IDENTIFICATION OF SUB-PHENOTYPES OF COVID-19 WITHIN PATIENT POPULATION

Project 13: ***COVID Sub phenotyping Project Proposal***

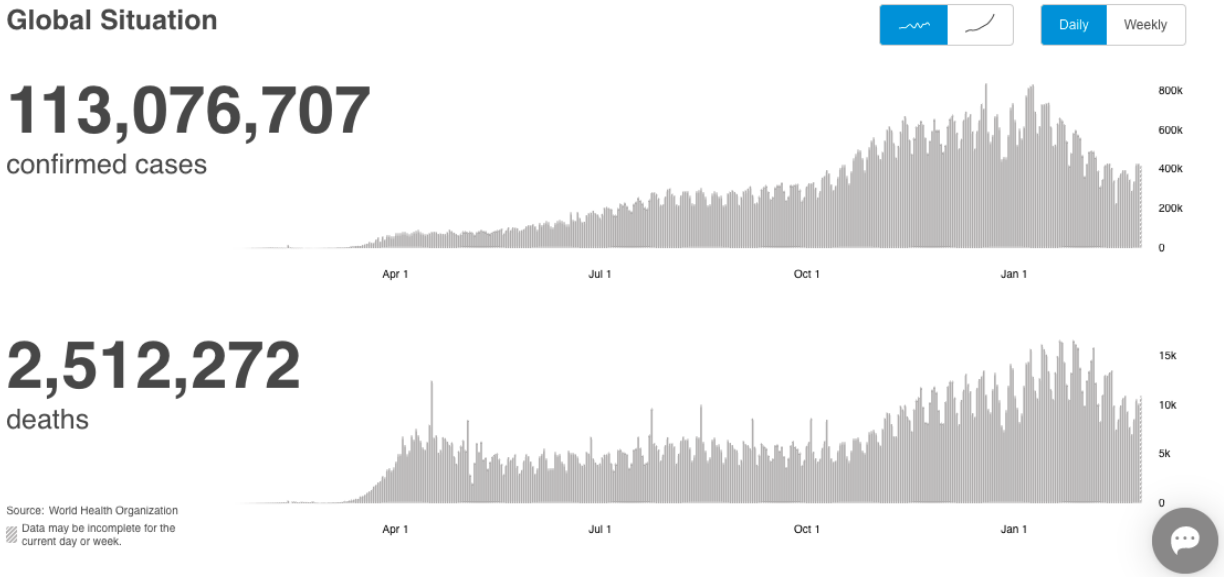
BMED 8813 BHI Presenter: **G-6**

Seonggeon Cho, Rohan Bhukar, Bryce Butler, Zhonghao Dai




Data driven clinical-decision making is required for better prognosis of disease

- Millions of deaths worldwide
- More than 12,000 mutations reported
- Variety of symptoms based on patients' preconditions and type of covid variants.
- Due to this variability, clinical-decision making is challenging



WHO, 2021

Identification of COVID-19 subphenotypes could lead to better understanding of the diverse host responses that result in these heterogeneous presentations.

Current challenges of COVID subphenotyping

Limited availability of COVID-19 patient data

- Lack Long term follow up
- Partial availability of medical health record
- Current models limited to hospitalized patients
- Current models limited to Age ≥ 60

Data Modalities: COVID-19 DREAM CHALLANGE

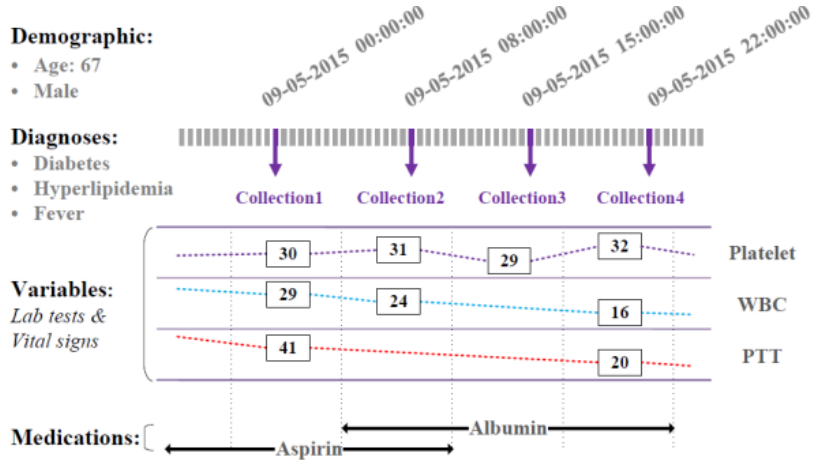
- **Time independent data**
  - Patient's demographic profile
    - Age
    - Gender
    - Race

- **Time dependent data**
  - Medical condition occurrence & timespan
    - Osteoarthritis, Lung fibrosis, bronchitis, etc.
  - Drug exposure & timespan of administration
  - Procedure occurrence & Device exposure

- **Short-term Time series data**
  - Vital signs measurement

- **Ground Truth**
  - Whether the patient was hospitalized within 3-week after positive diagnostic of COVID-19
    - 0 or 1 (Binary)

Characteristics	Total	Group 1	Group 2	Group 3	Group 4	*P* Value
No. (%)	696	139 (20)	97 (14)	277 (40)	183 (26)	...
Age, median (IQR), y	61 (47-73)	57 (42-71)	58 (49-73)	60 (44-72)	64 (50-78)	.04
Sex, male, No. (%)	355 (51)	77 (55.4)	54 (55.7)	133 (48)	91 (49.7)	.4
Race, No. (%)						.08
Black	588 (84.5)	121 (87.1)	81 (83.5)	235 (84.8)	151 (82.5)	...
White	44 (6.3)	6 (4.3)	6 (6.2)	19 (6.9)	13 (7.1)	...
Other	64 (9.2)	12 (8.6)	10 (10.3)	23 (8.3)	19 (10.4)	...
Comorbidity, No. (%)						...
Congestive heart failure	154 (22.1)	28 (20.1)	14 (14.4)	54 (19.5)	58 (31.7)	.002
Pulmonary disease	166 (23.9)	24 (17.3)	17 (17.5)	68 (24.5)	57 (31.1)	.01
Diabetes mellitus	92 (13.2)	20 (14.4)	11 (11.3)	38 (13.7)	23 (12.6)	.9
Hypertension	233 (33.5)	48 (34.5)	35 (36.1)	94 (33.9)	56 (30.6)	.8
Renal disease	41 (5.9)	7 (5)	4 (4.1)	14 (5.1)	16 (8.7)	.3
Liver disease	14 (2)	2 (1.4)	0 (0)	8 (2.9)	4 (2.2)	.3
BMI, kg/m^2	31 (10)	34 (11)	32 (8)	31 (10)	29 (8)	< .001



