

A Novel Application of Machine Learning to a New SEM Silicate Mineral Dataset

Benjamin Parfitt
Principal Research Scientist
Reiform
Washington, DC, USA
ben@reiform.com

Robert M. Welch
Department of Earth and Planetary Sciences
Harvard University
Cambridge, MA, USA
rwelch@g.harvard.edu

Abstract—Machine Learning (ML) continues to find applications in the geosciences, specifically in the classification of minerals from spectral or elemental data. We begin by exploring the use of four different methods for classification of elemental mineral samples from Scanning Electron Microscopy (SEM) and microprobe analysis in terms of structure, group, and subgroup. We create the most extensive silicate mineral group and subgroup classifiers available to the best of our knowledge, and achieve precision and recall values as high as the current state-of-the-art methods, which cover fewer groups and subgroups. Finally, we attempt to leverage the knowledge of structural families to improve classification performance on mineral groups, and reapply this process to improve performance on mineral subgroups. The train, test, validation split of data used in this paper will be posted online, along with the code and a webpage called *MINDicator* where anyone can use the new models easily.

Index Terms—machine learning, ensemble learning, mineralogy, silicates.

I. INTRODUCTION

Mineralogists group naturally occurring crystalline solids, called minerals, into seven different families. One of these is the silicate family, which makes up $\approx 90\%$ of all minerals in the earth's crust and is critical for rock formation [1]. Determining what silicate minerals are present in a sample is therefore crucial in determining rock forming processes, histories of metamorphism, and tectonic history, and has countless other applications [2], [3].

Geoscientists currently make mineral predictions by using petrographic microscopes and spectroscopy methods [4]. These methods have been invaluable for the field of mineralogy since its inception in the late 1800s. However, there are drawbacks. The process requires an in-depth knowledge of mineralogy to derive an accurate classification. Even with a high degree of mineralogical knowledge, the identification is not always reproducible. As the process stands today, the time to correctly identify a group of samples is directly related to how many samples one has since there is not a reliable, in-depth, automated method of classifying a wide range of silicate minerals through chemical analyses [5]. The identifications of mineral family, group, and subgroup are based on real elemental data, and are used to train our data-driven ML models. This type of data-driven model provides a high degree of automation and reproducibility, and a path to parallelize the process of silicate mineral identification.

Machine learning techniques offer an expressway between data collection and data analysis that is normally time consuming and requires a high degree of domain knowledge concerning mineral identification. It has been shown that machine learning can expedite the process of mineral identification from Raman Spectroscopy [6], X-ray fluorescence (XRF) [5], and Electron Microprobe Analysis (EMPA) analyses [7]. Additionally, some methods have used deep learning and Convolutional Neural Networks (CNN) to identify minerals using standard RGB images [8] and hardness measurements [9]. Machine learning provides rapid, reproducible methods for determining the mineral in question. Other methods also use random forest classifiers and ensemble learning to improve mineral identification within rock cross sections by using spectral signatures from SEM and hyperspectral analysis of rock cross-sections [10], [11].

Existing methods that operate on weight oxide data have yet to utilize a large dataset, which is necessary to capture the wide range of variability in the silicate group. We have developed several methods to determine structure, group, and subgroup classification that are state-of-the-art because we classify more classes with higher accuracies than previous methods, while utilizing a more robust and challenging dataset. To our knowledge, we are the first to classify family structure. Additionally, we classify 12 more groups and 7 more subgroups than any previous method using SEM data [5]–[7]. We achieve macro F1-scores (Section II-B) of 98.4%, 92.3%, and 90.7% on structural families, groups, and subgroups, respectively. Finally, we investigate the effectiveness of utilizing the explicit divisions of structures and groups in order to improve classification quality of groups and subgroups, respectively. These tests show promising results that warrant further investigation.

The remainder of the paper is structured as follows: Section II contains background information on mineralogy and machine learning, Section III contains an overview of related work, Section IV contains the details of our dataset, Section V describes our methods for classification, Section VI contains results, Section VII contains a discussion, Section VIII our methods for classification using the explicit divisions of structure and group, Section IX contains the results of that effort, Section X provides additional discussion, and Section XI provides concluding remarks.

