Automated Analysis of Mixed Sample Raman Spectra Using Feedforward Neural Networks and One-VS-All Decomposition

A. Atkinson[1], M.N. Abedin[2], G.D. Hines[2], A.T. Bradley[2], H. Elsayed-Ali[1] [1]Old Dominion University, Norfolk VA 23508, [2]NASA Langley Research Center, Hampton VA 23681, aatki011@odu.edu, m.n.abedin@nasa.gov

Introduction:

Interest in the use of Raman spectrometers has seen an increase in the fields of geology and planetary sciences due to the non-destructive insight Raman spectra may provide into the molecular makeup of a given sample. Advancements in Raman spectrometer hardware have allowed for compact instruments to have deployment capabilities directly on interplanetary missions, flexible usage conditions requiring no sample collection/preparation, and no need for daylight radiation shielding[1,2].

As the amount of science which can be collected from a Raman spectrometer in a given amount of time increases, a bottleneck will be created in data analysis which leaves a need for a faster method of spectral data classification. Recent studies[3,4] have shown that machine learning models are able to solve this problem by achieving high-accuracy classification. Liu et al[4] found the convolutional neural network (CNN) held the highest classification accuracy (96% top 5) for single sample Raman data.

Multi Label Classification:

Our experiments have shown that when multiple samples are struck by the same incident, Raman peaks from all samples may be present in its line profile[1]. In machine learning a case such as thus, when a single input has the potential to contain features from more than one learned class, is referred to as a multi-label problem[5]. In Liu et al a CNN was used for classification of single sample Raman data, however, due to the lack of spatial encoding in CNNs and the use of a softmax layer (SML), transferring the network to multi-label classification would not be feasible[6]. The weakness of SML for multi-label classification can be seen by visualizing the input weights of a SML which has been trained for pure sample classification, as shown in Figure 1.

Figure 1 a and c show low frequency region pure sample data from naphthalene and sulfur respectively which was used to train a SML for classification, collected at LaRC with a Kaiser Holoplex f/1.8 spectrometer, Big Sky Laser UltraCFR laser at 532nm, and Princeton Instruments PIMAX ICCD camera. Figure 1 b and c show the learned input weights of the two output nodes of the SML. It can be seen that the network not only learned to associate spectral features with a specific class (positive weights) but it also assumes that features are mutually exclusive from each other (negative weights), making multi-label classification impossible with this type of model.

