Algorithmic Detection of Elemental Biosignatures

National Aeronautics and Space Administration

Jesse Murray[1,2], Diana Gentry[3]

[1]NASA Internships, Fellowships & Scholarships, [2]University of Oxford, [3]NASA Ames Research Center

Introduction

Machine learning models that classify a planetary exploration sample as **non-indicative** or **indicative** of life can **play an important role in planning life-detection missions:**

1. They are based on clearly defined and consistent algorithms, regardless of sample type or origin.
2. They can reveal distinguishing features of life and suggest important measurements in a future mission.
3. They can be used to understand how combinations of different biosignatures affect overall confidence.

The need for this last capability was identified as a key gap in The Ladder of Life Detection (Neveu 2018).

Data Collection

- We selected **elemental composition** due to is availability across diverse samples measured in published literature.
- Selected samples had to meet these criteria:
    1. Clearly **non-indicative** or **indicative** of life.
    2. Analogous to a theoretical mission sample.
    3. Completely characterized by the elemental data.
- X-ray diffraction, mass spectrometry, etc. measurements were standardized to a simulated limit of detection.
- **Four clusters of samples** emerged in the principal component analysis (PCA) (**Fig. 1**) and boxplots (**Fig. 2**).

Sample Type	Number	Examples
Non-indicative	35	Lunar rock, basalt
Indicative mixed	19	Seawater, crop soil
Indicative non-alive	46	Coal, chalk, fossil
Indicative alive	10	Biofilm, bacteria

Modeling

Approach

- Classify a sample as **non-indicative** or **indicative** of life from its elemental composition.
- Apply a variety of common statistical models, as consensus among the models lends confidence.
- Use the Python scikit-learn software.

