithm selection and hyperparameter tuning produced model parameters that are comparable across the three sets of labels for each algorithm. However, the cross-validation accuracy differs. The results for the group-level cross validation are summarized in figure 6. Prediction results using the PCA-transformed features are systematically worse than those with original features (a difference of about the order of 10%) for raw data, and for the CLR-transformed features, the difference is usually less. Therefore, the remainder of the results do not utilize the PCA-transformed features.

The most accurate algorithms are Gaussian process and SVM, although the tree-based methods, such as AdaBoost and random forest, as well as the MLP algorithm (neuronet), are close (fig. 6). For example, a trained SVM or Gaussian process classifier with optimized hyperparameters predicts lithostratigraphic classes with excellent accuracy (>80%–90%) at the formation, group, and supergroup levels (fig. 6). Using the same workflow, we empirically evaluate the effect of the CLR transformation on predictive modeling. For this comparison, two parallel workflows are created that differed only in the use of CLR transformation—for one workflow, the features are raw data, and the other uses CLR-transformed data. The algorithm selection cross-validation scores for the best algorithm using raw data are identical to those for the CLR-transformed data at the supergroup level. However, there is a slight reduction of a few percent for the scores using raw data compared to those for the CLR-transformed data at the other geological levels. The final testing scores of 50 random train-test splits using the raw data are 0.89 (formation level), 0.90 (group level), and 0.94 (supergroup level), whereas for the CLR-transformed data, they are 0.83 (formation level), 0.90 (group level), and 0.94 (supergroup level). The differences between the 50 individual runs are shown in figure 7. Confusion matrices for 50 runs using raw data in the

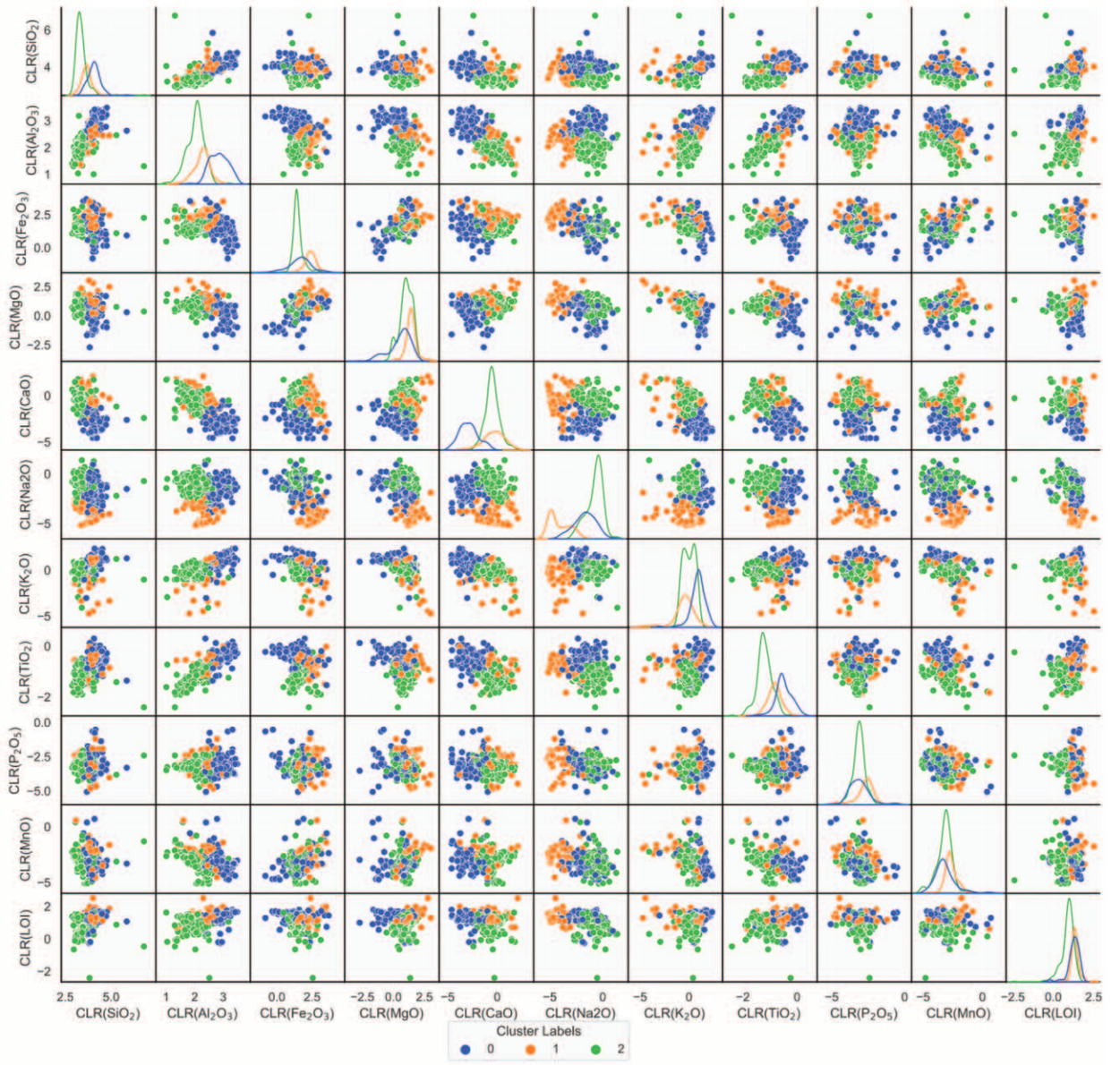

Figure 3. Scatter matrix of major-element chemistry of shale samples. Clusters are depicted as sample coloration. The upper diagonal features kernel density estimates of the data points on a per-cluster basis to depict overlap between the clusters. CLR = centered log ratio.

testing data set for each of the levels are shown in figures 8, 9 and 10 (the matrices are qualitatively similar for the CLR-transformed data). In general, predictions at the supergroup level are the most accurate, and the least accurate predictions occur at the formation level. As chemical differentiation appears to be largest at the formation level and smallest at the supergroup level, the cause of the variations in prediction accuracy at various levels is not insufficient chemical separation. The confusion matrices (figs. 8–10) at each level reveal that the most accurate predictions correlate with shale classifications that contain the highest number of samples (e.g., the Booysens and Roodepoort Formations and the West Rand and Central Rand Groups). An examination of prediction performance as a function of the number of samples clearly shows that, regardless of the level of shale classification or the algorithm used, the overall performance is strongly dependent on the number of available samples for a particular label in the data set (fig. 11). This also explains the general loss of

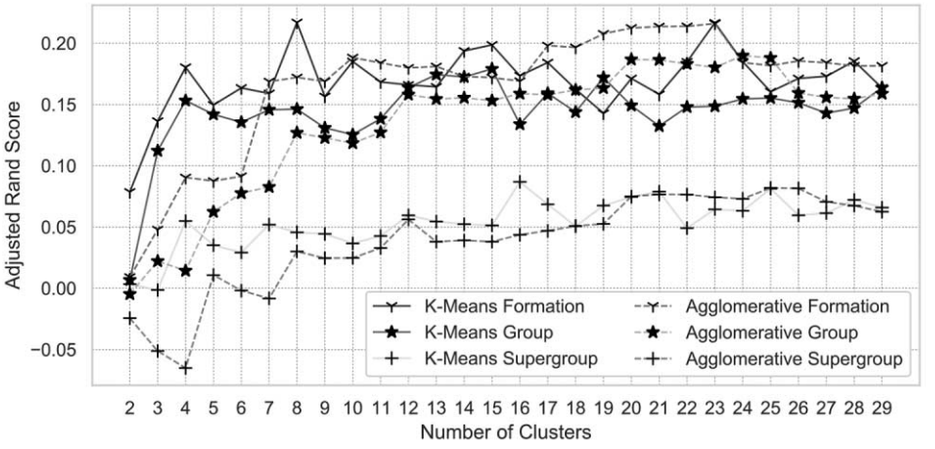

Figure 4. Adjusted Rand score of unsupervised clustering results at the formation, group, and supergroup levels. A color version of this figure is available online.

