Mineralium Deposita
https://doi.org/10.1007/s00126-021-01086-9

REGULAR ARTICLE

Machine learning for geochemical exploration: classifying metallogenic fertility in arc magmas and insights into porphyry copper deposit formation

Chetan L. Nathwani[1,2] · Jamie J. Wilkinson[1,2] · George Fry[3] · Robin N. Armstrong[1] · Daniel J. Smith[4] · Christian Ihlenfeld[3]

Received: 8 June 2021 / Accepted: 23 November 2021

Abstract

A current mineral exploration focus is the development of tools to identify magmatic districts predisposed to host porphyry copper deposits. In this paper, we train and test four, common, supervised machine learning algorithms: logistic regression, support vector machines, artificial neural networks (ANN) and Random Forest to classify metallogenic 'fertility' in arc magmas based on whole-rock geochemistry. We outline pre-processing steps that can be used to mitigate against the undesirable characteristics of geochemical data (high multicollinearity, sparsity, missing values, class imbalance and compositional data effects) and therefore produce more meaningful results. We evaluate the classification accuracy of each supervised machine learning technique using a tenfold cross-validation technique and by testing the models on deposits unseen during the training process. This yields 81–83% accuracy for all classifiers, and receiver operating characteristic (ROC) curves have mean area under curve (AUC) scores of 87–89% indicating the probability of ranking a 'fertile' rock higher than an 'unfertile' rock. By contrast, bivariate classification schemes show much lower performance, demonstrating the value of classifying geochemical data in high dimension space. Principal component analysis suggests that porphyry-fertile magmas fractionate deep in the arc crust, and that calc-alkaline magmas associated with Cu-rich porphyries evolve deeper in the crust than more alkaline magmas linked with Au-rich porphyries. Feature analysis of the machine learning classifiers suggests that the most important parameters associated with fertile magmas are low Mn, high Al, high Sr, high K and listric REE patterns. These signatures further highlight the association of porphyry Cu deposits with hydrous arc magmas that undergo amphibole fractionation in the deep arc crust.

Keywords Porphyry · Copper · Ore deposit · Mineral exploration · Magma fertility · Machine learning · Geochemistry

Introduction

Igneous rock suites associated with porphyry Cu deposits are typically characterised by a distinct whole-rock geochemical signature that has been developed as an indicator of metallogenic 'fertility', meaning that magmas with such signatures may be predisposed to form porphyry Cu mineralisation. Several hallmarks of magma fertility have been proposed including high Sr/Y, high La/Yb, high Eu/Eu*, high Sr/MnO and high Al_2O_3/TiO_2 (Baldwin and Pearce 1982; Richards 2011; Wilkinson 2013; Loucks 2014; Ahmed et al. 2020), which have been increasingly used in porphyry Cu exploration. The distinct chemistry is thought to originate from the processes that may be important in generating magmas that form porphyry Cu deposits, involving strongly compressional tectonic regimes that promote crustal thickening

Communicated by Editorial handling: C. N. Mercer.

✉ Chetan L. Nathwani
chetan.nathwani14@imperial.ac.uk

[1] London Centre for Ore Deposits and Exploration (LODE), Department of Earth Sciences, Natural History Museum, Cromwell Road, South Kensington, London SW7 5BD, UK

[2] Department of Earth Science and Engineering, Imperial College London, Exhibition Road, South Kensington Campus, London SW7 2AZ, UK

[3] Anglo American Plc, 20 Carlton House Terrace, London SW1Y 5AN, UK

[4] Department of Geology, University of Leicester, University Road, Leicester LE1 7RH, UK

Published online: 24 January 2022

and protracted magma storage at deep crustal levels (Sillitoe 2010; Richards 2011; Chiaradia and Caricchi 2017). This causes differentiation of magmas at high pressure and melt H_2O contents, stabilising amphibole ± garnet, in which Y, MREEs and HREEs are compatible, but suppressing plagioclase in which Sr and Eu are compatible. This leads to rocks having the characteristics noted above, which can then be discriminated from normal arc rocks using bivariate thresholds (e.g. Loucks 2014; Ahmed et al. 2020; Wells et al. 2021). Such classification methods are useful but limited because they ignore additional variables that may hold fertility signals and lead to false positive outcomes because other processes can generate similar geochemical signatures. Some parameters are also susceptible to modification by hydrothermal alteration. Furthermore, false negative results are also common in porphyry rocks, such as in less evolved compositions (< 65 wt% SiO_2; Loucks 2014) and in more variable tectonic settings associated with alkaline magmas and gold-rich porphyry deposits (Chiaradia 2020).

Machine learning is the science of using computers to learn from data. Over the last few decades, machine learning algorithms have been developed to identify patterns and trends in diverse datasets and make predictions (Alpaydın 2014; Hastie et al. 2009; Bell 2014; Kubat 2017). Supervised classification learning is a branch of this, where input data are assigned a class label and the machine is trained to predict the class label using the input data. A geological example is the classification of tectonic settings of basalts using whole-rock chemistry (Ueki et al. 2018). Such techniques have significant potential in mineral exploration because datasets are becoming increasingly large, with significant numbers of observations (i.e. analyses) and features (i.e. analytes). Previous work has highlighted the application of such techniques to a variety of data types to predict mineral prospectivity, such as hyperspectral mapping, lithological mapping, structural mapping, soil geochemistry and lithogeochemistry (Cracknell and Reading 2014; Carranza and Laborte 2015; Rodriguez-Galiano et al. 2015; Geranian et al. 2016). Many supervised machine learning algorithms exist, but the most successful and widely used algorithms in mineral exploration and geochemistry are logistic regression, decision trees (including Random Forest), support vector machines and artificial neural networks (Vermeesch 2006; Rodriguez-Galiano et al. 2015; Geranian et al. 2016; Zhao et al. 2016; Zhao et al. 2016; Gregory et al. 2019; Yeomans et al. 2020).

In this study, four supervised machine learning algorithms (logistic regression, artificial neural networks, support vector machines and Random Forest) were applied to classify the metallogenic fertility of rocks in a compilation of global whole-rock data in order to distinguish samples spatially and temporally associated with porphyry Cu deposits from samples not associated with known mineralisation. The aims of this study are to (i) demonstrate the potential of such techniques for porphyry Cu exploration in magmatic arcs and quantify any improvement with respect to existing, largely bivariate techniques; (ii) compare the performance of each classification technique; (iii) establish whether a high-performing classifier can be generated regardless of the magma affinity; and (iv) identify the most important discrimination parameters for magma fertility and discuss the implications of these for the formation of porphyry Cu deposits.

Methods

Data compilation and quality control

Whole-rock geochemical data from porphyry Cu deposits were compiled from the literature (Table 1). A range of deposit sizes and types were included to ensure that the machine learning models were capable of learning and predicting fertility, independent of deposit size or type. The data were randomly down-sampled to ensure that no single deposit accounted for greater than 100 observations in the dataset, to reduce bias from over-representation. In the compiled data, whole-rock analyses were assigned a magma affinity (calc-alkaline or high-K calc-alkaline to shoshonitic) to allow a comparison of their whole-rock chemistry. These magma affinities are derived from the literature, whole-rock geochemistry (Peccerillo and Taylor 1976) and reported nomenclature of igneous rocks, following the approach of Chiaradia (2020).

Data unrelated to porphyry Cu deposits were parsed from the GEOROC database (http://georoc.mpch-mainz.gwdg.de/georoc/) for the Andean, Sulawesi, Luzon, Banda and Solomon arcs. These arcs were selected because they are known to host porphyry deposits and prospects of various types and represent a range of tectonic settings and magma affinities. These datasets were filtered to exclude any observations labelled as sedimentary rocks, metamorphic rocks, peridotites or veined rocks.

Major and trace elements selected for the compiled data were Si, Al, Fe (calculated as Fe^{2+}), Mg, Ca, Na, K, Ti, Mn, Sr, Y, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb and Lu. These features were chosen based on (1) the knowledge that they exhibit high variance during petrogenetic processes; (2) the fact that they have previously been shown to effectively discriminate porphyry Cu fertile igneous rocks; and/or (3) their being commonly reported in literature studies. Additional elements (e.g. V, Sc, Nb and Zr; Loucks 2014; Wells et al. 2021) have previously been found to act as useful discriminants; however, here we restrict our element list to those that have the most complete records in

Table 1 Porphyry Cu deposits included in the whole-rock geochemical dataset used in the training dataset for the machine learning techniques. Magma affinities are derived from Chiaradia (2020). CA = associated with calc-alkaline magmas, K = associated with high-K calc-alkaline to shoshonitic magmas. Approximate age ranges of the igneous rocks in the dataset, tonnage and grades for deposits (Singer 2005), the number of observations present in the filtered dataset and references are given. The age ranges include magmatism that precedes mineralisation by > 2Myr

Deposit	Country	Magma affinity	Age range (Ma)	Tonnage	Cu grade (%)	Mo grade (%)	Au grade (g/t)	n	References
Altar	Argentina	K	12–10	802	0.42	-	0.06	10	Maydagan et al. (2014)
Almalyk	Uzbekistan	K	326–312	6080	0.39	0.0023	0.37	4	Cheng et al. (2018)
Andacolla	Chile	K	104	417	0.34	-	0.12	2	Richards et al. (2017)
Balsapamba	Ecuador	CA	22	-	< 0.1	-	-	9	Schütte et al. (2010a, 2010b)
Batu Hijau	Indonesia	K	15–13	1644	0.44	0	0.35	7	Cooke et al. (2005); Fiorentini and Garwin (2010)
Bingham Canyon	USA	K	38–37	3230	0.882	0.053	0.38	9	von Quadt et al. (2011); Grondahl and Zajacz (2017)
Canicapa	Ecuador	CA	20	-	-	-	-	3	Schütte et al. (2010a); Schütte et al. (2010b)
Chaucha	Ecuador	CA	24–10	363	0.4	0.03	0	7	Schütte et al. (2010a, 2010b)
Chuquicamata	Chile	CA	35–33	21,277	0.592	0.04	0.013	17	Ballard et al. (2002)
Corocco-huayco	Peru	CA	40–35	155	1.57	0	0.33	32	Chelle-Michou et al. (2014, 2015)
Cuellaje	Ecuador	CA	NA	-	-	-	-	3	Schütte et al. (2010a, 2010b)
Don Manuel	Chile	CA	4–3	-	-	-	-	14	Gilmer et al. (2018)
Dos Amigos-Tricolor	Chile	K	107	36	0.36			2	Richards et al. (2017)
El Abra	Chile	CA	63–37	1779.4	0.494	0.0058	0	85	Ballard et al. (2002)
El Salvador	Chile	CA	44–42	3836.3	0.447	0.022	0.1	5	Lee (2008)
El Teniente	Chile	CA	24–3	20,731	0.62	0.019	0.005	31	Stern and Skewes (1995); Reich (2001); Rojas (2003); Vry (2010); Stern et al. (2011)
Escondida	Chile	CA	268–37	11,158	0.769	0.0062	0.25	32	Richards et al. (2001)
Gaby-Papa Grande	Ecuador	CA	21	308	0.09	0.025	0.73	11	Schütte et al. (2010a, 2010b)

Table 1 (continued)

Deposit	Country	Magma affinity	Age range (Ma)	Tonnage	Cu grade (%)	Mo grade (%)	Au grade (g/t)	n	References
Junin	Ecuador	CA	9	319	0.71	0.026	0	1	Schütte et al. (2010a, 2010b)
Kadjaran	Armenia	K	34–21	1700	0.27	0.055	0.65	8	Rezeau et al. (2016, 2017)
La Colosa	Colombia	K	8	821	0.11	0.017	0.8	14	Gil-Rodriguez (2010); Naranjo et al. (2018)
Ministro Hales	Chile	CA	210–33	1249	0.68	-	-	7	Ballard (2001)
Northparkes	Australia	K	452–436	472	0.56	0	0.19	37	Pacey (2016)
Ok Tedi	Papau New Guinea	K	1.5–1.1	854	0.64	0.011	0.78	16	Pollard et al. (2020)
Pebble	Alaska (USA)	K	100–41	7510	0.416	0.024	0.33	43	Olson (2015); Olson et al. (2017)
Productora	Chile	K	130	214.3	0.48	0.0138	0.1	1	Richards et al. (2017)
Qulong	Tibet (China)	K	18–15	1517	0.52	0.032	0	22	Hu et al. (2015)
Radomiro Tomic	Chile	CA	39–34	4980	0.39	0.015	0	23	Ballard (2001); Cooke et al. (2005); Cabrera (2011)
Relincho	Chile	CA	100–64	581	0.43	0.018	0	88	Greenlaw (2014)
Rio Blanco-Los Bronces	Chile	CA	19–4	16,816	0.601	0.02	0	4	Skewes and Stern (1995); Toro et al. (2012)
Sarycheku	Uzbekistan	K	338–313	200	0.5	0	0.1	6	Cheng et al. (2018)
Tampakkan	Phillipines	K	14–0.2	2500	0.48	0	0.2	29	Rohrlach (2002)
Telimbela	Ecuador	CA	22	-	-	-	-	1	Schutte et al. (2010a, 2010b)

the database, and are least subject to incomplete analyses and missing values.

The use of composite datasets (e.g. GEOROC) that comprise numerous sources of data requires quality control because each dataset was acquired using different methods (different preparation and instrumentation). Rocks analysed will also contain varying degrees of hydrothermal alteration, particularly for those derived from porphyry Cu systems. To help ensure that the compilation contained acceptable data, selection criteria of < 3.5 wt% loss on ignition (LOI) and analytical totals of 97.5–101.5 wt% were set (c.f. Loucks 2014). Data were also filtered to remove analyses that were feldspathoid-normative or contained > 3% normative corundum (Verma et al. 2003; Williams et al. 2020), because these would likely indicate significant hydrothermal alteration, or, in the case of high normative corundum, could indicate plagioclase accumulation. After data filtering, the compilation from porphyry Cu deposits (not including GEOROC data) comprised 555 least altered observations from 41 porphyry Cu systems (Table 1; Electron Supplementary Materials (ESM), Fig. S1). The filtered GEOROC-derived database consisted of 3559 observations.

Treatment of outliers

Consideration of the handling of outliers is necessary when training machine learning algorithms as they can be highly sensitive to the range and distribution of data. However, deciding the method for handling outlying data points is challenging because outliers can have multiple origins. Outliers arising from analytical or human error would be unrepresentative of patterns in the dataset and are best

discarded. However, statistical outliers may also arise as a natural product of diverse geological processes and discarding them may ultimately bias the models. Common outlier identification methods, such as the standard deviation or Tukey method, rely on data being normally distributed, and so are not well suited to geochemical data that rarely exhibit normal or log-normal distributions (Reimann and Filzmoser 2000). For example, discarding data above the potential outlier and extreme outlier thresholds for Sr (i.e. data points that lie three times the inter-quartile range away from the median following the Tukey method) would remove many Sr observations that may reflect a less common, but nonetheless important, geological process. Because igneous rocks associated with porphyry deposits can be typified by high Sr, this would introduce a bias in the machine learning models. This may additionally reduce the capacity of the models to reveal geologically meaningful processes through feature analysis. Hence, we do not filter outliers from the data, with rare or extreme outliers treated as 'noise', to which the machine learning algorithms should attribute little weight.

