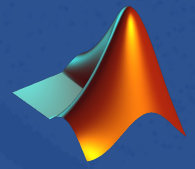
Compressing RF Signals for Rapid Machine Learning Development

Isaac Rodriguez-J

SONOMA STATE UNIVERSITY

Abstract



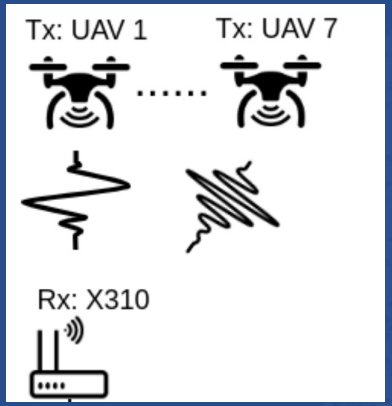
The purpose of this project is to research methods to reduce the dimensionality of the data and still correctly differentiate between 7 extremely similar devices based on radio frequency recordings. The data was taken from https://genesys-lab.org/hovering-uavs. They did their own research using the raw IQ data with preprocessing and data augmentation techniques followed by a convolutional neural network.

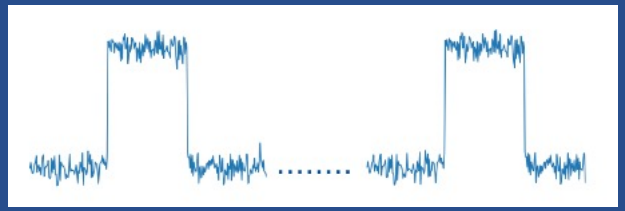
SONOMA
STATE
UNIVERSITY

Unmanned Aerial Vehicle (UAVs)

Study done by Northeastern University

Their Objective:
To independently classify which UAV transmitted to the receiver based on short bursts of received signals.





SONOMA STATE UNIVERSITY

Image Source:

https://www.dji.com/matrice100/features
https://genesys-lab.org/hovering-uavs

sigMF is a standardized way to record data of various signals

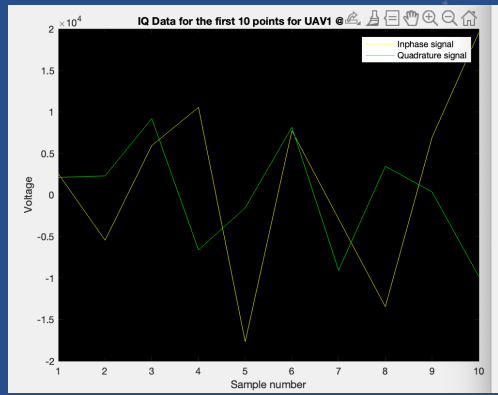
The Dataset

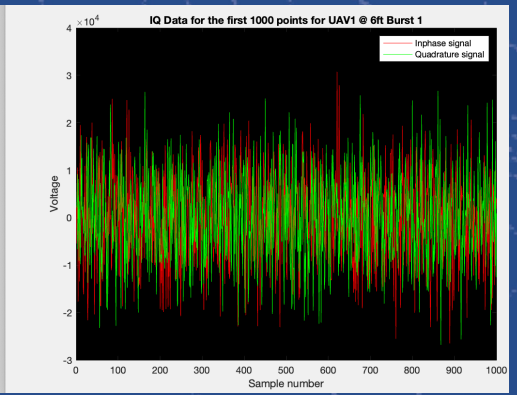
13,000 RF signals with an Average F_{center} 2.4GHz and BW 10MHz

Each RF recording is 2 seconds long with ~97,000 samples @ 0.8MB each.