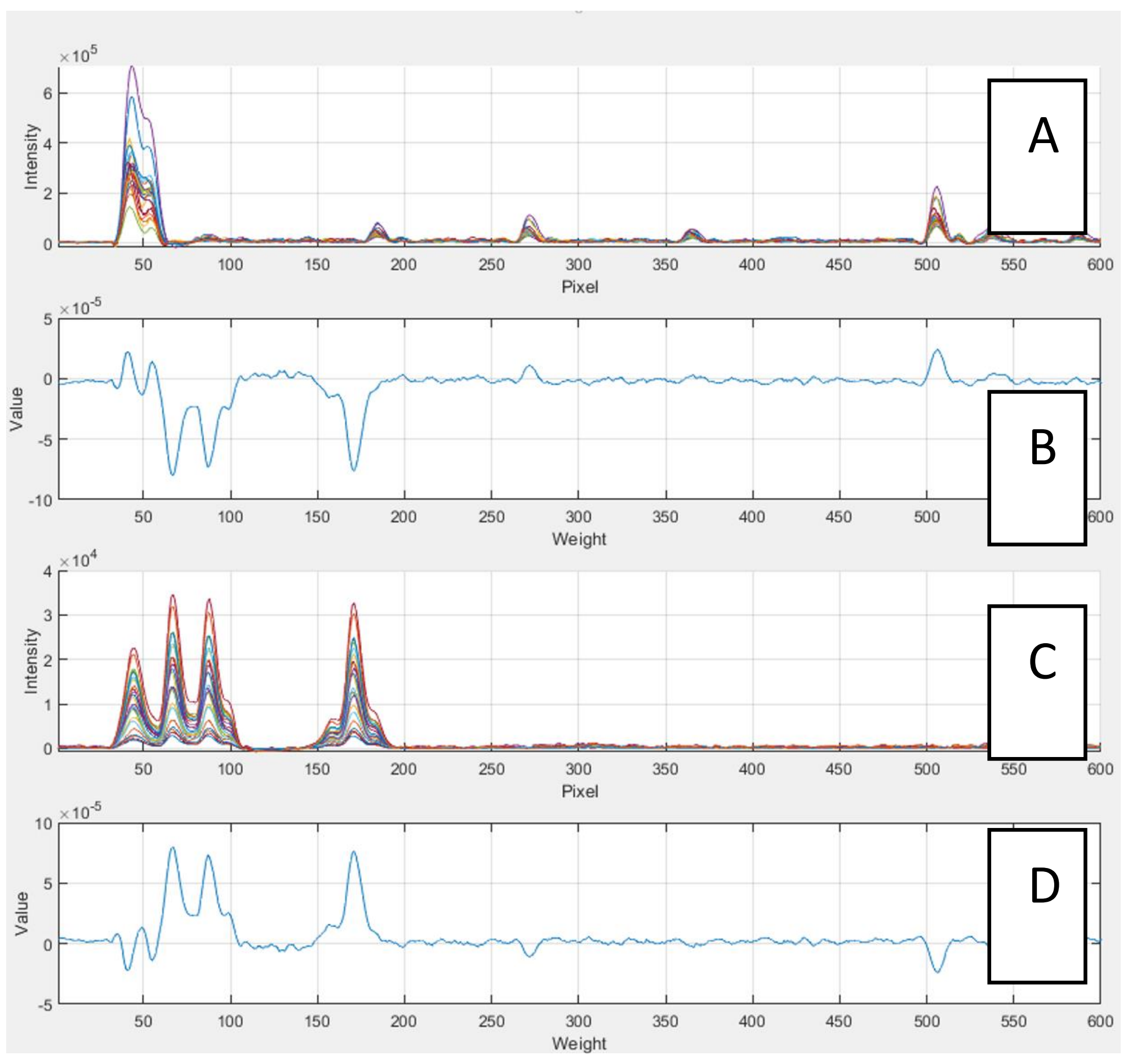

Figure 1 Pure Naphthalene and Pure sulfur training data (a,c) and their corresponding learned input weights for a trained SML (b,d).

Methodology:

Multi-label classification of mixed sample Raman spectra was achieved by implementing a model which avoids the issues found in CNN and SML by decomposing an *N*-class multi-label problem into "*N*" single class detection problems following the pseudocode in Procedure 1.[5]

Following Procedure 1 delivers the user a model consisting of multiple feedforward neural networks (fnn), each referred to as a "sample detector". Each sample detector is trained to detect spectral features from a single pure sample very well compared to all others. To reduce the effects of assumed mutual exclusivity the input lengths to each detector are limited to only areas of the line profile where spectral features of its respective pure sample are found, each region is referred to as a "zone of interest". The dataset used to train each detector consists only of spectral data extracted from its zones of interest and is decomposed to only have two classes.

Procedure 1 – Training Algorithm

```
for each pure sample i in training data
   detect regions of line profile containing spectral data
   initialize fnn with input length equal to the sum of all region lengths
   create a two class training dataset by extracting regional data
   train fnn using two class dataset
end
```

A mixed sample data point is classified by iteratively analyzing an input line profile with each trained sample detector, following Procedure 2. Line profile data from all sample detector's zones of interest is extracted. The extracted data is then forward propagated through all sample detectors. If a user-set activation threshold is reached, the pure sample associated with an arbitrary detector is labeled as present in the mixed sample being analyzed.

Procedure 2 – Analysis Algorithm

```
for each sample detector i in a trained model
   extract spectral data from the detector's regions of interest
   forward propagate extracted data through the detector's fnn
   if threshold activation reached
      report pure sample as detected
   else
      report pure sample as not detected
   end
end
```

The model can be run in two modes, the first mode returns a single binary value ("detected" or "not detected") for each pure sample in the training dataset. The second mode trains a fnn for every region of interest found for each pure sample, and outputs the amount of regions of interest which were activated by an input data point for each trained pure sample. It was experimentally found to be most effective to use both methods in parallel with each other. Graphical representation of the training and analysis procedures for the model used in this experiment are shown in Figure 2.

