

Methodology for Statistical Classification of Biosignatures Information (SCOBI)

Abdullah Shahid¹, Tao Sheng², Sunanda Sharma³, Diana Gentry⁴

¹North Carolina State University Raleigh, NC (amshahid@ncsu.edu), ²University of Pittsburgh, Pittsburgh, PA,

³Propulsion Laboratory, California Institute of Technology, Pasadena CA, ⁴NASA Ames Research Center, Mountain View, CA

Introduction: Recent advancements in astrobiology highlight the crucial role of Machine Learning (ML) in identifying biosignatures, including analyzing terrestrial analogues for planetary mapping and habitat identification [1] and differentiating modern and ancient biosignatures [2].

This project employs ML to create a binary classifier for detecting life from multiple instrument observations, using a large dataset from both biogenic and abiogenic systems. We combined elemental abundance, isotopic fractionation, VNIR reflectance, and Raman spectra with supervised learning algorithms like k-nearest neighbors (KNN), logistic regression (LR), support vector machines (SVM-PC), random forest (RF), Gaussian naïve Bayes (GNB) and a voting classifier integrating the first four. Our goal is to identify the most relevant feature combinations for life detection and assess feature importance. This research is significant for onboard data classification in space missions and can influence instrument selection for enhanced scientific return. This abstract focuses on the methodology of using diverse datasets and ML algorithms for effective life detection. The dataset is described in more detail in [3] also, and the results in [4].

Dataset Collection: Data was sourced from multiple public databases, such as the RRUFF Project, USGS Library, and others, along with publications and laboratory measurements. We compiled a database comprising 1277 measurements from 16 selected systems, categorized as either *indicative* or *non-indicative* of life. Indicative systems were further subdivided into categories: alive (microbes, vegetation), mixed (seawater, soil), and not-alive (bone, coal). Non-indicative examples included lunar rock and sand.

Feature Extraction: Elemental abundance data consisted of C, O, K, N, H, P, Mg, S, Ca, Na, Cl, Mn, Al, Si, Fe, and Ti. An artificial limit of detection of 1.5×10^{-5} was imposed as a threshold after multiple sensitivity tests. The isotopic fractionation data features were ^{13}C , ^{18}O and ^2H deltas; Absent fractionation data was replaced by the means of all data for that isotope. SciPy's signal package was used to extract features from the reflectance spectra systems between the ranges of 200-2100 nm. For the Raman spectra, a Hampel filter was used to remove outliers. Then SciPy's signal was used to derive summary parameters similar to the reflectance spectra. Raman spectra were split at 3000 cm^{-1} and features were calculated individually and later

recombined to avoid overweighting the O-H stretch region. For all representative systems data was initially collected as individual samples and then averaged for the respective representative systems which was then combined for the final dataset.

Data Exploration: Principal Components Analysis (PCA) was employed to understand the distribution and organization of both indicative and non-indicative systems, including the arrangement of mixed or alive indicative systems. To further explore these patterns, different groups of features were analyzed to observe their impact on the data organization.

Machine Learning Implementation: Due to a low representative systems sample size and high feature count we implemented the use of a Monte Carlo simulation to reduce overfitting via random sampling.

Accuracy and Metrics: 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). Additionally, proportion accurately identified, and evaluation of false positives and false negatives were also recorded.

Feature Importance: The voting classifier was iteratively run with each one of the 4 data types removed (ex. Raman features removed) to determine relative value of data combinations. This process was repeated with individual features iteratively removed (e.g. Ti elemental fraction removed).

Results: The overall accuracy of our voting classifier, with an AUC score of 0.79, demonstrates its effectiveness in distinguishing between indicative and non-indicative systems. Notably, while the voting classifier generally surpassed each individual algorithm in performance, the RF model showed a slightly superior accuracy (AUC = 0.81), particularly when utilizing all data types, followed by the KNN model (AUC = 0.80). In contrast, Machine SVM model had the lowest accuracy (AUC = 0.59). An informal evaluation involving ~10 subject matter experts further indicated that all algorithmic methods outperformed manual classification, underscoring the potential of machine learning approaches in this context.

References: [1] Warren-Rhodes et al. (2023) *Nat Astron.* doi:10.1038/s41550-022-01882-x [2] Cleaves, James II. (2023) *PNAS.* doi:10.1073/pnas.2307149120 [3] Sheng T. et al. (2024) (this conf.). [4] Gentry. et al. (2024) (this conf.).