prediction accuracy at more granular levels of shale classification, because the density of data diminishes as the number of classes increases (e.g., from the group to the formation level). However, the remarkable finding here is that shales, which are not traditionally classified by chemical composition alone and despite extensive chemical overlap at the formation, group, and supergroup levels, are highly predictable across all levels using only major-element concentration data. There is some prediction bias toward labels that contain the highest number of data points, such as the Witwatersrand Supergroup (fig. 10). Prediction bias is strongest at the supergroup level and weakest at the formation level (figs. 8–10), as the number of data points per label generally becomes less disparate with finer class divisions. In the context of an aggregation of historical data and the typical variability of geoscientific data, this is unavoidable, as eliminating relatively sparser labels or data points might be the best data science practice but is not a suitable geoscientific practice, especially for sample identification and mapping. However, with larger databases, for example, produced through accumulation, this may no longer be an issue.

Discussion

Chemostratigraphic Classification Schemes for Shale Using Unsupervised Machine Learning. Clustering results of chemical data for different shale units indicate that there are more geological classes than

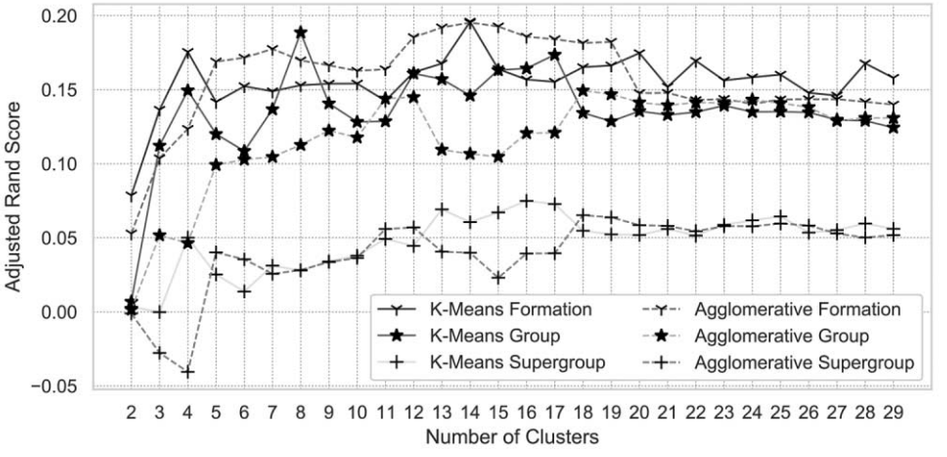

Figure 5. Adjusted Rand score of unsupervised clustering results of the five principal components of the features at the formation, group, and supergroup levels. A color version of this figure is available online.

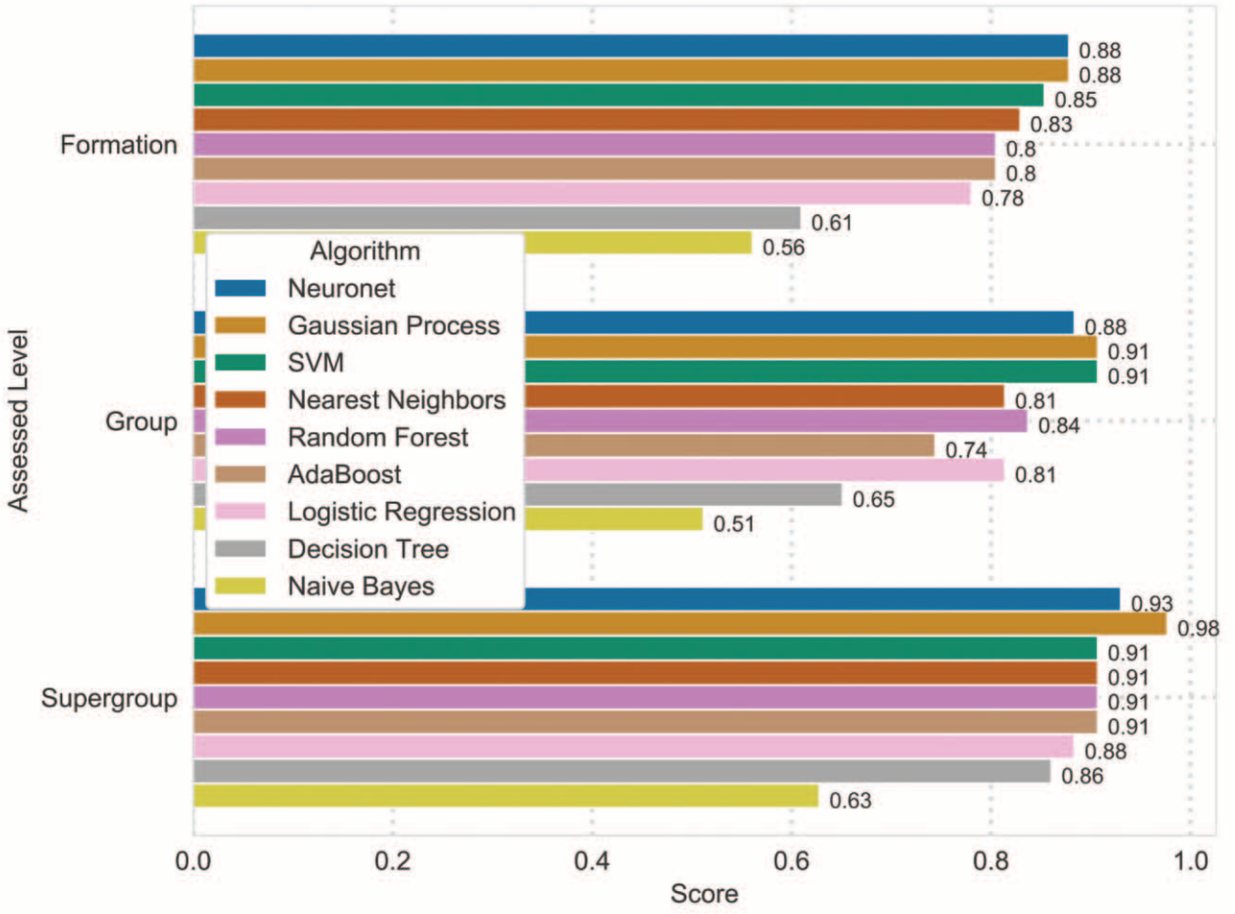

Figure 6. Algorithm selection results at all levels, with centered log ratio–transformed data.

