Integration of sparse and continuous datasets using machine learning for core mineralogy interpretation

Mayur Nawal[1], Bharath Shekar[*,1,2] and Priyank Jaiswal[3]

[1]Department of Earth Sciences, Indian Institute of Technology Bombay, Mumbai, India.

[2]National Centre of Excellence in Carbon Capture and Utilization, Indian Institute of Technology Bombay, Mumbai, India. E-mail: bshekar@iitb.ac.in.

[3] Boone Pickens School of Geology, Oklahoma State University, USA

In Earth Science, integrating non-invasive continuous data streams with discrete invasive measurements remains an open challenge. In this paper, we address such a problem, that of predicting whole-core mineralogy using discrete measurements with the help of machine learning (ML). Our targets are sparsely sampled mineralogy from X-Ray Diffraction (XRD) and features are continually sampled elemental oxides from X-Ray Fluorescence (XRF). Both datasets are acquired on a core cut from Mississippian age mixed siliciclastic-carbonate formation in the US mid-continent. The novelty of this paper is predicting multiple classes of output targets from input features in a small multi-dimensional data setting. Our workflow has three salient aspects. First, it shows how single output models are more effective in relating selective target-feature subsets than using a multi-output model for simultaneously relating the entire target-feature set. Specifically, we adopt a competitive ensemble strategy comprising three classes of regression algorithms - elastic-net (linear regression), XGBoost (tree-based), and feedforward neural networks (non-linear regression). Second, it shows that feature selection and engineering, when done using statistical relationships within the dataset and domain knowledge, can significantly improve target predictability. Thirdly, it incorporates k-fold cross-validation and grid-search-based parameter tuning to predict targets within 4-6% accuracy using 40% training data. Results open doors to generating a wealth of information in energy, environmental and climate sciences where remotely sensed data is cheap and abundant, while physical sampling may be limited due to analytical, logistic, or economic issues.

1. Introduction

Mineralogical data are used by petroleum engineers to characterize petroleum geology and deduce reservoir intervals (Selley, 2015; Jonge-Anderson *et. al*, 2022) and in the assessment of stimulation potential of reservoirs (Alsaif *et al.*, 2017). These data are also necessary for creating petrophysical models that can guide the drill bit. Ideally, such information should be continuous along the whole core. However, it is currently determined at discrete locations only due to the limitations of the analysis process. The most common method of determining mineralogy is through X-Ray Diffraction (XRD). In this method, rock samples are first ground into a powdered form followed by their illumination with X-rays. The resultant diffraction patterns are then interpreted for mineralogy (Buhrke *et al.*, 1997). The destructive nature of XRD sample preparation along with the time required to prepare them, restricts their extensive usage on cores, which are difficult and expensive to obtain in the first place. Non-destructive ways of remotely sensing mineral chemistry also exist. The X-Ray Fluorescence (XRF) method uses X-Ray to knock off an outer shell electron. Fluorescence radiation is emitted as the electron from an outer shell occupies the vacant position in an inner shell. The fluorescence energy, which is equal to the energy difference between the two shells, is characteristic of the atom. Thus, although XRF allows the determination of elemental composition in a continual manner along the core, it does not provide direct measurements of bulk mineralogy required for reservoir applications. By integrating XRD and XRF datasets, continuous chemistry along

the whole core can be obtained, which can be beneficial for reservoir development (Aoudia *et al.,* 2010, Meller and Ledésert, 2017). However, the problem is difficult because the two datasets often complement each other in non-intuitive ways. Previously, this integration has been attempted empirically, e.g., Kozlov *et al.* (2020), Hupp and Donovan (2018) and Lousber and Verryn (2008), as well as through novel instrumentation, e.g., Bortolotti *et al.* (2017), Vaniman *et al.* (1998) and Yellepeddi *et al.* (1996). While these methods have yielded good results, they are interpretive and occasionally require customized sample preparation. A robust and highly automated approach to integrate XRD-XRF data sets acquired with traditional instrumentation can help in the analysis of large volumes of existing datasets. Here, building upon the motivational work of Alnahwi and Loucks (2019), we have integrated XRD and XRF datasets using machine learning (ML) The novelty of this paper is predicting multiple classes of output targets from input features in a small multi-dimensional data setting.

Alnahwi and Loucks (2019) used artificial neural networks (ANN), which are one of the most basic ML models, to integrate XRD and XRF datasets and obtain continuous reservoir chemistry. Although the results were shown with limited data, the study opened doors to new opportunities and forms the basis of this paper. The generic field of ML aims to develop self-evolving and dynamically trainable models that can predict future outcomes by discovering hidden relationships between variables. Simple analytics tools such as regression, clustering, and support vector machines have been commonly used for exploring variable dependencies in geosciences. In the last decade, advances in computing power have popularized methods like neural networks and deep learning, which have rapidly matured in the hands of a prolific and rapidly increasing user base (Yu and Ma,

2021). Machine learning applications in Earth Sciences include sonic log prediction with stacked neural networks (Misra and He, 2019), convolutional neural networks (Kanfar *et al.*, 2020) and memory networks (Pham *et al.*, 2020) and lithofacies identification with support vector machines (Liu *et al.*, 2020) and Auto-ML approach (Nawal *et al.*, 2022). Deep learning techniques have also been applied in seismic inversion for elastic properties and in the integration of seismic and log data (Alfarraz *et al.*, 2019; Das *et al.*, 2019). Alabbad *et al.* (2021) predicted gas saturation from the core and well log data using four different supervised machine learning algorithms that included linear regression, random forests, feed-forward neural networks, and long short-term memory networks. Zeeshan *et al.* (2019) integrated core and log data to predict the static Poisson's ratio of carbonates using artificial intelligence algorithms. Kim et al. (2020) used deep neural networks to predict mineralogy from well log and XRD data.

Although recent efforts have focussed on physics-based simulations with neural networks (Raissi *et al.*, 2019), most implementations of ML algorithms are predictive in nature and are usually agnostic to underlying causative relationships. Therefore, their success has to be thoroughly validated to avoid the risk of it being coincidental. Risk is higher for small datasets where optimization and validation opportunities are limited, e.g., Karpatne *et al.* (2019) and Qi and Carr (2006). Olson *et al.* (2018) define a "small-data" problem as one where the measured data (e.g., features or target or both) lie between 10 and 10,000 in the count. The limiting number is somewhat arbitrary and is defined with respect to the complexity of the problem. For example, limited testing due to dataset size can lead to memorization and lack of generalization. Methods such as early stopping (Pasini, 2015), semi-supervised learning (Hady and Schwenker, 2013), ensemble, weighted loss functions

(Zhou, 2009), meta-learning (Hospedales *et al.*, 2020) and one/zero-shot methods (Hilprecht and Binnig, 2021) can be effective with a small data set. Uncertainties associated with small datasets have been addressed for random forests (Feng *et al.*, 2021a) and Bayesian neural networks (Feng *et al.*, 2021b). Interestingly, a number of routine measurements in earth science that physically involve rocks and sediments, e.g., grain size and isotopes, fall in the "small-data" category. Researchers have overcome the limitations posed due to limited data in several ways, for example, by feature selection and engineering (Dernoncourt et al., 2014), by data augmentation with fancy PCA (Kim et al., 2020), use of regularization (Zhang and Ling, 2018), with nested k-fold (Vabalas et al., 2019) and ensembling techniques (Rasolomanana and Onjaniaina, 2021). In this paper, drawing on peer research, we have used feature selection and engineering as well as the ensemble method to extract the most information from our dataset.

Alnahwi and Loucks (2019) generated their dataset on three cores totaling a footage of 212.1m. The cores were cut from the Eagle Ford shale formation, La Salle County, Texas. It comprised 35 XRD samples distributed non-uniformly and XRF measurements were generated uniformly at 2-in intervals along the core length. They constructed Feedforward Neural Network (FNN) models to predict six targets (XRD mineralogy) from six features (XRF elemental oxides). The features were magnesium (Mg), aluminum (Al), silicon (Si), sulfur (S), potassium (K), calcium (Ca), and iron (Fe) and the targets were calcite (CaCO3), dolomite (CaMg(CO3)2), quartz (SiO2), pyrite (FeS), feldspar (aluminosilicates), and clay (layered hydrated aluminosilicates). Subsets of features were related to individual targets through slim FNN models (up to two hidden layers with up to three neurons). For model construction, data were partitioned 60% for training, 20% for cross-validation and 20% for

testing. The trained network yielded high r^2 scores (0.85 – 0.95) between true and predicted targets. Although the data volume was at the limit of statistical significance, Alnahwi and Loucks (2019) showed that the integration was possible through careful feature selection and hyper-parameter tuning. An interesting aspect of their work was developing target-specific models through a common feature pool. Besides testing this idea for larger datasets, an obvious next-step is to make the search model-specific in addition to target-specific. For example, say if FNN, which is a non-linear regression model, is suitable for quartz, could an entirely different model, such as XGBoost which is a decision-tree model, be better for clay?