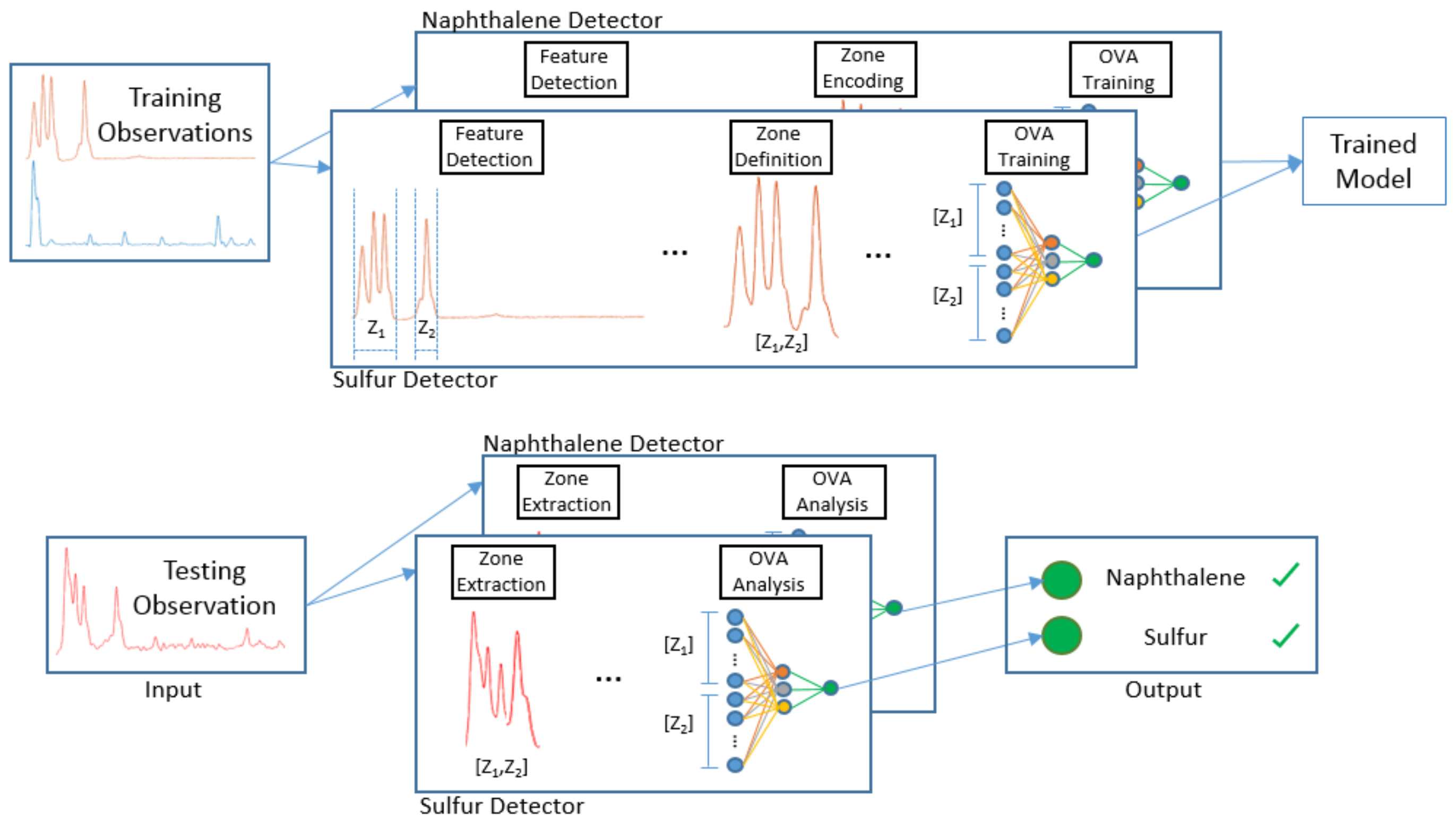

Figure 2 Dataflow diagram of training (top) and analysis (bottom) procedures for the two class multi-label model used in this experiment.

Results:

Figure 3 shows a front image of a sample composed of coarsely crushed naphthalene and sulfur. Each spot on the image overlay represents the center of the laser strike-point for a single data point. Between each exposure, the sample was carefully moved in steps of 1mm to ensure the mixture was not disturbed, data points were collected from a total area of 1cm^2. The color of each spot corresponds to the label(s) that a trained model applied to that data point. A red spot means features from both naphthalene and sulfur had activations, a yellow spot means only sulfur features had strong activation, and a green spot means only naphthalene features had strong activation.