clusters (three or four, depending on the algorithm used), which implies that the chemistry of different types of shale exhibits a substantial amount of overlap. Each cluster also exhibits intracluster variability. This immediately supports the idea that shale samples are not spatially or temporally homogeneous by chemistry at the observational scale represented by the data. For smaller data sets, it is reasonable to expect that the variability was too great to be adequately captured by the sampling methodology. However, for larger data sets, systematic differences are unlikely to be coincidental. Consequently, the causal geological processes must differ even within the same formation. Through a comparison of the *K*-means and agglomerative clustering results, it seems that *K*-means is better than agglomerative clustering at generating classes that are more chemically distinct. However, the match of either algorithm with geological labels is poor in general, and hence lithostratigraphic classifications do not conform well to data-driven chemostratigraphic classifications for shale in our data set. The greatest conformity between a purely chemostratigraphic and the existing lithostratigraphic classification schemes occurs at the level for which the chemical variability is the greatest, which is the formation level. The conformity decreases at the group and supergroup levels, as the chemical uniqueness at these levels diminishes. However, the separation of various clusters in chemical coordinates is interesting, as the greatest extent of chemical separation occurs only with a chemostratigraphic classification. Chemostratigraphic classification does not consider other characteristics of shale that would be considered in a lithostratigraphic classification, and thus it is clear that major-element concentrations are not always a discriminating factor between shale units at any geological classification level; however, it is clearly useful. Finally, it may be that some geologically derived class boundaries between rocks are not justifiable chemically, given the data in our database, although this would require additional hypothesis testing to verify. Hence, semisupervised learning may be a practical approach

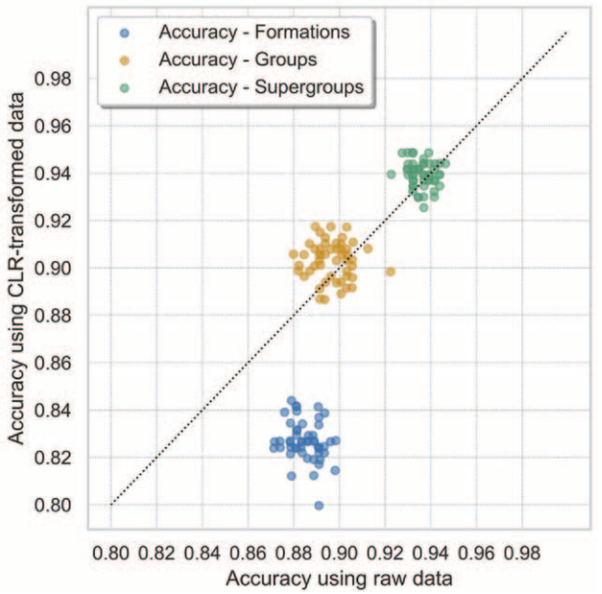

Figure 7. Comparison of results using centered log ratio (CLR)–transformed data versus raw data. The best algorithms for the CLR-transformed data were neuronet (formation level) and Gaussian process (group and supergroup levels). For the raw data, the best algorithms were support vector machine (SVM; formation and group levels) and Gaussian process (supergroup level). The 45-degree line depicts no difference in performance.

to sample classification and prediction if a stricter adherence to geological classification is desirable. Although the lack of significant conformity between the geological classification of shale and a purely chemical classification scheme (using major-element concentration data and unsupervised machine learning) may seem to suggest difficulties in the machine's ability to predict the geological classification of samples using only major-element concentration data, this is not the case.

Stratigraphic Prediction Using Supervised Machine Learning. Supervised machine-learning predictions of the shale samples at the formation, group, and supergroup levels were excellent despite significant multidimensional chemical overlap between geological classes. The accuracy generally declines with increasing granularity in the classification (e.g., from supergroups to groups) and is indicative of increasing data sparsity with an increasing number of labels. This implies that the performance of the predictions, especially at more granular levels of classification (e.g., formations) would most probably improve with more data. A visual comparison between the number of total samples per classification level and the mean prediction accuracy demonstrates a clear divide in the performance of the predictions at roughly 50 total samples, regardless of the assessed level of geological classification and the algorithm used (fig. 11). The heavily nonlinear behavior of the relationship between the number of training samples and the accuracy is reminiscent of a soft threshold, where a sharp increase in prediction accuracy occurs for geological classifications that contain over roughly 50 samples (fig. 11).

At classification levels less granular than the formation, aggregation of samples from various formations into groups and supergroups would imply that minimal additional samples are generally needed at these levels to reach high levels of prediction accuracy. The results at the group level are particularly interesting, especially for well-sampled units such as those of the Witwatersrand Supergroup, because there are almost an equal number of samples available for the Central Rand and the West Rand Groups, which constitute the bulk of the samples in this study. Indeed, all the studied shales, regardless of the supergroup, are easily separable using major-element data and are predicted with high levels of accuracy. This is interesting scientifically, as it indicates that the lithogeochemical contrast of the studied shales supports existing detrital zircon age spectra of the coarser-grained rocks and field-based evidence (Bickle and Nisbet 1993; Kositcin and Krapež 2004; Koglin et al. 2010; Toulkeridis et al. 2015). From the results obtained in this study, we infer that any formation hypothesis that can adequately model and predict the geochemistry of the rocks could exhibit a strong overall predictive quality and would likely be very useful in unraveling the geological history of rocks. From the success with the studied shales, we deduce that the same machine-learning method can be used to discriminate rocks from other cratons. In addition, the ability to predictively discriminate between various geological units and assess the discrimination performance as a function of the number of samples provides a notion of the practicality of the degree of chemical uniqueness of the units. In this case, it is possible to relate the degree of chemical separation between various units to their prediction performance. This is a quantitative and formalized method that is similar to the human's ability to discriminate between objects; where the characteristics are more different, the discrimination is more likely to be successful. Furthermore, it is clear that there exist a sufficient number of high-dimensional patterns within major-element (technically, major- and minor-element) concentration data that can be used to predict the geological classes of the shales.

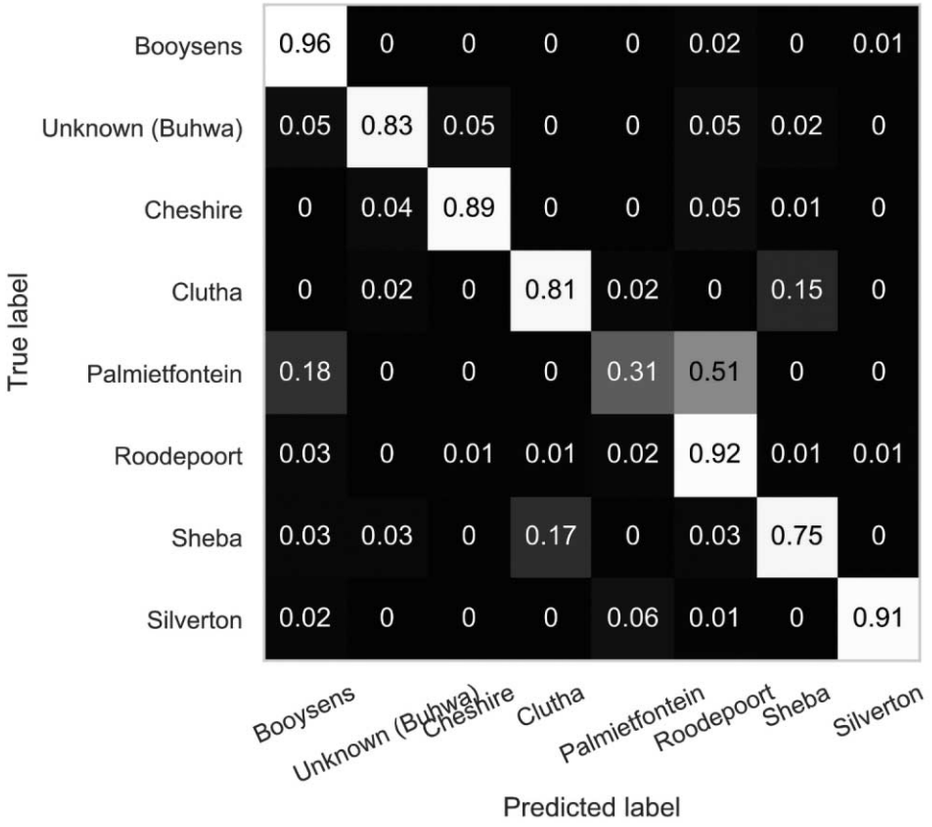

Figure 8. Confusion matrix at the formation level, using the support vector machine algorithm. A color version of this figure is available online.