Building on Alnahwi and Loucks (2019), we have attempted XRD-XRF integration using a target- and model-specific strategy and a more expanded dataset (52 XRD measurements; targets). Specifically, we create an ensemble of three competing models, i.e., elastic-net (regularized linear regression), XGBoost (tree-based), and FNN (non-linear regression). We have reinforced three ML "best practices''. First, we have examined the effectiveness of single output models for relating selective target-feature subsets as opposed to the popularly practiced multi-output model that can simultaneously relate the entire target-feature set. Second, we have examined the use of feature selection and engineering using statistical relationships within the dataset and domain knowledge to aid target predictability. Third, we have explored grid-search-based parameter tuning and k-fold cross-validation for ranking the models within the ensemble. The paper is organized as follows: we first describe data and their acquisition, followed by a description of the machine learning pipeline consisting of feature engineering and selection, individual ML models and the competitive ensemble strategy. Finally, we apply the models to our data

and discuss the results. Although shown in the context of XRD-XRF, the application can be extended to other small data problems in Earth Science requiring integration of geochemical (e.g., isotopes), sedimentological (grain size), and geophysical (e.g., acoustics and resistivity) datasets.

2. Data

A core from a reservoir rock from Canadian County, Oklahoma (Figure 1) was analyzed using X-ray fluorescence (XRF) and X-ray diffraction (XRD) experiments to generate the dataset used in this study (Raj, 2021). The main reservoir is of mixed siliciclastic-carbonate Mississippian age formation, and spans ~ 500 ft (10,400 – 10,900 ft). Following the standard practice, the core was longitudinally sliced to remove a slab that was a third of its thickness. From the slab, a number of plugs were extracted covering as many of the geological facies as possible. Approximately 0.2 inch portions from one end of the plugs were cut, powdered in a SPEX ball mill and with the use of a mortar-and-pestle and analyzed in a Rigaku MiniFlex Diffraction machine. Calibration of the machine was done on quartz. A powder diffraction file was used to match and identify minerals and the Rietveld refinement scheme was used to quantify the identified minerals using the commercially available software RIQAS (Ramkumar *et al.*, 2018).

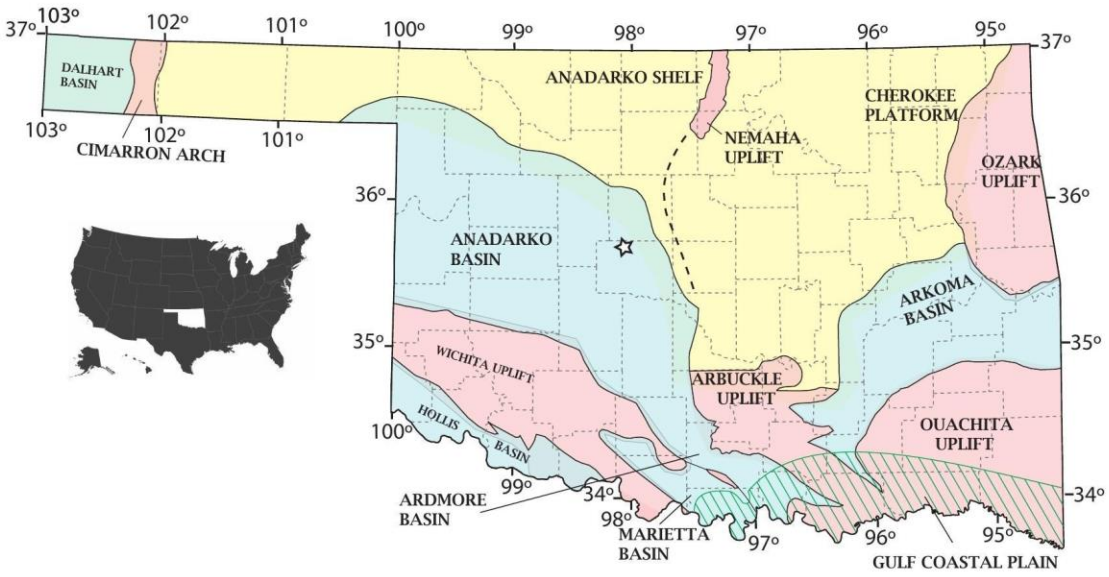

Figure1: Base map. State map of Oklahoma with major geological features (after Northcutt and Campbell, 1995). The location of the core used to generate the X-Ray Diffraction (XRD) and X-Ray Fluorescence (XRF) data used in this paper is marked with a star. Inset shows the location of Oklahoma within the United States of America.

The minerals identified using XRD were categorized into four groups for the purposes of this paper: quartz, carbonate, clay, and “others". The first three groups together constituted at least 85% of every sample (Figure 2a). Compositionally, quartz which is silica (SiO_2) and constituted 20 – 75% mass fraction of the samples. Carbonate comprised calcite ($CaCO_3$) and dolomite ($MgCO_3.CaCO_3$) with occasional siderite ($FeCO_3$) in fractional amounts. Carbonate was highly variable and constituted up to 60% in some samples. Calcite was the most dominant mineral in the carbonate group. Clay had the following minerals – kaolinite (<1%; $Al_2O_3\ 2SiO_2{\cdot}2H_2O$), chamosite (0 – 6%; $(Fe^{2+},Mg)_5Al(AlSi_3O_{10})(OH)_8$), illite (0 – 15%; $(K,H_3O)(Al,Mg,Fe)_2(Si,Al)_4O_{10}[(OH)_2,(H_2O)]$, smectite (0 – 1%; $[Na, Ca]_3{\cdot}[4\text{-}5]H_2O$

$[(Al_{1.5}Fe_2^{3+}Mg_3)Si_4O_{10}(OH)_2]^{-.3}$) and mixed illite-smectite layers. Finally, the grouping "others" consisted of muscovite (0 – 3%, $KAl_2(AlSi_3O_{10})(F,OH)_2$), microcline (0 – 4%; $KAlSi_3O_8$), albite (3 – 10%; $NaAlSi_3O_8$) and pyrite (1 – 6%; FeS_2).

The XRF data were acquired within the Mississippian formation using the hand-held Bruker T5 instrument at a 1-ft interval. Complementary data were also acquired using the Fourier Transform Infrared (FTIR) method, and the two datasets were together used to obtain the chemical signature of the core sample in terms of proportion of various oxides (as % of total mass) and elements (in ppm). In this dataset, six dominant oxides, SiO2, Al2O 3, Fe2O3, MgO, CaO, and K2O, (termed as "features"; Figure 2 b – g) contributed to the bulk of the mineralogy (termed as "targets"; Figure 3). The three targets, i.e. quartz, carbonate, and clay, had an increasing order of complexity. For example, the quartz target is comprised entirely of one feature, SiO2, the carbonate target is dominated by two features,

