Journal of Petroleum Science and Engineering

XRD-XRF Integration: Using machine learning for predicting continuous chemistry

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full Length Article
Keywords:	XRD; XRF; small data; machine learning; feature engineering; feature
Corresponding Author:	Bharath Shekar, PhD Indian Institute of Technology Bombay Mumbai, Maharashtra INDIA
First Author:	Mayur Nawal, M.Tech
Order of Authors:	Mayur Nawal, M.Tech
	Bharath Shekar, PhD
	Priyank Jaiswal, PhD
Abstract:	In Earth Science, integrating remotely-measured continuous data streams with discrete in-situ measurements remains an open challenge. In this paper, we address such a problem using machine learning. 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 work flow 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 is limited due to analytical, logistic or economic issues.
Suggested Reviewers:	Ahmed Alnahwi, PhD Geologist, Aramco: Saudi Arabian Oil Co ahmad.nahwee@aramco.com Dr. Alnahwi has worked on machine learning analysis of XRF data from the Eagle Ford
	Heather Bedle, PhD Assistant Professor, OU: The University of Oklahoma hbedle@ou.edu Dr. Bedle has published several papers on the application of machine learning to problems in petro physics.
	Manika Prasad, PhD Professor, Colorado School of Mines mprasad@mines.edu Prof. Manika Prasad is a well recognized expert on rock, sediment and fluid properties, multiphysics, quantitative nano- and microscale characterizations.
	Peder Aursand, PhD Senior Engineer in Machine Learning, Aker BP ASA peder.aursand@gmail.com Dr. Aursand is a machine learning expert and has organized several ML competitions, with the most recent one being the FORCE 2020 Lithology Prediction ML competition.

To
The Editor
Journal of Petroleum Science & Engineering

February 8, 2022

Subject- "Submission of a new manuscript"

Dear Editor,

I am herewith submitting our manuscript entitled "XRD-XRF Integration: Using machine learning for predicting continuous chemistry" by Mayur Nawal, Bharath Shekar and Priyank Jaiswal, for its possible publication in Journal of Petroleum Science & Engineering. We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere. We know of no conflicts of interest associated with this publication. As corresponding author, I confirm that the manuscript has been read and approved for submission by all the named authors.

Thanking you

Yours Sincerely
Bharath Shekar
IIT Bombay

XRD-XRF Integration: Using machine learning for predicting continuous chemistry

Mayur Nawal, Bharath Shekar, Priyank Jaiswal

Abstract

In Earth Science, integrating remotely-measured continuous data streams with discrete in- situ measurements remains an open challenge. In this paper, we address such a problem using machine learning. 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 set- ting. Our work flow 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 is limited due to analytical, logistic or economic issues.

XRD-XRF Integration: Using machine learning for predicting continuous chemistry

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

[1] **Department of Earth Sciences, Indian Institute of Technology Bombay, Mumbai, India. E-mail: bshekar@iitb.ac.in.**

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

(February 9, 2022)

Running head:

ABSTRACT

In Earth Science, integrating remotely-measured continuous data streams with discrete in-situ measurements remains an open challenge. In this paper, we address such a problem using machine learning. 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 work flow 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 is limited due to analytical, logistic or economic issues.

Keywords: XRD, XRF, small data, machine learning, feature engineering, feature selection

1. INTRODUCTION

Knowledge of reservoir chemistry is critical in oil and gas applications. It helps in selecting appropriate production (Sun *et al.*, 2020) and sequestration zones (Kelemen *et al.*, 2019) as well as in environmental assessment (Dutta *et al.*, 2017). It is 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 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 powered 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 time required to prepare them, restricts their extensive usage on cores, which are difficult and expensive to obtain at the first place. Non-destructive ways of remotely sensing mineral chemistry also exists. 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 determination of elemental composition in a continual manner along the core, it does not provide direct measurements of bulk mineralogy required for reservoir applications. The benefits of integrating XRD and XRD are obvious. 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.

