MACHINE LEARNING APPLIED TO MSL/CHEMCAM DATA. O. Forni[1], O. Gasnault[1], A. Cousin[1], R. B. Anderson[2], E. Dehouck[3], G. David[1], P. Pinet[1], C. Fabre[4], J. C. Bridges[5], R. C. Wiens[6], S. Maurice[1], P.-Y. Meslin[1], J. Lasue[1], N. Thomas[7]. [1]Institut de Recherches en Astrophysique et Planétologie, Toulouse, France, [2]USGS, Flagstaff, USA, [3]LGL-TPE, Lyon, France, [4]GeoRessources, Nancy, France, [5]University of Leicester, UK, [6]LANL, Los Alamos, USA, [7]Caltech, Pasadena, USA. [olivier.forni@irap.omp.eu]

Introduction: ChemCam is an active remote sensing instrument suite that has operated successfully on MSL since its landing in 2012 [1, 2]. It uses laser pulses to remove dust and to profile through weathering coatings of rocks up to 7 m away. Laser-induced breakdown spectroscopy (LIBS) obtains emission spectra of materials ablated from the samples in electronically excited states. The intensities of these lines are proportional to the abundance of the related element. ChemCam is sensitive to most chemical major elements as well as to a set of minor and trace elements such as Li, Sr, Ba, and Rb that are quantified. Qualitative and quantitative relationships between elements can be identified using univariate and multivariate techniques [3, 4]. One challenge of the data interpretation is to rapidly identify similar compositional and hence mineralogical phases or mixtures in the LIBS spectra. Some classification analyses have already been performed on these data using Independent Component Analysis (ICA) [5] or clustering techniques [6]. In this paper, we want to test and evaluate the performances of machine learning techniques. Here, we specifically focus on Artificial Neural Network (ANN) [7] .

ANN: In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The connections between artificial neurons are called 'edges'. Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times. In this application we use a multilayer perceptron (MLP) which is a family of feedforward artificial neural network. An MLP consists of, at least, three layers of nodes: an input layer, a hidden layer and an output layer. We use this configuration in our application. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. For the first two layers the activation function is the "Rectified Linear Unit" and "Softmax" for the last layer. The loss function which measures the inconsistency between the predicted value and the actual label is the categorical cross-entropy. In this application we will use the Keras package (https://keras.io/) which is a high-level neural networks API, written in Python and capable of running on top of TensorFlow (https://www.tensorflow.org/) that is an open-source machine learning library for research and production.

Data pre-processing: It is common to pre-process the data before feeding the model. Usual procedures include normalization of the data and/or achieving a dimensionality reduction like the projection on eigenvectors of the covariance matrix (Principal Component Analysis (PCA)). We choose instead to use the Independent Component Analysis (ICA) which has been demonstrated to have better selection properties than PCA [8]. After this operation we are left with 10 ICA score components which need to be compared to the 6144 channels of the ChemCam LIBS spectrum.

Results: As a preliminary test we will focus on three distinct families, two "simple," i.e., Ca-sulphates and Fe-rich, and one "complex," i.e., K-feldspar. In the first step we need to build the training database for these three families. We choose to take a known sample of each family and to correlate it with all the spectra. We retain the spectra that have a correlation coefficient larger than 0.95. The samples are "Rapitan-9", "Stark" and "Egg_Rock-1" for the Ca-sulphates, K-feldspar and Fe-rich respectively. With this procedure, we select 834, 91 and 26 samples for the Ca-sulphates, Fe-rich and K-feldspar respectively. The low number of samples in the K-feldspar family may potentially explain the difficulty of obtaining a good training set for this family. We run some cross-validation models having divided each set of families in four parts. On average, the Ca-sulphates are predicted in the test sets at 98%, the Fe-rich at 91% and the K-feldspar at 75%. This lower value for the K-feldspar may be explained in two ways: either due to the sparsity of the samples or the way we select the samples. Indeed the simple correlation method may select targets that are not K-feldspars. Interestingly two of those targets are always the same, namely Keith-4 and Weed_Creek-3 that, looking at the spectra and at their composition may question their relationship to the K-feldspar family

(Fig. 1). Using the model we have built we can now predict for all the 17251 ChemCam spectra (up to Sol 2247) to which family they belong and analyse how the model performs, i.e., do the spectra that are attributed to a family and that are not in the training set belong effectively to that family? The model returns a number between 0. and 1. for each sample and each class. To attribute a sample to a family we need to put a threshold above which the sample is said to belong to the family. Depending on this threshold, it can happen that a given sample belongs to two classes.