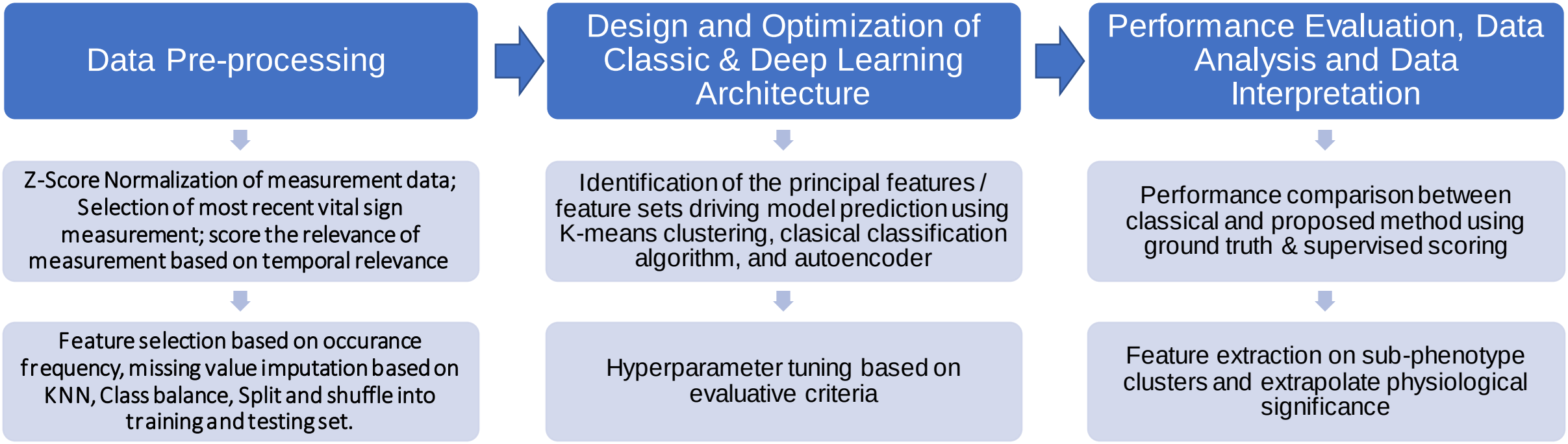
EHR Data : COVID-19 DREAM CHALLANGE

Data file	Training set	Evaluation set
Measurement data	197,498 x 20	88,996 x 20
Gold standard data	1251 x 2	536 x 2
Person data	1,251 x 18	536 x 18
Condition occurrence data	90,424 x 16	37,395 x 16
Device exposure data	27 x 15	10 x 15
Drug exposure	42,187 x 23	25,250 x 23
Observation data	26,674 x 18	12,794 x 18
Observation period	1,251 x 5	536 x 5
Procedure Occurrence data	1,420 x 14	781 x 5
Visit Occurrence	42,515 x 17	17,362 x 5
Total patients	1251	536

```
[7] df = pd.read_csv('/content/drive/MyDrive/bmed_8813/final_project/q2_synthetic_data_08-19-2020/release_08-19-2020/training/measurement.csv',sep=',')

[ ] df.shape

    (197498, 20)

[ ] df.head()
```

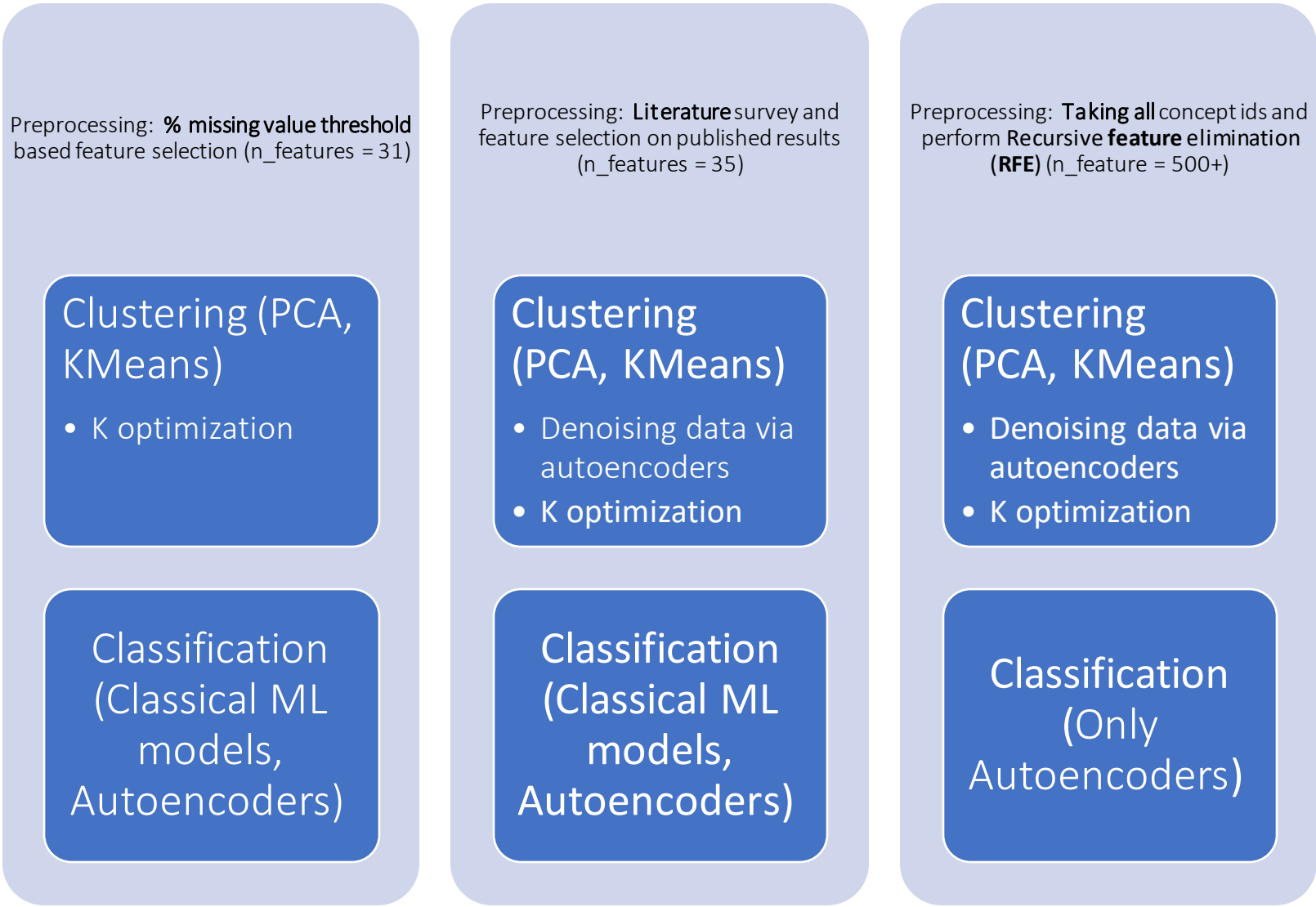
	person_id	measurement_id	measurement_concept_id	measurement_date	measurement_datetime	measurement_time	measurement_type_concept_id	operator_concept_id	value_as_
0	516	1	3000905	2015-11-14	2015-11-14 14:41:00	2018-07-14	44818702	4172703.0	
1	1193	2	3028288	2013-01-24	2013-01-24 14:41:00	2015-12-28	44818702	4172703.0	
2	949	3	3027114	2017-09-06	2017-09-06 14:41:00	2017-06-20	44818702	4172703.0	
3	1059	4	3012030	2018-12-23	2018-12-23 14:41:00	2019-02-26	44818702	4172703.0	
4	348	5	3016723	2012-10-26	2012-10-26 14:41:00	2019-03-01	44818702	4172703.0	