A case study of obtaining continuous reservoir chemistry though XRD-XRF integration was demonstrated by Alnahwi and Loucks (2019) using machine learning (ML). Although the results were shown with limited data, they 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. Recently, advanced methods like neural networks and deep learning have gained popularity and are rapidly reaching maturity in the hands of a prolific and rapidly increasing user base (Wang and Morra, 2020). Popular ML applications in Earth Sciences include sonic log prediction with stacked neural networks (Misra and Han, 2016), 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). 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) defines "small-data" problem as one where the measured data (e.g., features or target or both) lie between 10 and 10,000 in 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 help but uncertainties with small datasets are difficult to work around. 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.

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 generated uniformly at 5 cm (2-in) interval 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), Sulphur (S), Potassium (K), Calcium (Ca), and Iron (Fe) and the targets were Calcite ($CaCO_3$), Dolomite ($CaMg(CO_3)_2$), quartz (SiO_2), 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 practised 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 analysed using X-ray fluorescence (XRF) and X-ray diffraction (XRD) experiments to generate the

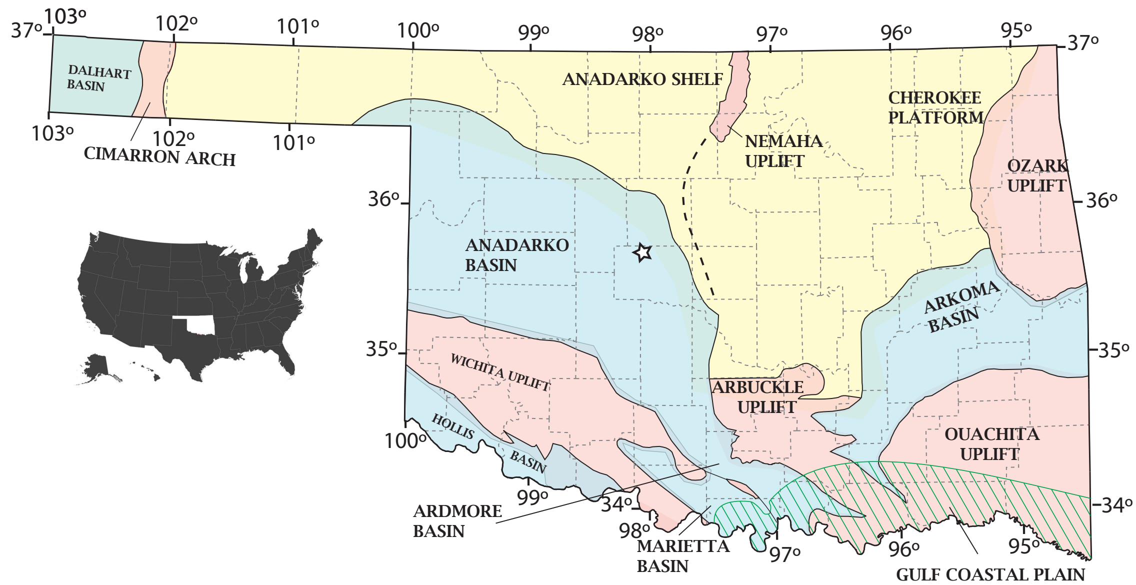

Figure 1: Base map. State map of Oklahoma with major geological features (after Northcutt and Campbell, 1995). The location of the core used to generate the X-Ray Diffraction (XRD) and X-Ray Fluorescence (XRF) data used in this paper is marked with a star. Inset shows location of Oklahoma within the United States of America.

dataset used in this study. The main reservoir is of mixed siliciclastic-carbonate Mississipian age formation, and spans $\sim$ 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 0.5 cm portion 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 Rigaku MiniFlex Diffraction machine. Calibration of the machine was done on quartz. Powder diffraction file was used to match and identify minerals and Rietveld refinement scheme was used to quantify the identified minerals using a commercially available software RIQAS (Ramkumar *et al.*, 2018).

