Machine Learning for Metal Identification in Water Samples using Laser-Induced Breakdown Spectroscopy (LIBS)

J. Gross, E. Orme, M. Fernandez, A. Carter, R. Silverstein, C. Ochatt, S. Pozsonyiova, L. Felipe, P.K. Diwakar

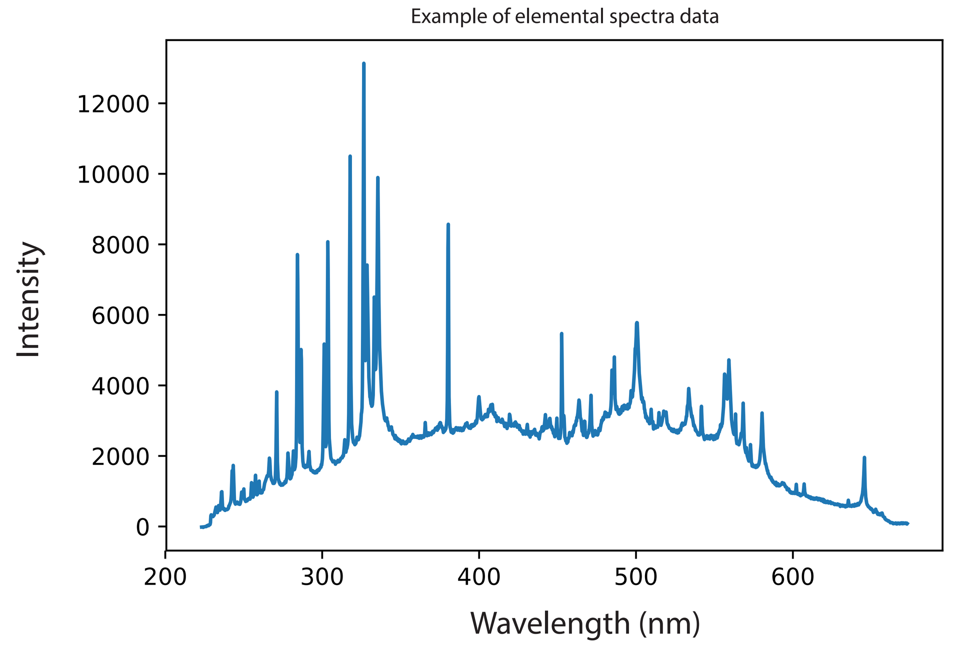
South Dakota School of Mines & Technology

Ransom Everglades School

Background

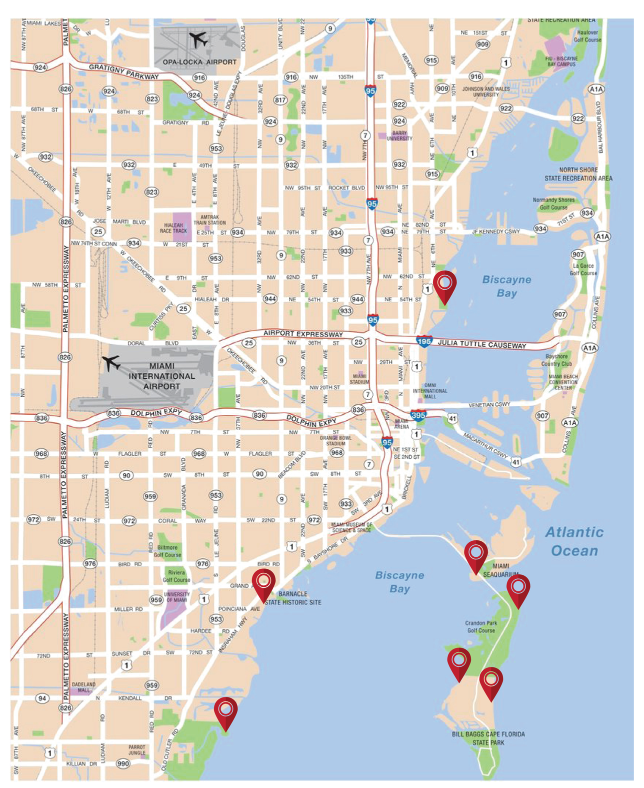
- Laser-induced breakdown spectroscopy (LIBS) is a multi-elemental analytical technique used for processing a variety of complex samples including solids, liquids, gases, and aerosols.
- Scans retrieve intricate patterns of spectral emissions.
- Their interpretation is time-consuming and challenging because the data is high dimensional and complex (example graph of data below).
- We needed a way to analyze LIBS data quickly, efficiently, and accurately.

