MAY 17 UPDATE

JUSTINE FILION, NEETHU GOPALAKRISHNA, SAISREE GR, SARA HALL

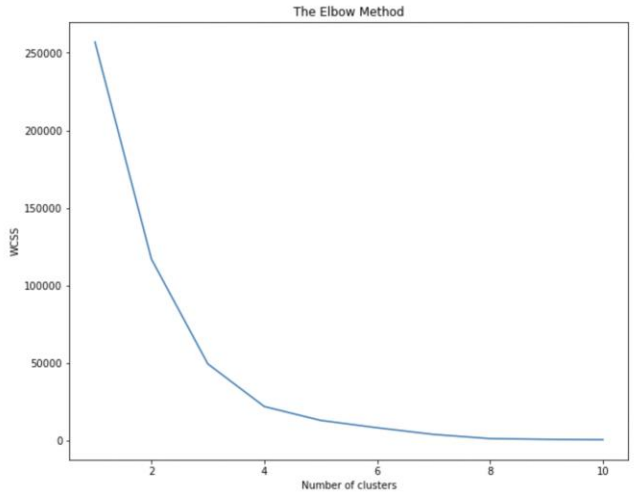
PROGRESS DURING THIS WEEK'S CYCLE
OVERALL PROGRESS

What's been tried

- Predictor wrangling
- Windowing
- Signal normalization/standardization
- Filtering
- Treating imbalanced dataset
- LOTS of research

PREDICTOR WRANGLING

- Remove duplicate IDs.
- Performed basic EDA on the dataset to check for null values and NAN
  - Certain columns contained only zeros or all the same value for all TestIDs
    - The column *TransDrift* is zero only when the test is successful.
- Plotted correlation matrix for the dataset and dropped predictors with correlation > 0.95.
- Performed k-means to see possible cluster formations with initialize at 3 clusters, as per elbow method.



PREDICTOR WRANGLING

- Performed random forest on the data to extract feature importance.
- Need to perform re-sampling of data to see if the clusters are formed better.

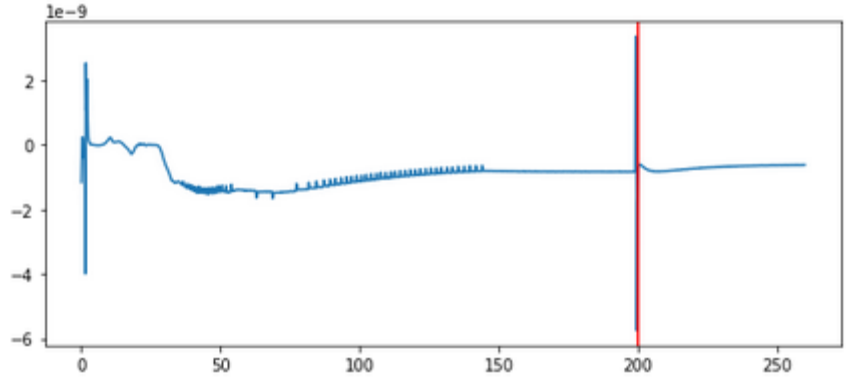
```
print(classification_report(Y_test, Y_pred1))
```

	precision	recall	f1-score	support
PinContact	0.66	0.45	0.53	51
UnSuccessful	0.99	0.99	0.99	1938
accuracy			0.98	1989
macro avg	0.82	0.72	0.76	1989
weighted avg	0.98	0.98	0.98	1989

```
feature_scores
```

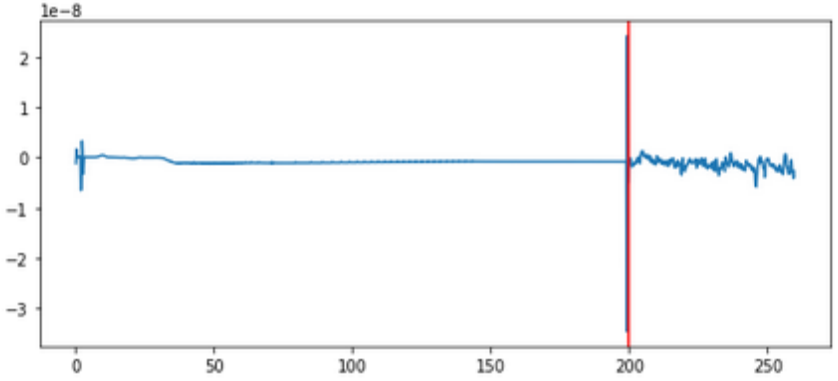
```
CExtrapolation       0.220742
CNoise               0.142294
SNoise               0.107944
CDrift               0.085921
SampleDetectTime     0.075425
PSecond              0.073891
SDrift               0.073530
CSecond              0.072283
TransDrift           0.071997
FluidNumber          0.045553
FluidType            0.014239
AFirst               0.013949
CWindowMovedBack     0.002231
dtype: float64
```