Each recording is saved as a chain of 97,000 complex numbers.

```
2524+2104i    -5460+2284i    5944+9184i    10536-6628i    -17680-1566i
```





SONOMA STATE UNIVERSITY

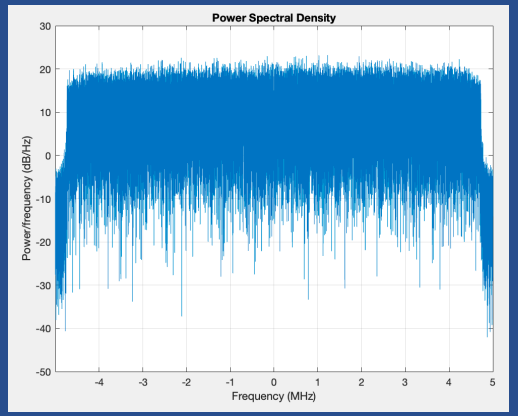
Complex IQ to RF Signal

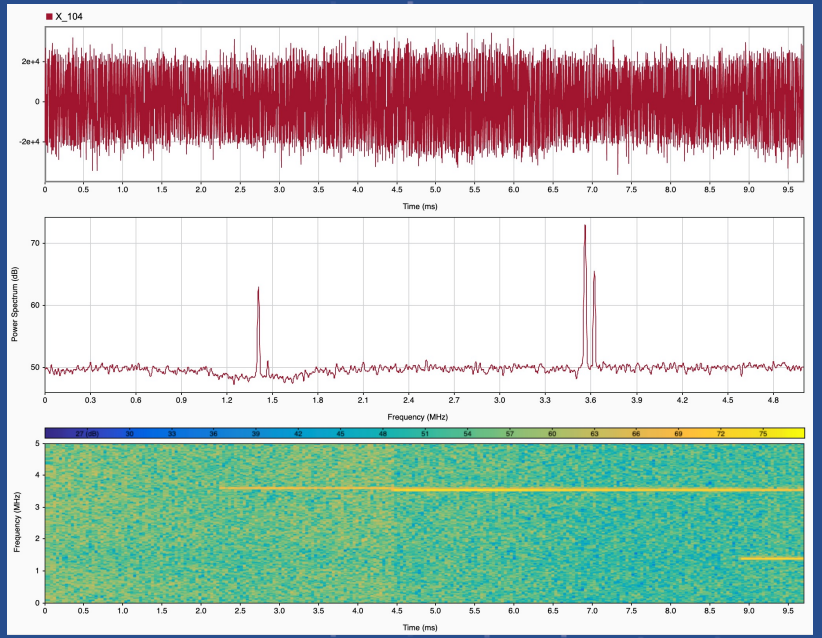
- I value
- Q value
- Center Frequency
- Sampling Frequency

where, $t = 0{:}\frac{1}{F_s}{:}\frac{1}{F_s} * 97{,}000$

$$\sqrt{I^2 + Q^2} * \cos(2\pi F_c t + atan(\frac{Q}{I}))$$

SONOMA STATE UNIVERSITY





What is Radio Frequency Fingerprinting?

The use of machine learning techniques to identify a device that transmits RF signals based on signal characteristics. These characteristics may stem from nonlinearities in the transmitting circuits and present operating stages.

SONOMA STATE UNIVERSITY

Signals are too heavy to work with...

Implement a way to reduce signal's impact on storage.

A proven method, is to use a spectrogram along with a convolutional neural network. The issue is that the images along with the implementation of a CNN will also take up a considerable amount of system resources.

My solution was to extract key features from each signal and uniquely identify which UAV is transmitting with considerably less values per data point.

SONOMA STATE UNIVERSITY

Key Features Extracted per Signal

Time Domain

- Mean
- Shape Factor
- Crest Factor
- Clearance Factor
- SNR
- Standard Deviation
- Peak Value
- Impulse Factor

Frequency Domain

- Mean Frequency
- Median Frequency
- Occupied Bandwidth
- Power Bandwidth
- Band Power

SONOMA STATE UNIVERSITY

Key Features Explained

Occupied Bandwidth:
measure of the bandwidth containing 99 % of the total integrated power for transmitted spectrum and is centered on the assigned channel frequency.

Shape Factor:
Shape factor is dependent on the signal shape while being independent of the signal dimensions.

Impulse Factor:
Compare the height of a peak to the mean level of the signal.

Crest Factor:
Indicates how extreme the peaks are in a waveform

Band Power:
Average power in the signal.

Clearance Factor:
Peak value divided by the squared mean value of the square roots of the absolute amplitudes.

