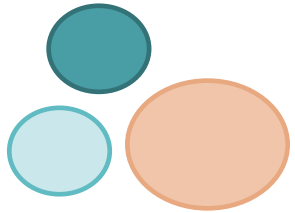
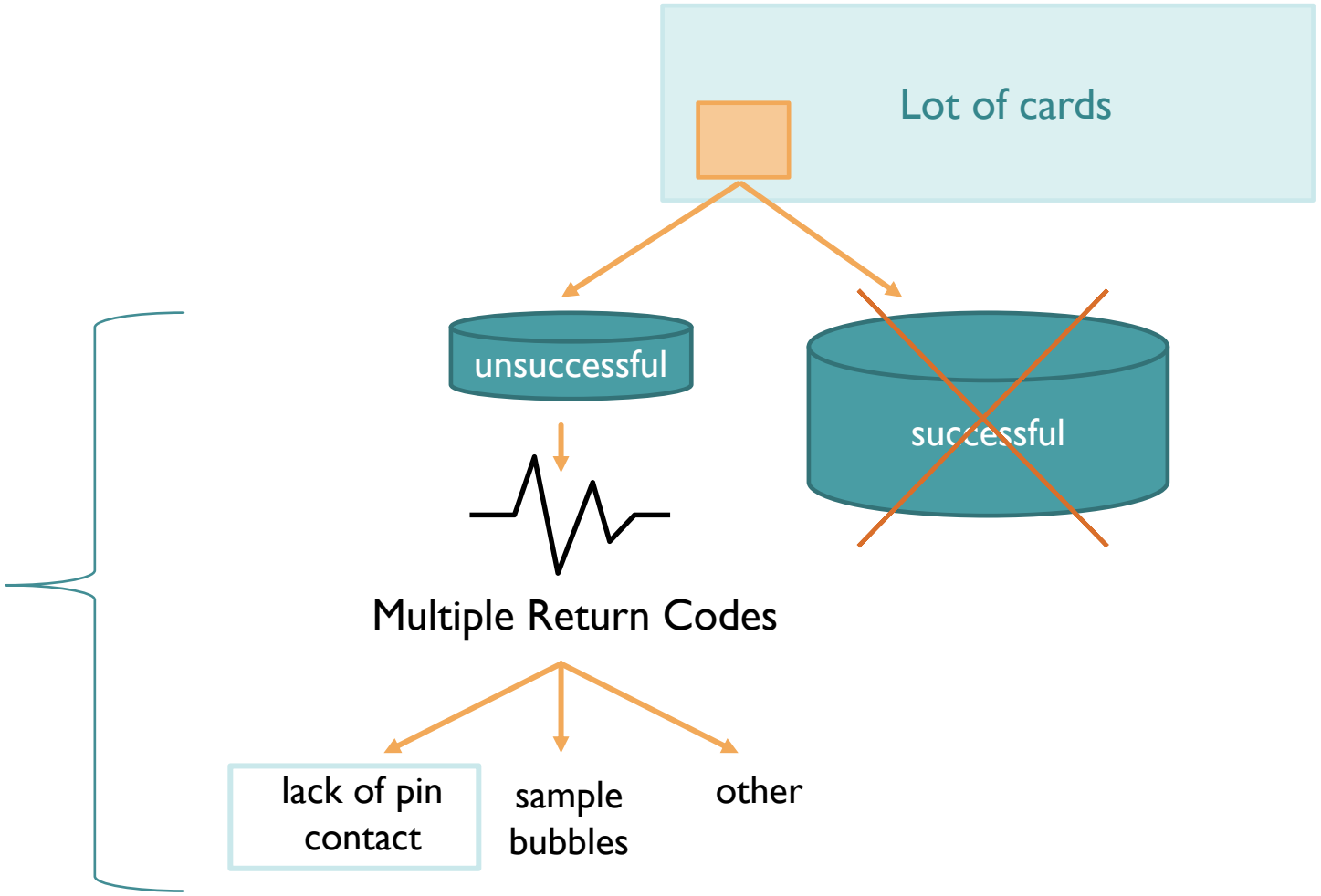
JUNE 7 UPDATE

JUSTINE FILION, NEETHU GOPALAKRISHNA, SAISREE GR, SARA HALL

RECAP: ANOMALY DETECTION IN BIOSENSOR WAVEFORMS

Can we find clusters?





RECAP: ANOMALY DETECTION IN BIOSENSOR WAVEFORMS

Research Questions:

Can we develop machine learning pipelines to cluster readings and isolate pin contact errors?

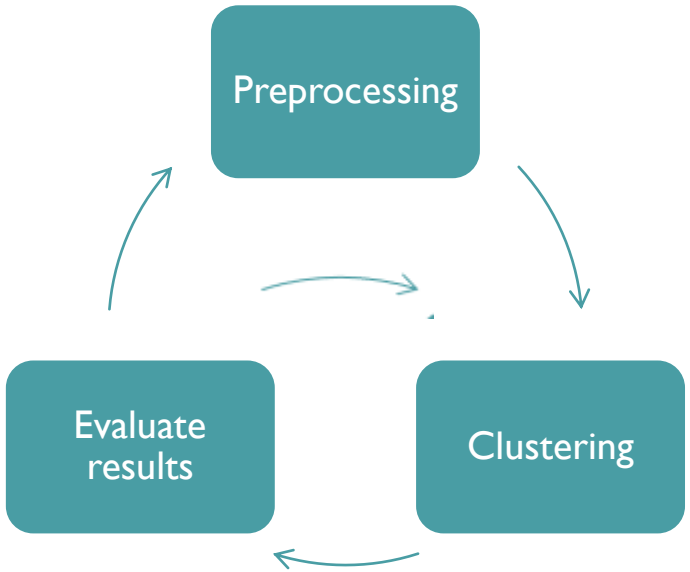
Determine which methods are effective and which are not for identifying anomalies in biosensor readings?

Deliverables:

- Well commented Python code for everything we have tried
  - Preprocessing
  - Clustering
  - Diagnostics
- A final report detailing our attempts
  - Will explain design decisions

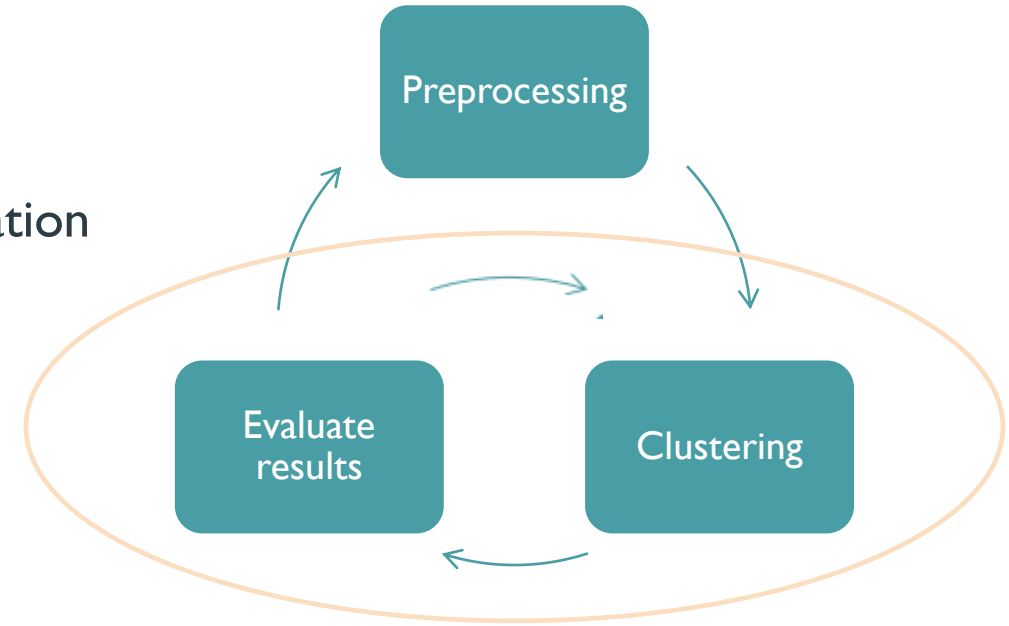
OVERALL PROGRESS

Two Iterations of:



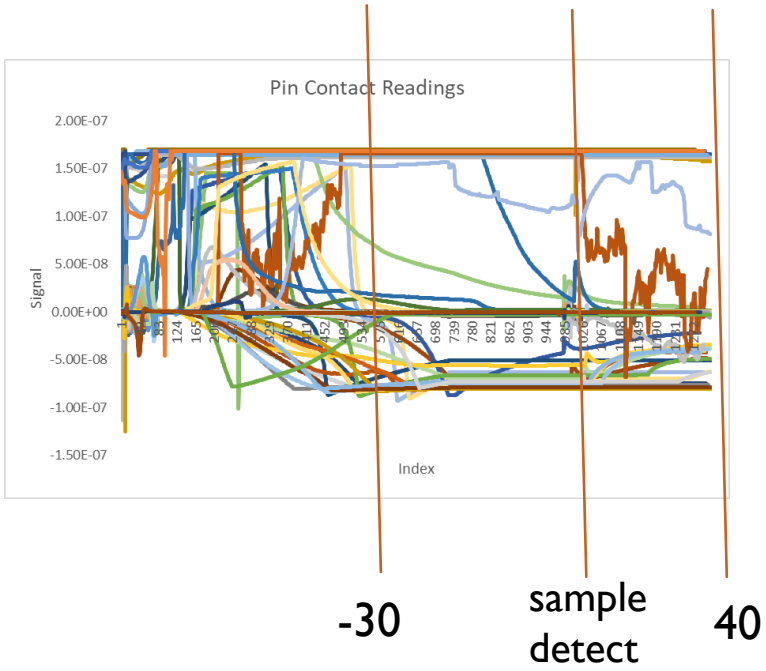
PROGRESS OVER LAST WEEKLY CYCLE

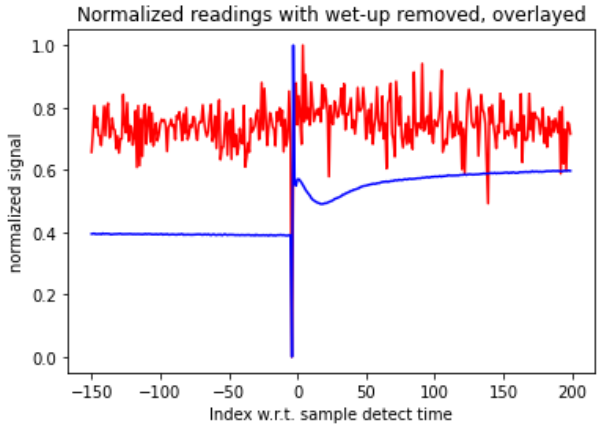
2nd Iteration



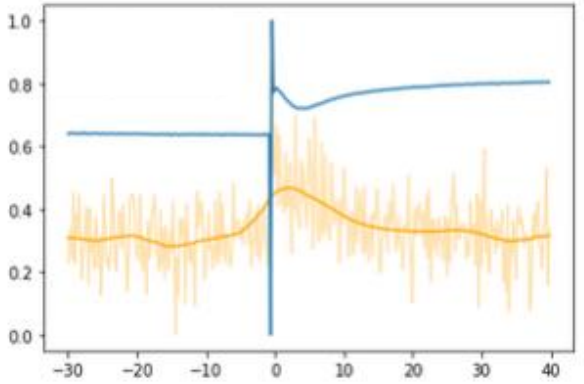
- KMeans
- SOM
- Autoencoder
- Kshape clustering

RECAP ON PREPROCESSING FOR ITERATION 2





Convolution



PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2

TSFRESH Predictors

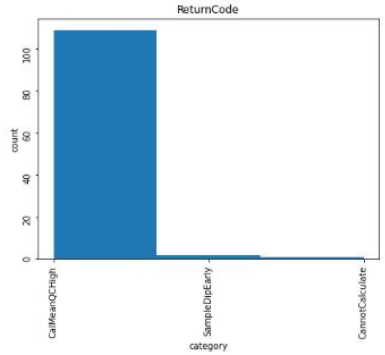
- Phase 1: Feature Extraction (~ 770 features)

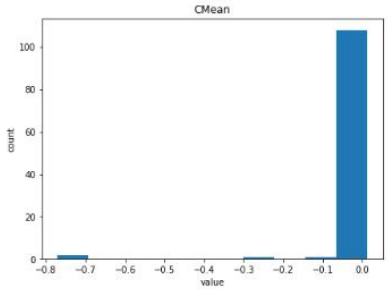
TestId	value__abs_energy	value__root_mean_square	value__absolute_sum_of_changes
8071094	0.177528	0.551571	3.345260
8078100	-0.181169	-0.009647	1.827179

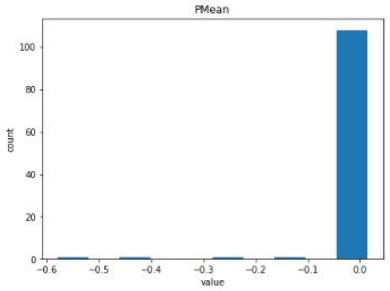
- Phase 2: Feature Significance Testing (~550 features)
- Phase 3: PCA for dimension reduction (~ 30 components)
  - 95 % accumulated amount of variance explained
- Phase 4: Clustered

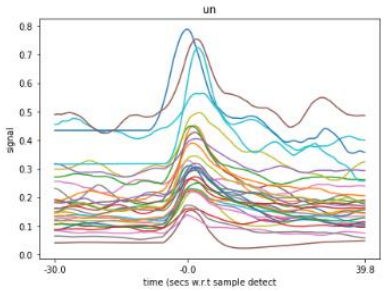
KMEANS

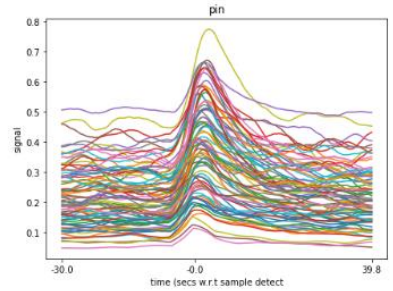
```
Cluster 7
-----------------------------------
Number of cont : 28
Number of synth : 0
Number of wild : 54
Number of tot_pins : 82
Number of un : 30
```





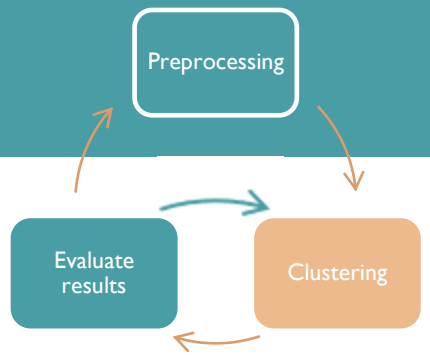




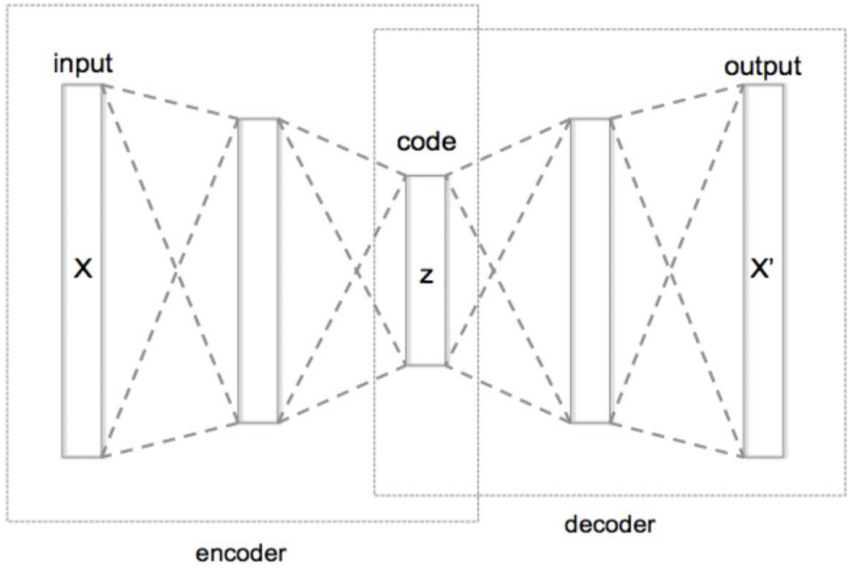


PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2

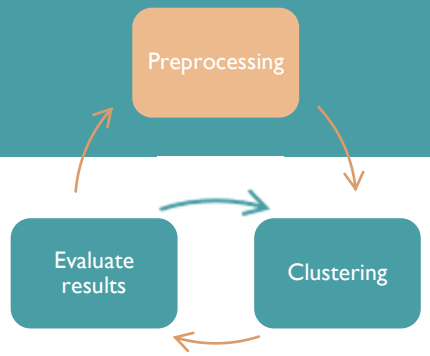


Autoencoder for feature extraction



PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2

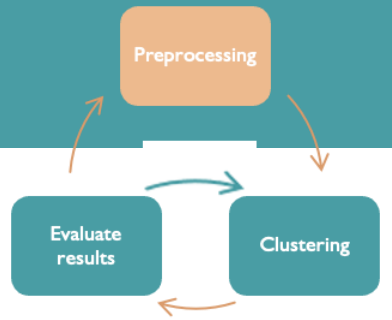


Autoencoder for feature extraction

- Extracted ~350 features with encoder
- Performed PCA on these features
  - ~3 components to account for ~95% of total variation
- Feature extraction from the predictor file using Random Forest

PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2

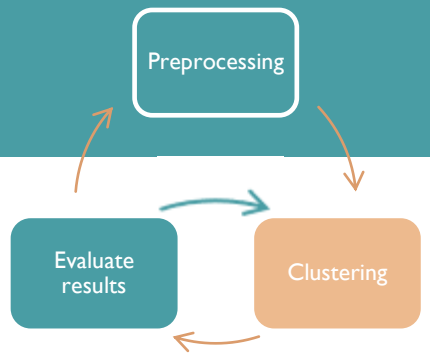


Random Forest Variable Importance:

CExtrapolation	0.220742
CNoise	0.142294
SNoise	0.107944
CDrift	0.085921
PSecond	0.073891
SDrift	0.073530
CSecond	0.072283
TransDrift	0.071997
AFirst	0.013949
CWindowMovedBack	0.002231

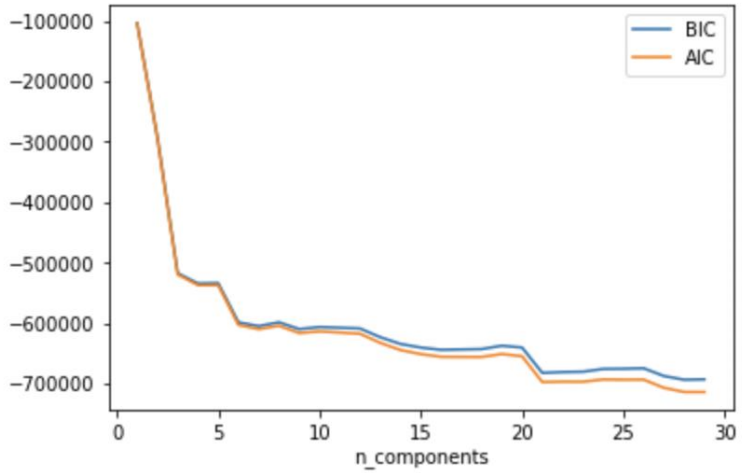
PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2



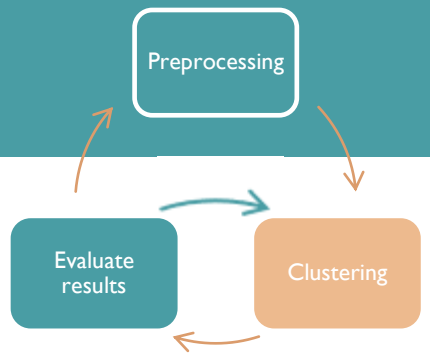
Clustering using 3 first PC + 10 predictors

1. Chose the number of clusters
   - 21 clusters



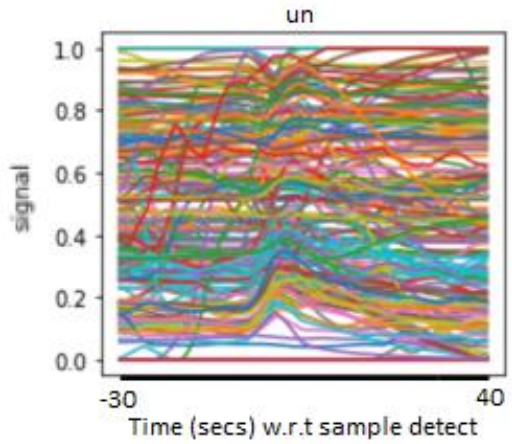
PRELIMINARY/INCREMENTAL RESULTS

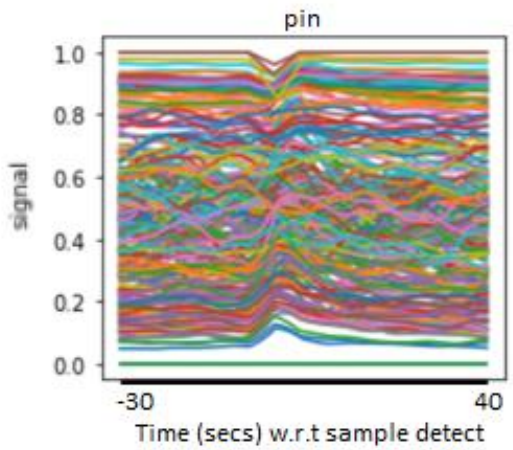
CLUSTERING – ITERATION 2



Clustering using 3 first PC + 10 predictors

1. Chose the number of clusters
2. Used Gaussian Mixture Modelling for clustering





Number of pins: 279

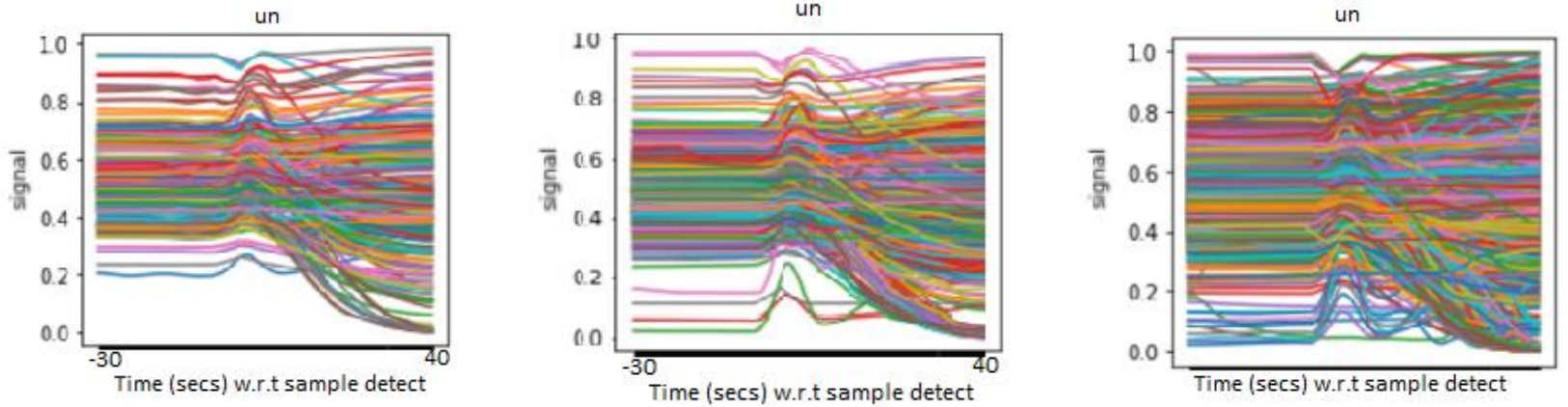
Number of unsuccessful: 230

PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2

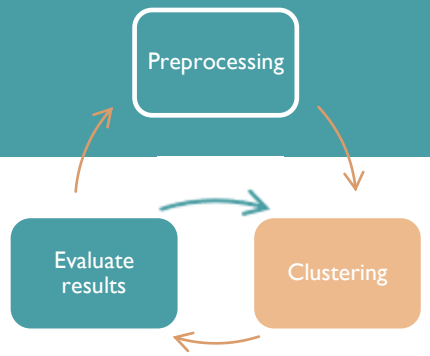
Clustering using 3 first PC + 10 predictors

Clusters containing only unsuccessful readings :



PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2

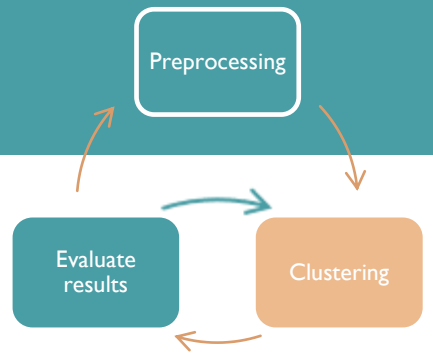


Clustering with KShape

- Uses the whole time series to cluster
- Shape based clustering
- Centroid based algorithm

PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2

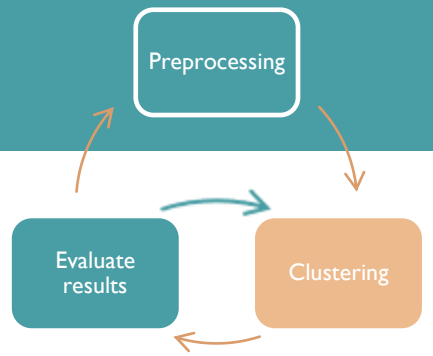


Whole timeseries clustering with KShape

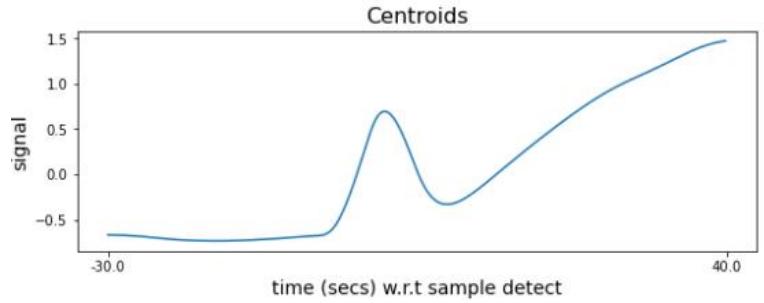
1. Randomly assign the readings in K different clusters

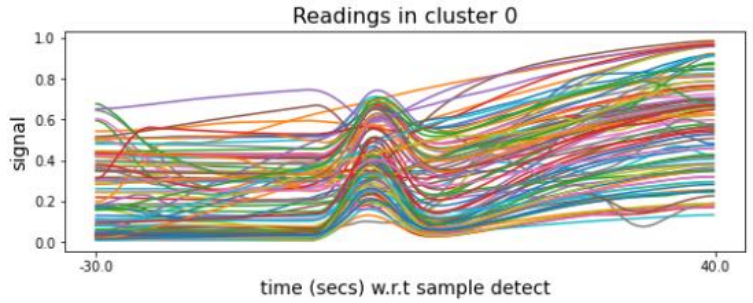
PRELIMINARY/INCREMENTAL RESULTS

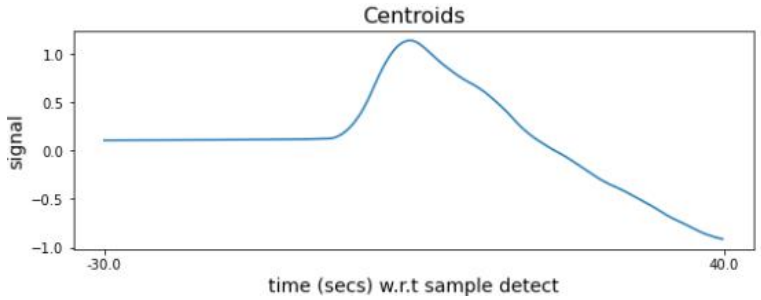
CLUSTERING – ITERATION 2

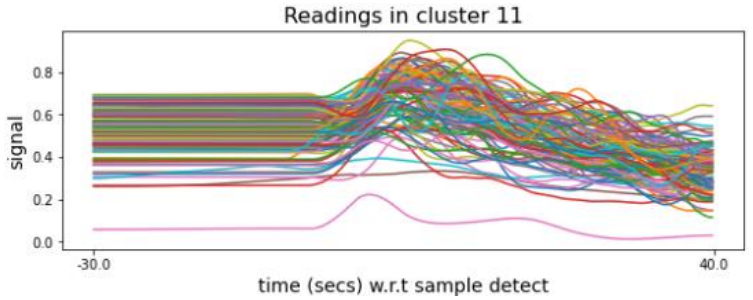


1. Randomly assign the readings in K different clusters
2. Calculate the centroid of each cluster



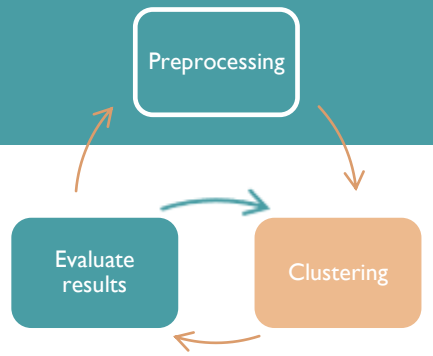




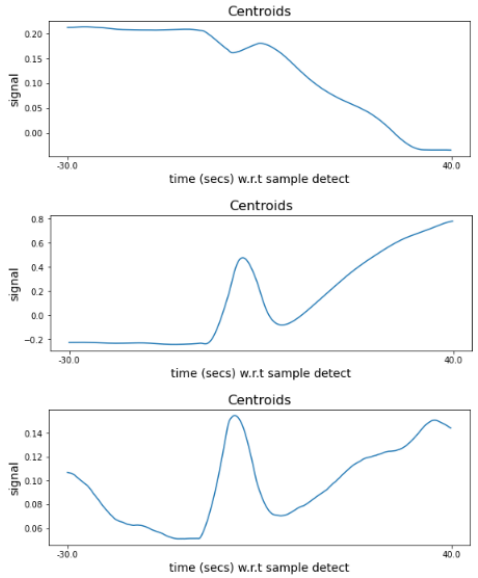


PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2

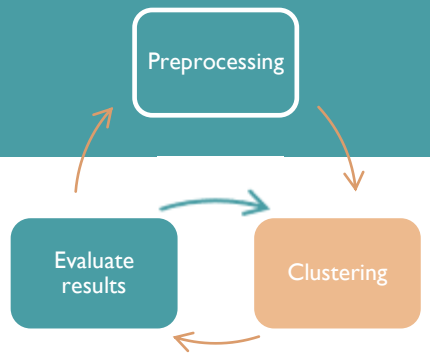


1. Randomly assign the readings in K different clusters
2. Calculate the centroid of each cluster
3. Assign each reading to the cluster with the most similar centroid



PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2

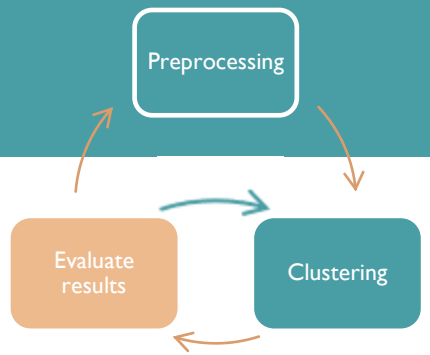


Whole timeseries clustering with KShape

1. Randomly assign the readings in K different clusters
2. Calculate the centroid of each cluster
3. Assign each reading to the cluster with the most similar centroid
4. Repeat step 2 and 3 until convergence

PRELIMINARY/INCREMENTAL RESULTS

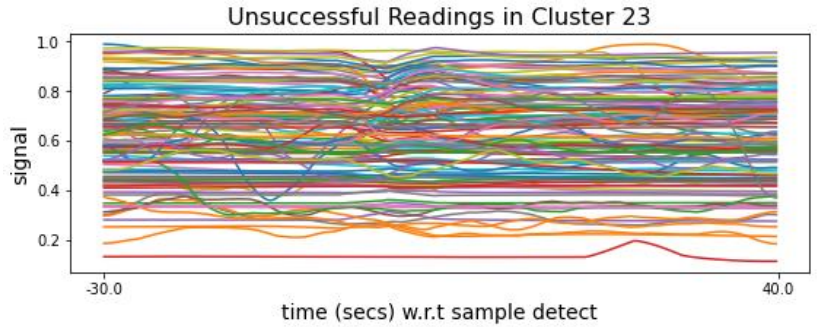
CLUSTERING – ITERATION 2

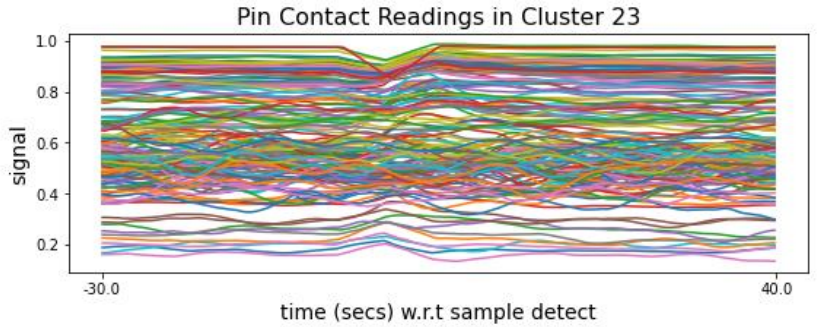


Whole timeseries clustering with KShape

Number of pins: 167

Number of unsuccessful: 135

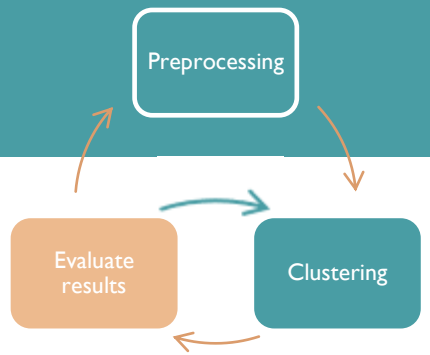




* Contains all of the synthetic pins

PRELIMINARY/INCREMENTAL RESULTS

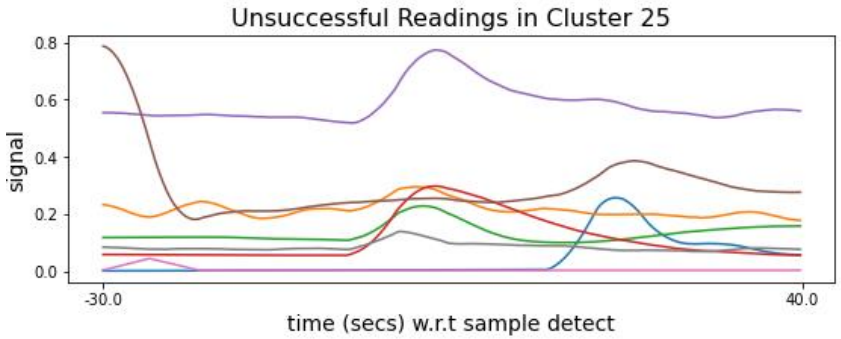
CLUSTERING – ITERATION 2

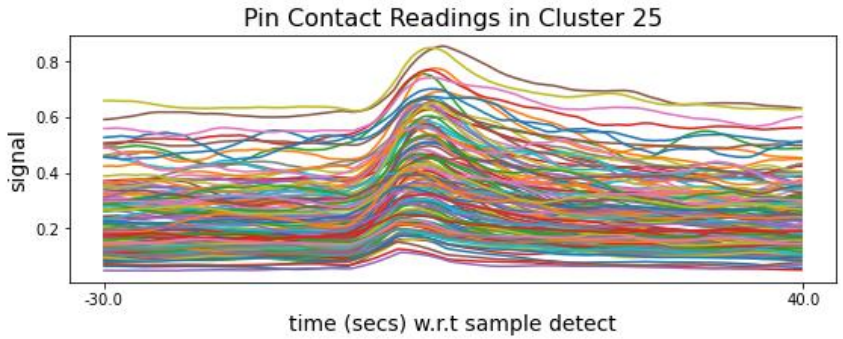


Whole timeseries clustering with KShape

Number of pins: 129

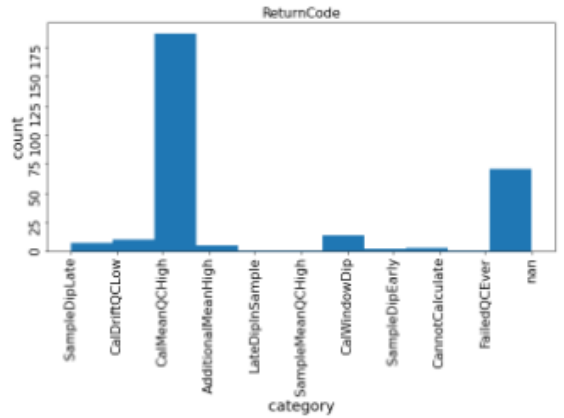
Number of unsuccessful: 8

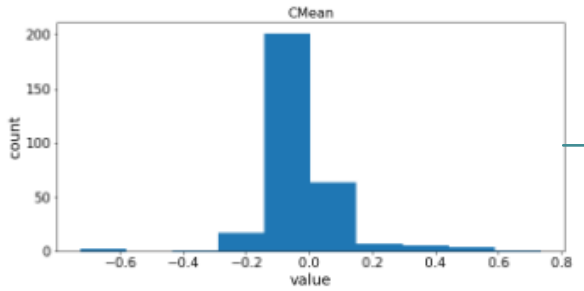


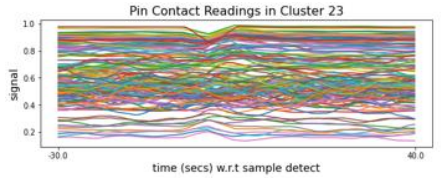


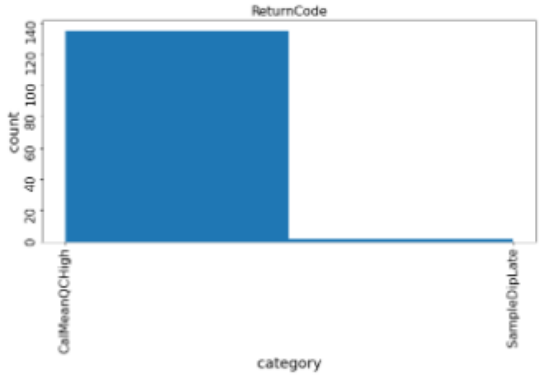
PRELIMINARY/INCREMENTAL RESULTS

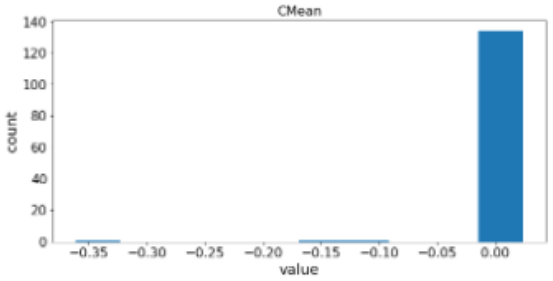
CLUSTERING – ITERATION 2

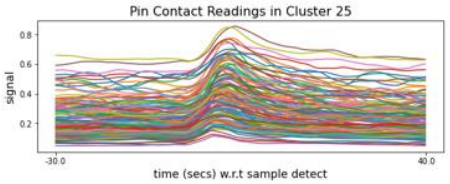






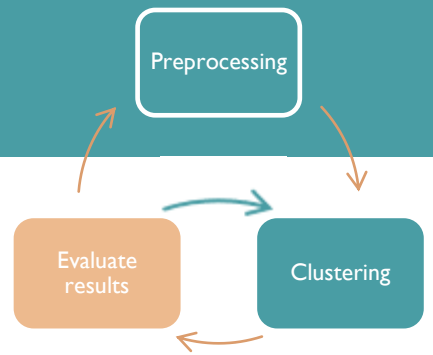




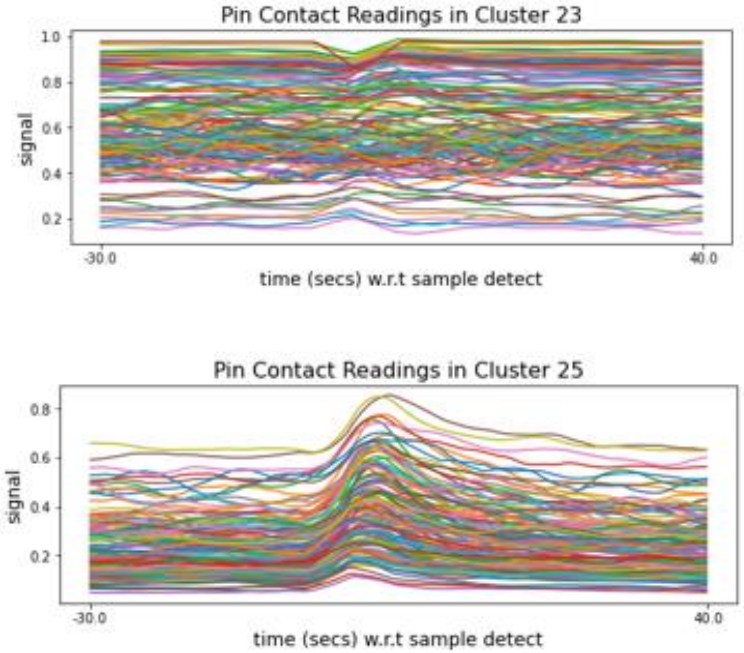


PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2



Whole timeseries clustering with KShape

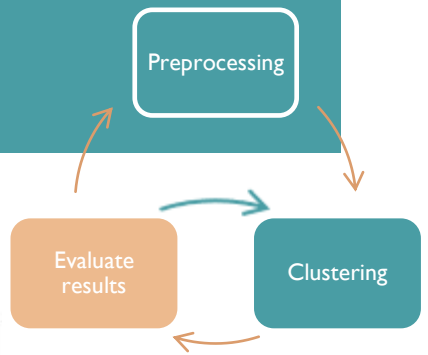


Proportion of pins:

$$\frac{129 + 167}{380} = \frac{296}{380} = 78\%$$

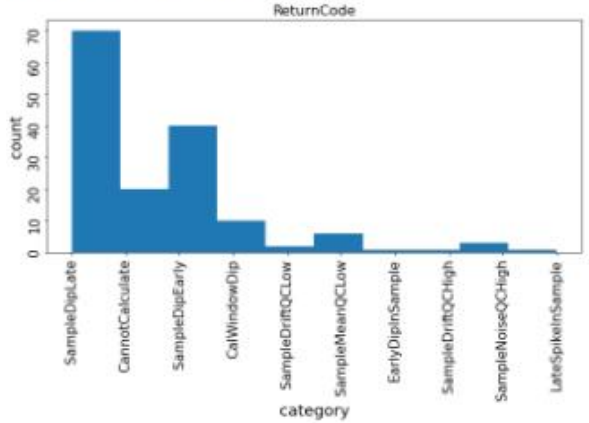
PRELIMINARY/INCREMENTAL RESULTS

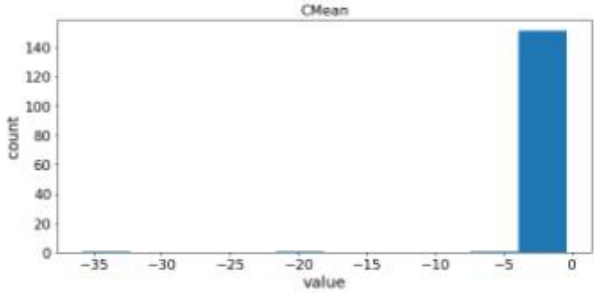
CLUSTERING – ITERATION 2

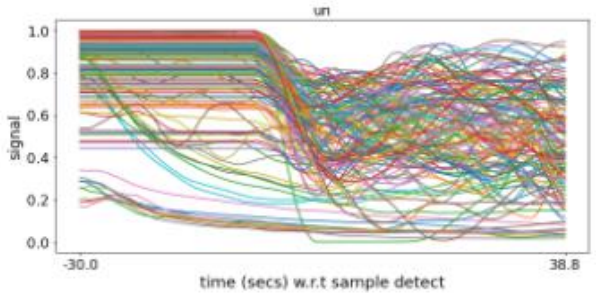


Clusters with only unsuccessful readings:

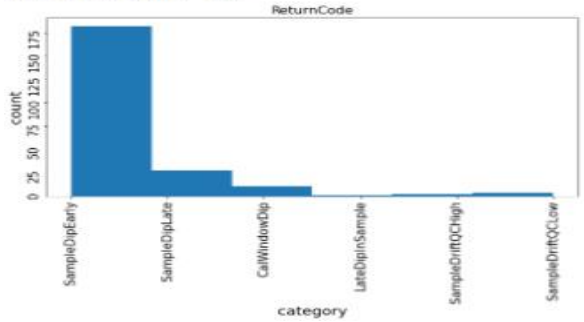
Number of un : 154

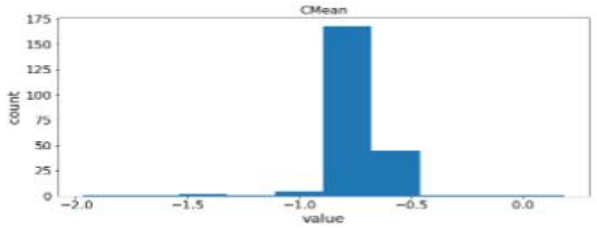


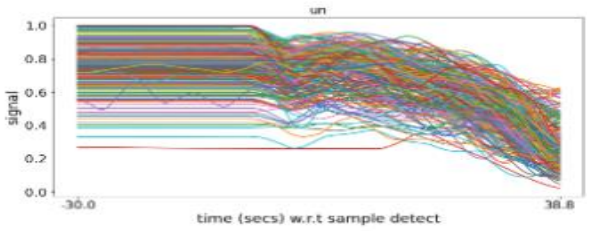




Number of un : 225







PRELIMINARY/INCREMENTAL RESULTS

CLUSTERING – ITERATION 2

Comparison between 2 promising pipelines:

Autoencoder, PCA (Timeseries), Predictors

- More preprocessing steps
- Clusters well less defined
- 1 cluster with 73 % of all pins

Kshape (Timeseries + Predictors)

- Less preprocessing steps
- Clusters are well defined
- 2 clusters with 78 % of all pins

PLANNING AND ACTIONS FOR THE NEXT CYCLE

Sara :

- Clean PCA and Preprocessing notebooks
- Waveform characterization
- Make a glossary to organize and describe code

Saisree :

- Clean the SOM notebook
- Work on cleaning up Github repo and making sure it's up to date
- Documenting the work done so far

Neethu :

- Clean the autoencoder and random forest feature importance notebook.
- Try running K-shape with some autoencoder features as additional predictors.
- Create slides for the next weekly cycle
- **Justine :**
- Clean the KShape notebook
- Work on characterizing the clusters using density plots for the predictors
- Upload our preprocessing csv files to the client's cloud
- Change all the directories in the notebooks so that they match the ones in the client's cloud

Team

- Create slides for and present work in meetings (split equally)
- Code review circle

ROADBLOCKS

- Make our code as reusable by the client as possible. This includes making a glossary of our various attempts.

DEVIATION FROM THE ORIGINAL PLAN/SCHEDULE

ACCOMPLISHED ALL LAST WEEK'S TASKS?

- Last week's tasks:

Sara

- Look into different ways of windowing and standardizing.
- Attempt whole time-series clustering/visualization with new windows.
- Dig into Fourier and other transforms more.
- Record meeting minutes for Siemens check-in on Tuesday.

Saisree

- Try other clustering algorithms on shape-based TS(DTW-SOM applied to different windows)
- Attempt feature extraction using 1-D CNN if the shape-based approach is not promising
- Record meeting minutes for Siemens check-in on Friday.

Neethu

- Using auto-encoder to extract features.
- Use the extracted features for clustering.
- Use LSTM for anomaly detection
- Record meeting minutes for Advisory committee on Tuesday.

Justine

- Do more research on anomaly detection in timeseries
- Record meeting minutes for check-in meeting with Capstone advisors.

➡ Started focussing on removing noise by comparison with standard waveform

DEVIATION FROM THE ORIGINAL PLAN/SCHEDULE

AHEAD/BEHIND/ON TRACK?

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.
June 6 - June 12	Tuning	• Focus on improving the most promising model(s). • If we have time, maybe look into data augmentation methods.
June 13 - June 19	Documenting	• Write the final report. • Generate slides and practice for the final presentation.

SUMMARY OF INTERACTIONS WITH THE CLIENT

- Exchanged a few emails throughout the week
  - The test ids
- Meeting on Monday, May 30th
  - Presented our findings in a data review meeting
- Meeting on Wednesday, June 1st
  - Discussed our plan for the week
  - Were confirmed that some of the unsuccessful readings in our pin contact cluster had been mislabelled
- Meeting on Friday , June 3rd
  - Presented them with the progress we made over the week and got feedback

SUMMARY OF INDIVIDUAL AND TEAM EFFORTS MAY 30 – JUNE 3

- Sara :
  - Meetings/Presentations: 11 hours
  - Coding/Debugging/Documenting: 31 hours
  - Administrative work: 1 hour
  - **Total : ind. + team = 43**

- Neethu :
  - Feature extraction and modelling : 22 hrs
  - Researching : 2 hrs
  - Administrative work : 2.5
  - **Total : ind. + team = 37.5**

Saisree :

- Researching : 7 hrs
- Clustering : 17 hrs
- Debugging/ Documenting: 13.5 hrs
- Administrative work: 2.5 hrs
- **Total : ind. + team : 39 hrs**

- Justine:
  - Coding : 24 hours
  - Presentations/Meetings : 18.25 hours
  - Research : 7.5 hours
  - **Total : ind. + team = 49.75 hours**

Team:

- **Time spent in meetings :** 11 hours

NEXT STEPS

1

Make more diagnostic plots from the predictor file

2

Clean and document our code thoroughly

3

Produce our final report detailing our design decisions

QUESTIONS OR FEEDBACK?

THANK YOU FOR YOUR TIME!