# Algorithmic Detection of Elemental Biosignatures

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

## Introduction

Machine learning models that classify a sample as ***indicative*** or ***non-indicative*** of life could play an important role in life-detection missions.

- Their predictions add redundancy to judgements based on human expertise.
- Their important features can reveal the most informative measurements within the operational constraints of a life-detection mission.

The Ladder of Life Detection (Neveu 2018) identifies a need to understand how combinations of biosignatures affect overall confidence. The present work provides a starting point to answer this need.

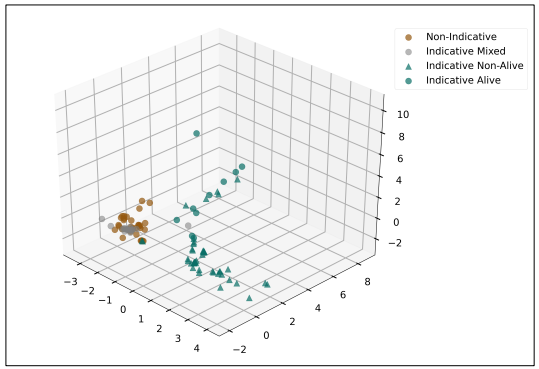


**Figure 1:** Principal Component Analysis reveals separation between indicative and non-indicative samples

## Methods

### Data Collection

Elemental abundance was chosen as a set of features due to its availability in diverse sample types:

- 35 **non-indicative**, e.g., lunar rock, basalt.
- 19 **indicative mixed**, e.g., seawater, crop soil.
- 46 **indicative non-alive**, e.g., coal, chalk.
- 10 **indicative alive**, e.g., biofilm, bacteria.

### Standardization

The samples were standardized to the same limit of detection of a simulated mission scenario (1,500 ppm).

### Modeling

Four classification models were used: k-nearest neighbors (**KNN**), logistic regression (**LR**), linear support vector machines (**SVM**), and Gaussian naïve Bayes (**GNB**).

### Training and Validation

The performances and feature importances of the six model variants were assessed with Monte Carlo simulations on 40:60 train to validation ratios.

### Feature Importances

To obtain feature importances, KNN was run on three principal components of the training data and LR and SVM were run with **L1 and L2 regularization**.

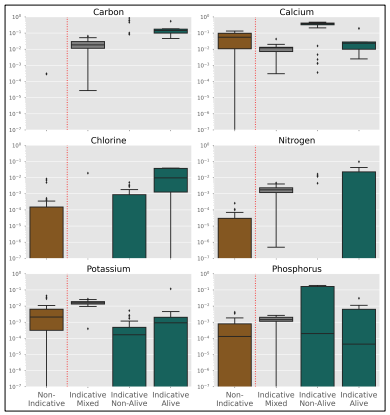


**Figure 2:** Abundances of **indicative** elements

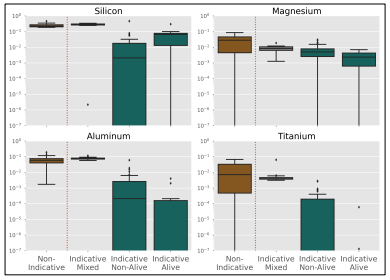


**Figure 3:** Abundances of **non-indicative** elements

## Results

### Indicative of Life Elements

- All models had C and Ca as strong, and Cl as medium.
- Most models had N, K, and P as medium.

### Non-Indicative of Life Elements

- All models had Si as strong.
- Most models had Mg, Al, and Ti as medium.

### Varied Elements

- Fe (slightly non-indicative), H (slightly indicative), O (varied widely), Na, Mn, and S.

Performance scores ranged from **82% - 94%**.

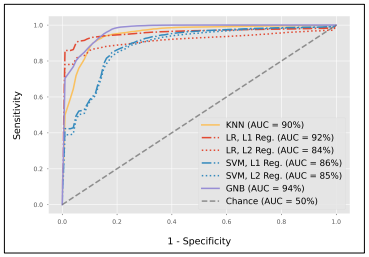


**Figure 4:** Receiver Operator Characteristics

## Future Work

1. Expand the data types to obtain even more predictive combinations of features, e.g. isotope fractionation.
2. Implement non-linear models, e.g. neural networks.

## Acknowledgements

- Aivaras Vilutis for early data collection.
- New Jersey Space Grant Consortium for stipend funding.