Random Forests

# Main Body Description

Random forests (RF) are an ensemble model consisting of a collection of decision trees, each trained on a random subset of the available features for either regression or classification. In this method, we follow the work of Willet et. al (2016) in applying RF to EPG and frame our EPG probe segmentation task as a classification task on non-overlapping, fixed-width windows of the unrectified probe signal where the target class of each window is the waveform type with the longest duration in that window. Features extracted from each window consisted of the mean and standard deviation along with the frequencies whose magnitudes were the largest after applying a Fourier transform to that window. These values allow us to characterize the signal appearance in each window, which can consist of hundreds of values depending on the window size, in fewer values which can then be used as features in the RF. Additionally, as the appearance of a waveform type is known to vary based on the settings used during EPG recording, we also pass in the impedance, voltage, and current type (alternating or direct) used on the EPG recording device while the signal was being measured. In this case of our dataset, this was a four-channel AC–DC electropenetrograph built by Andrew Dowell, where settings were varied across a range of values as described in Cooper et. al 2023. For details on our implementation, please refer to the supplementary materials.

# Supplementary Description

## Preprocessing

We begin by dividing the input unrectified EPG signal into equal width windows of length *w* and then computing the following features for each window:

1. Frequency Components: Using scipy’s Fast Fourier Transform implementation, we found the frequency of the *f* frequencies in each window with the largest magnitudes, which were oftentimes harmonics of each other.
2. Mean: The mean of the unrectified EPG signal in each window.
3. Standard deviation: The standard deviation of the unrectified EPG signal in each window.
4. Resistance: the impedance used on the EPG recording device while taking the measurements. This was the same for all windows in a given recording.
5. Volts: the voltage used on the EPG recording device while taking the measurements. This was the same for all windows in a given recording.
6. Current: the type of current used on the EPG recording device while taking the measurements, either alternating or direct current. This was the same for all windows in a given recording.
7. Label: the waveform type in each window with the longest duration. This was used as the training target. *Note that very short waveforms like Z are oftentimes never the longest waveform type in a window. This means that RF will inherently perform poorly on such waveform types.*

We found *w* = 3 seconds and *f* = 7 to be the best values for maximizing the cross-validated F1 score on our dataset.

# Model

We used sci-kit learn’s RandomForestClassifier class to implement the model, finding num\_estimators= 128 and max\_depth = 16 to be the best values for maximizing the cross-validated F1 score on our dataset.

# References

Willett, Denis S., Justin George, Nora S. Willett, Lukasz L. Stelinski, and Stephen L. Lapointe.

2016. Machine learning for characterization of insect vector feeding. PLOS Computational

Biology 12(11):e1005,158. doi:10.1371/journal.pcbi.1005158. URL https://app.dimensions.

ai/details/publication/pub.1014451934.