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 comprised entirely of 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%; $[Na,\ Ca]_3{\cdot}[4\text{-}5]H_2O$ $[(Al_{1.5}Fe_2^{3+}Mg_3)Si_4O_10(OH)_2]^{\cdot 3}$) and mixed Illite-Smectite layers. Finally, Others consisted of Muscovite (0 – 3%, $KAl_2(AlSi_3O_{10})(F,OH)_2$), Miroclene (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 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, SiO_2, Al_2O_3, Fe_2O_3, MgO, CaO, and K_2O, (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, while the quartz target comprised entirely of one feature, SiO_2, the carbonate target is dominated by two features, CaO and MgO, and the clay target essentially had all features, with SiO_2 and Al_2O_3 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.

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 target and features though 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 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 work flow 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,

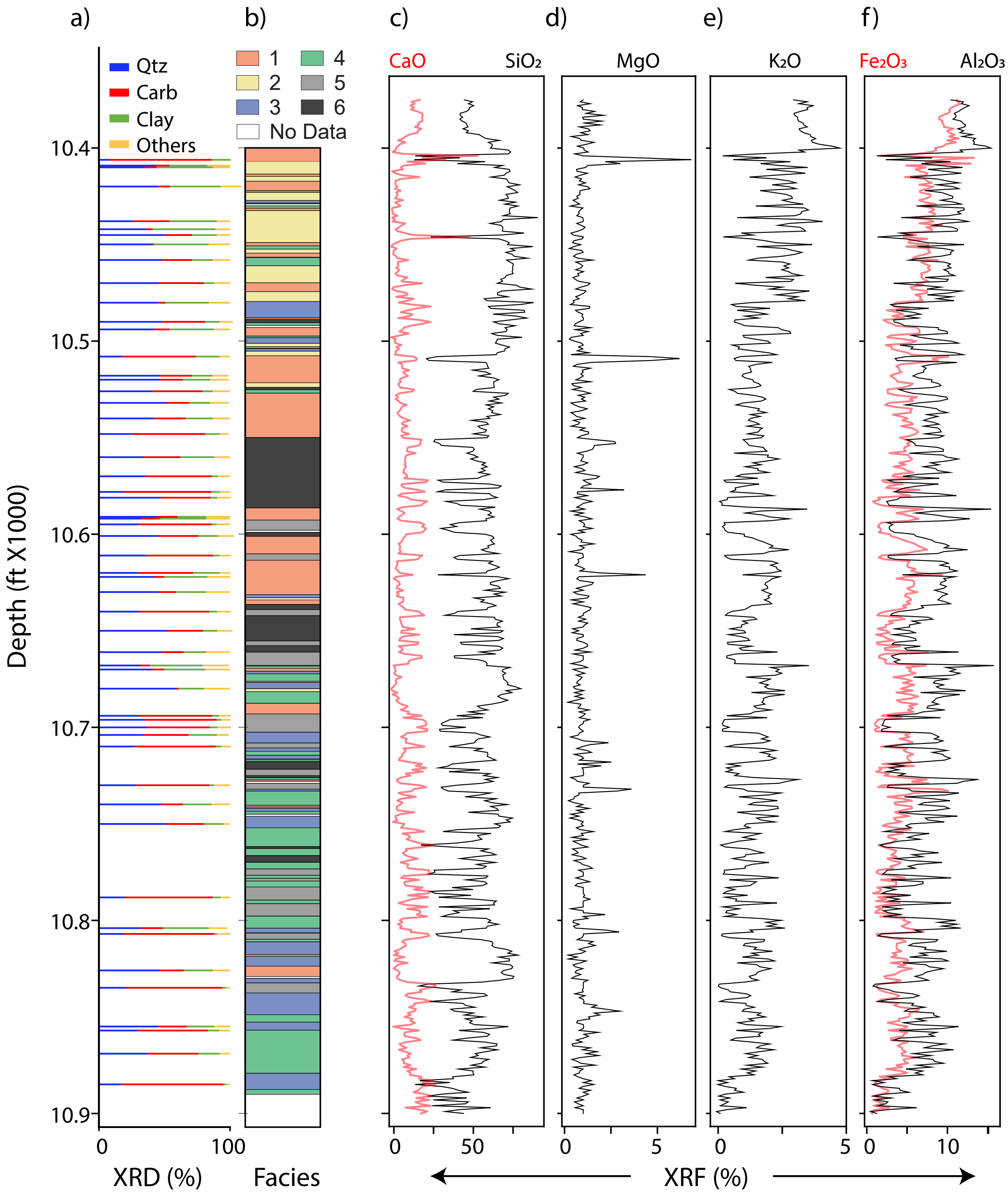

Figure 2: Data. (a) Targets: XRD generated mineralogy grouped into quartz, carbonates, and clay. (b) Facies: 1. Massive-bedded Mudstone-Siltstone, 2. Laminated Siltstone, 3. Burrowed Siltstone, 4. Bioturbated Siltstone, 5. Massive-bedded Packstone-Siltstone, and 7. Hummocky Cross-Stratied Planar-laminated Packstone-Grainstone. Features: XRF Elemental Oxides of (b) Calcium (red) and Silicon (black), (c) Magnesium, (d) Potassium and (e) 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.

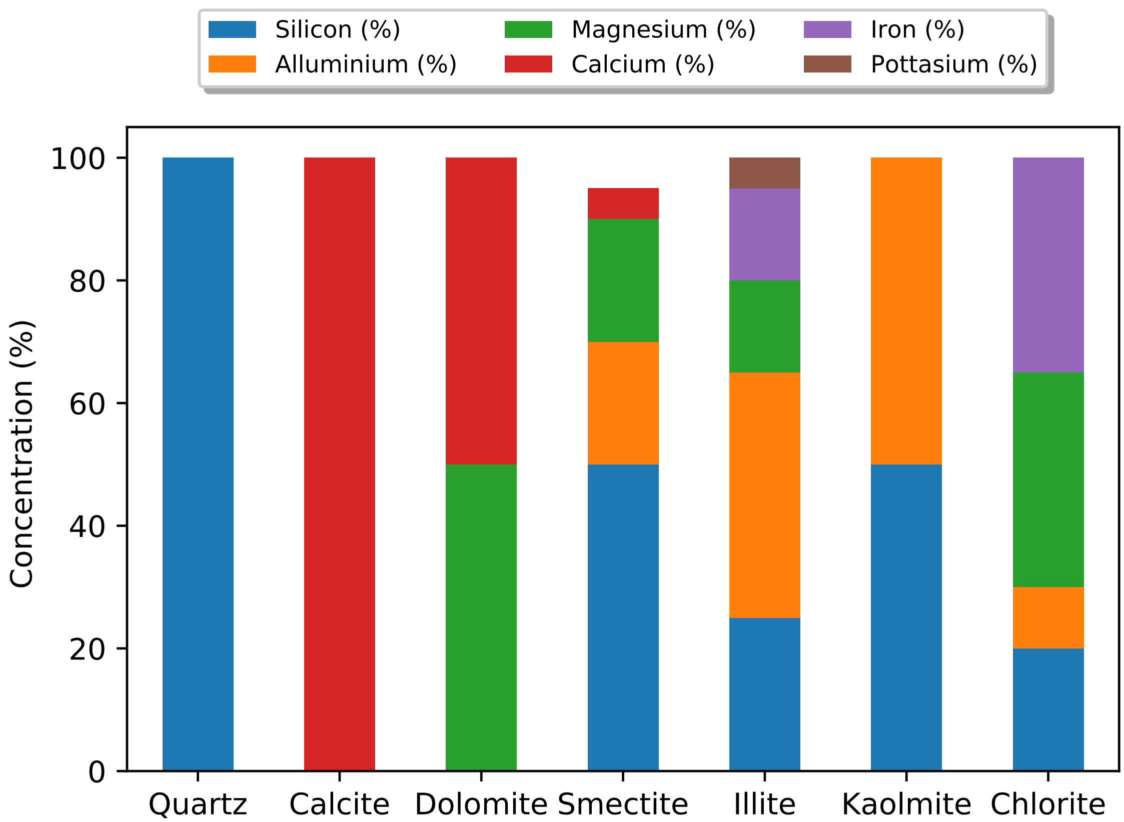

Figure 3: Distribution of features within the targets.