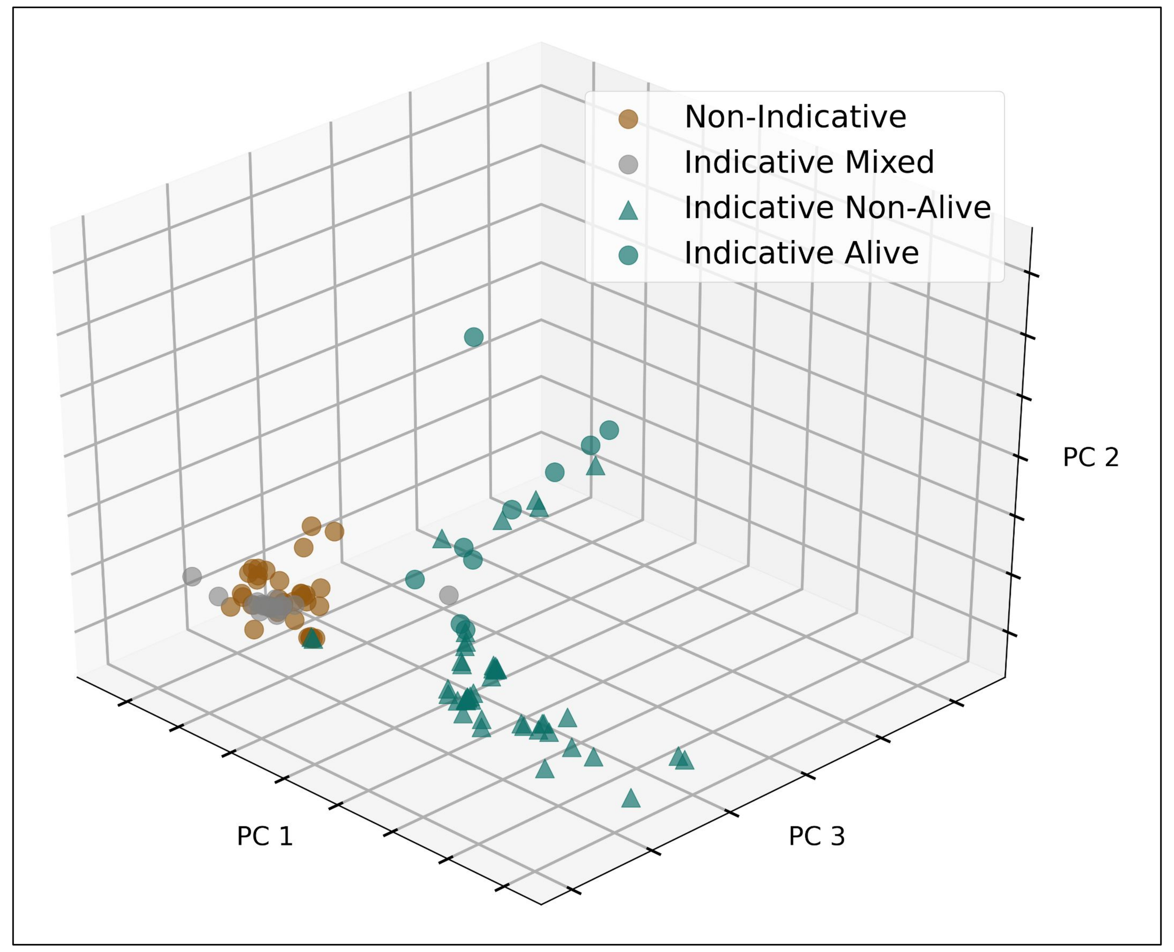

Figure 1: *PCA (reduces multi-dimensional elemental composition data into the most variable components), used for **KNN**.*

Model selection

These four classification algorithms offer different ways of making predictions.

- K-nearest neighbors (**KNN**)
- Logistic regression (**LR**)
- Linear support vector machines (**SVM**)
- Gaussian naïve Bayes (**GNB**)

Model training and testing

- The models were **trained** and **tested** on random splits of the data (40:60 splits, 1,000 each).

Figure 2: *Boxplots showing the abundance of elements in the four sample types.*

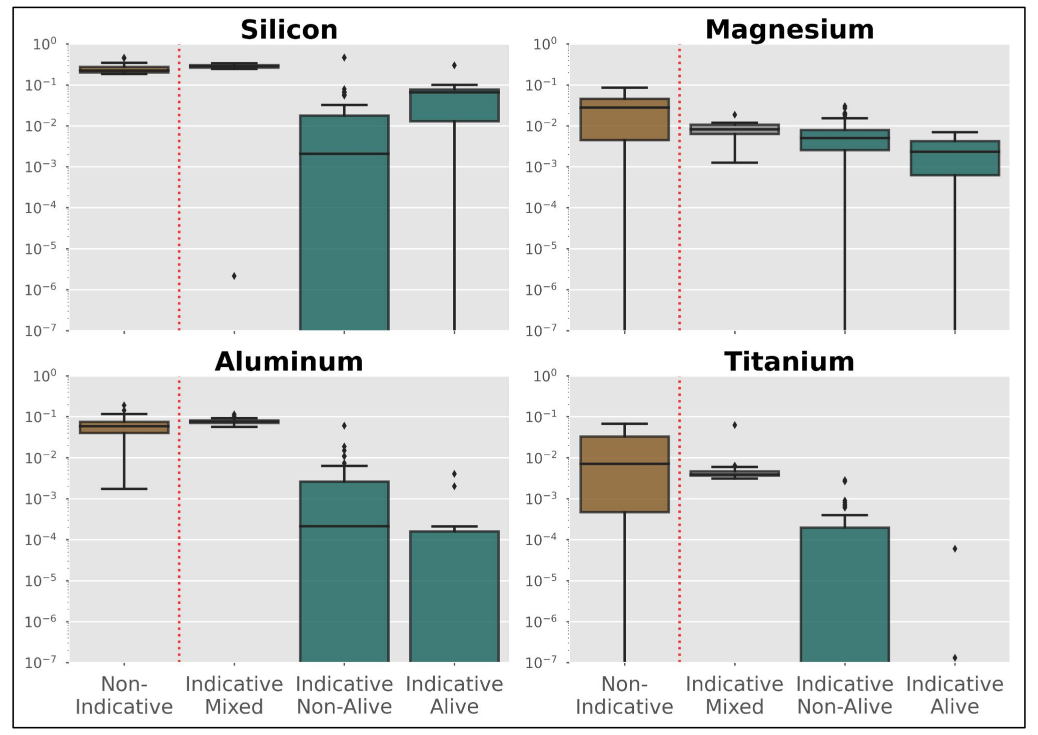

Figure 2a: *Elements found to be predictive of a **non-indicative of life** sample*