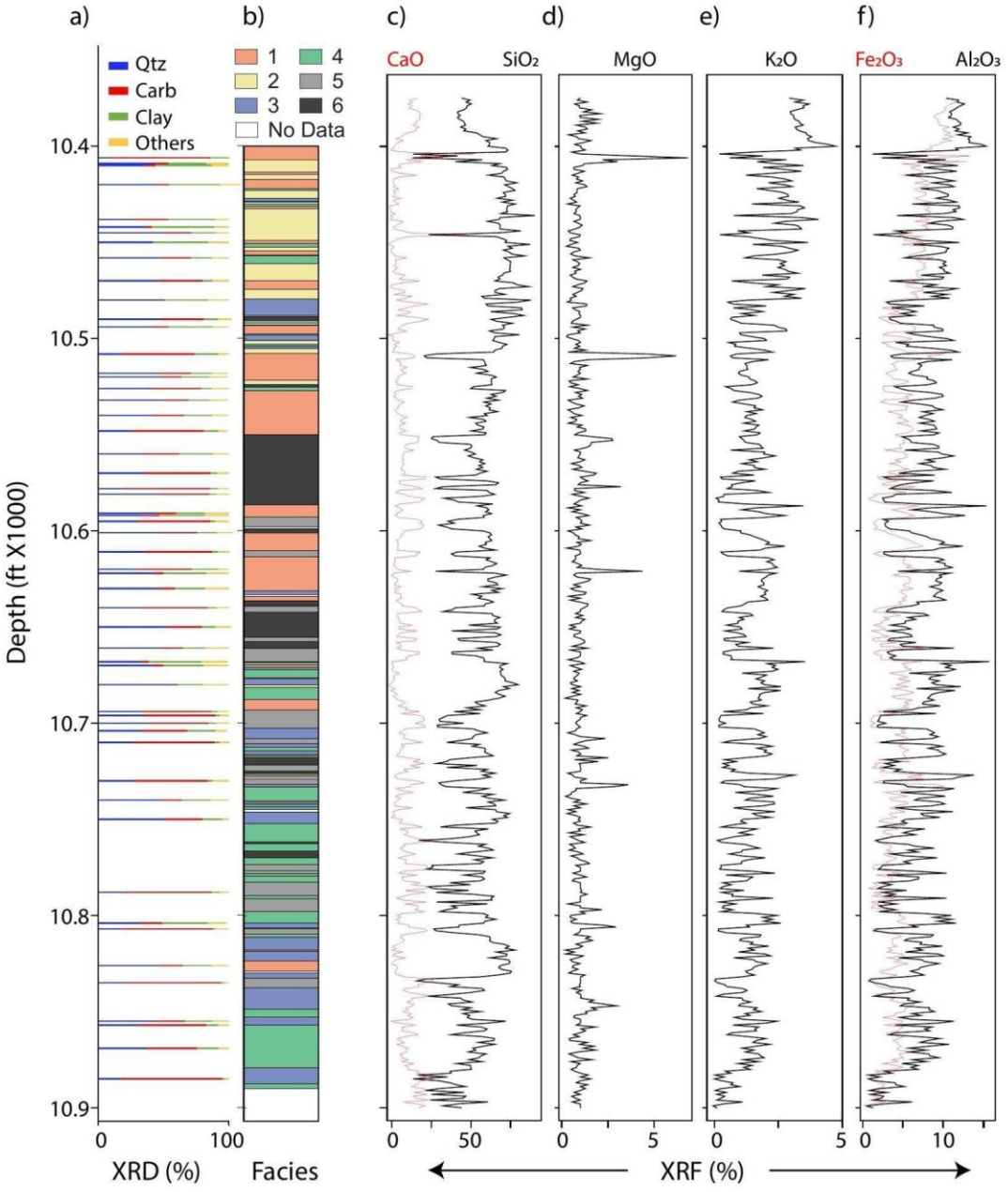

Figure2: Data. (a) Targets: XRD generated mineralogy grouped into quartz, carbonates, and clay. (b) Facies: 1. Massive-bedded Mudstone-Siltstone, 2. Laminated Siltstone, 3. Burrowed Siltstone, 4. Bioturbated Siltstone, 5. Massive-bedded Packstone-Siltstone, and 7. Hummocky Cross-Stratified Planar-laminated Packstone-Grainstone. Features: XRF Elemental Oxides of

(c) Calcium (red) and Silicon (black), (d) Magnesium, (e) Potassium and (f) Iron((red) and Aluminium (black). Depth sampling for XRF is regular (1 ft) and XRD is irregular but covers all the facies. Note the high negative correlation between SiO_2 and CaO and high negative correlation between K_2O, Fe_2O_3 and Al_2O_3. Correlation is further explored in section 3.1.

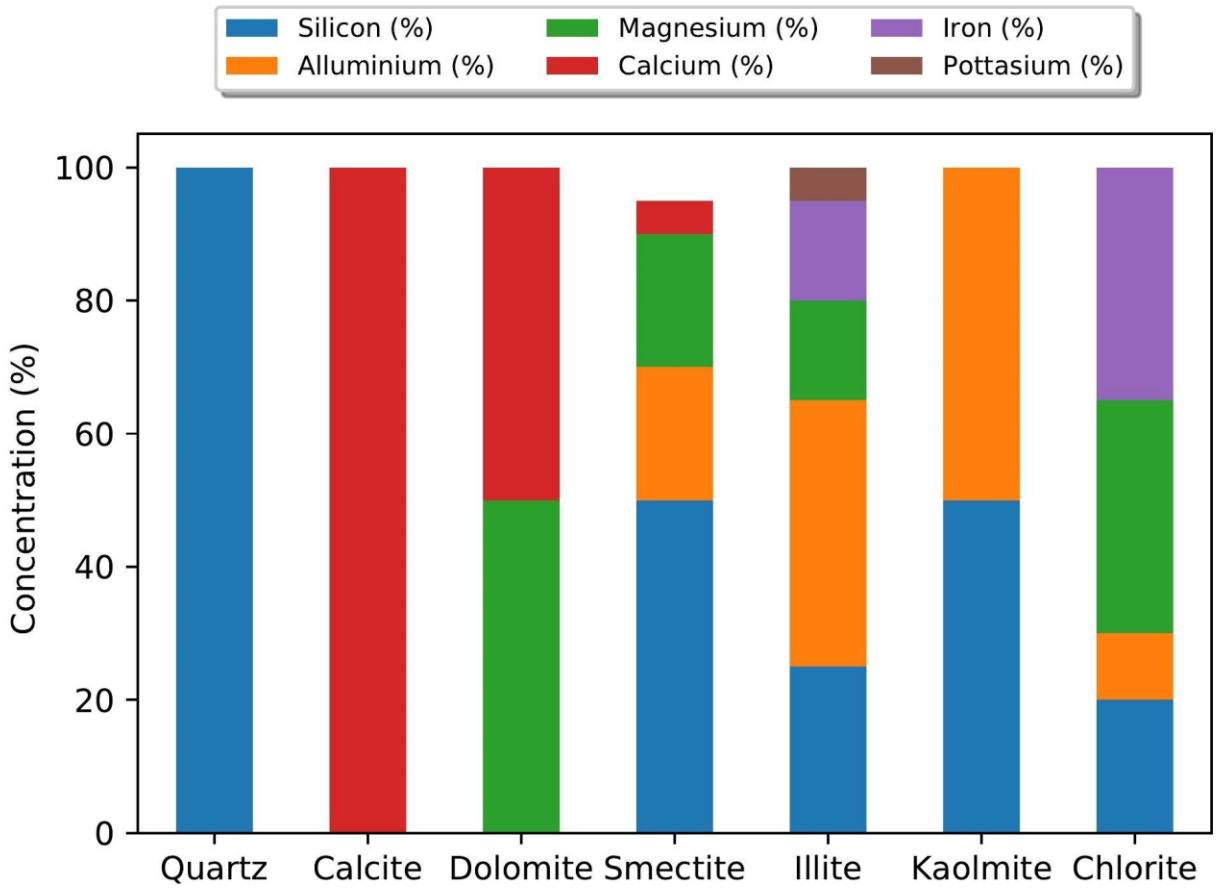

Figure3: *Distribution of features within the targets.*

CaO and MgO, and the clay target essentially had all features, with SiO2 and Al2O3 being the most dominant. It should be noted that XRF and XRD measurements were not concurrently performed. We considered XRD measurements that were within 2 inches of an XRF reading as being collocated. Neither XRD nor XRF data were interpolated in any manner. We have provided the facies information in Figure 2 to show that mineralogical

data were captured from all the main geological features of the reservoir, thus creating an unbiased framework for prediction.

3. Machine learning pipeline

The underlying philosophy of the pipeline developed in this paper is to allow flexibility and efficiency in testing various models and scenarios (training dataset percentages, etc.). It has three salient aspects: (i) feature engineering and selection are performed with a combination of domain knowledge and statistics, (ii) the best fitting model in every trial is automatically chosen with a competitive ensemble strategy, and (iii) the hyperparameters for all the networks chosen from a grid with k-fold cross-validation, thereby avoiding manual tuning.

3.1. Feature Engineering and Selection

The general practice in ML is to integrate all the targets and features through a deep and dense network that automatically selects target-specific features and engineers them for the best predictive ability. Such a network cannot be set up in a small data domain such as in this paper. Hence, we follow the alternative approach and identify and remove the features that do not contribute to a particular target (Heaton, 2016; Hua *et al.*, 2004) . As illustrated in Figure 3, the features are distributed within the targets. Although elemental oxides have a defined proportion within each mineral, accounting for them individually through a process of elimination is not possible. We have tried to set up a general workflow so that the idea can be applied to any combination of features (elements) and targets (minerals) by incorporating domain knowledge and statistical relationships.

Figure 2 suggests basic correlations between the features. For example, in this dataset, CaO and SiO_2, have a high negative correlation while Fe_2O_3, K_2O, and Al_2O_3 have a positive correlation. To further understand these relationships, we generated a Pearson correlation coefficient (PCC) for all the feature–target pairs (Figure 4). PCC is calculated as the ratio of covariance of the two variables with the product of their individual standard deviation (Rodgers and Nicewander, 1988):

$$\rho_{xy} = \frac{Cov(x, y)}{\sigma_x \, \sigma_y}, \qquad (1)$$

where ρ_{xy} is the PCC, $Cov(x, y)$ is the covariance between variables x and y, and σ_x and σ_y are the standard deviations of x and y, respectively. Pearson correlation coefficient measures the degree of linear association between the variables and takes values in the range $(-1,1)$, with the extreme values representing high linear correlation and 0 representing no correlation (Rodgers and Nicewander, 1988) . In that context, a correlation SiO_2 and carbonate (-0.91) may not seem logical because SiO_2 is absent in carbonate. However, we saw the need of exploring such relationships. For example, the absence of SiO_2-dominated targets, namely quartz, clay and “others", automatically implies presence of carbonates. Hence, relying only on constituent features for predicting the target would have limited the scope of our application. Thus, for this dataset, SiO_2 was the feature of choice for predicting carbonate. Likewise, solely based on the constituent features, discerning clay from "others'' might seem difficult as both the targets contained all the six features. However, a new line of reasoning emerged when we considered the dominant minerals in both targets. For example, illite and mixed-layer illite-smectite, which are both dominated by K_2O are dominant in clay. In “others”, K_2O is present in muscovite and

microcline. In terms of mass fraction, the K_2O-dominant minerals in the clay are present in 2 – 4 times higher concentration as compared to the K_2O-dominant minerals in "others". Thus, for this dataset, K_2O was the feature of choice for predicting clay.

