**Summary**

Random forests trained on spectral features of recordings achieved 84.2% accuracy on a test dataset for the task of labeling 1 second chunks. Although performance seems good overall, accuracy in underrepresented categories was poor and leaves lots of room for improvements. Next steps likely consist of using more bespoke methods to identify some of these underrepresented waveforms and the use of filtering to remove noise from existing label predictions.

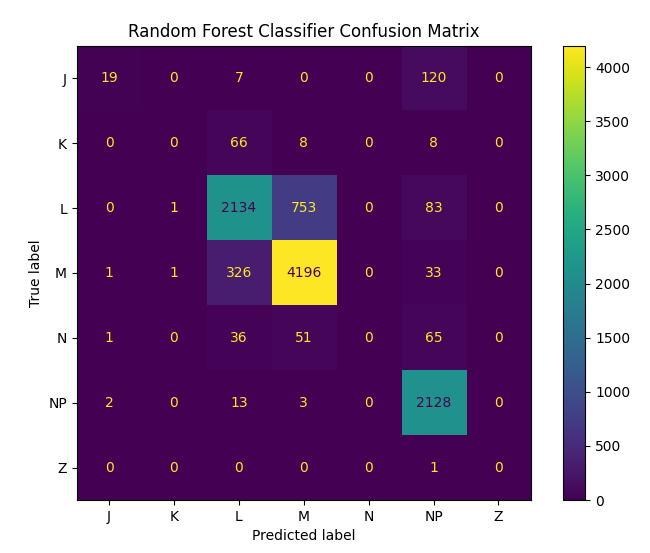
**Preprocessing**

This iteration of a Random Forests model was built using scikitlearn’s RandomForestClassifier with n\_estimators = 100, random\_state=42, and class\_weight=”balanced” parameters. Training data was produced by taking the unrectified input signal, chunking it into non-overlapping 1 second segments, and computing the following features:

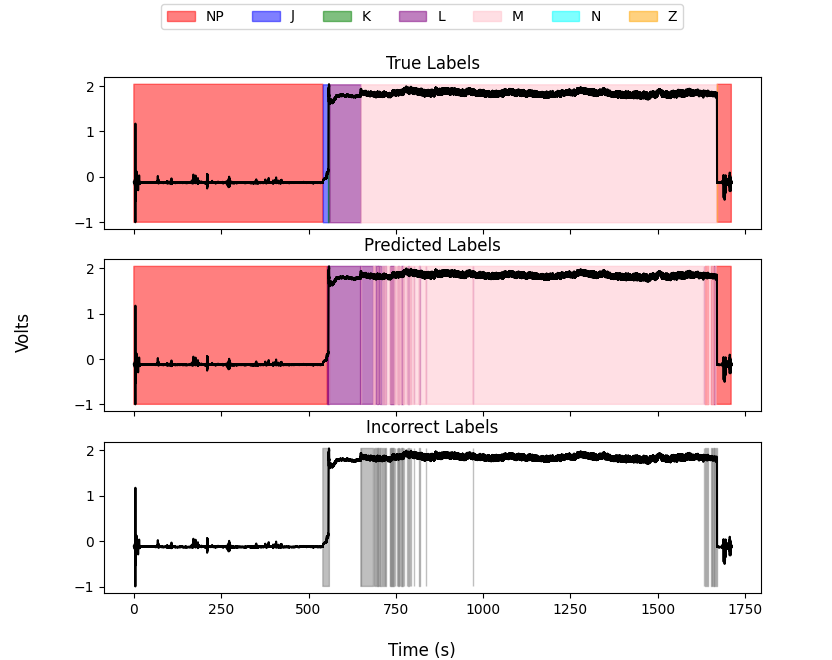
1. Largest frequency components: scipy’s fft function was used to get the six frequencies that had the highest magnitude in the frequency domain.
2. Mean: mean of the values in the chunk
3. Standard deviation: standard deviation of the values in the chunk
4. Resistance: the impedance used while taking the measurements. Same for all chunks from a given recording. *Note that inclusion of this feature was not found to improve performance.*
5. Volts: the input voltage used during measurements. Same for all chunks from a given recording.
6. Time: Timestamp of the first value in the chunk in seconds from the beginning of the recording
7. Current: the input current used during measurements, either AC or DC. Same for all chunks from a given recording.
8. Label: the most common label in the chunk. *Note that this may lead to loss of data in the case that a state is very short and the borders of the chunks work out so that the short label is essentially “gerrymandered” out of the dataset.*

