**PROJECT MIDWAY REPORT**

**OVERVIEW:**

**Analysis of related papers on mineral classification were done. We then extracted all the required spectral data, did the data preprocessing, feature extraction and applied ML-KNN - the machine learning algorithm for multi label classification,multi classifier SVM and Random forest classifier on our data, noted down the accuracy of the results. Finally algorithms were compared.**

**LITERATURE ANALYSIS:**

Although mineral detection wasn't done using machine learning algorithms much in the past, we were still able to find some related papers on multi label classification done on minerals. We did a thorough analysis on 2 such insightful papers and also found 3 useful related papers. The two insightful papers are:

**(i) Vision based metal spectral analysis using multi-label classification.**

Eranga Ukwatta et al.(2009) DOI: <https://doi.org/10.1109/CRV.2009.42>

Aim: Using computer vision and machine learning techniques on (LIBS) spectra, presented an in situ fully automatic method for detecting constituent elements in a sample specimen.

Overview: Unlike the traditional classification problems where classes are mutually exclusive, in spectral analysis a spectrum could contain emissions from multiple elements such that the disjointness of the labels is no longer valid. Here they casted the metal detection problem as a multi-label classification and thus enabling detection of elemental composition of the specimen. They applied both SVM and ANN to multiple metal classification and compared the performance and found that both machine learning approaches yield correct identification of metals to an accuracy of 99%.

The feature set that was to be taken was chosen based on an average similarity score obtained by correlating a pattern template present in the NIST database with each signature to be classified. As the metals were earlier known the features were taken only on the peaks of the corresponding metals.

For SVM they used a one-vs-rest classification approach i.e, they decomposed the problem into a set of binary classification problems and developed a classifier for each class which considers each training sample as positive to which it belongs and as negative to all others. To determine labels of a test sample, the binary classifiers thus developed were run individually on the test sample and every label for which the output of the classifier exceeds a predetermined threshold is selected as a label of the sample.They used Radial Basis Function (RBF) as the kernel and extend the SVM to multi-label classification.Six binary classifiers were used for each metal classes. The main advantage of SVM that was considered in the paper is that the classifier complexity was seen as the property of the support vectors and not the dimension of transformed space.

For ANN one output node per class was chosen for the multi label classification.Multi-layer perceptron with one hidden layer was considered.The feed forward ANN was trained by back propagation learning technique.The data was taken without rearrangement.

Limitations found include insufficient frequency response and non-uniform quantum efficiency characteristic of the camera used.

Similarity: The paper makes use of a multi-label classification approach to metal spectroscopy which enables multi-element analysis.We are also using the multi label classification approach to identify the minerals present in the spectra. Another similarity is, accurate elemental analysis is not required but rather a qualitative analysis. Our work also involves only qualitative analysis, and not any quantitative analysis.The use of SVM and ANN showed promising results and we are testing with SVM initially as ANN was seen to have overfitting.

**(ii)Multi-label classification for drill-core hyperspectral mineral mapping.**

I. C. Contreras et al.(2020) DOI: <https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-383-2020>

Aim: A multi-label classification concept is introduced for the mineral mapping task in drill-core hyperspectral data analysis using machine learning algorithms.

Overview:For generating the training set they used SEM-MLA data, which provides high-resolution mineralogical analysis For the multi-label classification, Classifier Chain method (CC) is implemented using the Random Forest (RF) algorithm as the base classifier. The quantitative and qualitative analysis of the obtained results shows that the multi-label classification approach provides meaningful and descriptive mineral maps and outperforms the single-label RF classification for the mineral mapping task.

Similarity: We are both working on hyperspectral data related to minerals. Both of us are using machine learning classification algorithms and using multi-label classification in the process.

Some other papers which we found useful to our topic include:

(iii) Su, Q., Zhao, Y., Yang, K., & Zhang, S. (2014). *Lunar terrain and mineral’s abundance automatic analysis. Optik - International Journal for Light and Electron Optics, 125(3), 1278–1282.* doi:10.1016/j.ijleo.2013.08.022

(iv)Yu, D., & Wang, J. (2020). *A survey on machine learning in chemical spectral analysis. Journal of Information Hiding and Privacy Protection, 2(4), 165-174.* doi:http://dx.doi.org/10.32604/jihpp.2020.010466

(v) Zhang, Y., Li, M., Han, S., Ren, Q., & Shi, J. (2019). *Intelligent Identification for Rock-Mineral Microscopic Images Using Ensemble Machine Learning Algorithms. Sensors, 19(18), 3914.* doi:10.3390/s19183914

First reflectance spectra of M-3 hyperspectral data was plotted. For training and testing our data, we acquired the necessary data from RELAB and PDS spectral data from Mineral and Rock Sample Database obtained from Centre For Terrestrial and Planetary Exploration (C-TAPE). We took the sample database for:

(i) spectral data for the mixture of orthopyroxene and clinopyroxene.