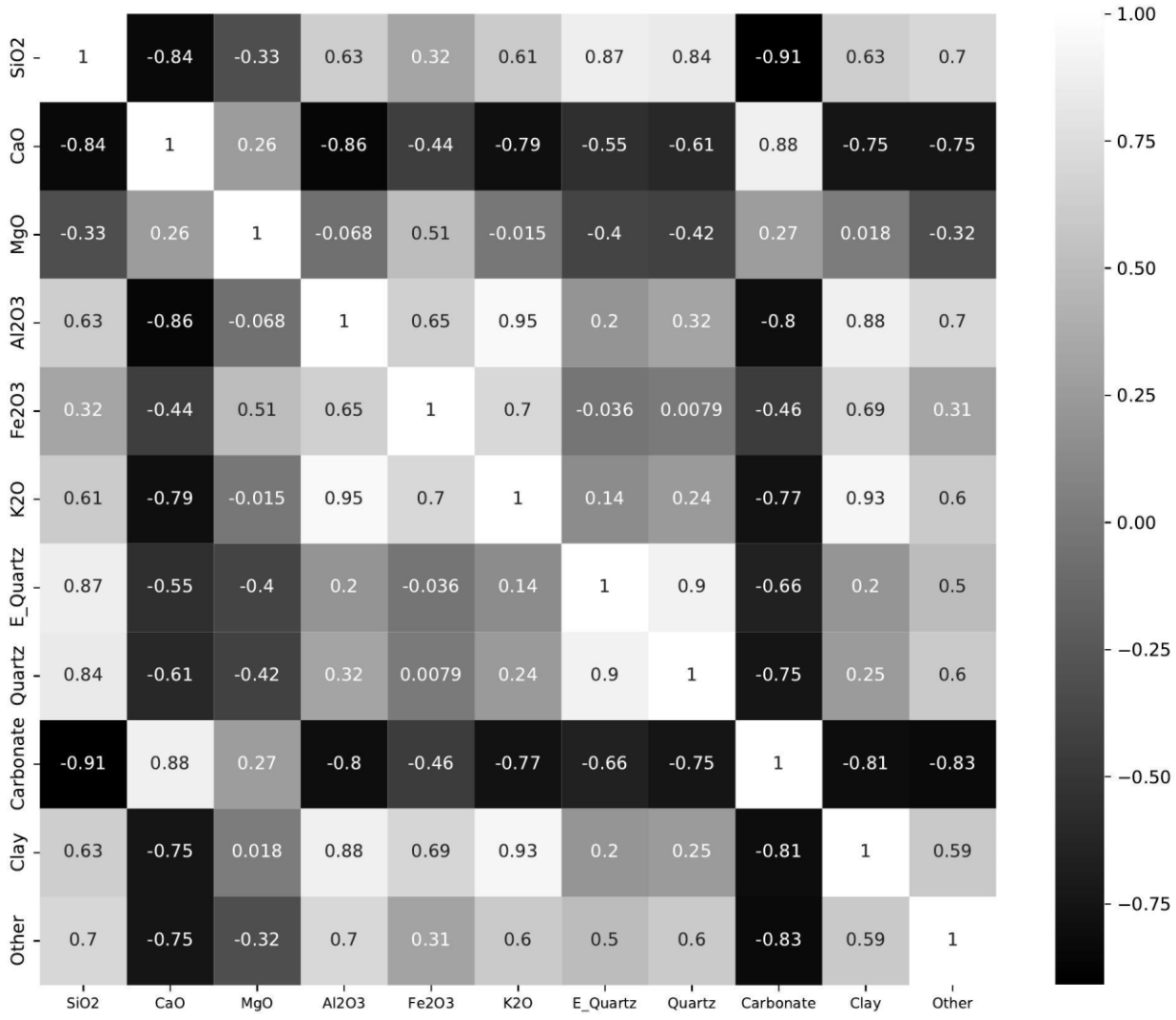

Figure 4: Pearson correlation coefficient computed for all possible pairs of features and targets using 40 % of data constituting an instance of the training set.

Predicting quartz was more challenging because its sole constituent (SiO_2) was also present in two other targets – clay and "others". For quartz, therefore, we attempted feature engineering and the goal was to come up with a generic expression of the form:

$$E\_quartz = SiO_2 - a \times K_2O, \qquad (2)$$

where the chemical symbols on the R.H.S. denote the percentages of oxides and the value of the parameter a is computed from a grid search such that the correlation between "E_quartz" and quartz is maximized. A new value of the parameter a is calculated for every trial. This of course, is finding the best feature in a linear sense.

3.2. Machine learning models and tuning

Equal importance should be given to both the ML algorithm and workflow in small data settings (Xin *et al.*, 2021) . Although the Pearson correlation coefficients (Figure 4) suggest linear relationships between a few targets and features, we explore three kinds of predictive models: (i) regularized linear regression as implemented by elastic-net (Hastie *et al.*, 2009) (ii) the tree-based XGBoost (Chen and Guestrin, 2016) based on regression trees and (iii) non-linear regression as implemented by feed-forward neural network (Goodfellow *et al.*, 2016). The key to the success of ML models lies in tuning their hyperparameters, which control the accuracy and robustness of algorithms. In general, hyperparameter optimization strategies are model-specific and decided by the statistics of the training data (Probst *et al.*, 2019). Here, we choose the hyperparameters with a grid search, with the optimum hyperparameters selected on the basis of k-fold cross validation. The cross-validations score should measure the similarity between the predicted and true

targets. There exist a variety of similarity measures including mean average error, mean squared error, normalized correlation, r^2 and correlation score, to name a few (Dhanani *et al.*, 2014) . We use the r^2 correlation score as it is robust in the presence of the outliers and provides an interpretable metric on the predictive ability of models (Chicco *et al.*, 2021). Table 1 summarizes the key parameters for the individual models, their significance and their range in the grid search. Next, we briefly describe the models while guiding the reader to the original references for details.

Elastic-net is a regularized linear regression method (Hastie *et al.*, 2009). It assumes a linear relationship between the input and output variables, and can be understood as a mapping between the input and output variables via a line or a multidimensional hyperplane. The loss function for the elastic net model is (Hastie *et al.*, 2009) :

$$L_{ENet} = \frac{1}{2 * m}(\sum_{i=0}^{m}(y_i - \hat{y}_i)^2 + \alpha(L1_{ratio}|w|_1 + 0.5(1 - L1_{ratio})||w||_2^2) \quad (3)$$

where m is the number of samples in the test data denoted by y, $\hat{y}$ is the predicted output, w are the weights of the elastic-net regression model. Thus, the loss function comprises three main components, corresponding to the norm of the residuals (first term on the R.H.S. of equation 3) and the regularization terms that penalize model complexity as quantified by a weighted average of the $l1$- and $l2$-norm of the weights of the linear model. The regularization factor α acts as a balancing term between minimization of residuals and model complexity, while the parameter $L1_{ratio}$ controls the relative weights of the $l1$- and $l2$-norm in the regularization term. The regularization terms help avoid overfitting and lead to better model generalization, in sparse and small data settings. Further, regularization

aids the loss function in approaching convexity, thereby increasing the chance of having a unique minimum (Hastie *et al.*, 2009). The values for $L1_{ratio}$ are found from a grid search over a range of values logarithmically spaced between 10^{-5} and 0.999 (Table 1). Lower values of $L1_{ratio}$ favor smaller models (with a smaller $l2$-norm), while higher values of $L1_{ratio}$ favor sparser models (with a smaller $l1$-norm).

XGBoost (XGB) is a scalable decision tree-based learning method (Chen and Guestrin, 2016). It is an efficient implementation of the gradient boosting method (Friedman, 2001). Owing to its decision tree roots, it is a widely accepted practice to use XGBoost along with the k-fold cross-validation method for robust results. XGBoost is implemented in a stacked manner with the results from one tree used in the prediction of the next tree. Further, in every tree, based on the values of maximum permissible depth (*max-depth*) and minimum remaining samples for splitting into further nodes (*min-sample-split*), the algorithm balances overfitting and underfitting. In this application, the values for *max-depth* were chosen between 2 to 12, while *min-sample-split* was chosen between 2 and 16 (Table 1) with a grid search. The maximum number of iterations (max-iter; Table. 1) is chosen by the grid search from a range of logarithmically spaced values between 100 to 20000 (Table 1).

Feedforward neural networks (FNN) are a class of non-linear regressive models with interconnected neurons that mimic connections in the human brain (Goodfellow *et al.*, 2016). Every neuron has an activation function that applies a transformation to the input obtained from the outputs of all the neurons from the previous layer with the last layer having an identity mapping. To avoid potential overfitting, we adopted a minimum-structure minimum-parameter approach. Table 1 details the grid of values supplied to

choose the number of hidden layers and neurons. We designed both multi- and single-output networks to contain less than 4 hidden layers, with the number of neurons descending by a factor of 2 with each layer (Sheela and Deepa, 2013). We use the rectified linear unit (ReLU) as the activation function in the hidden layers as it is well suited for regression problems (Goodfellow *et al.*, 2016) . The Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.01 is used to update the weights of the neural networks. The regularized loss function for the neural networks is given by:

$$L_{NN} = \frac{1}{m} \sum_{i=0}^{m} (y_i - \hat{y}_i)^2 + \alpha \parallel w \parallel_2^2, \qquad (4)$$

where m is the number of samples in the test data denoted by y, $\hat{y}$ is the predicted output, w are the weights of the neural network and α is the regularization parameter that penalizes the $l2$-norm of the weights. The range of values for the grid search for α are logarithmically spaced and range between 10^{-5} and 0.999.