II. BACKGROUND

A. Mineralogy

Minerals are naturally occurring crystalline solids with a repeatable pattern. Due to differences in chemistry and crystal structure, minerals are broken into seven families: silicates, oxides, sulfates, sulfides, carbonates, native elements, and halides [1]. The silicate family of minerals contains six structural groups which are also denoted as subfamilies. These are nesosilicates, cyclosilicates, sorosilicates, inosilicates, phyllosilicates, and tectosilicates/framework silicates, all of which have lattice differences [1]. Differences in crystal lattice configurations will alter the type of ions that can perform solid solution in a mineral, which alters the number of different elements present within a mineral group. For instance, clay minerals (a phyllosilicate) allow for a greater degree of solid solution than wollastonite (an inosilicate) due to lattice differences [1], [12]. Though silicate minerals cover a wide range of chemical variability, the structure of a mineral family, group, or subgroup creates a unique chemical "fingerprint" due to physical chemistry [13]. For example, this means that, in theory, the amphibole group is chemically unique from quartz or chlorite [4], [14], [15]. This creates a "fingerprint"; if an individual knows the chemistry, they can forecast the structure [13].

Geologists commonly determine mineral chemistry by a spectroscopy method, typically either Raman, SEM, Energy-Dispersive Spectrum X-Ray Fluorescence (EDS/XRF), or EMPA. These methods work by subjecting a mineralogical sample to a high intensity electromagnetic radiation source or electron beam. The released energies are then reported as weight percent oxide counts.

B. Machine Learning

Decision trees [16] are a simple tree structure in which training data are split at each tree node on some learned condition until only one class is remaining. The resulting tree is then used to classify unknown data samples. Extremely randomized trees (Extra Trees) [17] is an ensemble learning method that randomizes the data splitting at the nodes within the decision trees used. Both of these tree methods are computationally efficient. The K-Nearest-Neighbors (KNN) classifier finds the k closest known samples to some unknown sample (by euclidean distance), and uses the classes of the known samples to predict the unknown class. A drawback of this method is sensitivity to high dimensional spaces, both in accuracy and efficiency, making it more computationally expensive than the tree methods. The final method employed is the Support Vector Machine (SVM) [18], which finds a hyperplane that provides maximal separation for two sets of data. Unseen data points are then plotted and classified by their position relative to the hyperplane. The data can first be operated on by a kernel to map it to a different space, allowing non-separable data to be separated. In order to apply this method to our multiclass problem space, the problem is converted into several "one-vs-rest" binary classification problems.

One weakness of many publications in ML is the use of evaluation metrics that do not fully report the results [6]. Some background terminology: *True Positive* (TP) are all the correctly labeled samples from the class of interest; *False Positives* (FP) are samples labeled as the class of interest, but actually in another; *False Negatives* (FN) are samples labeled as another class, but actually in the class of interest. Often, the only metric used is *accuracy* over all points. Computed over all points, this is simply (all correct points)/(all points). However, if 90% of the points are from class A, and the classifier labels all data as A, then an accuracy of 90% will be reached, which is misleading. This is why using metrics such as recall and precision is important. *Recall* is defined as $(TP/(TP+FN))$. If we measure using per class metrics from our previous example, class A will have 100% recall, and all other classes will have 0% recall. At this point the average recall can be calculated, providing a realistic picture of the results (at best 50%). Another metric that is important is *precision*, defined $(TP/(TP+FP))$. However, it is often more cumbersome to report the pair of metrics for each class, so the *F1-score*, defined as $(2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$, is often used as a comprehensive metric. The *macro F1-score*, which is the average F1-score across all classes, is commonly used to evaluate models. Note that this is unaffected by the imbalance of data in the evaluated dataset. One final metric that is employed in the later sections of our evaluation is the *average top-3 recall*, (top-3 recall, for brevity). To calculate this metric, the probability vector V is taken from each classifier. Normally, the position of the largest value in V is used as the class identified, but instead the three largest values are observed and if any of those correspond to the correct class, the vector is considered "correct". Then, the average recall over all classes is computed as normal.

When large datasets are at hand in ML settings, the data are broken into three sets: train, validation, and test. The train set is used for training the model. The validation set is used to validate the model and tune the choices of various hyperparameters. The test set is used to test, or evaluate, the performance of the final trained model. By using these three sets it is ensured that the test data is not used to tune the hyperparameters, which would create a model that is tuned specifically for the test data and that creates misleading results.

III. RELATED WORK

Several ML methods, namely KNN, SVM, Extremely Random Trees, Weighted Neighbors [19], and CNNs, are reviewed as methods for mineral identification from Raman spectra [6]. The primary dataset used consisted of 3950 Raman spectra samples from 1214 mineral species, with some rather large class imbalances. A novel ML approach is introduced that achieves 89.31% accuracy, although precision and recall are not reported which leaves uncertainty when considered alongside the class imbalances. Further, the use of another novel method leveraging CNNs on the fusion of Raman, visible and near-infrared, and laser-induced breakdown spectroscopy data are explored as a method to improve accuracy of classification.