The addition of more dimensions of data, such as trace-element concentrations and isotope ratios, might appear to be useful at first glance; however, this is generally strongly dependent on the density of samples available. To achieve the same sample density in feature space with each increasing dimension, the amount of data must be increased exponentially. Thus, to achieve a reasonable sample density in feature space, data from multiple dimensions must be reduced into a smaller number of dimensions through dimensional reduction and/or feature selection techniques, with some loss of total information. In addition, for legacy geochemical data, issues such as data leveling and pervasive missing data for less frequently analyzed elements create serious challenges for data preprocessing. Discipline-specific feature engineering, such as crafting chemical discrimination criteria from multiple elements (e.g., elemental or isotopic ratios), would be highly useful in an environment that hybridizes the hypothesis- and data-driven approaches to insight generation. In this case, knowledge that is useful to the discipline is leveraged to craft discipline-specific features from the original data to provide a correspondingly discipline-specific application of classification and prediction, through either unsupervised or supervised machine-learning approaches. This hybrid approach might be unavoidable at present for those willing to adopt machine learning in geosciences, given that most of the geochemical databases that were intended for hypothesis testing are likely legacy and nonprimary and that issues such as data abundance, comparability, and quality create data-repurposing challenges. However, employing this approach reduces the amount of interesting and particularly unexpected insights that might be present in the data, as our knowledge of desirable discrimination criteria also biases the insights that we can discover toward the already known in the discipline. In any case, incorporation of extra dimensions in the feature space should be carefully balanced with potential performance loss through either sample density loss or dimensional reduction, and an actual feasible combination should be determined through cross validation. The use of raw compositional data

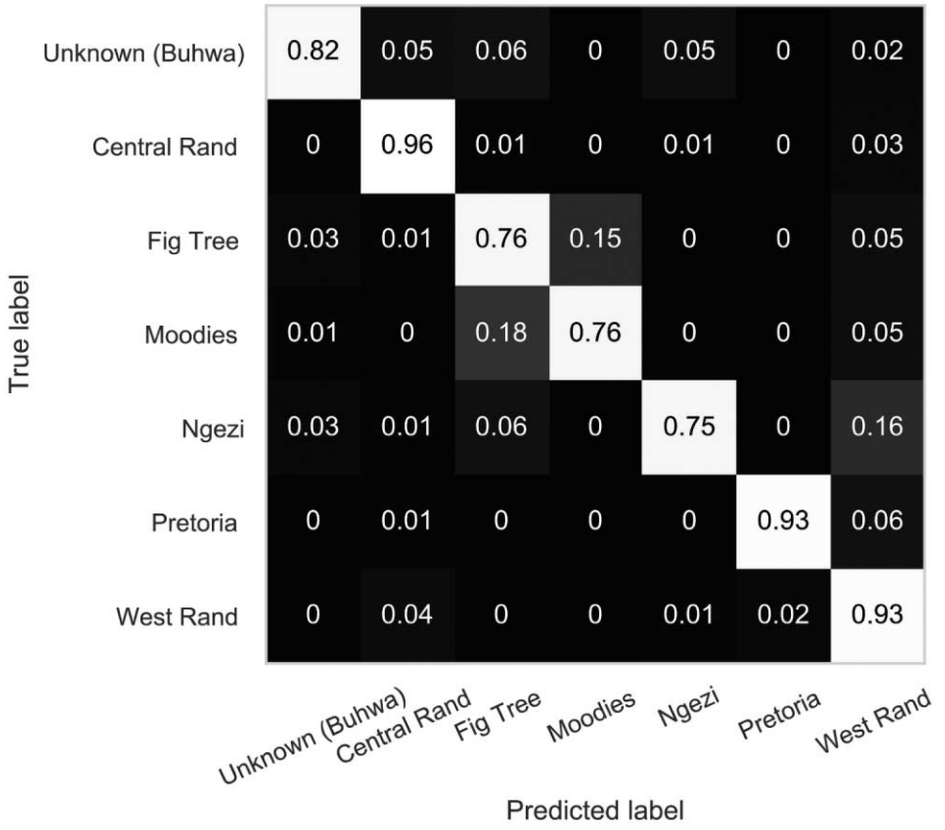

Figure 9. Confusion matrix at the group level, using the support vector machine algorithm. A color version of this figure is available online.

does not generally seem to affect algorithmic performance for the range of algorithms explored. For spatially aware algorithms that we have employed in this study, this may be due to the use of nonlinear decision boundaries in our algorithms, which seem to handle spatial geometric distortions better than linear decision boundaries. Indeed, several algorithms that are known to exhibit issues using raw compositional data are based on assumptions about linear properties of the feature space (e.g., linear transformations in PCA). However, the impact for some algorithms may be negligible. From our results, even for algorithms that are spatially aware and require properties (e.g., a distance metric) that are theoretically unsatisfied by the simplex geometry of compositional data, it is important to investigate the impact of data transformations such as log-ratio transforms using formalized performance assessment tools.

The ability to differentiate between various shale units with a high level of accuracy using solely major-element chemistry is geologically surprising, because shale is generally thought to be highly differentiable not by its chemical composition but rather by geological processes. The sedimentation process is a memoryless mixing process, and therefore the final mixture at equilibrium is not diagnostic of the process itself. Our results do not necessarily suggest that this thinking is wrong; instead, the takeaway is more nuanced. For example, it is highly probable that shale-forming geological processes are not homogeneous across time and space at every scale and/or that shales are generally disequilibrium mixtures, resulting in chemical fingerprints that can be used for shale correlation. Without knowing the scale of material and therefore the chemical heterogeneity in Archean shale and assuming that shales are not chemically differentiable in general (which is a strong assumption that has to be tested), it is moot to generalize the expectation of reliable predictions to all shales. Instead, the only line of inquiry would be to examine actual data and chemical variability of shales through more testing, for example, a data-driven approach. In this sense, machine learning could be merely highly useful, and the scientific value of the derived predictive models requires additional scrutiny. In theory, it is implausible for geological classes such as shale

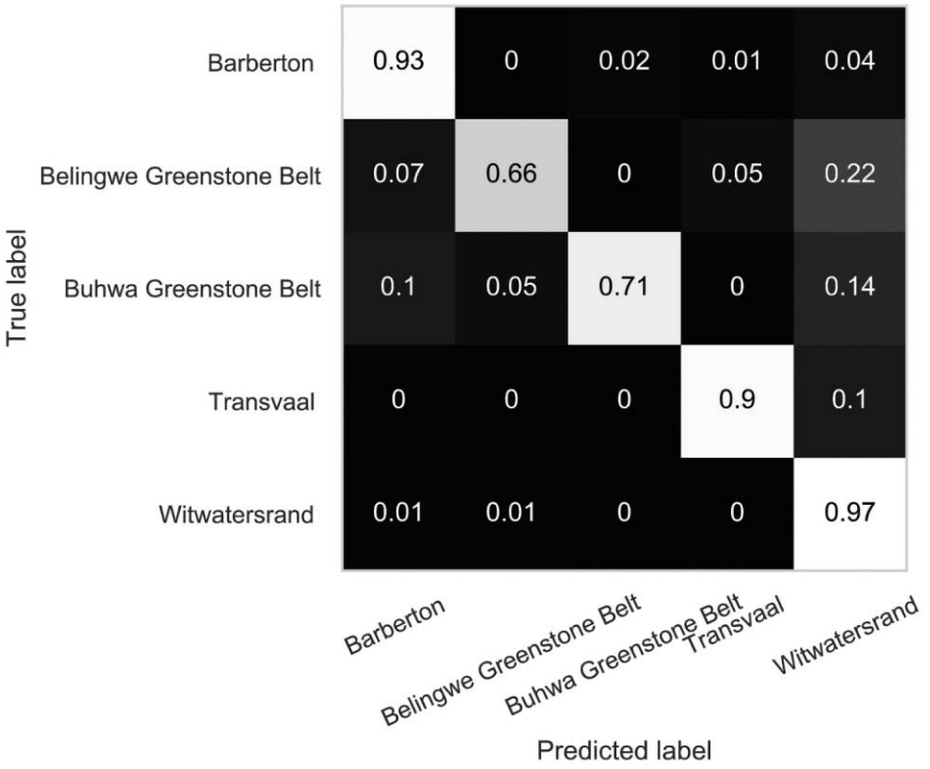

Figure 10. Confusion matrix at the supergroup level, using the Gaussian process algorithm. A color version of this figure is available online.