System Workflow

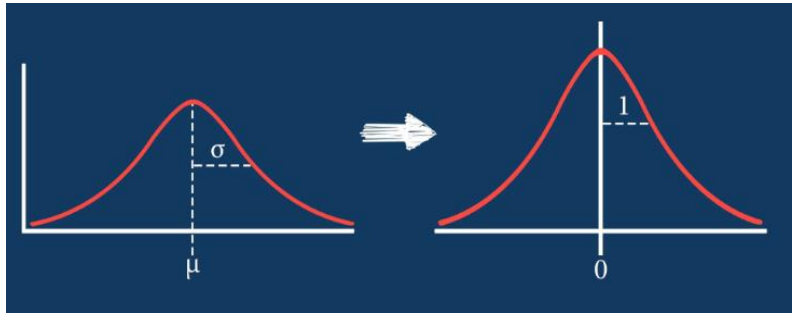


- Finally, evaluate the models' performance with Silhouette Coefficient, Dunn's Index – which are **Clustering Performance Evaluation Metrics for unsupervised scenario.**
- **Hospitalization status will taken as one critical variable for consideration to evaluating the results of sub-phenotyping.**

Current project workflow: 3 phase optimization



Data preprocessing



- **Z-score normalization of measurement data**
  - $Z = (x – \mu) / \sigma$
  - feature scaling to plot different variables on the same scale
    - Ex. Platelet count (100-300*10e9) vs IL-6 conc (0.1-5.9*10e-6)

- **Selection of most recent vital sign measurement**
  - Most recent vital sign measurement is more likely to represent patient's condition

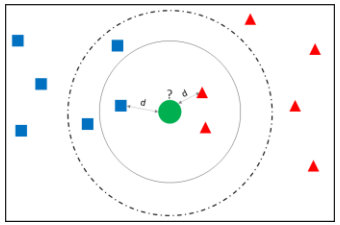
person_id	measure_concept_id	measurement_date	value_as_number	range_low	range_high	unit_source_value
1	3016723	4/1/2010	0.54	0.38	1.02	mg/dL
1	3016723	11/15/2010	0.68	0.2	1.1	mg/dL
1	3016723	4/7/2012	3.53	0.51	1.18	mg/dL
1	3016723	4/1/2014	0.7	0.38	1.02	mg/dL
1	3016723	4/7/2015	0.71	0.38	1.02	mg/dL
1	3016723	11/9/2015	0.8	0.51	1.18	mg/dL
1	3016723	9/8/2017	0.91	0.38	1.02	mg/dL
1	3016723	7/19/2019	1.45	0.51	1.18	mg/dL
1	3016723	8/10/2019	0.77	0.51	1.18	mg/dL
1	3016723	4/15/2020	0.89	0.51	1.18	mg/dL

- **Score the quality of measurement data based on temporal relevance**
  - Convert measurement data → temporal relevancy of the measurement data
  - Ex. Data measured in 2020 → 100% relevant; Data measured in 2018 → 80% relevant, etc.
  - Measurement data spam from January 2010 to December 2020, old measurement data are likely to be outdated.

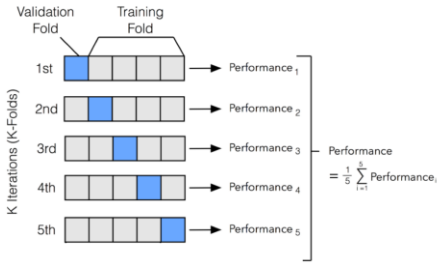
Data preprocessing (Cont.)

- **Measurement feature selection based on occurrence condition.**
  - If certain vital signs concepts are rarely measured, it is likely that these measurements does not drive model prediction and exist as noise.
  - 70% frequency is selected as cutoff, reducing feature dimensions from 490 to 64.

- **Missing Value imputation**
  - Each Patient is likely to lack measurement value of several vital signs used for after feature selection.
  - KNN imputation is used to fill in missing measurement values based of nearest neighbor estimated values.



- **Class Balance; Training and Testing set generation**
  - Class 1 (Hospitalized): 107 patients
  - Class 0 (Non-hospitalized): 1144 patients
  - Class 0 groups are randomly shuffled and 107 patients are selected to match balance the class ratio
  - 5-fold cross validation are performed on the 214 training sets, the generated model based of training set data will be applied on testing set without class balancing.

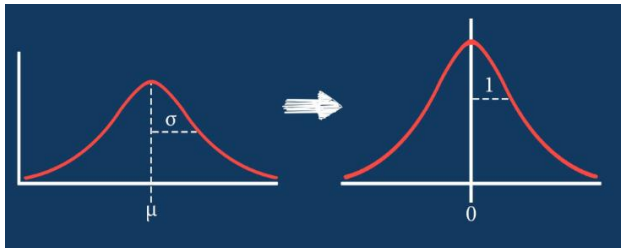


Data preprocessing

- **In total 38 features selected from literature survey (published Covid-19 papers)**
- **Features extracted from the multiple tables and data selected based on temporal relevance**
- **One-hot encoding the categorical data (sex, race, etc.)**
- **2 features were dropped due to very high missingness (> 80%)**
- **1 feature was dropped based on removal of quasi-constant features**
- **35 features for further processing**
- **Z-score normalization of data**
  - $Z = (x - \mu) / \sigma$
  - Dataset further used as input for clustering
- **Selection of most recent vital sign measurement**
  - Most recent vital sign measurement is more likely to represent patient's condition

Table1: Candidate features for diagnosis aid model

Groups	Candidate features				
Demographic information	Age	Gender			
Vital signs	Temperature (TEM)	Heart rate (HR)	Diastolic blood pressure (DIAS_BP)	Systolic blood pressure (SYS_BP)	
Blood routine values	White blood cell count (WBC)	Red blood cell count (RBC)	Hemoglobin (HGB)	Hematocrit (HCT)	Platelet count (PLT)
	Mean platelet volume (MPV)	Lymphocyte ratio (LYMPH%)	Lymphocyte count (LYMPH#)	Neutrophil ratio (NEUT%)	Neutrophil count (NEUT#)
	Eosinophil ratio (EO%)	Eosinophil count (EO#)	Monocyte ratio (MONO%)	Monocyte count (MONO#)	Basophil ratio (BASO%)
	Basophil count (BASO#)	Mean corpuscular volume (MCV)	Mean corpuscular hemoglobin content (MCH)	Mean corpuscular hemoglobin concentratio (MCHC)	Red blood cell volume distribution width (RDW-CV)
Clinical signs and symptoms on admission	Fever	Cough	Shortness of breath	Muscle ache	Headache
	Rhinorrhoea	Diarrhoea	Nausea	Vomiting	Chills
	Expectoration	Nasal congestion	Abdominal pain	Fatigue	Palpitation
	Sore throat	Shiver	Fever classification (FC)		
Infection-related biomarkers	C-reactive protein (CRP)	Interleukin-6 (IL-6)			
Others	Days from illness onset to first admission (DOA)				