This is shown to far outperform the use of a single dataset, reaching 92.76% accuracy. Again, neither precision nor any other more comprehensive metric is reported, leaving uncertainty of the performance of the model.

Random forest, SVM, and neural networks are used to predict mineral composition of eight different mineral classes for rock cross-sections by using hyperspectral imagery and SEM data [11]. The random forest method is able to achieve root-mean-squared error 0.02 on an unseen region of the rock section used to train the classifier. However, when moved to new samples the errors ranged from 0.03 to 0.12 for the lowest errors for each model.

A Classifier Chain Random Forest (CCRF) performs multi-class tasks by using a chain of binary classifiers, each of which is a random forest model, and where each step builds on the results of the previous classification. Applying this method to a hyperspectral image of a rock cross section achieved accuracies between 66.96% and 94.65% for six classes [10].

A decision tree is used to identify twelve different mineral groups from 4601 SEM-EDS analyses with a relatively balanced dataset [5]. This novel approach to determine minerals in thin sections reports 100% accuracy for the twelve minerals the study set out to identify. It should be noted that the dataset has a limited sample size of minerals from only igneous rocks and multiple samples are from the minerals identified to create their dataset. In this study, there is no exploration of options for solid solutions in mineral groups.

A set of 5 minerals from river sediment samples are identified using EMPA and EDS analyses [7]. Three different novel ML algorithms are used, with the most successful achieving $\approx 92\%$ accuracy. The authors show that ML algorithms can be used to classify geologic samples.

Many of these existing methods rely on additional spectral data, which is more computationally expensive to process and more expensive to obtain than EMPA or SEM data. Previous methods that operate on weight oxide data (such as SEM or EMPA) have yet to utilize a large dataset or classify the structural groups or mineral groups of samples. Our focus is on creating methods to determine structure, group, and subgroup for a larger number of classes than previously achieved, using a much larger dataset.

IV. THE DATASET

The dataset used is $> 99\%$ composed of data available from Earthchem [20], [21]. Each mineralogical sample contains the source DOI, location, methodology, sample ID, and chemical data. These analyses are chosen because they are the most common and accurate analysis types available to geoscientists. As dense as this source is, it contains only 10 viable clay mineral samples. To supplement clay mineral data, the other $< 1\%$ of our data were taken from two publications, one with microprobe analyses [22], and one with wet chemistry analyses [23].

In order to get only exact or near-exact data samples for training and evaluating our models, any EarthChem data that used “ $<$ ” in measurement is discarded, as this is a clear

indicator of uncertainty. All samples whose sum total weight oxide was not within 10% of 100%, that is, $|weight_{oxide} - 100| < 10$, are also discarded. We allow the room for error because there is inherent error within SEM or EMPA analyses due to a wide variety of inconsistencies within the sampling and preparation steps [24]. The dataset is split into train, test, and validation sets. The size of each set for subgroups, groups, and structure samples is provided in Table I. This shows that our dataset is much larger than datasets used in the past. Additionally, this new dataset has 17 subgroups and 20 groups (Table I), which is twelve groups and seven subgroups more than two previous datasets [5] [7], and is the only dataset to make the distinction between family structure, group, and subgroup.

V. PREDICTING THE STRUCTURE, GROUP, AND SUBGROUP WITH MACHINE LEARNING

We evaluate the performance of four ML models as applied to the tasks of classifying structure, group, and subgroup label. For each task, we evaluated the following 4 methods:

- 1) Decision Tree,
- 2) K-Nearest Neighbors algorithm (KNN),
- 3) Support-Vector Machine (SVM),
- 4) Extremely Randomised Trees (Extra Trees).