Treatment of missing values

Another issue is the treatment of missing values in datasets because most machine learning algorithms cannot be applied in such instances. Defining the optimal strategy for dealing with missing data is challenging since these can arise for a multitude of reasons. For example, in geochemical datasets, different analytical packages may have been used, which may not have all been capable of determining the full range of elements (missing values), or there may be data below the limit of detection (censored values).

A simple approach would be to remove observations that contain missing values. However, this can significantly reduce the size of the dataset because missing values for multiple elements are common in geochemical compilations with numerous, variable data sources. Furthermore, removing observations with missing data could introduce bias, for example because certain data sources (e.g. industry vs. academic) under-report specific elements. Another common method is to replace missing values ('impute' them) with a mean/median value derived from the rest of the data, or from within a given data subset. More sophisticated imputation methods exist such as those which use multi-variable regression, nearest neighbour approaches and non-parametric methods suitable for compositional data (e.g. Martín-Fernández 2003; van Buuren 2012). Significantly, classification methods now exist that are capable of handling missing values in datasets, such as the tree-based XGBoost system (Chen and Guestrin 2016). For simplicity, we did not include observations if they contained missing or censored values for the elements selected, except for partial gaps in rare earth element data which can be accurately interpolated from neighbouring rare earth elements. However, appropriately dealing with missing values is critical for industry-oriented classification tasks or with smaller training datasets where data loss is not an option.

Defining and classifying fertility

Arc magmas that are predisposed to generate porphyry Cu deposits are termed 'fertile' (e.g. Wilkinson 2013). However, district-scale studies have shown that fertile chemical signatures are not unique to the syn-mineralisation intrusions alone but can be found in rocks that closely pre- and post-date mineralisation, as well as rocks that formed several million years prior to mineralisation, such as host batholiths (Fig. 1, Ballard et al. 2002; Chiaradia et al., 2009; Nathwani et al. 2021). Ideally, therefore, 'fertility' should be treated as a probabilistic measure. However, for this initial analysis, we chose the simpler binary labelling ('fertile' or 'unfertile') approach, but this does require an objective definition in the training process to provide an effective classification model that is appropriate for the application for which it is being designed. We therefore classified any igneous rock from the

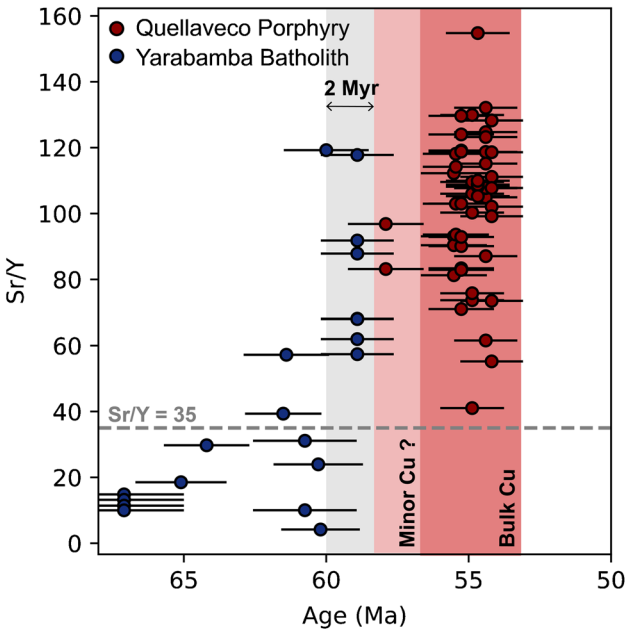

Fig. 1 Scatterplot showing whole-rock Sr/Y in the Quellaveco District, Southern Peru (data from Nathwani et al. 2021) as a function of emplacement age determined by zircon U–Pb LA ICP-MS data. Error bars are 2σ errors on the weighted mean with a propagated 2% systematic uncertainty. Dashed line indicates the 'fertile' threshold for Sr/Y proposed by Loucks (2014). Shaded regions indicate the time periods of bulk Cu mineralisation (darker red) and inferred minor Cu mineralisation (paler red) at Quellaveco (Simmons 2013; Nathwani et al. 2021). Plot indicates that high Sr/Y ('fertile') compositions were present in the ~ 2 Myr prior to the onset of Cu mineralisation (grey shaded region) and peaked during bulk Cu mineralisation

porphyry-related sample sets (from porphyry districts on the scale of ~ 10 km) that formed 2 Myr prior to, or after, mineralisation as 'fertile' and those from the same district that are older or younger than this window as 'unfertile'. This 2 Myr interval was selected because, according to recent studies, the development of 'fertile' magma chemistry occurs within 2 Myr of the onset of mineralisation in porphyry Cu districts, for example at Quellaveco, Peru (Fig. 1; Nathwani et al. 2021). This timescale is also equivalent to the maximum duration of porphyry Cu mineralisation in individual centres (Chelle-Michou and Rottier 2021). The temporal classification scheme was implemented using published geochronology for the ages of magmatic events and mineralisation for each included deposit (see Table 1). Based on this criterion, the porphyry Cu dataset consisted of 440 fertile observations and 115 unfertile observations. In the GEOROC-derived database, any observations that are from known porphyry systems were moved to the porphyry Cu database and the remaining observations were labelled as 'unfertile'. This age window for the porphyry district samples is not supposed to define the precise onset of the development of 'fertile' signals, which is difficult to be certain of based on current studies and may differ from site to site, but rather to provide an objective criterion for the training process. The outcomes of the machine learning models are discriminations that refer to this same 'fertility' definition, i.e. the models, if successful, should be able to recognise rocks that have a moderately close spatial and temporal relationship with a potential mineralisation event. This approach alone cannot predict whether a deposit actually formed since many additional factors that are not included in this approach may govern the formation of porphyry Cu mineralisation. These factors include the style and availability of structural conduits, the rheology and composition of the crust and the volume and duration of magmatic activity (Seedorff et al. 2005; Sillitoe 2010; Richards 2013; Chelle-Michou et al. 2017).

The compositional data problem

Geochemical data are an example of compositional data where each component is not an absolute value but are relative values that sum to a constant C (e.g. 100%). Compositional data are typically expressed as a sum of k individual components x_i, and for a given composition, if one component x_i were to increase, this would require a decrease in the other components x_{k-1} to retain the constant sum (Aitchison 1982). A classic example is whole-rock geochemical data that are typically expressed as percentages or in parts per million that sum to a constant (e.g. 100% or 10^6). The constant sum effect leads to spurious correlations, such as a bias towards negative correlations between components that are otherwise positively correlated (Chayes 1960). For example, a traditional 'Harker variation diagram' of SiO_2 vs Al_2O_3 for igneous rocks typically shows a negative correlation (ESM 1, Fig. S2), conventionally interpreted as indicating plagioclase fractionation. However, this trend may be partly or entirely spurious because an increase in SiO_2 from 50 to 75 wt% must be accompanied by a halving of other components to retain the constant sum of 100%. Additionally, igneous rock compositions typically yield non-normal distributions (Ahrens 1954; Reimann and Filzmoser 2000). These properties preclude the application of traditional statistical techniques on raw compositional data because many statistical techniques assume variables vary independently of each other and that they fulfil normal distributions.

The recognition of the interdependent and non-parametric nature of compositional data led to the introduction of log-ratio transformations (Aitchison 1986). The centred-log-ratio (*clr*) for an individual component of a composition x_i is obtained by dividing this component by the geometric mean, $g(x)$, of all the components of composition x and taking the natural logarithm of this ratio (Aitchison 1986, see ESM 1, Table S1 for an example calculation):

$$clr(x_i) = \ln\left(\frac{x_i}{g(x)}\right) \quad (1)$$

The *clr* transformation allows the redefined components to vary within unconstrained real space, rather than being constrained by a constant sum, and thus allows the application of multivariate statistic techniques on the transformed data. Such pre-processing is particularly valuable in this study where varying degrees of hydrothermal alteration may cause significant mass loss/mass gain, leading to additional spurious correlations in the dataset. Another useful characteristic of the *clr*-coordinates is that they are sensitive to relative rather than absolute variance. For example, in a suite of rocks where SiO_2 varies from 60 to 65 wt% and CaO from 1 to 6 wt%, both components have an absolute variance of 5 wt%, but relative variances of 1.08 and 5 respectively (Lipp et al. 2020). The latter is captured by *clr*-coordinates. Although Random Forest is non-parametric (i.e. can handle interdependent and non-normally distributed data), in this study, we have *clr*-transformed all data for consistency using the *pyrolite* package in Python (Williams et al. 2020).

It is common for datasets to be scaled before training machine learning models so that each feature has a mean of 0 and a standard deviation of 1. This can be advantageous in that it allows each feature to contribute proportionally during the training process and it can reduce the computation time required for certain techniques (e.g. gradient descent algorithms in artificial neural networks; Ioffe and Szegedy 2015). However, here we do not scale features, because all features have the same units and the removal of variance can lead to information loss (Lipp et al. 2020).

Class imbalance

Many classification problems involve classes that do not contain equal numbers of observations; this is likely to be common in mineral exploration applications because ore deposits represent rare geochemical anomalies in the Earth's crust. In this study, the proportion of data from porphyry systems (9%) is far lower than that from 'unfertile' igneous rocks (91%). This can lead to misrepresentation of classification accuracy because, in our case, if a classifier always predicted 'unfertile', then the classifier would misleadingly have a 91% classification accuracy, despite never correctly classifying any fertile observations. One method to avoid this problem is to use alternative performance metrics such as confusion matrices and receiver operating characteristic curves. However, class imbalance can still lead to models that favour prediction of the majority class and therefore have a higher probability of misclassification of the minority. It has been shown, particularly for smaller, complex datasets, that class imbalance can reduce the performance of decision trees, neural networks and support vector machines (Japkowicz and Stephen 2002).

Although the class imbalance in our dataset (91% unfertile, 9% fertile) is not extreme compared to that possible in other machine learning problems, a balanced class distribution is optimal (Weiss and Provost 2003). The simplest approaches to achieving this are to under-sample the majority class or over-sample the minority class to produce a class-balanced database (Kubat 2017). Although oversampling has been shown to produce better performance and avoids the data loss inherent in under-sampling (Japkowicz and Stephen 2002), under-sampling of the majority was employed in this study because over-sampling was found to lead to overfitting of the models (see ESM 1, Fig. S3).

Model generalisation

Ultimately, supervised machine learning models aim to perform well when predicting the output on data not encountered during the training process. This is attained when the model can generalise the training data, rather than overfitting or underfitting. Overfitting of training data occurs when the model captures variance or 'noise' in the dataset rather than the underlying data distribution, meaning it will fail to predict future observations reliably (ESM 1, Fig. S4). Such models are said to have high 'variance' error, meaning they are sensitive to small fluctuations in the training data. Underfitting occurs when a model does not fully capture the underlying data distribution because it is not complex enough (ESM 1, Fig. S4). Such models have high 'bias' error, meaning the model oversimplifies the problem due to incorrect assumptions during the training process. Decreasing variance (decreasing model complexity) causes an increase in bias, and vice versa (known as the 'bias-variance tradeoff'); hence, an optimal balance must be sought to reduce the total error in the model. If one considers a two-dimensional fictive dataset comprising two data classes (ESM 1, Fig. S4), the optimal classifier will better classify unseen test data when compared to an overfit classifier which models noise or an underfit classifier which oversimplifies. Feature selection, dimensionality reduction and hyperparameter tuning are examples of methods that can be used during the training process to mitigate against overfitting and underfitting, and thus reduce total error.

Dimensionality reduction and principal component analysis

Although the focus of this study is supervised machine learning, it is common for supervised techniques to be preceded by an unsupervised step for dimensionality reduction. Unsupervised techniques are those that learn patterns from unlabelled data. Dimensionality reduction reduces a dataset into a smaller number of parameters that can represent covariances within the original dataset. The technique is valuable because, with increasing numbers of features (such as compositional variables), complete data tend to become increasingly sparse as the Euclidean distance between data points increases. Sparsity generally increases exponentially with increasing features, requiring extremely large numbers of observations to cover the high dimension space. It can be challenging to apply supervised machine learning techniques to such datasets and the generated models are highly prone to overfitting. Although such effects may be small with a dataset of the size considered here, and similarly accurate results are produced with and without dimensionality reduction (ESM 1, Fig. S5), we prefer to include this stage as best practice. Furthermore, geological interpretability also can be derived from this stage (principal component loadings and scores).

Here, we use principal component analysis (PCA) on the *clr*-transformed data to reduce the features in the geochemical database to a smaller number of features that are representative of the covariance structure of the dataset. Principal components (PCs) are linear combinations of the original variables. The first PC is computed using least squares fitting to find a plane that accounts for the maximum amount of variance in the dataset. The second PC is consequently orthogonal to the first component to show the dimension of second-most variance. Further PCs are calculated in the same manner. These PCs increase the signal/noise ratio in a dataset which can aid in more effective classification. Also, they are orthonormal (i.e. statistically independent) and reflect linear processes which can be attributed to geological processes (Grunsky and Caritat 2019). Each observation is assigned a factor score for each PC which represents the

degree of variance of the observation in the dimension of the PC. These factor scores are used as the inputs for machine learning. PCA was implemented using the PCA function in *scikit learn* in Python (Pedregosa et al. 2011). Only the first 6 PCs, which account for ~ 90% of the variance of the training dataset (ESM 1, Fig. S6), are included in the models. Removing the further PCs reduces noise in the training dataset and therefore makes the models less prone to overfitting (see ESM 1, Fig. S7).

Supervised machine learning techniques

The descriptions of the machine learning techniques used here are a brief summary of those found in Hastie et al. (2009) Alpaydın (2014), Bell (2014) and Kubat (2017). For a complete discussion of these techniques, the reader should refer to these texts and references cited therein. All machine learning techniques were applied using *scikit learn* (Pedregosa et al. 2011), a machine learning package coded in Python.

Logistic Regression

Logistic regression is one of the simplest supervised machine learning techniques. It uses a logistic (sigmoid) function to predict a binary class label based on a linear combination of one or more independent variables. For an example dataset (Fig. 2a), the probability of an observation X (e.g. a whole-rock composition) belonging to a class Y (where Y can either be fertile or unfertile) would be modelled using the logistic function to calculate a probability between 1 and 0:

$$P(Y|X) = \frac{1}{1 + e^{-f(x)}} \tag{2}$$

where the linear combination of features $f(x)$ is the sum of each individual feature x (e.g. each component in a bulk rock composition) multiplied by a weight coefficient w with an added bias term, or intercept b:

$$f(x) = \sum_{i=1}^{n} w_i x_i + b \tag{3}$$

A numerical optimisation algorithm (which minimises an objective function) is used to select the values for the weight coefficients for each feature to minimise the classification error during the training process. The derived 'best fit' logistic function can then be used to predict the probability of an unknown observation belonging to class Y (e.g. the probability of being fertile), and if the probability is higher than a 'threshold' (typically 0.5—Fig. 2a), it is classified as positive (e.g. fertile). Logistic regression has been used across a range of geological studies, with relevant examples including prospectivity mapping (Carranza and Hale 2001; Porwal et al. 2010), hydrothermal alteration mapping using lithogeochemical data (Mokhtari 2014) and detrital provenance studies (Itano et al. 2020).