The machine learning models described above and the grid-search-based tuning of hyper parameters was implemented using the Scikit-learn library (Pedregosa *et al.*, 2011).

Model	Parameter	Description	Range
	α	Weight of the regularization term (equation 3). Helps balance the bias-variance trade-off.	8 log-spaced values between 10^{-3} and 0.999.

Elastic Net	$L1_{ratio}$	Relative weighting of $l1$- and $l2$-norm in the regularization term (equation 3).	7 log-spaced values between 10^{-5} and 0.999
XGBoost	*max-depth*	Integral number of nodes in the tree.	2n, n ∈ [1, 2,...6]
	min-sample-split	Minimum integral number of samples required to split an internal node.	2n, n ∈ [1, 2,...8]
Feedforward Neural Networks	Hidden Layers and Neurons	Nonlinear transformation of the features	Number of layers k ranges between 1 and 4. Number of neurons in i^{th} layer is 2^{k-i}.
	α	Weight of the regularization term (equation 4). Helps balance the bias-variance trade-off.	8 log-spaced values between 10^{-6} and 300
Common parameter to all models	*max-iter*	Maximum number of iterations.	10 log-spaced values between 100 to 20000

Table 1: Parameterization of machine learning models The range of hyperparameter for elastic net, XGBoost and feed forward neural network are inspired from the work by Liu et al., (2018), Ørebæk *and Geitle (2021) and Sheela and Deepa (2013), respectively.*

3.3. Model selection: Competitive Ensemble

We implemented a competitive strategy for selecting the best model for any feature-target subset. We used the r^2 score on the cross-validation set to rank the models and choose the model that yields the highest score for a particular training set:

$$m_i = max(\{r^2(m)\}) \quad (5)$$

where i represents an instance of the training dataset, m represents the possible models amongst elastic-net, XGBoost and feedforward neural networks. The simplest model was selected in the instances where the r^2 scores were the same. We considered elastic-net to be the simplest model followed by XGBoost, single-output and multi-output neural networks.

4. Application and Results

Our application had dual goals. The first was to construct a model with the best repeatability and prediction accuracy. The second was to ensure that within the limited amount of data in hand with only 54 discrete targets the models were trained and tested appropriately. We devised two model variations *within each* elastic-net, XGBoost and feedforward neural networks:

Variation I: a single model that takes all the features as input to predict all targets at once

Variation II: three different models to separately predict each target using a combination of input data and engineered features.

Table 2 lists the four machine learning models devised across the two variations along with a description of the input and outputs for each model.

Variation	Model name	Description
I (No Feature Engineering)	M1	Predict *all targets* with *all raw features* as input with a *single* model.
II (With Feature Engineering)	M2	Predict *individual targets* with *all features* as input with *separate* models for each target.
	M3	Predict *individual targets* with *top 3 features* as input with *separate* models for each target.
	M4	Predict *individual targets* with *the best feature* as input with *separate* models for each target.

Table 2: Machine learning models. The top three features and the best feature are chosen according to the highest absolute value of the Pearson correlation coefficient (equation 1) computed using the training set (for example Figure 4).

For choosing the optimum amount for training, we examined the average cross-validation score from 10 random subsets (see Figure A.1, Appendix A) and concluded that 40% (21 samples) was adequate. Thus, we used 40 % of the data for training the models and reserved the remaining 60 % for testing the performance of the models. We used k-fold cross-validation with k=4 to select the optimum hyper parameters for the models with a grid search on the range of parameters as listed in Table 1. For a given instance of training data and model, 4-fold cross-validation involves training the model 4 times, and the best performing model is chosen as the one that yields the highest r^2 score on the cross-validation set, following accepted practice in machine learning (Goodfellow *et al.*, 2016) . Thus, for a given instance of training data, we obtain four models each for elastic-net, XGBoost and feedforward neural network (see Table 2) that are optimal for the dataset.

The optimal machine learning implementation for each model was ranked according to the r^2 scores on the cross-validation sets and the best model chosen: this is the competitive ensemble strategy described in section 3.3. However, to further ensure robustness and repeatability, we performed 11 trials with different draws of training and test data sets. The average r^2 scores for the 11 trials with 4-fold cross-validation (amounting to a total of 44 trials) for the models M1–M4 and the competitive ensemble are listed in Table 3. For quartz, the competitive ensemble method yielded an average r^2 score of 0.65, while the best performing individual model (average r^2 score of 0.63) was the feed-forward neural

network implementation of the model M2, *i.e.*, the model with all input features to predict the proportion of quartz. It should be noted that the input features included the six compound concentrations from XRF (Figure 2) and the engineered feature"E_quartz" calculated as per equation 2. Leaving out the engineered feature led to a reduction of around 0.1 in the r^2 scores for quartz, highlighting the importance of feature engineering in the prediction of quartz.

In the prediction of the proportion of carbonates, the Elastic Net implementation of the M3 model, *i.e.*, the model with the top 3 features to predict carbonate, performed best with an average r^2 score of 0.81. The competitive ensemble chose the Elastic Net implementation of the M3 model for most trials, thereby resulting in a similar average r^2 score as that of the individual model. The top 3 features used in the prediction of carbonates were SiO_2, CaO and Al_2O_3. The competitive ensemble strategy predicted clay with an average r^2 score of 0.77. The Elastic Net implementation of M4 model, *i.e.*, the model with the best feature to predict clay, was the best performing individual model with an average r^2 score of 0.76 and close to that achieved by the competitive ensemble. The proportion of K_2O was the top feature for predicting the proportion of clay. It should be noted that Table 2 reveals that individual models with feature selection (M3 and M4) perform better than models that are fed *all* features for carbonate and clay. We next discuss the performance of the thus chosen models on the test set.

	Quartz				Carbonate				Clay			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
Elastic-net	0.10	0.53	0.51	0.51	0.52	0.79	**0.81**	0.71	0.37	0.67	0.72	**0.76**
XGBoost	0.20	0.44	0.46	0.38	0.30	0.74	0.72	0.69	0.30	0.56	0.52	0.65
Neural Network	0.20	**0.63**	0.58	0.60	0.40	0.64	0.64	0.59	0.30	0.71	0.66	0.73
Competitive Ensemble	0.30	**0.65**	0.62	0.62	0.55	0.8	**0.81**	0.71	0.40	0.76	0.74	**0.77**

Table 3: Average r^2 scores of models (Table 2) and the competitive ensemble computed for the cross-validation sets populated with 4-fold cross-validation for 11 independent draws of 40 % training data. The best performing score amongst the individual models and the competitive ensemble are reported in bold.

In general the models M2, M3 and M4 perform better than M1. This is most-likely a dataset dependent phenomenon. In this dataset, we found that K_2O has a strong relationship with clay and therefore M4 clearly stood out. The targets “carbonates” and “quartz” do not have such a clear relationship with any one target element and there the performance of M2 and M3 versus M1 does not stand out as clearly. As intuitively expected, this relationship will

change from one dataset to another. We further discuss these aspects of model performance vis-à-vis mineral chemistry in the "Discussion" section.

Figure 5 displays the *test* data predictions from the best performing individual models for 11 trials with different draws of training and test data sets. The average r^2 scores of the test results from the chosen models are 0.71, 0.86 and 0.82 for quartz, carbonate and clay, respectively. The median predictions match the test data measurements with an average error of 3.2 %, 4.3 % and 4.5 % for quartz, carbonate and clay, respectively. The r^2 values for the test data are higher than the training cross validation scores in Table 3 since the training cross validation is a two-tiered average score, (*i.e.,* the average of the 4-fold cross-validation and 11 independent draws of training data), whereas the test score is a single-tier average (*i.e.* the average of the r^2 scores from 11 independent trials). The predictions of mineral proportions from the competitive ensemble strategy are similar to those shown in Figure 5 and have not been plotted. The r^2 scores on the test dataset for the models chosen by the competitive ensemble for each of the 11 trials are listed in Table A1 in Appendix A.