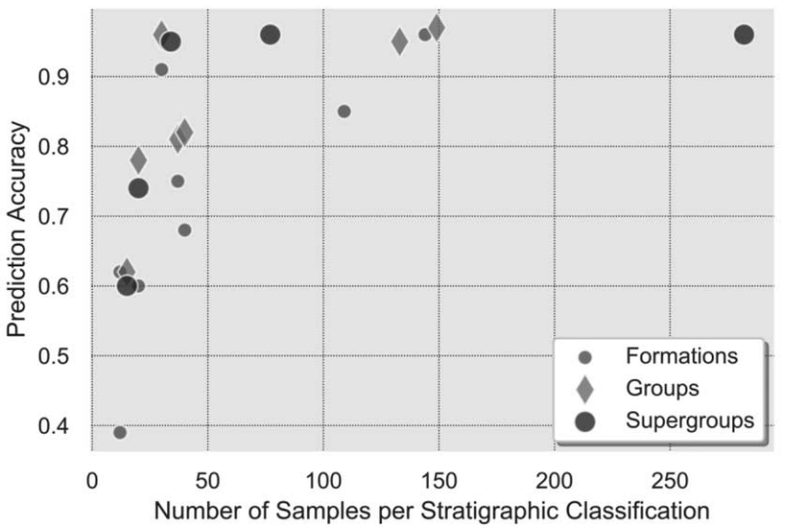

Figure 11. Prediction accuracy versus number of total samples per shale classification. Note that as the number of samples increases (e.g., more than ~50), the prediction accuracy increases, regardless of the granularity in the level of geological classification. A color version of this figure is available online.

units to have completely identical chemistry. The data-driven perspective is to ensure that for greater feature overlaps between various predicted labels, more data are available, so as to be able to reach a given level of prediction accuracy. Shale units predicted using chemistry are no different. It should be possible to distinguish between geological classes on the basis of chemistry using machine learning and, if necessary, to adopt additional lines of evidence as necessary or available with some caveats (see below). Therefore, the cost to utilize supervised machine learning to identify shale samples is rather minimal, because major-element concentrations are apparently sufficient to differentiate shale samples such as ours to a high degree of accuracy and the minimum number of samples per classification could be as low as 50. Nevertheless, data-driven approaches can lead to highly reproducible results, as they can use a large amount of data, and the addition of data generally improves the machine's ability to predict (and therefore correlate) stratigraphy.

Considerations for Data-Driven Classification and Prediction of Shale. Classification of rocks in traditional geology is becoming increasingly complex and can involve lines of evidence that range from mineralogy to texture, lithogeochemistry, isotopic composition, field relations, geophysical properties, and so on. Therefore, chemostratigraphic classification of shale is unlikely to be fully congruent with any current geological classification. There are two main methods to leverage machine learning for classification of rocks. The first is an unsupervised approach similar to the approach we used, which seeks to partition the geochemical composition of the rock samples into a number of clusters such that the intracluster variance is maximized. In this scheme, chemical overlap between various classes is minimized. However, as lithostratigraphic classification of shale units exhibits a great extent of chemical overlap, this approach indeed did not produce classes that are highly congruent with the sample's geological classes at any assessed level (formation, group, and supergroup). Nevertheless, given that geological classifications are increasingly based on higher-dimensional data and that the reasoning process strives to retain logical consistency, it is reasonable to expect that unsupervised machine learning will be increasingly applied to samples by using data that contain roughly the same types of evidence used for geological classifications, given sufficient data density. Alternatively, it is also possible to use this technique to detect previously unknown geological classifications or conclusively reject outdated and unwarranted classifications using large amounts of data. The limitation to this technique is mainly the availability of data, as increasing dimensionality (e.g., the addition of other lines of geoscientific data) necessitates an exponential increase in the amount of data required to conserve sample density in feature space. This is clearly unfeasible at present, and instead, the most likely approach would be dimensional reduction of the feature space (e.g., regular or kernel PCA, or elemental ratios), which requires additional application-specific considerations and performance analysis. The second method attempts to predict geological classes of rock samples using supervised learning. As we have demonstrated in this study, even in the case of shale classification, which is typically a complex multidimensional task and involves criteria beyond chemical composition, supervised machine-learning algorithms can accurately predict sample classification at a variety of levels using solely major-element concentration data. This is a surprisingly useful and practical result. Lower costs of chemical analyses also have a significant potential effect to allow disaggregated data to break through silos, and standardizing the type of data collected (e.g., major elements) ensures data comparability and therefore promotes reuse. This is supportive of general trends in the society toward open data, open science, open information, and citizen science. In this regard, the explanatory power and cost of individual data points are reduced in exchange for more abundant and comparable data, combined with more automated data processing and predictive modeling through machine learning. The intent is to reduce the high cost associated with geoscientific data, and therefore the threshold for entry into scientific inquiry, while simultaneously leveraging modern data-driven techniques to increase or preserve accuracy for current uses. This would permit the collection of larger amounts of data and, through the use of proper databases and data sharing, improve the readiness of the discipline to adopt digitization techniques, such as artificial intelligence, machine learning, automation, and modern data analytics.

Implication for Stratigraphic Classification and Exploration for Economic Potential. Our study has a direct implication for stratigraphic correlation in shale-rich basins, both with and without economic potential. Historically, the use of field observations, petrography, age dating, and long-distance stratigraphic correlations using geophysical methods (e.g., three-dimensional reflection seismics), lithomarkers, and biomarkers was a major paradigm for the classification of rock units and sequences. This approach appears to work well in many areas and