Artificial Neural Networks

Artificial neural networks are named due to their analogous architecture to animal brains. The individual unit of an artificial neural network is a neurone, which, like a biological neurone, converts a series of input signals to an output signal. A neurone first takes the weighted sum of all the inputs (in which each input element is assigned a weight w depending on its desired impact on the output) and adds a bias term b (Eq. 3). This then passes through a function f (such as a sigmoid, ReLu or tanh function), termed the activation function, which converts the value to an output that determines how 'active' the neurone is (Fig. 2c inset). An example neurone with x_n input features with corresponding w_n weights and a bias term b could use a sigmoid activation function like logistic regression (Eq. 3). Essentially, this converts the input to a value between a set range (e.g. between 0 and 1, where 1 is fully active and 0 is inactive). Feed-forward, multi-layer neural networks consist of layers, where each layer is composed of many neurones (Fig. 2c). The data are first fed from an input layer (Fig. 2c) and then pass through one, or several, 'hidden layers' before arriving at an output layer where a classification can be made (Fig. 2c; fertile or unfertile). During training, the model aims to determine the values for w and b in each neurone that produces the most successful classification result. When an untrained ANN first receives training data and produces an initial classification, it computes the classification error using the cost function. The algorithm works backwards through the network by a process known as back-propagation to adjust the weights and biases to minimise the cost. It does so by computing a gradient for the cost function and adjusting the weights and biases towards minima for the cost function (using the steepest gradient) by gradient descent.

The back-propagation process has a substantial number of weights and biases that must be adjusted in each neurone to achieve a low classification error. ANNs can produce overly complex models, making them prone to overfitting, particularly where the number of hidden neurones is large, or where noisy datasets are involved, thus regularisation methods to reduce overfitting are generally required (e.g. dropout; Srivastava et al. 2014). Neural networks have been the focus of a number of studies of mineral prospectivity mapping using a variety of data including soil geochemistry, geological map information and geophysics (Porwal et al. 2003; Rodriguez-Galiano et al. 2015; Zhang et al. 2019; Li et al. 2020, 2021).

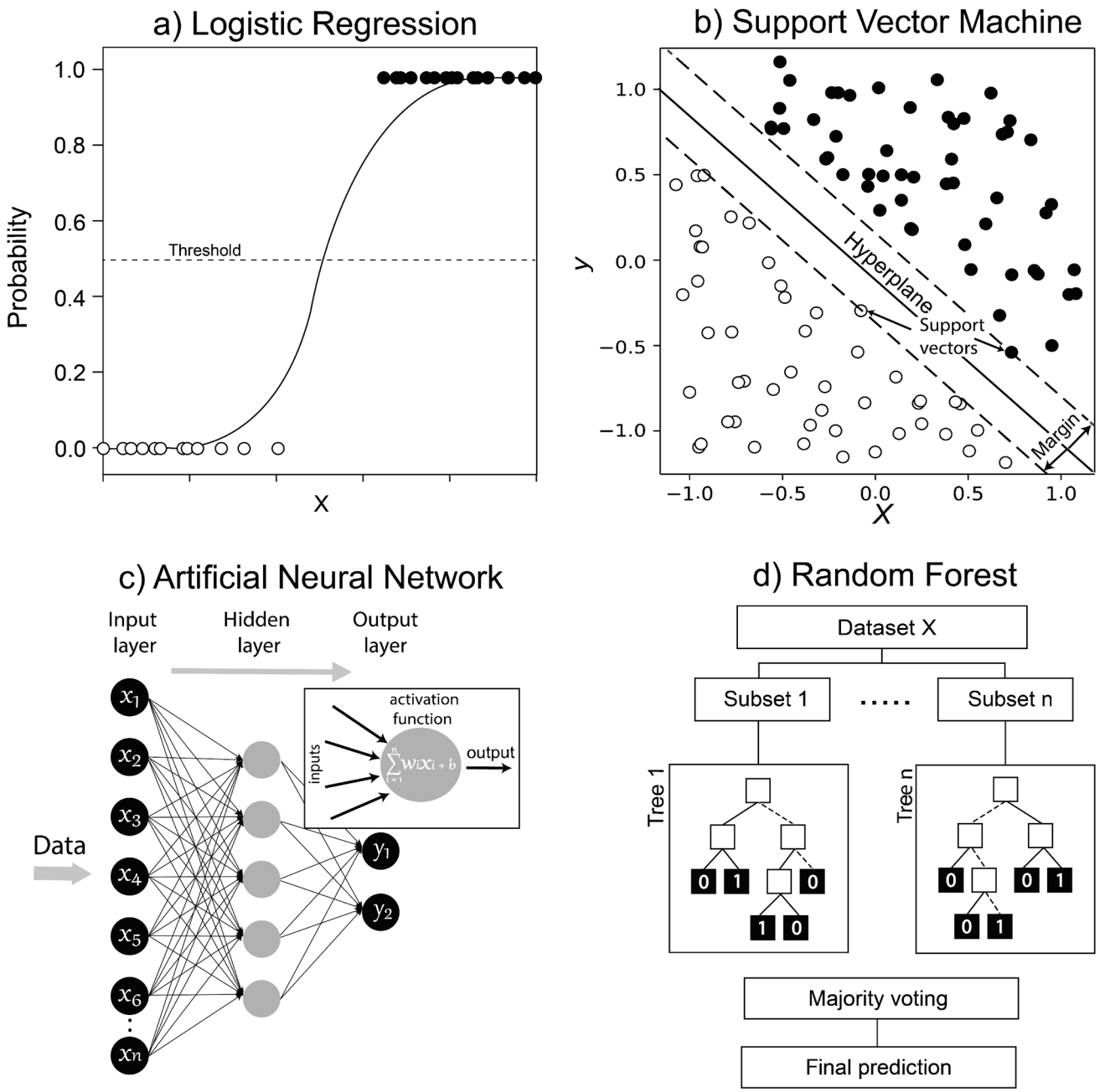

Fig. 2 Schematic illustrations of the four supervised machine learning methods used in this study. **a** Logistic regression showing a logistic function fit to binary data based on a variable *x*; **b** support vector machine (SVM) showing linear hyperplane between binary classes on a bivariate plot, showing the locations of the support vectors and the margin; **c** artificial neural network (ANN) showing the input layer consisting of *n* different features (i.e. elements), which are translated through a 'hidden layer' composed of five neurones, producing a binary output *y* (i.e. fertile or unfertile). The inset figure shows an individual neurone which contains a logistic function as the activation function; **d** Random Forest—where a dataset is sub-sampled *n* times and a classification tree is built on each subset to predict the label (i.e. fertility). The average result of all trees is taken as a final prediction

Support vector machines

Support vector machines are a supervised machine learning method used in the classification of high-dimensional, non-linear data. The method transforms data into high dimension space with the aim of separating them using a high-dimensional decision surface, termed a hyperplane (Cortes and Vapnik 1995). In a simple case, two classes of bivariate data (Fig. 2b) can be separated by a line, with the data points closest to this line being the support vectors. The optimal

classification position of the line is that which is furthest from the support vectors and therefore has the maximum margin.

In many real datasets, it is impossible to completely separate data classes, thereby requiring the formulation of a soft margin. Here, a penalty is incurred for data that are misclassified or are within the margin, with misclassified data points further from the hyperplane being assigned a larger penalty. This is important for noisy datasets; however, this can also lead to overfitting where the model becomes too biased towards noise in the training data. The user can specify the importance given to classification mistakes (C), where a low importance focuses on maximising the margin (i.e. the distance between the support vectors), whereas a high importance focuses on reducing misclassifications — at the expense of reducing the margin.

In many classification problems, it may not be possible to use a linear or planar function to classify the data. In these situations, linearly inseparable data are mapped to a higher-dimension space using a kernel function by which they can be better separated. This process of transforming variables into higher dimensions for classification is more computationally expensive; consequently, support vector machines use the *kernel trick* which allows the algorithm to operate in high dimension space without data transformation.

A notable example of a geochemical application of support vector machines is the tectonic discrimination of volcanic rocks using whole-rock geochemistry and isotopes by Petrelli and Perugini (2016). The authors showed that model performance improves when models are trained on an increasing number of dimensions (analytes) and by using a non-linear kernel function instead of a linear one. Support vector machines have also been shown to be effective in mineral prospectivity mapping, lithological classification and alteration facies discrimination (Zuo and Carranza 2011; Abedi et al. 2012; Abbaszadeh et al. 2015; Geranian et al. 2016).

Classification trees

Classification trees, a subset of decision trees, use observations of an item to make a prediction of its class value using a repetitive set of binary partitions based on threshold values of the observations (Breiman et al. 2017). These are represented as 'trees' in which classification of a dataset X initiates at the 'root node' and passes down the tree where, at each split, or 'node', a partition is made in X into two descendent subsets based on the split in one variable (greater than or less than a given concentration value for a given element). Eventually, the item reaches a 'terminal node' where it is assigned a class value (e.g. fertile or unfertile). Each node aims to choose a variable and value that best splits the set of items, minimising the probability of misclassification (i.e. the impurity). This can be quantified by the Gini impurity G, which is the probability p of incorrectly classifying a randomly chosen observation in the dataset if it were randomly labelled according to the class distribution in the dataset, where C is the total number of classes:

$$G = \sum_{i=1}^{C} p(i) \times (1 - p(i)) \tag{4}$$

Classification trees offer some advantages compared to other machine learning methods. The first, which is a particular advantage for studying compositional data, is that classification trees are non-parametric and therefore do not assume that the data conform to a certain distribution, unlike logistic regression, artificial neural networks and some support vector machines. A second advantage is that the interpretability of classification trees is far greater because they can be visualised in two dimensions. However, decision-tree learners can create over-complex trees that do not generalise the data sufficiently (overfitting). Mechanisms such as pruning (removing sections of trees that are non-critical or redundant) and setting the maximum depth of the tree are necessary to avoid this problem. Vermeesch (2006) demonstrated that classification trees built on 51 major, minor and trace elements and isotopic ratios could successfully classify the tectonic affinity of basalts with 89% probability, and described several useful properties of this approach compared to typical bi- or tri-variate tectonic discrimination diagrams.

Ensemble methods and Random Forest

Random Forest is an example of an ensemble classification method that uses a combination of many predictors (classification trees) and selects the majority prediction as the final output (Breiman 2001). Random Forest reduces the variance of the averaged outcome compared to a single decision tree, and therefore greatly reduces the error rate. The variance is reduced in two ways that aim to minimise the correlation between individual trees. First, each tree in the Random Forest is built on a random sub-selection of observations (a bootstrap) in the data (Fig. 2d), in which sub-selection occurs by replacement. This means that observations may be repeated in numerous bootstraps. Second, at each node in a tree, the split is made using a random sub-selection of features in the data. An observation is classified by each of these de-correlated trees in the Random Forest, and the majority outcome is the predicted class. This process of aggregating several multiple versions of a predictor from sub-selections of the dataset is known as bagging (Breiman 1996). A further advantage of the bagging process is that the observations that are not incorporated into the sub-selection for each tree, the 'out-of-bag' data, can be used to assess the

classification error as trees are added to the forest, meaning a validation dataset is not necessary.

Random Forest is particularly effective for datasets with many features, even when a considerable proportion is unimportant or where there are limited or noisy observations. The approach is also thought to reduce overfitting of the training data (Breiman 2001) because, as more trees are added to the forest, the error function converges towards a minimum, making them much more robust than a single decision tree. For example, Petrelli et al. (2020) used a number of machine learning algorithms and clinopyroxene-melt chemistry to determine temperatures and pressures of crystallisation of a series of igneous rocks, and their tree-based ensemble methods (including Random Forest) generally produced lower errors than a single decision tree model.

Random Forest is being increasingly used for mineral exploration problems. For example, Gregory et al. (2019) used it to effectively classify ore deposit types using 11 different trace elements in pyrite. A popular use of Random Forest is for mapping lithologies and mineral prospectivity using a combination of geochemical and geophysical data (Carranza and Laborte 2015; Harris and Grunsky 2015; Rodriguez-Galiano et al. 2015).

Model validation

Testing of models on unseen data is critical in assessing the ability of supervised machine learning models to generalise. Models are typically tested by withholding a portion of the dataset from the data used to train the model and using the withheld data to test the model performance. The performance of models developed here was validated using a tenfold cross-validation technique (Fig. 3). This involves splitting the data into tenfold (or subsets), with ninefold used to train the model, and the withheld fold used to test the model. This is repeated 10 times, until every fold has appeared once as the test set, and an average of the metric scores is taken to give the model performance. The benefit of this approach, as opposed to a single train/test set, is that it reduces the possibility of high bias that may arise from a single train/test set and helps to ensure models generalise better on unseen data. A further 'test' dataset, which is never encountered by the algorithm during the training process, is kept aside to further test model performance. It is important to ensure that pre-processing steps such as PCA are fit to only the training data after cross-validation splits, rather than to the entire dataset, and then applied to both the training and validation sets. This is because pre-processing steps should be learnt from only the training dataset; otherwise, the training dataset will be transformed based on information held in the validation dataset, ultimately biasing the cross-validation process. We therefore implement pre-processing steps using a 'pipeline' which sequentially applies data transformations

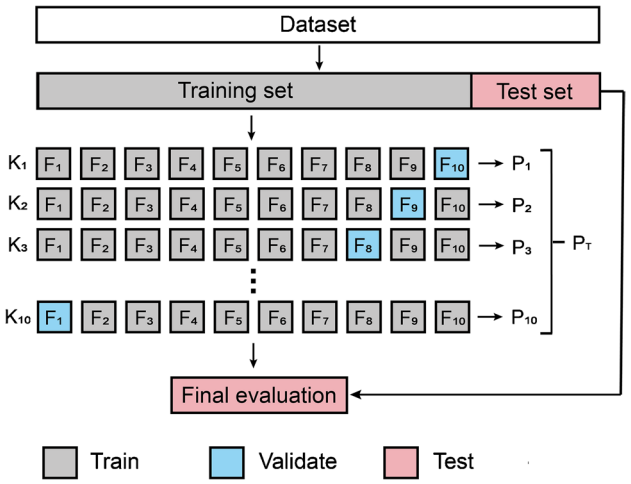

Fig. 3 Schematic illustration of a tenfold cross-validation workflow. The dataset is split into a training set and test set. The training set is further split into tenfold upon which the cross-validation process is carried out 10 times, with each step involving training of the model with ninefold and testing with the excluded fold. Performance metrics (P) are calculated for each fold and the mean metric (P_T) is calculated as the overall performance. Hyperparameter tuning occurs during this cross-validation process. A final evaluation is then performed on the tuned model using the test set, where at this stage no further tuning of the model occurs. Figure modified after Pedregosa et al. (2011)

and the estimator for each cross-validation step (Pedregosa et al. 2011).