SONOMA STATE UNIVERSITY

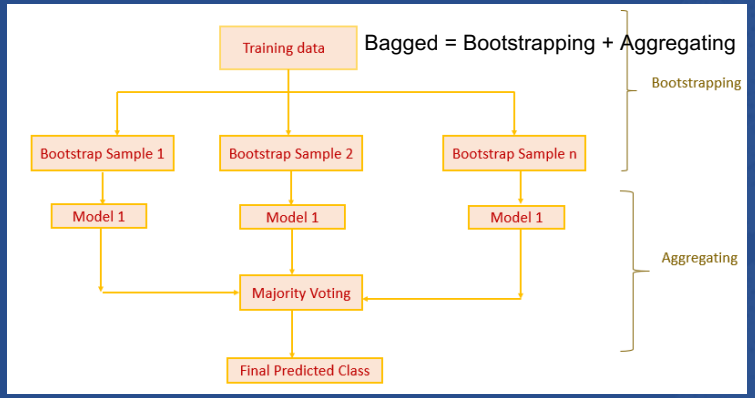
Results

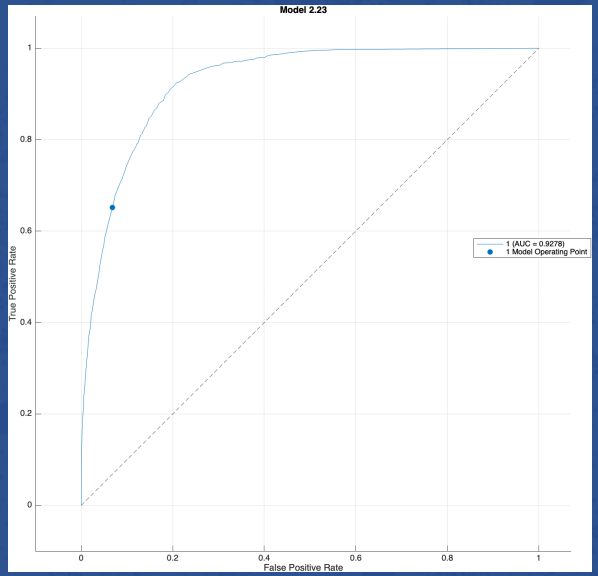
Total dataset storage reduction from 10GB to 1.2 MB. (97000 to 13 points)

Having this dataset loaded on MATLAB has no impact on performance.

The machine learning model used to train on this dataset was an ensemble of bagged decision trees. Total training time was 69 seconds, with a validation accuracy of 80.1%.

A Boostrap Sample is a small random sample from the original dataset. A single decision tree trains on this data and provides its own predictions based on its given data. Another decision tree will be given another random subset of data. All the decision trees have a majority vote on classifying the UAV.