regions but has also led to misclassification and incorrect correlation of certain sequences (Wagener 1972; SACS 1980, 2006). In certain instances, poor stratigraphic correlation has often led to missed economic mineralization potential, for example, the Flatreef PGE deposit in the Bushveld Igneous Complex (South Africa; Grobler et al. 2019). In the Witwatersrand Supergroup, where shale horizons dominate the lower stratigraphy, several schools of thought were developed for both stratigraphic correlation of the rocks and basin-type classification. For example, even though the Witwatersrand Supergroup has long been subdivided into the more shale-dominated shallow marine to distal fluviodeltaic West Rand Group and the coarser-grained, mainly fluvial to fluviodeltaic Central Rand Group (Frimmel 2014), for many years a debate on whether the entire Witwatersrand Supergroup represents a foreland basin or a combination of a passive basin (i.e., West Rand Group) and a foreland basin (i.e., Central Rand Group) or different types of basins prevailed (Burke et al. 1986; Stanistreet and McCarthy 1991). This also led to the development of basin analysis models that regard the Witwatersrand Supergroup as a representative of either one slowly evolving basin (Burke et al. 1986; Catuneanu 2001) or an erosional remnant of stacked sediments deposited in a variety of tectonic settings ("successor basin"; Frimmel and Minter 2002). Additional evidence on the architectural setting of the Witwatersrand Supergroup were provided by Kositcin and Krapež (2004), who observed the complexity of zircon provenance age spectra and noticed tectonic differences between the lower Witwatersrand Supergroup (i.e., the West Rand Group) and the upper Witwatersrand Supergroup (i.e., the Central Rand Group). A follow-up study by Koglin et al. (2010) and a review by Frimmel (2019) supported the differences between the Central Rand and West Rand Groups, which led to the conclusion that the Witwatersrand Supergroup is best explained by a foreland basin (i.e., the Central Rand Group) superimposed on older passive-margin deposits (i.e., the West Rand Group). Given that the shales of the Witwatersrand formed under varying surface processes, as evident from the chemical variations (Nwaila and Frimmel 2019; fig. 3) and physical properties (e.g., laminated shale are interpreted to represent tidalites, while magnetic shales represent distal marine deposits), it is important to ensure that they are assigned into correct stratigraphic classes for any usage. Our results indicate that the sediment compositions of the Central Rand and West Rand Groups are clearly sufficiently distinct as to allow for their discrimination at a very high accuracy (>94%) with only a few hundred samples. Any formation hypothesis that can explain, model, and predict this difference in sediment composition in detail would be well supported by our results. Using machine-learning algorithms, our study shows that major-element concentration data are very useful to aid lithostratigraphic classification and correlation and hypothesis testing.

Although some formations within groups are still poorly defined (e.g., the Parktown and Brixton Formations in the West Rand Group), multidimensional predictive modeling is easily able to differentiate between them and therefore clearly indicates that the sediment compositions of such geological units are, and in general could be, highly distinct. This corroborates field-based and U-Pb dating of detrital zircon evidence from the studied supergroups in both the Kaapvaal and Zimbabwe Cratons. This ability of the machine to examine high-dimensional data is suitable to extract minute and multidimensional insights and/or highly complex differences that are unnoticeable or easily missed by humans using traditional approaches. Therefore, the ability to perform hypothesis testing absolutely should not rest on traditional approaches alone but should incorporate data-driven approaches. Interpretations that are derived from traditional approaches and/or using smaller data sets should be closely scrutinized using larger data sets, and the reproducibility of scientific findings should always be tested in a data-driven manner. If the data do not fit any particular hypothesis, however grandiose or elegant it may be, then the hypothesis should be openly reconsidered.

Challenges to the adoption of modern data analytics and predictive analytics such as machine learning are discipline specific. Within the geosciences, the availability of data is heavily constrained by the high cost of data. Other issues, such as nonopen data and a lack of data management capabilities at the researcher and institution levels, leave valuable data in isolated silos that are not conducive to modern usages of data. However, the specifics of this reality are geoscience specific and are not easily remediated, as the production chain of data in geosciences is multistaged, with lots of feedbacks, and ultimately is difficult to reform, but it could be refined iteratively. Certain types of geoscientific data, such as chemistry, have the potential to become the "omics" of the geosciences, by simultaneously leveraging enormous data volumes, machine learning for automated data processing and insight extraction, and globally standardized and openly accessible databases. At present, hypothesis testing is still the dominant mode of data

production in geosciences, and data-driven science is comparatively much rarer to nonexistent. However, there are increasing signs that sufficient quantity and quality of data, where they exist, have been leveraged to demonstrate the value of data-driven approaches in geosciences (e.g., Nwaila et al. 2020; Zhang et al. 2021).

Conclusion

At the exploration stage for basin analysis and new mining projects, stratigraphic classification of rock units usually makes use of long-range stratigraphic correlation between new sites and sites that are well explored. Correct discrimination of Archean shales is often crucial for provenance evaluation, paleoenvironmental reconstructions, basin analysis, and exploration for economic potential. The main idea proposed in this study is the application of machine-learning algorithms to aggregated data for the classification of Archean shales. Machine-learning algorithms applied to major-element chemical data proved to be highly appropriate for the discrimination and prediction of Archean shales. Although shales physically appear to have uniform composition, their chemistry is highly variable. Such complex and multidimensional relationships are naturally ideal for machine-learning and data-driven approaches. The accuracy of machine learning–based predictions of shale classification depends strongly on the number of samples and is not troubled by geological complexity or chemical variability. When our data-driven approach is contrasted with traditional geochemical discrimination plots, it becomes evident that our approach is accurate. Our approach considers multiple major-element concentration data that are routinely analyzed for shale characterization and does not require expensive or highly specialized analyses. In the case of the Witwatersrand Supergroup, a comparison between the results of the West Rand and Central Rand Groups strongly supports the notion that they were deposited in different basin settings. The results illustrate that data and machine-learning algorithms are able to provide accurate predictions of stratigraphic classifications and robust fingerprinting of rocks that are conventionally difficult to discriminate. In addition, the benefit of data and data-driven approaches are numerous in primarily hypothesis-driven scientific disciplines, such as increasing objectivity and reproducibility, lowering the cost of data and the threshold of entry to conducting scientific research, and building constructive debates that are anchored in abundant data and objective analytics.

ACKNOWLEDGMENTS

The support of the Department of Science and Innovation–National Research Foundation Centre of Excellence for Integrated Mineral and Energy Resource Analysis (DSI-NRF CIMERA) toward this research is hereby acknowledged. We would also like to acknowledge the anonymous reviewers and the journal editors for their insightful input, which helped to increase the quality of our work.

REFERENCES CITED

Aitchison, J. 1982. The statistical analysis of compositional data. J. R. Stat. Soc. B 44:139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

Aitchison, J.; Barceló-Vidal, C.; Martín-Fernández, J. A.; and Pawlowsky-Glahn, V. 2000. Logratio analysis and compositional distance. Math. Geol. 32:271–275. https://doi.org/10.1023/A:1007529726302.

Bickle, M., and Nisbet, E., eds. 1993. The geology of the Belingwe Greenstone Belt, Zimbabwe: a study of the evolution of Archaean continental crust. Geol. Soc. Zimbabwe Spec. Publ. 2. Rotterdam, Balkema.

Breiman, L. 1996. Bagging predictors. Mach. Learn. 24: 123–140.

Burke, K.; Kidd, W. S. F.; and Kusky, T. M. 1986. Archean foreland basin tectonics in the Witwatersrand, South Africa. Tectonics 5:439–456. https://doi.org/10.1029/TC005i003p00439.

Catuneanu, O. 2001. Flexural partitioning of the late Archaean Witwatersrand foreland system, South Africa. Sediment. Geol. 141–142:95–112. https://doi.org/10.1016/S0037-0738(01)00070-7.

Chen, L.; Wang, L.; Miao, J.; Gao, H.; Zhang, Y.; Yao, Y.; Bai, M.; Mei, L.; and He, J. 2020. Review of the application of big data and artificial intelligence. J. Phys. Conf. Ser. 1684:012007. https://iopscience.iop.org/article/10.1088/1742-6596/1684/1/012007.

Chen, S.; Hattori, K.; and Grunsky, E. C. 2018. Identification of sandstones above blind uranium deposits using multivariate statistical assessment of compositional data, Athabasca Basin, Canada. J. Geochem. Explor. 188: 229–239. https://doi.org/10.1016/j.gexplo.2018.01.026.