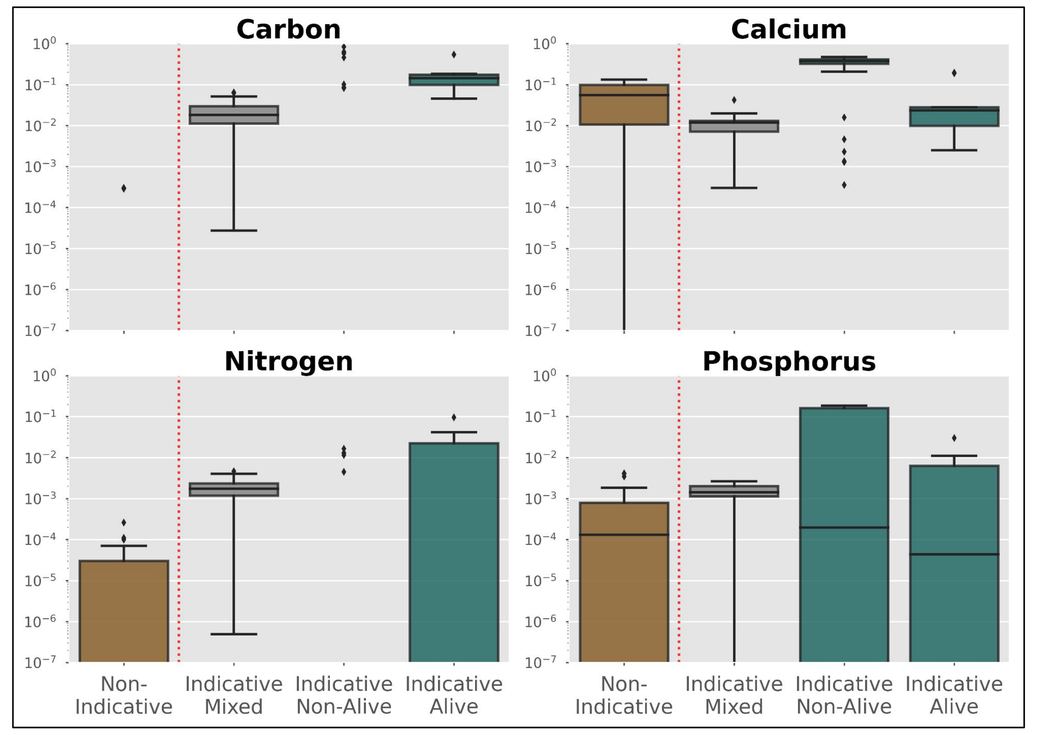

Figure 2b: *Some elements found to be predictive of an **indicative of life** sample.*

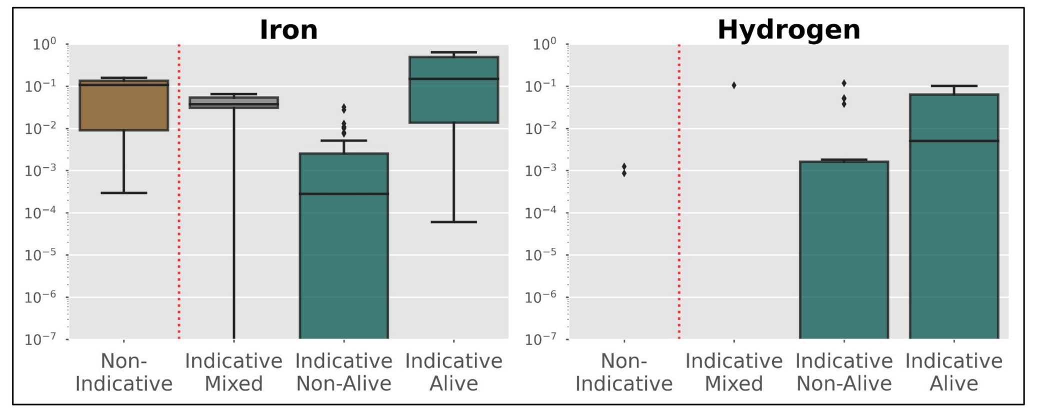

Figure 2c: *Some elements with varied predictions.*

Results

Elements predictive of a non-indicative of life sample

- All models found Si abundance to be a strong predictor.
- Most models found Mg, Al, and Ti as moderate predictors.