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 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):

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

where ρ_{xy} is the PCC, $\mathrm{Cov}(x, y)$ is 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 imply presence of carbonate. 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 Ilite-Smectite, that 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

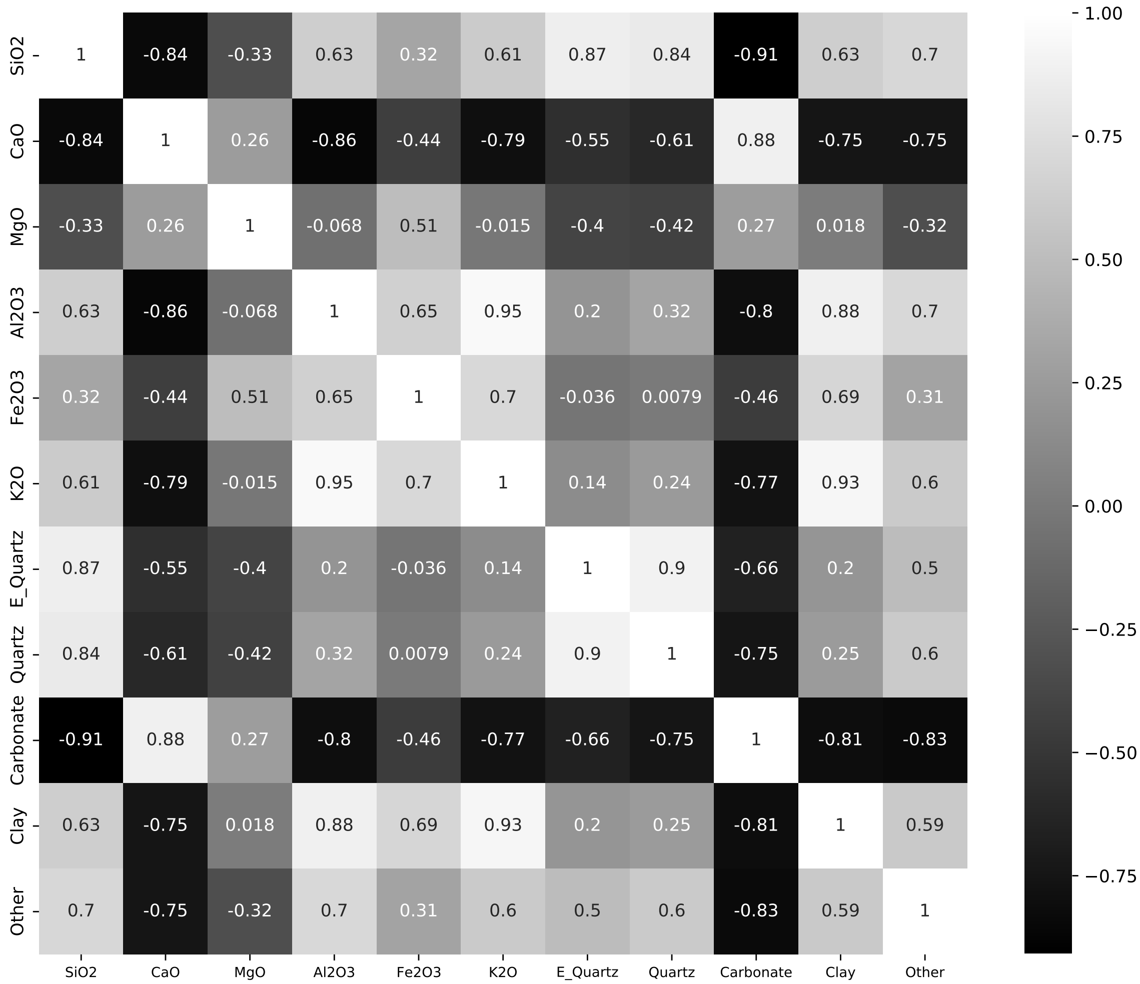

Figure 4: Pearson correlation coefficient computed for all possible pairs of features and targets using 40 % of data constituting an instance of the training set.

