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

J. Gross[1], *E. Orme*[1&3], *M. Fernandez*[1,4], *A. Carter*[2], *R. Silverstein*[2], *C. Ochatt*[1], *S. Pozsonyiova*[5], *L. Felipe*[1], *P.K. Diwakar*[6a]

1. *Ransom Everglades School, Coconut Grove, Miami, FL 33133, USA*
2. *Miami Waterkeeper, Miami, FL 33145, USA*
3. *Augustana University, Sioux Falls, SD, 57197, USA*
4. *Florida International University – College of Engineering and Computing, Miami, FL 33174*
5. *Macalester College, St. Paul, MN, 55105, USA*
6. *South Dakota School of Mines and Technology, Rapid City, SD, 57701, USA*

a. *Corresponding author*

Laser-induced breakdown spectroscopy (LIBS) is a multi-elemental analytical technique used for detecting a variety of complex samples including solids, liquids, gases, and aerosols. The approach uses a pulsed laser and ablates the samples while producing plasma, resulting in a sample's elemental spectra. However, the scans retrieve intricate patterns of spectral emission that make their interpretation time-consuming and challenging. The key to overcoming this challenge can be achieved by analyzing LIBS data using supervised machine learning models that can acquire input variables (x) and output variables (y) to create a mapping function from input to output. With sufficient training data, these models can learn from the data, identify patterns, to eventually make predictions without excessive programming. In this study, data collected from six different metals in both liquid and solid form - Cu, Fe, Ni, Sn, W, and Zn - were used for training and testing. The data was normalized so neither the intensity nor the state of matter would affect the model's training and prediction. LIBS spectral data was analyzed using classification methods including k-nearest neighbors (K-NN), random forests, support vector machines (SVM), neural networks (MLP), Gaussian naïve Bayes (GNB) and gradient boosted decision trees. Models were used to identify metals present in data taken from Biscayne

Bay coastal sites through Miami Waterkeeper. The algorithm developed compared different binary classifiers for each of the six metals and selected the best performing ones, thus producing six binary results for each water sample. The union of these results were then used to identify the metals present in water samples from coastal areas with around 92% accuracy. Although overfitting is possible, the use of 3-fold cross-validation to determine accuracy tested with random samples shows that it is unlikely. Prospects of training the machine learning models with labeled data allowed for simplifying needs of standards for each target element, increased accuracy, and reduced time needed for LIBS data analysis. Our results suggest that machine learning algorithms show potential to improve LIBS data interpretation with ML-driven predictability.