Several metrics exist to evaluate the classification performance of binary classifiers. The most popular metric is accuracy, which is simply the proportion of correct predictions made by the classifier. However, because accuracy is less useful for class imbalanced datasets and multi-class classification problems, additional metrics are often used which provide better insight into model performance. The true positive rate (TPR), also known as sensitivity or recall, is a measure of how many of the positive observations were correctly classified (true positive, TP) compared to positive observations that were incorrectly classified (false negative, FN). The false positive rate (FPR) is a measure of how many negative observations were incorrectly classified (false positive, FP) compared to the number of true negative observations (true negative, TN). The precision (PPV) is a measure of how many positive observations were correctly classified. Commonly, the harmonic mean of recall and precision, known as the F1 score, is used as an alternative metric to accuracy. These parameters are defined as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{6}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

$$\mathrm{F1} = 2 \times \frac{\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} \tag{8}$$

A common method for evaluating binary classification performance is to use a receiver operating characteristic (ROC) curve. Binary classifiers will provide a probability that an observation belongs to a class and if this probability is greater than the threshold (typically 0.5) it will classify the observation as belonging to the positive class. A ROC curve plots the true positive rate versus the false positive rate (FPR) as the discrimination threshold is varied for the classifier. A classifier with no skill would show an equal TPR and FPR as the threshold is varied, whereas a skilful classifier would have a high true positive rate and low false positive rate. The area under the ROC curve for a classifier is known as the area under curve (AUC) and is commonly used as a metric for the performance of a classifier. AUC indicates the probability that a classifier will rank a randomly chosen positive observation higher than a randomly chosen negative observation (Fawcett 2006), where AUC = 1 would indicate a perfect classifier, and AUC = 0.5 would be a classifier with no skill.

Model optimisation

Parameters that are not directly learnt during machine learning, but which control how the algorithm itself constructs the model and learns from the data are known as hyperparameters and are specified 'up-front' during initialisation of the model. Examples of hyperparameters are number and maximum length of trees in the Random Forest algorithm, the number and size of layers in an artificial neural network, or the regularisation factor and kernel function for a support vector machine. To select the optimal hyperparameters for each algorithm, a 'grid search' is used that fits the model to the dataset using different combinations of specified hyperparameters and returns the optimal hyperparameters for the model using tenfold cross-validation. A full list of tuned hyperparameters from this study is provided in ESM 1, Table S2.

Results and discussion

Bivariate discrimination plots

Understanding the petrogenesis of arc magmas associated with porphyry deposits has typically relied on bivariate plots, primarily involving Sr, Y and REE ratios (e.g. Richards and Kerrich 2007; Loucks 2014). Bivariate plots can partly separate the dataset compiled here, demonstrating differences in the chemistry of fertile arc rocks compared to unfertile arc rocks (Figs. 4 and 5), consistent with previous studies (high Sr/Y, high Al_2O_3/TiO_2, high Sr/MnO and high La/Yb). Such plots may be useful in exploration for terranes predisposed to host porphyry Cu deposits. However, an issue with such bivariate classification schemes is that many misclassifications exist due to overlap between the two classes. For example, separate kernel density plots for the classes in the dataset demonstrate that the bivariate fertility signals are not useful for less evolved compositions (< 60 wt%; Fig. 4). Furthermore, the discriminating features are more pronounced for porphyry Cu deposits found in thicker arcs and associated with calc-alkaline magmas (typically Cu- and Mo-rich porphyries), than for many porphyry deposits that are found in thinner arcs and which are generally associated with high-K calc-alkaline and shoshonitic magmas, and are typically Au-rich (Figs. 4, 5 and 6; Sillitoe 1997; Müller and Groves 2019; Chiaradia 2020). Thus, false negative fertility signals are frequently observed in igneous rocks associated with Cu–Au porphyry systems. Additionally, false positives can be common using such classification schemes partly because bivariate signatures such as high Sr/Y and high La/Yb are not exclusive to magmas forming porphyry Cu deposits and can form through a variety of petrological processes (e.g. Bourdon 2002; Topuz et al. 2005; Macpherson et al. 2006; Chiaradia, 2009; Moyen 2009).

We tested the performance of the Loucks (2014) SiO_2 vs Sr/Y and Ahmed et al. (2020) Sr/Y vs Sr/MnO schemes on our dataset prior to implementing supervised machine learning techniques. Loucks (2014) suggested that igneous complexes having Sr/Y > 35 at SiO_2 > 57 wt% should be considered Cu-fertile (Fig. 5). Based on our dataset, this criterion produces an accuracy score of 69% for porphyry Cu deposits associated with calc-alkaline magma suites (TPR = 77%, FPR = 35%), and an accuracy of 65% for porphyry Cu deposits related to high-K calc-alkaline and shoshonitic magma suites (TPR = 72%, FPR = 34%). Classifying the data compiled here using the Sr/Y vs. Sr/MnO fertility classification diagram (Ahmed et al. 2019; Fig. 6) returns a 48% classification accuracy (TPR = 68%, FPR = 25%) for porphyry Cu deposits associated with calc-alkaline magma suites and a 45% accuracy (TPR = 56%, FPR = 26%) for porphyry Cu deposits related to high-K calc-alkaline and shoshonitic magma suites (Fig. 6). These tests demonstrate that such schemes are useful but can be limited. Lower classification performance is seen for the alkaline rocks in both cases, and the former test yields approximately 1 in 3 false positives. We suggest that this is mainly a consequence of the small number of elements being used to classify the data, which do not account for the full variance of the datasets and therefore cannot capture the full, underlying differences between the

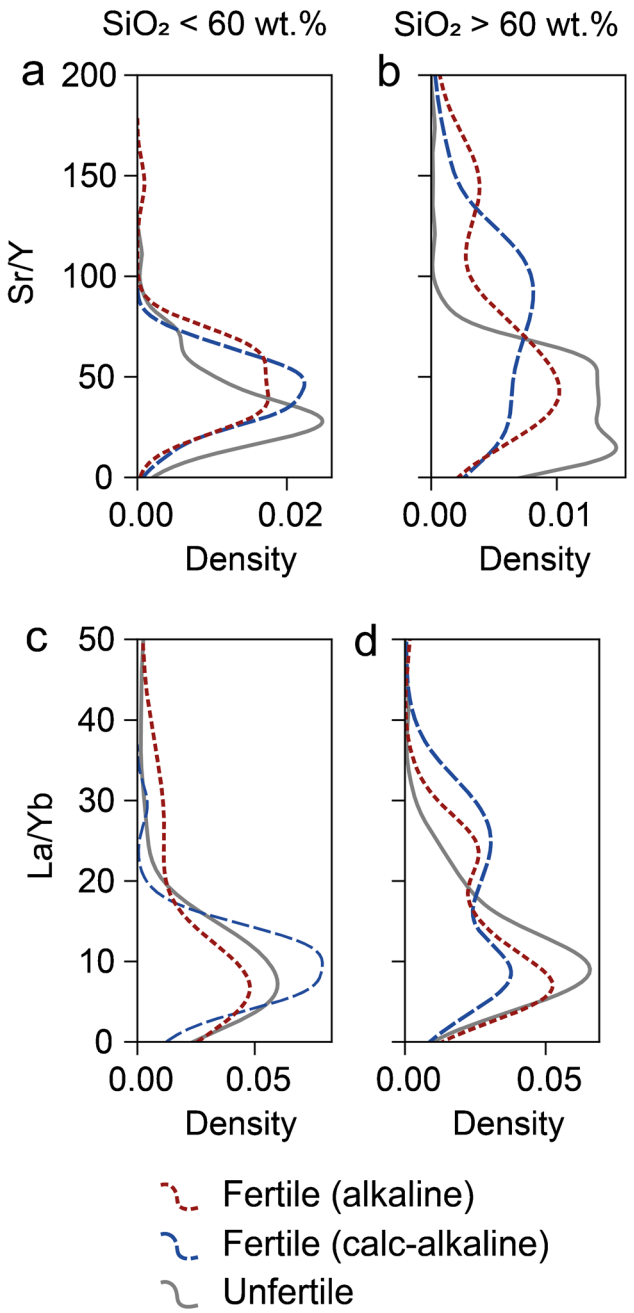

Fig. 4 Kernel density plots showing the frequency distribution of Sr/Y (a and b) and La/Yb (c and d) in the training dataset. Separate plots are given for less evolved (< 60 wt% SiO_2) and more evolved (> 60 wt%) compositions to illustrate that previously defined 'fertility indicators' produce more false negatives in less evolved compositions. The separate curves for calc-alkaline (CA)- and alkaline-related (K) porphyry systems demonstrate that some signatures (e.g. high Sr/Y) are less pronounced in alkaline-related porphyry systems. The small bump at ~ 150 Sr/Y for alkaline-related porphyry systems is caused by the Qulong porphyry Cu–Mo deposit which display atypical high-Mg, ultrapotassic and high Sr/Y compositions

populations. These data populations may instead be better separated in high dimension space, where a larger proportion of the data variance can be modelled.

Bivariate plots can still be used to visualise the higher dimension space of the data by plotting PC scores. The PC loadings that relate to these scores can reflect key geological processes that may be important fingerprints of magma evolution and porphyry copper deposit potential (Fig. 7 and ESM 1, Fig. S8). A plot of PC1 vs PC2 for the entire dataset (Fig. 7a) allows 70% of the dataset variance to be represented which, in this case, is the maximum possible on a bivariate plot. The related PC loadings plot (Fig. 7b) shows which elements exert control over each PC, enabling a visualisation of data variance related to geological processes. Here, magma differentiation appears to be the process controlling PC1 because Mg, Ca, Fe, Mn and Ti have large positive loadings, whereas K, LREEs and Si have negative loadings. PC2 appears to predominantly reflect mineral-specific fractionation processes typical of the lower arc crust (e.g. garnet, amphibole and lack of plagioclase) because MREEs, HREEs and Y (compatible in amphibole and garnet) have high positive loadings but Sr, Al, Na, Si and Eu (compatible in plagioclase) exhibit negative loadings.

A few key observations can be drawn from Fig. 7. First, the fertile magmas fall further along an igneous differentiation trend (lower PC1 scores) than unfertile rocks, consistent with the association of porphyry Cu deposits with intermediate-felsic arc magmas. Second, fertile magmas have experienced fractionation deeper in the crust (lower PC2 scores) compared to unfertile magmas in accordance with their high Sr/Y and La/Yb ratios and common association with arcs where the crust is thick. We find that porphyry Cu deposits associated with calc-alkaline magmas evolved at deeper levels in the crust (lower PC2 scores) than those associated with alkaline magmas. This is in line with the link between porphyry deposits in thicker arc crust (typically Cu-rich and Au-poor) with calc-alkaline magmas, and between porphyry deposits in thinner arc crust (typically Au-rich) with more alkaline magmas (Sillitoe 2000; Richards 2009; Chiaradia 2020; Park et al. 2021).

Machine learning results

Average performance metrics for each classifier are reported from the tenfold cross-validation process (Table 2), with individual metrics for each fold reported in ESM 1, Table S3. All algorithms show similarly strong performance for precision (0.84–0.88), recall (0.75–0.78), F1 score (0.79–0.80), FPR (0.10–0.14) and accuracy (0.81–0.83). These average scores are a significant improvement compared to those obtained from testing the classifiers of Loucks (2014) and Ahmed et al. (2019). For each supervised machine learning

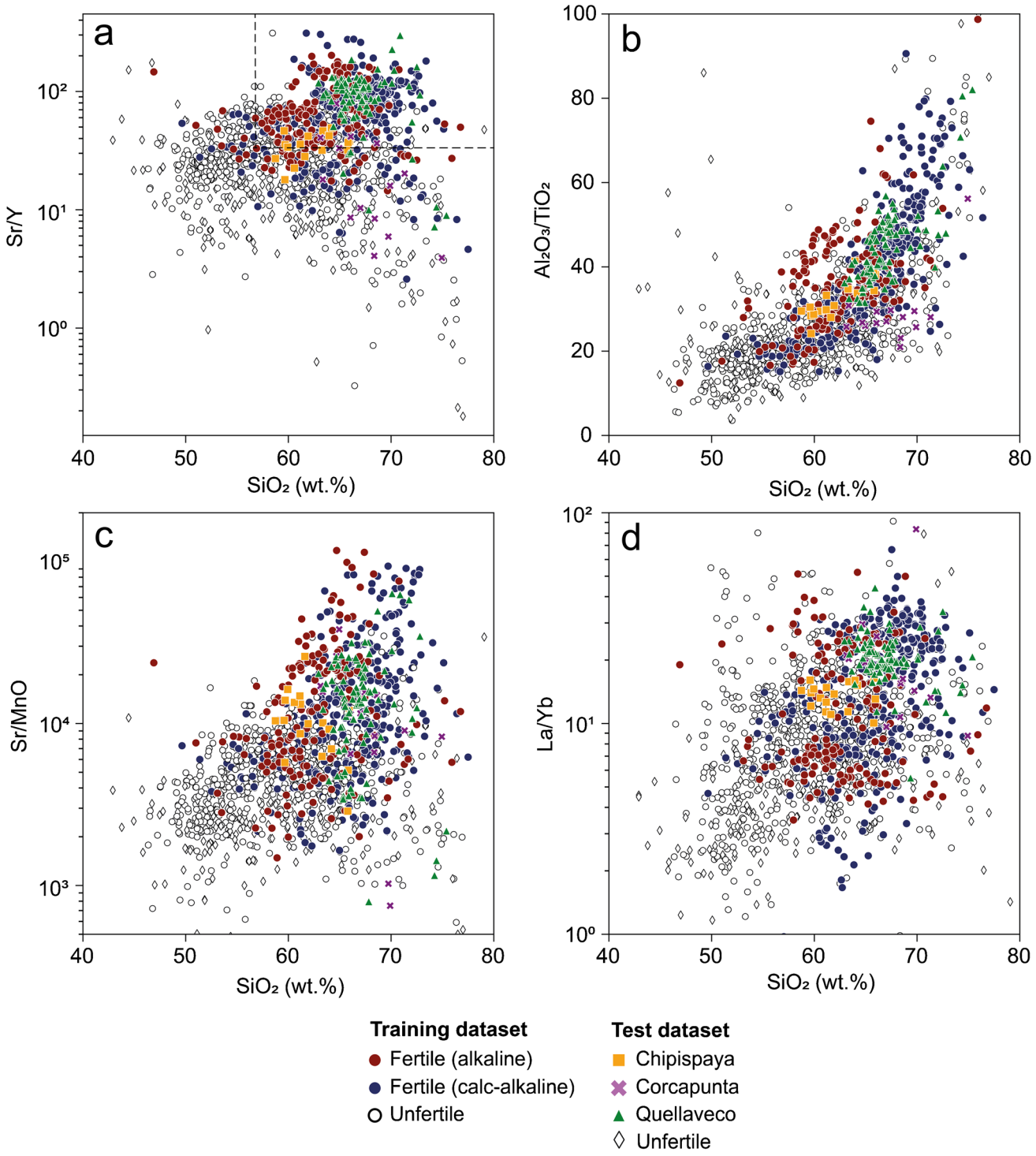

Fig. 5 Scatterplots of trace element ratios vs. SiO_2. (a) Sr/Y plot from Loucks (2014) with a dashed line indicating the Cu-fertile criterion of Sr/Y > 35 at SiO_2 > 57 wt%. (b) Al_2O_3/TiO_2 (Loucks 2014), (c) Sr/MnO (Ahmed et al. 2019). (d) La/Yb. Porphyry Cu rocks are separated based on their magma affinity where alkaline refers to rocks that classify as high-K calc-alkaline to shoshonitic compositions (Peccerillo and Taylor 1976). Compositions of the test dataset (Quellaveco, Chipispaya, Corcapunta and Coastal Batholith) are included for comparison