(ii) spectral data for the mixture of orthopyroxene and olivine.

(iii) pure spectral data of olivine.

(iv)pure spectral data of orthopyroxene.

(v) pure spectral data of clinopyroxene.

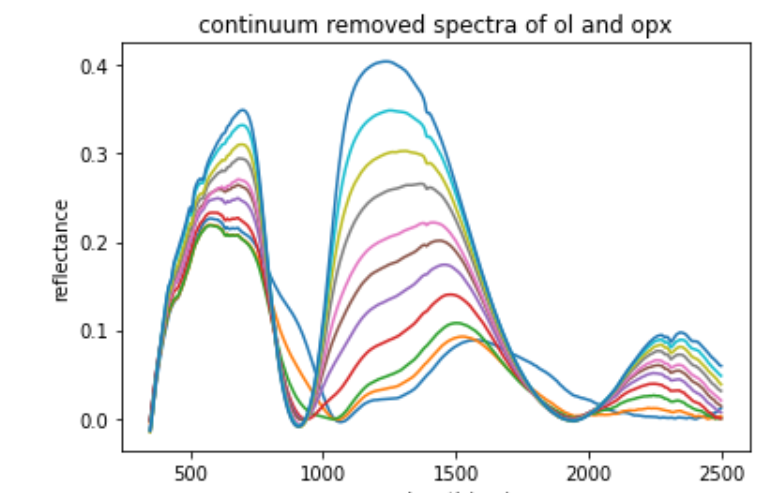
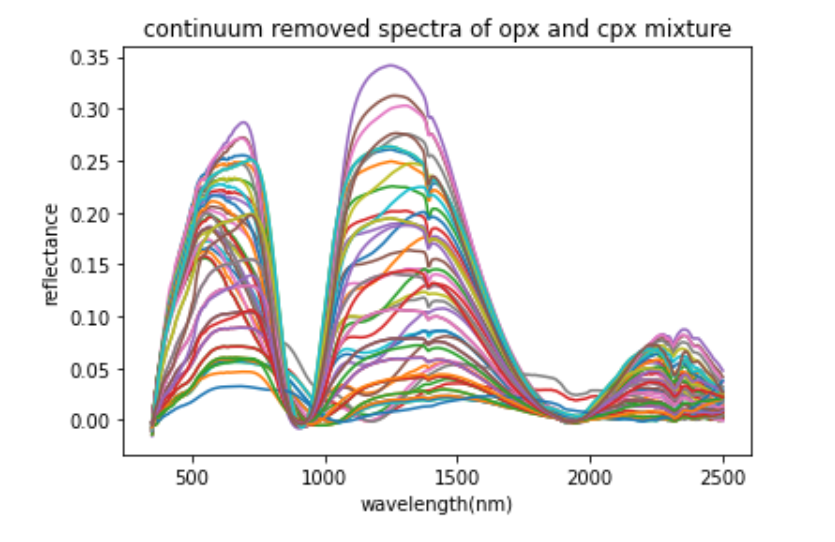
We converted the data into .xlsx format.