**Results**

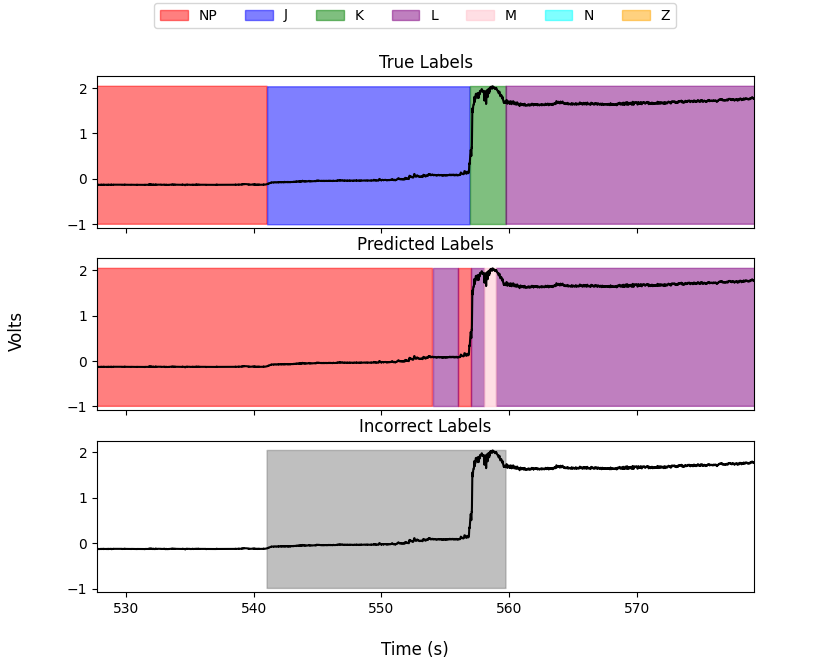
Holding out 12 of the recordings and training on the other 50, random forests produced a model with 84.2% accuracy. A confusion matrix for its performance across labels is below.



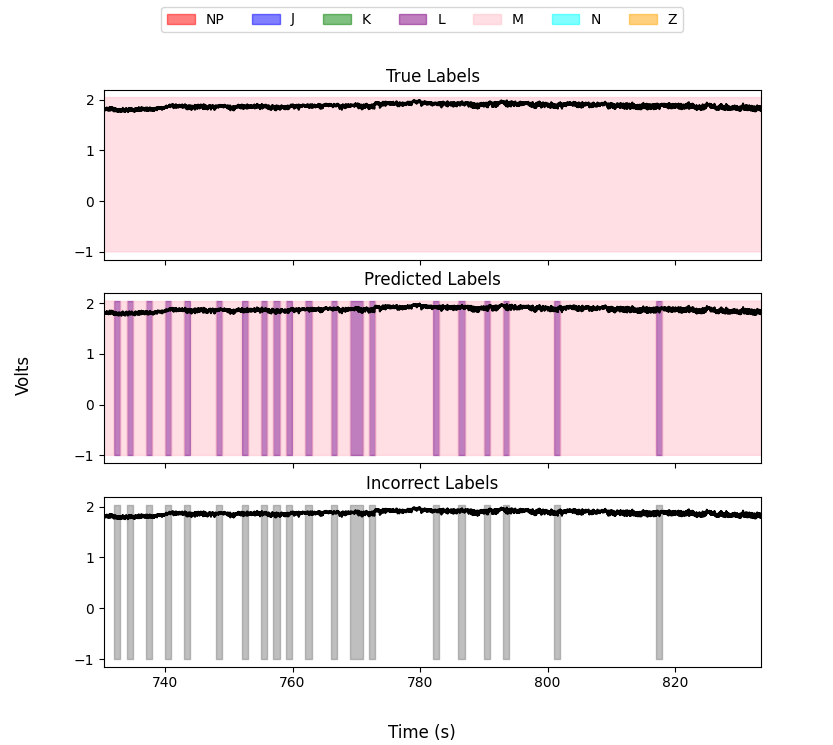
More interestingly, here are some examples of how the labels looked compared to the actual recordings:



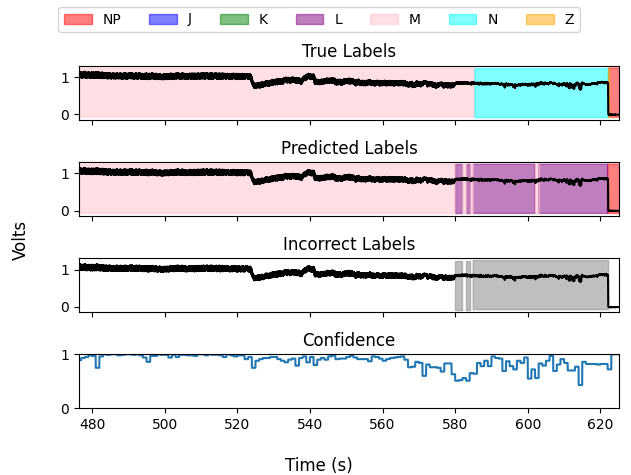
This one is pretty good and demonstrates most of the shortcomings of random forests so far. Among these are completely ignoring the J and K waveforms (zoomed image below). This is likely partially due to a lack of training data on these labels and also because they are not particularly distinctive, especially J. It may also be a result of this method’s very narrow field of view. Simpler techniques that look for sharp increases after the NP label may be all that’s necessary to improve finding of these waveforms as we know that they always occur between NP and L.



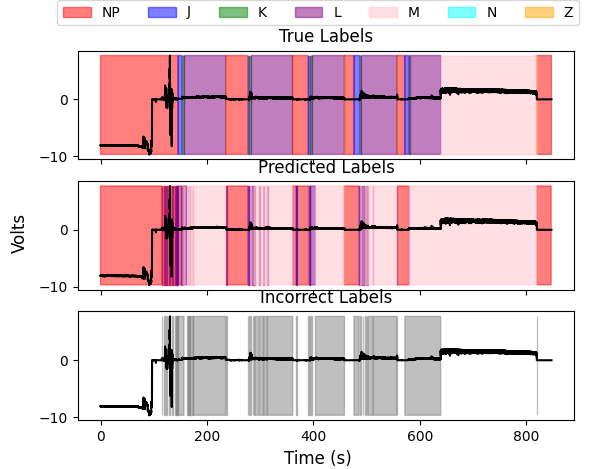
Another common source of error is the spurious labeling of L or M waveforms in the middle of larger intervals of waveforms of the other type (example below). This “barcoding” problem is probably easily solved by applying some sort of filter after labeling has taken place to replace very short intervals of a given label with the label that surrounds them.



Another example of error can be seen in the recording below in which N is completely mislabelled. In fact, looking at the confusion matrix we can see that *not a single* second of any test recording got the N label. Seeing as there were only 2.5 minutes of it total and knowing that N is used as the “all encompassing” label for anything unknown, this waveform likely requires special handling although it is unclear how to best approach that because it is irregular by definition. It may prove useful to extract a confidence from the random forests and use that to classify intervals as N if they are unable to confidently predict the waveform label. Using an enhanced plot that includes confidence (i.e. the proportion of decision trees that voted in favor of the predicted category), we see that the confidence is lower than the M that precedes it but it isn’t obviously different from the probabilities on some other waveforms in the test dataset (plot below).



The final case in which the random forests appear to perform poorly is when mosquitos make multiple probes like in the recording below. While NP is identified quite well, the L states between are picked up very poorly. The exact reason for this is unclear and this recording is one of the ones that the model performed the worst on.



**Conclusions**

Although this model performs decently well from a raw accuracy standpoint, it stands to improve in its performance on less-represented waveform types as well as through eliminating small errors. Larger issues, like completely incorrect classification of minutes of data, remain a starting point for further investigation. Other improvements can likely be had by making use of a more “global” view of the recordings. Whether this means introducing new features or perhaps an ensemble of different methods remains to be seen.