Example of elemental spectra data



Objectives

- To develop a trained machine learning model capable of analyzing data collected by LIBS with the goal of identifying the presence of heavy metals in Biscayne Bay water samples.
- To streamline a way that increases accuracy and reduces time needed for LIBS data analysis.



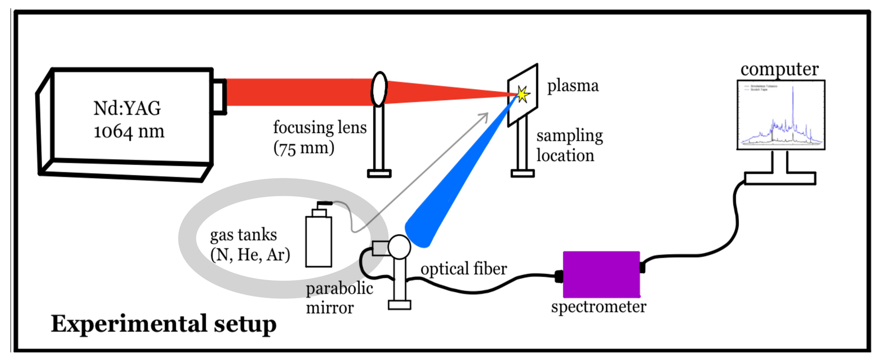
Approach

Brainstorm → Data Collection → Data Pre-Processing → Training/Testing → Validation → Application

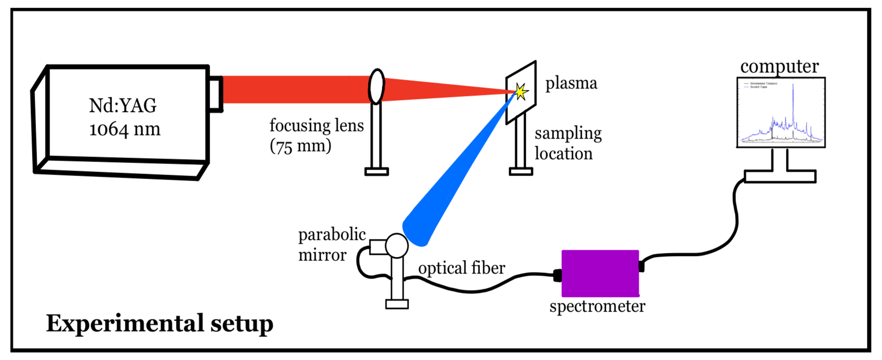
Experimental Setup

- Data was collected in both solid and liquid form using a pulsed laser.

