Random Forest Classification for Identifying Bacteria with Mass Spectrometry in Mixed Samples

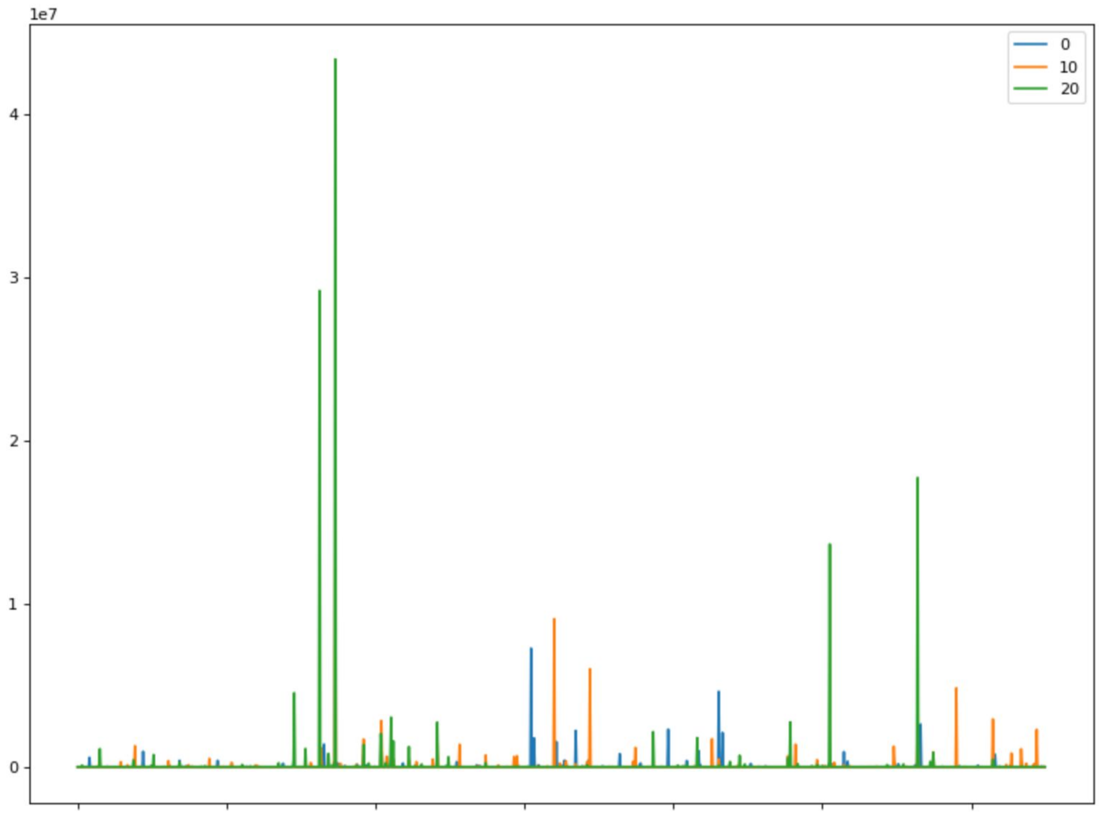
David Gray

July 1, 2019

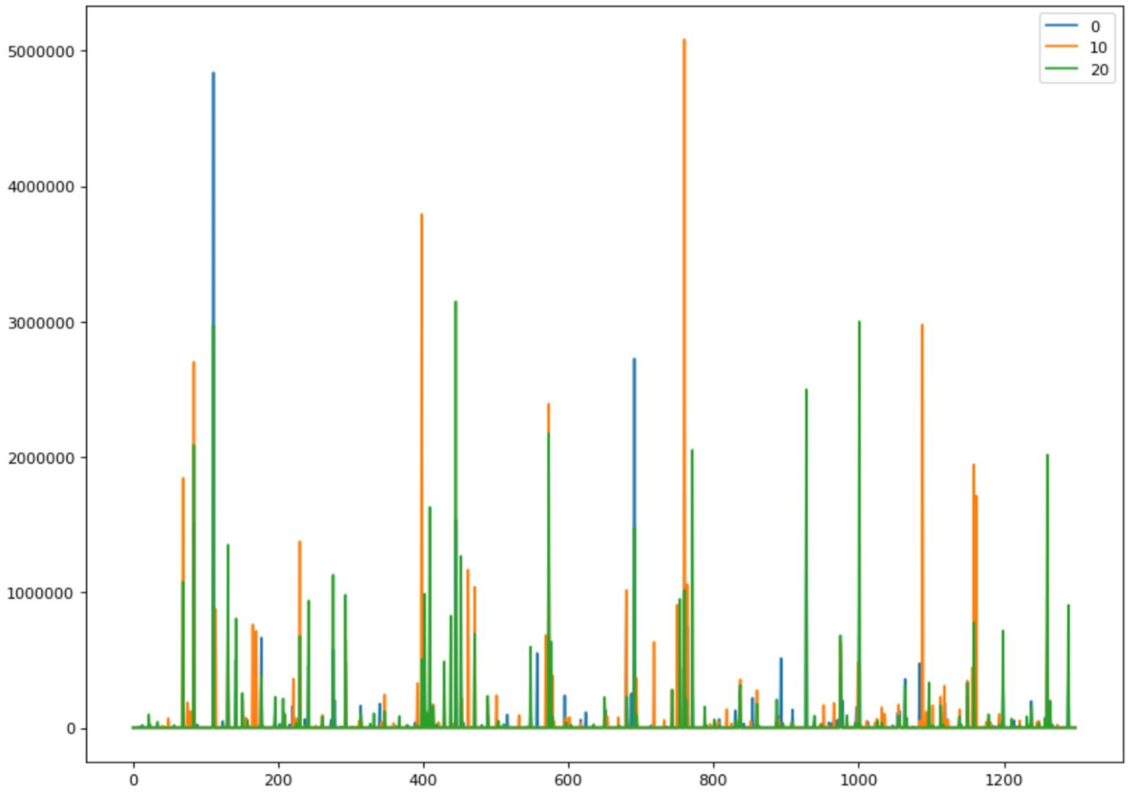
The dataset used in this study is described at the UCI machine learning repository as “a dataset to explore machine learning approaches for the identification of microorganisms from mass-spectrometry data.” The prerequisite for matrix-assisted laser desorption/ionization time-of-flight mass spectrometry for the purpose of identifying bacterial species is an isolated colony of the microorganism. The problem is that MS-based microbial identification is not feasible routinely. One of the issues is that polymicrobial samples can will yield a ‘mixed’ MS fingerprint. The purpose of this work is to use machine learning to try to decipher identity from samples with two species of bacteria.

Information about the MALDI-TOF mass-spectrometry data set analyzed in this project was provided in a read-me text. The dataset includes 931 MALDI-TOF mass spectra. This includes 571 spectra of pure, individual samples as a reference. The “mixed dataset” consists of 360 spectra. So far, I have not found any need to perform cleaning steps, address missing values, or handle outliers.