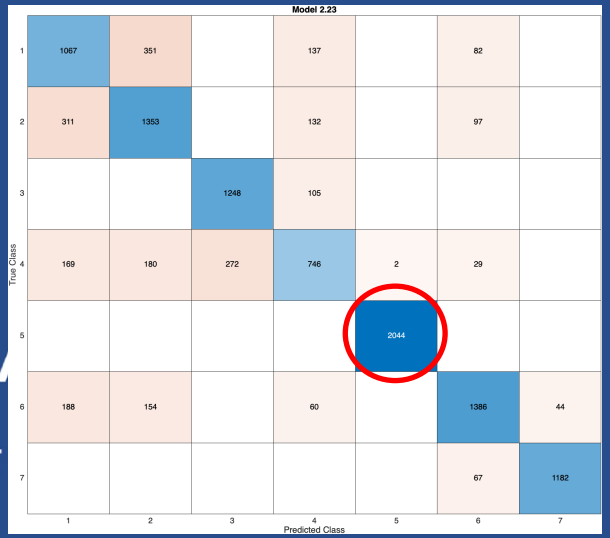


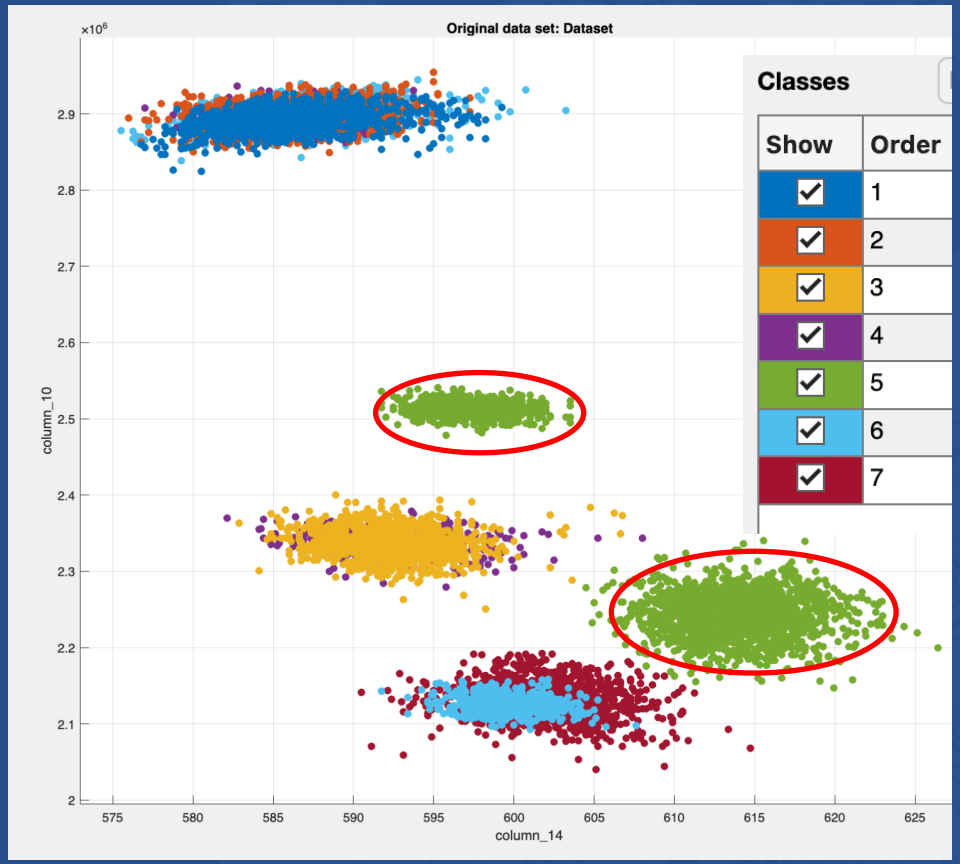


SONOMA STATE UNIVERSITY

Results

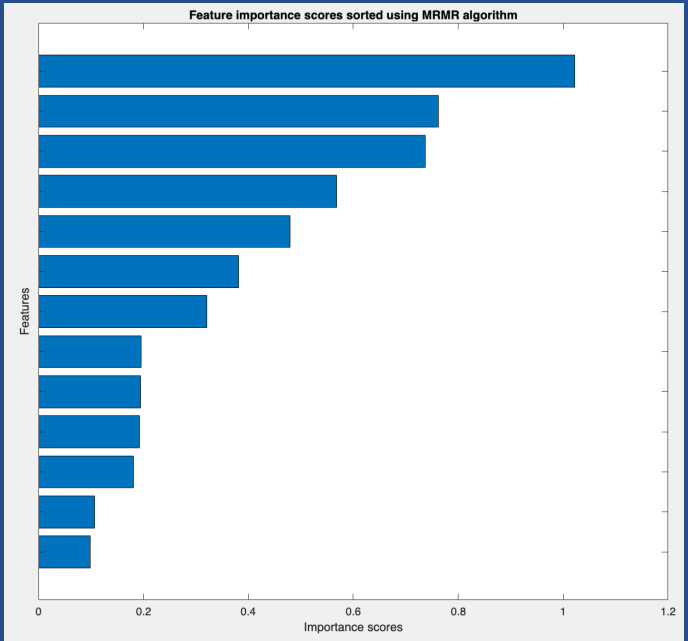
Looking at the data, we can see why UAV5 was the easiest to classify. Its' data does not overlap with other UAV's. Its also easy to see that UAV 1 and 2 overlap, UAV 3 and 4 overlap, And UAV 6 and 7 overlap. This means that there is the possibility to engineer new features that could differentiate between UAV's.





SONOMA STATE UNIVERSITY

Results



Here we have an issue. Typically importance scores are between 1 and 0. Our most influential feature has a score of 1.02. Not only that but half of the other features have abnormally high importance scores. This could mean a few things.

