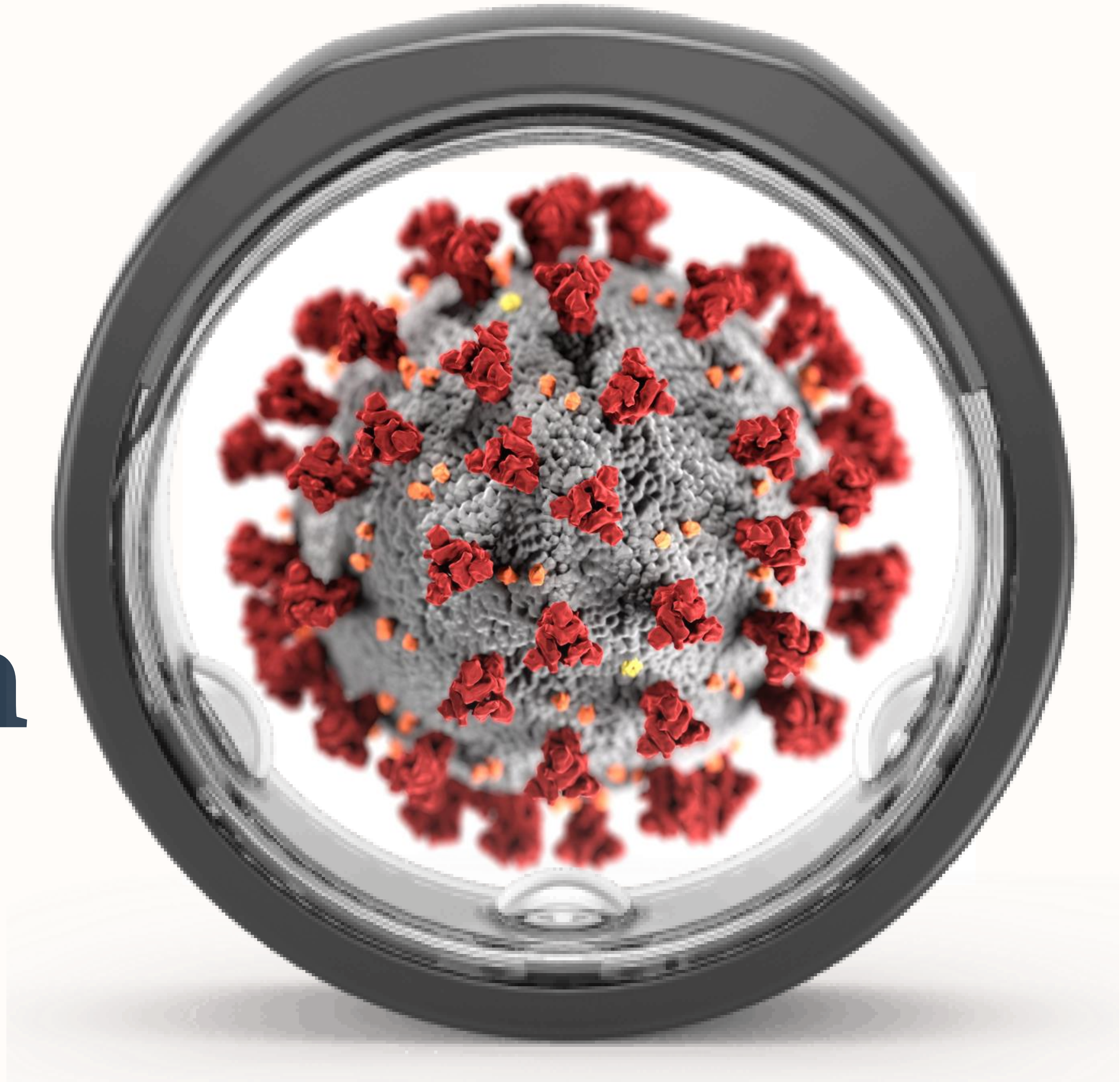
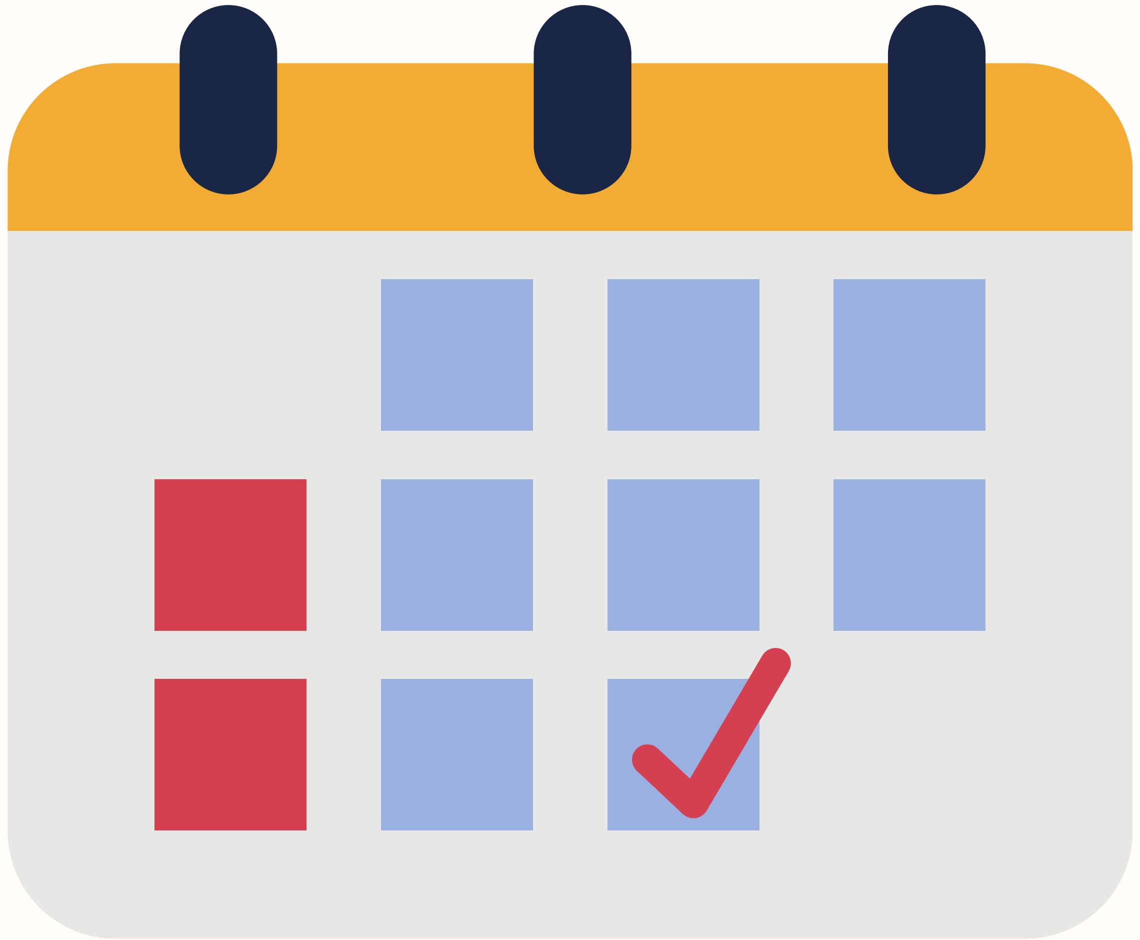
IoT Wearable Data-
Fever Analysis &
COVID Onset Detection



Wearables - Vital role to predict vital Information

Agenda

- Team
- Project Overview
- Solution Architecture
- EDA (Exploratory Data Analysis)
- Data Preparation
- Modeling & Evaluation
- Scalability
- Visualization
- Video - Demo
- Key Findings
- Future Use



Team

Swetha Varadharajan

Solution Architect

Yogesh Bansal

Model Expert

Sasi Mahalingam

Visualization Architect

Venu Mamidala

ML Expert

Raj Krishnan

Data Specialist

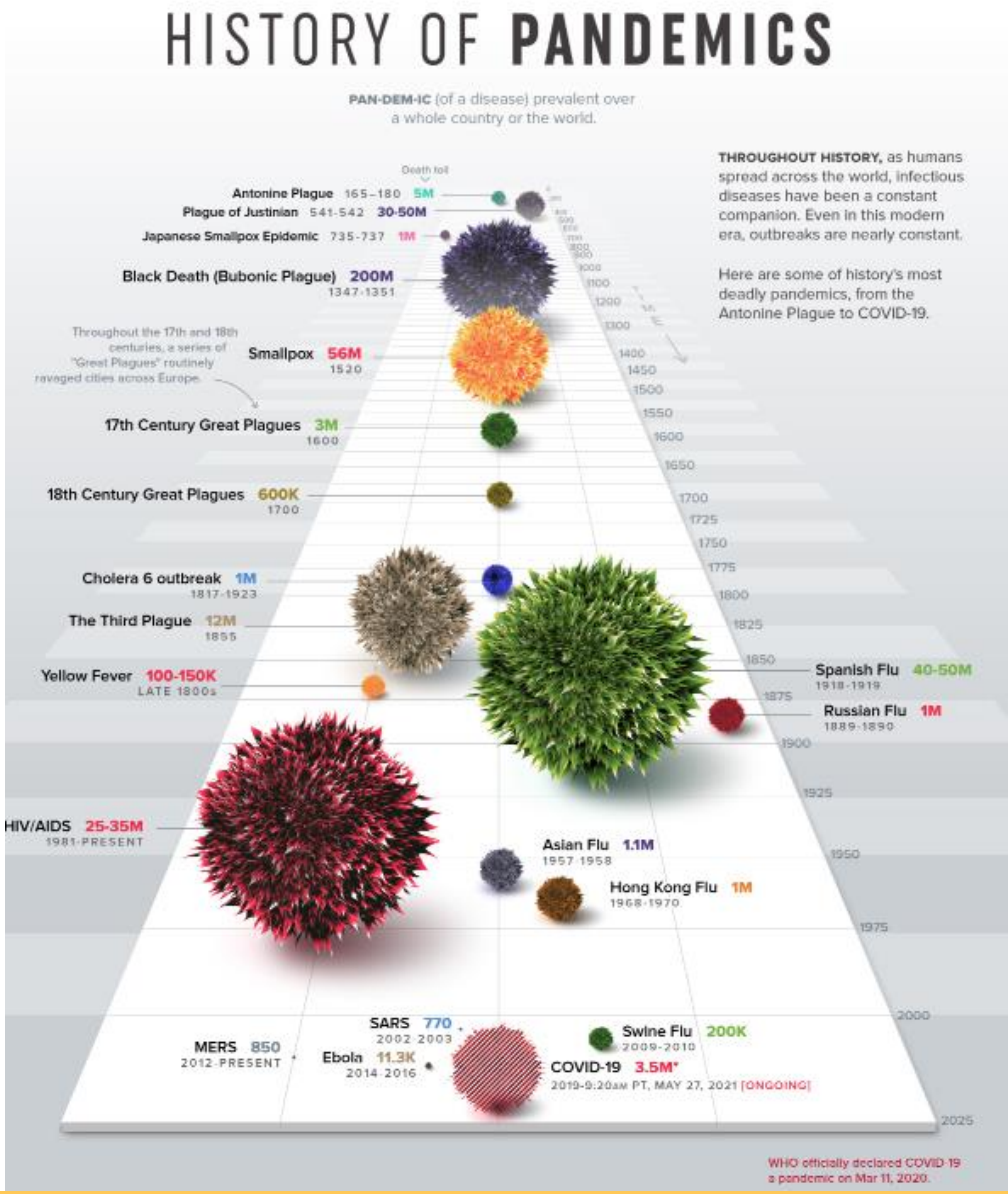
Prof. Benjamin Smarr

Advisor

Project Overview

Problem Statement

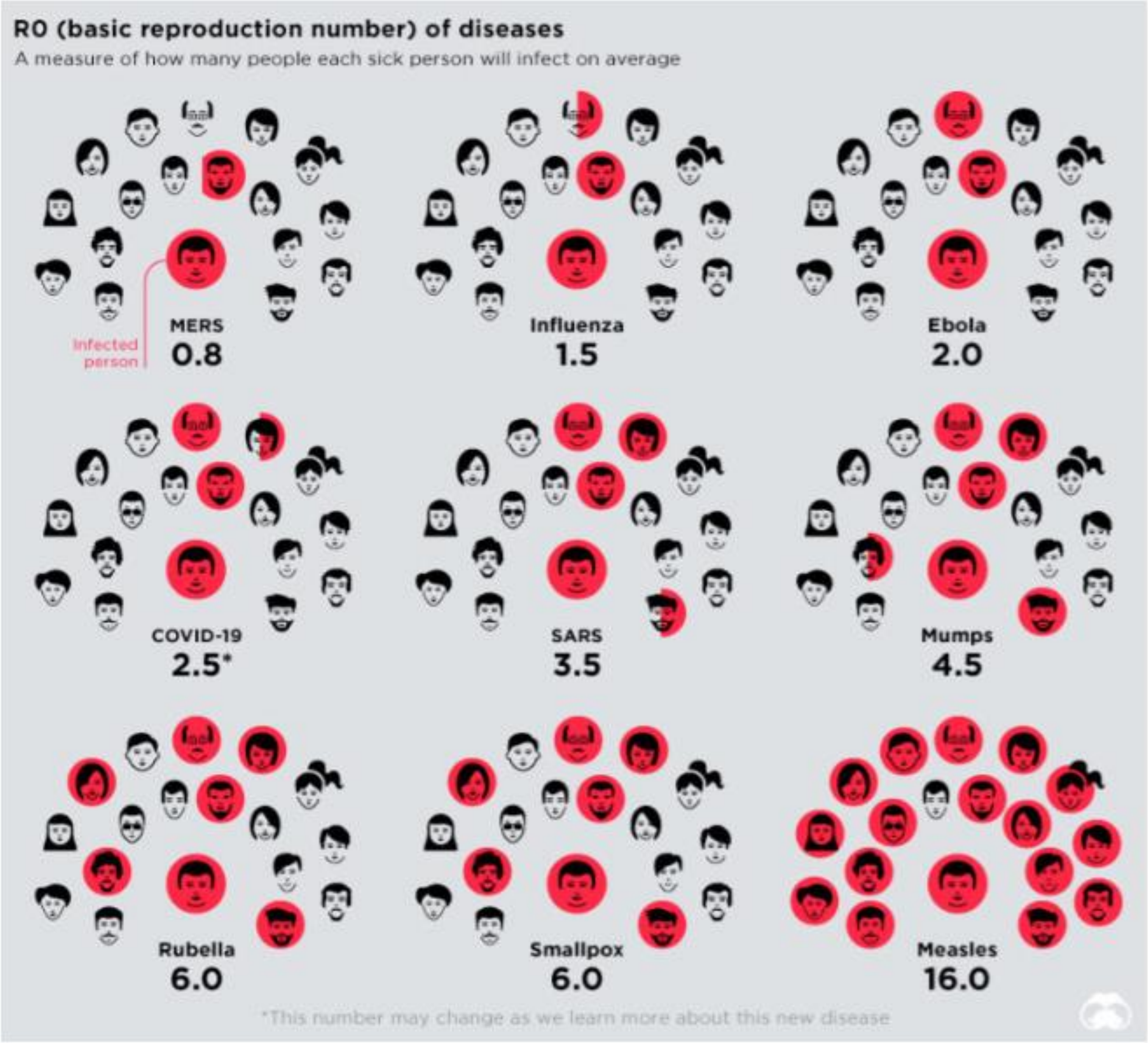
Pandemics are not new to mankind



Source: https://www.visualcapitalist.com/

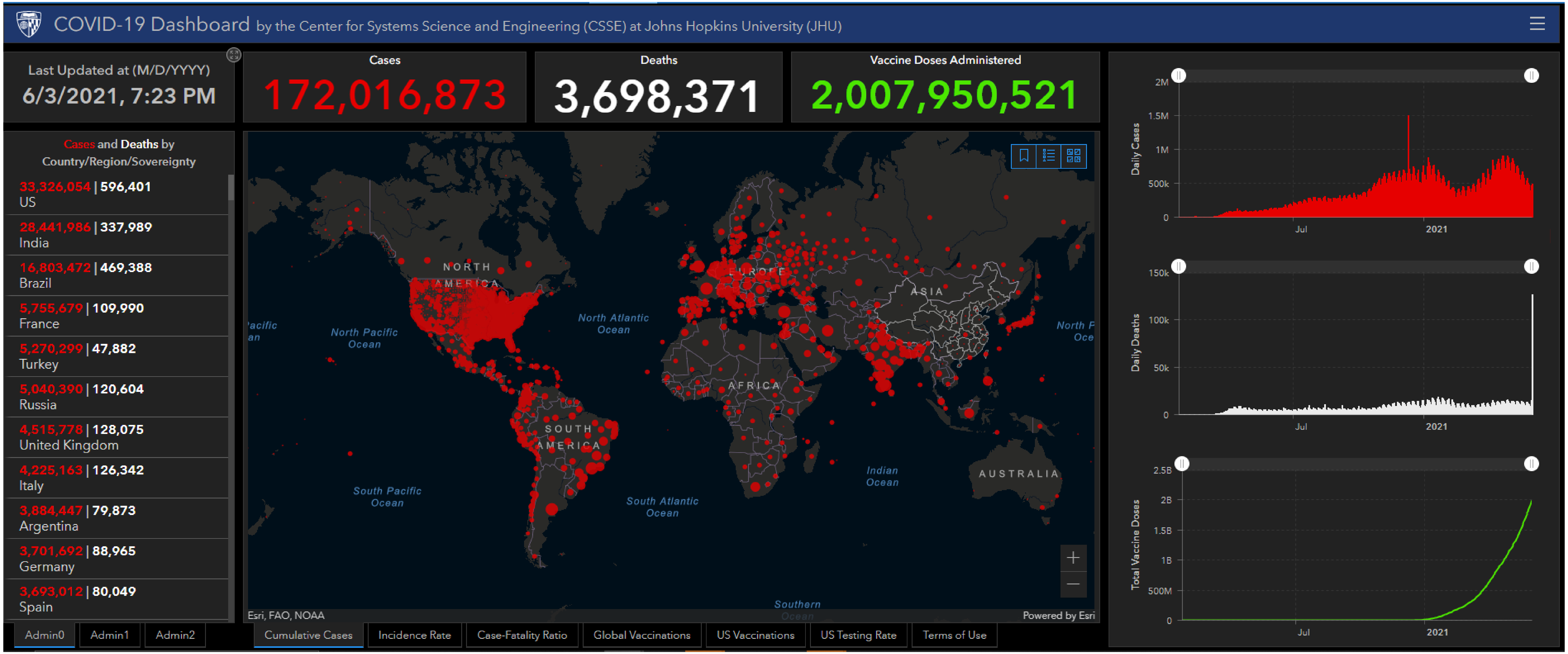
Problem Statement

Everyone has the responsibility to control the spread



Source: https://www.visualcapitalist.com/

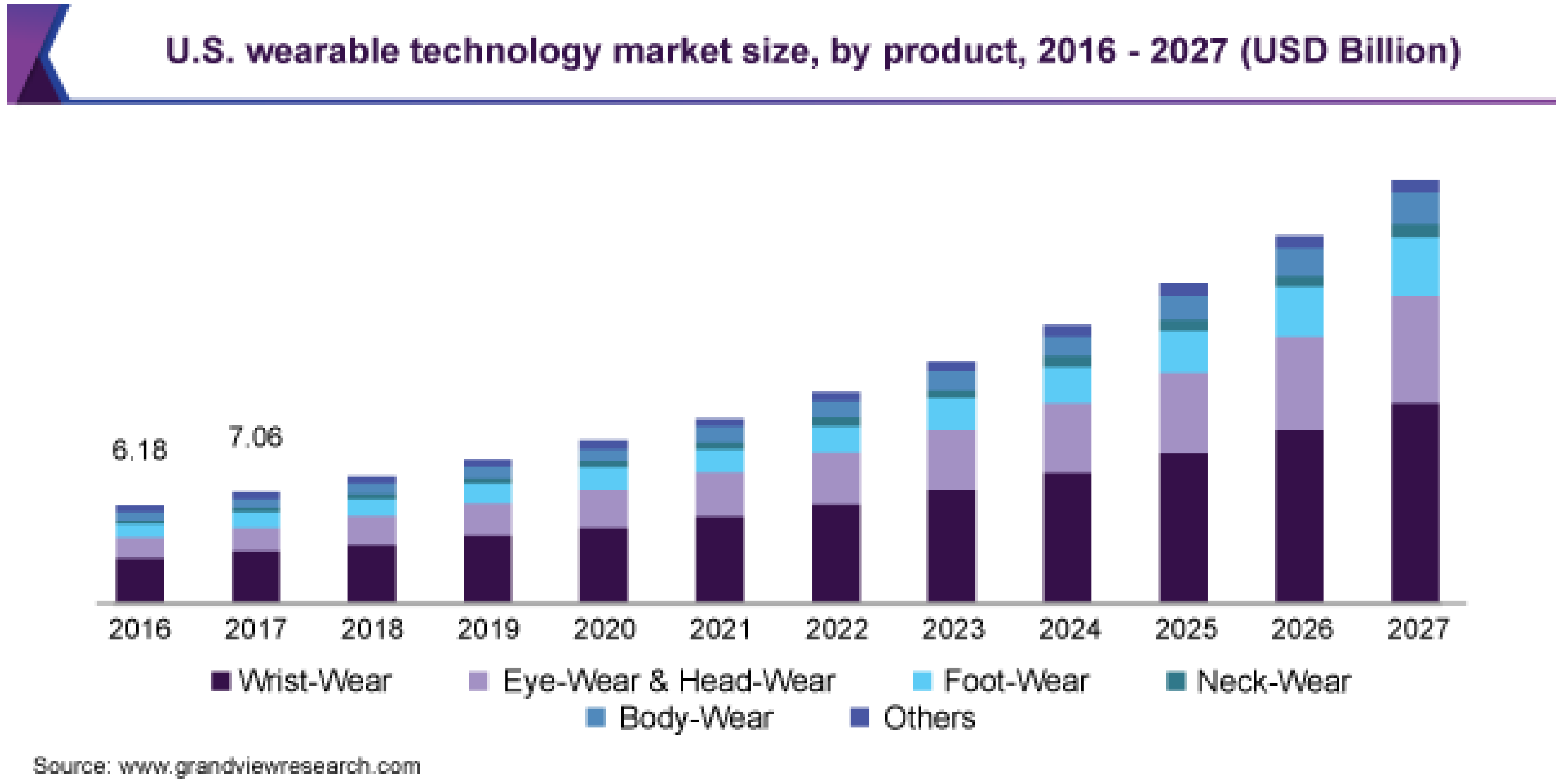
Corona Virus Status



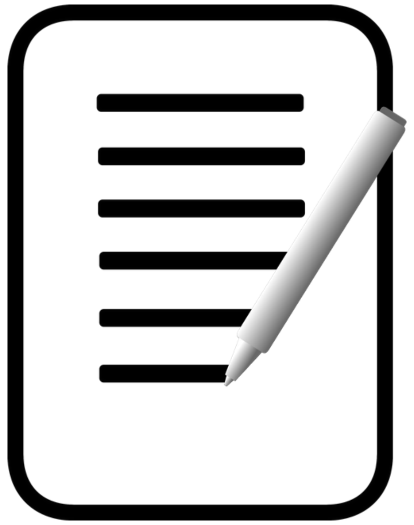
Source:https://coronavirus.jhu.edu/map.html

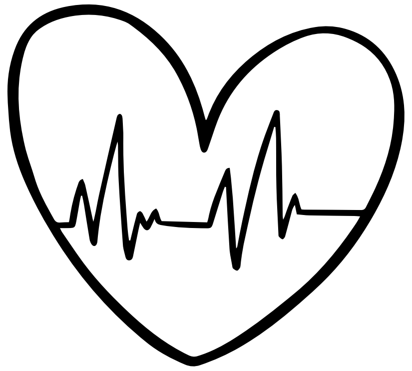
Data Resources

- Pervasive Wearables Industry



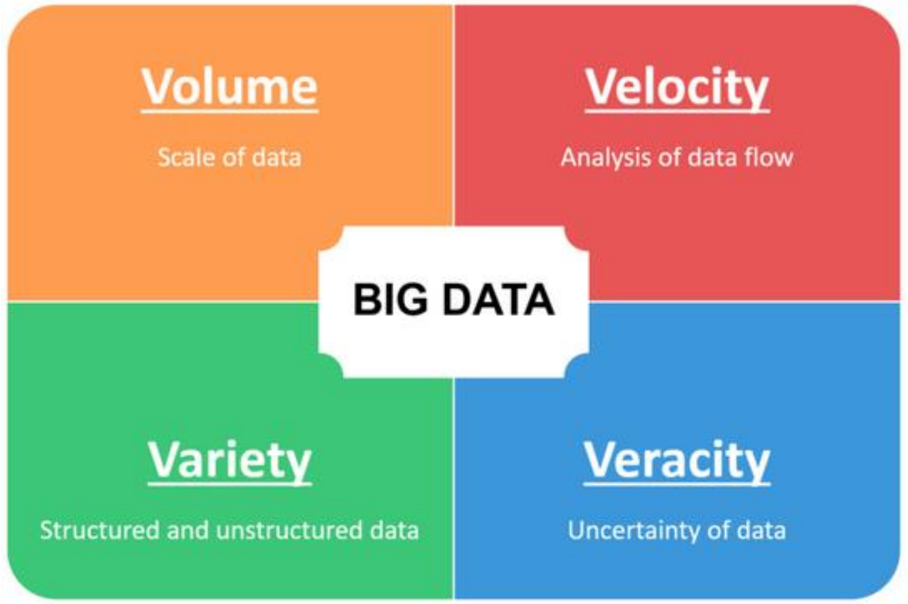
- Physiological Signals from Oura Ring





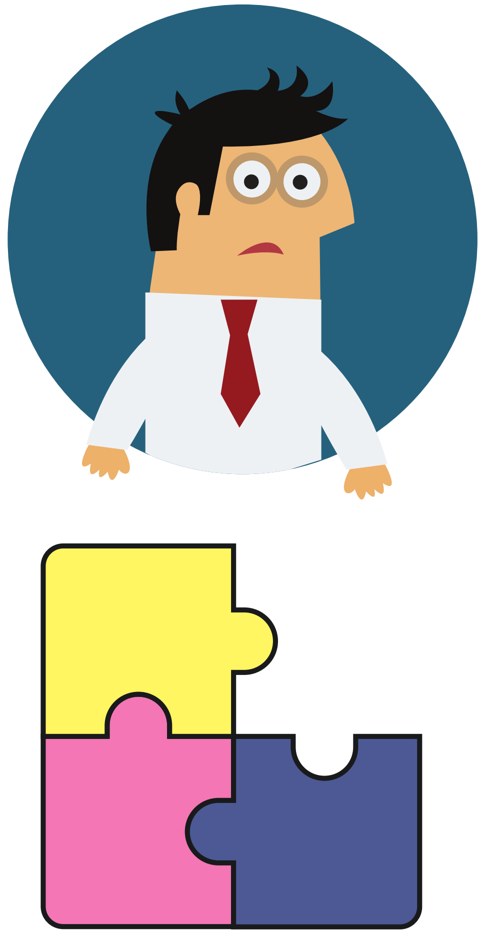
Challenges

Missing Architecture

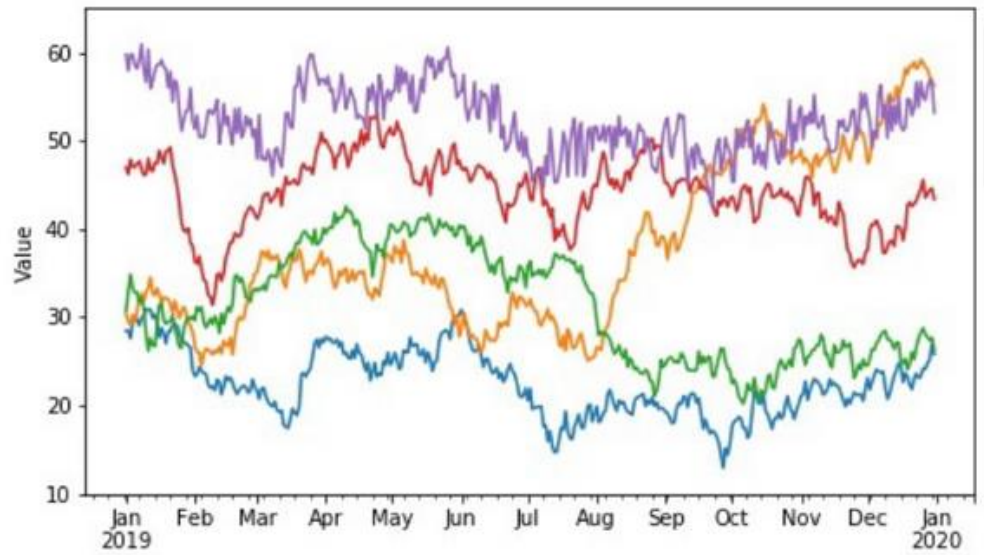


Source: https://chartio.com/

Big Data Management



Missing Piece

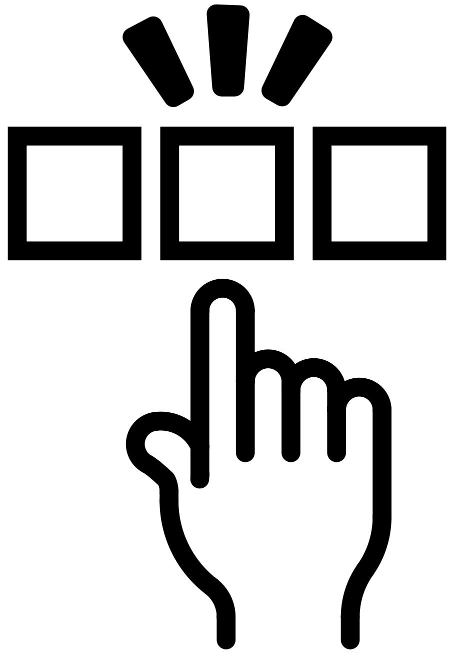


Source: https://www.kdnuggets.com/

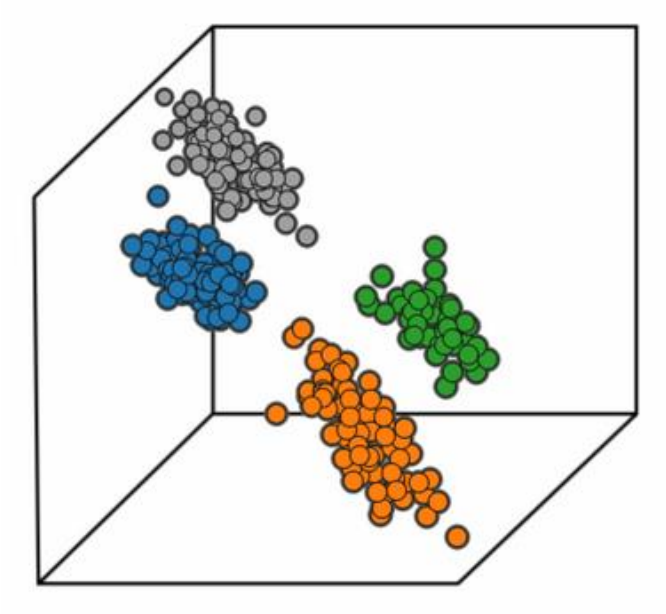
Time Series Data

Opportunities

Architecture that enables rapid exploration at scale



Feature ranking & selection



Signal based Clustering

Dynamic healthy baseline

Covid Onset detection

Customers

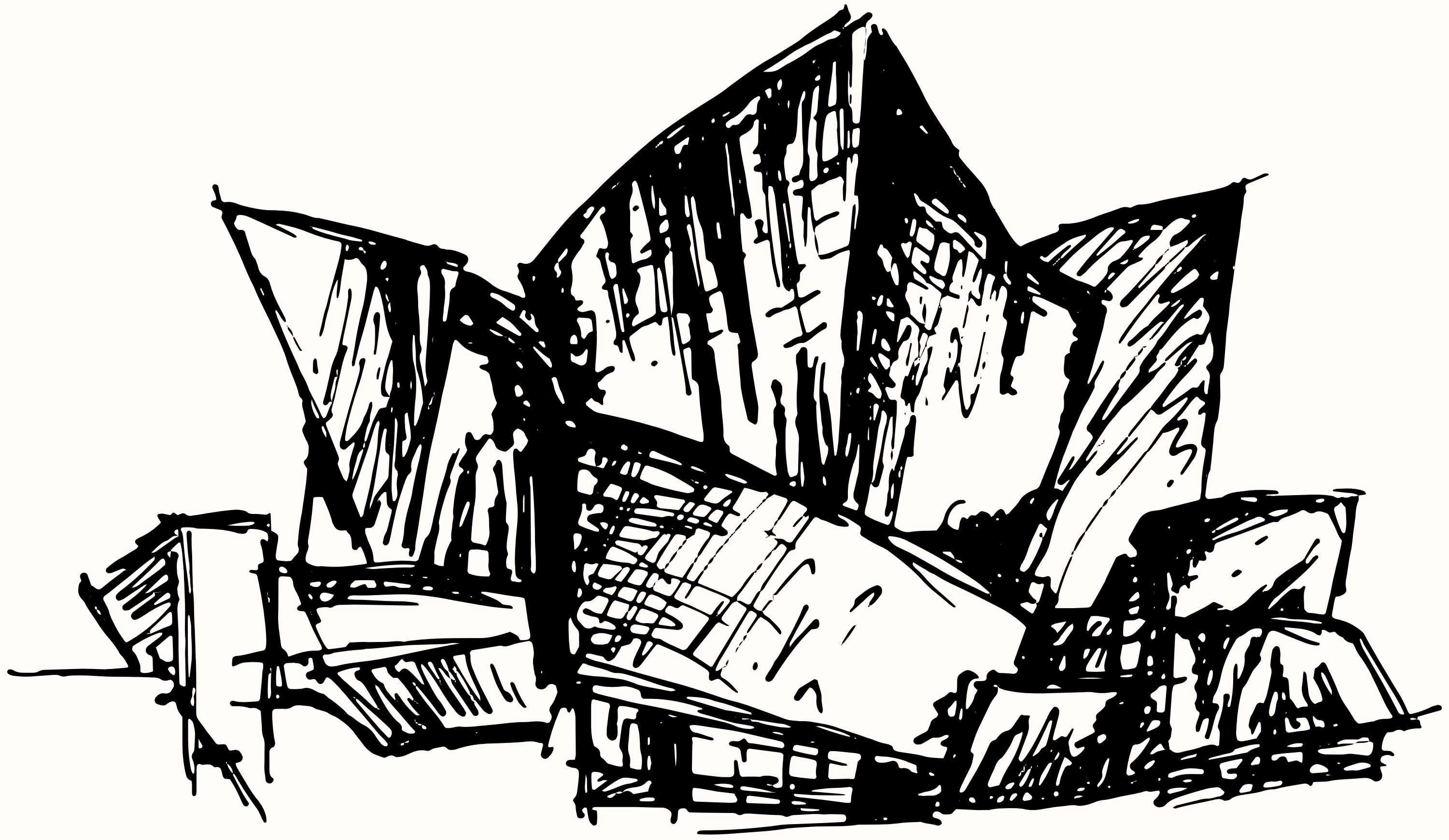
- Covid and Other Medical Researchers.
  i.e. Tempredict

- Wearable Health Tracker users.

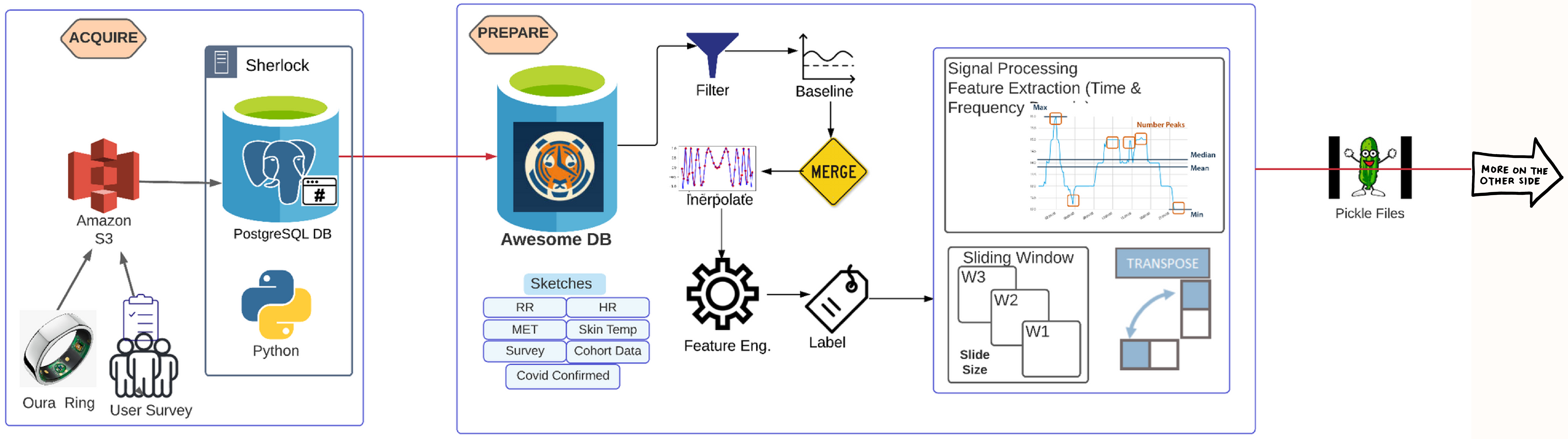
- Data Teams working on harnessing physiological data.

- Domain Experts, Clinicians - analyzing health sensor data.

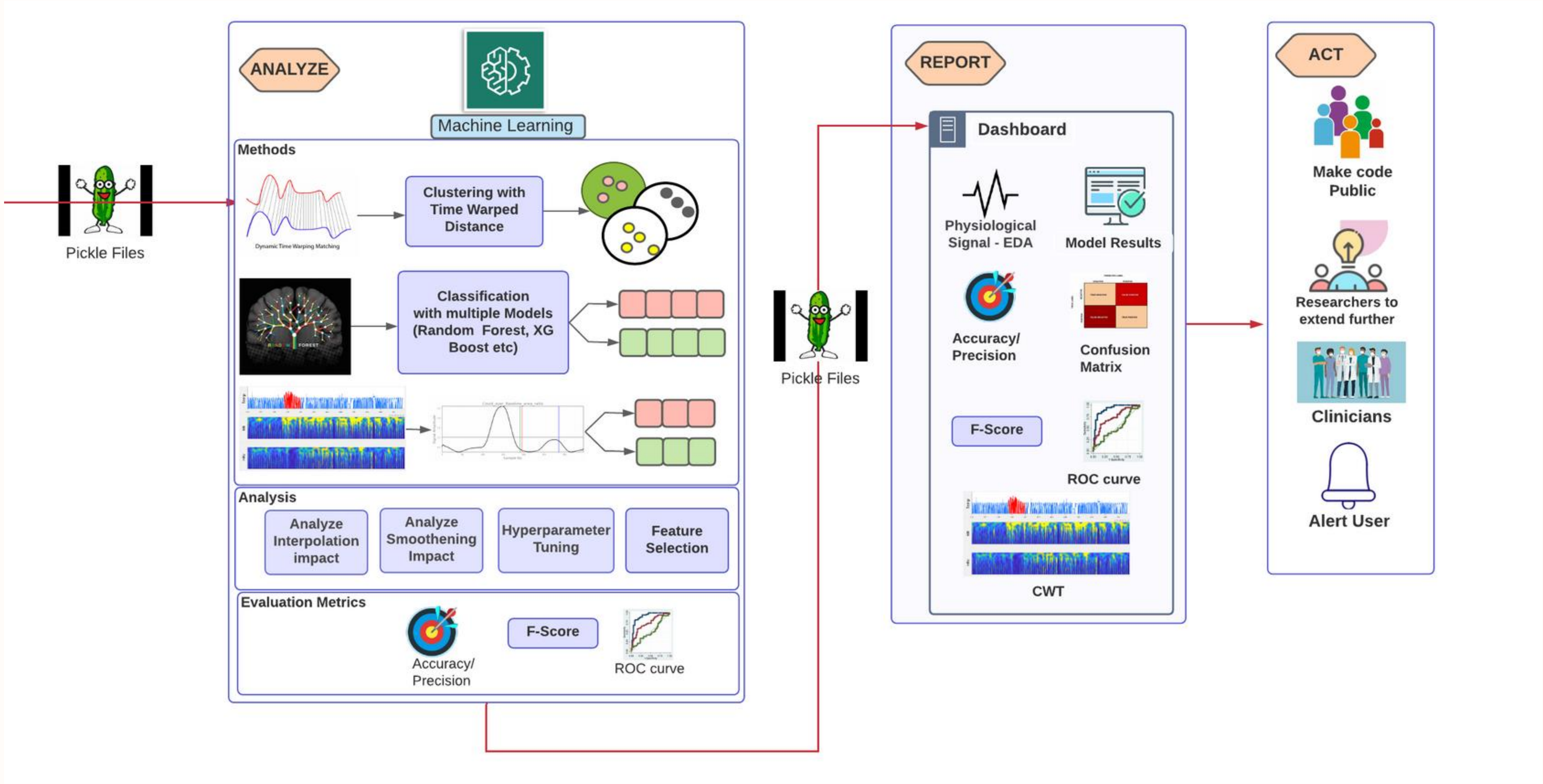
Solution Architecture



Solution Architecture



Solution Architecture



EDA

Data Sources

- **Physiological Data from Oura ring**

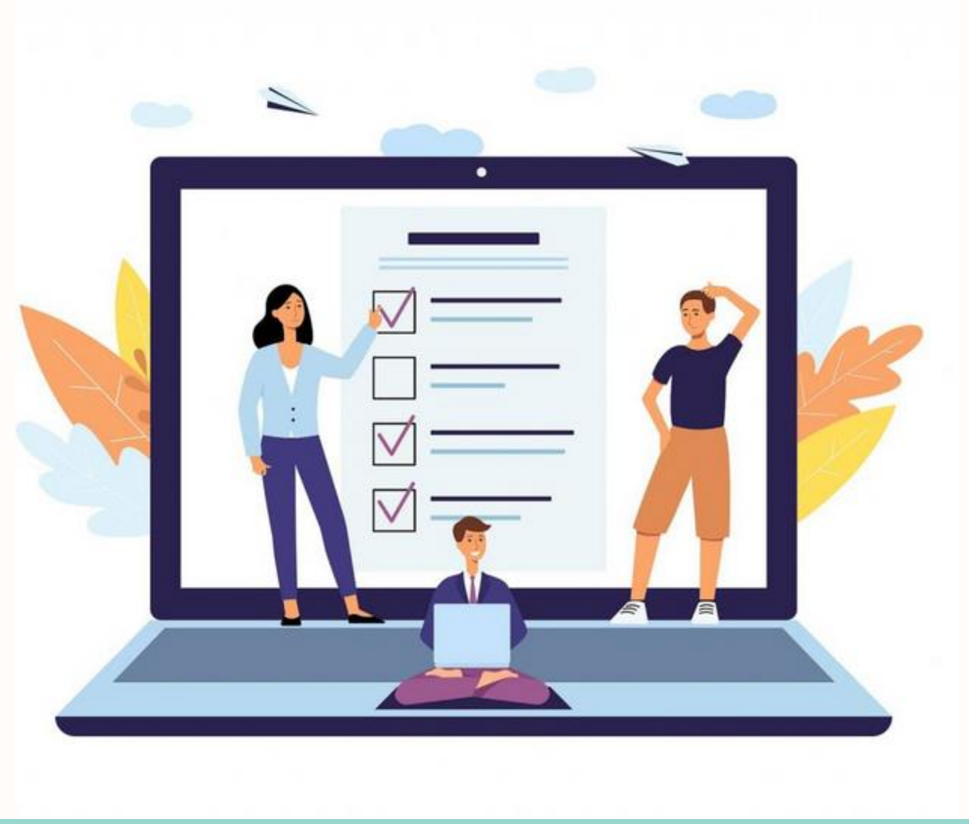
  [Every minute]

  - Skin Temperature
  - Respiratory Rate (RR)
  - Heart Rate - HR (IBI -InterBeat Interval)
  - Metabolic Equivalent for Task - MET
  - Sleep & Wake Pattern

- **Survey Data from Users**

  [Onboarding/Daily & Monthly]

  - Demography, comorbidities, Prescribed medications
  - Symptoms [Dry Cough/Shortness of breath/Headache]
  - Covid Test information (type and date)
  - Other infections (flu, common cold, etc)



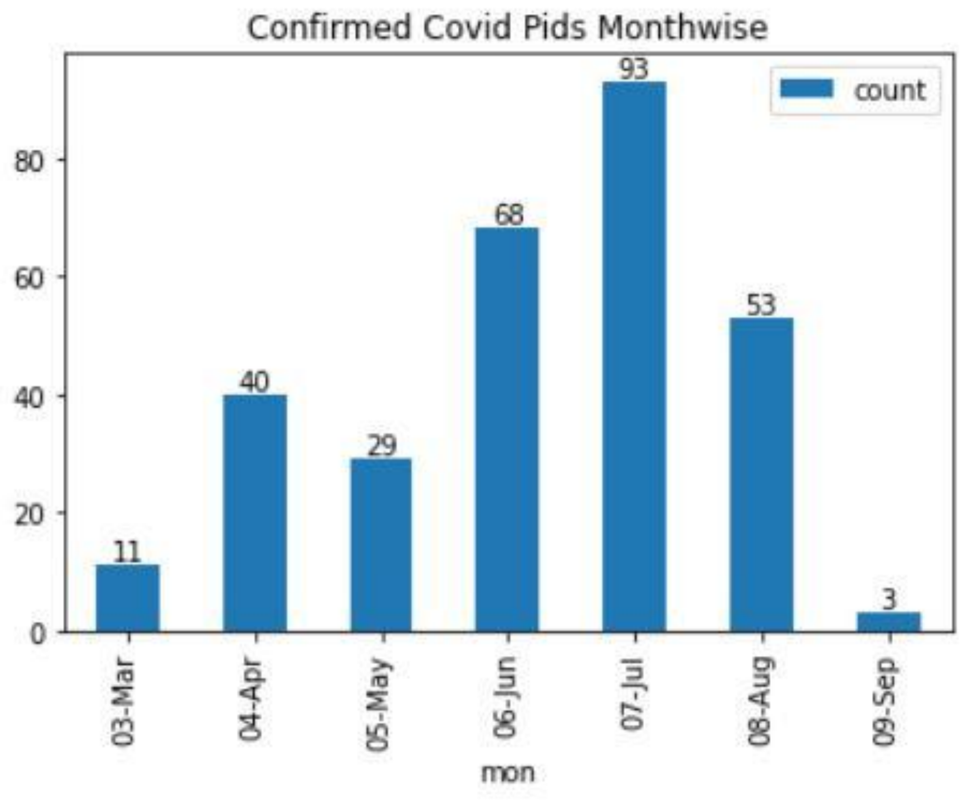
Data Statistics

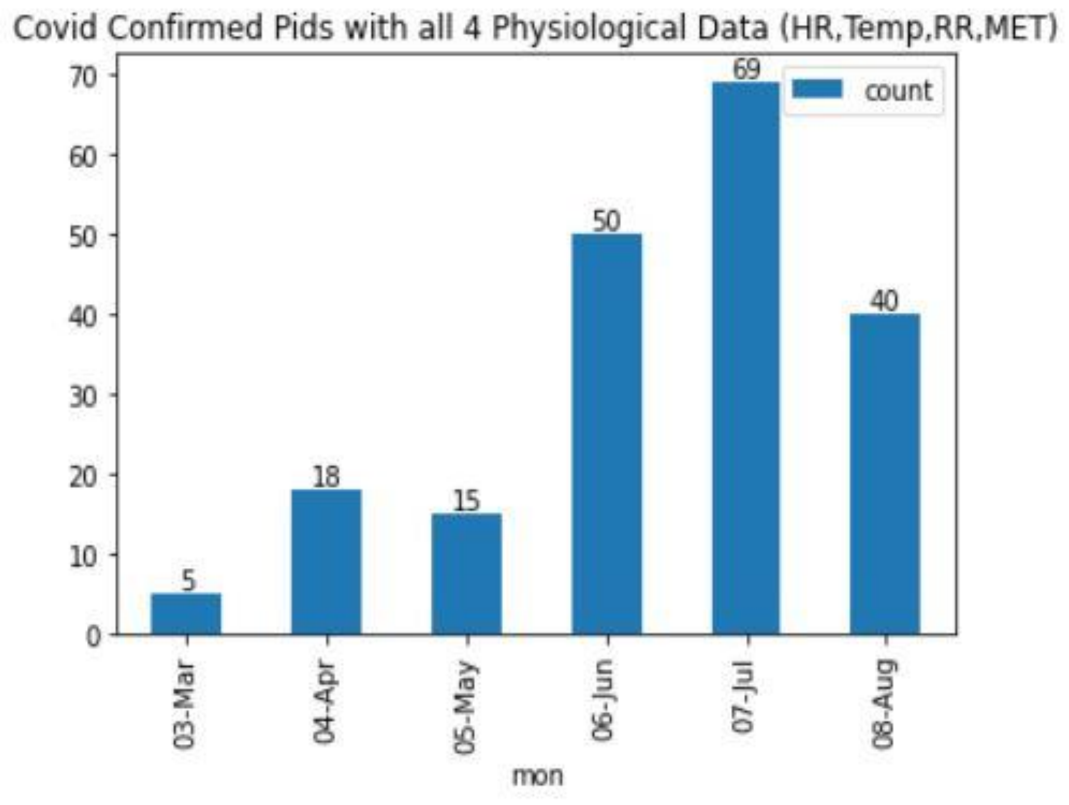
- Survey Data:
  3 Million surveys from 64K persons

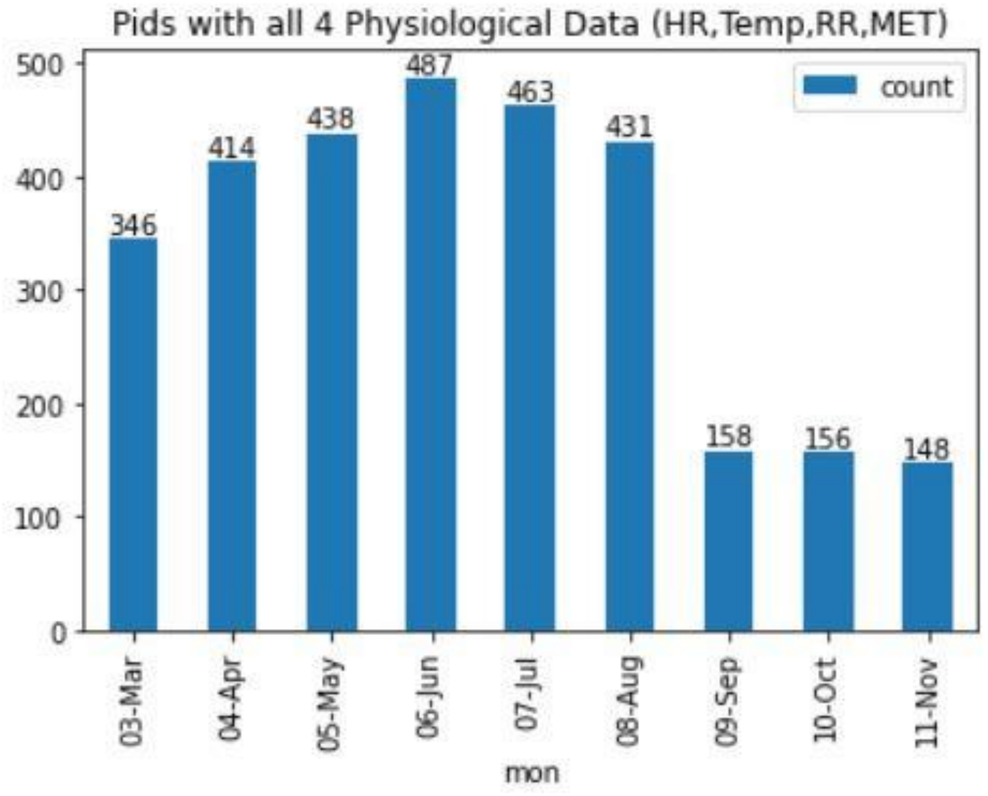
- Device data from Oura ring:
  Minute by minute raw physiological data for 10 months

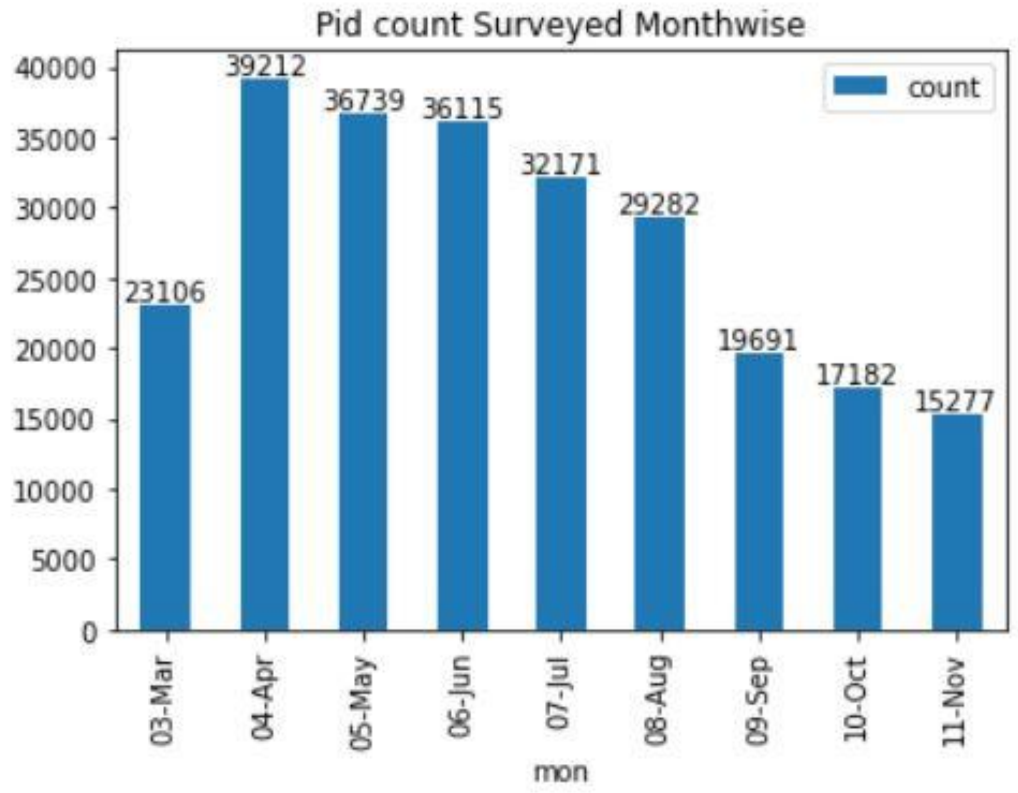
- PCR Confirmed COVID Cases:
  295 Individuals

- Golden Set - 147:
  Based on data availability and signal rhythmicity





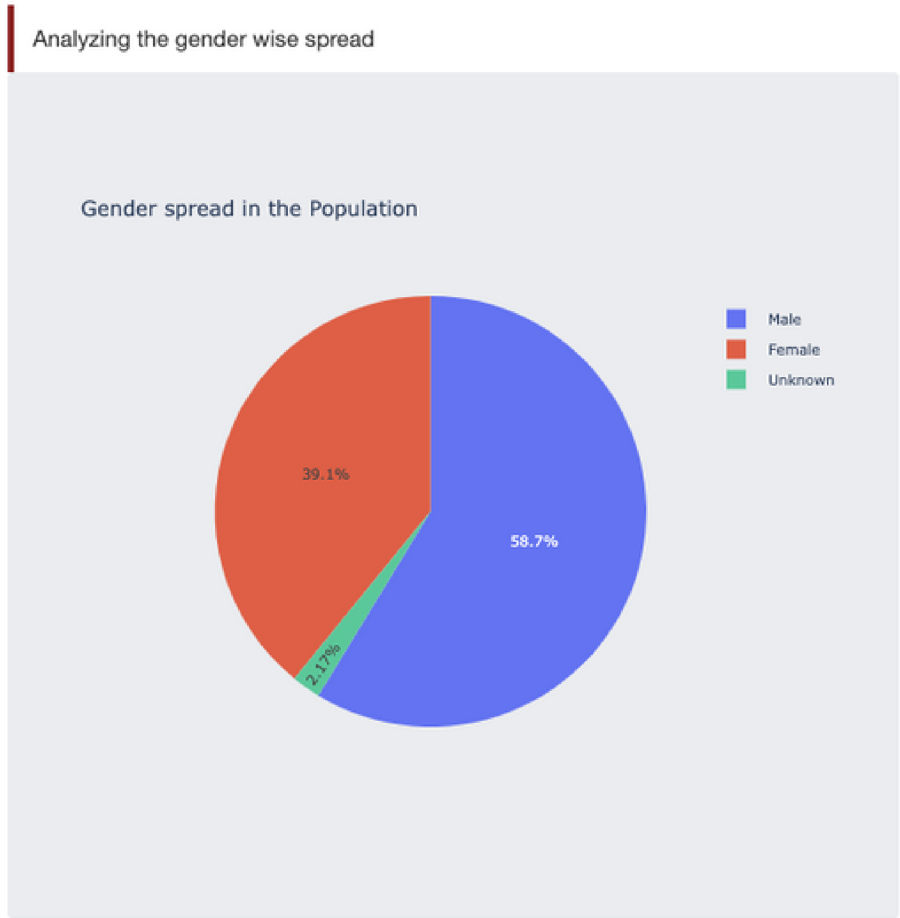


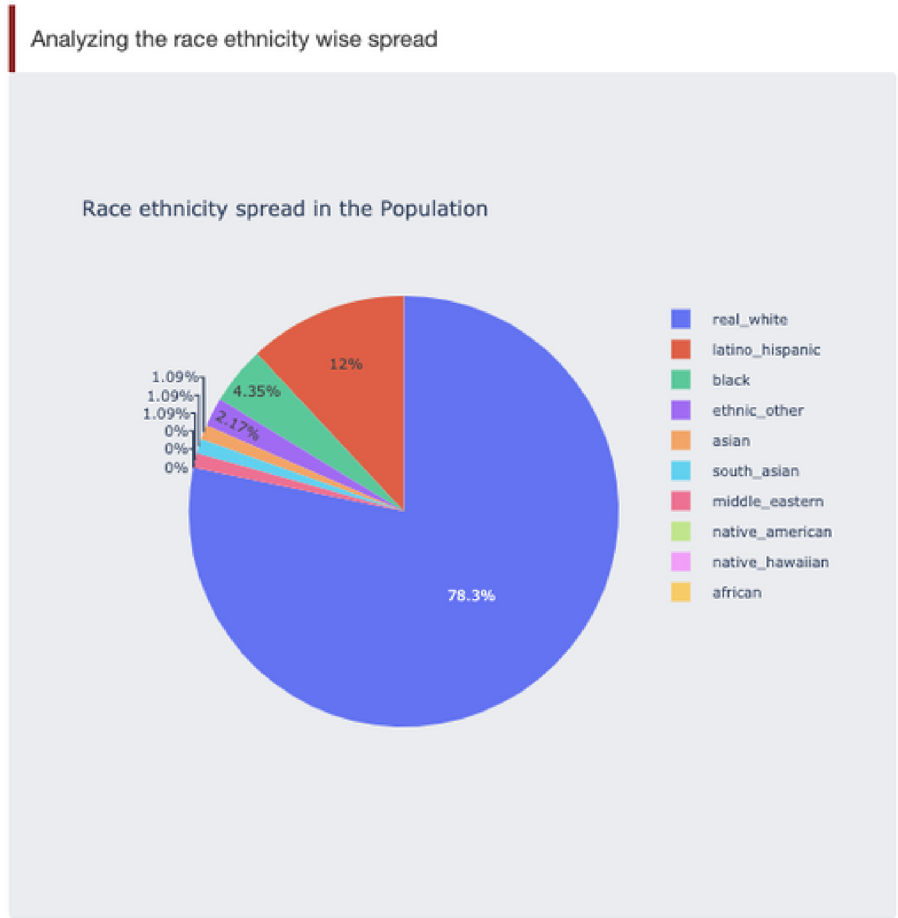


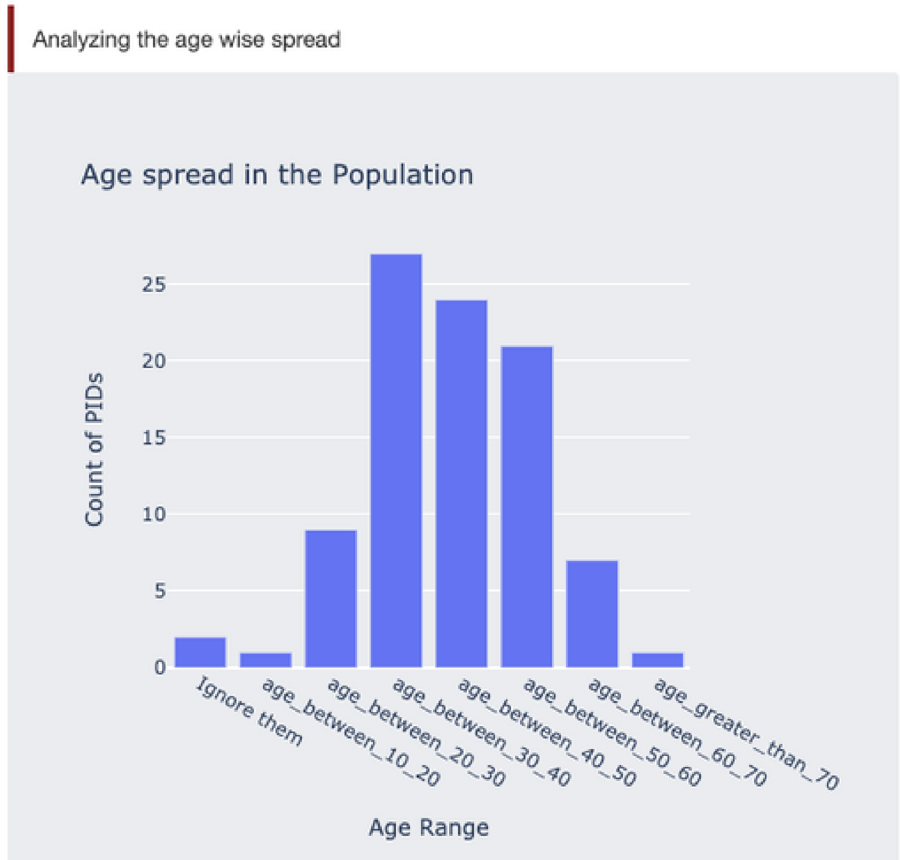
Data Statistics

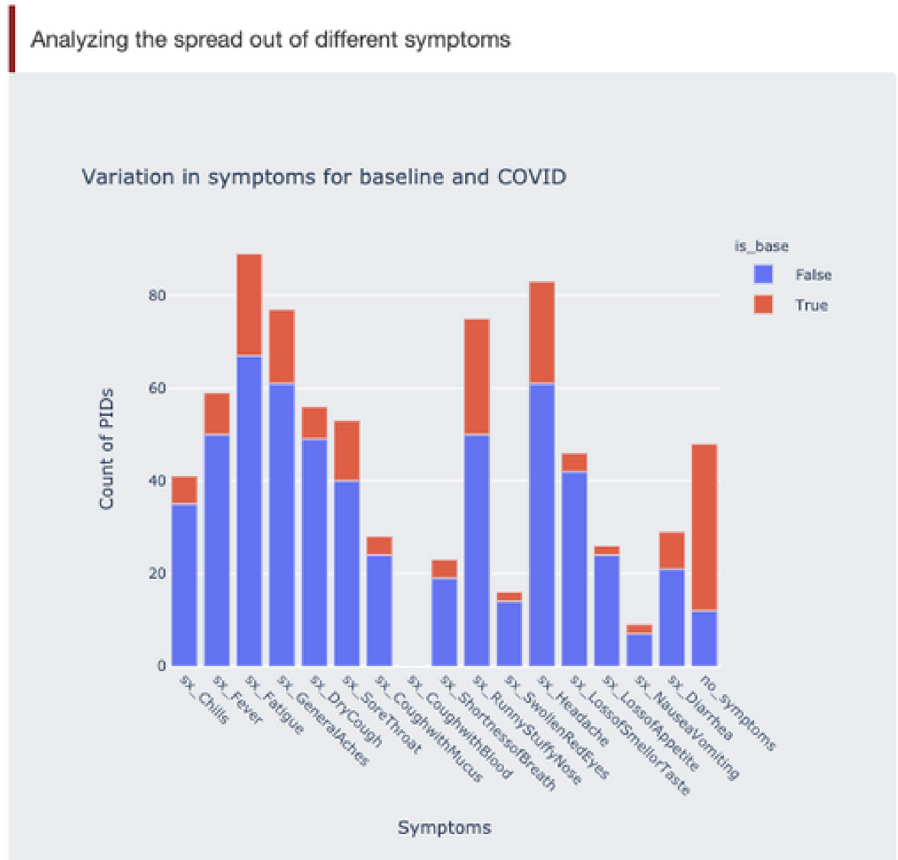
Demography

- More Male Individuals than Female
- Majority Age range between 20 and 50
- Majority from the white population
- Presence of symptoms in both Covid and Baseline window









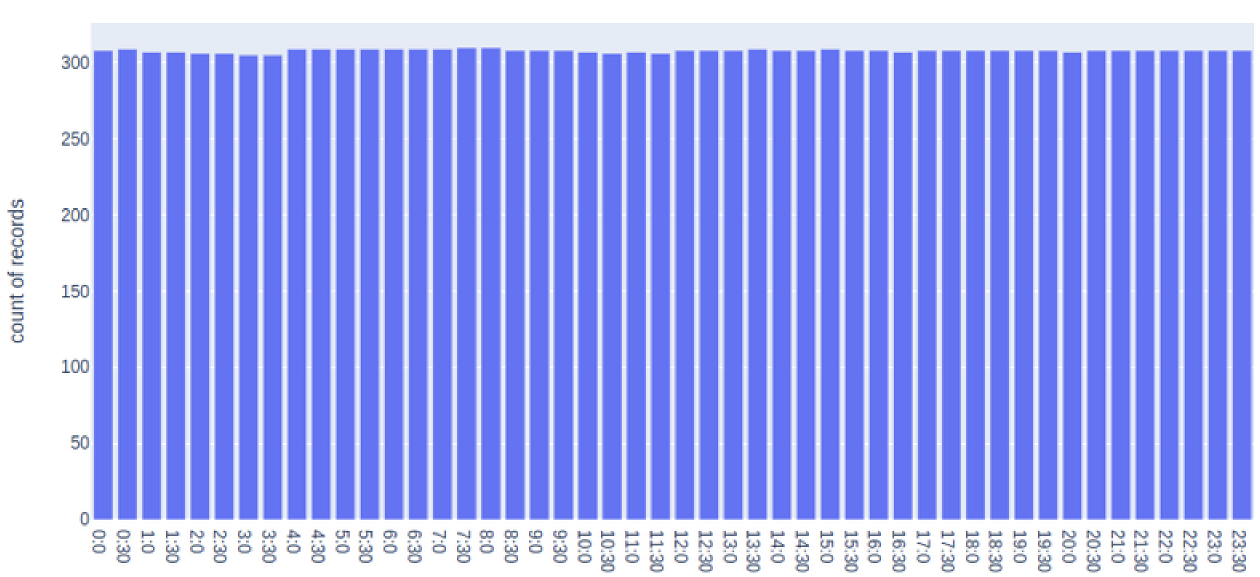
EDA

Data Availability of an Individual

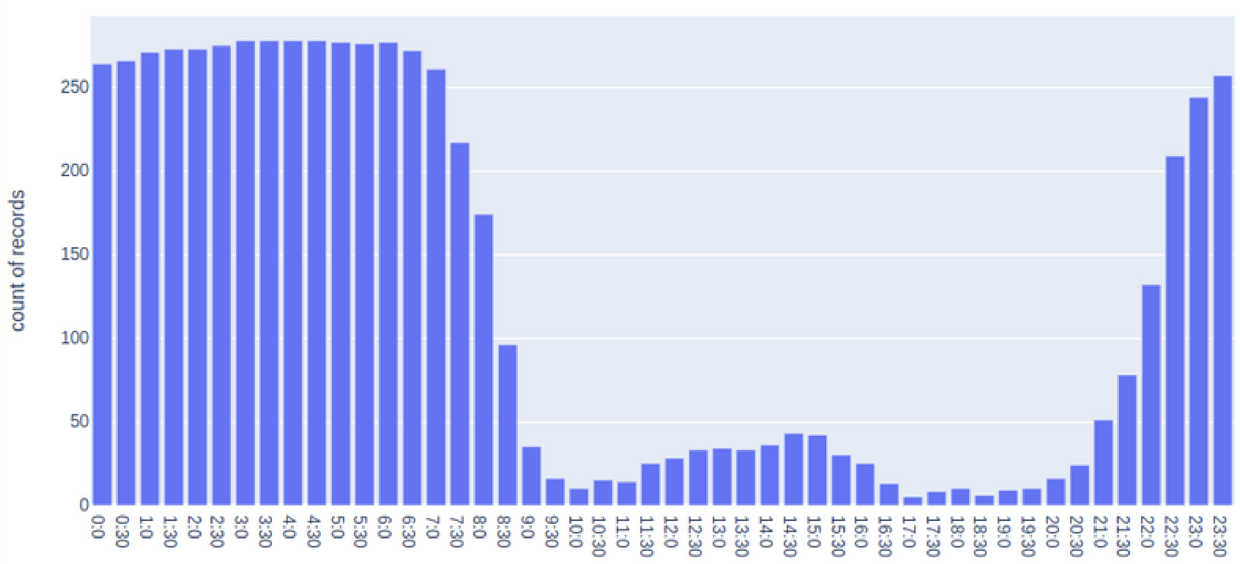
- MET/ Temperature -
  24 hrs Data available

- RR/HR -
  Data mostly available only during sleep
  due to memory limitation in the device

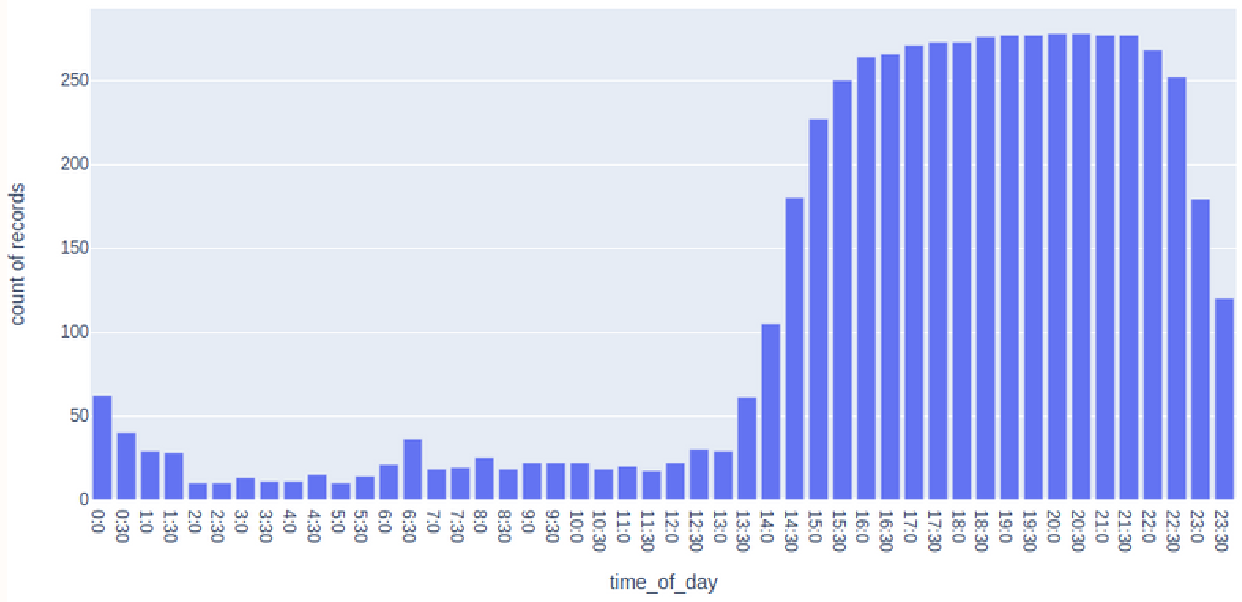
MET /
SKIN_TEMP



HR/RR
Day time
person

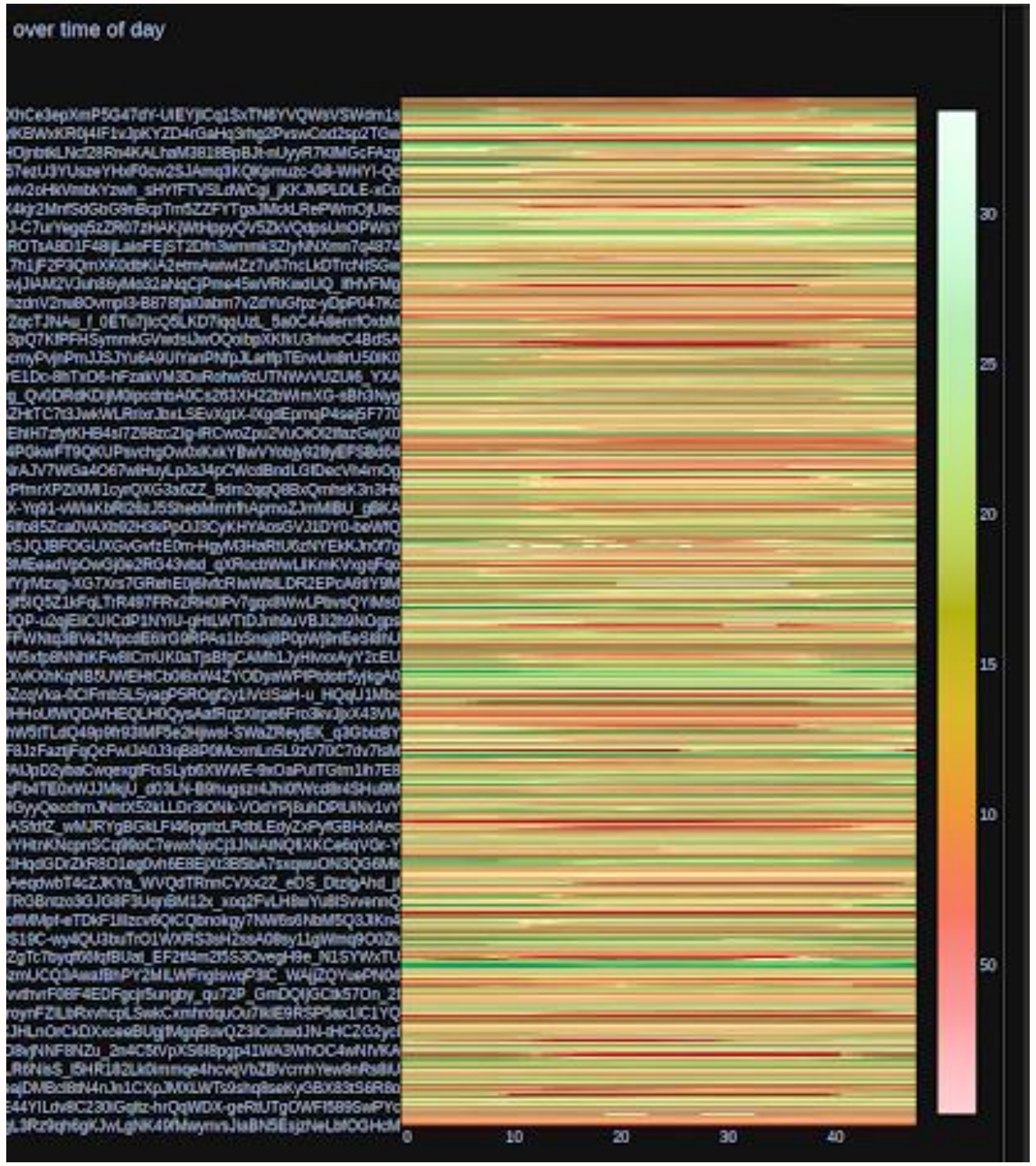


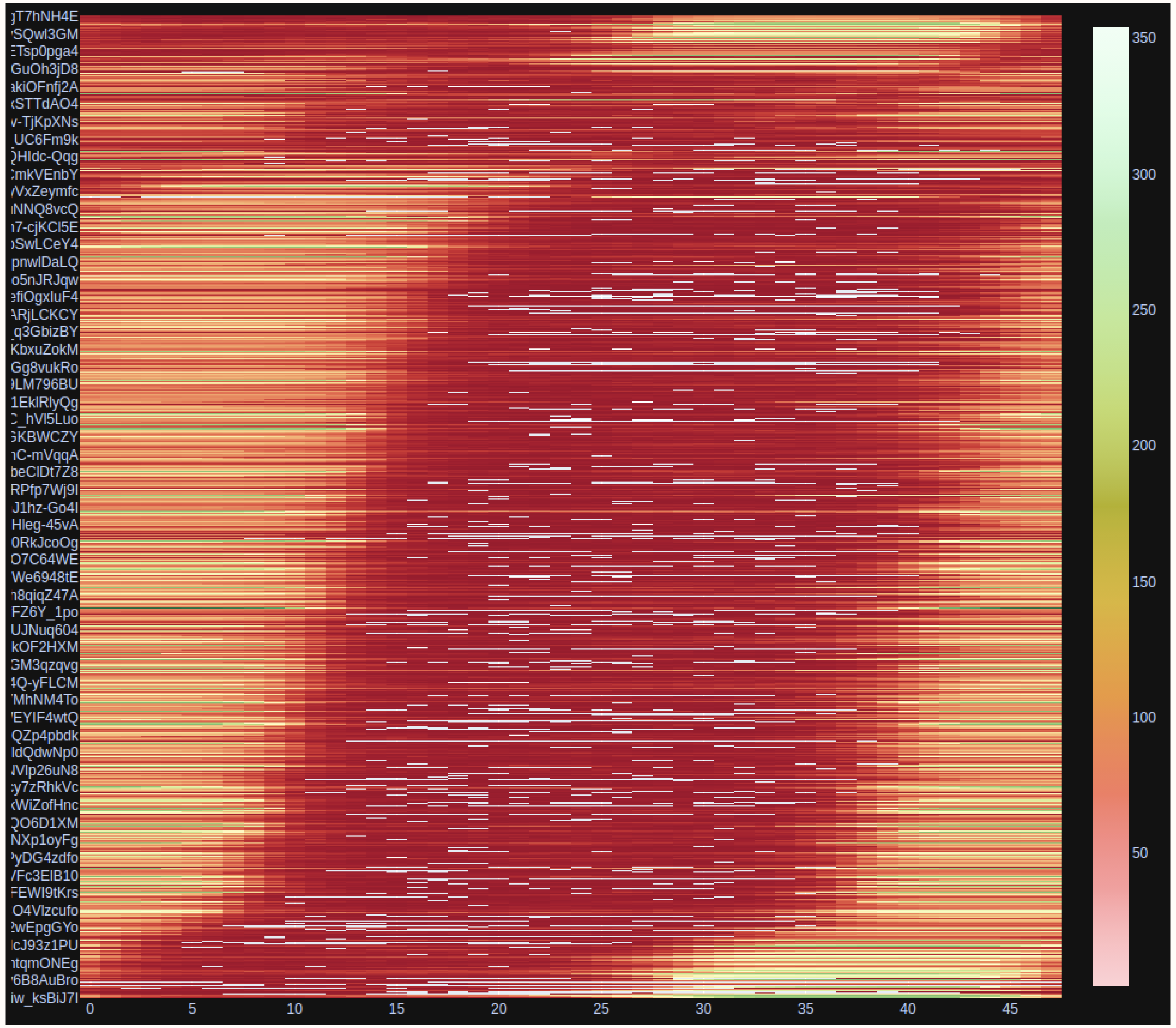
HR/RR
Night time
person



EDA

Data Availability of all individuals in the full dataset



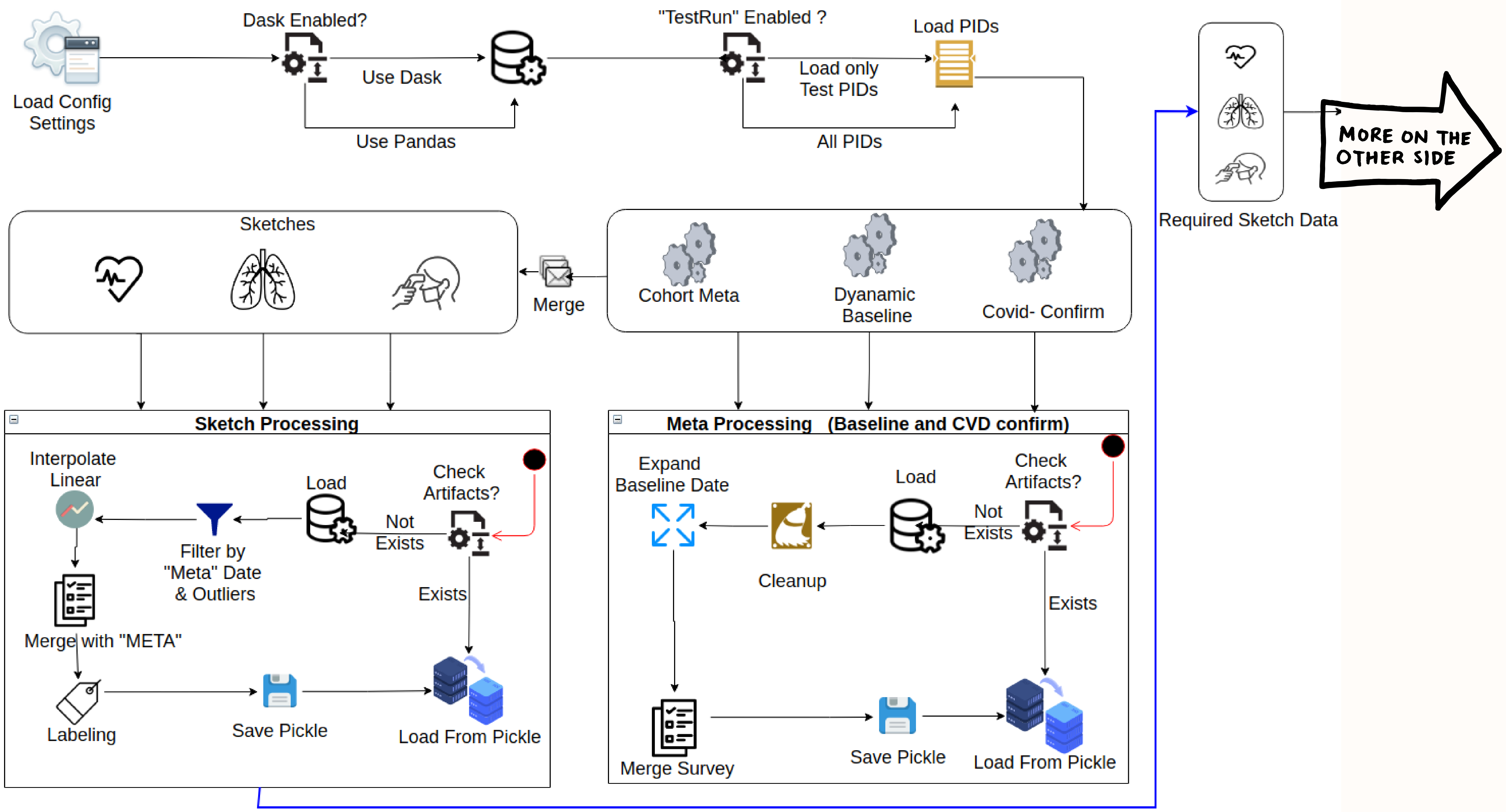


EDA

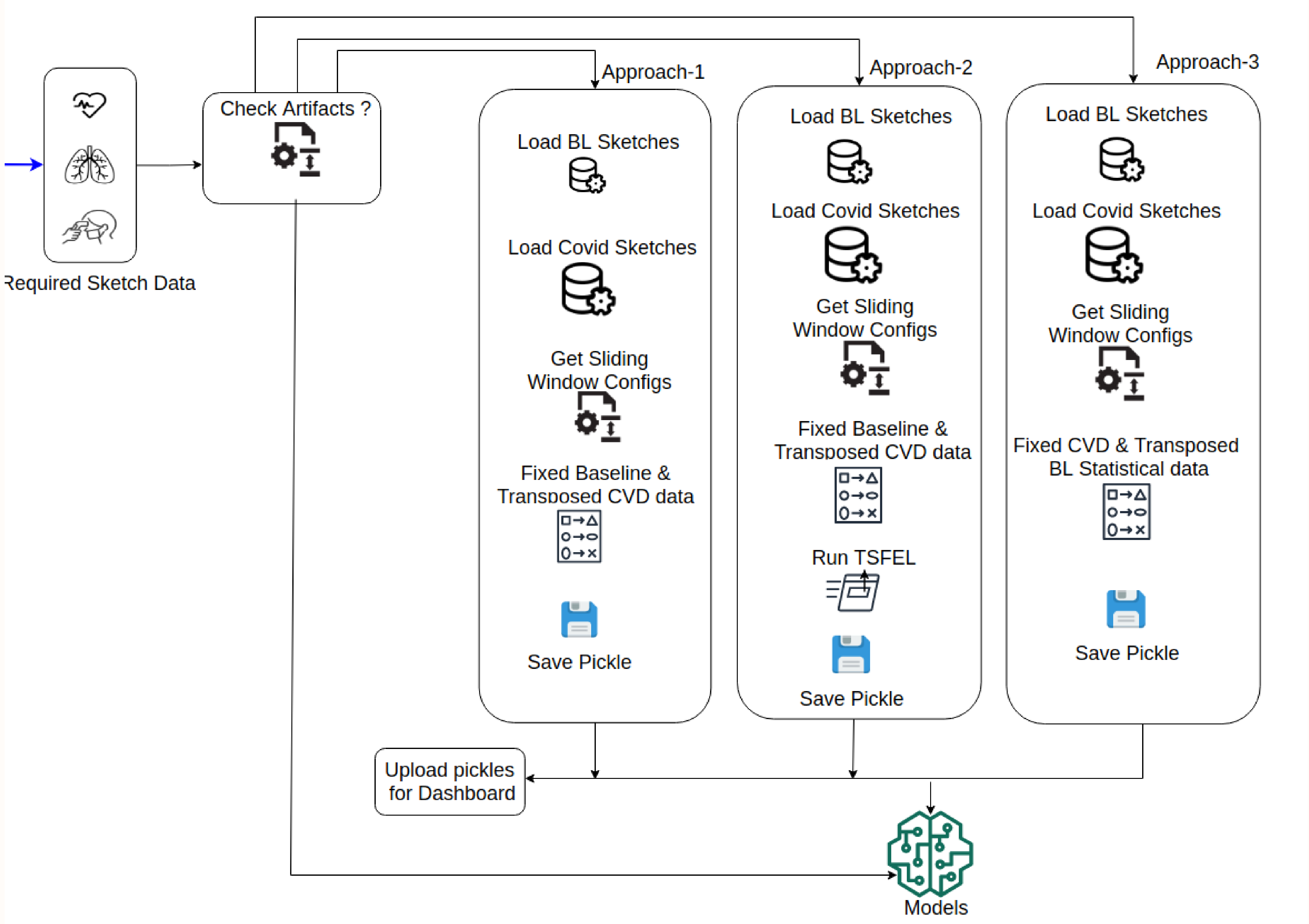
- **Baseline Window**
  The time period when the person is healthy
- **Duration**
  3 weeks to account for Weekly and daily rhythmicity
- **Covid Window**
  The time period when the person is infected or to detect infection
- **Duration**
  3 weeks

Data Preparation

Data Preparation



Data Preparation



Data Preparation

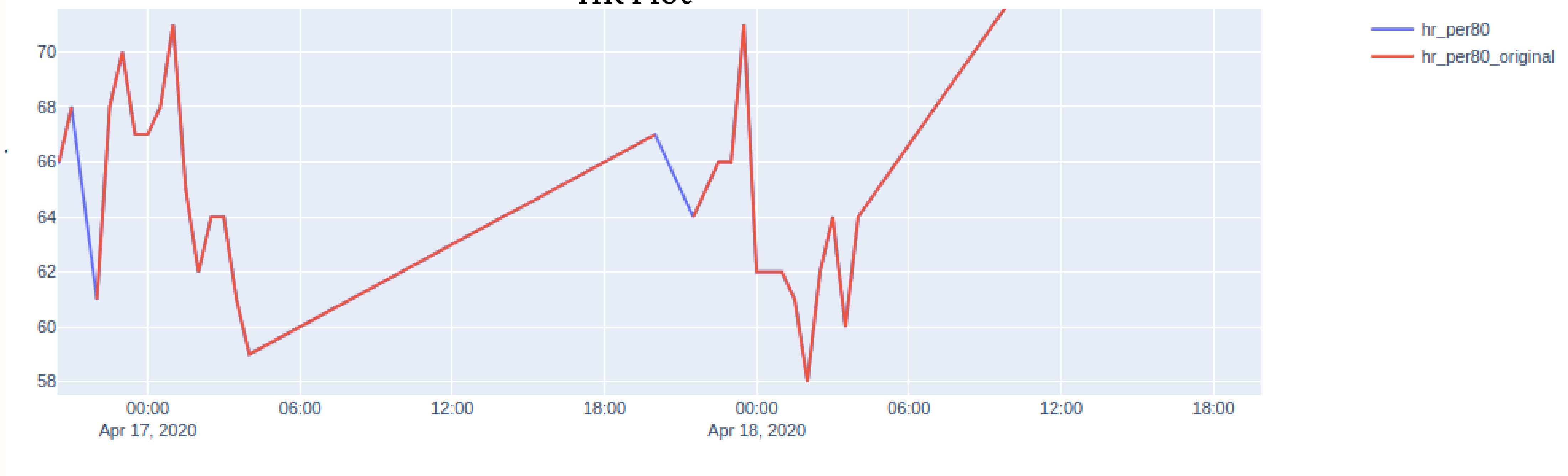
- Interpolation

  Linear interpolation over time for the missing values preserves the rhythm of the physiology data.

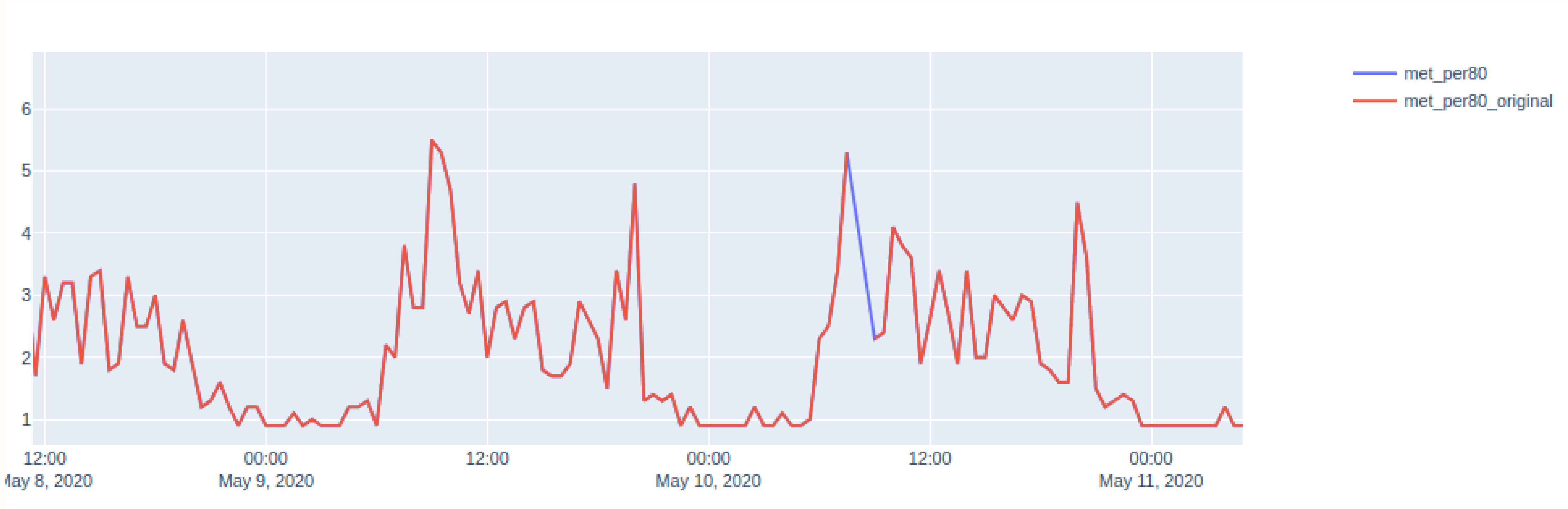
- Filter Outliers

  Remove Nonwear times
  Remove unrealistic recordings

HR Plot



MET Plot



Data Preparation

Baseline - 21 Days (3 weeks)

- By Days of the week -
  3 Sundays, Mondays, etc..
- All Available continuous 7 days
- 30, 60, 90 days baselines

Clustering

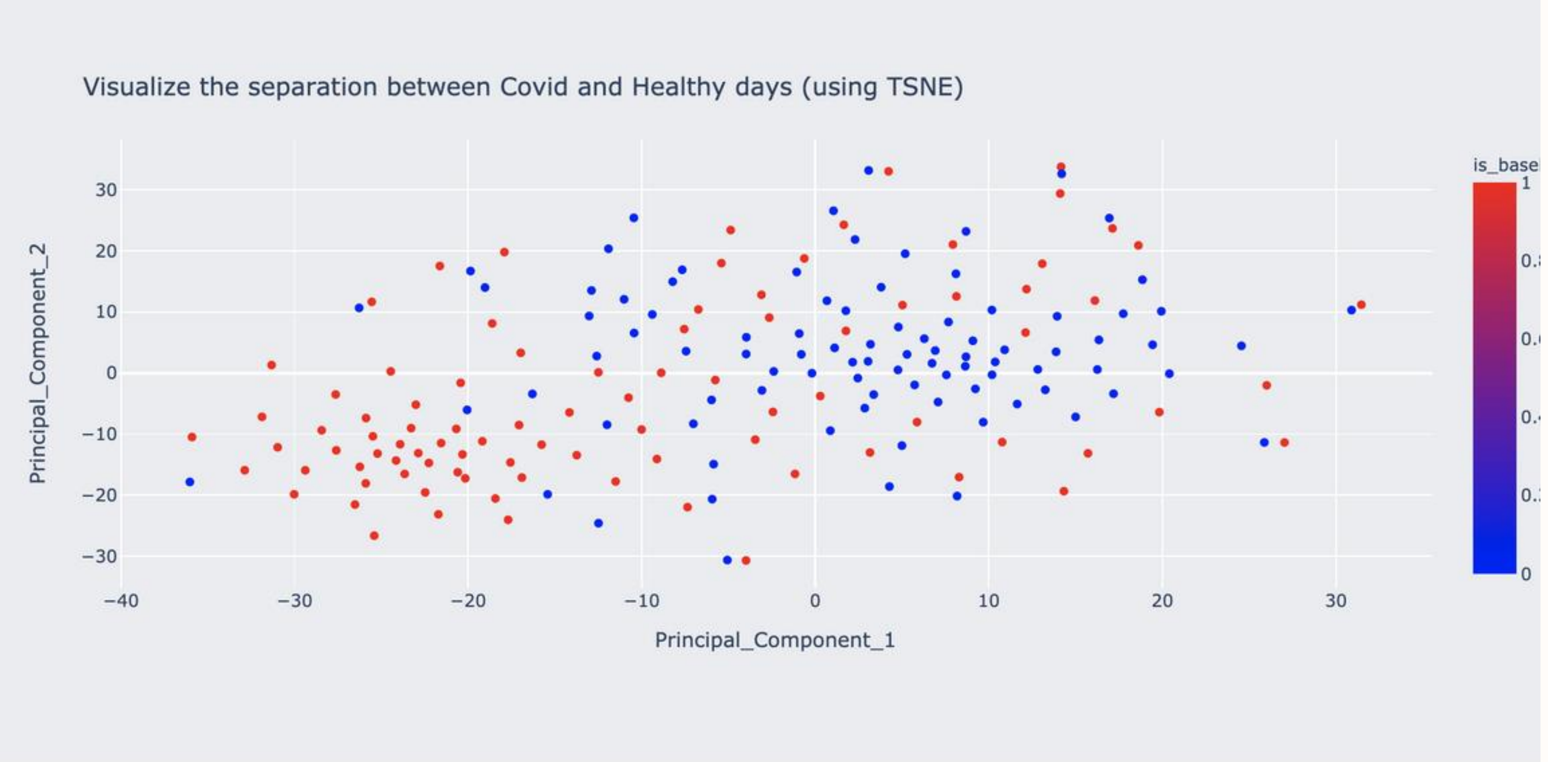
Cluster individuals based on variances in Temp, MET, HR, RR, and RMSSD. Transposing them into 3200 features

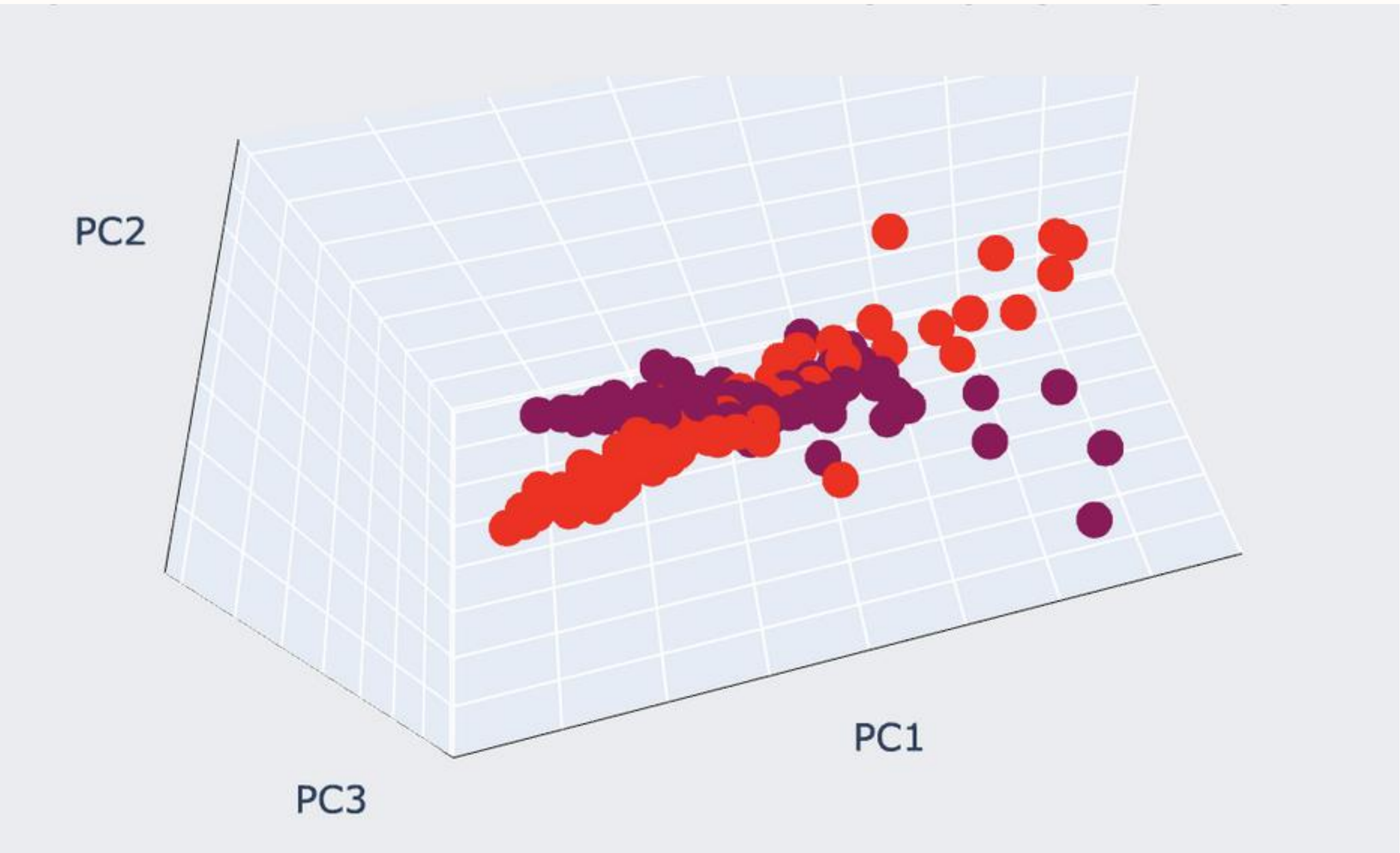
3200 Features were reduced to 450 using PCA

Gaussian distribution clustering with cluster size 4 to produce true covid and baseline clusters.

To visualize in 2D/3D the cluster input (450 dimensions) is further reduced to 2/3 using TSNE and PCA.

60% are clustered in different clusters during covid and baseline





Labeling

- Sx Date:
  Symptom Onset Date

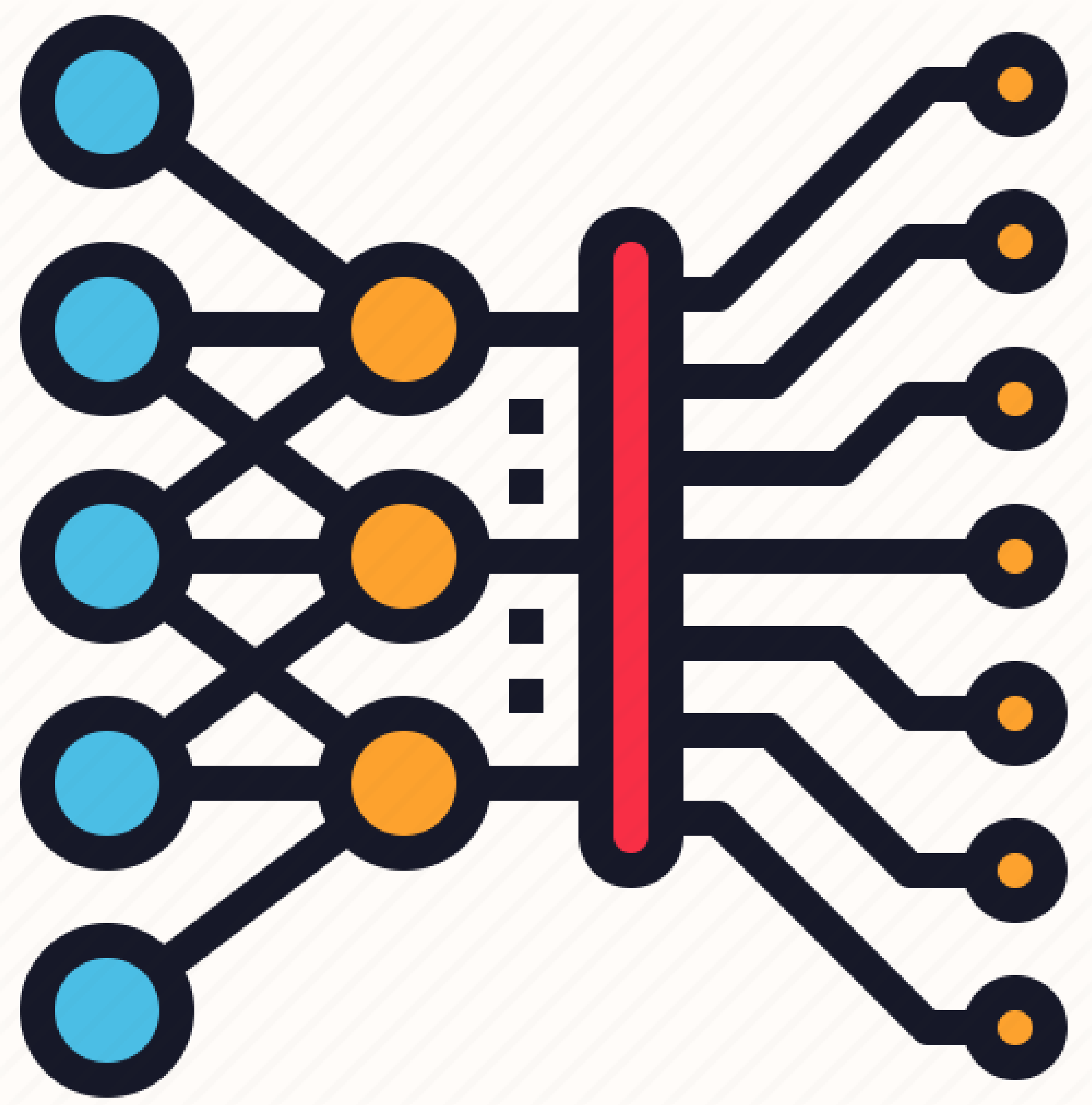
- Dx Date:
  Diagnosis Date

- Px Date:
  Median Date between
  Max of HR Variation date and
  Max of RR Variation date

- Target Labeling

  +/- 2 days between Px date and Dx date (Diagnosis date)

Modeling



Modeling

Approach-1: Transposed Baseline & Sliding Covid Window

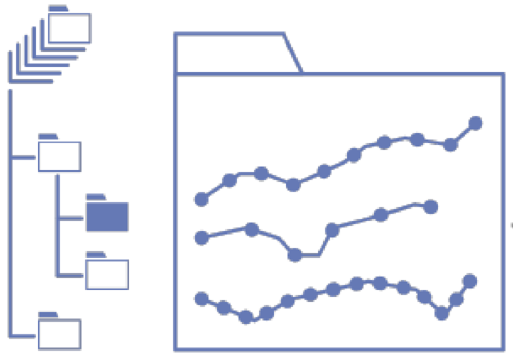
For each physiological data, 21 days baseline, a sliding Covid window of 3 days was taken.

Data was transposed to columns and passed as input and train the model

Target : Covid Onset (early detection)

Input data shape : 1530 * 10518
Total Population : 90
Train Test split : 80/20

Time Series Signal



Original Data

PID	Base_date_time	HR_Mean
x1	Mon_b1	v1
x1	Tue_b2	v2
x1	Wed_b3	v3
x1	Mon_c1	v4
x1	Tue_c2	v5
x1	Wed_c3	v6
x1	Thurs_c4	v7

Baseline Records

Covid Records

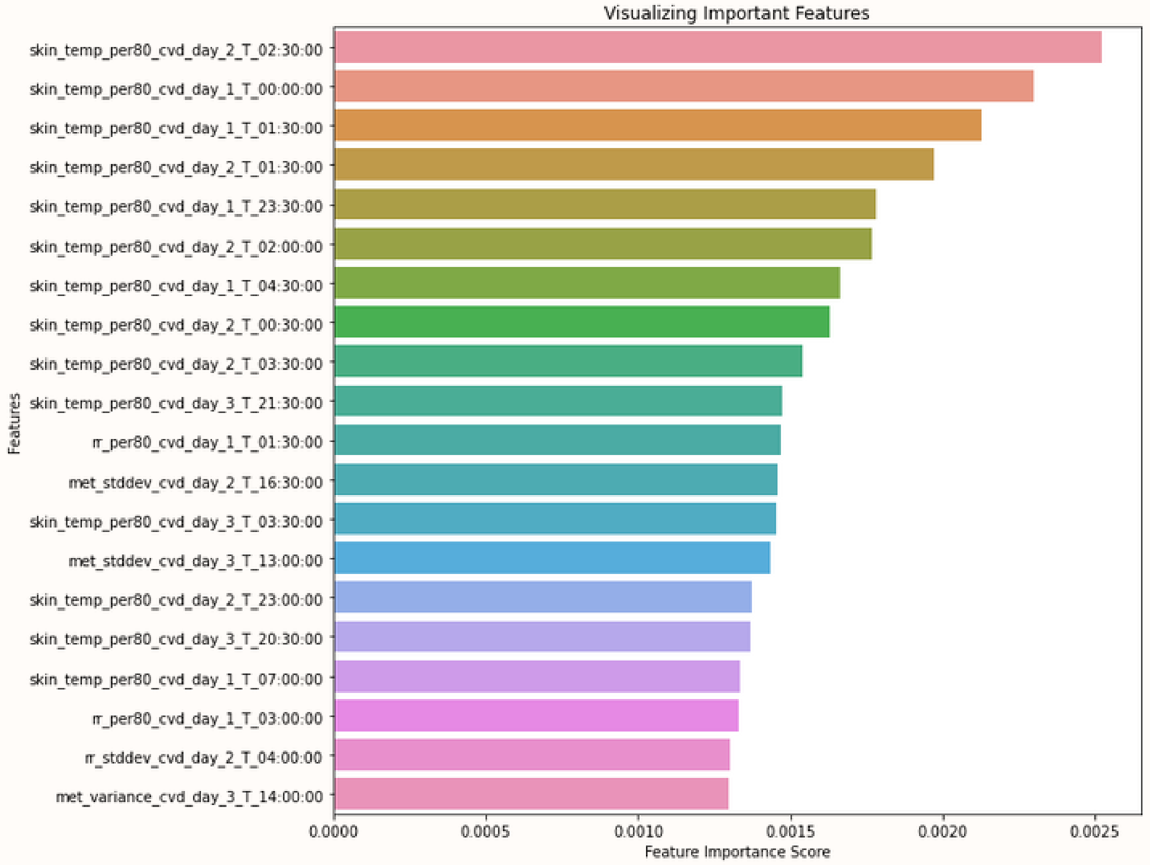
Transposed Data

	Baseline Records (21 days)			Moving covid window		
PID	Mon_b1	Tue_b2	Wed_b3	cvd1	cvd2	cvd3
x1	v1	v2	v3	v4	v5	v6
x1	v1	v2	v3	v5	v6	v7

Modeling

Approach-1: Feature Importance & Results

Feature Importance



Comparison between different algorithms

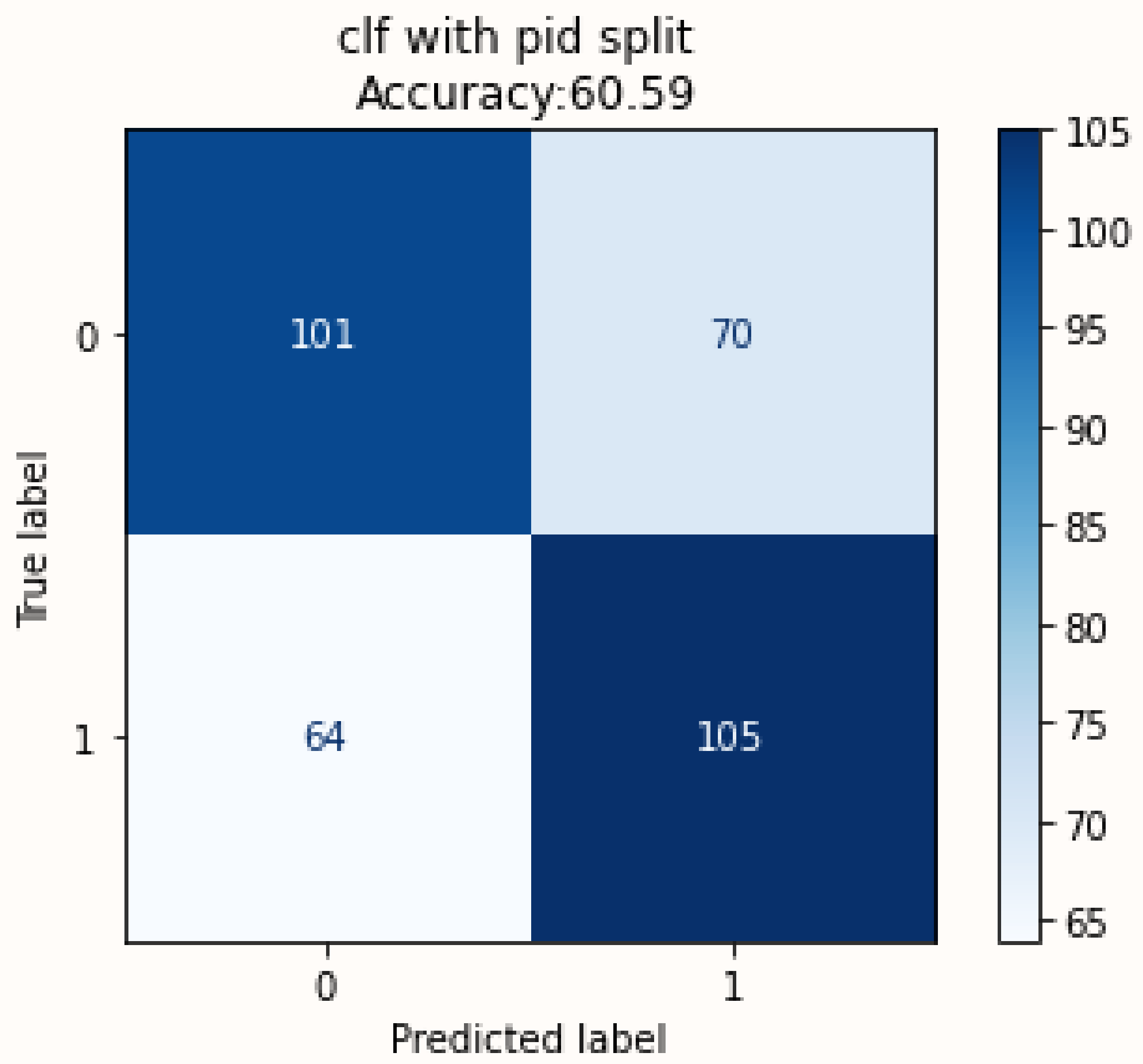
APPROACH-1					
Test train PID split	Model/Classifier	Accuracy	Precision	Recall	Fscore
(80:20)	RandomForest	57.84%	0.5893	0.5784	0.5752
(80:20)	XGB	55.23 %	0.5635	0.5522	0.5477
(80:20)	Bagging	56.86 %.	0.5854	0.5686	0.5611
(80:20)	GradientBoosting	53.27 %.	0.5465	0.5326	0.5241
(80:20)	AdaBoost	53.92 %.	0.5477	0.5392	0.5362

Modeling

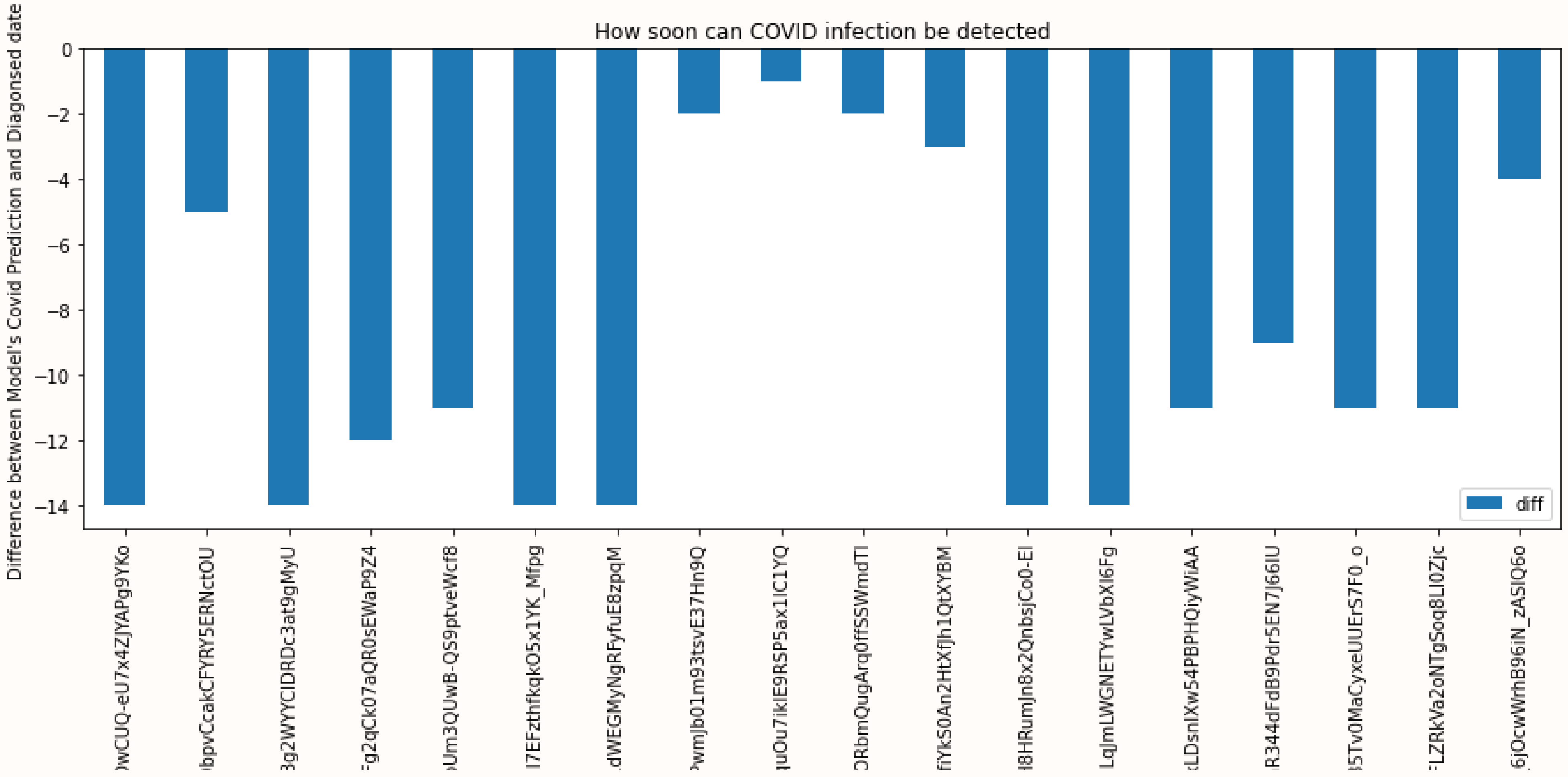
Approach-1: Evaluation

90%

Confusion Matrix



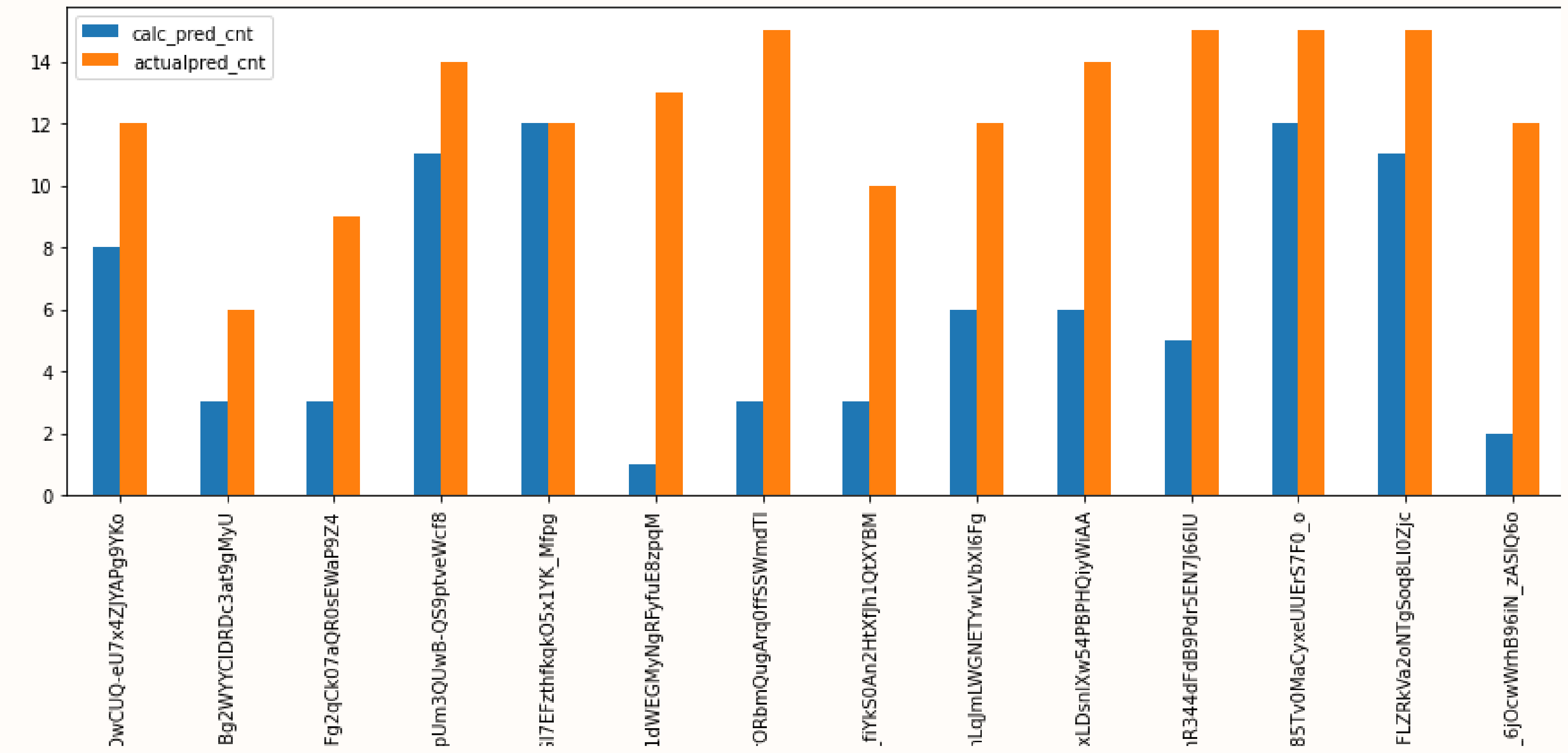
How soon can the onset be detected



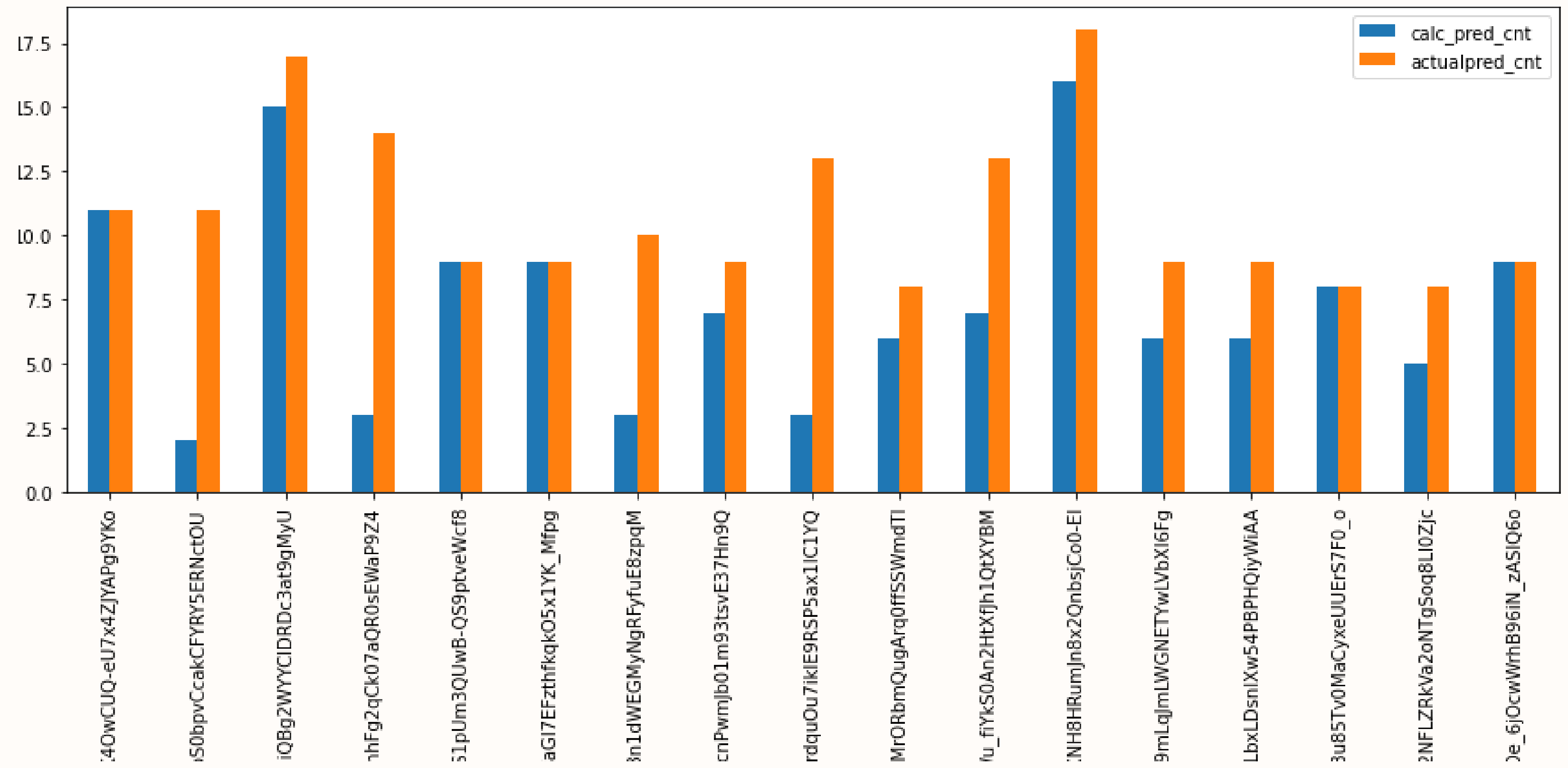
Modeling

Approach-1: Evaluation

Analysis True Positives (Day Level per PID)

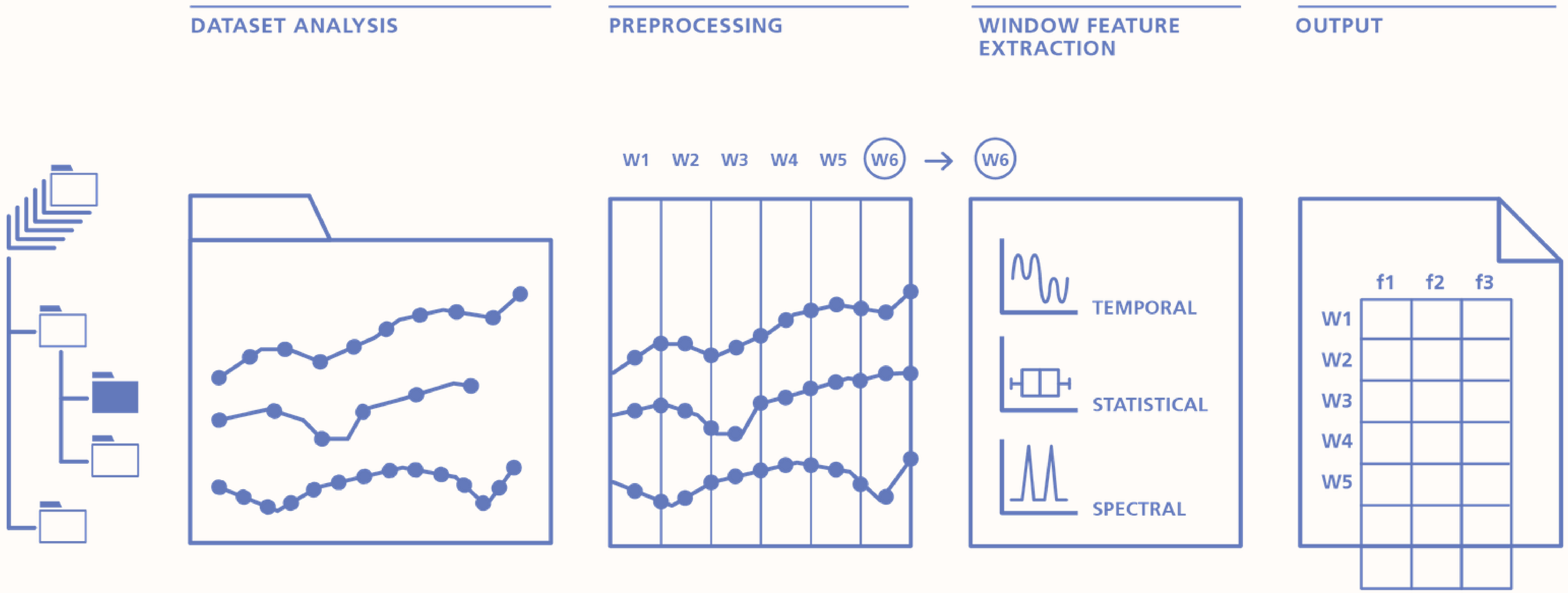


Analysis False Positives (Day Level per PID)



Modeling

Approach-2 : Time Series Feature Extraction Library



Source: https://tsfel.readthedocs.io/

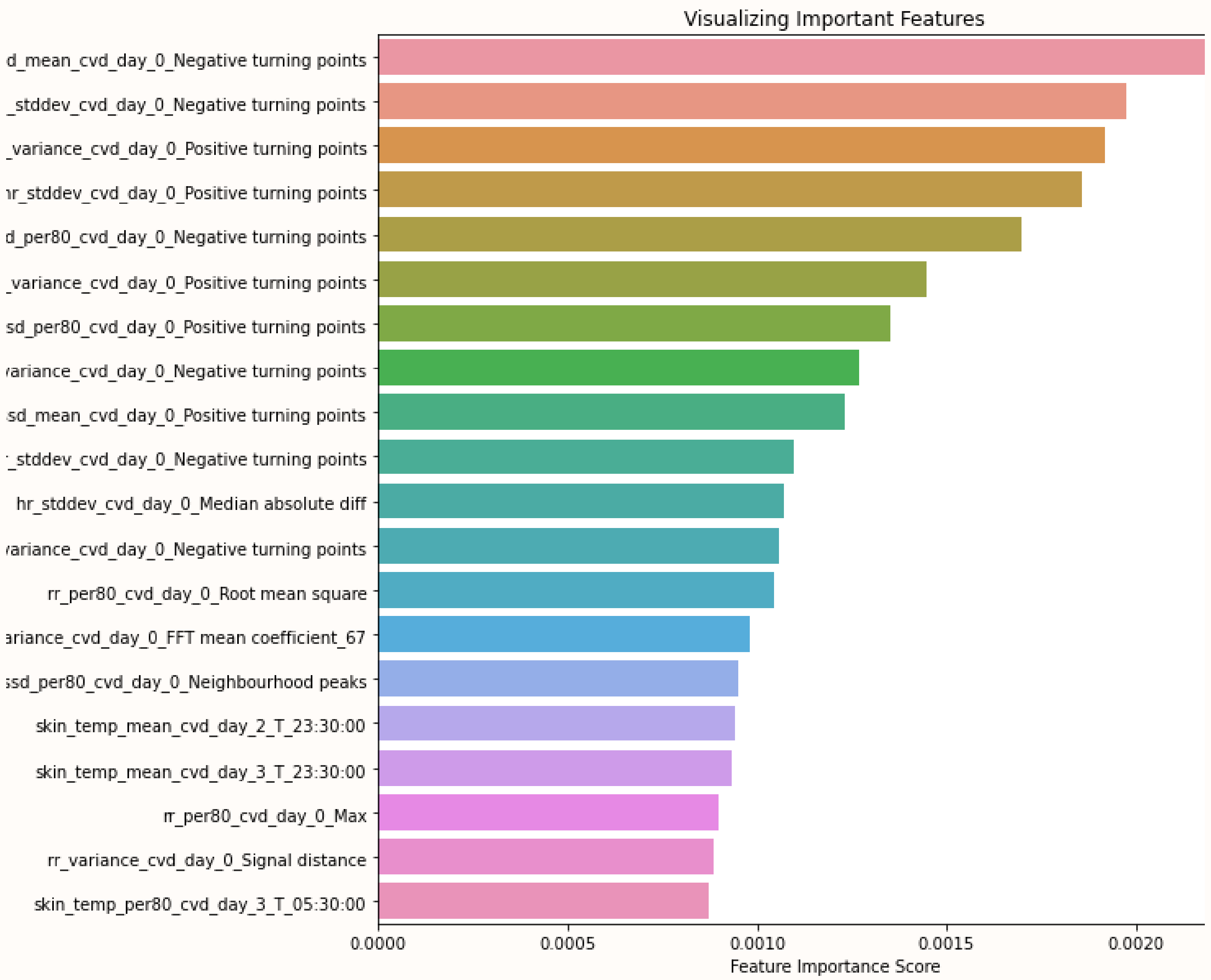
Input data shape : 1530 * 11388
Total Population : 90
Train Test split : 80/20

Modeling

Approach-2: Feature Importance & Results

TSFEL

Feature Importance



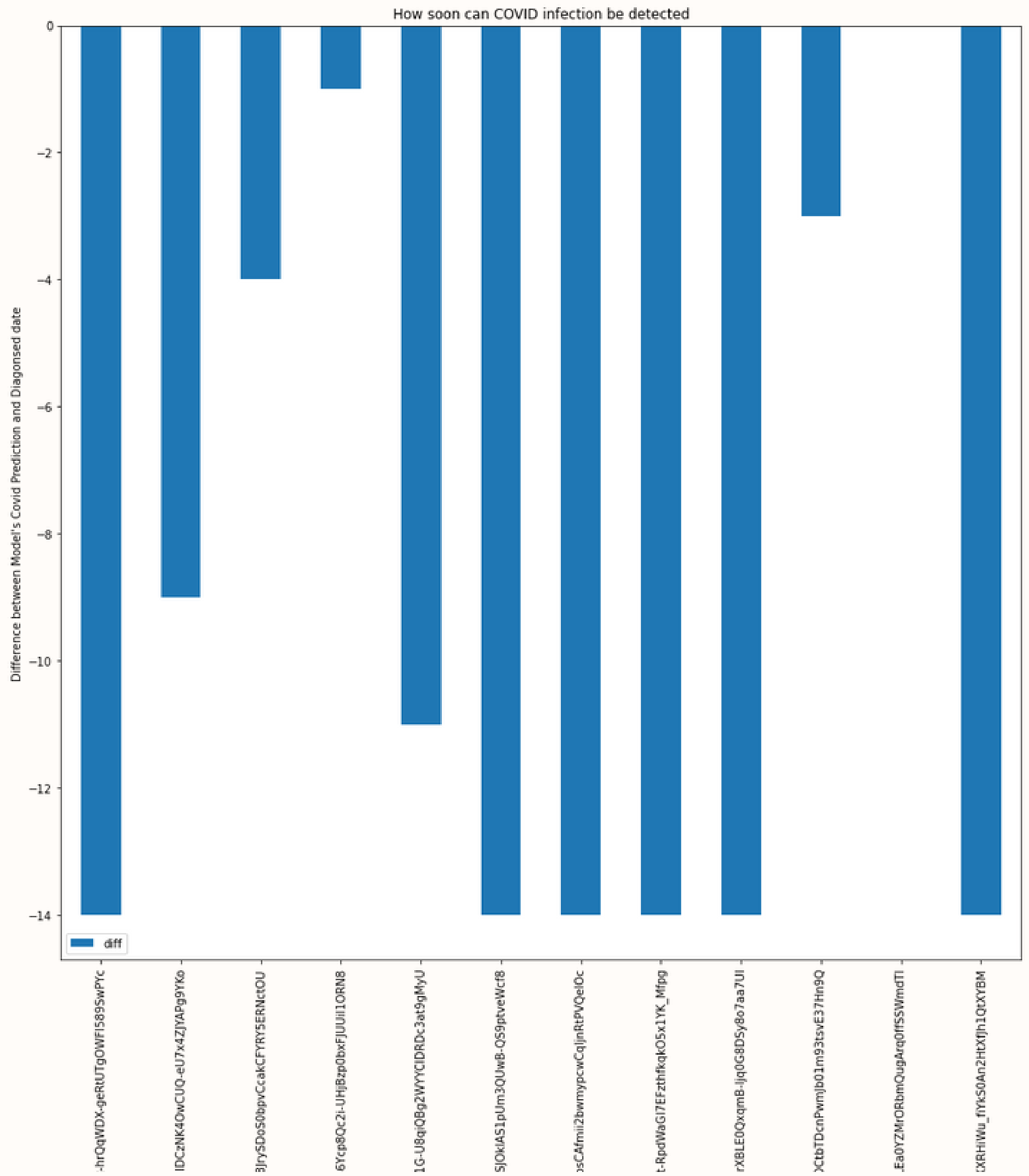
Comparison between different algorithms

Test train PID split	Model/Classifier	Accuracy	Precision	Recall	Fscore
(80:20)	RandomForest	58.82 %	0.6051	0.5882	0.5802
(80:20)	XGB	53.59 %	0.5893	0.5784	0.5752
(80:20)	Bagging	58.17 %	0.5991	0.5816	0.5727
(80:20)	GradientBoosting	57.84 %.	0.5843	0.5784	0.5768
(80:20)	AdaBoost	58.5 %	0.5897	0.5849	0.584

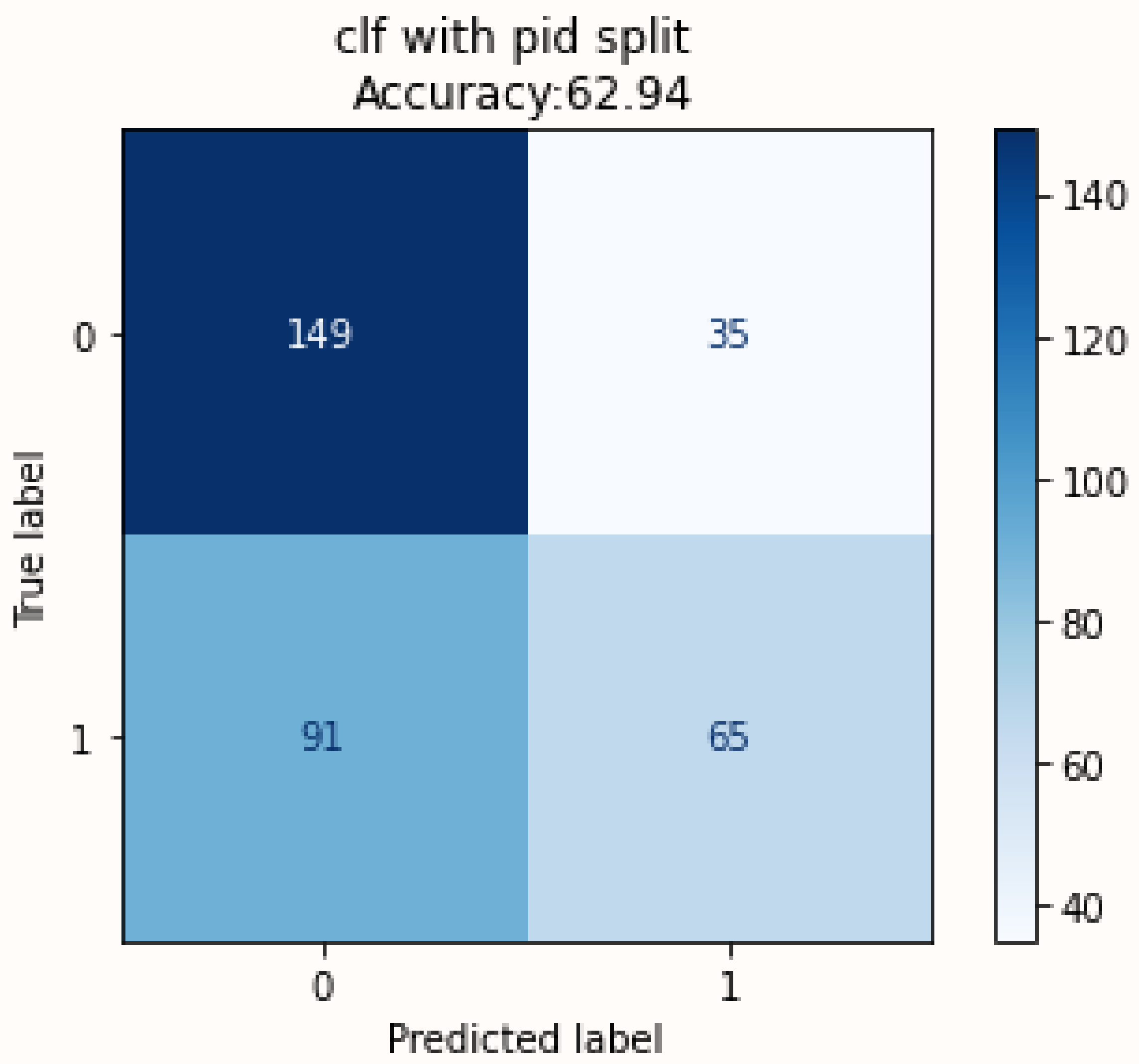
Modeling

Approach-2: Evaluation

How soon can the onset be detected



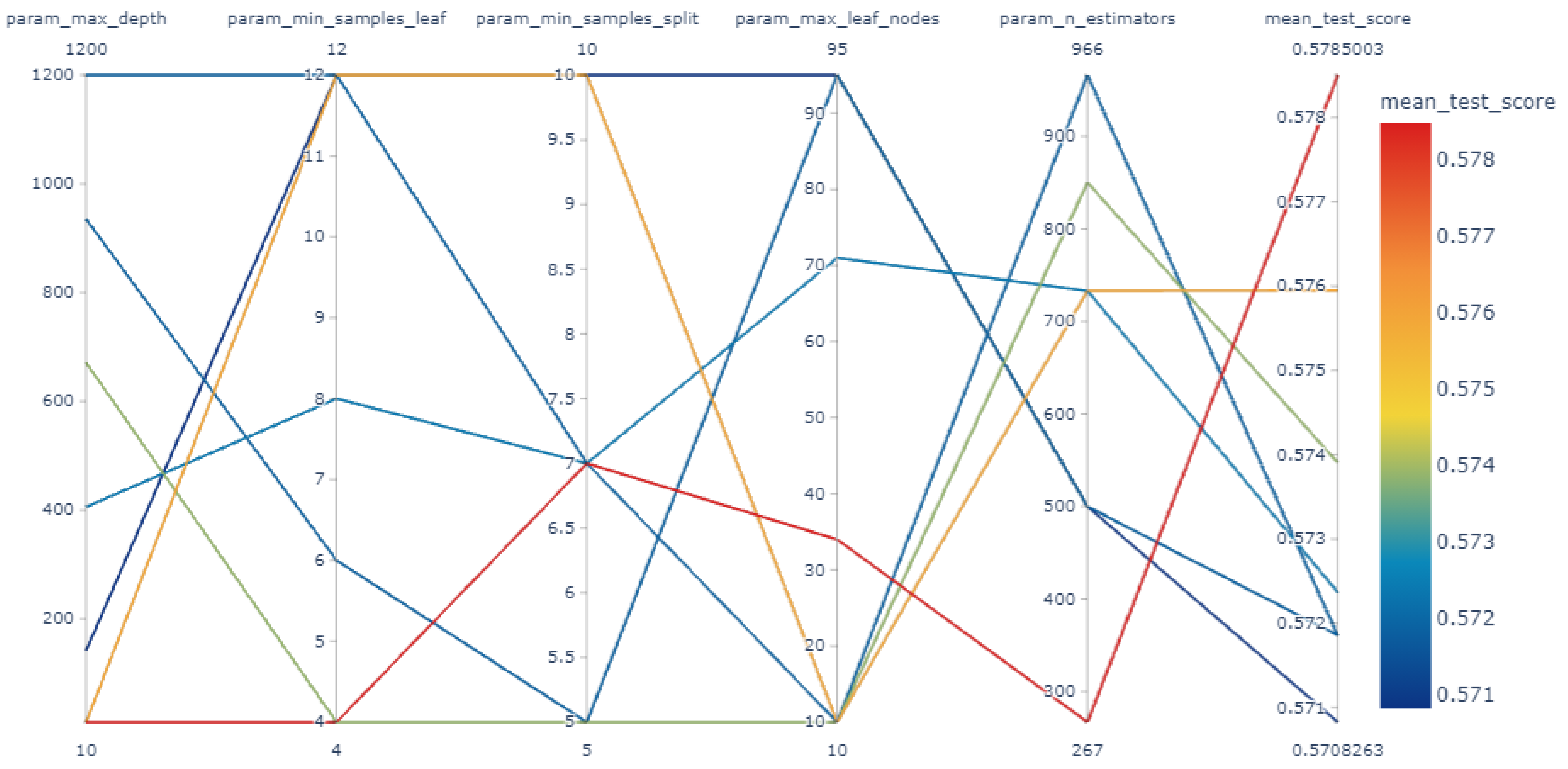
Confusion Matrix



Modeling

Approach-2: Evaluation

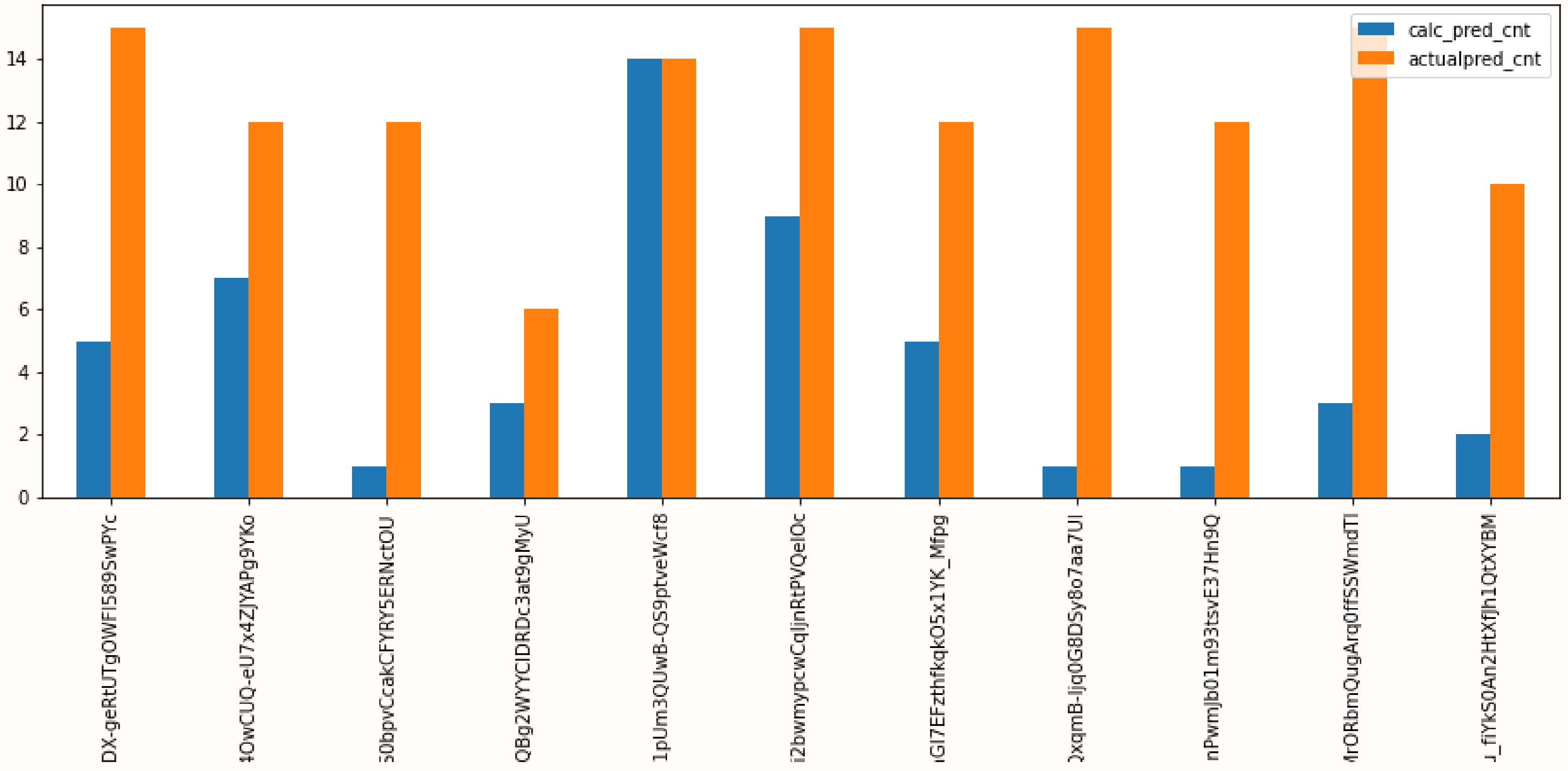
Parallel Coordinate Plot to select best parameters



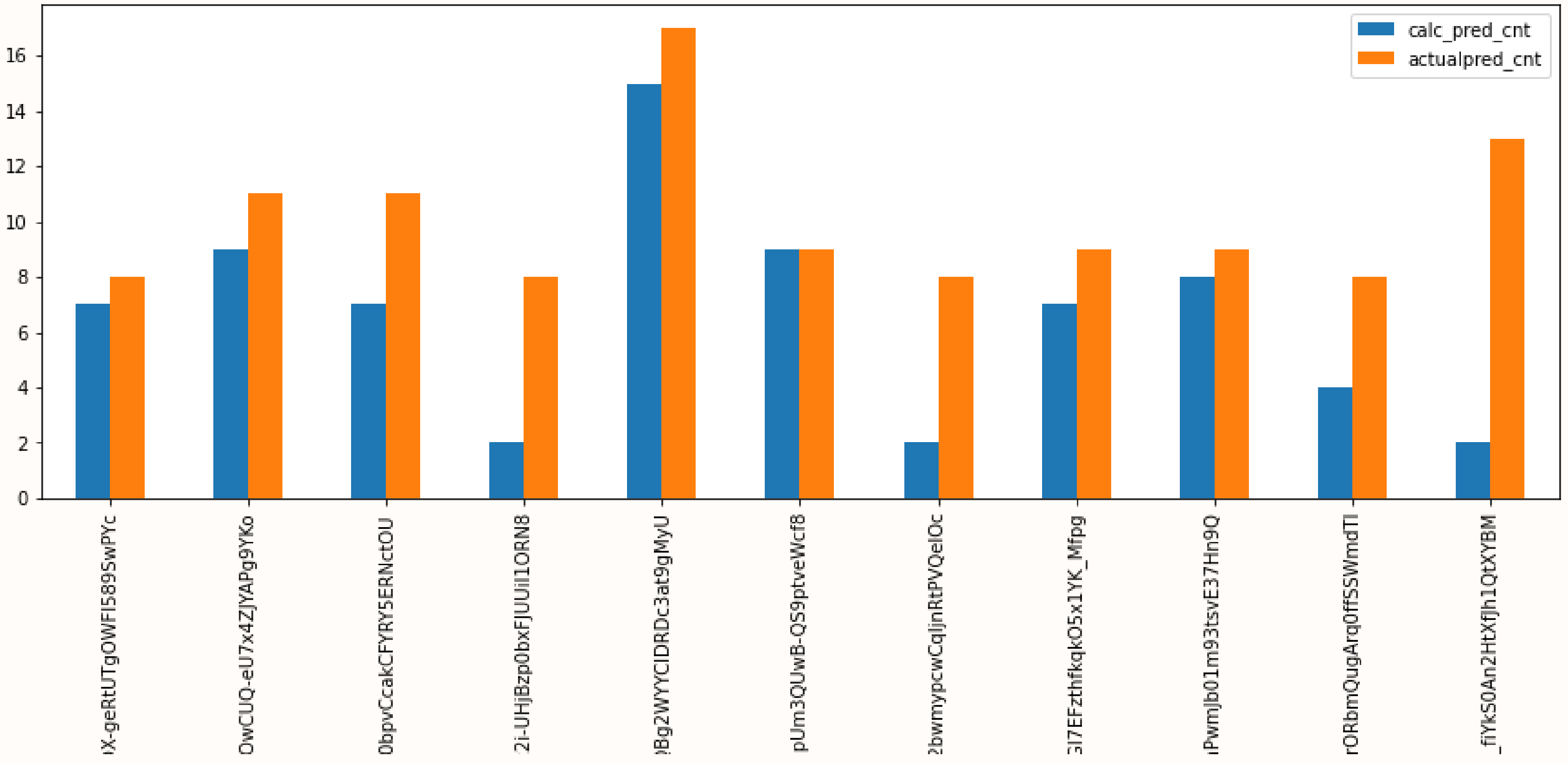
Modeling

Approach-2: Evaluation

Analysis True Positives (Day Level per PID)



Analysis False Positives (Day Level per PID)



Modeling

Approach-3: Aggregated Baseline Statistics

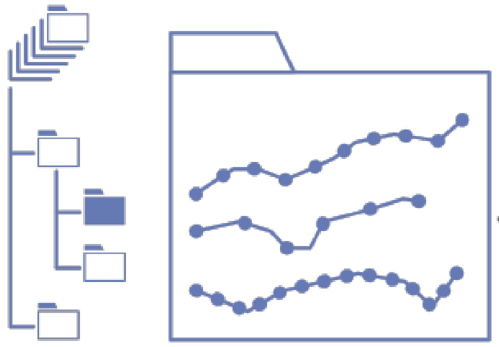
Append the day based baseline's statistical feature

Higher order features
Ratio of temp & met
Ratio of HR & RMSSD

Deviation from Baseline

Input data shape : 92671, 60
Total Population : 90
Train Test split : 80/20

Time Series Signal



Original Data

PID	Base_date_time	HR_Mean
x1	Mon_b1	v1
x1	Mon_b2	v2
x1	Mon_b3	v3
x1	Tue_b1	v4
x1	Tue_b2	v5
x1	Tue_b3	v6
x1	Mon_c1	v7
x1	Tue_c2	v8

Baseline Records

Covid Records

Baseline Aggregated

Covid Records as is

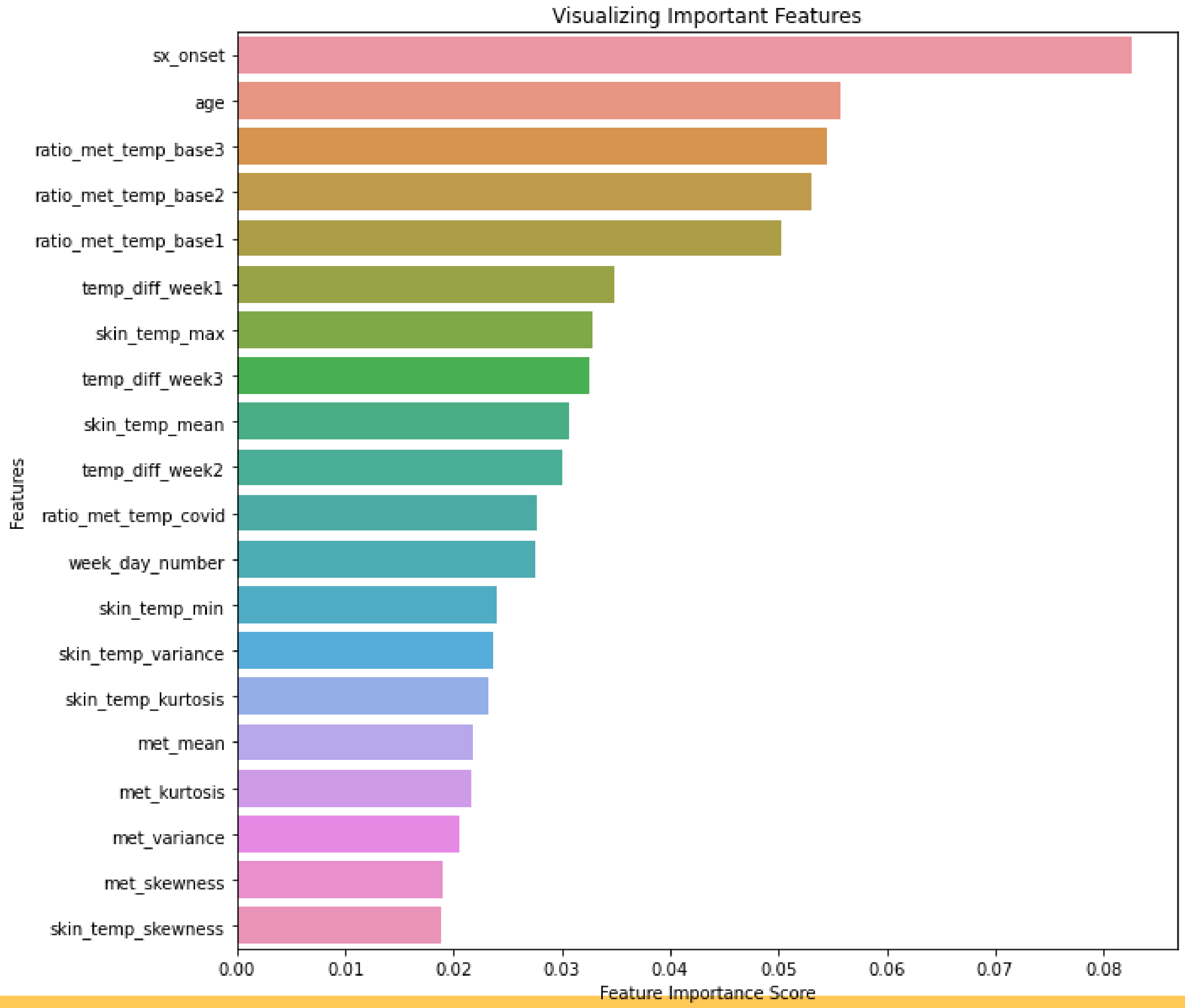
Aggregated Baseline Records

PID	Base_date_time	HR_Mean	Agg_b1	Agg_b2	Agg_b3
x1	Mon_c1	v7	mean(v1)	mean(v2)	mean(v3)
x1	Tue_c2	v8	mean(v4)	mean(v5)	mean(v6)

Modeling

Approach-3: Feature Importance & Results

Feature Importance



Comparison between different algorithms

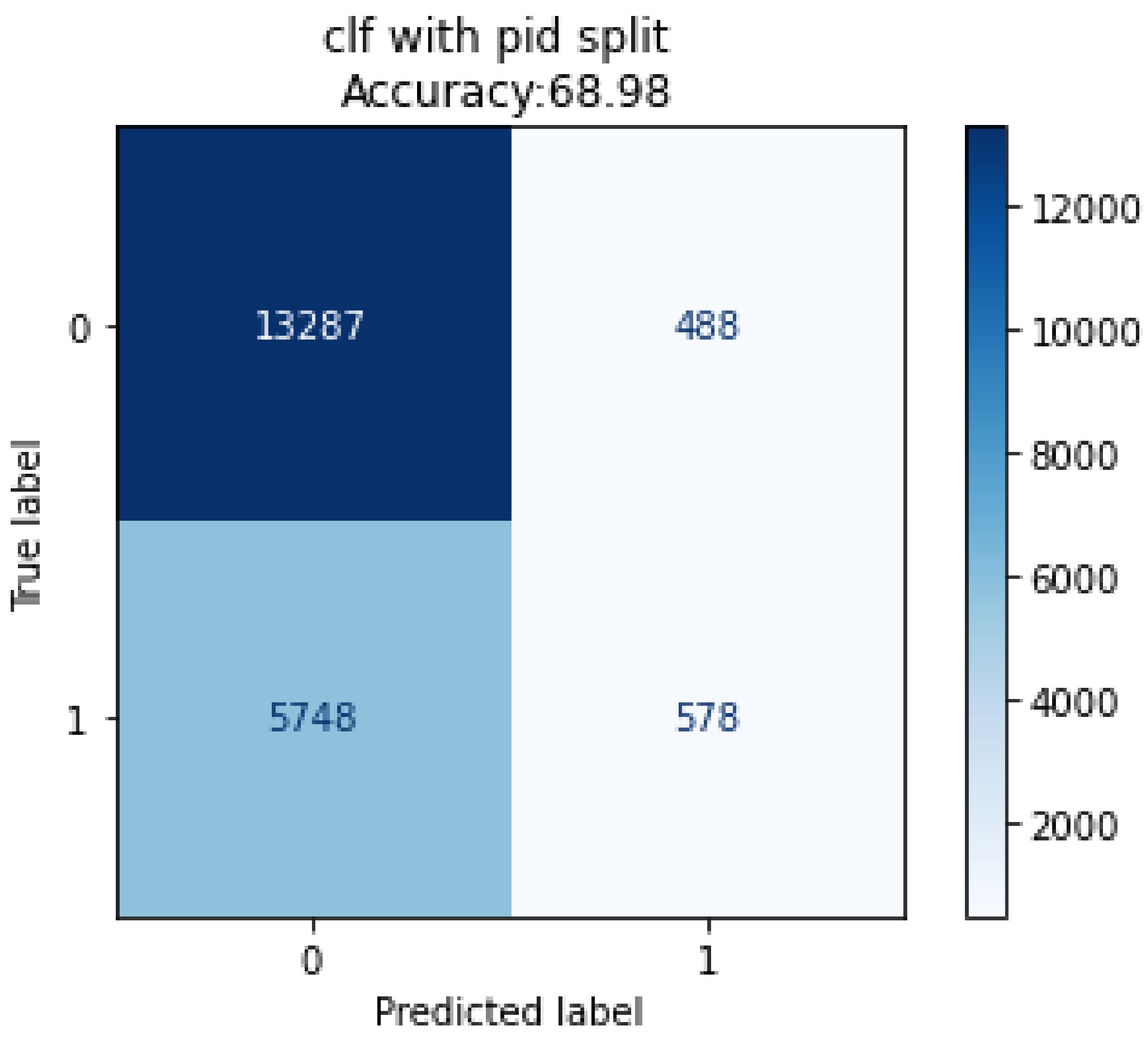
APPROACH-3					
Test train PID split	Model/Classifier	Accuracy	Precision	Recall	Fscore
(80:20)	RandomForest	69.17%	0.6626	0.6915	0.5936
(80:20)	Bagging	68.94 %	0.6531	0.6894	0.5889
(80:20)	GradientBoosting	55.79 %.	0.5826	0.5578	0.5679
(80:20)	AdaBoost	65.84 %	0.6041	0.6584	0.6091

Modeling

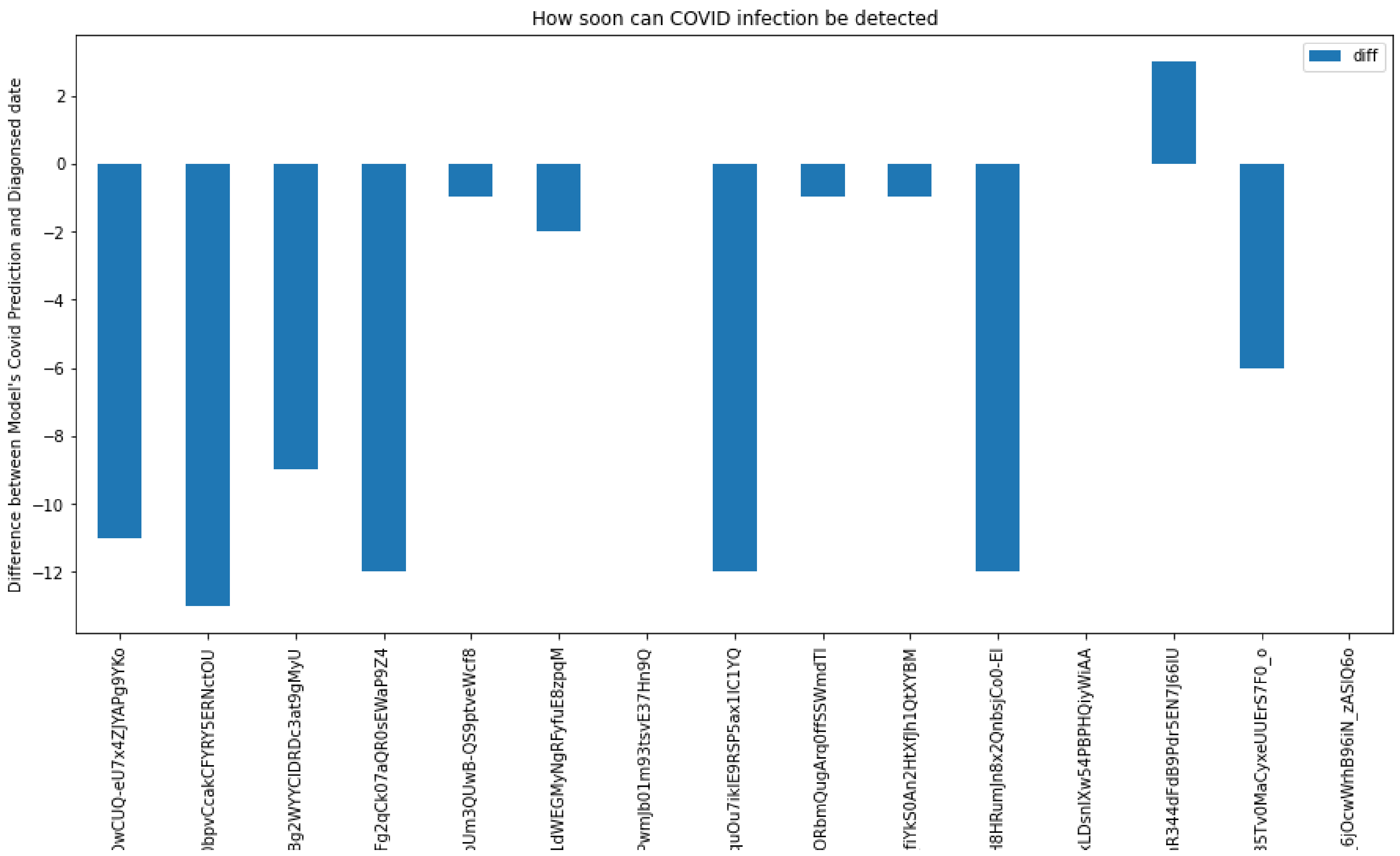
Approach-3: Evaluation

75%

Confusion Matrix



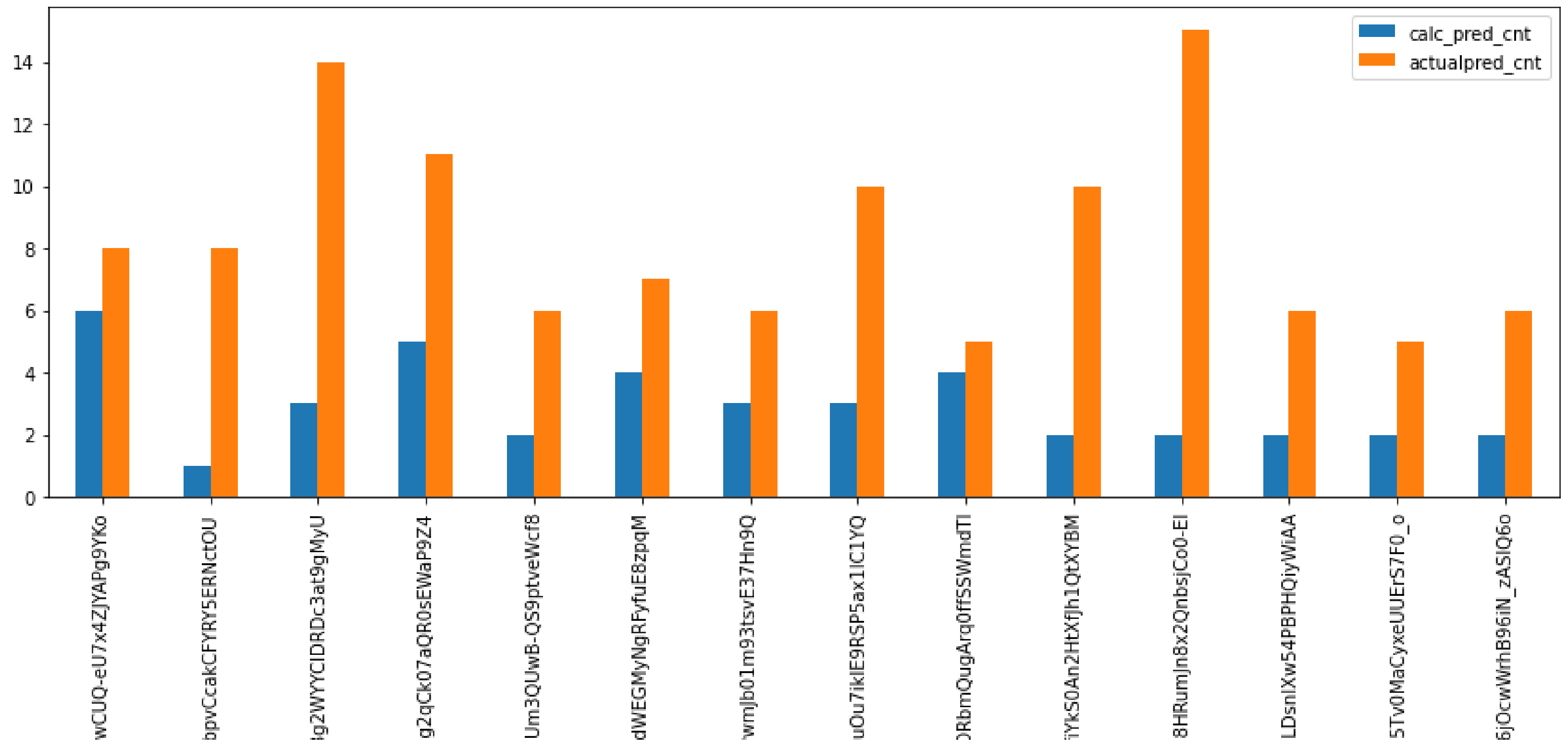
How soon can the onset be detected



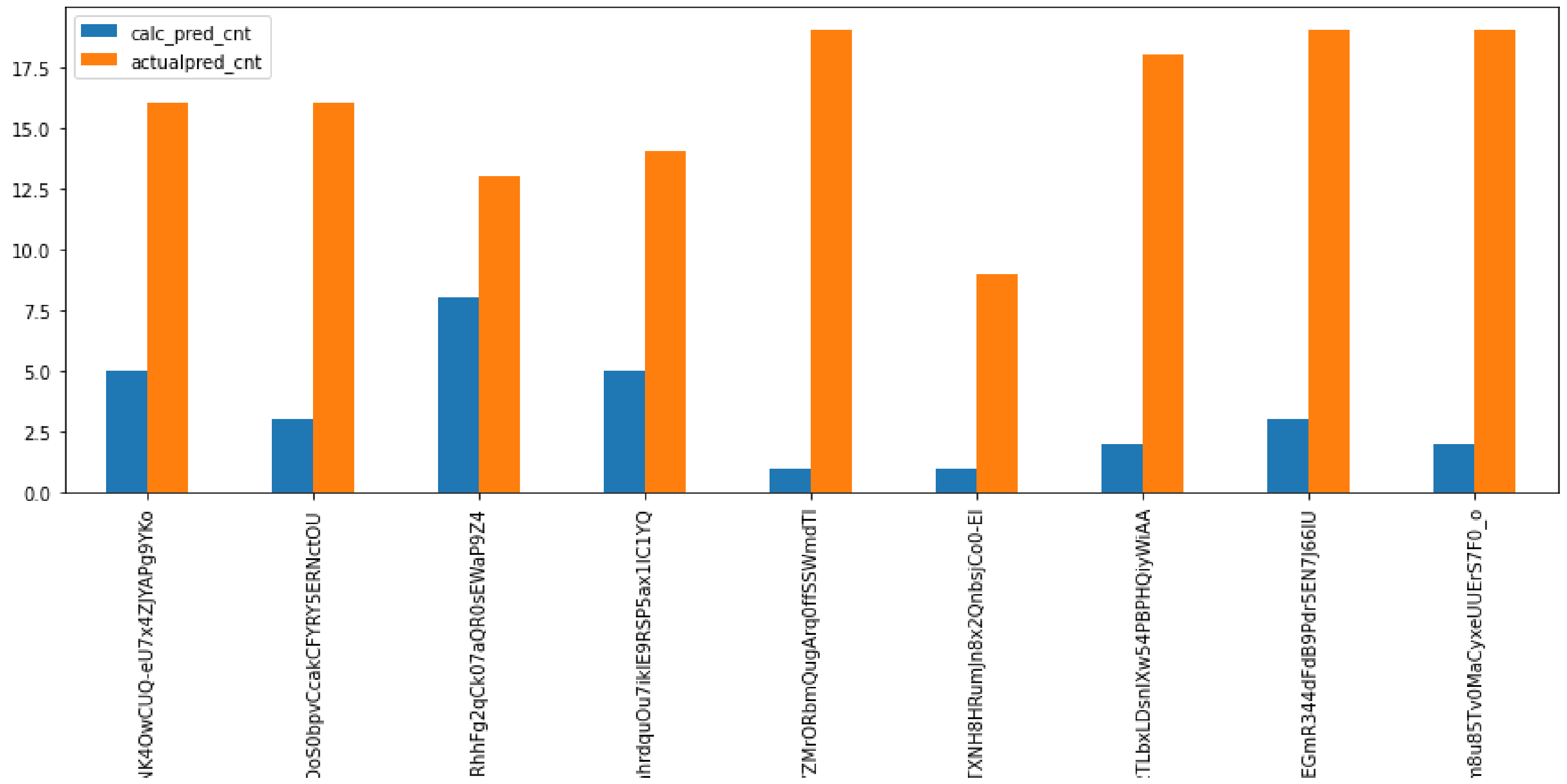
Modeling

Approach-3: Evaluation

Analysis True Positives (Day Level per PID)



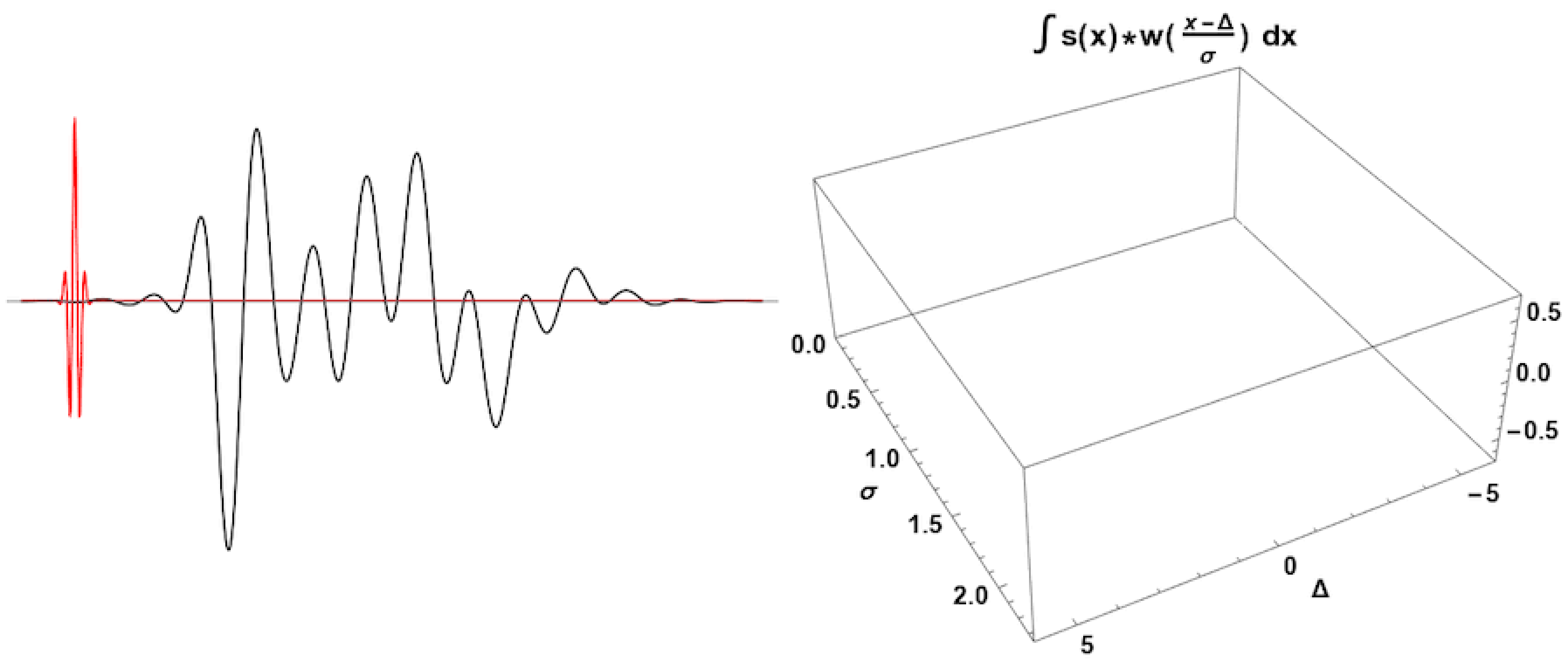
Analysis False Positives (Day Level per PID)



Modeling

Approach-4: Pattern Recognition in Frequency Domain

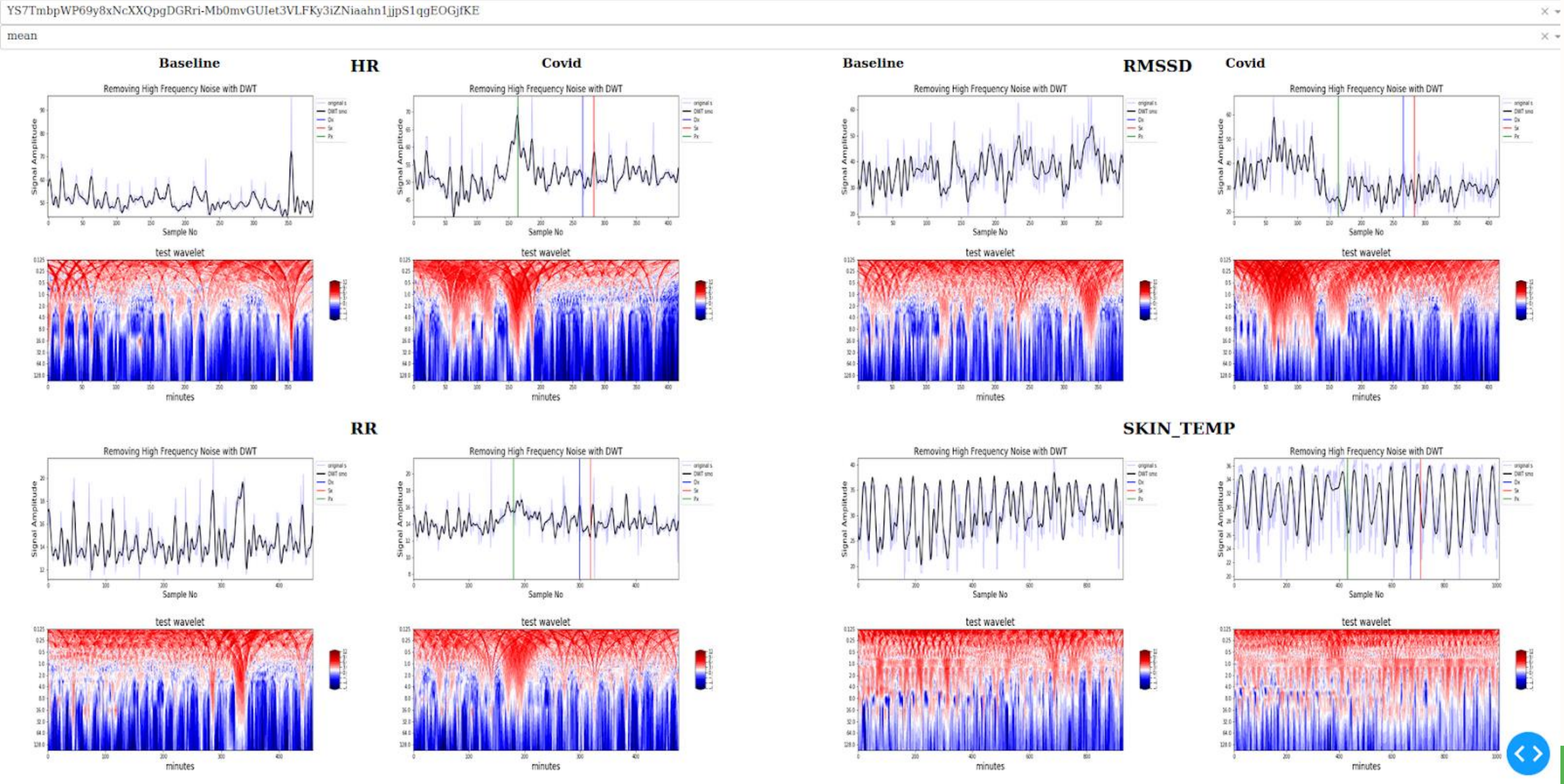
Continuous Wavelet Transform (CWT)



Source: wikipedia

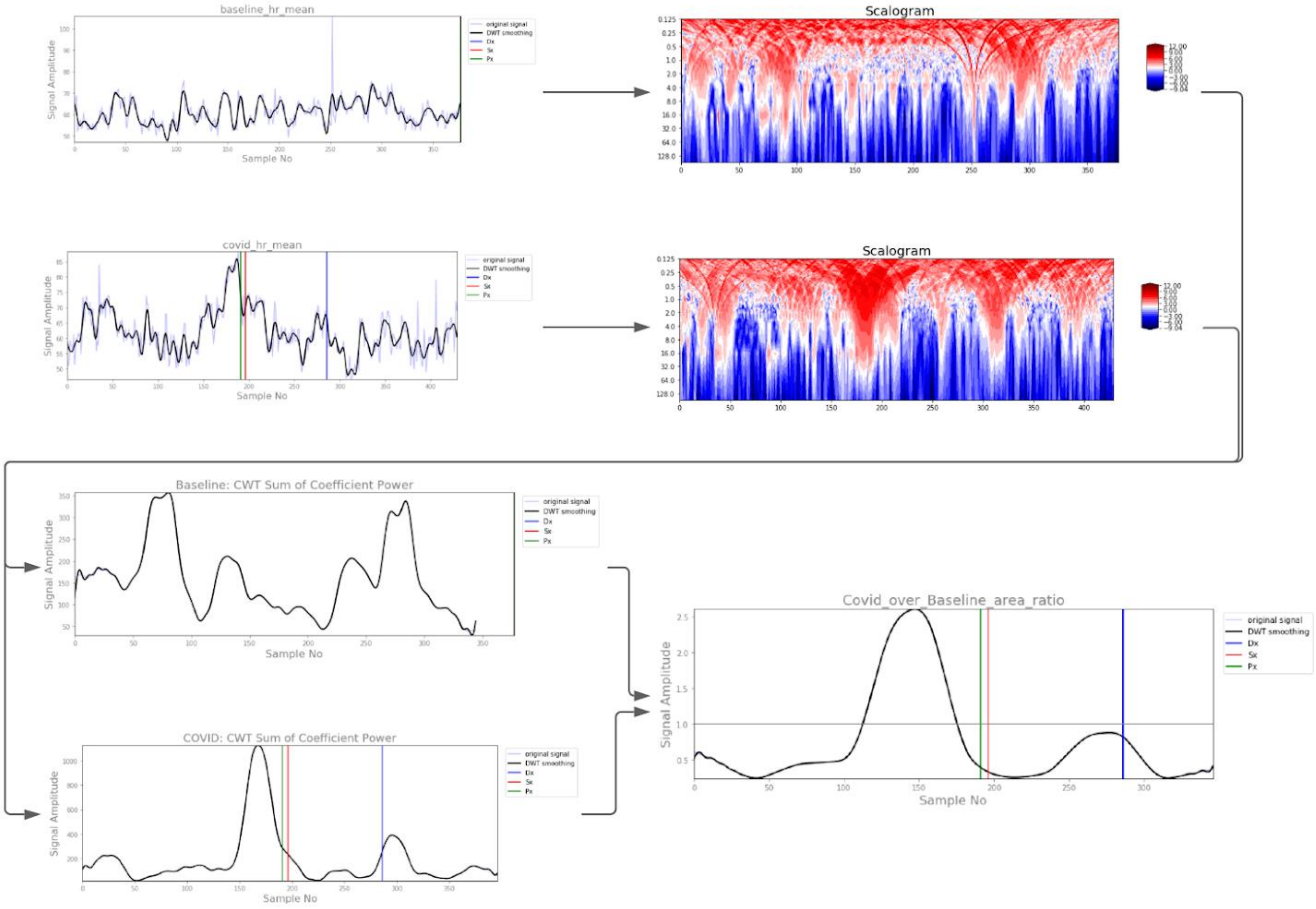
Modeling

Approach-4: EDA - CWT



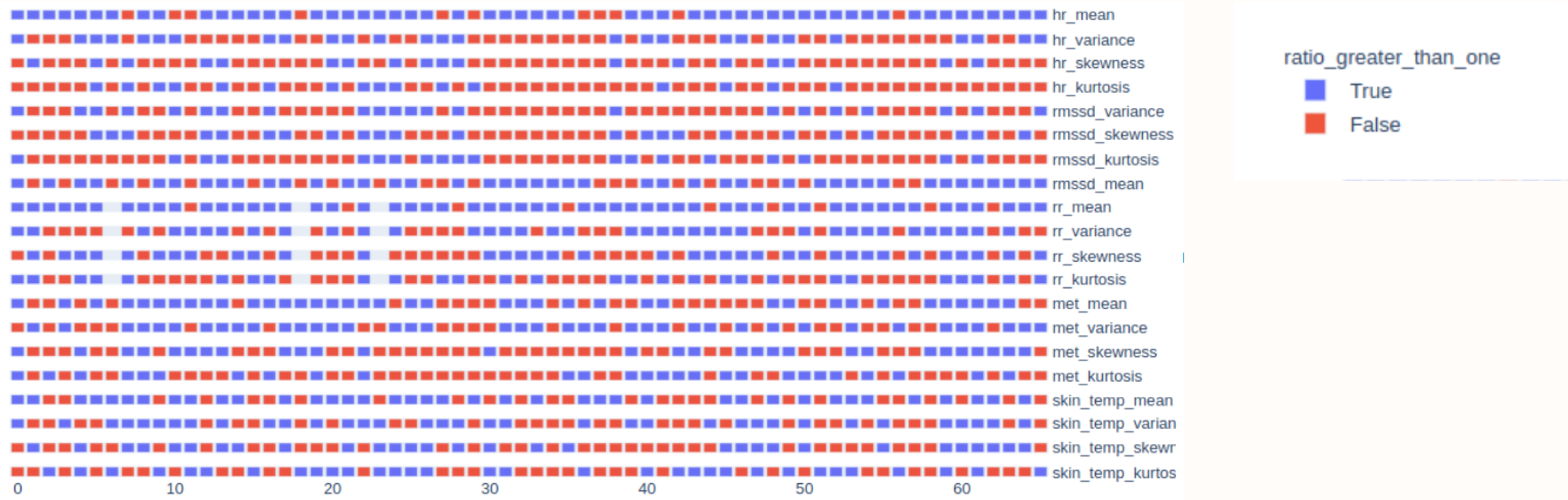
Modeling

Approach-4: CWT- Model



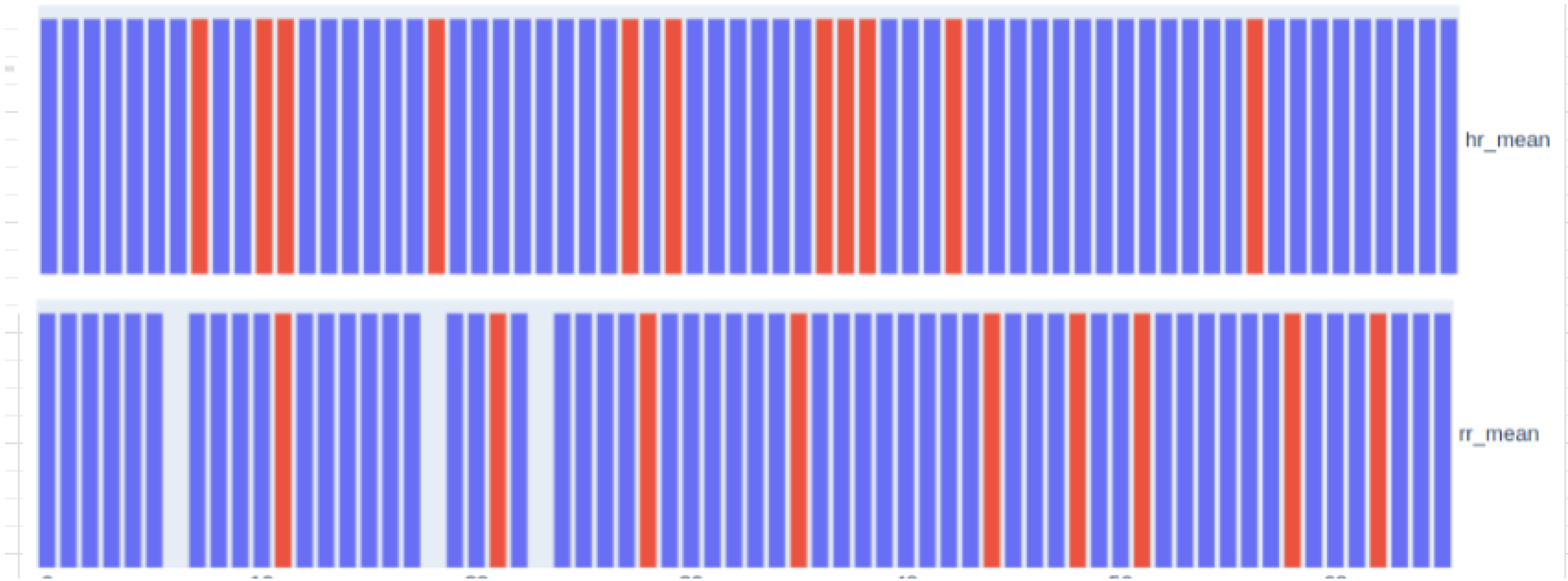
Modeling

Approach-4: CWT- Output Interpretation



Modeling

Approach-4: CWT- Parameter Selection

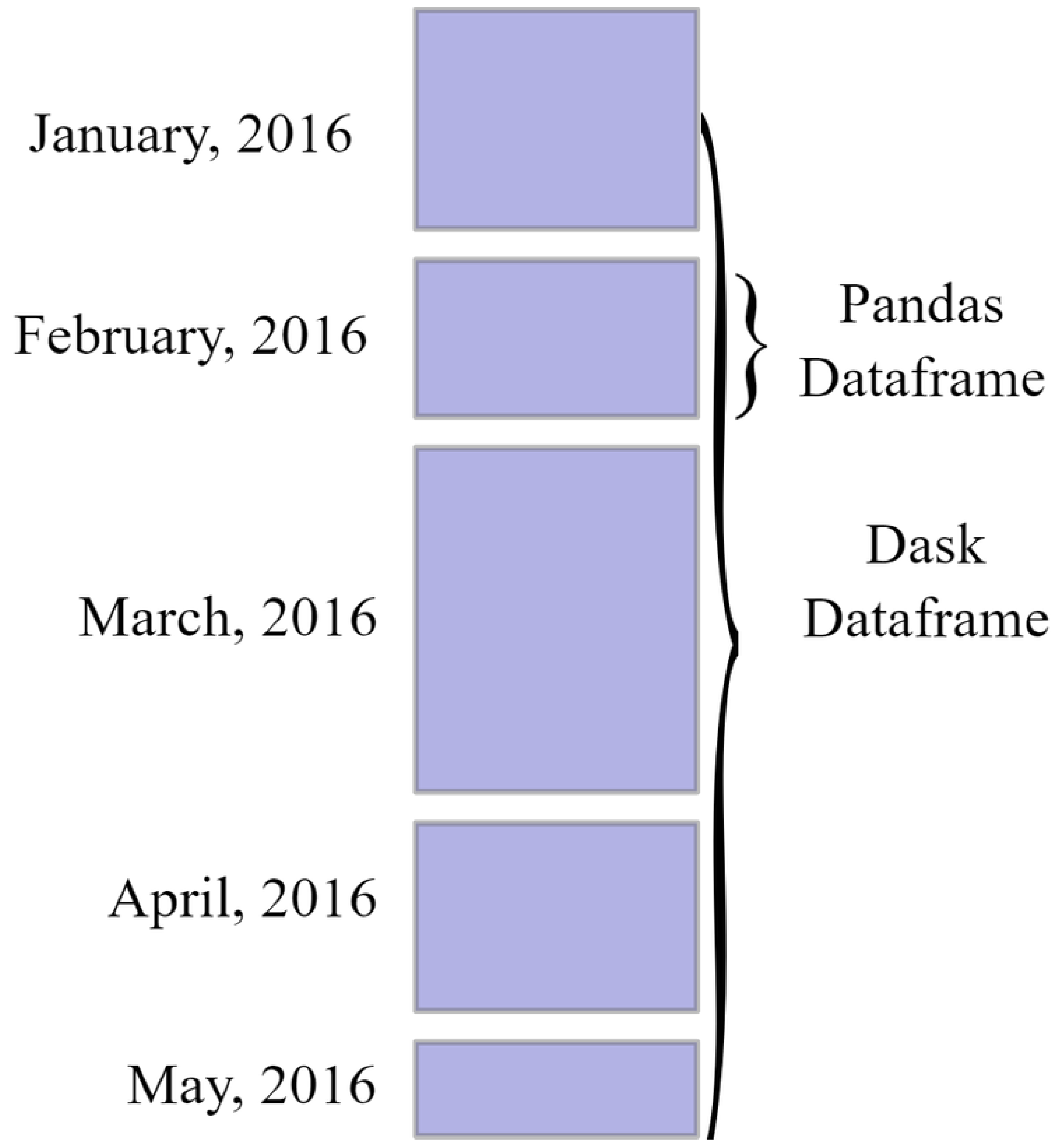


Scalability

Scalability

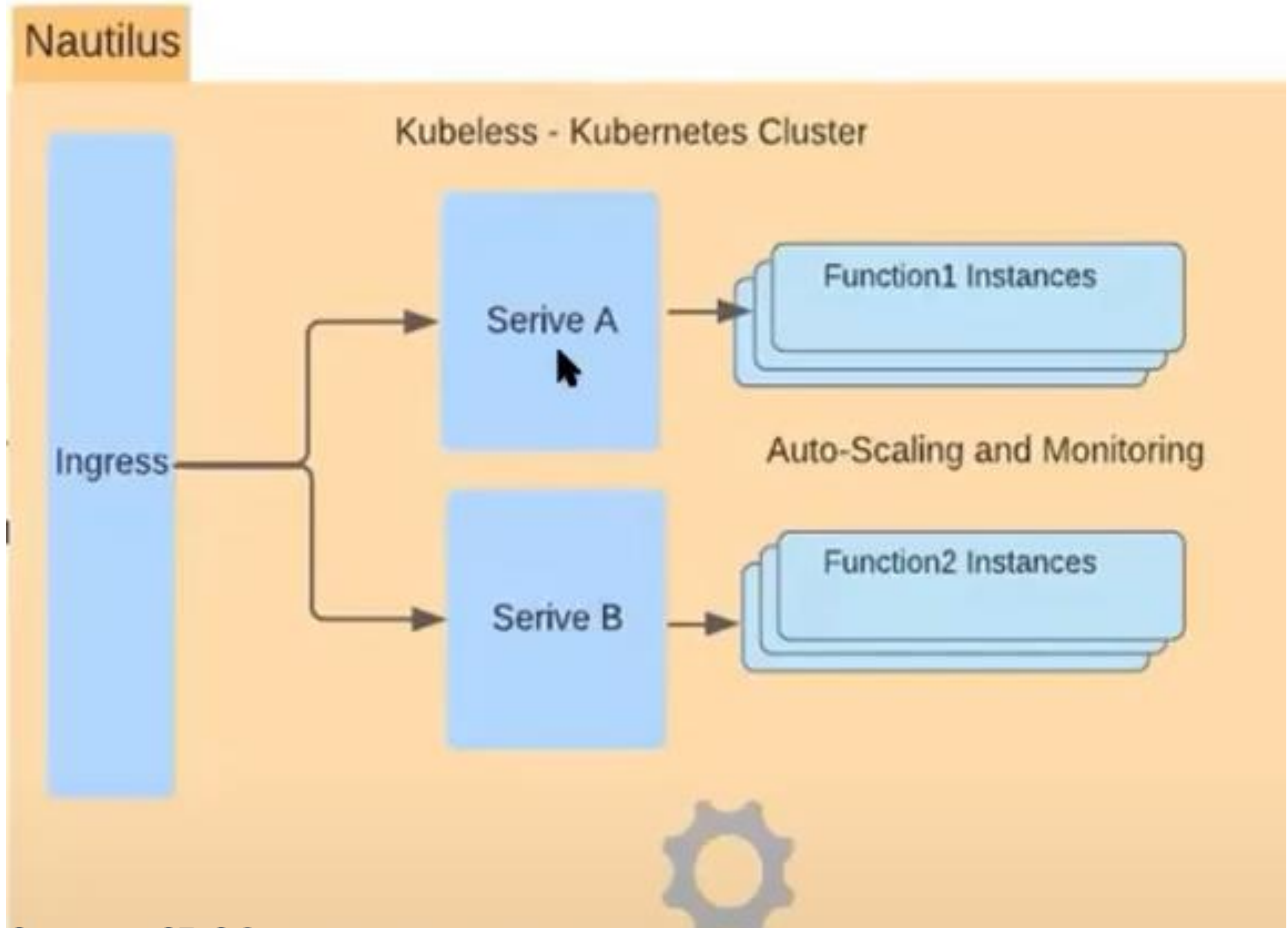
Dask

- Dask is a flexible library for parallel computing in Python
- Dask DataFrame is a large parallel DataFrame composed of many smaller Pandas DataFrames.
- Dask-DataFrames may live on disk for a single machine, or on many different machines in a cluster.
- Dask DataFrame uses the multi-threaded scheduler that exposes parallelism



Scalability

- Nautilus environment is built with Kubeless Kubernetes cluster architecture.
- Leveraging the existing auto-scalable infrastructure.
- When an instance is loaded beyond 80% by a service, a new instance will be automatically spun, and the load will be shared with the new instance.



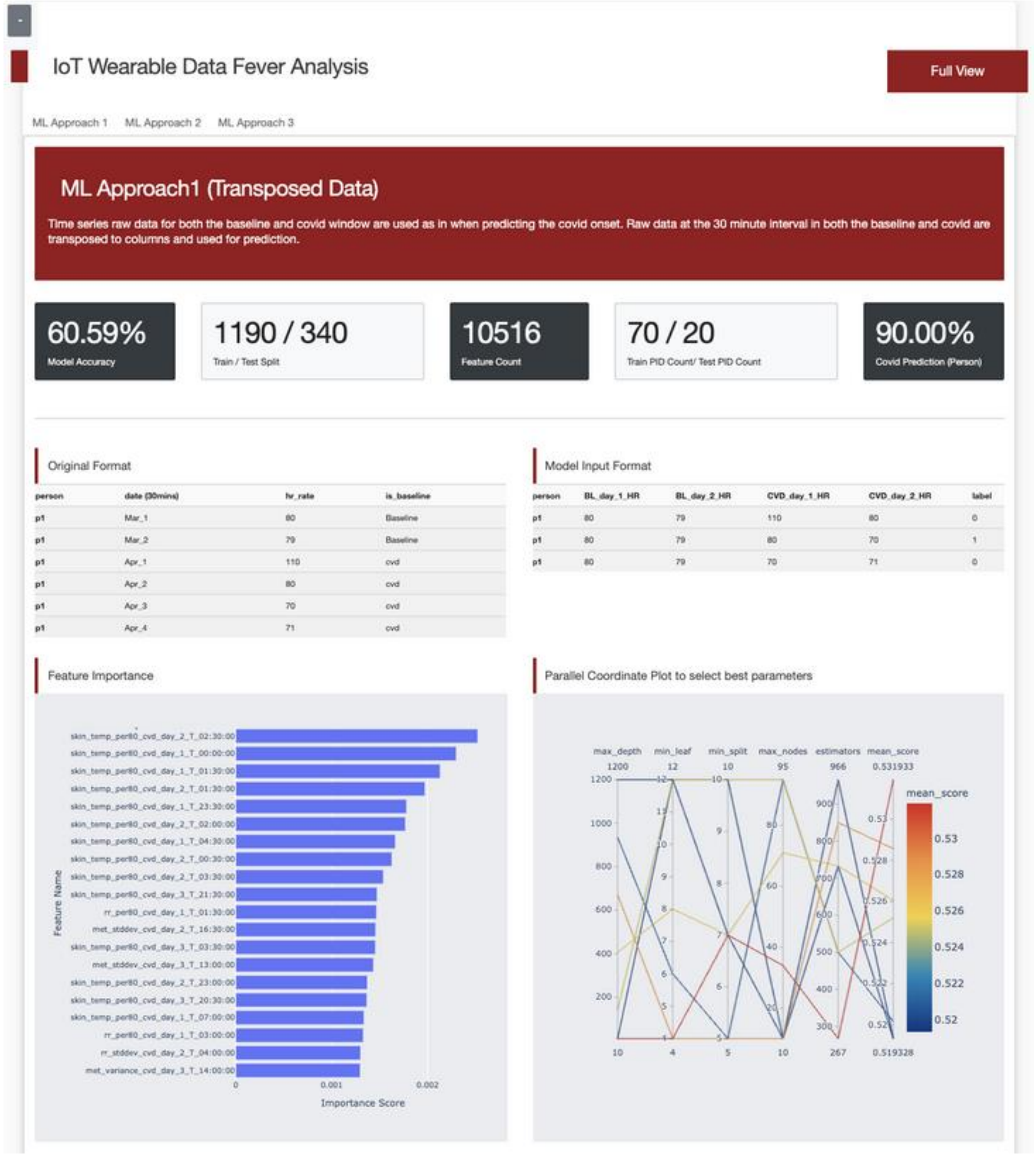
Source: SDSC

Visualization

Visualization

Tools & Techniques

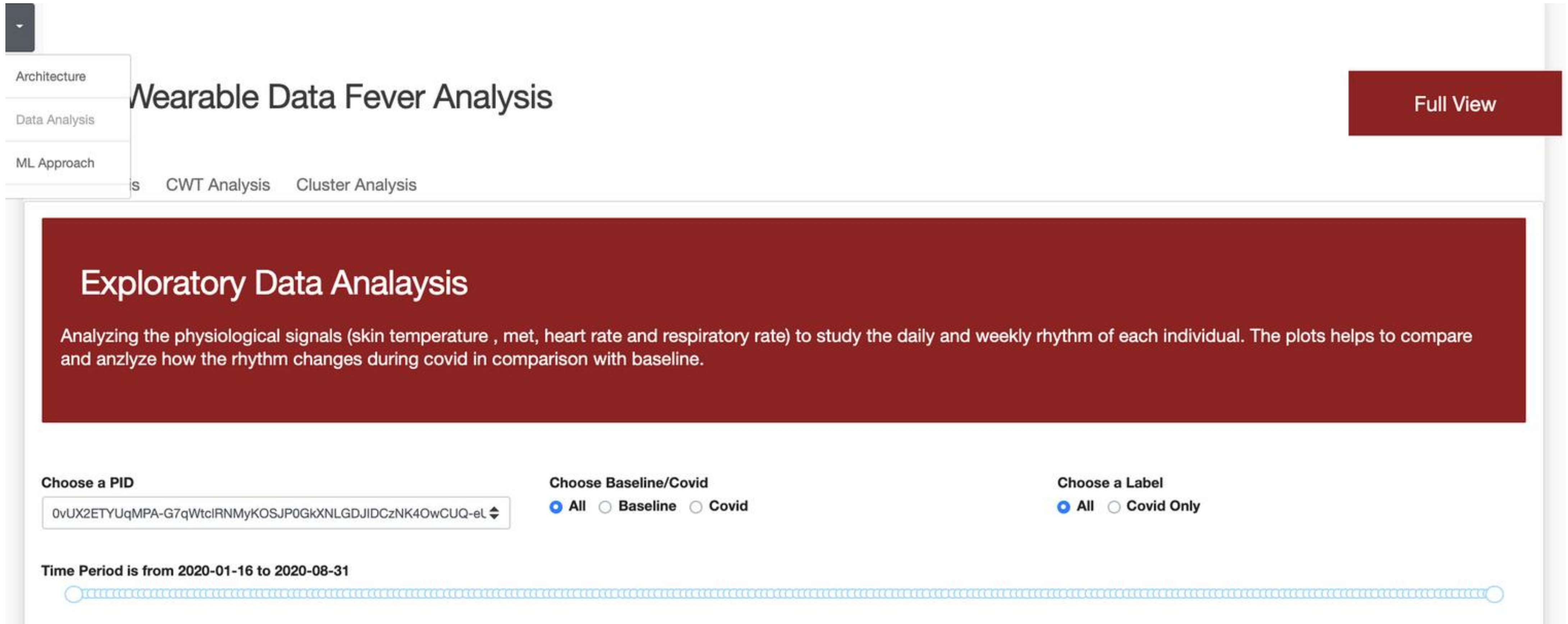
- Load Pickles from nautilus
- Ploty plots
- Plots integrated with DASH
- HTML,CSS and Dash-Bootstrap

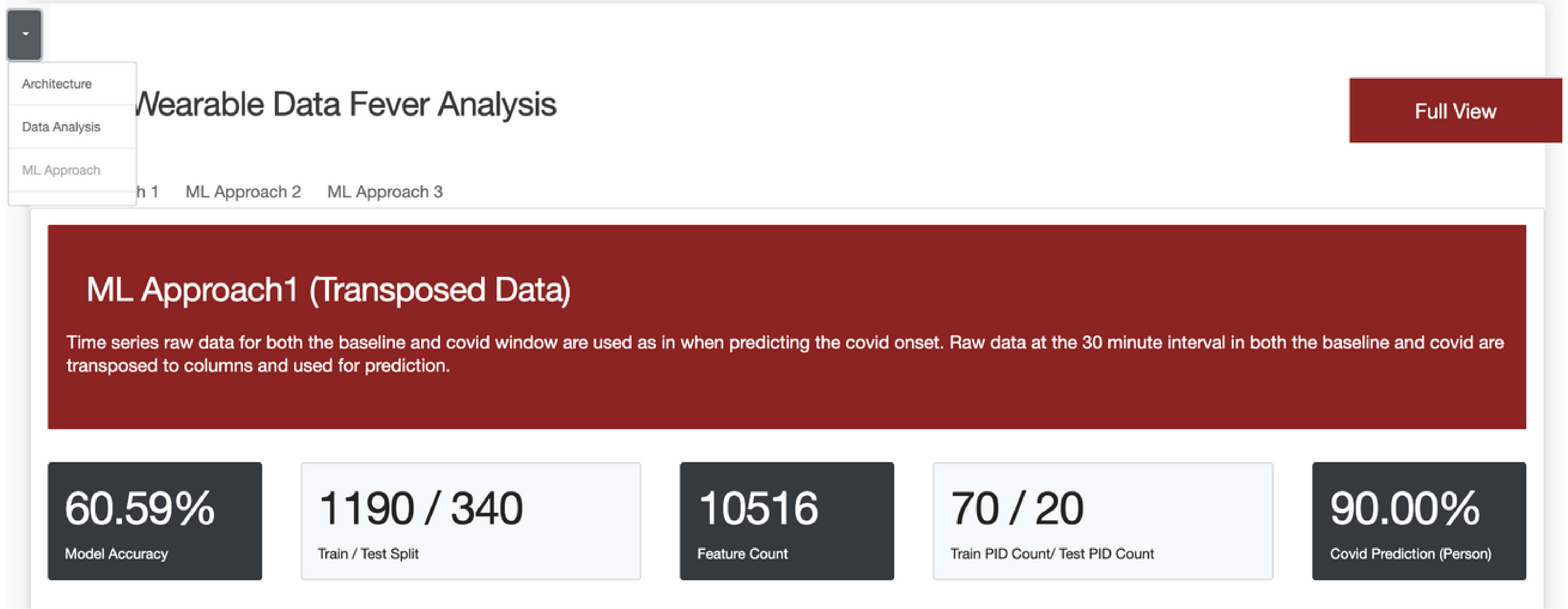


Visualization

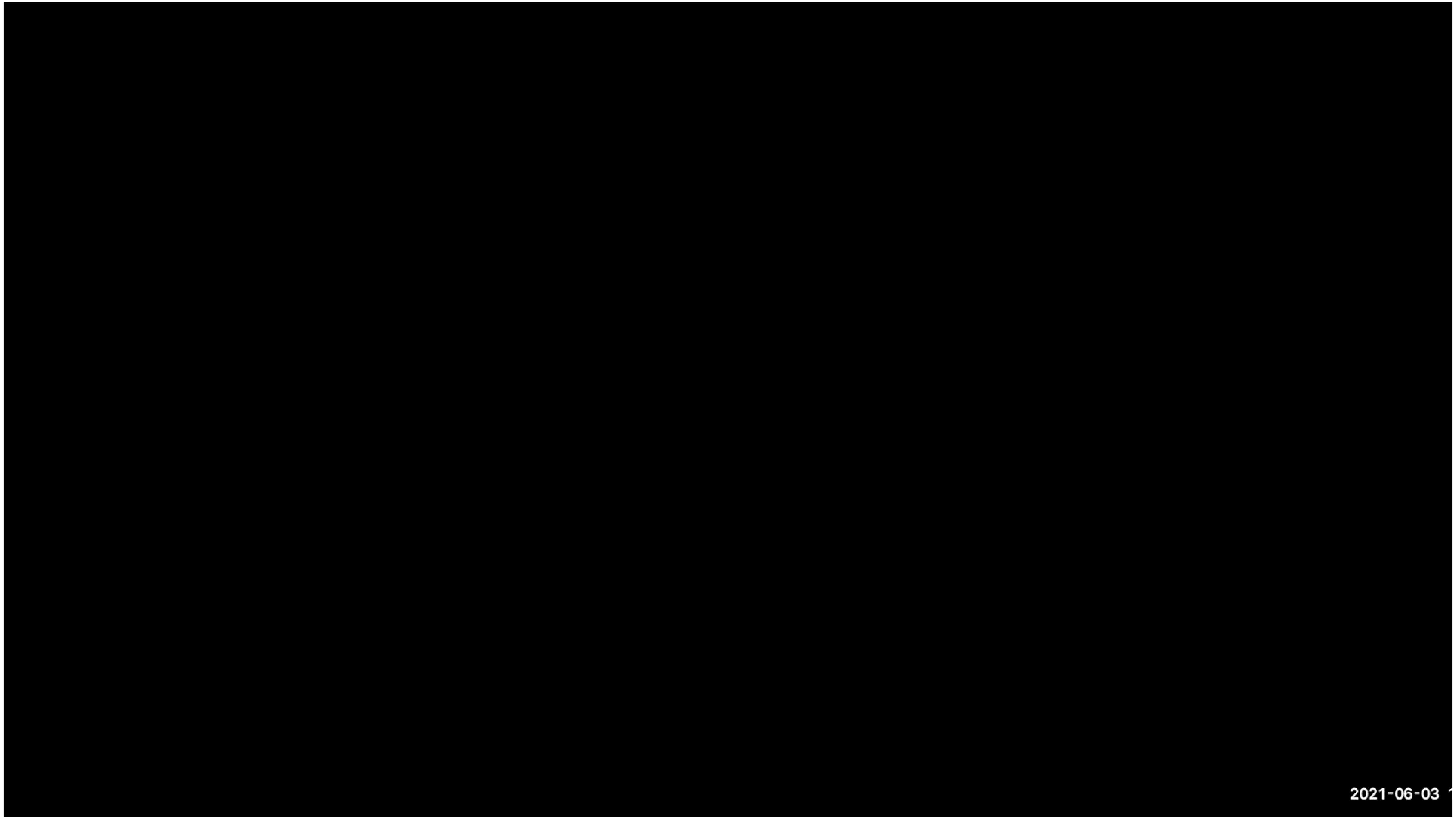
Setup & Extensibility

- Virtual Environment
  - Why
  - How
- Easy to Extend for more approaches
- Easy to add more plots





Demo



Key Findings & Future Use

Key Findings

- Skin temperature is the key feature predicting Covid in the time domain
- Heart rate and Respiratory rate are important features in the frequency domain
- The model performs better when all three variables are combined.
- Skin temperatures are higher for female than male (healthy and covid window)
- Higher-order features of baseline are the key features in the baseline aggregation approach
  Ratio : skin temperature / met, HR / RMSSD
  Deviation: the difference between covid to corresponding three baselines

Key Findings

- Baseline Data was helpful in reducing false positive results
- TSFEL approach signifies entropy (amount of uncertainty) and negative turning points are the important features for the predictions
- Most of the models' true prediction is around the Symptoms onset date or the calculated px date (Calculated physiological max date)

Future Use

- The framework allows easy integration of new physiological data. Ex. Sleep Summary
- The framework facilitates the development and evaluation of new model approaches or selecting dynamic baseline
- Extend the research for other Medical diagnosis like pregnancy prediction
- Ease of collaboration among different stakeholders (for eg. clinicians, algorithm developers & physicians)

Acknowledgment

Prof. Ilkay Altintas
Program Advisor

Prof. Benjamin Smarr
Academic Advisor

Joseph Natale
postdoctoral researcher

Subhasis Dasgupta
Computational and Data Science Researcher Specialist level 4

Varun Viswanath
Ph.D. student, Dept. of Electrical and Computer Engineering

Shweta Purawat
Computational and Data Science Researcher Specialist level 4

Paul Norton
Security specialist, San Diego Supercomputer Center

Danielle Whitehair
Technical project manager, San Diego Supercomputer Center

Shakti Davis
Engineer, MIT Lincoln Lab

Thank you!

Questions