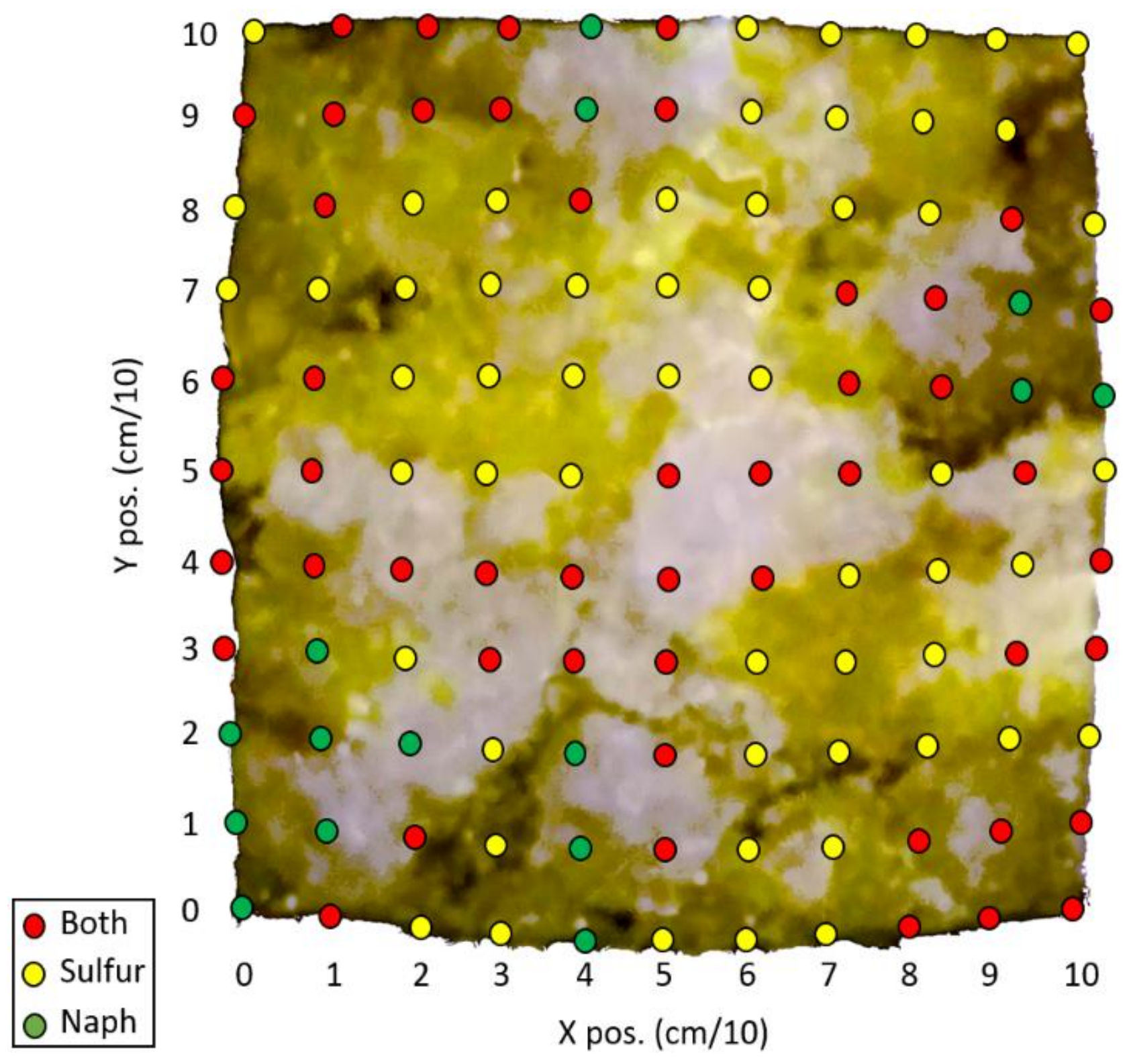

Figure 3 High contrast image of mixed naphthalene and sulfur. Red = mixed classification, yellow = sulfur classification, green = naphthalene classification.

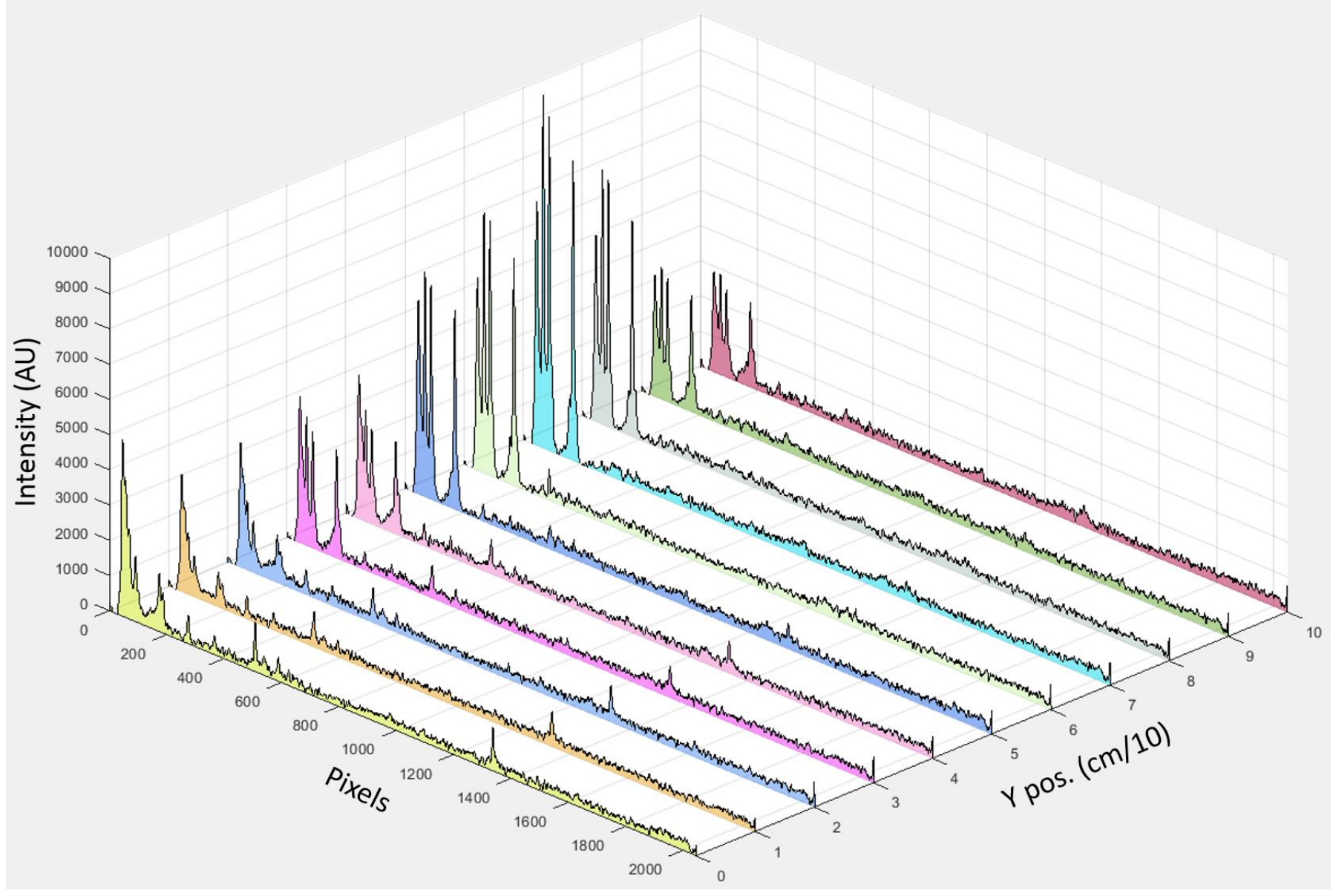

Figure 4 Line profiles of data points (0,0) – (0,10).

Summary:

- Advancements in Raman hardware are creating a need for faster analysis methods.
- A model has been developed which can analyze mixed Raman samples after being trained with only pure sample data.

Acknowledgements:

This research has been supported in part by NASA LaRC Internal Research and Development (IRAD) program. We would like to thank all involved for inspiration and critique.

References:

[1] Abedin, M.N. et al, (2018), Applied Optics, 57, 62. **[2]** Mars.nasa.gov, (2017), SHERLOC For Scientists - Mars 2020 Rover. **[3]** Carey, C. et al, (2015), Journal of Raman Spectroscopy, 46, 894-903. **[4]** Liu, J. et al, (2017), The Analyst, 142, 4067- 4074 **[5]** Pushpa, M., Karpagavalli, S., Procedia Computer Science, 115, 572-579. **[6]** Chatterjee, S., (2017), Saama Technologies, https://www.saama.com/blog/capsule-networks-limitations-cnns/.