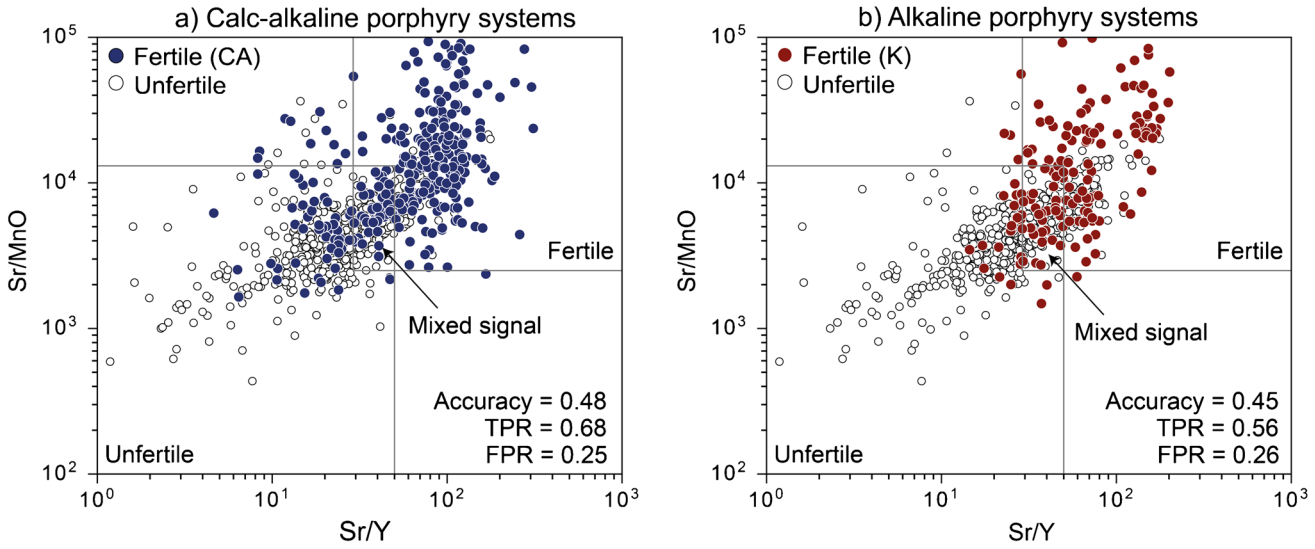

Fig. 6 Whole-rock geochemical data compilation plotted on the bivariate fertility classification scheme of Ahmed et al. (2019) for porphyry deposits associated with calc-alkaline magmatic suites (left) and high-K calc-alkaline-shoshonitic magma suites (right). Magma affinities are from literature, whole-rock geochemistry (Peccerillo and Taylor 1976) and reported nomenclature of igneous rocks, following the approach of Chiaradia (2020). The accuracy, true positive rate (TPR) and false positive rate (FPR) are calculated based on this classification scheme. These metrics are calculated using the same data used to train the machine learning models, hence allowing comparison between bivariate and machine learning classifiers. Data within the overlapping 'mixed zone' are assigned as misclassifications

technique studied here, the ROC curve was computed for each fold completed in the tenfold cross-validation process (ESM 1, Fig. S9). An average ROC curve for each classifier was determined using vertical averaging (Fawcett 2006), allowing comparison of the performance of the four classifiers (Fig. 8). For all four classification techniques used, the AUC varied between 0.87 and 0.89, indicating that there is an 87–89% probability that the classifiers rank a randomly selected fertile rock higher than a randomly selected unfertile rock (Fawcett 2006).

Independent tests of model performance

A limitation of using such model validation is that the testing of the classification techniques may use data from deposits or locations that the models have been trained on. To further test the strength of the classification techniques, the models were tested on four independent datasets containing data from three porphyry Cu systems that do not appear in the training dataset at all, plus additional withheld unfertile GEOROC data. The three porphyry Cu deposits tested are of varying size, type and tectonic setting, all found in the Peruvian Andes: (i) Quellaveco—a giant Palaoecene-Eocene porphyry Cu–Mo deposit (3000 Mt at 0.57% Cu); (ii) Corcapunta—a Miocene porphyry Cu–Mo prospect; and (iii) Chipispaya—a Miocene porphyry Cu–Au prospect. Whole-rock data for Corcapunta and Chipispaya are provided in ESM 2 and data for Quellaveco are from Nathwani et al. (2021). The data from Corcapunta and Chipispaya were acquired during the same analytical runs described in Nathwani et al. (2021), and quality control of data can be found therein. Distributions of elements in the training and test datasets were monitored to ensure no biases were present due to the different analytical methods (ESM 1, Fig. S10).

Classification of the fertile test datasets against the withheld unfertile data, using the four trained classification models, produced good classification performance (Fig. 9a; ESM 1 Table S4) with models giving accuracy scores of 0.93–0.97 (TPR = 0.95–1.00 and FPR = 0.06–0.13), except for the support vector machine which yielded an accuracy of 0.46 since it could not correctly classify any of the fertile observations (TPR = 0). All models gave AUC scores of 0.91–1.00.

A limitation of such model testing is that using GEOROC data as the negative class may not be wholly realistic because this database comprises mostly volcanic rocks, many of which are mafic-intermediate compositions not typically associated with porphyry Cu deposits. As a further test, we evaluated the ability of the models to discriminate the porphyry Cu test datasets from granitoids from the Coastal Cordillera, Chile (Jara et al. 2021), which arguably more realistically represent the unfertile lithologies that would be encountered in porphyry Cu exploration. For this test, classification performance is weaker, with accuracy = 0.68–0.71, TPR = 1.00 and FPR = 0.36–0.54. This indicates that all fertile rocks were correctly classified as fertile, but many false positives (or assumed false positives given that there may be undetected fertile systems present in the dataset) are present.

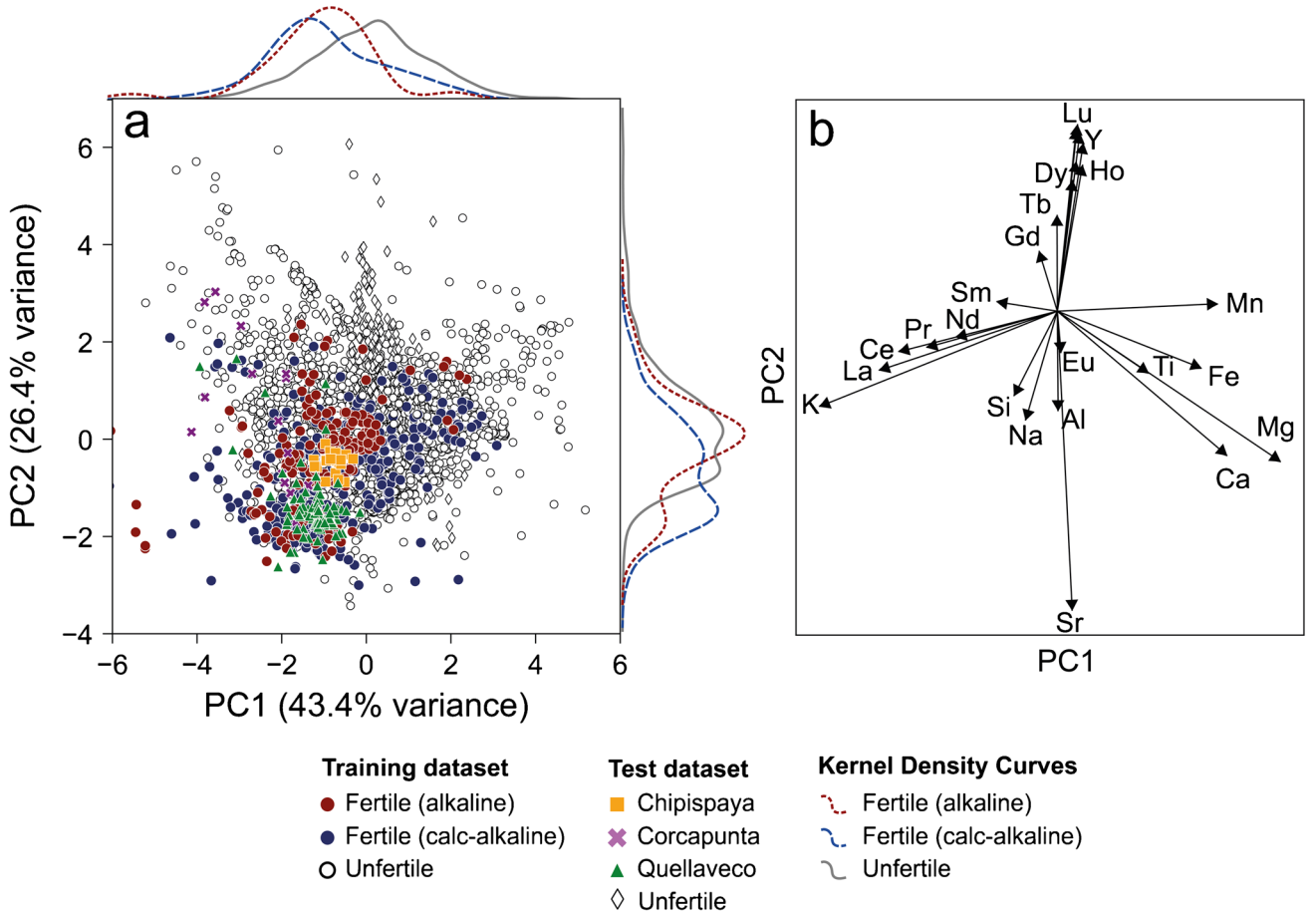

Fig. 7 Principal component analysis plots of the training and test (Chipispaya, Corcapunta and Quellaveco) datasets. (a) PC1 vs PC2 scores which illustrate the maximum variance possibly represented on a 2D plot (70% variance). (b) PC loadings plot for PC1 vs PC2 where vectors for each *clr*-transformed element show their relative loadings in each of the two PCs shown in (a)

Table 2 Mean performance metrics after a tenfold cross-validation of the training dataset (SVM = support vector machine, ANN = artificial neural network, LR = logistic regression, RF = Random Forest). Values in parentheses are 1-sigma standard deviations

	Accuracy	F1 Score	Precision	TPR (Recall)	ROC-AUC	FPR
SVM	0.81 (0.11)	0.79 (0.14)	0.86 (0.06)	0.75 (0.21)	0.88 (0.11)	0.12 (0.05)
ANN	0.83 (0.09)	0.80 (0.12)	0.88 (0.05)	0.76 (0.18)	0.89 (0.10)	0.10 (0.04)
LR	0.81 (0.09)	0.80 (0.12)	0.84 (0.06)	0.77 (0.16)	0.87 (0.09)	0.14 (0.04)
RF	0.82 (0.10)	0.80 (0.13)	0.87 (0.05)	0.76 (0.18)	0.89 (0.11)	0.11 (0.04)

However, apart from the support vector machine (SVM), the AUC scores are still high (0.97–1.00), indicating that despite the binary classification deteriorating, the models are still able to robustly predict a higher probability of fertility for the porphyry Cu-associated rocks compared to the unfertile data. This emphasises that the probabilities extracted from such models may be most useful, rather than the binary output alone. Overall, the strong classification performance validates the inference that magmatic processes associated with porphyry Cu formation are distinct from processes in typical magmatic arcs (e.g. Wilkinson, 2013), illustrating the additional knowledge that can be extracted by process validation (Grunsky and Caritat 2019).

Significantly, there is little difference in classification metrics between the various porphyry Cu deposits in the test dataset (ESM 1, Fig. S11); hence, the models could be used regardless of the magma affinity or deposit size/type. We note that many exploration programmes aim to discriminate fertile rocks from specific lithologies, in which case the training dataset could be refined, or sample weights could

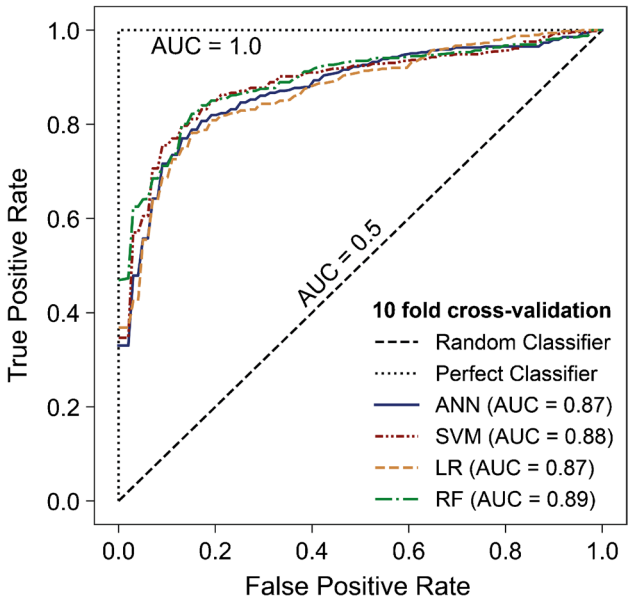

Fig. 8 Receiver operating characteristic (ROC) curves for each machine learning technique, showing the true positive rate versus false positive rate as the classification threshold is varied between 0 and 1. The ROC curve for each model is an average of 10 curves from the tenfold cross-validation, determined by the trapezoid rule. For reference, a random (AUC = 0.5) and perfect classifier (AUC = 1.0) are shown

be assigned during the training process whereby higher weighted observations more strongly control the position of the decision boundary.