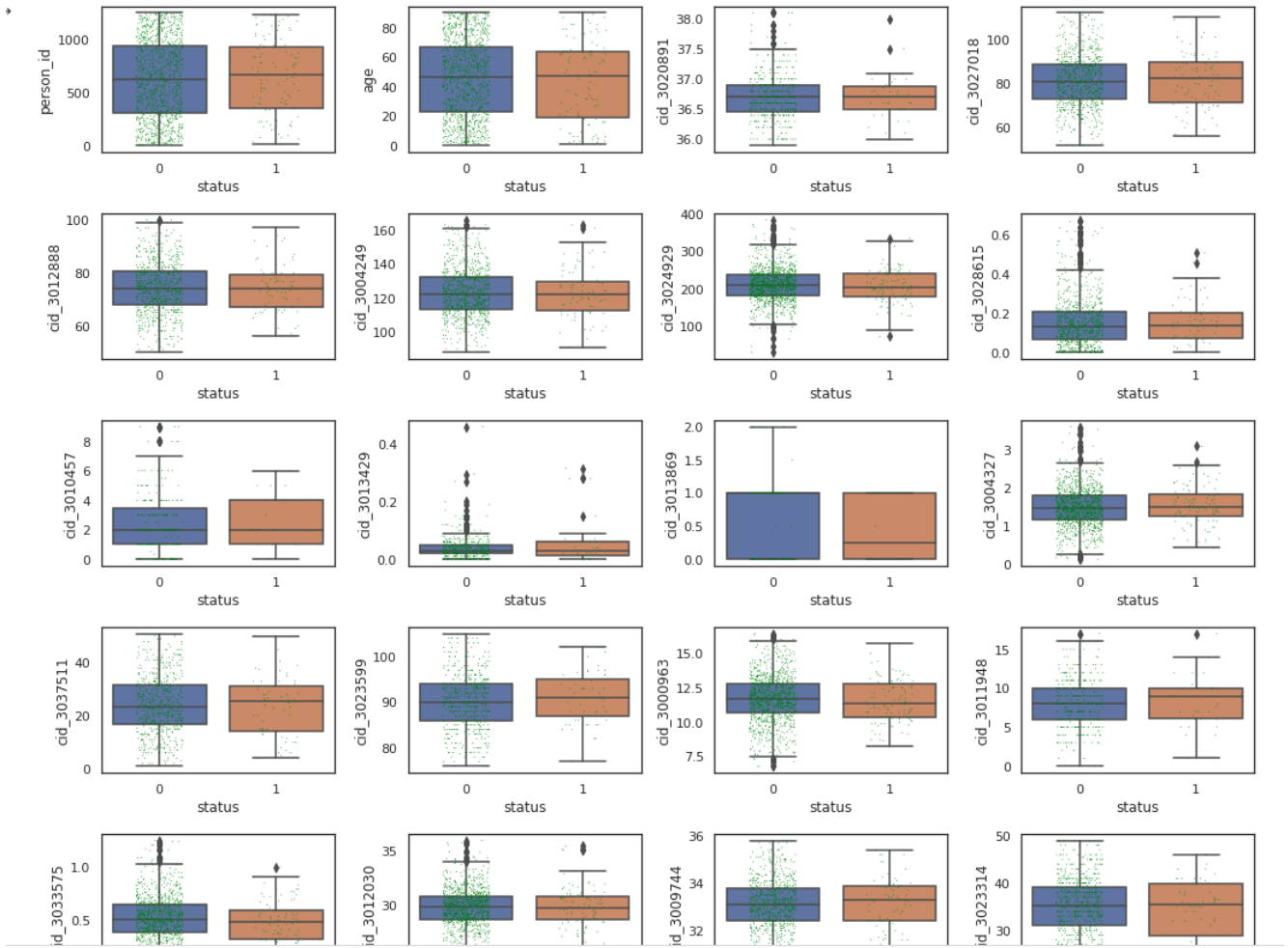


person_id	measure_concept_id	measurement_date	value_as_number	range_low	range_high	unit_source_value
1	3016723	4/1/2010	0.54	0.38	1.02	mg/dL
1	3016723	11/15/2010	0.68	0.2	1.1	mg/dL
1	3016723	4/7/2012	3.53	0.51	1.18	mg/dL
1	3016723	4/1/2014	0.7	0.38	1.02	mg/dL
1	3016723	4/7/2015	0.71	0.38	1.02	mg/dL
1	3016723	11/9/2015	0.8	0.51	1.18	mg/dL
1	3016723	9/8/2017	0.91	0.38	1.02	mg/dL
1	3016723	7/19/2019	1.45	0.51	1.18	mg/dL
1	3016723	8/10/2019	0.77	0.51	1.18	mg/dL
1	3016723	4/15/2020	0.89	0.51	1.18	mg/dL

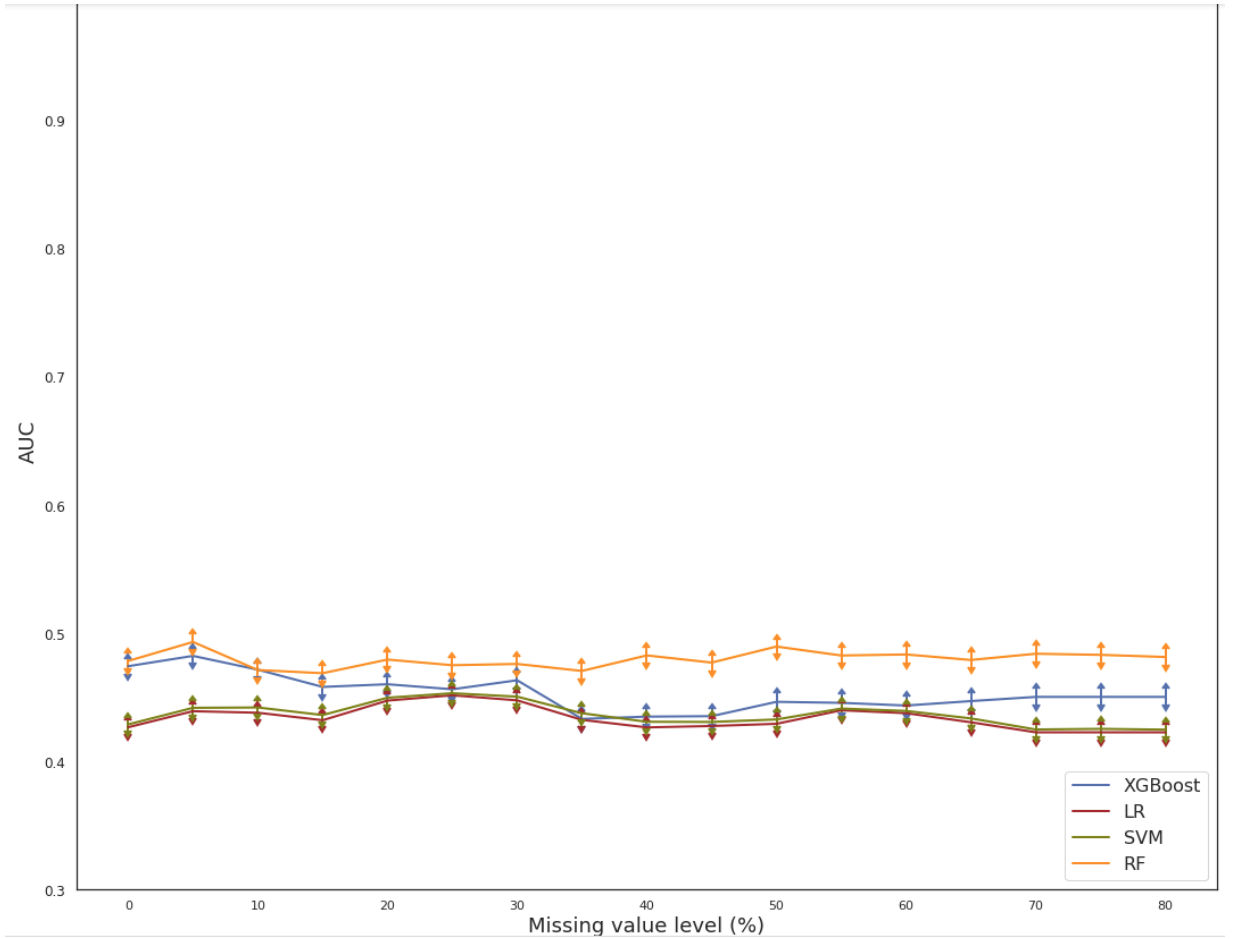
- **Score the quality of measurement data based on temporal relevance**
  - Convert measurement data → temporal relevancy of the measurement data
- For Classification,
  - Split dataset into train and validation sets, then feature scaling and finally imputation using KNN (mean and median were tried but resulted in sub-optimal performance)