Elements predictive of an indicative of life sample

- All models found C and Ca as strong, and Cl as moderate.
- Most models found N, K, and P as moderate.

Elements with varied prediction directions

- Fe (slightly non-indicative), H (slightly indicative), O (varied widely), Na, Mn, and S.
- Some of these elements were not present in enough samples to be important predictors.

Model Performance

- Mean accuracy scores were similar, averaging 88% ± 4%.
- More false positives than false negatives (preferred tradeoff).

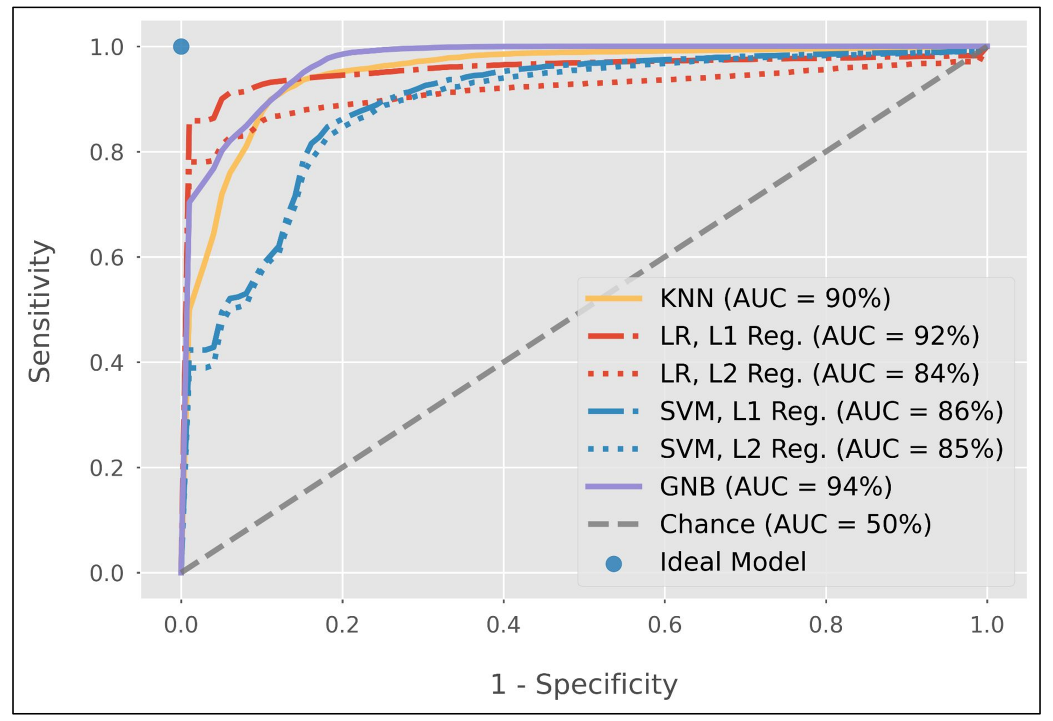

Figure 3: *Receiver Operator Characteristics (**LR** and **GNB** are shown to be the best performing models).*

Future Work

- Expand data to include other types of data, e.g. isotope fractionation, free energy, spectral information, etc.
- Implement non-linear models, e.g. neural networks.

Acknowledgements

Aivaras Vilutis for early data collection. T. Stucky, M. Furlong, J. Koehne, D. Mauro, and A. Schramm project conception in 2018 NASA APEX program. New Jersey Space Grant Consortium for stipend funding. ARC intern coordinators, especially Abel Morelos.