- Not enough features.
- Too much overlapping data.
- Features are too similar.

In any case, we will explore another method that uses more features.

SONOMA STATE UNIVERSITY

Using Audio Signal Techniques

The techniques implemented by the researchers on the bottom right are commonly used in audio signals. In their case the power spectral density(2 class, 100% accuracy), Mel-Frequency Cepstral Coefficients(MFCC) (4 class, 98.67% accuracy) and Linear Frequency Cepstral Coefficients (LFCC)(10 class, 95.15% accuracy). Their research and dataset are similar, but they used UAVs from different manufactures for their study.

In this case, we will focus on the Linear Frequency Cepstral Coefficients. LFCC is very similar to MFCC. MFCC focuses on the audio spectrum, with more emphasis on the lower frequencies. Similarly, LFCC focuses on the audio spectrum but has equal preference over the whole spectrum.

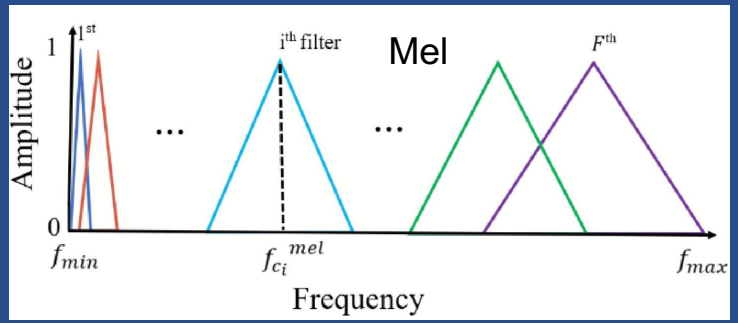
In the research below, they extracted the cepstral coefficients and did further processing. In our case, we extract the coefficients and consolidate them into one feature per data point.

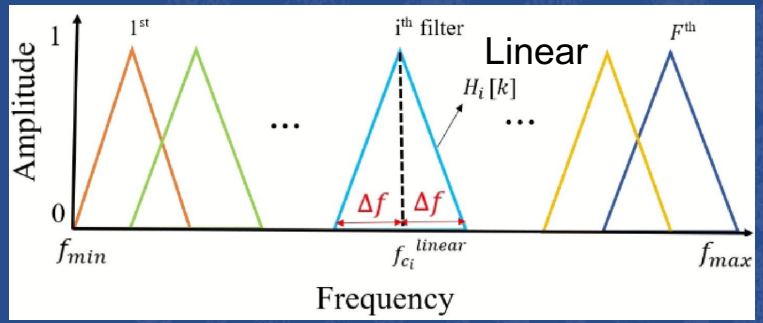
SONOMA STATE UNIVERSITY

Rabiye Kılıç, Nida Kumbasar, Emin Argun Oral, Ibrahim Yucel Ozbek, Drone classification using RF signal based spectral features, Engineering Science and Technology, an International Journal, Volume 28,2022, 101028, ISSN 2215-0986, https://doi.org/10.1016/j.jestch.2021.06.008. (https://www.sciencedirect.com/science/article/pii/S2215098621001403)

Linear Frequency Cepstrum

Based on the Mel-frequency cepstrum (MFC), the LFC are derived from a linear scale. The MFC can be seen as a short-term power spectrum for a sound signal (typically human speech). The MFC gathers more data at lower frequencies and less at higher in the audible spectrum. The linear frequency cepstrum gathers data evenly through out the given spectrum.



