APPLICATION OF MACHINE LEARNING METHODS FOR IDENTIFICATION OF SURFACE COMPOSITION THROUGH THE EXOMARS PANCAM INSTRUMENT. G. R. L. Kodikara and L. J. McHenry, Dept. of Geosciences, University of Wisconsin- Milwaukee, Milwaukee, WI 53211, gayantha@uwm.edu, lmchenry@uwm.edu.

Introduction: Here we demonstrate the application of advanced Machine Learning (ML) algorithms for the identification of mineral assemblages using the Panoramic Camera (PanCam) instrument onboard on the upcoming ExoMars Rover mission. The PanCam camera uses a dedicated "geology" filter set consisting of twelve narrowband filters of pre-determined wavelength between 440 nm and 1000 nm to map the mineralogy of the surface [1]. 200 reflectance spectra covering five mineral groups with twenty mineral species, resampled by the Geological filters on the Exomars PanCam instruments, were used to demonstrate the capability of ML methods to identify mineralogy using twelve spectral bands in the visible to near infrared wavelength region.

Methods: We use 200 reflectance spectra covering twenty mineral species in four mineral groups to help select the most suitable ML algorithm. The mineral groups includ, mafic minerals, ferric minerals, sulfate minerals, phyllosilicate minerals, and carbonate minerals. Each group was composed of forty spectra covering ten spectra of four mineral species (Table 1). Spectra were taken from RELAB and USGS spectral databases.

Table 1: Mineral reflectance spectra used for this study

Mineral Group	Mineral Species	# Spectra
Mafic	Olivine	10
	CPX	10
	OPX	10
	Plagioclase	10
Ferric	Hematite	10
	Goethite	10
	Magnetite	10
	Ferrihydrite	10
Sulfate	Gypsum	10
	Alunite	10
	Jarosite	10
	Copiapite	10
Phyllosilicate	Nontronite	10
	Montmorillonite	10
	Saponite	10
	Serpentine	10
Carbonate	Magnesite	10
	Dolomite	10
	Calcite	10
	Aragonite	10
Total	20	200

All spectra were resampled using Gaussian spectral response functions defined by the fwhm (full-width-half-maximum) values of the geologic filters of the PanCam instrument to compare directly with PanCam multispectral image data [1]. We created a Spectral Resampling Bandpass Filter. Eleven band indices were initially chosen for this analysis based on the literature (Table 2) [2,3].

Table 2: Calculated spectral parameters (band indices).

Index	Description
BS01	$(R_{670} - R_{440})/(670 - 440)$
BS02	$(R_{610} - R_{530})/(610 - 530)$
BS03	$(R_{1000} - R_{740})/(1000 - 740)$
BS04	$(R_{1000} - R_{950})/(1000 - 950)$
BR01	R_{670}/R_{440}
BR02	R_{1000}/R_{740}
BR03	R_{1000}/R_{950}
BD01	$1 - (R_{530}/[(0.530 * R_{500}) + (0.470 * R_{570})]$
BD02	$1 - (R_{900}/[(0.455 * R_{840}) + (0.545 * R_{950})]$
BD03	$1 - (R_{610}/[(0.600 * R_{570}) + (0.400 * R_{670})]$
BD04	$1 - (R_{950}/[(0.500 * R_{900}) + (0.500 * R_{1000})]$

Feature selection is an effective way to identify the most important spectral parameters in a dataset and discard others as irrelevant or redundant. Here we use the Learning Vector Quantization (LVQ) model to estimate the feature importance and the best feature combination was then used as the input features for the ML algorithms [4]. We used stratified random sampling with proportional allocations to split the entire dataset into two sets as training (to train the models) and validation (to evaluate their performance). Training sets include 80% of the observations (160 observations) and the rest are assigned as validation (40 observations) and kept aside to measure the accuracy of the winning ML algorithm/s. We adopted nine ML algorithms, including Linear Discriminant analysis (lda), Generalized Linear Models (glmnet), Partial Least Squares (pls), Support Vector Machine (svm), Naïve Bayers, Neural Network (nnet), Random Forest (rf), C5.0 and Boosted Trees in the R statistical software package [5, 6]. We used the k-fold cross validation method to estimate the test error associated with each machine learning to evaluate their performance on the training dataset [7]. For that, the training dataset was split into ten parts, nine to train and one for test, and program runs were conducted for all combinations of train-test splits. We also repeated the

same process three times for each algorithm to achieve the most reliable results. Finally, model performance was measured using the validation data set with two statistical measures, Overall accuracy and Kappa.

Results: We found that BD02 (band depth at 900 nm wavelength) was the best band index to classify the selected mineral groups, except for the sulfate mineral group (Figure 1).

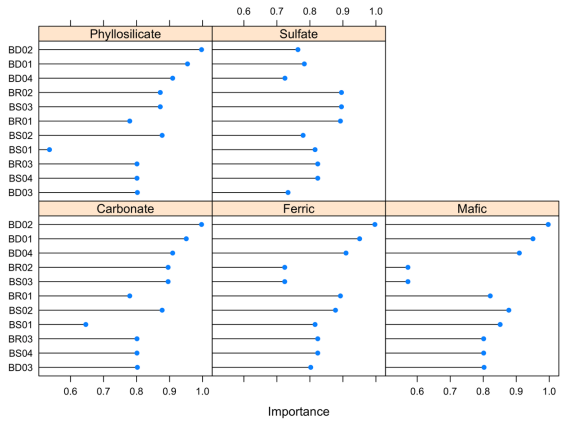

Figure 1: Importance scores of each band index for each mineral group (Importance are in 0-1 scale).

BR02 (ratio of bands at 1000 nm and 740 nm) was the most important band index for classifying the sulfate mineral group, while it was the least important for the Mafic mineral group. BD01 (band depth at 530 nm wavelength) was the second most important spectral parameter for classifying these mineral groups. Combination of spectral parameters BD02, BD01, and BR02 show the highest performance after calculating different feature combinations. Random forest and C5.0 ML algorithms were the best ML algorithms to identify those mineral groups using these three spectral parameters (Figure 2). Accuracy tells us the percentage of observations that the model classified correctly, while the kappa statistics tell us how well two evaluators can classify an observation correctly.

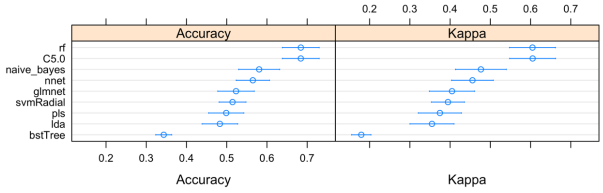

Figure 2: Accuracy and interrater reliability of adopted Machine learning methods. The different approaches yielded accuracies between 50 and 60%.

Future Work: In the work presented here, we only use the mineral group as the classification parameter. We are continuing our research to identify the best spectral parameter/s to identify different mineral species using the PanCam resampled mineral spectra. The best ML algorithm can be used to help map mineral compositions using the PanCam image data.

References: [1] Vago J. L. et al. (2017) Astrobiology, 17 (6). 471-510. [2] Harris, J. K. et al. (2015) Icarus, 252. 284-300. [3] Cousins, C. R. et al. (2012) Planetary and Space Science, 71. 80-100. [4] Kohonen, T. (2001) Self-Organizing Maps, 501 pp. [5] Kuhn, M. (2008) Journal of Statistical Software, 28(5), 26 pp. [6] Kuhn M. and Johnson K. (2016) Applied Predictive Modeling, 600 pp. [7] James G. W. et al. (2017) An Introduction to Statistical Learning with Applications in R, 426 pp.