in 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. 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:

$$\text{E\_quartz} = \text{SiO}_2 - a \times \text{K}_2O, \tag{2}$$

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 maximised. 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, normalised correlation, r^2 correlation score, to name a few (Dhanani *et al.*, 2014). We use the $r2$ correlation score as it is robust in the presence of the outliers (Dhanani *et al.*, 2014). 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):

$$\mathcal{L}_{\text{ENet}} = \frac{1}{2*m}\sum_{i=0}^{m}(y_i - \hat{y}_i)^2 + \alpha\left(\text{L1}_{\text{ratio}}\|\mathbf{w}\|_1 + 0.5\,(1 - \text{L1}_{\text{ratio}})\|\mathbf{w}\|_2^2\right) \tag{3}$$

where m is the number of samples in the test data denoted by $\mathbf{y}$, $\hat{\mathbf{y}}$ is the predicted output, $\mathbf{w}$ 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 $\ell 1$- and $\ell 2$-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 $\ell 1$- and $\ell 2$-norm in the regularization term. The regularization terms help avoid overfitting and lead to better model generalization, in sparse and small data setting. 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 $L1_{\mathrm{ratio}}$ are found from a grid search over a range of values logarithmically spaced between 10^{-5} and 0.999 (Table 1). Lower values of $L1_{\mathrm{ratio}}$ favour smaller models (with a smaller $\ell 2$-norm), while higher values of $L1_{\mathrm{ratio}}$ favour sparser models (with a smaller $\ell 1$-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 k-fold cross validation method for robust results. XGBost 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 over fitting 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. 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:

$$\mathcal{L}_{\mathrm{NN}} = \frac{1}{m} \sum_{i=0}^{m} (y_i - \hat{y}_i)^2 + \alpha \left\| \mathbf{w} \right\|_2^2, \tag{4}$$

where m is the number of samples in the test data denoted by $\mathbf{y}$, $\hat{\mathbf{y}}$ is the predicted output, $\mathbf{w}$ are the weights of the neural network and α is the regularization parameter that penalizes the $\ell 2$-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 were implemented using the Scikit-learn library (Pedregosa *et al.*, 2011).

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:

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

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

Model	Parameter	Description	Range
Elastic Net	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
	$\mathrm{L1}_{\mathrm{ratio}}$	Relative weighting of $\ell 1$- and $\ell 2$-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.	$2\,n, n : [1\,..\,6]$
	min-sample-split	Minimum integral number of samples required to split an internal node.	$2\,n, n : [1\,..\,8]$
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: Parametrization of machine learning models.

Variation	Model name	Description
I	M1	Predict *all targets* with *all features* as input with a *single* model.
II	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).

be simplest model followed by XGBoost, single-output and multi-output neural networks.

4. APPLICATION AND RESULTS

Our application had dual goals. First was to construct a model with best repeatability and prediction accuracy. Second was ensure that within the limited amount of data in hand with only 54 discrete targets the models were trained and tested appropriately. The two goals, which is at the core of small data problems, had to be met simultaneously. We devised two model variations *within each of* 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 separate 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.

We used 40 % of the data for training the models and reserved the remaining 60 % of the data for testing the performance of the models. We use 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 yield 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 can in turn be 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 proportion of carbonates, the Elastic Net implementation of the M3 model, *i.e.*, the model with 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

	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.

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.

Figure 5 displays the *test* data predictions from the competitive ensemble strategy 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 predictions of mineral proportions from the best individual models for each particular mineral, specifically, feed forward neural network implementation of M1 for quartz, elastic-net implementation of M3 for carbonate and elastic-net implementation of M4 for clay, were 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

4.

DISCUSSION

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. However, others kinds of model such as naive Bayes, support vector machines and kernel-based solutions can also be included in the ensemble. 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 network 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 the network. The selection of the neural network architectures is an open problem, although there are certain heuristics are applied to fix the upper limit on both the total number of neurons and layers, primarily based on the