Experimental Setup with Gasses

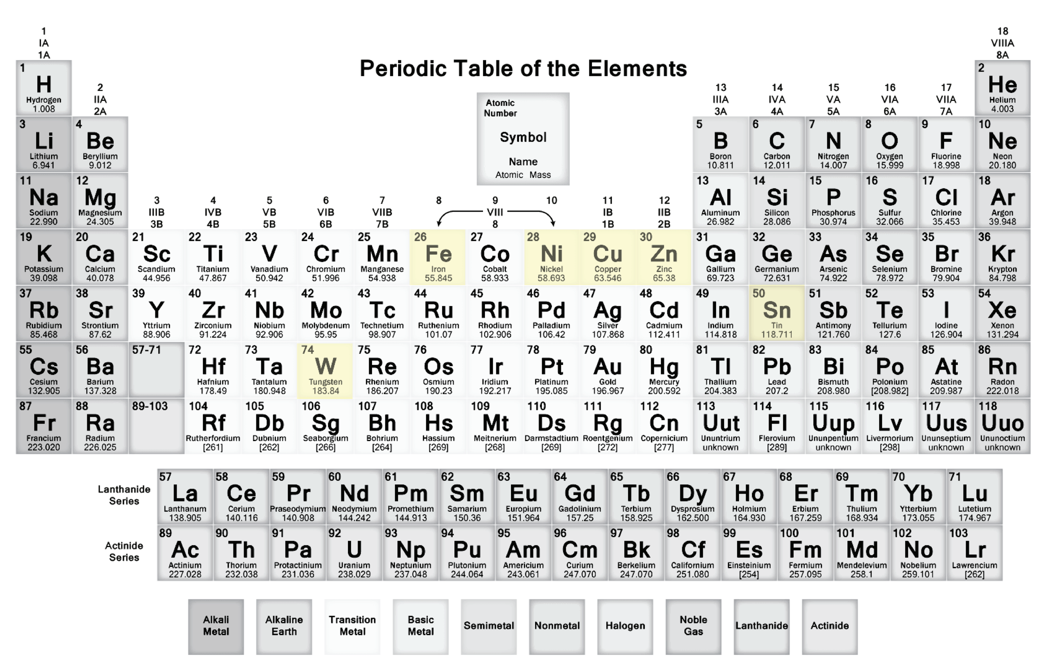


Experimental Setup without Gasses



Data Collection

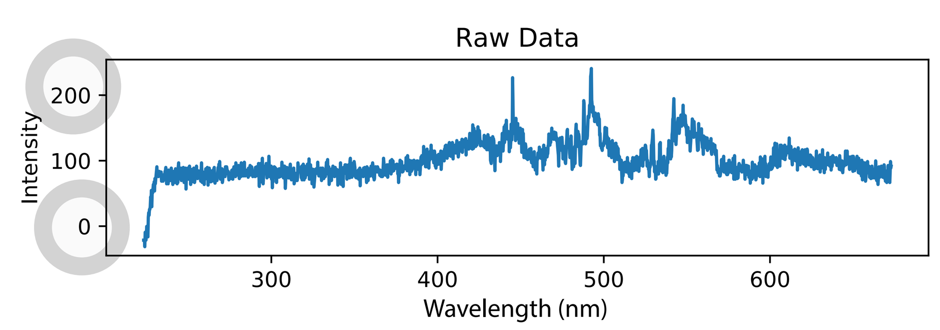
- In this study, training data was collected from six different metals.
- The data collected was used to train, test, and validate the model.

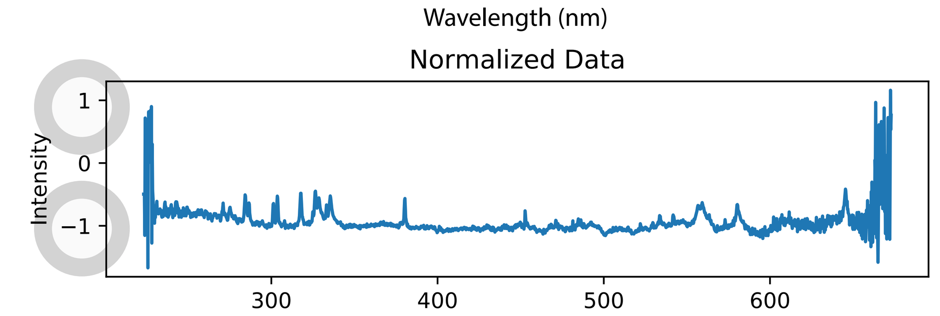


Cu Fe Ni Sn W Zn

Data Pre-Processing

- The data was normalized so that niether the intensity nor the state of matter would affect the model's training and prediction.
- Normalization of data allows the machine learning algorithms to evaluate the data on the same scale and increases model accuracy as well as computational complexity.

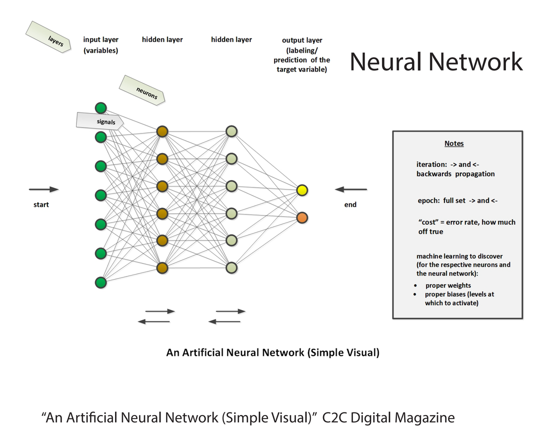




Training

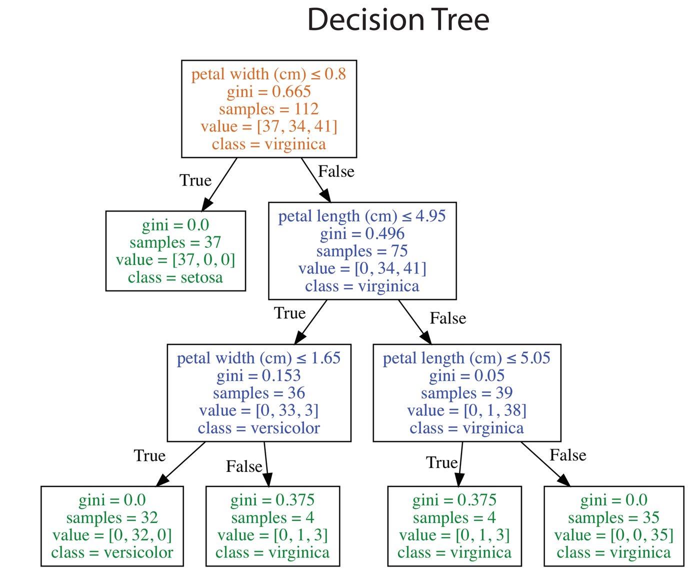
- Seven classifiers were trained for each element and the best performing one was selected:
  1) K - Nearest Neighbors
  2) Decision Tree
  3) Random Forest
  4) Multi-Layer Perceptron (Neural Network)
  5) Ada Boosted Decision Trees
  6) Gaussian Naïve Bayes
  7) Support Vector Machine