The implementation from sklearn in Python [25] was used for all methods. These 4 methods are chosen because they all have been shown to be effective for solving mineral classification tasks in the past [5]–[7], and we hope to build on those successes with our much larger dataset and new set of tasks. The hyperparameters and metadata for most methods are constant throughout the experiments. All parameters for respective methods are listed below:

- 1) `random_state=1, criterion='gini'`
- 2) `Neighbors: (Subgroup, 20), (Group, 25), (Structure, 25)`
- 3) `decision_function_shape='ovo', kernel='linear', C=18`
- 4) `random_state=42, criterion='gini'`

The k chosen for the KNN algorithm is lower for subgroups because the lowest represented classes in those subsets have fewer points than the lowest represented classes in the group or structure subsets. The train set is used to train the four classifiers for each method, and validation set is used to choose the best of the four classifiers. This is done because the choice of the best model is considered a form of parameter tuning, which is the purpose of the validation data. The best classifier is then evaluated on the test data. The macro-F1 metric is

TABLE I
THE NUMBER OF SAMPLES IN EACH OF THE TRAIN, TEST, AND VALIDATION SETS, FOR THE SUBGROUP, GROUP, AND STRUCTURE CLASSIFICATION TASKS.

Set	Subgroup	Group	Structure
Train	145161	282213	282213
Test	29032	56442	56442
Validation	19354	37629	37629
Classes	17	20	6

used to evaluate all models across all classes to choose the best model.

VI. RESULTS OF PREDICTION ON VALIDATION DATA

As shown in Table II, the KNN classifier performs best on the validation data in every task. The performance of the best model from each task on the test data is then evaluated, with results shown in Table III. The difference between the average recall and average precision of the subgroup classifier, about 7%, is the largest such gap of any of the classifiers. The subgroup, group, and structure classifiers all break 90% for the macro F1-score, with the structure classifier reaching 98.4%.

The disparity between the smallest and largest classes when considering the number of training points per class is quite large, and is reported in detail for the group classification dataset in Table IV. The class with the most points is olivine, and the class with the least is wollastonite, which has 99.96% fewer points than olivine.

The effect of having very few training points on the F1-score of the model is reported in Figure 1. All classes that achieve an F1-score lower than 90% have a number of data points 99% lower than the number of data points contained in the class with the most points. The worst performing model has a number of data points more than 99.9% lower than the class with the most points. Interestingly, the class with the fewest points, wollastonite, achieves a perfect F1-score of 100%. This is discussed in detail below.

VII. A BRIEF DISCUSSION OF THE RESULTS

As can be seen in Figure 1, it is not true that having few samples will prevent a model from classifying the class correctly, but rather that having many samples will increase the likelihood of high performance on a class. Also, all of the classes that receive worse than 0.9 as the F1-score have very few training samples.

The varying F1-scores for the low-sample classes are due to two factors: the uniqueness of crystal structure and the number of data-points per class.

Amphiboles and pyroxenes are both inosilicates that differ in structure, but have similar chemistry (which is the data used to classify the samples) [14] [26]. If a decrease in accuracy was solely due to similar chemistry, it would be apparent in these two classes. This is not the case, as amphiboles and pyroxenes have approximately the same accuracy, demonstrating that the

TABLE II

THE MACRO F1-SCORE OBTAINED ON THE VALIDATION DATA AFTER TRAINING EACH ML ALGORITHM ON THE DATA FOR SUBGROUPS, GROUPS, AND STRUCTURE DATASETS.

ML Algorithm	Macro F1-score		
	Subgroup	Group	Structure
DecisionTree	88.428	90.905	97.672
KNN	91.084	92.278	98.279
ExtraTree	86.518	87.952	96.695
SVM	86.844	88.421	96.675

TABLE III

THE PRECISION, RECALL, AND F1-SCORE FOR THE BEST CLASSIFIER FOR EACH TASK FROM SUBGROUP, GROUP, AND STRUCTURE, AS INDICATED IN TABLE II ON THE TEST DATASET.

Metric	Subgroup (KNN)	Group (KNN)	Structure (KNN)
Precision	95.327	93.295	97.845
Recall	88.551	92.186	98.994
F1-score	90.764	92.332	98.404

high number of training points allows us to discern one from another (Figure 1).

Inversely, wollastonite has a far greater F1-score than zeolite. While they both have relatively few training points, they have drastically different F1-scores. This is most likely caused by the uniqueness of the wollastonite lattice structure and,

TABLE IV

THE F1-SCORE, TRAINING DATA POINTS, TRAINING DATA POINTS AS A FRACTION OF THE LARGEST CLASS, AND TRAINING DATA POINTS AS A FRACTION OF THE MEAN POINTS IN ALL CLASSES, FOR EACH GROUP.