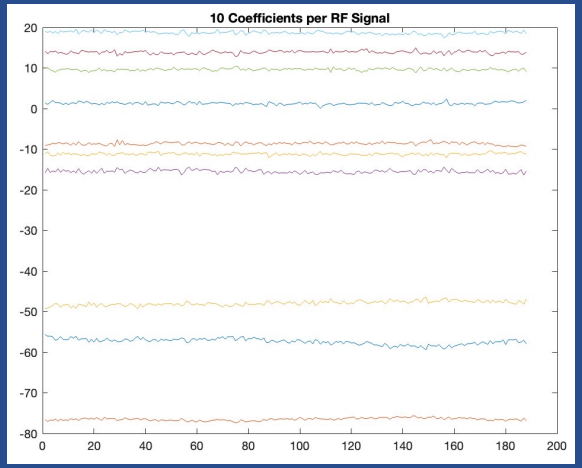


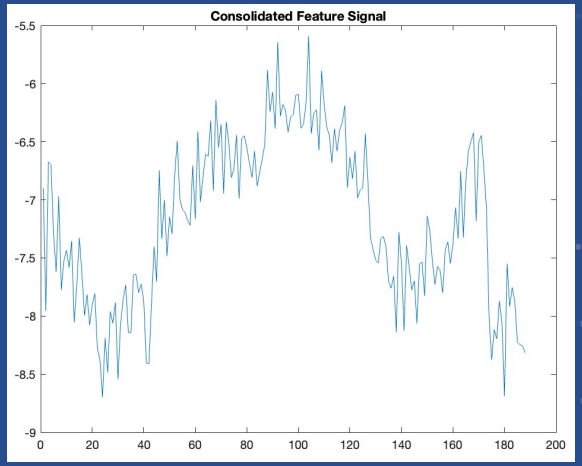
SONOMA STATE UNIVERSITY

The Mel-frequency filter bank on the left has compressed bins on the left and stretched on the right. The Linear frequency filter bank has equal sized bins through out the spectrum. Compressing the bins mean more emphasis is taken on that part of the spectrum while stretched bin has less emphasis.

Linear Frequency Cepstrum Coefficients

The image on the left has 10 coefficients extracted from 1 RF signal. I consolidated these 10 signals into 1, effectively compressing each data point from 97000 to 188 points.





The researchers used 12 coefficients in their study. They took the 12 signals and further preprocessed them and fed it to their machine learning model.

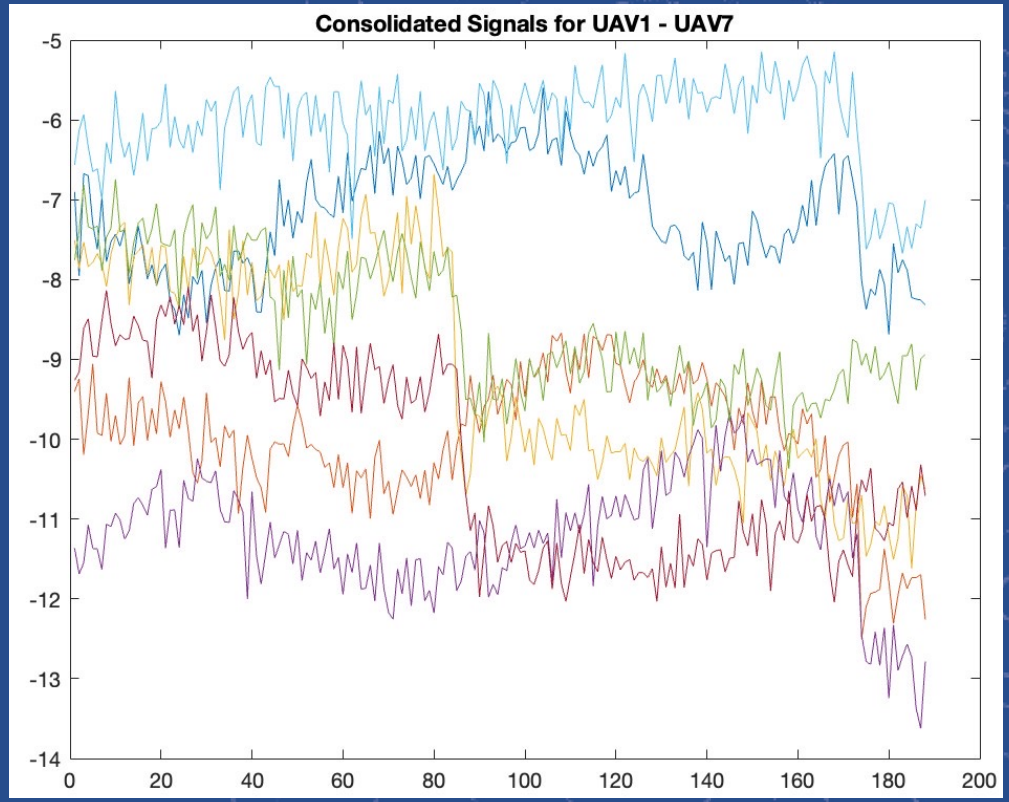
I combined 10 coefficients into 1 signal and took each data point in this signal as a feature. Using this method left me with 188 features.