Though Random Forest, logistic regression and artificial neural networks show comparably strong performance, we suggest Random Forest provides the most accessible and effective tool for geochemical exploration out of the classifiers studied here. The non-parametric properties of Random Forest, the in-built error prediction, feature importance evaluation and the ease of visualisation are all desirable properties. In Random Forest, feature importance can be assessed using an average of the mean decrease in the impurity, i.e. how much each variable reduces uncertainty when classifying data in a tree (Fig. 10a). Compositional components that appear at the top of classification trees have a higher importance as they produce the largest decrease in impurity. A second method is where features in a test dataset are permuted, and the test data are reclassified; any drop in the model performance after the feature is permuted is indicative of the feature importance (Fig. 10b). Lastly, Random Forest shows the least variability in prediction accuracy during hyperparameter tuning and performs strongly using the default hyperparameters (Probst et al. 2019), making it effective across a range of problems 'out of the box'.

Petrogenetic implications

A commonly cited limitation of supervised machine learning techniques is their 'black box' nature, in which the

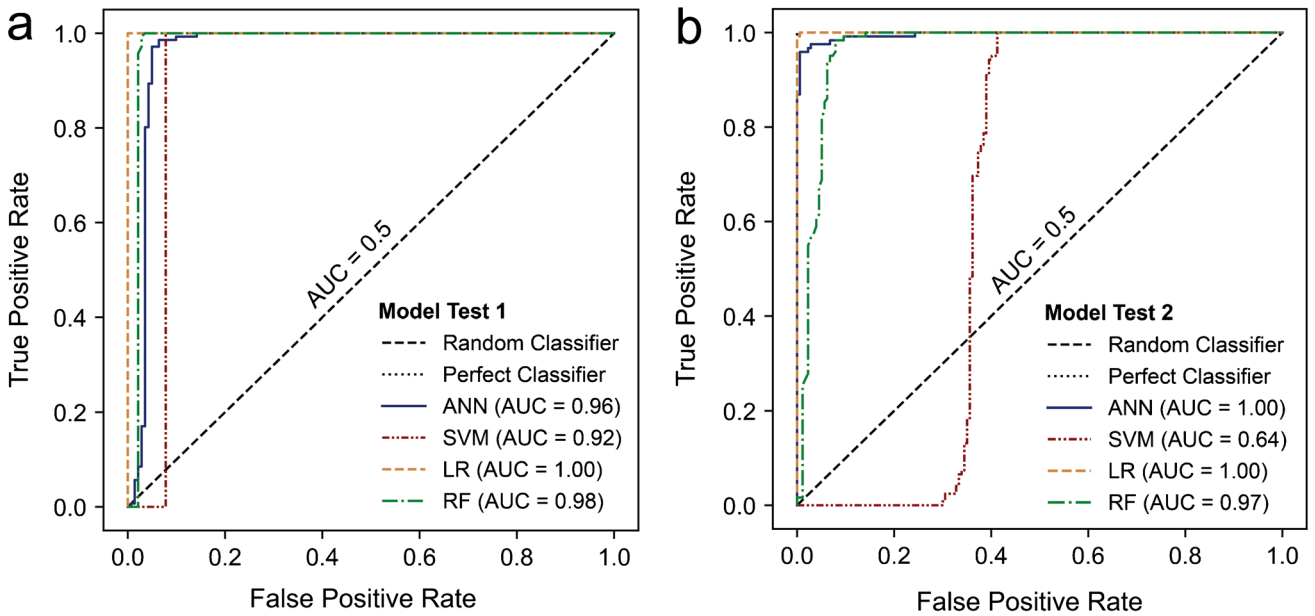

Fig. 9 Receiver operating characteristic (ROC) curves for model tests that were performed on data unseen in the training process. **a** ROC curves for each model when discriminating data from the Quellaveco, Chipispaya and Corcapunta porphyry Cu deposits from GEOROC data that was randomly sampled and withheld prior to model training. **b** ROC curves for each model when discriminating data from the Quellaveco, Chipispaya and Corcapunta porphyry Cu deposits from data from the Coastal Cordillera (Chile)

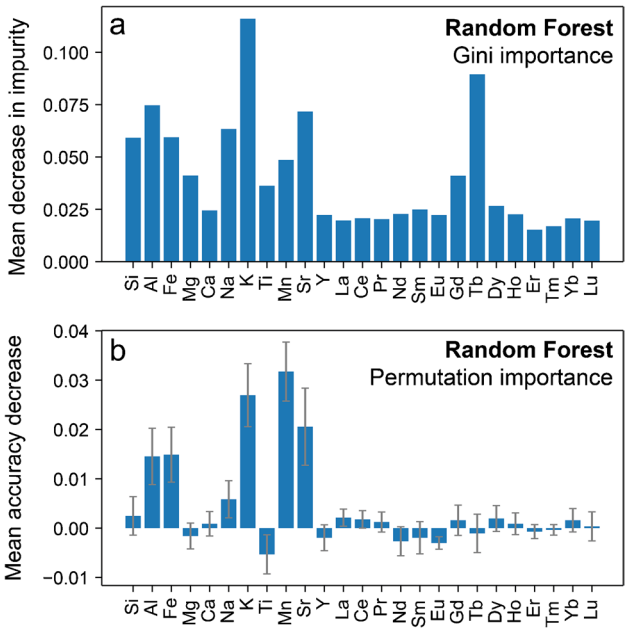

Fig. 10 Normalised feature importance scores generated by the Random Forest algorithm when discriminating fertile and unfertile data. **a** Mean decrease in impurity for each element. **b** Permutation importance where data from the test dataset were re-classified by the Random Forest model with each feature permuted to evaluate their influence on the output. Features were permuted 10 times each and the average decrease in accuracy is shown, where the error bar indicates the 1-sigma standard deviation. Negative mean accuracy indicates permuting the feature improved the model performance

procedure used to obtain the classification result is opaque (e.g. Papernot et al. 2017; Rudin 2019). This can lead to unexplainable methodologies that do not improve knowledge of the processes contributing towards data separation. Supervised machine learning models, which are 'black box', can make informed revision of imperfect classifications difficult and can impact the wider adoption of derived models. The use of feature importance analysis is one way to address this problem, where scores are assigned to features based on their importance in predicting the target variable. Quantified feature importance derived from petrological data using machine learning has been shown to provide insights into petrogenetic processes (Ueki et al. 2018; Petrelli et al. 2020; Lindsay et al. 2021).

Feature importance scores for the dataset interrogated here were calculated using Python's *sci-kit learn* implementation of Random Forest. This reveals that the most important features used to classify magma fertility in arc rocks by the trained model are K, Tb, Sr and Mn for Gini importance (Fig. 10a), and Mn, K, Sr and Fe for permutation importance (Fig. 10b).

The other classifiers used in this study do not contain inbuilt functions for feature importance. Fortunately, model explainability libraries such as SHAP (Lundberg and Lee 2017) and LIME (Ribeiro et al. 2016) can derive feature importance scores from many supervised machine learning techniques. Here we use SHAP (Shapely Additive exPlanations) to explain individual predictions using the *shap* library for Python (Lundberg and Lee 2017). SHAP values are calculated using a coalition game theory in which different coalitions of the feature set (i.e. numerous iterations of the models with all possible element combinations) are used to re-estimate the class prediction, and the difference in prediction when a specific feature is observed versus excluded is averaged. Individual compositional parameters are input into the SHAP model rather than the PC scores to allow better recognition of the relative importance of each parameter. Inputting elements as *clr*-coordinates instead of PC scores will produce slightly different models to those that were validated/tested; however, we prefer the additional interpretability in having feature importance scores assigned to individual chemical components and do not anticipate large differences between these models. We emphasise that such feature importance scores reflect the importance of a feature to the model, rather than the direct importance of this feature in nature.

Feature importance modelling is particularly challenging for features with high multicollinearity. As an example, because REEs are collinear, a model might predominantly use Tb rather than its neighbouring REEs since they do not provide any further information and thus rendering them redundant. This can lead to large differences in feature importance for such collinear features.

Using the SHAP library, feature importance scores were calculated for each model during classification of the test dataset and averages of these are reported (Fig. 11). Generally, the five highest ranked features are similar between the classifiers: support vector machine (Mn, Sr, Al, Tb and K), artificial neural network (Sr, Mn, Al, Ca and Tb), logistic regression (Al, Tb, Mn, Sr, Ce and La) and Random Forest (Al, K, Mn, Sr and Ti). A key advantage of SHAP is that feature importance scores can be calculated for individual compositions (ESM 1, Figs. S12-15). Analysis of these scores indicates that the components which display high concentrations in fertile rocks compared to the unfertile rocks are Al, Sr, K, LREEs and Ti, whereas components that are low in fertile rocks are Mn, MREE-HREEs and Ca.

The derived features of importance can be used to interpret petrogenetic processes that are key to forming magmas parental to porphyry Cu deposits. Low Mn is consistent with previous work which suggested that early fractionation of phases such as amphibole and garnet, in which Mn is compatible during high pressure magma differentiation in thickened arcs (e.g. $D^{amphibole/melt}_{Mn} = 1–28$; Nandedkar et al. 2016), can produce low-Mn derivative melts (Tang et al. 2020). Alternatively, low Mn may result from hydrothermal fluid loss during porphyry emplacement (Baldwin

and Pearce 1982). High Al and high Sr in porphyry-related magmatic suites have also been previously observed and related to suppression of plagioclase fractionation in hydrous melts at high pressures (Feiss 1978; Mason and Feiss 1979; Loucks 2014).

As noted above, interpreting individual feature importance scores for REEs is challenging due to multicollinearity. However, all models have identified Tb (and moderate importance for Dy and Ho) as the most useful REEs for classification, which broadly coincides with the peak amphibole-melt partition coefficients for the REEs (Nandedkar et al. 2016). This implicates amphibole (±titanite) fractionation in the lower crust during the generation of fertile arc magmas, in agreement with the numerous studies that have identified listric REE curves in igneous rocks associated with porphyry Cu deposits (Richards and Kerrich 2007; Richards, 2011; Loucks 2014; Nathwani et al. 2021). Our models have not identified a strong role for the HREEs in classifying fertile magmas which could suggest a lesser role of garnet (±zircon) in producing these magma compositions. Garnet is stable in magmas at higher pressures (> 0.8 GPa) and/or higher melt water contents (> 8 wt% H_2O) (Alonso-Perez et al. 2009) which are conditions that are often linked with porphyry Cu fertile magma evolution in the lower crust (e.g. Lee and Tang 2020). The inferred greater importance of amphibole fractionation in generating fertile magma compositions, as derived from our models, may reflect the most common intra-crustal conditions (pressure and melt water contents) under which parental magmas for porphyry Cu deposits are formed. Overall, the most important geochemical features are consistent with previous petrological studies of porphyry Cu deposit formation which emphasise crustal thickening fostered by strongly compressional tectonic settings and consequent magma evolution at depth (Rohrlach et al. 2005; Chiaradia et al., 2009; Richards et al. 2012; Chelle-Michou et al. 2015; Nathwani et al. 2021).

Machine learning biases and additional influences

The design, evaluation and application of machine learning algorithms in mineral exploration requires consideration of biases that may skew decisions towards a particular group. Many of these biases are data biases, since large datasets are often heterogeneous with many subgroups (Mehrabi et al. 2019). For example, the porphyry Cu deposits within the dataset explored here may be globally unrepresentative because studies tend to focus on the largest deposits in mature exploration terranes such as the Andes. Additionally, our models assume equal meaning and importance of a range of deposit styles, sizes and economic importance, in varying tectonic regimes (i.e. accumulation bias), which are not equally represented in the training dataset (see Table 1). There are also key differences between the two data classes. The fertile dataset is largely based on known mineral occurrences, where sampling

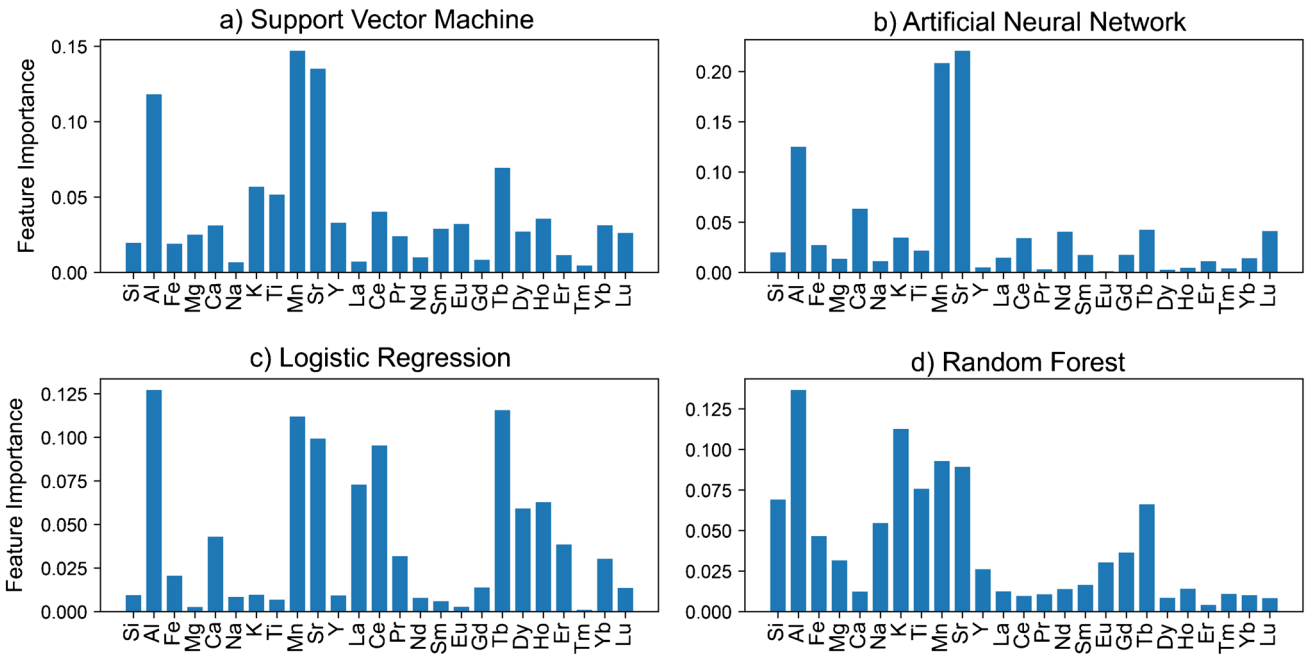

Fig. 11 Normalised feature importance (SHAP) scores for classification of the test datasets by each classifier. Higher feature importance scores indicate a larger importance in discriminating the data from unfertile/fertile arc rocks, as determined by coalition game theory. Asides from Random Forest, these were calculated based on a random subset of 10 observations from each test dataset because the SHAP calculations are more computationally expensive for these models

in the original studies focused on the weak to strongly altered porphyritic intrusions, mostly emplaced at < 6-km depth (Seedorff et al. 2005), with dense spatiotemporal sampling. By contrast, the unfertile dataset, mostly derived from GEOROC, contains a spectrum of mostly unaltered igneous rock types and compositions with a sparse geographical distribution, emplaced at a range of depths or erupted at surface. These biases are non-exhaustive, partly unavoidable and of varying importance based on the application, but can be partly mitigated with more rigorous sampling, sub-sampling and weighting of sub-groups during the training process.