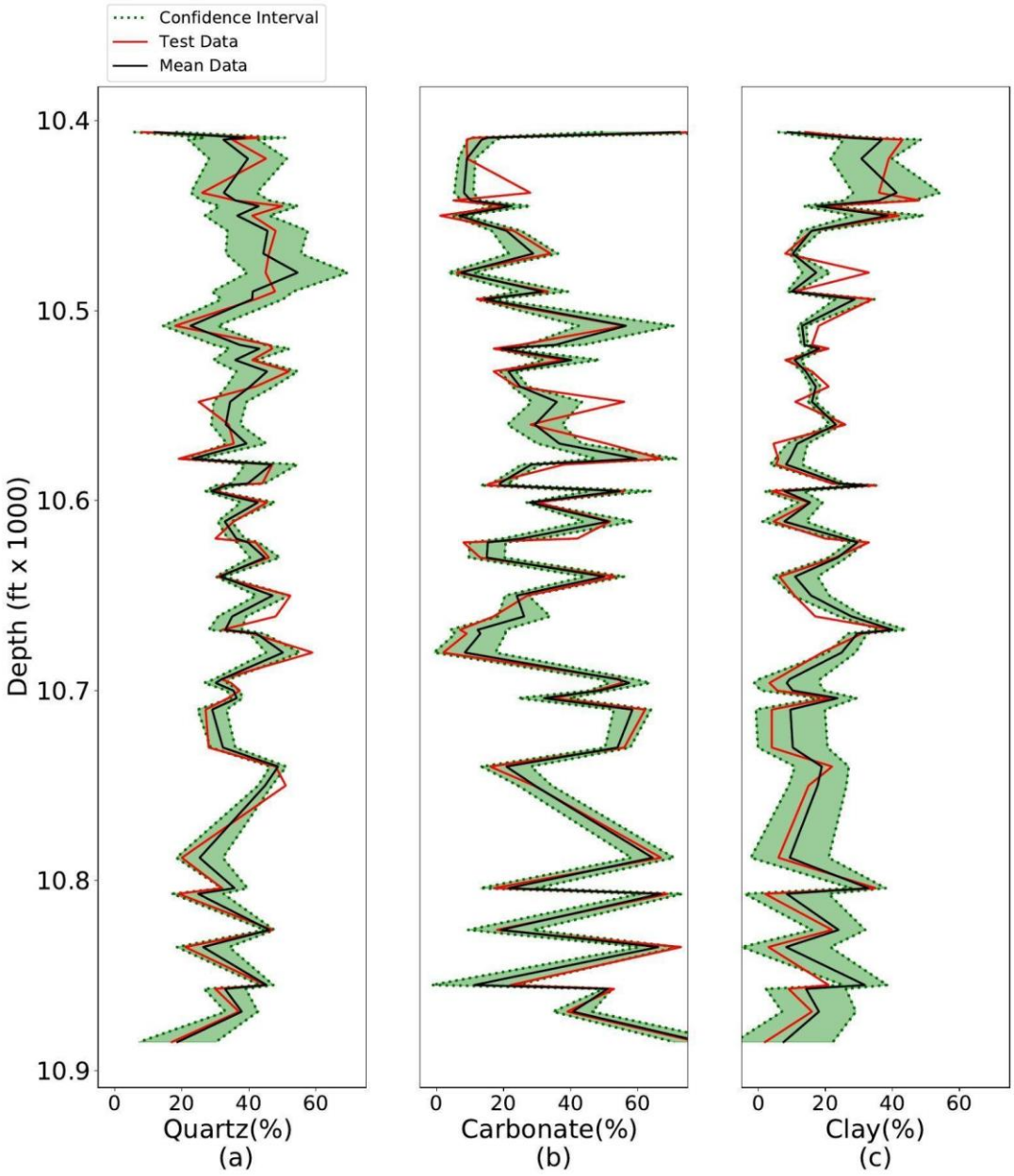

Figure5: Prediction of XRD test targets by the competitive ensemble for 11 trials. The red solid line represents the mineral proportion as measured by XRD, while the black line is the mean prediction from the competitive ensemble. The dashed green lines represent one

standard deviation and the shaded green area corresponds to the 1-sigma confidence interval. The average r^2 scores for quartz, carbonate, and clay are 0.71, 0.86, and 0.82, respectively.

5. Discussion

The mean r^2 scores of 11 trials obtained from the competitive ensemble for quartz, carbonate, and clay were 0.71, 0.86, and 0.82, respectively, within $\pm$ 5% variance (95% of the predicted targets were within 5% of the actual targets). The models were able to predict the phase (relative change from the previous sample) more accurately than the magnitudes. While the scatter in predictions is rooted in the statistical nature of the ML models, the input data also has a significant role to play. The XRD and XRF datasets used in this paper were not collocated. For example, the core plug which was used for generating the XRD sample was 3 inches in diameter, implying an averaging in mineralogy. The XRF data were acquired using a handheld device. The core section was first split into 1ft intervals and boxed. The XRF measurements were taken on the boxed sections at roughly the same location within the foot. While the approach may seem very unscientific, it is representative of the general practice and has been used to generate the enormous volume of XRF data available with the oil and gas operators, although not necessarily at 1ft interval. The reason behind such a sampling approach is that traditionally, XRD and XRF data are not collected with the intent of integrating them, although it is becoming a growing trend. Geologists use XRD-generated mineralogy primarily for interpreting the broader context of deposition. In addition, XRD also provides the total content of organic matter (TOC), which is essential for understanding the petroleum system. On the other hand, XRF mainly serves

as a tool for geochemical exploration, where elemental ratios rather than their absolute values provide more insight into the deposition environment. XRF is also a quick way to test for the presence of trace metals that have known associations with TOC. Merging XRD and XRF, as shown in this paper, serves a dual purpose. While the independent benefits of the two kinds of data are retained, relative changes in bulk mineralogical groups, e.g., quartz, carbonate and clays, provide insight into how depositional environment, e.g., distance from the shoreline, runoff, rainfall, provenance, etc., have changed (Yuretich *et al.,* 1999, Naidu *et al.,* 2014). Thus, although the r^2 scores reported in this paper are not very high, the results are geologically valuable and motivate practitioners to acquire XRD and XRF more comprehensively than the current practice.

The competitive ensemble strategy allows automated selection of the best fitting model for a given feature-target combination (Dietterich, 2000) . Hansen and Salamon (1990) suggest using diverse kinds of models within an ensemble for an exhaustive search of the solution space. The competitive ensemble strategy statistically outperforms individual models and can help avoid local minima. Our ensemble comprises three kinds of models: linear regression, tree-based and neural network. Within each predictive model family, we explored multiple algorithms. However, for the sake of brevity, we have presented results from representative algorithms only. For example, we have not presented results from tree-based algorithms like the random forest, ADABoost, and decision trees, as they produced results similar to those from XGBoost. The ensemble strategy is flexible and can be modified to include a diversity of algorithms, for instance, naive Bayes, support vector machines and kernel-based algorithms. The idea of grid-based tuning and cross-validation will remain unchanged. Within the ensemble, we have tested two variations of models: (I)

multi-input models to simultaneously predict all outputs and (II) models to predict each output separately using all features or selected features using domain knowledge and Pearson correlation coefficient computed from the training data. In our case, model variation II with separate models for each mineral performed better than the commonly used variation I with a single model that digests all input data to predict all the outputs simultaneously. Further, regularized linear regression as implemented by elastic-net was more successful in predicting the proportions of carbonate and clay, while feed forward neural networks best predicted the quartz target. Prior to addressing this problem, we had expected a feed forward neural network that predicts all targets with all compounds as input features to perform the best. However, this was not borne out by the results primarily due to the depth of the network. The selection of the neural network architectures is an open problem, although there are certain heuristics to fix the upper limit on both the total number of neurons and layers, primarily based on the amount of data (Bishop, 2006). The small size of our dataset did not allow the creation of a deep network, which in turn restricted its ability to simultaneously predict multiple targets within acceptable accuracy. We also explored sequential feed forward neural networks, wherein the output of one neural network was input to the next network. However, such an approach led to a compounding of errors and generally poor predictions, partly owing to the small size of the dataset. Furthermore, we also explored multiple predicting strategies, such as stacking (Misra and He, 2019) wherein we added predicted targets to the feature pool to predict subsequent targets. One of the strategies involved the prediction of quartz, followed by the utilization of the predicted quartz proportion along with other features to predict carbonate, and so on. However, the results from the stacking strategy were not

promising. Feature selection and engineering play a crucial role in the performance of the models in predicting the mineralogy. Feature selection and engineering are open-ended problems by themselves, and to the best of our knowledge, they are best addressed for a small data set up with a combination of statistical analysis (e.g. correlation coefficients, cross plots) and domain knowledge.

Although the total number of XRD data points, 54, may seem too small for ML, this number is at the higher end of measurements made for reservoir studies. XRD measurements are time intensive and destructive and therefore generated without redundancies. A small increase in XRD measurements, for example, an increase in 17 measurements in this study as compared to Alnahwi and Loucks (2019), may provide disproportionately better constraints on geology. Even though the dataset in this paper is at the limit of statistical significance, it is still reasonable for computation of the Pearson coefficient (David, 1938; Bonett and Wright, 2000), as it is used only to guide feature selection and engineering and *not* directly in prediction. Additionally, Figure A1 (Appendix A) shows that a relatively small proportion (> 30%) of data is optimal for training, possibly due to linear relationships between the features and targets. For small data ML applications, the importance of domain knowledge is paramount. In this paper, feature selection was adequate for two targets, carbonate and clay, and feature engineering was only necessary only for quartz. This might appear counterintuitive as the quartz target was simplest in terms of the feature content in that it only comprised one feature – SiO_2. The challenge here was that SiO_2 was dominant in clay (as well as “others"). On the other hand, carbonate and clay, which comprises more than one feature, have feature advantages. For example, CaO is present in both carbonate and clay but its concentration in carbonate far exceeded its

concentration in clay, *i.e.*, the XRF-measured CaO proportion was almost exclusively sensitive to carbonate. Likewise, K_2O is present in both clay and “others”, but its concentration in clay was higher than its concentration in “others”, implying that XRF K_2O was more sensitive to clay than “others". The relative sensitivity of features to target is also reflected in the r^2 scores. For example, the r^2 score for carbonate predicted using the CaO feature was higher than the r^2 score for clay predicted using K_2O. Unfortunately, since SiO_2 is not unique or dominant to quartz, an engineered feature had to be derived. Since SiO_2 is also dominant in clay, we subtracted a proportion of SiO_2 expected in clay by using K_2O, which is exclusive to clay. This decomposition was heuristic and it is easy to see how feature selection and engineering strategy will change for different mineralogy.