Condie, K. C.; Des Marais, D. J.; and Abbott, D. 2001. Precambrian superplumes and supercontinents: a record in black shales, carbon isotopes, and paleoclimates?

Precambrian Res. 106:239–260. https://doi.org/10.1016/S0301-9268(00)00097-8.

Coveney, R. M., Jr.; Watney, W. L.; and Maples, C. G. 1991. Contrasting depositional models for Pennsylvanian black shale discerned from molybdenum abundances. Geology 19:147–150. https://doi.org/10.1130/0091-7613(1991)019<0147:CDMFPB>2.3.CO;2.

Cover, T., and Hart, P. 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13:21–27. https://doi.org/10.1109/TIT.1967.1053964.

Cramer, J. S. 2004. The early origins of the logit model. Stud. Hist. Philos. Sci. C 35:613–626. https://www.sciencedirect.com/science/article/pii/S1369848604000676.

Curry, H. B. 1944. The method of steepest descent for non-linear minimisation problems. Q. Appl. Math. 2:258–261. https://doi.org/10.1090/qam/10667.

Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. 2:303–314. https://doi.org/10.1007/BF02551274.

Domingos, P. 2012. A few useful things to know about machine learning. Commun. ACM 55:78–87. https://doi.org/10.1145/2347736.2347755.

Dramsch, J. S. 2020. 70 years of machine learning in geoscience in review. *In* Moseley, B., and Krischer, L., eds. Machine learning in geosciences. Adv. Geophys. 61:1–55. https://doi.org/10.1016/bs.agph.2020.08.002.

Feng, R., and Kerrich, R. 1990. Geochemistry of fine-grained clastic sediments in the Archean Abitibi greenstone belt, Canada: implications for provenance and tectonic setting. Geochim. Cosmochim. Acta 54:1061–1081. https://doi.org/10.1016/0016-7037(90)90439-R.

Fix, E., and Hodges, J. L., Jr. (1951) 1989. An important contribution to nonparametric discriminant analysis and density estimation. Int. Stat. Rev. 57:238–247. https://doi.org/10.2307/1403796.

Forgy, E. W. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics 21:768–769.

Freund, Y., and Schapire, R. E. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. *In* Vitányi, P., ed. Computational learning theory, EuroCOLT 1995. Lecture Notes in Computer Science 904. Berlin, Springer, p. 23–37. https://doi.org/10.1007/3-540-59119-2_166.

Frimmel, H. E. 2014. A giant Mesoarchean crustal gold-enrichment episode: possible causes and consequences for exploration. *In* Kelley, K. D., and Golden, H. C., eds. Building exploration capability for the 21st century. Soc. Econ. Geol. Spec. Publ. 18:209–234.

———. 2019. The Witwatersrand Basin and its gold deposits. *In* Kröner, A., and Hofmann, A., eds. The Archaean geology of the Kaapvaal Craton, southern Africa. Cham, Switzerland, Springer, p. 255–275. https://doi.org/10.1007/978-3-319-78652-0_10.

Frimmel, H. E., and Minter, W. E. L. 2002. Recent developments concerning the geological history and genesis of the Witwatersrand gold deposits, South Africa. *In* Goldfarb, R. J., and Nielsen, R. L., eds. Integrated methods for discovery: global exploration in the twenty-first century. Soc. Econ. Geol. Spec. Publ. 9:17–45.

Gazley, M. F.; Collins, K. S.; Robertson, J.; Hines, B. R.; Fisher, L. A.; and McFarlane, A. 2015. Application of principal component analysis and cluster analysis to mineral exploration and mine geology. *In* AusIMM New Zealand Branch Annual Conference 2015. Dunedin, Australasian Institute of Mining and Metallurgy (AusIMM), New Zealand Branch, p. 131–139.

Grobler, D. F.; Brits, J. A. N.; Maier, W. D.; and Crossingham, A. 2019. Litho- and chemostratigraphy of the Flatreef PGE deposit, northern Bushveld Complex. Miner. Depos. 54:3–28. https://doi.org/10.1007/s00126-018-0800-x.

Grunsky, E. C., and de Caritat, P. 2019. State-of-the-art analysis of geochemical data for mineral exploration. Geochem. Explor. Environ. Anal. 20:217–232. https://doi.org/10.1144/geochem2019-031.

Grunsky, E. C.; Mueller, U. A.; and Corrigan, D. 2014. A study of the lake sediment geochemistry of the Melville Peninsula using multivariate methods: application for predictive geological mapping. J. Geochem. Explor. 141:15–41. https://doi.org/10.1016/j.gexplo.2013.07.013.

Gu, A.; Sala, F.; Gunel, B.; and Ré, C. 2019. Learning mixed-curvature representations in product spaces. *In* International Conference on Learning Representations, 7th (New Orleans), Proc. https://openreview.net/pdf?id=HJxeWnCcF7.

Harris, J. R.; Grunsky, E.; Behnia, P.; and Corrigan, D. 2015. Data-and knowledge-driven mineral prospectivity maps for Canada's North. Ore Geol. Rev. 71:788–803. https://doi.org/10.1016/j.oregeorev.2015.01.004.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. The elements of statistical learning: data mining, inference, and prediction. New York, Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. *In* Proceedings: 2015 IEEE International Conference on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Soc., p. 1026–1034. https://doi.org/10.1109/ICCV.2015.123.

Ho, T. K. 1995. Random decision forests. *In* Proceedings of the third International Conference on Document Analysis and Recognition, Los Alamitos, CA: IEEE Computer Soc., p. 278–282. https://doi.org/10.1109/ICDAR.1995.598994.

Hsu, C. W., and Lin, C. J. 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 13:415–425. https://doi.org/10.1109/72.991427.

Jiang, S. Y.; Yang, J. H.; Ling, H. F.; Chen, Y. Q.; Feng, H. Z.; Zhao, K. D.; and Ni, P. 2007. Extreme enrichment of polymetallic Ni-Mo-PGE-Au in lower Cambrian black shales of South China: an Os isotope and PGE geochemical investigation. Palaeogeogr. Palaeoclimatol. Palaeoecol. 254:217–228. https://doi.org/10.1016/j.palaeo.2007.03.024.

Johnson, S. C.; Large, R. R.; Coveney, R. M.; Kelley, K. D.; Slack, J. F.; Steadman, J. A.; Gregory, D. D.; Sack, P. J.;

and Meffre, S. 2017. Secular distribution of highly metalliferous black shales corresponds with peaks in past atmosphere oxygenation. Miner. Depos. 52:791–798. https://doi.org/10.1007/s00126-017-0735-7.

Karatzoglou, A.; Meyer, D.; and Hornik, K. 2006. Support vector machines in R. J. Stat. Softw. 15:9.

Karpatne, A.; Ebert-Uphoff, I.; Ravela, S.; Babaie, H. A.; and Kumar, V. 2018. Machine learning for the geosciences: challenges and opportunities. IEEE Trans. Knowl. Data Eng. 31:1544–1554. https://doi.org/10.1109/TKDE.2018.2861006.

Kaufman, L., and Rousseeuw, P. 1990. Finding groups in data: an introduction to cluster analysis. New York, Wiley.

Kim, W.; Paola, C.; Swenson, J. B.; and Voller, V. R. 2006. Shoreline response to autogenic processes of sediment storage and release in the fluvial system. J. Geophys. Res. Earth Surf. 111:F04013. https://doi.org/10.1029/2006JF000470.

Koglin, N.; Zeh, A.; Frimmel, H. E.; and Gerdes, A. 2010. New constraints on the auriferous Witwatersrand sediment provenance from combined detrital zircon U-Pb and Lu-Hf isotope data for the Eldorado Reef (Central Rand Group, South Africa). Precambrian Res. 183:817–824. https://doi.org/10.1016/j.precamres.2010.09.009.