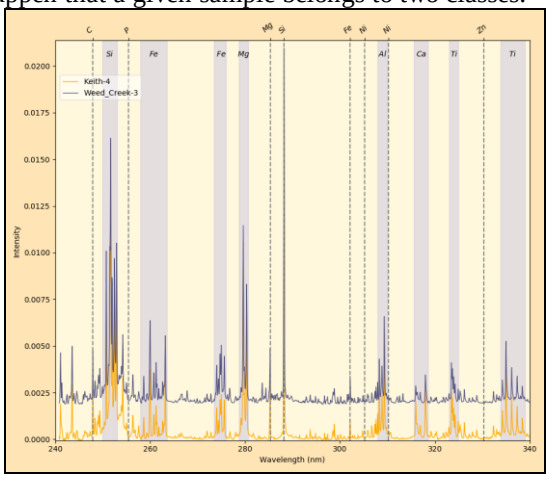

Figure 1: Keith-4 and Weed-Creek-3 spectra showing the high Si content but also large Ti lines

In the Ca-sulphate family, two points are found to be a fluorite and an apatite and not a Ca-sulphate, namely Alvord-1 and Yampi-7 because they exhibit the characteristic CaF molecular emission line [9]. It is worth noting that many others large CaF emission bearing spectra are not selected by the model like the two others Alvord points or Epworth-5 demonstrating the relatively good robustness of the model. In the Fe-rich family, nodules points isolated in the bedrock have been detected like Grange [10]. Interestingly, some mixtures of Fe-rich with silicates are identified like Kilpfonteinheuvel_ccam-1, as well as Ca-sulphate-rich targets like Quarry_Haven-4, in which target points #3, #9 and #10 are flagged as Ca-sulphate. Newmachar-DRT-9 is shown to also belong to the Ca-sulphate family (fig. 2). The K-feldspar family is interesting because the model finds almost all the K-feldspars that have been previously identified [11], along with some others that weren't identified before, like some blind targets (CC_BT_386a, CC_BT_434a, shot at a fixed angle from the rover) [12]. In the K-feldspar class the Askival target, which looks like a cumulate and exhibits light-toned (major) phenocrysts, dark-toned well crystallized fine-grained patches, and gray-toned and dark-toned veins [13], shows up strikingly (Fig 3.)

Summary: Preliminary results using machine learning techniques are encouraging. Further improvements can be considered, like the construction of reliable training sets using unsupervised classification techniques. From the neural network point of view, other choices of the optimizer should be tested as well as number of layers and their neuron contents. An obvious development of this work will be to include more classes.

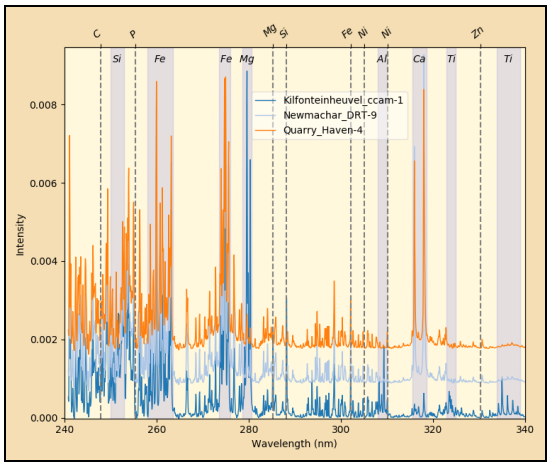

Figure 2: Fe-rich samples detected by the model exhibiting the Ca lines in Newmachar-DRT-9 and Quarry_Haven-4 and Si lines in Klipfonteinheuvel_ccam-1.

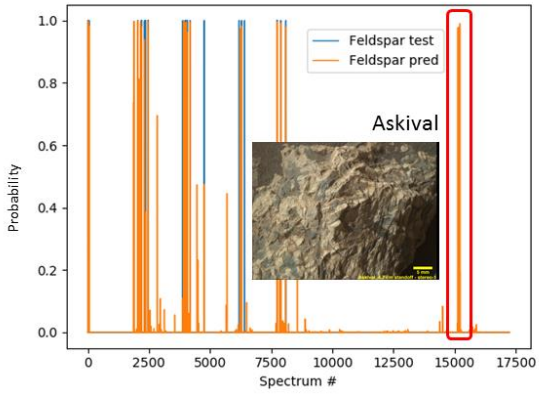

Figure 3. Detection of K-feldspars The Askival target is strongly selected.

References : [1] Maurice S. et al. (2012) *SSR, 170*, 95 [2] Wiens R.C. et al. (2012) *SSR, 170*, 167 [3] Clegg S.M. et al., (2017) *SCAB, 129*, 64. [4] Payré V. et al.. (2017) *JGR, 122*, 650 [5] Forni O. et al. (2013) *LPS XLIV*, Abstract #1262. [6] Gasnault O. et al. (2015) *LPS XLVI*, Abstract #2789. [7] Anderson R .B. *Icarus, 215*, 608-627.[8] Forni O. et al. (2013) *SCAB, 86*, 31-41. [9] Forni et al. (2015) *GRL, 42*, 1020-1028. [10] David et al. (2018) *LPS XLIX*, Abstract 2079 [11] Cousin et al., (2017) *Icarus 288*, 265-283. [12] Cousin A. et al. (2014) *LPS XLV*, Abstract 1278 [13] Bridges et al (2019), this meeting.