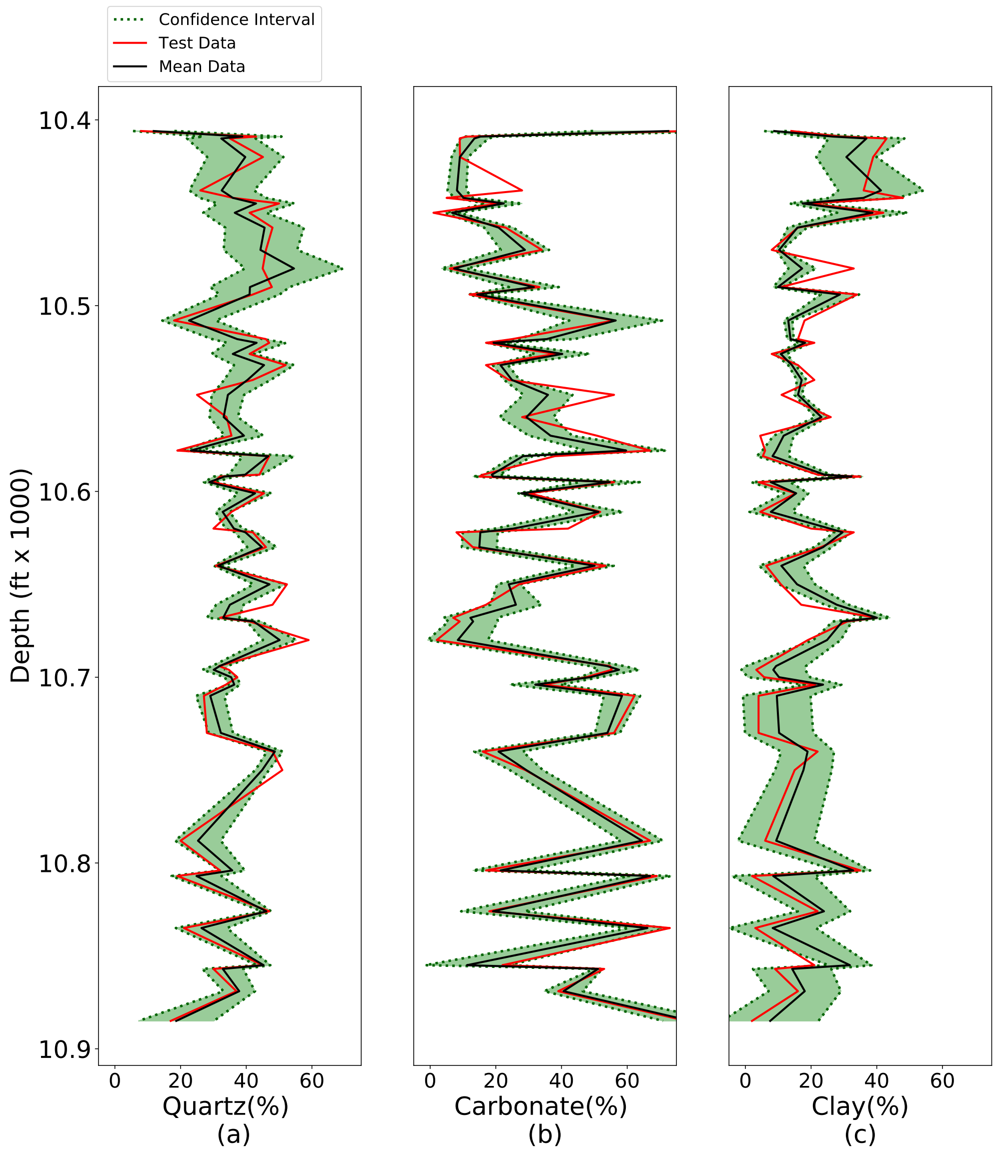

Figure 5: Prediction of XRD *test* targets by competitive ensemble for 11 trials. The red solid line represents the mineral proportion as measured by XRD, while the black line is the mean prediction from the competitive ensemble. The dashed green lines represent one standard deviation and the shaded green area corresponds to the 1-sigma confidence interval. (a) quartz, average r^2 score: 0.71, (b) carbonate, average r^2 score: 0.86 and (c) clay, average r^2 score: 0.82.

amount of data. The small size of our dataset did not allow creating 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. 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 (for e.g. correlation coefficients, cross plots) and domain knowledge.