Kontinen, A.; Huhma, H.; Lahaye, Y.; and O'Brien, H. 2013. New U-Pb zircon age, Sm-Nd isotope and geochemical data on Proterozoic granitic rocks in the area west of the Oulunjärvi Lake, central Finland. *In* Hölttä, P., ed. Current research: GTK Mineral Potential Workshop, Kuopi, May 2012. Geol. Surv. Finland Rep. Investig. 198:70–74.

Kositcin, N., and Krapež, B. 2004. Relationship between detrital zircon age-spectra and the tectonic evolution of the late Archaean Witwatersrand Basin, South Africa. Precambrian Res. 129:141–168. https://doi.org/10.1016/j.precamres.2003.10.011.

Kotsiantis, S. B. 2007. Supervised machine learning: a review of classification techniques. Informatica 31:249–268. https://www.informatica.si/index.php/informatica/article/viewFile/148/140.

———. 2014. Bagging and boosting variants for handling classifications problems: a survey. Knowl. Eng. Rev. 29:78–100. https://doi.org/10.1017/S0269888913000313.

Large, R. R.; Gregory, D. D.; Steadman, J. A.; Tomkins, A. G.; Lounejeva, E.; Danyushevsky, L. V.; Halpin, J. A.; et al. 2015. Gold in the oceans through time. Earth Planet. Sci. Lett. 428:139–150. https://doi.org/10.1016/j.epsl.2015.07.026.

Lehmann, B.; Nägler, T. F.; Holland, H. D.; Wille, M.; Mao, J.; Pan, J.; Ma, D.; and Dulski, P. 2007. Highly metalliferous carbonaceous shale and early Cambrian seawater. Geology 35:403–406. https://doi.org/10.1130/G23543A.1.

Lemaréchal, C. 2012. Cauchy and the gradient method. *In* Grötschel, M., ed. Optimization stories. Doc. Math. Extra Vol.:251–254.

Leventhal, J. S. 1991. Comparison of organic geochemistry and metal enrichment in two black shales: Cambrian Alum Shale of Sweden and Devonian Chattanooga Shale of United States. Miner. Depos. 26:104–112. https://doi.org/10.1007/BF00195256.

Lloyd, S. P. (1957) 1982. Least square quantization in PCM. IEEE Trans. Inf. Theory 28:129–137.

Lundervold, A. S., and Lundervold, A. 2019. An overview of deep learning in medical imaging focusing on MRI. Z. Med. Phys. 29:102–127. https://doi.org/10.1016/j.zemedi.2018.11.002.

Lyons, T. W.; Reinhard, C. T.; and Scott, C. 2009. Redox redux. Geobiology 7:489–494. https://doi.org/10.1111/j.1472-4669.2009.00222.x.

Meyer, K. M., and Kump, L. R. 2008. Oceanic euxinia in Earth history: causes and consequences. Annu. Rev. Earth Planet. Sci. 36:251–288. https://doi.org/10.1146/annurev.earth.36.031207.124256.

Nwaila, G. T., and Frimmel, H. E. 2019. Highly siderophile elements in Archaean and Palaeoproterozoic marine shales of the Kaapvaal Craton, South Africa. Mineral. Petrol. 113:307–327. https://doi.org/10.1007/s00710-018-0650-3.

Nwaila, G. T.; Frimmel, H. E.; and Minter, W. E. L. 2017. Provenance and geochemical variations in shales of the Mesoarchean Witwatersrand Supergroup. J. Geol. 125:399–422. https://doi.org/10.1086/692329.

Nwaila, G. T.; Zhang, S. E.; Frimmel, H. E.; Manzi, M. S. D.; Dohm, C.; Durrheim, R. J.; Burnett, M.; and Tolmay, L. 2020. Local and target exploration of conglomerate-hosted gold deposits using machine learning algorithms: a case study of the Witwatersrand gold ores, South Africa. Nat. Resour. Res. 29:135–159. https://doi.org/10.1007/s11053-019-09498-1.

Rasmussen, C. E., and Williams, C. K. I. 2006. Gaussian processes for machine learning. Cambridge, MA, MIT Press.

Rokach, L. 2005. A survey of clustering algorithms. *In* Rokach, L., and Maimon, O., eds. Data mining and knowledge discovery handbook. Boston, Springer, p. 269–298. https://doi.org/10.1007/978-0-387-09823-4_14.

Rosenblatt, F. 1961. Principles of neurodynamics: perceptrons and the theory of brain mechanisms. Washington, DC, Spartan.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning internal representations by error propagation. *In* Rumelhart, D. E.; McClelland, J. L.; and PDP research group, eds. Parallel distributed processing: explorations in the microstructure of cognition. Volume 1: foundations. Cambridge, MA, MIT Press, p. 318–362.

Russell, S. J., and Norvig, P. 2010. Artificial intelligence: a modern approach (3rd ed.). Upper Saddle River, NJ, Prentice Hall.

SACS (South Africa Committee for Stratigraphy). 1980. Stratigraphy of South Africa. Part 1: lithostratigraphy of the Republic of South Africa, South West Africa/Namibia and the Republics of Bophuthatswana, Transkei, and Venda. Geol. Surv. S. Afr. Handb. 8. Pretoria, Geol. Surv. S. Afr.

———. 2006. A revised stratigraphic framework for the Witwatersrand Supergroup. Lithostratigr. Ser. 42. Pretoria, Counc. Geosci., 7 p.

Stanistreet, I. G., and McCarthy, T. S. 1991. Changing tectono-sedimentary scenarios relevant to the development of the late Archaean Witwatersrand Basin. J. Afr. Earth Sci. 13:65–81. https://doi.org/10.1016/0899-5362(91)90044-Y.

Toulkeridis, T.; Clauer, N.; Kröner, A.; and Todt, W. 2015. Mineralogy, geochemistry and isotopic dating of shales from the Barberton Greenstone Belt, South Africa: provenance and tectonic implications. S. Afr. J. Geol. 118:389–410.

Vapnik, V. 1998. Statistical learning theory. New York, Springer.

Vine, J. D., and Tourtelot, E. B. 1970. Geochemistry of black shale deposits; a summary report. Econ. Geol. 65:253–272. https://doi.org/10.2113/gsecongeo.65.3.253.

Wagener, G. F. 1972. Suggestions for the revision of the existing Witwatersrand stratigraphic classification and nomenclature. Trans. Geol. Soc. S. Afr. 75:77–84.

Ward, J. H., Jr. 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58:236–244.

Witten, I. H., and Frank, E. 2005. Data mining: practical machine learning tools and techniques (2nd ed.). San Francisco, Morgan Kaufman.

Wronkiewicz, D. J., and Condie, K. C. 1987. Geochemistry of Archean shales from the Witwatersrand Supergroup, South Africa: source-area weathering and provenance. Geochim. Cosmochim. Acta 51:2401–2416. https://doi.org/10.1016/0016-7037(87)90293-6.

Xu, L.; Lehmann, B.; and Mao, J. 2013. Seawater contribution to polymetallic Ni-Mo-PGE-Au mineralization in early Cambrian black shales of South China: evidence from Mo isotope, PGE, trace element, and REE geochemistry. Ore Geol. Rev. 52:66–84. https://doi.org/10.1016/j.oregeorev.2012.06.003.

Zalasiewicz, J.; Smith, A.; Brenchley, P.; Evans, J.; Knox, R.; Riley, N.; Gale, A.; et al. 2004. Simplifying the stratigraphy of time. Geology 32:1–4. https://doi.org/10.1130/G19920.1.

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