6. Conclusion

This paper integrates irregularly spaced and invasively sampled sparse XRD data with regularly spaced, continuous, and non-invasive XRF measurements using machine learning. The work involved predicting three targets and six features. The targets were broad categories of minerals - quartz, carbonate, and clay - inferred from XRD measurements. The features, direct measurements from XRF, were elemental oxides MgO, FeO, CaO, SiO, K_2O and Al_2O_3. The novelty of the paper is in designing a competitive ensemble strategy to determine which of the three popular machine learning model classes - linear regression (elastic-net), tree-based (XGBoost), and non-linear regression (feedforward neural networks) were most appropriate for predicting a particular target using the most relevant subset of features. This implementation led to a realization that while targets such as clay

and carbonate could be predicted using raw features, predicting quartz required feature engineering. Using a carefully implemented approach for hyperparameter tuning, we were able to train the models only with 40 % of the data. The robustness of the predictions was demonstrated by repeating the experiments on eleven varied training sets. Elastic-net was found to be most appropriate for predicting clay and carbonate, while the neural network was most appropriate for quartz. Our workflow is generic and applicable to any XRF-XRD companion datasets.

7. Acknowledgment

Devon Energy Corporation, USA, provided the core access, the XRF dataset and a part of the XRD dataset. The machine learning models were trained at the computational geophysics lab, Department of Earth Sciences, Indian Institute of Technology, Bombay. The authors would like to thank Heather Bedle and the Editors for constructive feedback.

8. Code and Data

The codes and data are available on https://github.com/mayurnawal123/XRF_XRD_Integration.

9. References

Alabbad, M., Prasad, M., and Ge Jin. (2021). Machine learning model evaluation: A case study for core guided petrophysical analysis. SEG Expanded Abstracts, 1656–60. https://doi.org/10.1190/segam2021-3582984.1.

Alfarraj, M., and AlRegib, G. (2019). Semi-supervised learning for acoustic impedance inversion. SEG Expanded Abstracts, 2398–2301. https://doi.org/10.1190/segam2019-3215902.1

Alnahwi, A. and Loucks, R. (2019). Mineralogical composition and total organic carbon quantification using x-ray fluorescence data from the Upper Cretaceous Eagle Ford Group in southern Texas. AAPG Bulletin, 103:2891–2907. https://doi.org/10.1306/04151918090

Alsaif, N. A., Hage, A. R. and Hamam, H. H. (2017). Mineralogy and Geomechanical Analysis for Hydraulic Fracturing: An Integrated Approach to Assess Rock Fracability in Sandstone Reservoirs: Abu Dhabi International Petroleum Exhibition & Conference, SPE-188606-MS.

Aoudia, K., Miskimins, J.L., Harris, N.B. and Mnich, C.A. (2010). Statistical analysis of the effects of mineralogy on rock mechanical properties of the Woodford shale and the associated impacts for hydraulic fracture treatment design. In 44th US Rock Mechanics Symposium and 5th US-Canada Rock Mechanics Symposium. OnePetro.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Bonett, D. G. and Wright, T. A. (2000). Sample size requirements for estimating pearson, kendall and spearman correlations. Psychometrika 65, 23–28. https://doi.org/10.1007/BF02294183

Bortolotti, M., Lutterotti, L., and Pepponi, G. (2017). Combining XRD and XRF analysis in one Rietveld-like fitting. Powder Diffraction, 32(S1), 1–6. https://doi.org/10.1017/S0885715617000276

Buhrke, V. E., Jenkins, R., and Smith, D. K., editors (1997). A Practical Guide for the Preparation of Specimens for X-Ray Fluorescence and X-Ray Diffraction Analysis. Wiley. https://doi.org/10.1021/ed076p762

Burnaev, E., Erofeev, P., and Papanov, A. (2015). Influence of resampling on accuracy of imbalanced classification. In Eighth international conference on machine vision (ICMV 2015). International Society for Optics and Photonics, 9875, 987521. https://doi.org/10.1117/12.2228523

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ. Computer science, 7, e623. https://doi.org/10.7717/peerj-cs.623

Das, V., Pollack, A., Wollner, U., and Mukerji, T., Convolutional neural network for seismic impedance inversion. Geophysics, 84 (6), R869-R880. https://doi.org/10.1190/geo2018-0838.1

David, F.N. (1938). Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples. Cambridge: Cambridge University Press.

Jonge-Anderson, I., Ma, J., Wu, X. and Stow, D. (2022). Determining reservoir intervals in the Bowland Shale using petrophysics and rock physics models: Geophysical Journal International, 228 (1), 39–65.

Dernoncourt, D., Hanczar, B., Zucker, J.D. (2014). Analysis of feature selection stability on high dimension and small sample data. Computational Statistics & Data Analysis, 71 681–93. https://doi.org/10.1016/j.csda.2013.07.012.

Dhanani, A., Lee, S. Y., Phothilimthana, P., and Pardos, Z. (2014). A Comparison of Error Metrics for Learning Model Parameters in Bayesian Knowledge Tracing. Technical Report UCB/EECS-2014-131, EECS Department, University of California, Berkeley.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple Classifier Systems, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Feng, R., Grana, D. and Balling, N. (2021a). Imputation of missing well log data by random forest and its uncertainty analysis. Computers & Geosciences 152 (2021): 104763.

Feng, R., Grana, D. and Balling, N. (2021b). Variational inference in Bayesian neural network for well-log prediction. Geophysics, 2021, 1–45.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5):1189–1232.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT Press.

Hady, M. F. A. and Schwenker, F. (2013). Semi-supervised Learning, pages 215–239. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hansen, L. and Salamon, P. (1990). Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12:993–1001.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Series in Statistics.

Heaton, J. (2016). An empirical analysis of feature engineering for predictive modeling. In SoutheastCon. IEEE.

Hilprecht, B. and Binnig, C. (2021). One Model to Rule them All: Towards Zero-Shot Learning for Databases. arXiv preprint arXiv:2105.00642.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. (2020). Meta-learning in neural networks: A survey. arXiv preprint arXiv:2004.05439.

Hua, J., Xiong, Z., Lowey, J., Suh, E., and Dougherty, E. R. (2004). Optimal number of features as a function of sample size for various classification rules. Bioinformatics, 21(8):1509–1515.

Hupp, B. N. and Donovan, J. J. (2018). Quantitative mineralogy for facies definition in the Marcellus Shale (Appalachian Basin, USA) using XRD-XRF integration. Sedimentary Geology, 371(C).

Kanfar, R., Shaikh, O., Yousefzadeh, M., and Mukerji, T. (2020). Real-time well log prediction from drilling data using deep learning. In International Petroleum Technology Conference.

Karpatne, A., Elbert-Uphoff, I., Ravela, S., Babaie, H., and Kumar, V. (2019). Machine Learning for the Geosciences: Challenges and Opportunities. IEEE Transactions on Knowledge and Data Engineering, 31(8):1544–1554.

Kim, D., Choi, J., Kim, D., and Byun, J., (2020). Predicting mineralogy by integrating core and well log data using a deep neural network. J. Pet. Sci. Eng. 195, 107838

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Kozlov, E., Fomina, E., and Khvorov, P. (2020). Factor Analysis of XRF and XRPD Data on the Example of the Rocks of the Kontozero Carbonatite Complex (NW Russia). Part II: Geological Interpretation. Crystals, 10(10).

Liu, J., Liang, G., Siegmund, K. D., and Lewinger, J. P. (2008) Data integration by multi-tuning parameter elastic net regression. BMC Bioinformatics 19: 369. https://doi.org/10.1186/s12859-018-2401-1.

Liu, X.-Y., Zhou, L., Chen, X.-H., and Li, J.-Y. (2020). Lithofacies identification using support vector machine based on local deep multi-kernel learning. Petroleum Science, 17:954–966.

Lousber, M. and Verryn, S. (2008). Combining XRF and XRD analyses and sample preparation to solve mineralogical problems. South African Journal of Geology, 111:229–238.

Misra, S. and He, J. (2019). Stacked neural network architecture to model the multifrequency conductivity/permittivity responses of subsurface shale formations. In Machine Learning for Subsurface Characterization, edited by Misra, S., Li H., and He, J. Gulf Professional Publishing.

Meller, C. and Ledésert, B. (2017). Is there a link between mineralogy, petrophysics, and the hydraulic and seismic behaviors of the Soultz-sous-Forêts granite during stimulation? A review and reinterpretation of petro-hydromechanical data toward a Better Understanding of Induced Seismicity and Fluid Flow. Journal of Geophysical Research: Solid Earth, 122(12), pp.9755-9774.

Naidu, P.D., Singh, A.D., Ganeshram, R. and Bharti, S.K. (2014). Abrupt climate-induced changes in carbonate burial in the Arabian Sea: Causes and consequences. Geochemistry, Geophysics, Geosystems, 15(4), 1398-1406.

Nawal, M., Kumar, S., and Shekar, B. (2022). LithoBot: An AutoML approach to identify lithofacies. Second International Meeting for Applied Geoscience & Energy Expanded Abstracts.