Group	F1-Score(%)	[Train]	[Train]/Max(Train)%	[Train]/Mean(Train)%
Aenigmatite	92.68	188	00.21	01.33
Amphibole	96.82	12054	13.40	85.42
Analcime	80.00	95	00.11	00.67
Chlorite	92.59	140	00.16	00.99
Clay Mineral	81.08	99	00.11	00.70
Cordierite	98.18	142	00.16	01.01
Epidote	85.39	211	00.23	01.50
Feldspar	99.81	56443	62.74	400.00
Feldspathoid	98.07	1294	01.44	09.17
Garnet	99.62	27471	30.54	194.68
Kyanite Group	95.08	161	00.18	01.14
Melilite	97.08	412	00.46	02.92
Mica	97.53	10870	12.08	77.03
Olivine	99.88	89961	100.00	637.54
Prehnite	96.00	58	00.06	00.41
Pyroxine	99.34	82301	91.49	583.25
Quartz	100.00	113	00.13	00.80
Serpentine	82.93	91	00.10	00.64
Wollastonaite	100.00	36	00.04	00.26
Zeolite	54.55	73	00.08	00.52

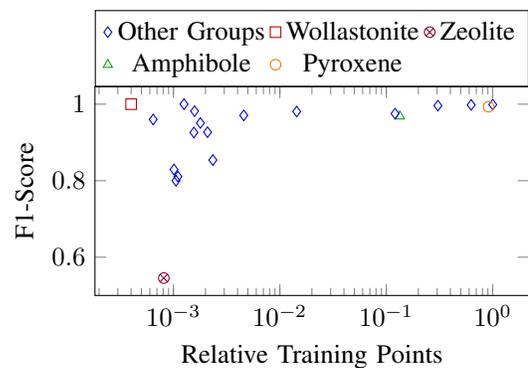


Fig. 1. The relative number of training data points per each group class (points in class/max(points in classes)) versus the F1-score of the best model for the group task (from Table I) on test data.

therefore, chemical fingerprint. The dichotomy between the accuracy of the zeolite and wollastonite groups show that sample count is not the sole indicator for performance of a class. Wollastonite is the only silicate mineral in the pyroxenoid group in our dataset. As noted by [27], pyroxenoids have a unique structure from the pyroxenes and other mineral groups. This uniqueness is compounded with that fact that $\approx 40\text{-}50\%$ of wollastonite is captured by CaO in our dataset. This value is two orders of magnitude larger than any other CaO data-point in the dataset. It is this uniqueness of wollastonite that allows us to determine and discern from other silicate minerals without a robust dataset. As the other members of this low data-point cluster do not share the same uniqueness as wollastonite, it can be observed that their low sample count hinders their accuracy. Zeolites in particular suffer from this issue. Zeolites accept a far greater range ion exchange than that of wollastonite [12]. This wide range of chemical impurities decreases the uniqueness of zeolites.

The reasoning for the difference between wollastonite and zeolite holds for the other groups with relative training points less than 10^{-2} , demonstrating that with a low number of relative training points, a mineral's lattice structure dictates the accuracy.

VIII. IMPROVING GROUP AND SUBGROUP PREDICTION

The silicate structural groups and mineral groups provide explicit divisions to create subsets of the data, and we aim to leverage this knowledge to improve classification results for the group and subgroup classification tasks. In order to utilize these divisions to attempt to improve the classifications of mineral groups, a high performing classifier of structural class, say S , is used. For a given data point v , which is a vector of elemental weights, the probability vector that is the output of $S(v)$ is appended to the end of v . This is the initial augmentation step. At this point, a classifier is trained to classify mineral groups using the augmented data points. The structural classifier S is chosen based on the average top-3 recall rather than the F1-score in order to increase the chances that the correct class is indicated strongly in the probability vector returned by $S(v)$, even if it is not the highest value in the vector. This choice was supported when testing using a subset of the data for the augmented group classification task. The results are omitted for brevity. In this way, additional data that the model can use to separate the various classes is explicitly provided by leveraging domain specific knowledge. The use of multiple models is an adaptation of ensemble learning, which demonstrated success in [6]. The macro-F1 metric is used to evaluate the resulting models across all classes in order to choose the best model to use on the test data for each task.

IX. RESULTS OF PREDICTION USING AUGMENTED MODELS

The best models for the group and structure tasks, as evaluated by the top-3 recall metric, are both the KNN classifiers.

This is shown in Table V. These are also the best models as evaluated by F1-score, shown in Table II.