Despite these limitations, we suggest our approach is valid because the signatures of magma fertility (Sr/Y, La/Yb and Sr/MnO) have been found both in volcanic rocks associated with porphyry deposits (e.g. Behnsen et al. 2021) and in the deeper source granitoids of porphyry systems (e.g. Ahmed et al., 2020). A broad background dataset is necessary in supervised machine learning for geochemical exploration to allow the classifiers to recognise the range of rock types that may be encountered during application. Re-running our models, where the GEOROC dataset was refined to only include plutonic rocks, only led to a small (~0.05 decrease in ROC-AUC) reduction in model performance and no systematic changes to feature importance (Figs. S16 and S17).

Despite the data filtering techniques and dimensionality reduction used to prepare the dataset, an additional bias arises because many of the rocks from porphyry deposits have some hydrothermal alteration. The addition and removal of elements during hydrothermal alteration associated with mineralisation could mean that the classifiers are partly discriminating differences in degree of alteration instead of differences in primary igneous geochemistry. Generally, a partly hydrothermally influenced classification model is still empirically useful, but clouds the underlying process understanding and may diminish classification accuracy because of the potential for significant spatial variability in alteration effects (Ulrich and Heinrich 2002; Cooke et al. 2014). Furthermore, these effects can partly to strongly obscure certain igneous signatures; for example, the mobility of Sr can impede its use in fingerprinting magma fertility (Wells et al., 2021). These hydrothermal effects are more likely for the most fluid mobile components, such as K, Sr, Mn, Ca and Fe. Despite this, we find most of the variance displayed by these elements is explainable by magmatic processes, as evidenced by PCA (Fig. 7 and ESM 1, Fig. S8).

Conclusions

We have demonstrated that four supervised machine learning techniques (logistic regression, artificial neural networks, support vector machines and Random Forest) can be trained to discriminate igneous rocks spatially and temporally associated with porphyry copper deposits from those unrelated to mineralised systems with high performance, regardless of magma affinity deposit type or size. This methodology is superior to more traditional bivariate classification schemes. Our generalised approach can be adapted for more bespoke exploration applications, such as targeting a more specific deposit type or size, using a more refined training dataset, or by applying weightings during the training process. Many of these techniques require critical pre-processing steps to account for many properties of geochemical datasets such as compositional data effects, sparsity, high multicollinearity and class imbalance. Random Forest potentially provides the most transparent and simple model, requiring little pre-processing of data.

Feature importance scores derived from the classifiers provide a degree of interpretability that can help to encourage model adoption. These scores also provide useful insights into petrogenetic processes associated with the formation of magmas parental to porphyry Cu deposits. The most important components that help to discriminate fertile arc rocks from unfertile rocks are high Al, low Mn, high Sr, high K and listric REE patterns. Several of these (Al, Mn, MREEs and Sr) are consistent with geochemical signatures produced during hydrous, high-pressure magmatic evolution in the lower crust where plagioclase fractionation is suppressed, and amphibole (± garnet) are abundant fractionating phases. Thus, the derived features of importance further validate previous petrological studies of porphyry Cu deposit–related magmas. Overall, our approach demonstrates the power of utilising the high dimension space of geochemical data for making informed classifications for efficient mineral exploration.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00126-021-01086-9.

Acknowledgements This work was supported by a Science and Solutions for a Changing Planet doctoral studentship, funded by the Natural Environment Research council (grant NE/L002515/1) and Anglo American. J. J. W., R. N. A. and D. J. S. acknowledge funding under Natural Environment Research Council grant (NE/P017452/1) 'From arc magmas to ores (FAMOS): A mineral systems approach'. We are grateful to Alex Lipp, Simon Large, Tom Matthews, Matt Loader, Julian Pearce, Frances Jenner and the FAMOS research consortium for discussions related to this work. We thank Matt Loader for the assistance with data compilation. This paper benefited greatly from constructive reviews by Eric Grunsky and Massimo Chiaradia, and we thank Celestine Mercer and Georges Beaudoin for the editorial handling.

Data availability All data used to support this study are referenced in the text and new data are provided in the Supplementary Material.

References

Abbaszadeh M, Hezarkhani A, Soltani-Mohammadi S (2015) Classification of alteration zones based on whole-rock geochemical data using support vector machine. J Geol Soc India 85:500–508. https://doi.org/10.1007/s12594-015-0242-3

Abedi M, Norouzi G-H, Bahroudi A (2012) Support vector machine for multi-classification of mineral prospectivity areas. Comput Geosci 46:272–283. https://doi.org/10.1016/j.cageo.2011.12.014

Ahmed A, Crawford AJ, Leslie C et al (2020) Assessing copper fertility of intrusive rocks using field portable X-ray fluorescence (pXRF) data. Geochemistry: Exploration Environment, Analysis 20:81–97https://doi.org/10.1144/geochem2018-077

Ahrens LH (1954) The lognormal distribution of the elements (A fundamental law of geochemistry and its subsidiary). Geochim Cosmochim Acta 5:49–73. https://doi.org/10.1016/0016-7037(54)90040-X

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

Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London, New York

Alonso-Perez R, Müntener O, Ulmer P (2009) Igneous garnet and amphibole fractionation in the roots of island arcs: experimental constraints on andesitic liquids. Contrib Mineral Petrol 157:541–558. https://doi.org/10.1007/s00410-008-0351-8

Alpaydın E (2014) Introduction to machine learning, 3rd edn. MIT Press, Cambridge, Mass

Baldwin JA, Pearce JA (1982) Discrimination of productive and non-productive porphyritic intrusions in the Chilean Andes. Econ Geol. https://doi.org/10.2113/gsecongeo.77.3.664

Ballard JR (2001) A comparative study between the geochemistry of ore-bearing and barren calc-alkaline intrusions. https://doi.org/10.25911/5D78DB47E57F9

Ballard JR, Palin MJ, Campbell IH (2002) Relative oxidation states of magmas inferred from Ce(IV)/Ce(III) in zircon: application to porphyry copper deposits of northern Chile. Contrib Mineral Petrol 144:347–364. https://doi.org/10.1007/s00410-002-0402-5

Behnsen H, Spandler C, Corral I, et al (2021) Copper-gold fertility of arc volcanic rocks: a case study from the Early Permian Lizzie Creek Volcanic Group, NE Queensland, Australia. Economic Geology. https://doi.org/10.5382/econgeo.4806

Bell J (2014) Machine learning: hands-on for developers and technical professionals. Wiley, Indianapolis, Indiana

Bourdon E (2002) Adakite-like lavas from antisana volcano (Ecuador): evidence for slab melt metasomatism beneath Andean Northern Volcanic Zone. J Petrol 43:199–217. https://doi.org/10.1093/petrology/43.2.199

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

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

Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) Classification and regression trees, 1st edn. Routledge

Cabrera J (2011) Estudio petrografico y petrologico de los porfidos alimentadores del districto mina Radomiro Tomic II Region, Chile. MSc thesis, Universidad de Concepción

Carranza EJM, Hale M (2001) Logistic regression for geologically constrained mapping of gold potential, Baguio District, Philippines. Explor Min Geol 10:165–175. https://doi.org/10.2113/0100165

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

Chayes F (1960) On correlation between variables of constant sum. J Geophys Res 1896–1977(65):4185–4193. https://doi.org/10.1029/JZ065i012p04185

Chelle-Michou C, Chiaradia M, Béguelin P, Ulianov A (2015) Petrological evolution of the magmatic suite associated with the Coroccohuayco Cu(–Au–Fe) Porphyry-Skarn Deposit, Peru. J Petrology 56:1829–1862. https://doi.org/10.1093/petrology/egv056

Chelle-Michou C, Chiaradia M, Ovtcharova M et al (2014) Zircon petrochronology reveals the temporal link between porphyry systems and the magmatic evolution of their hidden plutonic roots (the Eocene Coroccohuayco deposit, Peru). Lithos 198–199:129–140. https://doi.org/10.1016/j.lithos.2014.03.017

Chelle-Michou C, Rottier B Transcrustal magmatic controls on the size of porphyry Cu systems–State of knowledge and open questions. Society of Economic Geologists Special Publication 1:87–100. doi: https://doi.org/10.5382/SP.24.06

Chelle-Michou C, Rottier B, Caricchi L, Simpson G (2017) Tempo of magma degassing and the genesis of porphyry copper deposits. Sci Rep 7:40566. https://doi.org/10.1038/srep40566

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco California USA, pp 785–794

Cheng Z, Zhang Z, Chai F et al (2018) Carboniferous porphyry Cu–Au deposits in the Almalyk orefield, Uzbekistan: the Sarycheku and Kalmakyr examples. Int Geol Rev 60:1–20. https://doi.org/10.1080/00206814.2017.1309996

Chiaradia M (2020) Gold endowments of porphyry deposits controlled by precipitation efficiency. Nat Commun 11:248. https://doi.org/10.1038/s41467-019-14113-1

Chiaradia M (2009) Adakite-like magmas from fractional crystallization and melting-assimilation of mafic lower crust (Eocene Macuchi arc, Western Cordillera, Ecuador). Chem Geol 265:468–487. https://doi.org/10.1016/j.chemgeo.2009.05.014

Chiaradia M, Caricchi L (2017) Stochastic modelling of deep magmatic controls on porphyry copper deposit endowment. Sci Rep 7:44523. https://doi.org/10.1038/srep44523

Chiaradia M, Merino D, Spikings R (2009) Rapid transition to long-lived deep crustal magmatic maturation and the formation of giant porphyry-related mineralization (Yanacocha, Peru). Earth Planet Sci Lett 288:505–515. https://doi.org/10.1016/j.epsl.2009.10.012

Cooke DR, Hollings P, Walshe JL (2005) giant porphyry deposits: characteristics, distribution, and tectonic controls. Econ Geol 100:801–818. https://doi.org/10.2113/gsecongeo.100.5.801

Cooke DR, Hollings P, Wilkinson JJ, Tosdal RM (2014) Geochemistry of porphyry deposits. In: Treatise on Geochemistry. Elsevier, pp 357–381

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1023/A:1022627411411

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

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27:861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Feiss PG (1978) Magmatic sources of copper in porphyry copper deposits. Econ Geol 73:397–404. https://doi.org/10.2113/gsecongeo.73.3.397

Fiorentini ML, Garwin SL (2010) Evidence of a mantle contribution in the genesis of magmatic rocks from the Neogene Batu Hijau district in the Sunda Arc, South Western Sumbawa, Indonesia. Contrib Mineral Petrol 159:819–837. https://doi.org/10.1007/s00410-009-0457-7

Geranian H, Tabatabaei SH, Asadi HH, Carranza EJM (2016) Application of discriminant analysis and support vector machine in mapping gold potential areas for further drilling in the Sari-Gunay Gold Deposit, NW Iran. Nat Resour Res 25:145–159. https://doi.org/10.1007/s11053-015-9271-2

Gilmer AK, Sparks RSJ, Blundy JD et al (2018) Petrogenesis and assembly of the Don Manuel Igneous Complex, Miocene-Pliocene Porphyry Copper Belt, Central Chile. J Petrol 59:1067–1108. https://doi.org/10.1093/petrology/egy055

Gil-Rodriguez J (2010) Igneous petrology of the Colosa gold-rich porphyry system (Tolima, Colombia). PSM/EG thesis, Tucson (Arizona), USA, University of Arizona, 35p

Greenlaw L (2014) Surface lithogeochemistry of the Relincho porphyry copper-molybdenum deposit. Atacama Region, Chile 10(14288/1):0167019

Gregory DD, Cracknell MJ, Large RR et al (2019) Distinguishing ore deposit type and barren sedimentary pyrite using laser ablation-inductively coupled plasma-mass spectrometry trace element data and statistical analysis of large data sets. Econ Geol 114:771–786. https://doi.org/10.5382/econgeo.4654

Grondahl C, Zajacz Z (2017) Magmatic controls on the genesis of porphyry Cu–Mo–Au deposits: the Bingham Canyon example. Earth Planet Sci Lett 480:53–65. https://doi.org/10.1016/j.epsl.2017.09.036

Grunsky EC, de Caritat P (2019) State-of-the-art analysis of geochemical data for mineral exploration. Geochemistry: Exploration. Environment, Analysis 20:217–232. https://doi.org/10.1144/geochem2019-031

Harris JR, Grunsky EC (2015) Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. Comput Geosci 80:9–25. https://doi.org/10.1016/j.cageo.2015.03.013

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York, NY

Hu Y, Liu J, Ling M et al (2015) The formation of Qulong adakites and their relationship with porphyry copper deposit: geochemical constraints. Lithos 220–223:60–80. https://doi.org/10.1016/j.lithos.2014.12.025

Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. PMLR, pp 448–456

Itano K, Ueki K, Iizuka T, Kuwatani T (2020) Geochemical discrimination of monazite source rock based on machine learning techniques and multinomial logistic regression analysis. Geosciences 10:63. https://doi.org/10.3390/geosciences10020063

Japkowicz N, Stephen S (2002) The class imbalance problem: a systematic study1. IDA 6:429–449. https://doi.org/10.3233/IDA-2002-6504

Jara JJ, Barra F, Reich M et al (2021) Geochronology and petrogenesis of intrusive rocks in the Coastal Cordillera of northern Chile: insights from zircon U-Pb dating and trace element geochemistry. Gondwana Res 93:48–72. https://doi.org/10.1016/j.gr.2021.01.007

Kubat M (2017) An introduction to machine learning, 2nd ed. 2017. Springer International Publishing : Imprint: Springer, Cham

Lee C-TA, Tang M (2020) How to make porphyry copper deposits. Earth Planet Sci Lett 529:115868. https://doi.org/10.1016/j.epsl.2019.115868

Lee RG (2008) Genesis of the El Salvador porphyry copper deposit, Chile and distribution of epithermal alteration at Lassen Peak, California. PhD thesis, Oregon State University

Li H, Li X, Yuan F et al (2020) Convolutional neural network and transfer learning based mineral prospectivity modeling for geochemical exploration of Au mineralization within the Guandian-Zhangbaling area, Anhui Province. China Applied Geochemistry 122:104747. https://doi.org/10.1016/j.apgeochem.2020.104747

Li T, Zuo R, Xiong Y, Peng Y (2021) Random-drop data augmentation of deep convolutional neural network for mineral prospectivity mapping. Nat Resour Res 30:27–38. https://doi.org/10.1007/s11053-020-09742-z

Lindsay JJ, Hughes HSR, Yeomans CM et al (2021) A machine learning approach for regional geochemical data: platinum-group element geochemistry vs geodynamic settings of the North Atlantic Igneous Province. Geosci Front 12:101098. https://doi.org/10.1016/j.gsf.2020.10.005

Lipp AG, Shorttle O, Syvret F, Roberts GG (2020) Major element composition of sediments in terms of weathering and provenance: implications for crustal recycling. Geochemistry, Geophysics, Geosystems 21:e2019GC008758. https://doi.org/10.1029/2019GC008758