https://www.medrxiv.org/content/10.1101/2020.03.19.20039099v1

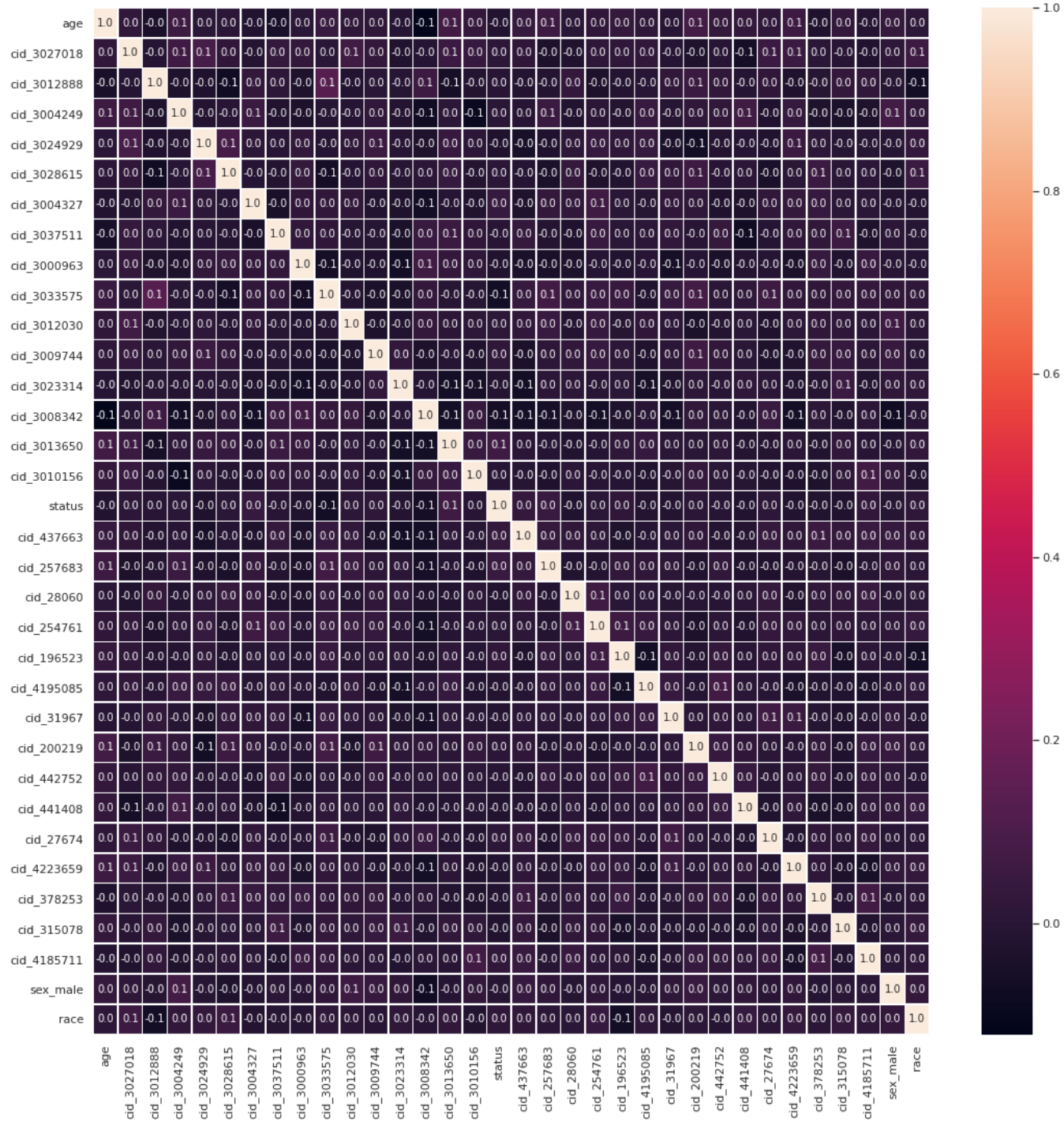
Data preprocessing (Cont.)