SONOMA STATE UNIVERSITY

$$\sum_{n=1}^{10}(-1)^{1-n}coef(n)$$

Feature Engineering

Using 10 coefficients from the linear frequency cepstrum is enough to recreate data points where each UAV has a distinguishable transmitting signal.



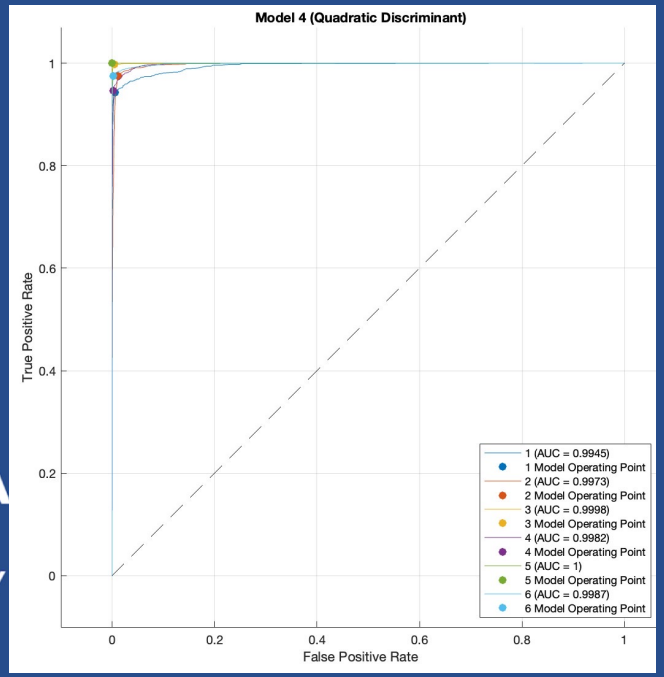
SONOMA STATE UNIVERSITY

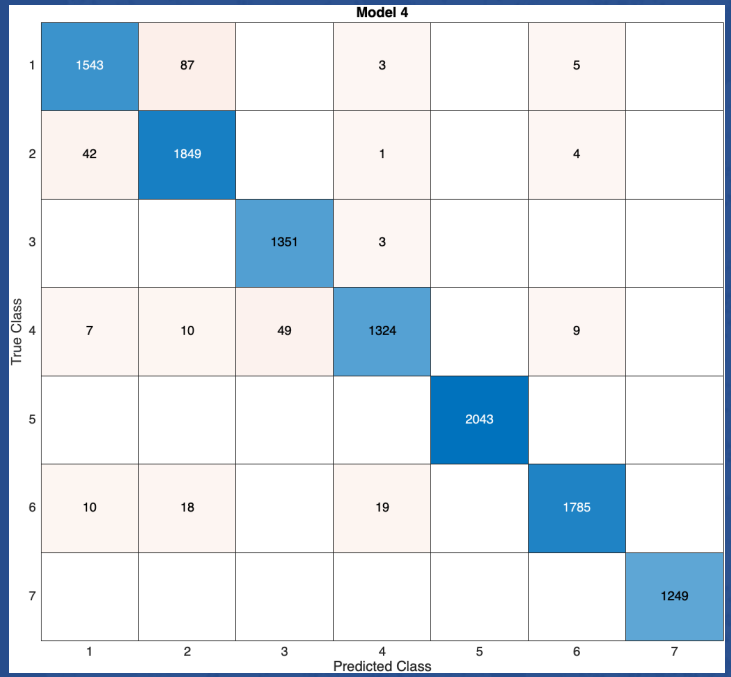
Results

In this case, the dataset has reduced in size from 10 GB to 8.1 MB (97000 to 188 points).

The machine learning model used for this study was a Quadratic Discriminant.

Total training time 20.47 seconds, with validation accuracy of 97.7%.