Loucks RR (2014) Distinctive composition of copper-ore-forming arcmagmas. Aust J Earth Sci 61:5–16. https://doi.org/10.1080/08120099.2013.865676

Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. arXiv:170507874[cs, stat]

Macpherson CG, Dreher ST, Thirlwall MF (2006) Adakites without slab melting: high pressure differentiation of island arc magma, Mindanao, the Philippines. Earth Planet Sci Lett 243:581–593. https://doi.org/10.1016/j.epsl.2005.12.034

Martín-Fernández JA (2003) Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Math Geol 35:253–278. https://doi.org/10.1023/A:1023866030544

Mason DR, Feiss PG (1979) On the relationship between whole rock chemistry and porphyry copper mineralization. Econ Geol 74:1506–1510. https://doi.org/10.2113/gsecongeo.74.6.1506

Maydagan L, Franchini M, Chiaradia M et al (2014) The altar porphyry Cu-(Au-Mo) deposit (Argentina): a complex magmatic-hydrothermal system with evidence of recharge processes. Econ Geol 109:621–641. https://doi.org/10.2113/econgeo.109.3.621

Mehrabi N, Morstatter F, Saxena N, et al (2019) A survey on bias and fairness in machine learning. arXiv:190809635 [cs]

Mokhtari AR (2014) Hydrothermal alteration mapping through multivariate logistic regression analysis of lithogeochemical data. J Geochem Explor 145:207–212. https://doi.org/10.1016/j.gexplo.2014.06.008

Moyen J-F (2009) High Sr/Y and La/Yb ratios: The meaning of the "adakitic signature." Lithos 112:556–574. https://doi.org/10.1016/j.lithos.2009.04.001

Müller D, Groves DI (2019) Potassic igneous rocks and associated gold-copper mineralization, 5th ed. 2019. Springer International Publishing : Imprint: Springer, Cham

Nandedkar RH, Hürlimann N, Ulmer P, Müntener O (2016) Amphibole–melt trace element partitioning of fractionating calc-alkaline magmas in the lower crust: an experimental study. Contrib Mineral Petrol 171:71. https://doi.org/10.1007/s00410-016-1278-0

Naranjo A, Horner J, Jahoda R et al (2018) La Colosa Au porphyry deposit, Colombia: mineralization styles, structural controls, and age constraints. Econ Geol 113:553–578. https://doi.org/10.5382/econgeo.2018.4562

Nathwani CL, Simmons AT, Large SJE et al (2021) From long-lived batholith construction to giant porphyry copper deposit formation: petrological and zircon chemical evolution of the Quellaveco District. Southern Peru Contrib Mineral Petrol 176:12. https://doi.org/10.1007/s00410-020-01766-1

Olson NH (2015) The geology, geochronology, and geochemistry of the Kaskanak Batholith, and other late Cretaceous to Eocene magmatism at the Pebble porphyry Cu-Au-Mo deposit, SW Alaska. MSc thesis, Oregon State University

Olson NH, Dilles JH, Kent AJR, Lang JR (2017) Geochemistry of the Cretaceous Kaskanak Batholith and genesis of the pebble porphyry Cu-Au-Mo deposit, Southwest Alaska. Am Miner 102:1597–1621. https://doi.org/10.2138/am-2017-6053

Papernot N, McDaniel P, Goodfellow I, et al (2017) Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. ACM, Abu Dhabi United Arab Emirates, pp 506–519

Park J-W, Campbell IH, Chiaradia M et al (2021) Crustal magmatic controls on the formation of porphyry copper deposits. Nat Rev Earth Environ 2:542–557. https://doi.org/10.1038/s43017-021-00182-8

Peccerillo A, Taylor SR (1976) Geochemistry of eocene calc-alkaline volcanic rocks from the Kastamonu area, Northern Turkey. Contr Mineral and Petrol 58:63–81. https://doi.org/10.1007/BF00384745

Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Petrelli M, Caricchi L, Perugini D (2020) Machine learning thermobarometry: application to clinopyroxene-bearing magmas. J Geophys Res Solid Earth 125. https://doi.org/10.1029/2020JB020130

Petrelli M, Perugini D (2016) Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data. Contrib Mineral Petrol 171:81. https://doi.org/10.1007/s00410-016-1292-2

Pollard PJ, Jongens R, Stein H, et al (2020) Rapid formation of porphyry and skarn copper-gold mineralization in a postsubduction environment: Re-Os and U-Pb geochronology of the Ok Tedi Mine, Papua New Guinea. Economic Geology

Porwal A, Carranza EJM, Hale M (2003) Artificial neural networks for mineral-potential mapping: a case study from Aravalli Province, Western India. Nat Resour Res 12:155–171. https://doi.org/10.1023/A:1025171803637

Porwal A, González-Álvarez I, Markwitz V et al (2010) Weights-of-evidence and logistic regression modeling of magmatic nickel sulfide prospectivity in the Yilgarn Craton, Western Australia. Ore Geol Rev 38:184–196. https://doi.org/10.1016/j.oregeorev.2010.04.002

Probst P, Boulesteix A-L, Bischl B (2019) Tunability: importance of hyperparameters of machine learning algorithms. J Mach Learn Res 20:1–32

Reich M (2001) Estudio petrográfico, mineraloquímico y geoquímico de los cuerpos intrusivos de Sewell y La Huifa, Yacimiento El Teniente, VI Región. Memoria de Título, Universidad de Concepción, Chile

Reimann C, Filzmoser P (2000) Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environ Geol 39:1001–1014. https://doi.org/10.1007/s002549900081

Rezeau H, Moritz R, Leuthold J et al (2017) 30 Myr of Cenozoic magmatism along the Tethyan margin during Arabia-Eurasia accretionary orogenesis (Meghri–Ordubad pluton, southernmost Lesser Caucasus). Lithos 288–289:108–124. https://doi.org/10.1016/j.lithos.2017.07.007

Rezeau H, Moritz R, Wotzlaw J-F et al (2016) Temporal and genetic link between incremental pluton assembly and pulsed porphyry Cu-Mo formation in accretionary orogens. Geology 44:627–630. https://doi.org/10.1130/G38088.1

Ribeiro MT, Singh S, Guestrin C (2016) "Why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Francisco California USA, pp 1135–1144

Richards JP (2011) High Sr/Y arc magmas and porphyry Cu ± Mo ± Au deposits: just add water. Econ Geol 106:1075–1081. https://doi.org/10.2113/econgeo.106.7.1075

Richards JP (2013) Giant ore deposits formed by optimal alignments and combinations of geological processes. Nat Geosci 6:911–916. https://doi.org/10.1038/ngeo1920

Richards JP (2009) Postsubduction porphyry Cu-Au and epithermal Au deposits: products of remelting of subduction-modified lithosphere. Geology 37:247–250. https://doi.org/10.1130/G25451A.1

Richards JP, Boyce AJ, Pringle MS (2001) Geologic evolution of the Escondida Area, Northern Chile: a model for spatial and temporal localization of porphyry Cu mineralization. Econ Geol 96:271–305. https://doi.org/10.2113/gsecongeo.96.2.271

Richards JP, Kerrich R (2007) Special paper: adakite-like rocks: their diverse origins and questionable role in metallogenesis. Econ Geol 102:537–576. https://doi.org/10.2113/gsecongeo.102.4.537

Richards JP, López GP, Zhu J-J et al (2017) Contrasting tectonic settings and sulfur contents of magmas associated with cretaceous porphyry Cu ± Mo ± Au and intrusion-related iron oxide Cu-Au deposits in Northern Chile *. Econ Geol 112:295–318. https://doi.org/10.2113/econgeo.112.2.295

Richards JP, Spell T, Rameh E et al (2012) High Sr/Y magmas reflect arc maturity, high magmatic water content, and porphyry Cu ± Mo ± Au potential: examples from the tethyan arcs of Central and Eastern Iran and Western Pakistan. Econ Geol 107:295–332. https://doi.org/10.2113/econgeo.107.2.295

Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M (2015) Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol Rev 71:804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Rohrlach BD (2002) Tectonic evolution, petrochemistry, geochronology and palaeohydrology of the Tampakan porphyry and high sulphidation epithermal Cu-Au deposit Mindanao, Phillipines. xxvi, 499, [139] leaves. https://doi.org/10.25911/5D7638DDAB226

Rohrlach BD, Loucks RR, Porter TM (2005) Multi-million-year cyclic ramp-up of volatiles in a lower crustal magma reservoir trapped below the Tampakan copper-gold deposit by Mio-Pliocene crustal compression in the southern Philippines. Super Porphyry Copper and Gold Deposits: A Global Perspective: Adelaide, PGC Publishing 2:369–407

Rojas A (2003) Porfido Teniente: Dos fases intrusivas caracteristicas geológicas, petrográficas y geoquímicas, yacimiento El Teniente [abs]. Concepción, Chile, 2003, Abstract Volume 9

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215. https://doi.org/10.1038/s42256-019-0048-x

Schutte P, Chiaradia M, Beate B (2010a) Petrogenetic evolution of arc magmatism associated with Late Oligocene to Late Miocene porphyry-related ore deposits in Ecuador. Econ Geol 105:1243–1270. https://doi.org/10.2113/econgeo.105.7.1243

Schütte P, Chiaradia M, Beate B (2010b) Geodynamic controls on Tertiary arc magmatism in Ecuador: constraints from U-Pb zircon geochronology of Oligocene-Miocene intrusions and regional age distribution trends. Tectonophysics 489:159–176. https://doi.org/10.1016/j.tecto.2010.04.015

E Seedorff JH Dilles JM Proffett et al 2005 Porphyry Deposits: Characteristics and Origin of Hypogene Features 10.5382/AV100.1

Sillitoe RH (2010) Porphyry copper systems. Econ Geol 105:3–41. https://doi.org/10.2113/gsecongeo.105.1.3

Sillitoe RH (1997) Characteristics and controls of the largest porphyry copper-gold and epithermal gold deposits in the circum-Pacific region. Aust J Earth Sci 44:373–388. https://doi.org/10.1080/08120099708728318

RH Sillitoe 2000 Gold-Rich Porphyry Deposits: Descriptive and Genetic Models and Their Role in Exploration and Discovery 10.5382/Rev.13.09

Simmons AT (2013) Magmatic and hydrothermal stratigraphy of Paleocene and Eocene porphyry Cu-Mo deposits in southern Peru. PhD Thesis, University of British Columbia

Skewes MA, Stern CR (1995) Genesis of the Giant Late Miocene to Pliocene copper deposits of Central Chile in the Context of Andean Magmatic and Tectonic Evolution. Int Geol Rev 37:893–909. https://doi.org/10.1080/00206819509465432

Srivastava N, Hinton G, Krizhevsky A et al (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958

Stern CR, Skewes MA (1995) Miocene to present magmatic evolution at the northern end of the Andean Southern Volcanic Zone, Central Chile. Andean Geology 22:261–272

Stern CR, Skewes MA, Arévalo A (2011) Magmatic evolution of the Giant El Teniente Cu–Mo Deposit, Central Chile. J Petrol 52:1591–1617. https://doi.org/10.1093/petrology/egq029

Tang M, Lee C-TA, Rudnick RL, Condie KC (2020) Rapid mantle convection drove massive crustal thickening in the late Archean. Geochim Cosmochim Acta 278:6–15. https://doi.org/10.1016/j.gca.2019.03.039

Topuz G, Altherr R, Schwarz WH et al (2005) Post-collisional plutonism with adakite-like signatures: the Eocene Saraycık granodiorite (Eastern Pontides, Turkey). Contrib Mineral Petrol 150:441–455. https://doi.org/10.1007/s00410-005-0022-y

Toro JC, Ortúzar J, Zamorano J et al (2012) Protracted magmatic-hydrothermal history of the Río Blanco-Los Bronces district, Central Chile: development of world's greatest known concentration of copper. Society of Economic Geologists Special Publication 16:105–126

Ueki K, Hino H, Kuwatani T (2018) Geochemical discrimination and characteristics of magmatic tectonic settings: a machine-learning-based approach. Geochem Geophys Geosyst 19:1327–1347. https://doi.org/10.1029/2017GC007401

Ulrich T, Heinrich CA (2002) Geology and alteration geochemistry of the porphyry Cu-Au deposit at Bajo de la Alumbrera, Argentina. Econ Geol 97:1865–1888. https://doi.org/10.2113/gsecongeo.97.8.1865

van Buuren S (2012) Flexible imputation of missing data, 0 edn. Chapman and Hall/CRC

Verma SP, Torres-Alvarado IS, Velasco-Tapia F (2003) A revised CIPW norm. Swiss Bulletin of Mineralogy and Petrology 83:197–216

Vermeesch P (2006) Tectonic discrimination of basalts with classification trees. Geochim Cosmochim Acta 70:1839–1848. https://doi.org/10.1016/j.gca.2005.12.016

von Quadt A, Erni M, Martinek K et al (2011) Zircon crystallization and the lifetimes of ore-forming magmatic-hydrothermal systems. Geology 39:731–734. https://doi.org/10.1130/G31966.1

Vry VH (2010) Geological and hydrothermal fluid evolution at El Teniente, Chile. PhD Thesis, Imperial College London

Weiss GM, Provost F (2003) Learning when training data are costly: the effect of class distribution on tree induction. jair 19:315–354. https://doi.org/10.1613/jair.1199

Wells TJ, Meffre S, Cooke DR et al (2021) Assessment of magmatic fertility using pXRF on altered rocks from the Ordovician Macquarie Arc, New South Wales. Aust J Earth Sci 68:397–409. https://doi.org/10.1080/08120099.2020.1782471

Wilkinson JJ (2013) Triggers for the formation of porphyry ore deposits in magmatic arcs. Nature Geosci 6:917–925. https://doi.org/10.1038/ngeo1940

Williams M, Schoneveld L, Mao Y, et al (2020) pyrolite: Python for geochemistry. JOSS 5:2314. https://doi.org/10.21105/joss.02314

Yeomans CM, Shail RK, Grebby S et al (2020) A machine learning approach to tungsten prospectivity modelling using knowledge-driven feature extraction and model confidence. Geosci Front 11:2067–2081. https://doi.org/10.1016/j.gsf.2020.05.016

Zhang S, Xiao K, Carranza EJM et al (2019) Integration of auto-encoder network with density-based spatial clustering for geochemical anomaly detection for mineral exploration. Comput Geosci 130:43–56. https://doi.org/10.1016/j.cageo.2019.05.011

Zhao J, Chen S, Zuo R (2016) Identifying geochemical anomalies associated with Au–Cu mineralization using multifractal and artificial neural network models in the Ningqiang district, Shaanxi, China. J Geochem Explor 164:54–64. https://doi.org/10.1016/j.gexplo.2015.06.018

Zuo R, Carranza EJM (2011) Support vector machine: a tool for mapping mineral prospectivity. Comput Geosci 37:1967–1975. https://doi.org/10.1016/j.cageo.2010.09.014