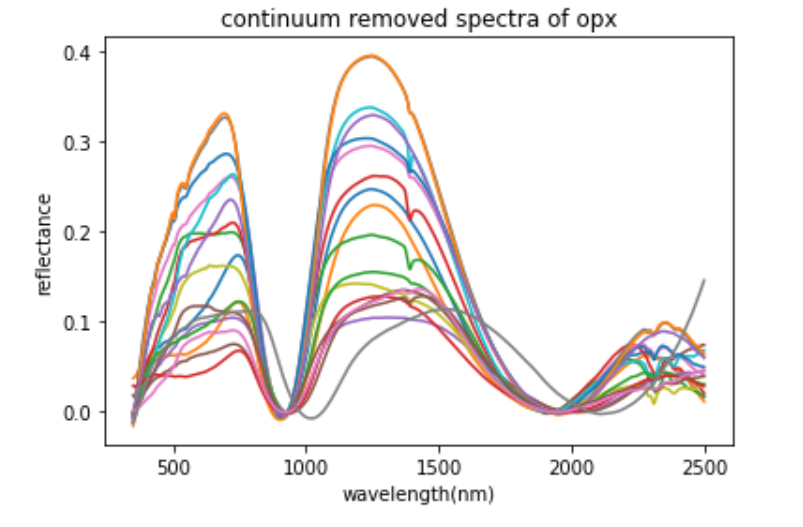
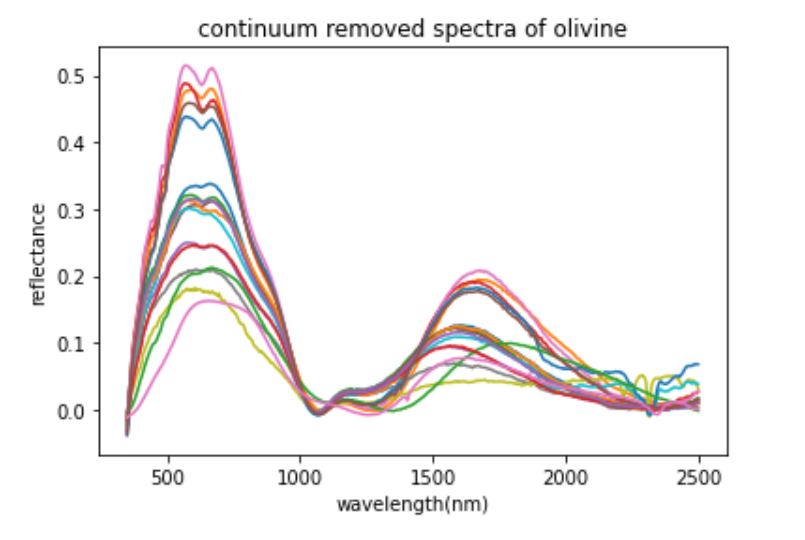
We then did the experiment using the four dataset above and compared the accuracy results obtained via the three machine learning algorithms.

**Data preprocessing:**

* **Data cleaning** was done. We found and deleted the rows and columns with missing data.
* We also made sure no duplicated data was found and the noisy data were also cleared.
* We then proceeded with the **data transformation**. This included normalization of the data, and that was done by continuum removal.

The spectra after continuum removal is as shown below:





**Feature Engineering:**

* **Feature engineering** includes Pre-processing of existing features,adding new features, and selecting the best features or combination of features based on feature importance.
* As for our dataset, we considered one of the important features as the viewing geometry, and only those with viewing geometry of 30 degree were considered.
* We also made sure that the data of the same wavelengths are being considered and compared. And that the range between the wavelengths noted are the same.
* **Feature selection** was done by plotting the number of components vs variance, using the **Principal Component Analysis** (PCA) technique. The resulting data frame was saved.

This data was then split into training and testing data by the 80/20 ratio taking 20%.

We then applied some machine learning algorithms on this data and compared the F1 score and the accuracy score among them. The machine learning algorithms applied were SVM, ML-KNN, and Random Forest algorithm. The observations obtained are as given below:

| **Classes** | **SVM** | **ML - KNN** | **Random Forest** |
| --- | --- | --- | --- |
| **F1-score of:**  Olivine  Orthopyroxene  Clinopyroxene | 0.80  1.00  0.98 | 0.89  0.97  0.86 | 0.36  0.86  0.69 |
| **Score** | 0.8823529411764706 | 0.8095238095238095 | 0.8571428571428571 |

**EXPERIMENT RESULTS:**

* Feature selection using PCA reduced features and gave good results.
* All the three machine learning algorithms gave satisfactory results and thus proved machine learning techniques can be used to solve the mineral detection problems.
* K-fold cross validation showed the training and testing data split by the 80/20 ratio gives good results
* Among the three ML algorithms, SVM is found to have the best result for this experiment.

<div style="position: relative; width: 100%; height: 0; padding-top: 56.2500%;

padding-bottom: 48px; box-shadow: 0 2px 8px 0 rgba(63,69,81,0.16); margin-top: 1.6em; margin-bottom: 0.9em; overflow: hidden;

border-radius: 8px; will-change: transform;">

<iframe loading="lazy" style="position: absolute; width: 100%; height: 100%; top: 0; left: 0; border: none; padding: 0;margin: 0;"

src="https:&#x2F;&#x2F;www.canva.com&#x2F;design&#x2F;DAEtAee0xRA&#x2F;view?embed">

</iframe>

</div>

<a href="https:&#x2F;&#x2F;www.canva.com&#x2F;design&#x2F;DAEtAee0xRA&#x2F;view?utm\_content=DAEtAee0xRA&amp;utm\_campaign=designshare&amp;utm\_medium=embeds&amp;utm\_source=link" target="\_blank" rel="noopener">Automatic Mineral Detection on Lunar Surface</a> by Insha Malayil