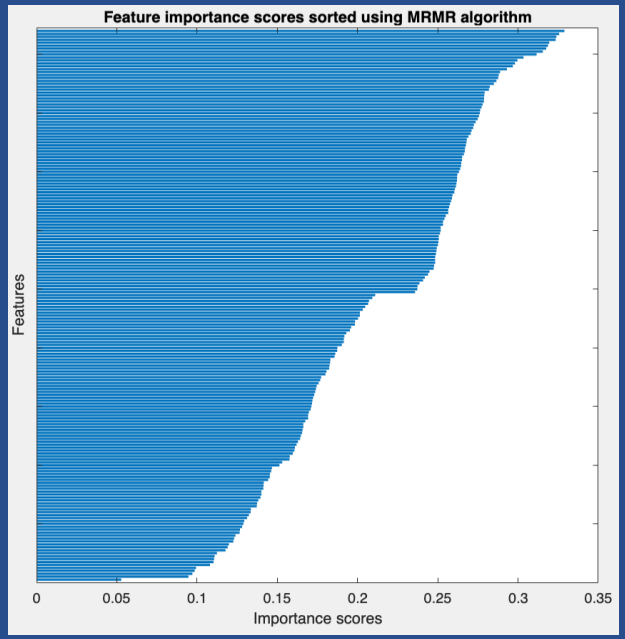


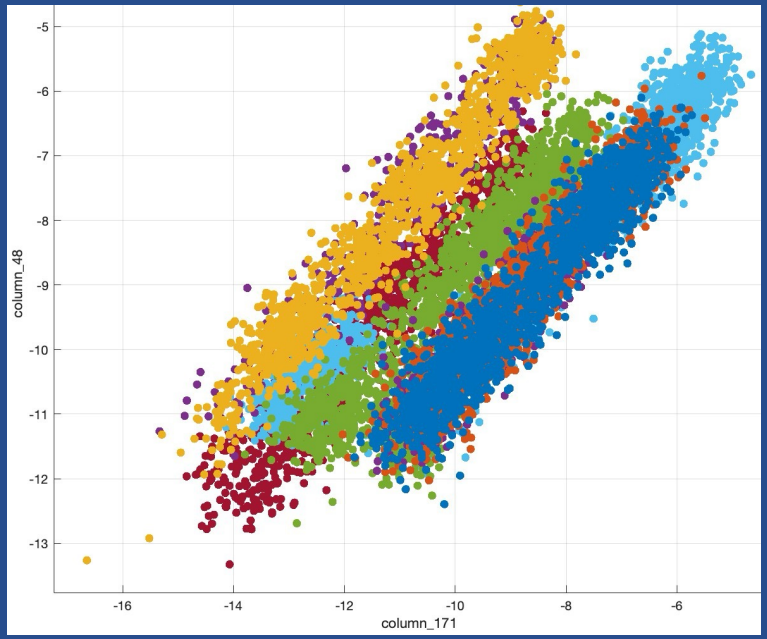


SONOMA STATE UNIVERSITY

Results

In this slide we can see how there is a lot of overlapping data on the bottom right. In this case, it would seem like any machine learning model would have a hard time predicting correctly. The important thing to notice is that each feature slowly builds up to capture most of the sample space of possible outcomes. In other words, there are many graphs like the one on the right that capture different angles of data that allow for an accurate machine learning model.





SONOMA STATE UNIVERSITY

Reference

Nasim Soltani, Guillem Reus-Muns, Batool Salehi, Jennifer Dy, Stratis Ioannidis, and Kaushik Chowdhury, "RF Fingerprinting Unmanned Aerial Vehicles with Non-standard Transmitter Waveforms," Accepted in IEEE Transactions on Vehicular Technology, Nov. 2020.

https://genesys-lab.org/papers/UAV-TVT-20.pdf

https://genesys-lab.org/hovering-uavs

https://www.mathworks.com/help/predmaint/ug/signal-features.html

https://rfmw.em.keysight.com/rfcomms/refdocs/tdscdma/Occupied_Bandwidth_(OBW)_Measurement.htm#:~:text=Occupied%20Bandwidth%3A%20the%20bandwidth%20that,frequency%20and%20the%20lower%20frequency.

SONOMA STATE UNIVERSITY