Nie, Feiping, Hu Zhanxuan, and Xuelong Li. “An Investigation for Loss Functions Widely Used in Machine Learning.” Communications in Information and Systems 18 (January 1, 2018): 37–52. https://doi.org/10.4310/CIS.2018.v18.n1.a2.

Northcutt, R. and Campbell, J. (1995). Geological Provinces of Oklahoma. Technical report, Oklahoma Geological Survey, Norman, Oklahoma.

Olson, M., Wyner, A. J., and Berk, R. (2018). Modern Neural Networks Generalize on Small Data Sets. In 32nd Conference on Neural Information Processing Systems

Pasini, A. (2015). Artificial neural networks for small dataset analysis. Journal of thoracic disease 7 (5): 953-60. doi:10.3978/j.issn.2072-1439.2015.04.61

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.

Pham, N., Wu, X., and Naeini, E. (2020). Missing well log prediction using convolutional long short-term memory network. Geophysics, 85(4):WA159–WA171.

Probst, P., Boulesteix, A.-L., and Bischl, B. (2019). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. Journal of Machine Learning Research, 20:1–32.

Qi, L. and Carr, T. R. (2006). Neural network prediction of carbonate lithofacies from well logs, Big Bow and Sand Arroyo Creek fields, Southwest Kansas. Computers & Geosciences, 32(7):947–964.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378:686–707.

Raj. R. (2021). Nanopore architecture and elastic velocities in the Mississippian-age carbonate-siliciclastic reservoirs of the US mid-continent (Doctoral dissertation, Oklahoma State University, Stillwater, U.S.A) Retrieved from https://shareok.org/handle/11244/335729Ramkumar, T., Selvakumar, M., Vasanthsankar, R., Sathishkumar, A. S., Narayanasamy, P., and Girija, G. (2018). Rietveld refinement of powder X-ray diffraction, microstructural and mechanical studies of magnesium matrix composites processed by high energy ball milling. Journal of Magnesium and Alloys, 6(4).

Rasolomanana, O. M. 'Ensemble Neural Network Using A Small Dataset For The Prediction Of Bankruptcy : Combining Numerical And Textual Data'. Graduate School of Economics and Business Administration, Hokkaido University. https://ideas.repec.org/p/hok/dpaper/361.html.

Rodgers, J. L. and Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. The American Statistician, 42(1): 59–66.

Selley, R. C. (2015). Elements of Petroleum Geology. Academic Press.

Sheela, K. G. and Deepa, S. N. Review on Methods to Fix Number of Hidden Neurons in Neural Networks: Mathematical Problems in Engineering 2013 (20 Ιούνιος 2013): 425740. https://doi.org/10.1155/2013/425740.

Spiess, A.-N., and Neumeyer, N. (2010) An Evaluation of R2 as an Inadequate Measure for Nonlinear Models in Pharmacological and Biochemical Research: A Monte Carlo Approach. BMC Pharmacology 10 (1): 6. https://doi.org/10.1186/1471-2210-10-6.

Tariq, Z., Mahmoud, M. & Abdulraheem, A. (2019). Core log integration: a hybrid intelligent data-driven solution to improve elastic parameter prediction. Neural Computing and Applications 31, 8561–8581. https://doi.org/10.1007/s00521-019-04101-3

Vabalas, A., Gowen, E., Poliakoff, E., Casson, A. J. Machine learning algorithm validation with a limited sample size. PLOS ONE 14 (11 2019): 1–20. https://doi.org/10.1371/journal.pone.0224365.

Vaniman, D., Bish, D., Blake, D., Elliott, S. T., Sarrazin, P., Collins, S. A., and Chipera, S. (1998). Landed XRD/XRF analysis of prime targets in the search for past or present Martian life. Planets, 103(E13):31477–31489.

Xin, D., Miao, H., Parameswaran, A., and Polyzotis, N. (2021). Production MachineLearning Pipelines: Empirical Analysis and Optimization Opportunities. In International Conference on Management of Data, 2639–2652.

Yellepeddi, R., Bapst, A., and Bonvin, D. (1996). Applications of an Integrated XRF-XRD Spectrometer. Journal de Physique IV Proceedings, 06(C4): C4–781–C4–788.

Yuretich, R., Melles, M., Sarata, B. and Grobe, H. (1999). Clay minerals in the sediments of Lake Baikal; a useful climate proxy. Journal of Sedimentary Research, 69(3), 588-596.

Zhang, Y., Ling, C. (2018) A strategy to apply machine learning to small datasets in materials science. npj Comput Mater 4, 25 . https://doi.org/10.1038/s41524-018-0081-z

Zhou, Z.-H. (2009). Ensemble Learning, pages 270–273. Springer US, Boston, MA.

Ørebæk, O-E. and Geitle, M. Exploring the Hyperparameters of XGBoost Through 3D Visualizations. In Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.

Appendix A: Detailed Results

The models chosen by the competitive ensemble strategy and the corresponding test r^2 scores are listed in Table A1. The input and output features for the models M1-M4 for each of the minerals are listed in Table A2. The M2 model, *i.e.*, the model that uses all input features along with the feature-engineered quartz to predict a single mineral, is most suited to predict quartz, with a neural network implementation. The M3 model, *i.e.*, the model that uses the top three features (as measured by the Pearson correlation coefficient) to predict a single mineral, is most suited to predict carbonate. Clay was best predicted using the M4 model that took K_2O as the only input to predict the proportion of the mineral. Elastic-net implementation was most suitable for both carbonate and clay. It should be noted that the

best-suited models are chosen by the competitive ensemble strategy based on 4-fold cross-validation scores (Table 3), *prior* to their application on test data.

Iteration	Quartz Model	r^2	Carbonate Model	r^2	Clay Model	r^2
1	Neural network, M2	0.70	Elastic-net, M3	0.89	Elastic-net, M4	0.78
2	Elastic-net, M2	0.69	Elastic-net, M1	0.86	Elastic-net, M4	0.82
3	Elastic-net, M2	0.75	Elastic-net, M1	0.89	Elastic-net, M4	0.79
4	Neural network, M2	0.79	Elastic-net, M3	0.91	Elastic-net, M4	0.82
5	XGBoost, M2	0.66	Elastic-net, M3	0.85	XGBoost, M4	0.90
6	Neural network, M2	0.70	Elastic-net, M3	0.85	Elastic-net, M4	0.85
7	XGBoost, M2	0.69	Elastic-net, M3	0.91	Elastic-net, M4	0.87
8	Neural network, M2	0.72	Elastic-net, M3	0.84	Elastic-net, M4	0.74
9	Elastic-net, M2	0.71	XGBoost, M3	0.63	Elastic-net, M4	0.79
10	Neural network, M2	0.75	Elastic-net, M3	0.87	Elastic-net, M4	0.80
11	Neural network, M2	0.76	Elastic-net, M3	0.91	Elastic-net, M4	0.83

Table A1: *Models chosen by the competitive ensemble and the corresponding test* r^2 *scores for 11 independent draws of 40 % training data. The input and output features for models M1-M4 are listed in Table A2.*

Model	Quartz	Carbonate	Clay
M1	All raw features	All raw features	All raw features
M2	All input raw features except that SiO2 is replaced by E_quartz	All raw features	All raw features
M3	E_quartz, CaO, MgO	Al2O3, CaO, SiO2	K2O, Al2O3, CaO
M4	E_quartz	SiO2	K2O

Table A2: Input and output features for the models M1-M4 used to predict the mineralogy.

Figure A1 shows the variation of average cross-validation scores for the best performing models as a function of the percentage of training data. It can be observed that the cross-validation score rapidly increases up to the point where 30% of the data are used for training. The cross-validation score does not exhibit significant variation for training sets larger than 30%, indicating that a proportion greater than 30% may be optimal for training the machine learning models. This observation combined with the fact that elastic-net was the best-performing model for clay and carbonate is suggestive of inherent linearity between features and targets.

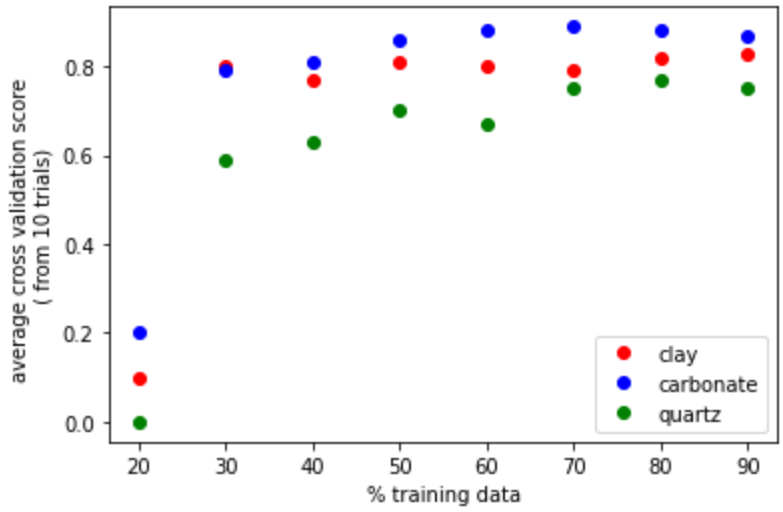

Figure A1: The average cross validation scores for the best performing models plotted as a function of the percentage of training data.