The results of evaluating (on the validation data) the best models for the subgroup and group tasks, trained with data augmented by the best top-3 models, are displayed in Table VI. The change in F1-score from the old models to the new models is also displayed. The best models are KNN for both subgroup and group. The KNN subgroup classifier has a higher F1-score than the best models from the non-augmented training, while the new group classifier performs worse. The largest improvement for both the subgroup and group models is in the SVM, with an improvement of over 3% in both cases. All subgroup models showed improvement, while only the ExtraTree and SVM were improved of the group classification models, while the KNN group model F1-score and the decision tree F1-score decreased by 0.029% and 2.046%, respectively.

When evaluated on the test data, both the subgroup and group models performed worse across precision, recall, and F1-score than the non-augmented methods, as shown in Table VII. The F1-score for the subgroup model decreased by just under 0.06%, while the F1-score for the group model decreased by more than 2%. This is also shown in Table VII. The greater decrease in accuracy from the group model reflects the worse performance on the validation set. The only subgroup classes that performed worse in the new model were clinopyroxene, kaersutite, and phlogopite, by -0.004%, -1.001%, and -0.007%, respectively.

The per-class change in F1-score is displayed in Figure 2 for the groups only, as this experienced the larger decrease. The classes experience both increases and decreases in F1-score. The largest increase of just under 10% is for Analcime, the worst performing class in the original model. The largest

TABLE V
TOP-3 RECALL OF THE CLASSIFIERS SHOWN IN TABLE II (EVALUATED ON THE VALIDATION DATA) FOR GROUP AND STRUCTURE CLASSIFICATION. THE BEST RESULTS ARE IN BOLD.

Task	Top-3 Recall (%)	
	Group	Structure
DecisionTree	93.168	98.860
KNN	97.883	99.770
ExtraTree	88.466	97.427
SVM	97.179	99.568

TABLE VI
THE RESULTS OF EVALUATION ON THE VALIDATION DATA AFTER TRAINING THE SUBGROUP AND GROUP CLASSIFIERS ON THE DATASET AUGMENTED WITH THE HIGHEST TOP-3 RECALL FROM THE PREVIOUS CLASSIFIER. "CHANGE" INDICATES THE CHANGE IN ACCURACY FROM THE CLASSIFIERS TRAINED WITH THE ORIGINAL DATA TO THOSE TRAINED WITH THE AUGMENTED DATA. THE BEST RESULTS ARE IN BOLD.

ML Algorithm	SubGroup (%)		Group (%)	
	Macro F1	Change	Macro F1	Change
DecisionTree	89.286	0.858	88.860	-2.046
KNN	91.107	0.023	92.249	-0.029
ExtraTree	86.888	0.370	88.765	0.813
SVM	90.255	3.411	91.778	3.358

TABLE VII

THE RESULTS OF RUNNING THE BEST CLASSIFIERS (AS INDICATED IN TABLE VI) ON THE TEST DATASETS AFTER AUGMENTING THEM WITH THE PROBABILITIES FROM THE CLASSIFIERS WITH THE HIGHEST TOP-3 RECALL. AVERAGE PRECISION, RECALL, AND F1-SCORE OVER ALL CLASSES ARE REPORTED.

Metric	SubGroup (%)		Group (%)	
	KNN	Change	KNN	Change
Precision	95.204	-0.123	92.057	-1.237
Recall	88.546	-0.005	90.462	-1.724
F1-score	90.705	-0.060	90.302	-2.030

decrease of just over 25% is for wollastonite. Zeolite also experienced a substantial decrease in F1-score, of over 10%. All three of the most affected classes had less than 1% of the mean points in all group classes used for training (Table IV).

X. DISCUSSION OF THE RESULTS OF THE AUGMENTED MODELS

The F1-scores from training with the augmented data were higher for several of the classifiers than those from the original models on the validation data for both the subgroup and group tasks, which shows promise for the method. However, the results on the test data were worse in both cases, which indicates that there are considerable improvements to be made before this technique can be viable. The ultimate decrease in accuracy for the KNN classifiers and the large increases shown by both SVM classifiers are likely due to the increase in dimensionality of the feature vectors. This increase, which is caused by augmenting the vectors with the data regarding either the structure or group classification, could be mitigated by the use of some dimension-reducing method, such as Principal Component Analysis (PCA) [28]. The changes of the group classification results are discussed in depth, as the changes are much larger than those for the subgroup classification.