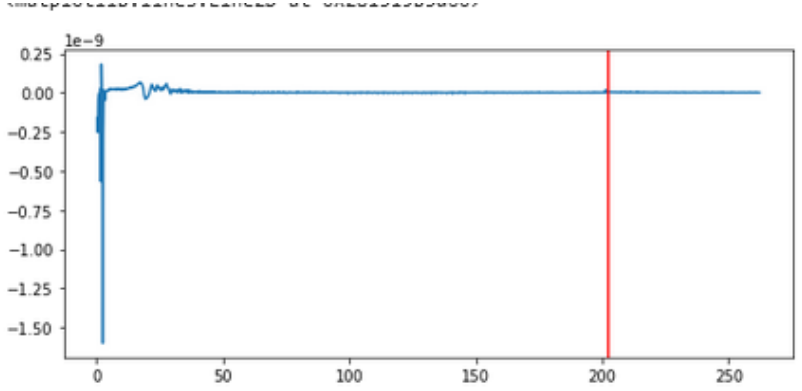
WINDOWING – EXAMPLE TRACES



Successful

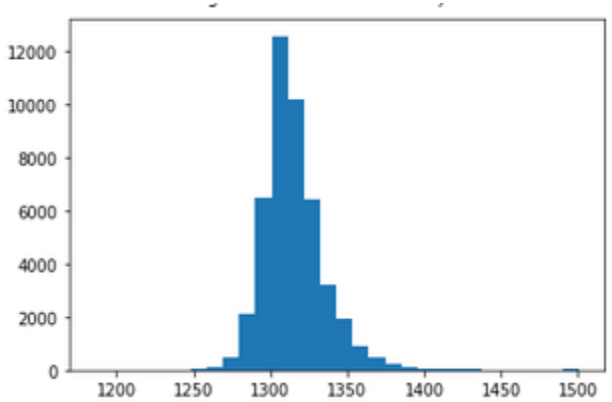
Unsuccessful





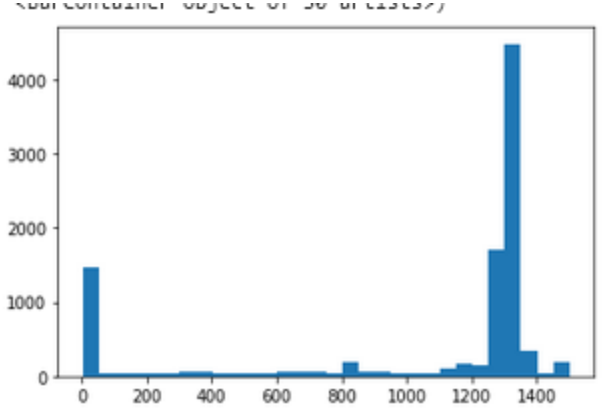
Pin Contact

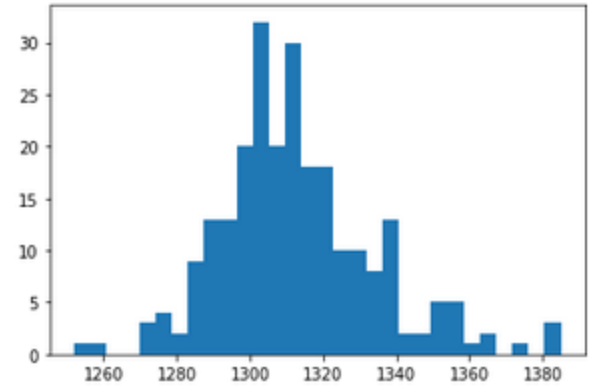
LENGTHS OF READINGS



Successful

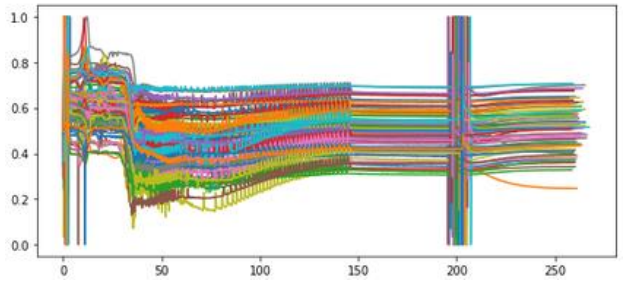
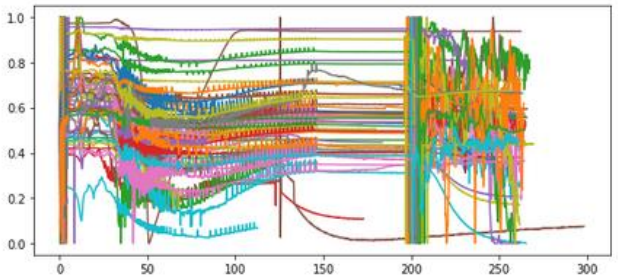
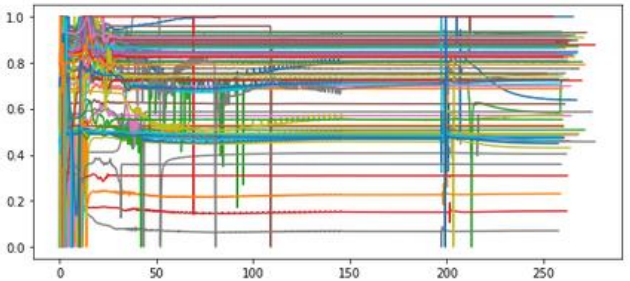
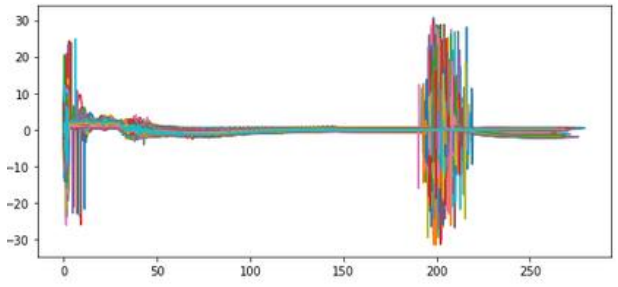
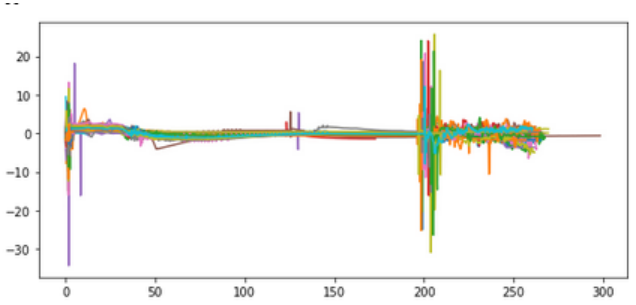
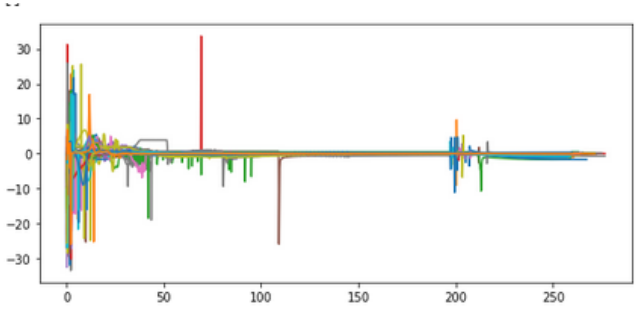
Unsuccessful



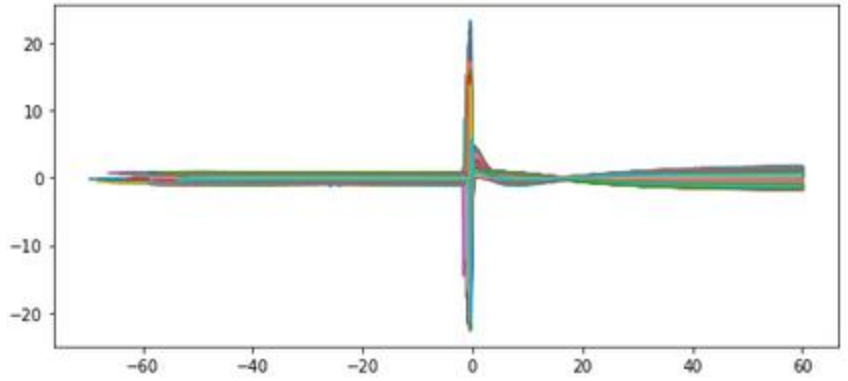


Pin Contact

SIGNAL NORMALIZATION

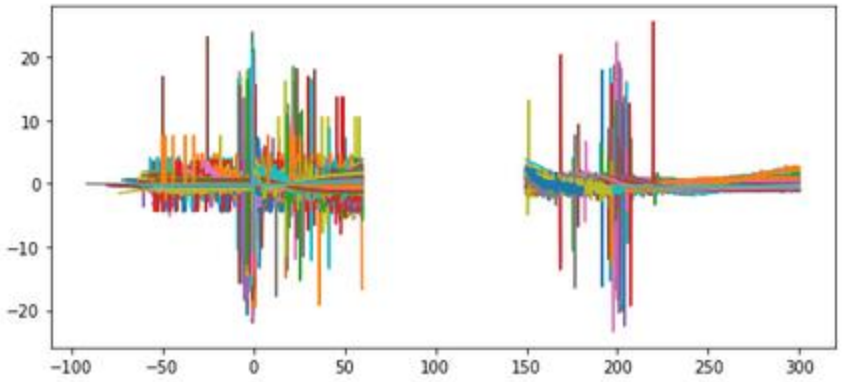
	Successful	Unsuccessful	Pin Contact
Normalization			
Standardization			

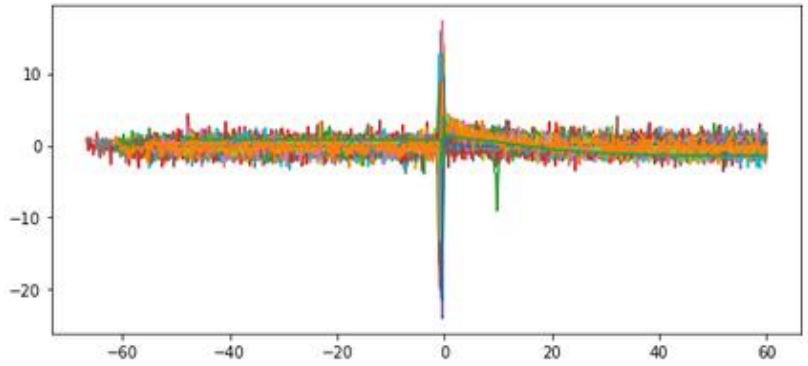
STANDARDIZATION WITH WET-UP REMOVED



Successful

Unsuccessful

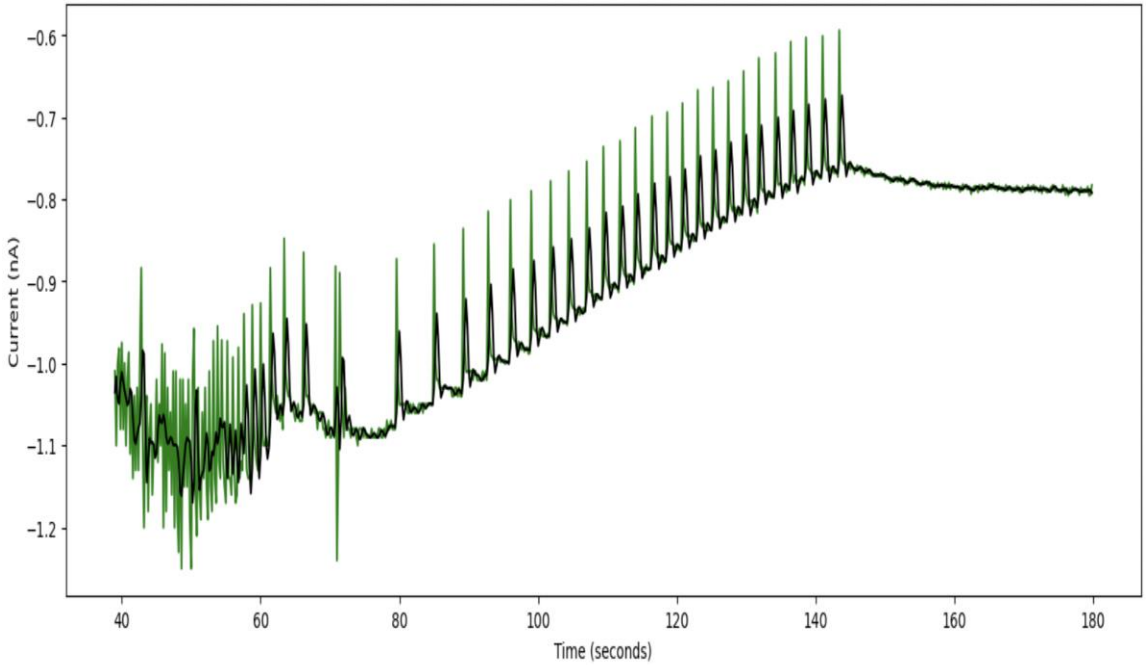




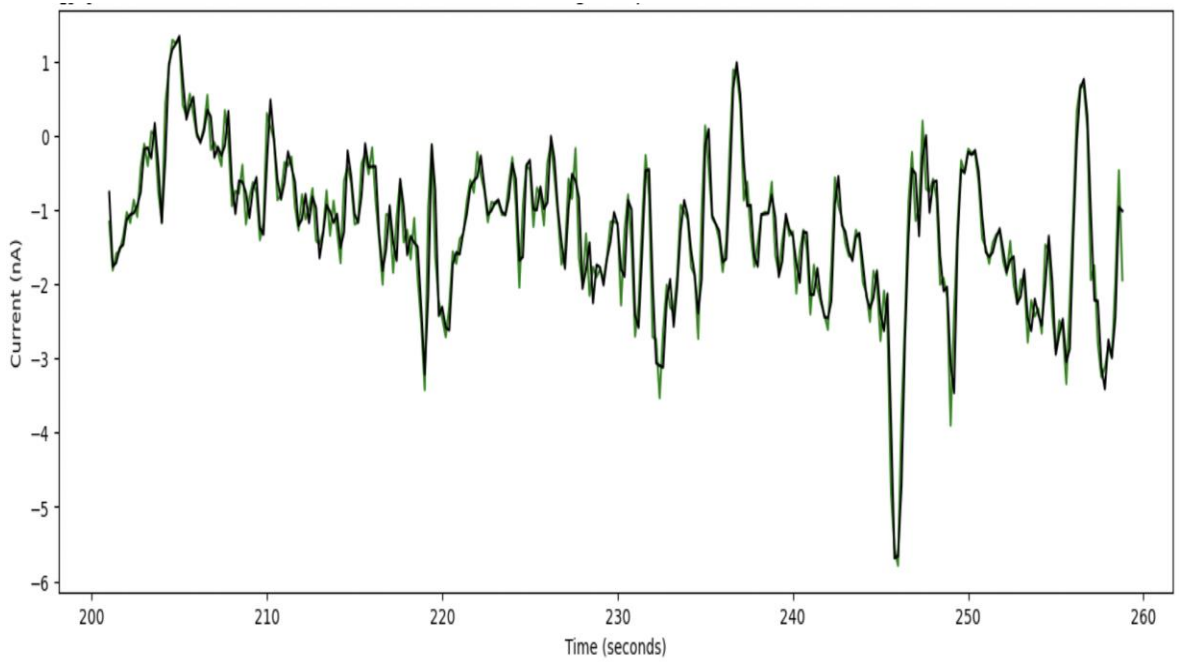
Pin Contact

LOW PASS FILTERING

- Filter that lets low frequencies pass and removes high frequencies from waveform
- This causes waveforms to smoothen out



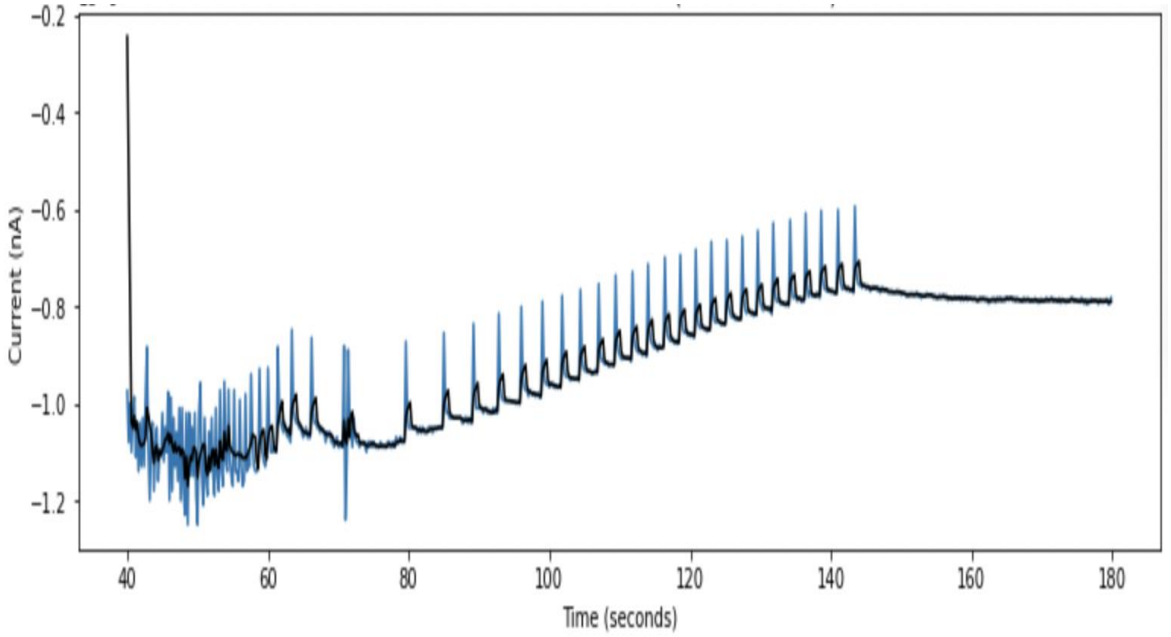
creatinine sensor noise



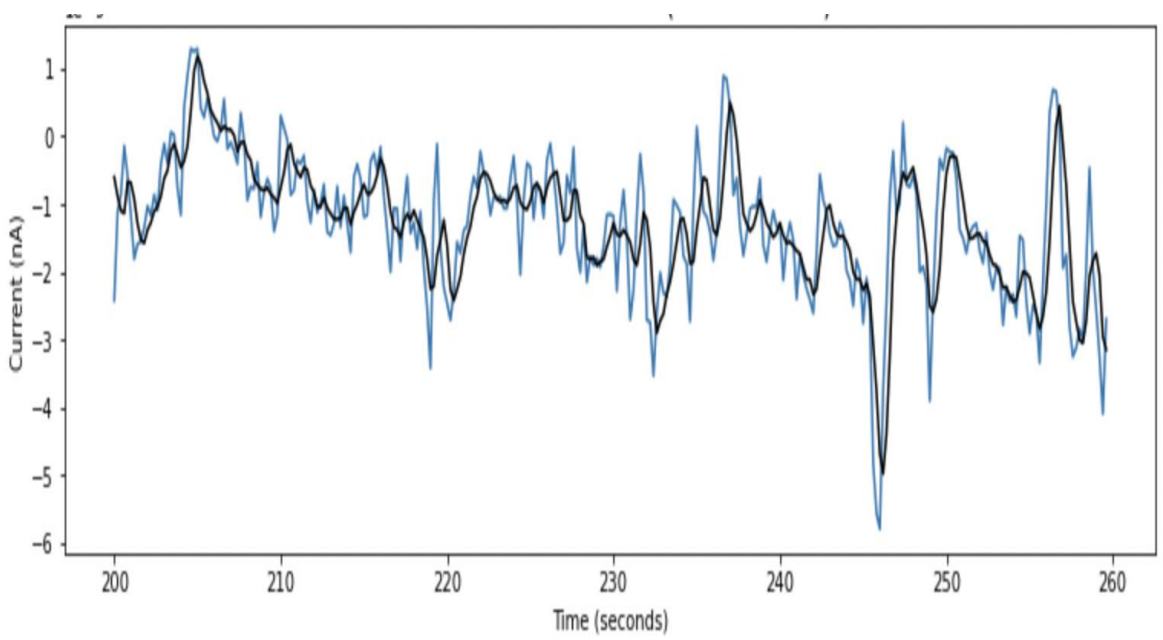
noise after sample detection

BAND PASS FILTERING

- Filter that lets band of frequencies to pass through



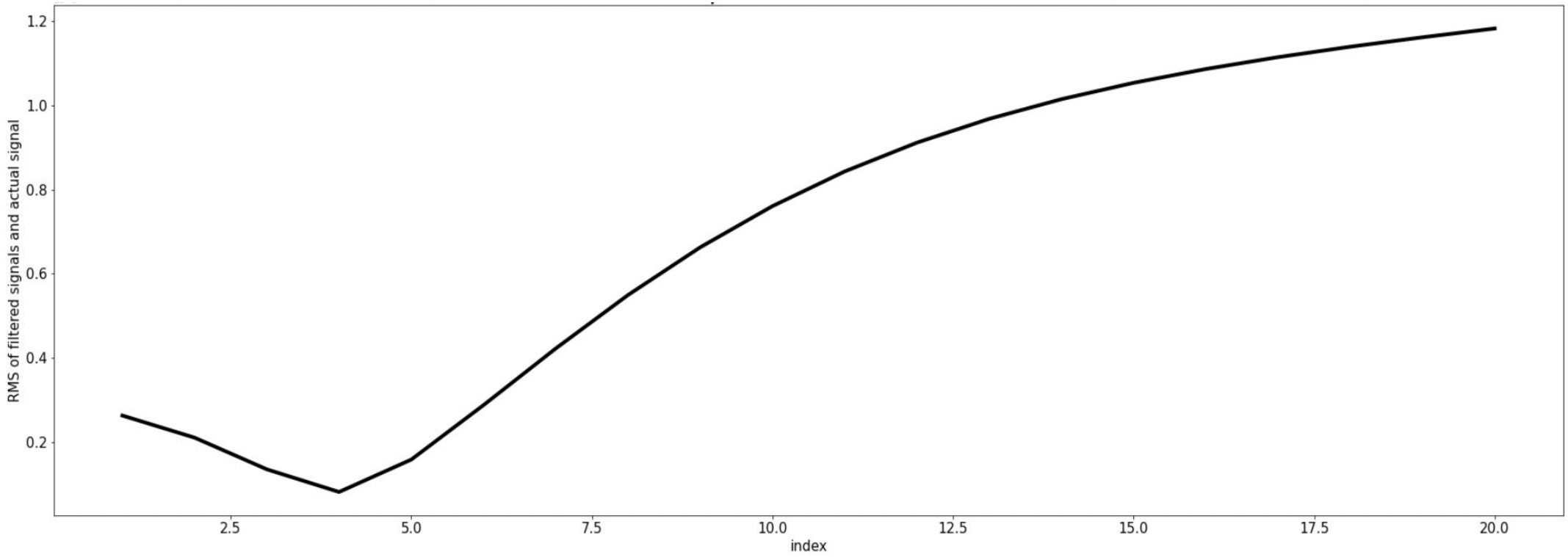
creatinine sensor noise



noise after sample detection

COMPARISON OF FILTERED SIGNALS

Used Root Mean Square (RMS) of filtered signal of different bands vs the actual signal



IMBALANCED DATASET (IDEAS)

- 1) Bootstrapping
  - i.e. Use ~175 unsuccessful and ~175 pin contact errors → find clusters
  - Use 175 new unsuccessful and the same 175 pin contact errors → find clusters
  - Continue until all unsuccessful readings have been seen
  - Q : How are we going to aggreagate the results of all the clusters. Form clusters of similar clusters?

IMBALANCED DATASET (IDEAS)

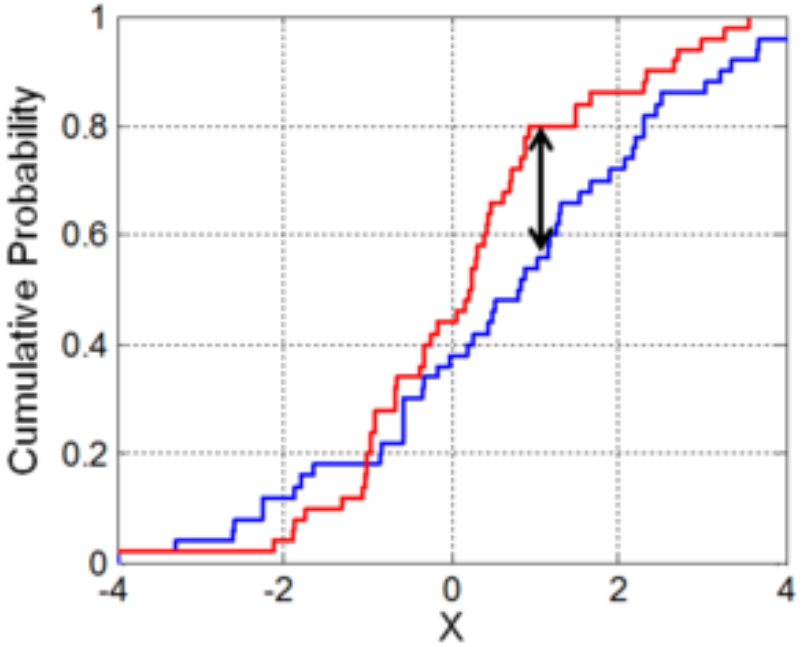
- 2) Find a representative sample of successful/unsuccessful readings
  - 2.1) Using the predictor file & the two-sample Kolmogorov-Smirnov test sampling
  - 2.2) Using the waveforms

IMBALANCED DATASET (IDEAS)

- 2.1) Using the predictor file & the two-sample Kolmogorov-Smirnov test sampling

H0 : The 2 distributions are identical
H1 : The 2 distributions are not identical

If the statistic is small → p-value is high → fail to reject H0 → 2 distributions are identical



https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

KOLMOGOROV-SMIRNOV TEST

```
kolmo_test(successful2, 1000, 0.05)
```

```
{'CExtrapolation': [0.8710775759659111, True],
 'CMean': [0.5462690104731439, True],
 'CDrift': [0.9448864294305432, True],
 'CNoise': [0.8282820223027797, True],
 'CSecond': [0.6197319185124562, True],
 'SExtrapolation': [0.8748995423746171, True],
 'SMean': [0.8748995423746171, True],
 'SDrift': [0.6426228831198901, True],
 'SNoise': [0.47694027276459017, True],
 'SSecond': [0.29286473932060386, True],
 'PMean': [0.626379092277919, True],
 'PDrift': [0.049077427225685866, False],
 'PNoise': [0.6528067173449178, True],
 'PSecond': [0.44705345833275956, True],
 'AExtrapolation': [0.4992275143554922, True],
 'AMean': [0.7273140115362452, True],
 'ADrift': [0.6384056074953328, True],
 'ANoise': [0.0264090541521258, False],
 'ASecond': [0.4726403627581633, True],
```

```
kolmo_test(successful2, 2000, 0.05)
```

```
{'CExtrapolation': [0.6773138530642135, True],
 'CMean': [0.7824266720606812, True],
 'CDrift': [0.15086290670995495, True],
 'CNoise': [0.2853004289760581, True],
 'CSecond': [0.4187037081549171, True],
 'SExtrapolation': [0.3681144732252517, True],
 'SMean': [0.3681144732252517, True],
 'SDrift': [0.6400129926074813, True],
 'SNoise': [0.6074274767898193, True],
 'SSecond': [0.5432887182528332, True],
 'PMean': [0.7344248375725215, True],
 'PDrift': [0.21148647134115584, True],
 'PNoise': [0.569395907340357, True],
 'PSecond': [0.22505757104979496, True],
 'AExtrapolation': [0.5502870153497481, True],
 'AMean': [0.9440205582057362, True],
 'ADrift': [0.8047662338408447, True],
 'ANoise': [0.06940811307775663, True],
 'ASecond': [0.9824432744141767, True],
```

KOLMOGOROV-SMIRNOV TEST – WHY I DON'T TRUST IT

0 0 0 0 10 → mean of 2
2 2 2 2 2 → mean of 2

- Predictors are aggregate measures that don't necessarily tell us about the distribution of the waveforms

IMBALANCED DATASET (IDEAS)

- 2) Find a representative sample of successful/unsuccessful readings
  - 2.1) Using the predictor file & the two-sample Kolmogorov-Smirnov test sampling
  - 2.2) Using the waveforms

PRELIMINARY/INCREMENTAL RESULTS

- Wrangled the data files
  - Number of readings in the timeseries and the predictors match
  - Removed columns that were not useful
- Standardized the waveforms
- Filtered the waveforms (low pass and band pass)
- Decided to use all unsuccessful and pin contact readings

PLANNING AND ACTIONS FOR THE NEXT CYCLE

Sara

- Finish wrangling the time series data and separate windows.
- Try different clustering methods on the windowed data.
- Read into longest common subsequence as a distance measure

Saisree

- Find DTW distance matrix (applied to specific windows)
- Build SOM clustering model.

Neethu

- Apply discrete wavelet transforms for feature extraction
- Use features for clustering algorithms (applied to specific windows)

Justine

- Create feature matrix based on the raw waveforms
- Use various clustering algorithms (applied to specific windows)

DEVIATION FROM THE ORIGINAL PLAN/SCHEDULE

ACCOMPLISHED ALL LAST WEEK'S TASKS?

- Will focus solely on timeseries data
  - Split in 2 teams of 2, according to feature-based clustering and shape-based clustering
- Will find clusters only within the unsuccessful readings
- Last week's tasks:

LITERATURE REVIEW/DATA PRE-PROCESSING	
Statistical test to use subset of successful	Justine
Building training set with bootstrap	Justine
Filtering (noise reduction)	Sara, Saisree
Windowing of time series	Sara, Saisree
Cleaning and wrangling the predictor file	Neethu

DEVIATION FROM THE ORIGINAL PLAN/SCHEDULE

AHEAD/BEHIND/ON TRACK?

May 9 - May 15	Data Preprocessing and research on time series analysis	• Figure out ways to perform noise reduction. • Look into ways to deal with unbalanced data (determine how different the successful readings are from each other to see if we can justify using fewer samples). • Use visualizations to see how we can divide the time series into different windows. • Research methods for time series clustering. • Clean and wrangle the data. • Figure out how to build our training/testing sets.
May 16 - June 5	Modelling	• Try to build various machine learning pipelines to figure out what works and what doesn't in terms of clustering different types of unsuccessful readings • If the unsupervised pipelines are unsuccessful, we will try building some supervised pipelines to classify successful, unsuccessful, and pin contact. • Midterm presentation May 31.

SUMMARY OF INTERACTIONS WITH THE CLIENT

- Exchanged a few emails throughout the week
  - Our data was updated
- Meeting on Friday, May 13th
  - Presented them with the progress we made over the week
  - Asked questions to get a better understanding of how our pipeline would be useful to them
  - They also asked us questions to better understand our thought process

SUMMARY OF INDIVIDUAL AND TEAM EFFORTS MAY 9 - 15

- Sara :
  - Data wrangling: 9 hours
  - Data exploration: 10
  - Literature review: 11.5
  - Administrative work/Meetings : 8.5
  - **Total : ind. + team = 39**

- Neethu :
  - Data wrangling and cleaning : 14 hrs
  - Researching : 6 hrs
  - Modelling on the wrangled data to find insights : 6 hr
  - Others : (virtual env, minutes , slides) : 5
  - **Total : ind. + team = 37.5**

Saisree :

- Researching : 12 hrs
- Wrangling, filtering : 10 hrs
- Fourier transformation , DTW: 8 hrs
- Administrative work/Meetings : 9 hrs
- **Total : ind. + team : 39 hrs**

- Justine:
  - Researching : 20 hours
  - Kolmogorov-Smirnov Test, resampling function, feature extraction : 4.5 hours
  - Others (setting up virtual environment, writing minutes, slides for presentations, generating data, meetings etc.) : 15 hours
  - **Total : ind. + team = 39.5**
- Team:
  - **Time spent in meetings : 6.5 hours**