Neural Network



"An Artificial Neural Network (Simple Visual)" C2C Digital Magazine

Decision Tree



Testing

- To train the model, 60% of data was used for training and 40% for testing with the 60/40 split assigned randomly.
- Once the model was trained with 60% of the data, the remaining 40% was used to make the predictions upon which the model is evaluated.
- Accuracy$_i$ = total predicted correct / total predictions
- This process occured three times and the accuracies were averaged to get the total accuracy. This process is known as three-fold cross validation.

Cross Validation - Fold 1

Partition 1 Training Data	Partition 2 Training Data	Partition 3 Testing Data

Cross Validation - Fold 2

Partition 1 Training Data	Partition 2 Testing Data	Partition 3 Training Data

Cross Validation - Fold 3

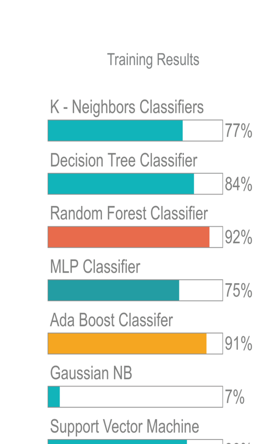
Partition 1 Testing Data	Partition 2 Training Data	Partition 3 Training Data

Model Evalutation

- Total Accuracy of a model = Product of each of the six model's accuracy

$$A_{tot} = \prod_{i=0}^{n} A_i$$

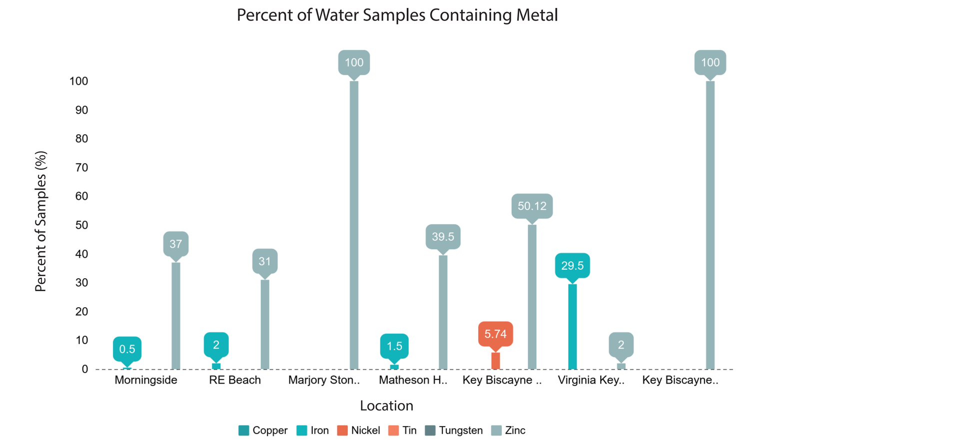
- Random Forests produced the best overall results



Application with Biscayne Bay Water Samples

- The trained random forest model was used to identify the metals present at seven coastal locations.
  Below are the obtained results:

Percent of Water Samples Containing Metal



Percent of Water Samples Containing Metal

Location	Copper	Iron	Nickel	Tin	Tungsten	Zinc
Morningside	0.0	0.5	0.00	0.0	0.0	37.00
RE black pebble beach	0.0	2.0	0.00	0.0	0.0	31.00
Marjory Stoneman Biscayne Nature Center	0.0	0.0	0.00	0.0	0.0	100.00
Matheson Hammock	0.0	1.5	0.00	0.0	0.0	39.50
Key Biscayne Yacht Club	0.0	0.0	5.74	0.0	0.0	50.12
Virginia Key RSMAS	0.0	29.5	0.00	0.0	0.0	2.00
Key Biscayne Beach Club	0.0	0.0	0.00	0.0	0.0	100.00

Conclusions

- Machine learning (ML) provides a faster, more efficient method of analyzing data collected through LIBS.
- In this study, we reduced the time needed while increasing the accuracy of LIBS data Analysis.

Typical LIBS Data Analysis	ML for LIBS Data Analysis
Accurate	Highly Accurate
Inefficient	Efficient
Time Consuming	Fast
Complex	Simple

Contact Info

Email: 20jgross@ransomeverglades.org
Linkedin: https://www.linkedin.com/in/joseph-gross-b141aa143
Github: Joseph-Gross

QR Codes

Github Repository

Abstract