The groups whose performance improves by more than 0.1% after augmentation are serpentine, cordierite, aenigmatite, analcime, epidote, and the kyanite group. These improvements are likely a result of the new model, as the

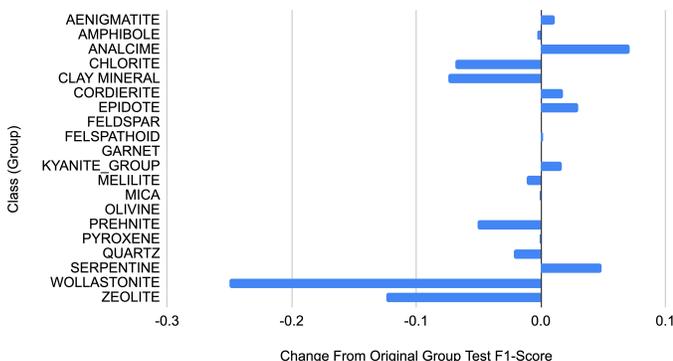


Fig. 2. The difference in classification F1-score from the original Group classifier to the Group classifier with structure data. Higher indicates better performance from the Group classifier with structure data.

train, validation, and test datasets are all constant between experiments. These mineral groups are from different structural sub-families, which disallows the possibility of crystal lattice structures dictating the change in performance after augmentation. The groups whose F1-scores decrease after augmentation are also from different structural subfamilies, maintaining this pattern.

Wollastonite and zeolite experience the most significant decrease in performance, and have the fewest and fourth fewest training points of all classes, respectively (Table IV). The low point count contributes to this observed performance loss. Additionally, the other groups that lose performance all have point counts less than 100, which is 7% of the mean points per class and 0.12% of the max point in any class. Also, by augmenting the feature vectors with resulting structural predictions, the uniqueness of the wollastonite samples that allowed them to perform well with few training samples may be reduced. The inverse holds for groups with a point count greater than 100. Mineral groups over this relative point count either experience very little change, or a notable increase in F1-score. This coupling between low point counts and a decrease in performance after augmentation leads us to believe that one can successfully apply structure augmentation to groups with high counts of relative training points.

XI. CONCLUSION AND FUTURE WORK

Previously, [5]–[7] demonstrated success using ML as a tool for mineral identification. We have expanded on their ideas by not only using a larger dataset of silicate minerals, but by being able to identify more classes at the same level of accuracy as the studies before us. We also note that although we have created the most extensive silicate mineral classifiers to the best of our knowledge, classes with fewer instances could still be improved. This leads us to believe that the development of future models that incorporate even more data, along with deeper structural information, will perform even better. Additionally, the use of methods to create synthetic data to balance the classes could be beneficial in future efforts. It is also anticipated that the use of dimension reduction will greatly improve classification results. To aid in the collection and development of the models, we will be offering free use of these models through a website [29], and the ability to upload labeled data to improve the model.

ACKNOWLEDGMENT

We thank Jack Hay and Mia Ratino for their time spent in discussions, reading drafts, and providing feedback. We also thank Reiform for funding the compute resources for this project and hosting the open source ML models on their servers.

REFERENCES

[1] W. A. Deer, R. A. Howie, and J. Zussman, *An Introduction to the Rock-Forming Minerals*. Mineralogical Society of Great Britain and Ireland, 01 2013. [Online]. Available: <https://doi.org/10.1180/DHZ>

[2] T. Koljonen and L. Carlson, "Behaviour of the major elements and minerals in sediments of four humic lakes in south-western Finland," *Fennia*, no. 137, pp. 5–47, 1975.

