Statistical Classification of Biosignature Information using Multiple Instrument Observations

Abdullah Shahid, North Carolina State University Raleigh, Raleigh, United States, Tao Sheng, University of Pittsburgh, Computational Biology, Pittsburgh, United States, Sunanda Sharma, Jet Propulsion Laboratory, Pasadena, United States and Diana Gentry, NASA Ames Research Center, Biospheric Science Branch, Moffett Field, United States

Abstract Text: The accurate identification of biosignatures (indications of life) from data taken from remote or in situ planetary exploration is one of the most important challenges in astrobiology, the interdisciplinary field examining habitability and the potential for extraterrestrial life. This study employs machine learning algorithms to optimize the identification of biosignatures, with an emphasis on those which are agnostic to a specific biochemical basis. We exploit the wealth of terrestrial data available from biogenic and abiogenic systems to enhance efficient feature prioritization. Our dataset, pulled from public databases and laboratory recorded measurements, includes elemental abundance, isotopic fractionation, and VNIR/Raman spectra The data curation process included standardization for detection limits and ranges. Subsequent feature extraction yielded detailed inputs for machine learning, including combinations of elemental content, isotopic ratios, and parameters of spectral peaks and troughs. Feature significance was evaluated across diverse machine learning methodologies, such as k-nearest neighbors, logistic regression, Random Forest, support vector machines, and Gaussian Naïve Bayes, along with a combined voting classifier. We utilized Receiver Operating Characteristic Area Under the Curve (ROC AUC) across 2,000 50% test-train splits as a robust metric of model performance. Results revealed a promising ROC AUC of 0.853 for the combined voting classifier. Removing elemental abundance data notably reduced model accuracy (13% decrease in AUC), highlighting its critical role in biosignature detection. Several other individual data features exhibited significance within their respective data types, offering additional granularity. This research fortifies the relevance of machine learning to astrobiology, potentially enhancing life detection missions by allowing algorithmic prioritization of high-interest samples for further investigation. Future work will refine data standardization, expand the dataset to include more terrestrial systems, and incorporate convolutional neural networks for spectral feature extraction. The potential for public data sharing is also under exploration, reinforcing our commitment to collective scientific advancement.