Boxplot: for each feature we found no extreme values

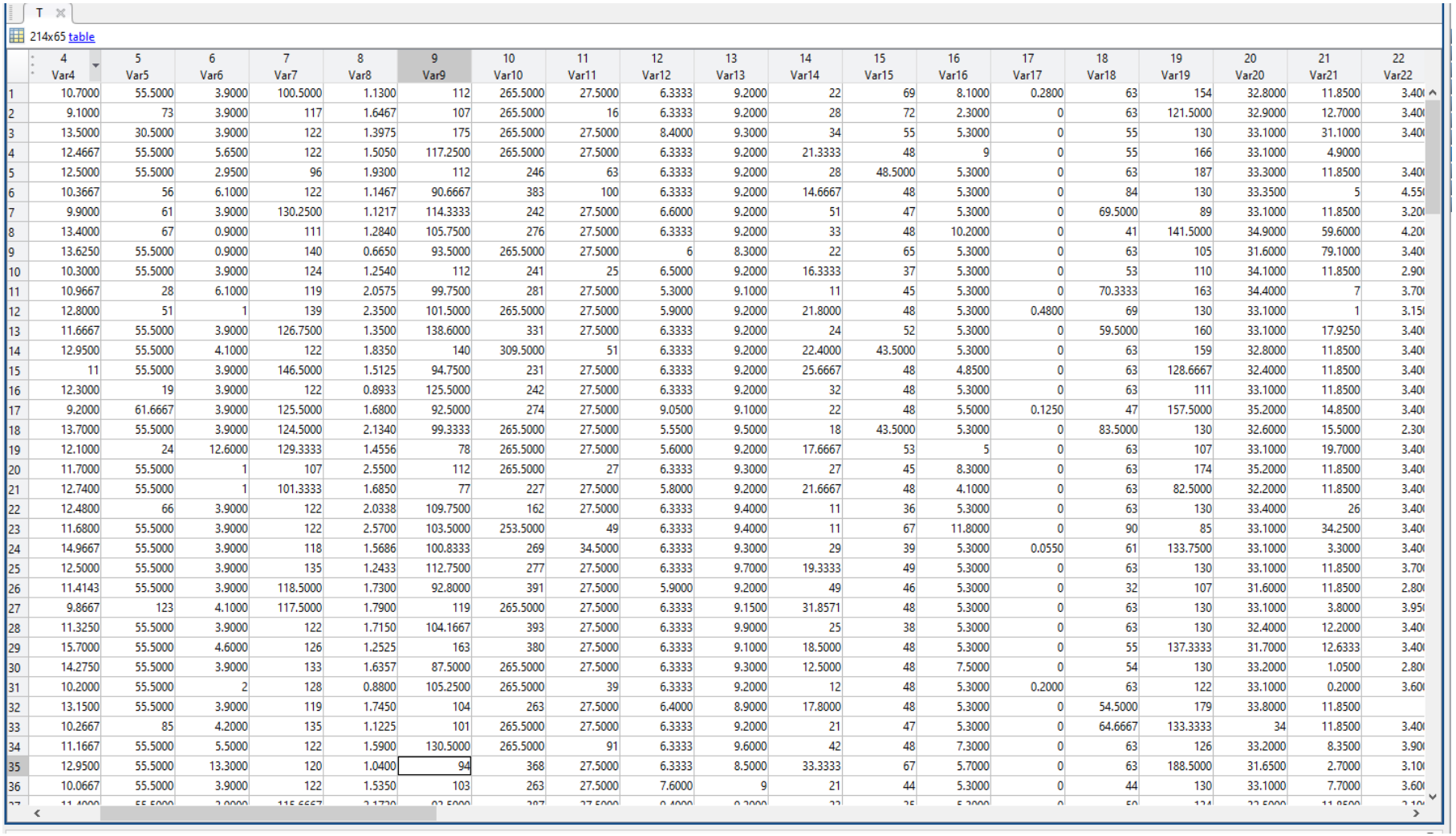


Upon imputation at different thresholds of missing value, found that highest classification AUCs achieved at 50% missing value imputation.
Dataframe with <=50% missing feature values was created and further imputed using KNN. Additional, column of Body_temp was added manually due very small margin of missingness.



Correlation matrix plot:
None of the selected and processed features were found to be correlated.

Data format post data pre-processing



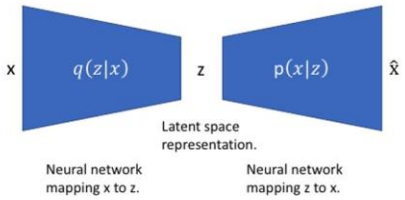
T

214x65 table

	4 Var4	5 Var5	6 Var6	7 Var7	8 Var8	9 Var9	10 Var10	11 Var11	12 Var12	13 Var13	14 Var14	15 Var15	16 Var16	17 Var17	18 Var18	19 Var19	20 Var20	21 Var21	22 Var22
1	10.7000	55.5000	3.9000	100.5000	1.1300	112	265.5000	27.5000	6.3333	9.2000	22	69	8.1000	0.2800	63	154	32.8000	11.8500	3.40
2	9.1000	73	3.9000	117	1.6467	107	265.5000	16	6.3333	9.2000	28	72	2.3000	0	63	121.5000	32.9000	12.7000	3.40
3	13.5000	30.5000	3.9000	122	1.3975	175	265.5000	27.5000	8.4000	9.3000	34	55	5.3000	0	55	130	33.1000	31.1000	3.40
4	12.4667	55.5000	5.6500	122	1.5050	117.2500	265.5000	27.5000	6.3333	9.2000	21.3333	48	9	0	55	166	33.1000	4.9000	
5	12.5000	55.5000	2.9500	96	1.9300	112	246	63	6.3333	9.2000	28	48.5000	5.3000	0	63	187	33.3000	11.8500	3.40
6	10.3667	56	6.1000	122	1.1467	90.6667	383	100	6.3333	9.2000	14.6667	48	5.3000	0	84	130	33.3500	5	4.55
7	9.9000	61	3.9000	130.2500	1.1217	114.3333	242	27.5000	6.6000	9.2000	51	47	5.3000	0	69.5000	89	33.1000	11.8500	3.20
8	13.4000	67	0.9000	111	1.2840	105.7500	276	27.5000	6.3333	9.2000	33	48	10.2000	0	41	141.5000	34.9000	59.6000	4.20
9	13.6250	55.5000	0.9000	140	0.6650	93.5000	265.5000	27.5000	6	8.3000	22	65	5.3000	0	63	105	31.6000	79.1000	3.40
10	10.3000	55.5000	3.9000	124	1.2540	112	241	25	6.5000	9.2000	16.3333	37	5.3000	0	53	110	34.1000	11.8500	2.90
11	10.9667	28	6.1000	119	2.0575	99.7500	281	27.5000	5.3000	9.1000	11	45	5.3000	0	70.3333	163	34.4000	7	3.70
12	12.8000	51	1	139	2.3500	101.5000	265.5000	27.5000	5.9000	9.2000	21.8000	48	5.3000	0.4800	69	130	33.1000	1	3.15
13	11.6667	55.5000	3.9000	126.7500	1.3500	138.6000	331	27.5000	6.3333	9.2000	24	52	5.3000	0	59.5000	160	33.1000	17.9250	3.40
14	12.9500	55.5000	4.1000	122	1.8350	140	309.5000	51	6.3333	9.2000	22.4000	43.5000	5.3000	0	63	159	32.8000	11.8500	3.40
15	11	55.5000	3.9000	146.5000	1.5125	94.7500	231	27.5000	6.3333	9.2000	25.6667	48	4.8500	0	63	128.6667	32.4000	11.8500	3.40
16	12.3000	19	3.9000	122	0.8933	125.5000	242	27.5000	6.3333	9.2000	32	48	5.3000	0	63	111	33.1000	11.8500	3.40
17	9.2000	61.6667	3.9000	125.5000	1.6800	92.5000	274	27.5000	9.0500	9.1000	22	48	5.5000	0.1250	47	157.5000	35.2000	14.8500	3.40
18	13.7000	55.5000	3.9000	124.5000	2.1340	99.3333	265.5000	27.5000	5.5500	9.5000	18	43.5000	5.3000	0	83.5000	130	32.6000	15.5000	2.30
19	12.1000	24	12.6000	129.3333	1.4556	78	265.5000	27.5000	5.6000	9.2000	17.6667	53	5	0	63	107	33.1000	19.7000	3.40
20	11.7000	55.5000	1	107	2.5500	112	265.5000	27	6.3333	9.3000	27	45	8.3000	0	63	174	35.2000	11.8500	3.40
21	12.7400	55.5000	1	101.3333	1.6850	77	227	27.5000	5.8000	9.2000	21.6667	48	4.1000	0	63	82.5000	32.2000	11.8500	3.40
22	12.4800	66	3.9000	122	2.0338	109.7500	162	27.5000	6.3333	9.4000	11	36	5.3000	0	63	130	33.4000	26	3.40
23	11.6800	55.5000	3.9000	122	2.5700	103.5000	253.5000	49	6.3333	9.4000	11	67	11.8000	0	90	85	33.1000	34.2500	3.40
24	14.9667	55.5000	3.9000	118	1.5686	100.8333	269	34.5000	6.3333	9.3000	29	39	5.3000	0.0550	61	133.7500	33.1000	3.3000	3.40
25	12.5000	55.5000	3.9000	135	1.2433	112.7500	277	27.5000	6.3333	9.7000	19.3333	49	5.3000	0	63	130	33.1000	11.8500	3.70
26	11.4143	55.5000	3.9000	118.5000	1.7300	92.8000	391	27.5000	5.9000	9.2000	49	46	5.3000	0	32	107	31.6000	11.8500	2.80
27	9.8667	123	4.1000	117.5000	1.7900	119	265.5000	27.5000	6.3333	9.1500	31.8571	48	5.3000	0	63	130	33.1000	3.8000	3.95
28	11.3250	55.5000	3.9000	122	1.7150	104.1667	393	27.5000	6.3333	9.9000	25	38	5.3000	0	63	130	32.4000	12.2000	3.40
29	15.7000	55.5000	4.6000	126	1.2525	163	380	27.5000	6.3333	9.1000	18.5000	48	5.3000	0	55	137.3333	31.7000	12.6333	3.40
30	14.2750	55.5000	3.9000	133	1.6357	87.5000	265.5000	27.5000	6.3333	9.3000	12.5000	48	7.5000	0	54	130	33.2000	1.0500	2.80
31	10.2000	55.5000	2	128	0.8800	105.2500	265.5000	39	6.3333	9.2000	12	48	5.3000	0.2000	63	122	33.1000	0.2000	3.60
32	13.1500	55.5000	3.9000	119	1.7450	104	263	27.5000	6.4000	8.9000	17.8000	48	5.3000	0	54.5000	179	33.8000	11.8500	
33	10.2667	85	4.2000	135	1.1225	101	265.5000	27.5000	6.3333	9.2000	21	47	5.3000	0	64.6667	133.3333	34	11.8500	3.40
34	11.1667	55.5000	5.5000	122	1.5900	130.5000	265.5000	91	6.3333	9.6000	42	48	7.3000	0	63	126	33.2000	8.3500	3.90
35	12.9500	55.5000	13.3000	120	1.0400	94	368	27.5000	6.3333	8.5000	33.3333	67	5.7000	0	63	188.5000	31.6500	2.7000	3.10
36	10.0667	55.5000	3.9000	122	1.5350	103	263	27.5000	7.6000	9	21	44	5.3000	0	44	130	33.1000	7.7000	3.60