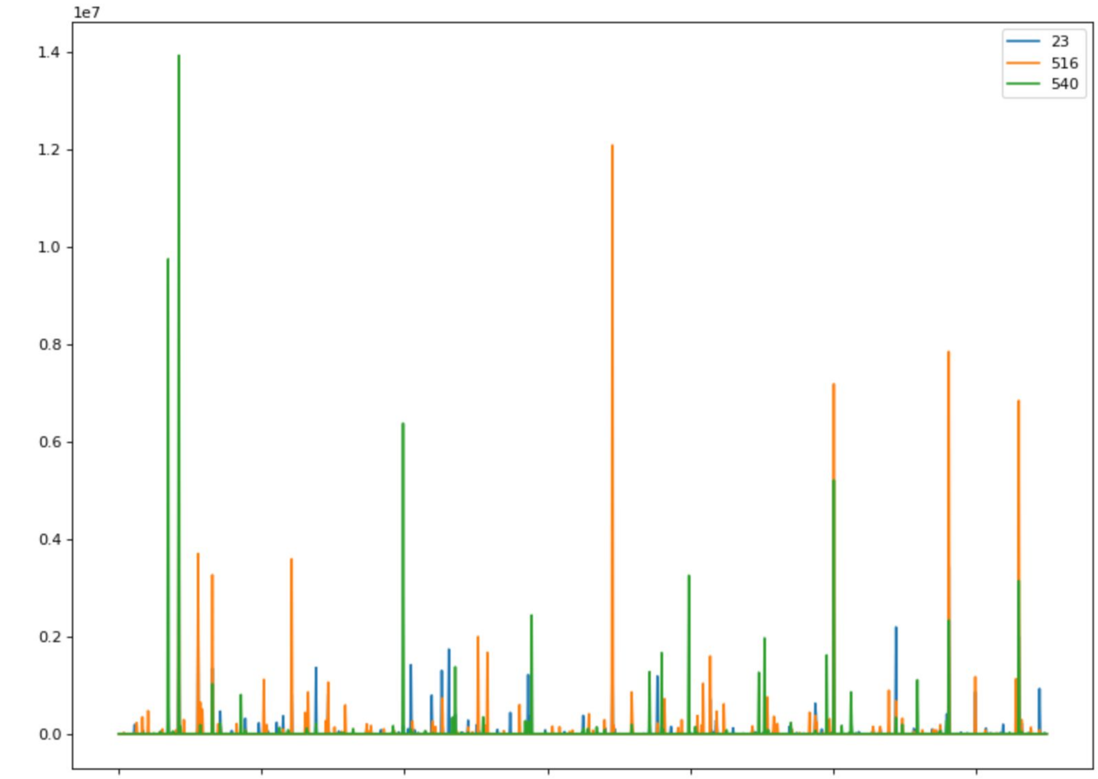
The initial findings are that there is some overlap between the pure spectra and the mixed spectra. Also, the pure spectra and mixed spectra are unique depending on the samples.



Pure Spectra of Samples 0, 10, and 20



Mixed Spectra of Samples 0, 10 and 20



Mixed (23) and Pure (516 and 540) Spectra. There is overlap among all three of these samples

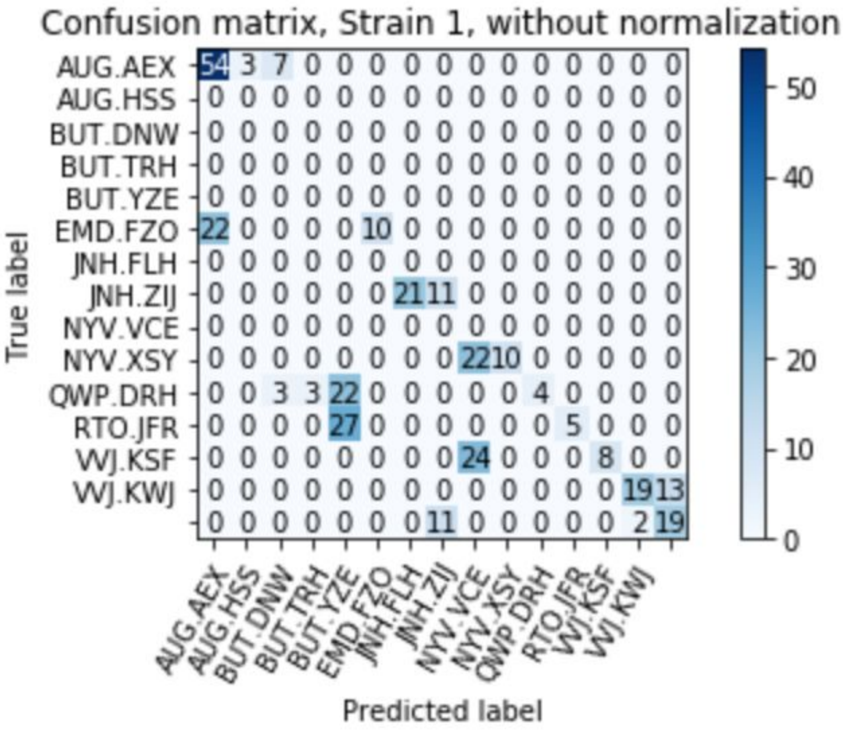
A random forest classifier was used to identify one of the strains of the mixture of two bacteria. The number of estimators was one hundred with a max depth of “None.” There are cases where the ratio of samples was zero to one (i.e., only one type of bacteria was present in the sample. When naively identifying the sample (not taking into account that one of the samples was not present), the accuracy was:

Accuracy of Predicting Strain 1 (Naive)	39.2%
Accuracy of Predicting Strain 2 (Naive)	58.1%

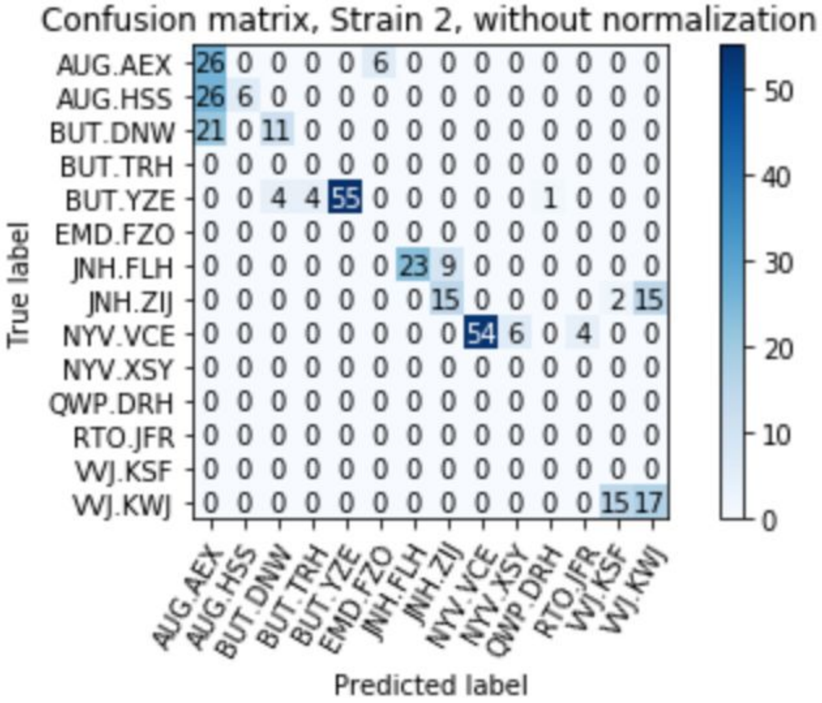
Taking into account the zero proportion of some samples, the accuracy improved:

Accuracy of Predicting Strain 1	43.8%
Accuracy of Predicting Strain 2	64.7%

Here are two heat maps demonstrating where the matches were made:



Confusion Matrix for Predicting Strain 1



	AUG.AEX	AUG.HSS	BUT.DNW	BUT.TRH	BUT.YZE	EMD.FZO	JNH.FLH	JNH.ZIJ	NYV.VCE	NYV.XSY	QWP.DRH	RTO.JFR	VWJ.KSF	VWJ.KWJ
AUG.AEX	26	0	0	0	0	6	0	0	0	0	0	0	0	0
AUG.HSS	26	6	0	0	0	0	0	0	0	0	0	0	0	0
BUT.DNW	21	0	11	0	0	0	0	0	0	0	0	0	0	0
BUT.TRH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BUT.YZE	0	0	4	4	55	0	0	0	0	0	1	0	0	0
EMD.FZO	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JNH.FLH	0	0	0	0	0	0	23	9	0	0	0	0	0	0
JNH.ZIJ	0	0	0	0	0	0	0	15	0	0	0	0	2	15
NYV.VCE	0	0	0	0	0	0	0	0	54	6	0	4	0	0
NYV.XSY	0	0	0	0	0	0	0	0	0	0	0	0	0	0
QWP.DRH	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RTO.JFR	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VWJ.KSF	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VWJ.KWJ	0	0	0	0	0	0	0	0	0	0	0	0	15	17

Confusion Matrix for Predicting Strain 2

This analysis shows that there are a few strains that are particularly well predicted especially from Strain 2 samples.

The authors of the paper pioneering analysis on this data and simulated data reported: "Few spectra were misidentified and mixtures were always at least partially identified. More than 60% of the mixtures were detected and correctly identified."

Our model would need to be modified to try to predict both bacteria types in the mixed samples as there is only one species in this work predicted per mixed sample.

It will be interesting to compare the samples correctly predicted that are in a higher or lower dilution. Comparing the average correctly predicted with different techniques or parameters will help refine the method. It would be interesting to look at other statistics to evaluate the model beyond accuracy.