- [3] M. Hopkins, T. M. Harrison, and C. E. Manning, "Low heat flow inferred from > 4 Gyr zircons suggests Hadean plate boundary interactions," *Nature*, vol. 456, no. 7221, pp. 493–496, 2008.
- [4] W. A. Deer, R. A. Howie, and W. S. Wise, *Rock Forming Minerals; Single Chain Silicates*, 2nd ed. London: Geological Society of London, 1997.
- [5] E. Akkaş, L. Akin, H. Evren Çubukçu, and H. Artuner, "Application of Decision Tree Algorithm for classification and identification of natural minerals using SEM-EDS," *Computers and Geosciences*, vol. 80, pp. 38–48, 2015.
- [6] P. Jahoda, I. Drozdovskiy, S. J. Payler, L. Turchi, L. Bessone, and F. Sauro, "Machine learning for recognizing minerals from multispectral data," *Analyst*, vol. 146, pp. 184–195, 2021. [Online]. Available: <http://dx.doi.org/10.1039/D0AN01483D>
- [7] H. Hao, X. Jiang, Y. Sun, W. Dou, and Q. Gu, "A Method for Classification of Heavy Mineral Based on Machine Learning," *Proceedings - 2020 IEEE 22nd International Conference on High Performance Computing and Communications, IEEE 18th International Conference on Smart City and IEEE 6th International Conference on Data Science and Systems, HPCC-SmartCity-DSS 2020*, pp. 991–998, 2020.
- [8] Y. Zhang, M. Li, S. Han, Q. Ren, and J. Shi, "Intelligent identification for rock-mineral microscopic images using ensemble machine learning algorithms," *Sensors*, vol. 19, no. 18, 2019. [Online]. Available: <https://www.mdpi.com/1424-8220/19/18/3914>
- [9] X. Zeng, Y. Xiao, X. Ji, and G. Wang, "Mineral identification based on deep learning that combines image and mohs hardness," *Minerals*, vol. 11, no. 5, 2021. [Online]. Available: <https://www.mdpi.com/2075-163X/11/5/506>
- [10] I. C. Contreras, M. Khodadadzadeh, and R. Gloaguen, "Multi-label classification for drill-core hyperspectral mineral mapping," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 43, no. B3, pp. 383–388, 2020.
- [11] L. Tuşa et al., "Drill-core mineral abundance estimation using hyperspectral and high-resolution mineralogical data," *Remote Sensing*, vol. 12, no. 7, 2020. [Online]. Available: <https://www.mdpi.com/2072-4292/12/7/1218>
- [12] W. A. Deer, R. A. Howie, W. S. Wise, and J. Zussman, *Rock-forming minerals. Volume 4B. Framework silicates: silica minerals. Feldspathoids and the zeolites*, 2nd ed. London: Geological Society of London, 2004.
- [13] D. H. Brouwer, S. Cadars, J. Eckert, Z. Liu, O. Terasaki, and B. F. Chmelka, "A general protocol for determining the structures of molecularly ordered but noncrystalline silicate frameworks," *Journal of the American Chemical Society*, vol. 135, no. 15, pp. 5641–5655, 2013.
- [14] F. C. Hawthorne et al., "Ima report: Nomenclature of the amphibole supergroup," *American Mineralogist*, vol. 97, no. 11-12, pp. 2031–2048, 2012.
- [15] P. Bayliss, "Nomenclature of the trioctahedral chlorites," *Canadian Mineralogist*, vol. 13, pp. 178–180, 1975.
- [16] W. A. Belson, "Matching and Prediction on the Principle of Biological Classification," *Journal of the Royal Statistical Society Series C*, vol. 8, no. 2, pp. 65–75, June 1959. [Online]. Available: <https://ideas.repec.org/a/bla/jorssc/v8y1959i2p65-75.html>
- [17] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.
- [18] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.
- [19] R. J. Samworth, "Optimal weighted nearest neighbour classifiers," *The Annals of Statistics*, vol. 40, no. 5, pp. 2733 – 2763, 2012. [Online]. Available: <https://doi.org/10.1214/12-AOS1049>
- [20] "Earthchem." [Online]. Available: <http://portal.earthchem.org/>
- [21] K. Lehnert, Y. Su, C. H. Langmuir, B. Sarbas, and U. Nohl, "A global geochemical database structure for rocks," *Geochemistry, Geophysics, Geosystems*, vol. 1, no. 5, 2000.
- [22] J. Środoń, E. Zeelmaekers, and A. Derkowski, "The charge of component layers of illite-smectite in bentonites and the nature of end-member illite," *Clays and Clay Minerals*, vol. 57, no. 5, pp. 649–671, 2009.
- [23] C. S. Ross and S. B. Hendricks, "Minerals of the montmorillonite group their origin and relation to soils and clays." *Geological survey professional paper 205-b*, vol. 23-79, pp. 23–79, 1943.
- [24] T. O. Ziebold, "Precision and Sensitivity in Electron Microprobe Analysis," *Analytical Chemistry*, vol. 39, no. 8, pp. 858–861, 1967.
- [25] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [26] N. Morimoto, "Nomenclature of Pyroxenes," *Mineralogy and Petrology*, vol. 39, no. 1, pp. 55–76, 1988.
- [27] B. E. Warren and J. Bischof, "The crystal structure of the monoclinic pyroxenes," *Zeitschrift für Kristallographie - Crystalline Materials*, vol. 80, no. 1-6, pp. 391–401, 1931.
- [28] I. Jolliffe, *Principal Component Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1094–1096. [Online]. Available: https://doi.org/10.1007/978-3-642-04898-2_455
- [29] J. Hay and B. Parfitt, Jun 2021. [Online]. Available: <https://mindicator.reiform.com/>