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 counter intuitive 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 comprised more than one feature have feature advantage. 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.

The mean r2 scores of 11 trials obtained from 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 are 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. Further, the samples collection itself had associated uncertainties. 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 cut into 1 ft 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 1 ft 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 geology. 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 deposition environment, e.g., distance from the shoreline, runoff, rainfall, provenance, etc., have changed. Thus, although the r^2 scores reported in this paper are not very high, results are geologically valuable and motivate practitioners to acquire XRD and XRF more comprehensively than the current practice. For example, handheld XRF guns can be mounted for precise movements and XRD sampling could be based on mineral clustering (e.g., Vaidyanathan, 2001) or conducted iteratively (Burnaev *et al.*, 2015), or driven by XRF itself.

CONCLUSION

This paper integrates irregularly spaced and physically sampled sparse XRD data with regularly spaced and remotely sampled using continuous 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 an automatic 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 % data. Robustness of the predictions were demonstrated by repeating the experiments on eleven varied training sets. Elastic-net was found to be most appropriate for predicting clay and carbonate, while neural network was most appropriate for quartz. Our work flow is generic and applicable on any XRF-XRD companion datasets. The nature of the problem addressed in this paper is typical of small-data problems in geosciences.

ACKNOWLEDGMENT

Devon Energy Corporation, USA, provided the core access, the XRF dataset and a part of the XRD dataset. Winton Cornell, U. Tulsa, Oklahoma and Rohit Raj, formerly of Oklahoma State University helped in generating XRD data. The machine learning models were trained at the computational geophysics lab, Department of Earth Sciences, Indian Institute of Technology, Bombay.

REFERENCES

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.

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

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.

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., volume 9875, page 987521.

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.

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.

Dutta, M., Saikia, J., Taffarel, S. R., Waanders, F. B., de Medeiros, D., Cutruneo, C. M., Silva, L. F., and Saikia, B. K. (2017). Environmental assessment and nano-mineralogical characterization of coal, overburden and sediment from Indian coal mining acid drainage. Geoscience Frontiers, 8(6):185–1297.

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.

Kelemen, P., Benson, S. M., Pilorgé, H., Psarras, P., and Wilcox, J. (2019). Challenges of CO_2 Storage in Minerals and Geological Formations. Frontiers in Climate, 1(9).

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, 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 Han, Y. (2016). Petrophysical Interpretation of Multi-Frequency Electromagnetic Measurements in Clay and Conductive-Mineral-Rich Mudrocks. In Unconventional Resources Technology Conference.

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):953–960.

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.

Ramkumar, 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).

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

Sun, T., Li, H., Wu, K., Chen, F., Zhu, Z., and Hu, Z. (2020). Data-Driven Predictive Modelling of Mineral Prospectivity Using Machine Learning and Deep Learning Methods: A Case Study from Southern Jiangxi Province. Minerals, 10(2).

Vaidyanathan, P. (2001). Generalizations of the sampling theorem: Seven decades after nyquist. IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 48(9):1094–1109.

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.

Wang, H. and Morra, G. (2020). Artificial Intelligence in Geosciences. Artificial Intelligence in Geosciences, 1:52–53.

Xin, D., Miao, H., Parameswaran, A., and Polyzotis, N. (2021). Production Machine

Learning Pipelines: Empirical Analysis and Optimization Opportunities. In International Conference on Management of Data, pages 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.

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

APPENDIX A: RESULTS IN DETAIL

The models chosen by the competitive ensemble strategy and the corresponding test r^2 scores are listed in Table 4. 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.

	Quartz		Carbonate		Clay	
Iteration	Model	r^2	Model	r^2	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 4: Models chosen by the competitive ensemble and the corresponding test r^2 scores for 11 independent draws of 40 % training data.

Highlights

XRD-XRF Integration: Using machine learning for predicting continuous chemistry

Mayur Nawal, Bharath Shekar, Priyank Jaiswal

- Predict bulk mineralogy (XRD) from compound (XRF) data
- Machine learning with small dataset
- Feature engineering and selection helps improve predictability
- Multiple classes of machine learning models
- Competitive ensemble to automatically select the models

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: