STATISTICAL CLASSIFICATION OF BIOSIGNATURE INFORMATION: IMPROVING LIFE-DETECTION CONFIDENCE USING MACHINE LEARNING

Abdullah Shahid[1,2], Tao Sheng[3,4,5], Jesse B. Murray[7], Aarya Mishra[5,6], Ellen Czaplinski[9],Sunanda Sharma[8,9], Diana Gentry[10]
NASA Ames Research Center – OSTEM Internship Program[1], North Carolina State University[2], NASA Ames Research Center - Space Life Sciences Training Program / KBR Inc.[3], University of Pittsburgh[4], NASA Ames Research Center - Volunteer Internship Program [5], University of San Francisco[6], University of Oxford[7], Massachusetts Institute of Technology[8], NASA Jet Propulsion Laboratory[9], NASA Ames Research Center[10]

Introduction: Astrobiology, and the search for signs of past or present life in the universe, are a core priority for the upcoming decade of space exploration [1], especially the need to understand how multiple observations of a system (i.e., multiple potential biosignatures) can be used to increase confidence in life detection [2]. This is particularly true for agnostic biosignatures, or those not specific to a biochemical basis or mechanism [3], in contrast to otherwise highly specific biosignatures such as DNA, chlorophyll, ATP, etc. This project aims to use the extensive amounts of terrestrial data available from biogenic and abiogenic systems to create a binary classifier for life detection. The preliminary data set is limited to measurements that have been previously suggested as agnostic biosignatures, including elemental abundance and distribution, isotopic fractionation, VNIR reflectance spectra, and Raman spectra. This work aims to determine which combinations of features across these data types are most relevant to life detection through assessing feature importance in multiple machine learning algorithms. Our work will help establish which data types and features are most valuable for planning future life detection missions.

Materials and Methods: *Data selection:* Data types (elemental abundance, isotopic fractionation, reflectance spectra, and Raman spectra) were collected from various public databases, publications, and lab-recorded measurements to create a representative-systems dataset. The representative systems, such as basalt, bone, or biofilm, were chosen to be unambiguously classifiable as indicative or non-indicative of life -- edge cases such as prions, protobionts, and technological devices were not included. The indicative systems were further tagged as indicative alive (microbes, vegetation, etc.), indicative mixed (seawater, soil, etc.) and indicative not-alive (bone, coal, etc.) to track whether some could be more effectively classified than others. Examples of non-indicative systems were lunar rock, sand or basalt. Overall, 15 indicative and 8 non indicative representative systems were created from a total 77 indicative samples and 255 non indicative samples.

Data Collection: Elemental and isotopic fractionation data was collected for each representative system using various publications reporting results from X-ray diffraction and laser-induced breakdown spectroscopy techniques. Multiple measurements of the elemental distribution and isotopic fraction were aggregated into one metric for each system through geometric and arithmetic means respectively. Databases such as ECOSTRESS [4], USGS [5], RELAB [6], and PDS-CRISM [7] were used for the reflectance spectroscopy data. Public sources such as the RRUFF [8] and others were used to collect Raman spectroscopy data for a majority of the systems. Lab recorded measurements of reflectance and Raman spectroscopy data of soil, ice, and seawater were used to complete the dataset.

Standardization: Measurements were standardized to a common limits of detection, range, etc. Elemental abundance measurements were set to an artificial limit of detection of 1.5×10^{-5}. Absent isotopic fractionation data for a specific isotope was replaced with the mean of all data for that isotope. Raman spectra with incident wavelengths of 532 and 514.5 nm and similar instrumentation were used. Finally, for reflectance spectroscopy, only 200-2100 nm data points were used after ordering the spectra from lowest to highest wavelength.

Feature Extraction: The elemental features used for all the representative systems include fractional content of C, O, K, N, H, P, Mg, S, Ca, Na, Cl, Mn, Al, Si, Fe, and Ti. The isotopic fractionation data features were Carbon-13, Oxygen-18 and deuterium deltas between the ratios of heavier and lighter isotopes. Numerical analysis using SciPy signal [9] was done to determine the number and position of peaks, troughs, peak widths, mean reflectance, and mean peak widths; these served as features for the reflectance spectroscopy data. Similarly, the Raman spectroscopy data features extracted were mean intensity value, number of peaks, number of troughs, broadest peak, and mean peak width. These features were used in the following algorithmic implementation.

Machine Learning Implementation: The data was trained and tested on *k*-nearest neighbors (KNN), logistic regression with L2 regularization (LR), Random Forest (RF), support vector machines (SVM), logistic regression, and Gaussian Naïve Bayes (GNB), and then with a voting classifier combining the output from all of the above. Additionally, Principal Component Analysis was used an unsupervised learning method to supplement findings from the supervised learning models. The classification performance was evaluated with 2,000 50% train-test splits with Monte Carlo simulations. The

voting classifier was then iteratively run with each one of the 4 data types removed to allow assessment of the relative value of each combination of data types. After this, feature importance was tested by removing individual features to assess relative feature importance.

Results: *Accuracy:* Ranging from zero to one, the Receiver Operating Characteristic Area Under the Curve (ROC AUC) is a standard metric for classification accuracy (0.5 is equivalent to random guessing). The ROC AUC of the combined voting classifier was 0.853 and this was similar to the performance of the model with only elemental distribution (AUC = 0.859) and only Raman spectroscopy (AUC = 0.86). However only using isotopic fractionation (AUC = 0.716) and only using reflectance spectroscopy (AUC = 0.474) led to a significant decrease in the accuracy of the model.

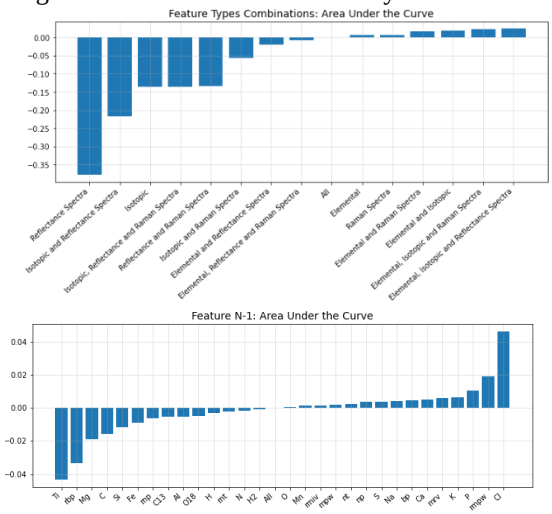

Figure 1*: Accuracy changes resulting from removal of groups of data feature types (top) and individual data features (bottom).*

Data Type Feature Testing: When evaluating the importance of groups of data features, removal of the elemental data type causes the most significant decrease in the model's accuracy. All significant decreases of 13% and more are present when elemental abundance is removed from the model training data (Figure 1). Raman spectroscopy displays strong performance when tested individually; however, it shows a greater decrease in accuracy (>5%) when combined with data types other than elemental abundance. Isotopic fractionation and reflectance spectroscopy exhibit a low performance when used individually and this translated to the combination test where, when combined with each other, a significant decrease in performance was seen (>20%); accuracy only increased when Raman spectroscopy or elemental abundance was added to the combination.

Individual Feature Testing: The most significant decrease in accuracy was found with the removal of the element abundance data feature titanium (Δ AUC of -0.0432) and the Raman spectroscopy data feature broadest peaks (Δ AUC of -0.0334). Conversely, when the elemental abundance data feature chlorine is removed, we see an increase in accuracy (Δ AUC of 0.0463). The results suggest that the individual data features do not cause extreme variations in accuracy. Therefore, overall no general trend is seen where one individual data feature is substantially more valuable. However, certain data features are more significant than others in context of data types; for example, Raman broadest peaks leads to greatest decrease in accuracy and hence it is the most significant feature within the Raman spectra data type.

Conclusions and Discussion: The results suggest that elemental abundance distribution and Raman spectral features are valuable in assessing whether a sample may be indicative or not indicative of life. For example, these two data types and algorithm could be used to automatically exclude likely abiogenic (low interest) samples while prioritizing samples for further robotic investigation or planned sample return. The model AUC of 0.853 is very good in context of a relatively small and broad collected dataset and implementation, and thus serves as a proof of concept for further development with additional data, which could be specifically selected to represent a mission target environment such as an ocean world, polar ice, or subsurface aquifer. The ability to quantitatively assess how different instruments increase science return is particularly valuable.

Future Work: Work is ongoing to improve the data standardization, particularly the handling of non-reported values, as well as to add more systems and new data types. Work is also underway investigating a convolutional neural network to extract features from the Raman and VNIR spectra, a more general approach than the features manually selected in this preliminary effort. We are evaluating venues to make the standardized data publicly available.

Acknowledgements: NASA Internships, Fellowships and Scholarships (NIFS) Program

References:[1] NASEM (2022), https://doi.org/10.17226/26522 [2] Neveu et al., (2018), https://doi.org/10.1089/ast.2017.1773. [3] NASEM, (2019) https://doi.org/10.17226/25252. [4] Meerdink et al., (2019), https://doi.org/10.1016/j.rse.2019.05.015 [5] Kokaly et al., (2017), https://doi.org/10.3133/ds1035 [6] PDS Geosciences Node. RELAB spectral database, 2014. http://www.planetary.brown.edu /relabdocs/relab.htm [7] PDS Geosciences Node. CRISM spectral library. https://speclib.rsl.wustl.edu, [8] Lafuente et al., (2015), https://doi.org/10.1515/9783110417104-003 [9] Virtanen et al., (2020), https://doi.org/10.1038/s41592-019-0686-2