Design and Optimization of Classic & Deep Learning Architecture

- 3 processing pipelines are implemented:
  - Hosptialization classification based on classical architecture
    - PCA, LDA, Naïve Bayes, SVM, KNN & Ensemble
  - Hospitialization classification based on autoencoder architecture



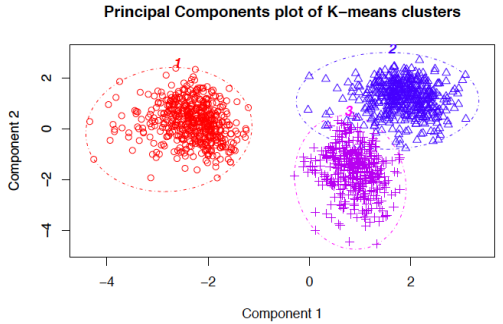
Contractive autoencoder:

$$\mathcal{J} = \sum_{x \in D} L(x, \tilde{x}(\phi, \psi, x)) + \lambda \left\| J_{\psi}(x) \right\|_F^2$$

Loss term

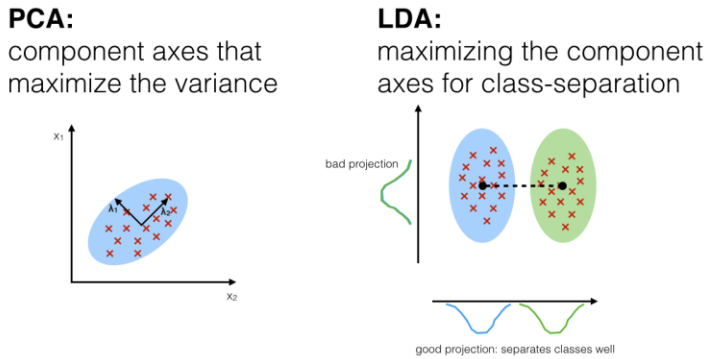
Regularization term

- Patient subpopulation clustering based on K-means clustering

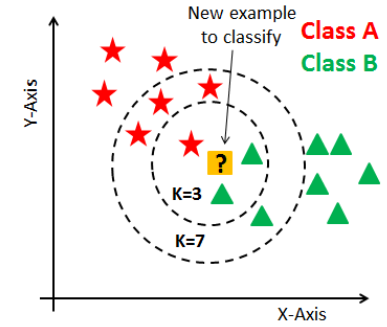


Hosptialization classification based on classical architecture

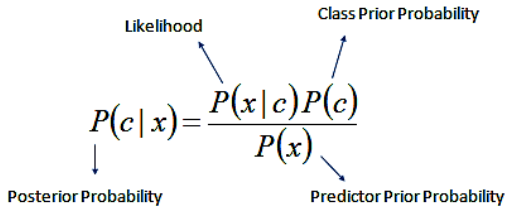
- PCA & LDA:



- KNN:



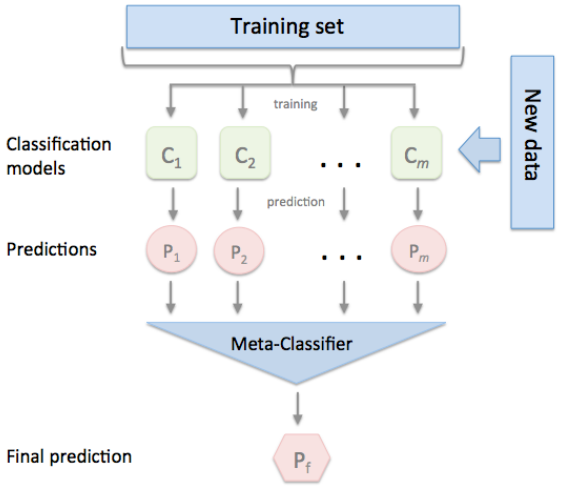
- Naïve bayes:



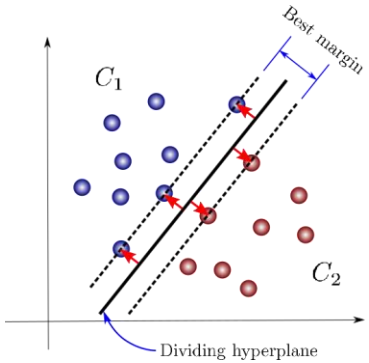
$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- Ensemble:



- SVM:



Classical Method Result

Data Browser

History

1.7 Naive Bayes Last change: Kernel Naive Bayes	Accuracy: 50.0% 64/64 features
1.8 SVM Last change: Linear SVM	Accuracy: 50.5% 64/64 features
1.9 SVM Last change: Quadratic SVM	Accuracy: 51.9% 64/64 features
1.10 SVM Last change: Cubic SVM	Accuracy: 50.9% 64/64 features
1.11 SVM Last change: Fine Gaussian SVM	Accuracy: 50.0% 64/64 features
1.12 SVM Last change: Medium Gaussian SVM	Accuracy: 50.0% 64/64 features
1.13 SVM Last change: Coarse Gaussian SVM	Accuracy: 50.0% 64/64 features
1.14 KNN Last change: Fine KNN	Accuracy: 53.7% 64/64 features
1.15 KNN Last change: Medium KNN	Accuracy: 50.0% 64/64 features
1.16 KNN Last change: Coarse KNN	Accuracy: 50.0% 64/64 features
1.17 KNN Last change: Cosine KNN	Accuracy: 50.0% 64/64 features
1.18 KNN Last change: Cubic KNN	Accuracy: 50.0% 64/64 features
1.19 KNN Last change: Weighted KNN	Accuracy: 50.0% 64/64 features
1.20 Ensemble Last change: Boosted Trees	Accuracy: 50.5% 64/64 features
1.21 Ensemble Last change: Bagged Trees	Accuracy: 50.0% 64/64 features
1.22 Ensemble Last change: Subspace Discriminant	Accuracy: 50.9% 64/64 features
1.23 Ensemble Last change: Subspace KNN	Accuracy: 55.1% 64/64 features
1.24 Ensemble Last change: RUSBoosted Trees	Accuracy: 47.2% 64/64 features

Current Model

Model 1.23: Trained

Results

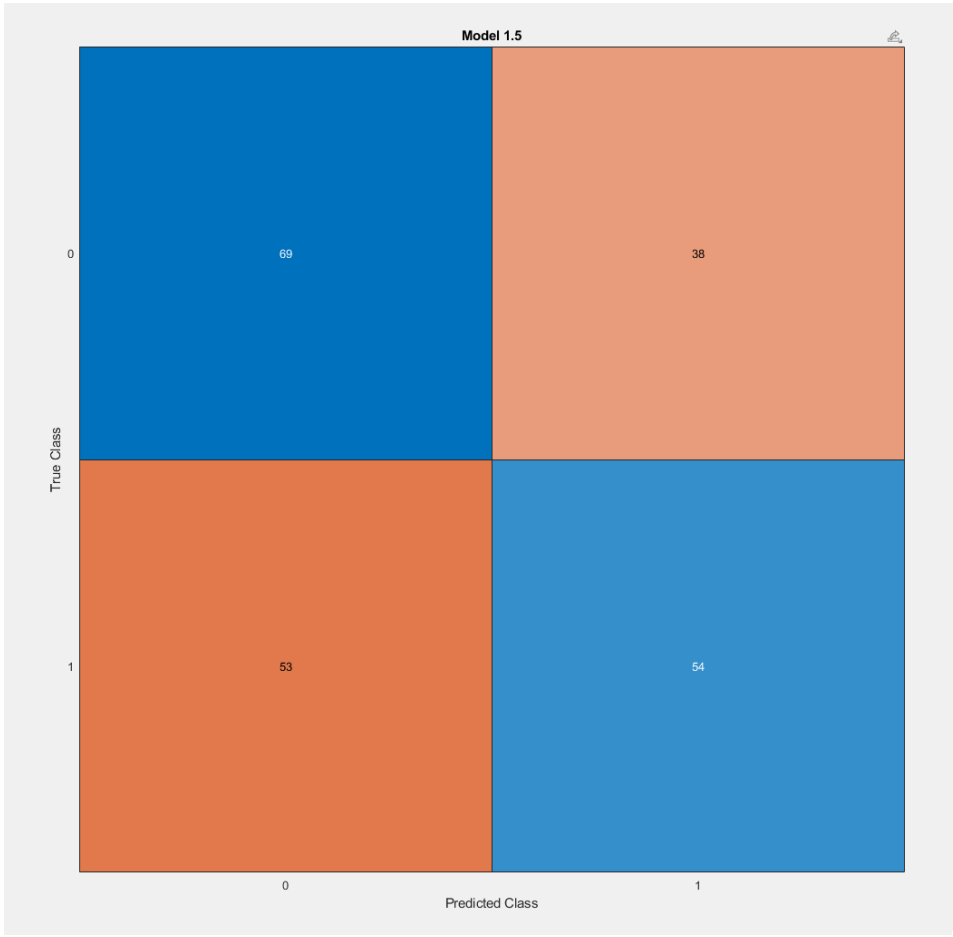
Accuracy	55.1%
Total misclassification cost	528
Prediction speed	~1600 obs/sec
Training time	1.1006 sec

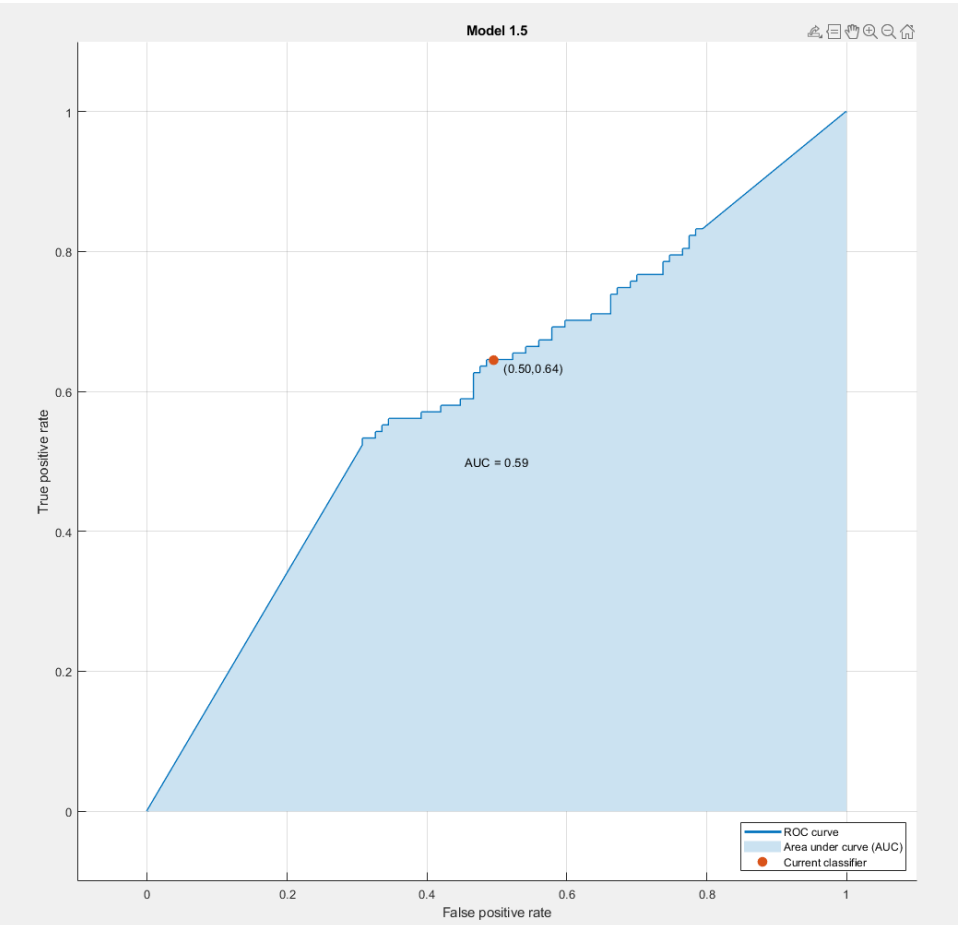
Model Type
Preset: Subspace KNN
Ensemble method: Subspace
Learner type: Nearest neighbors
Number of learners: 30
Subspace dimension: 32

Optimizer Options
Hyperparameter options disabled

Feature Selection
All features used in the model, before PCA

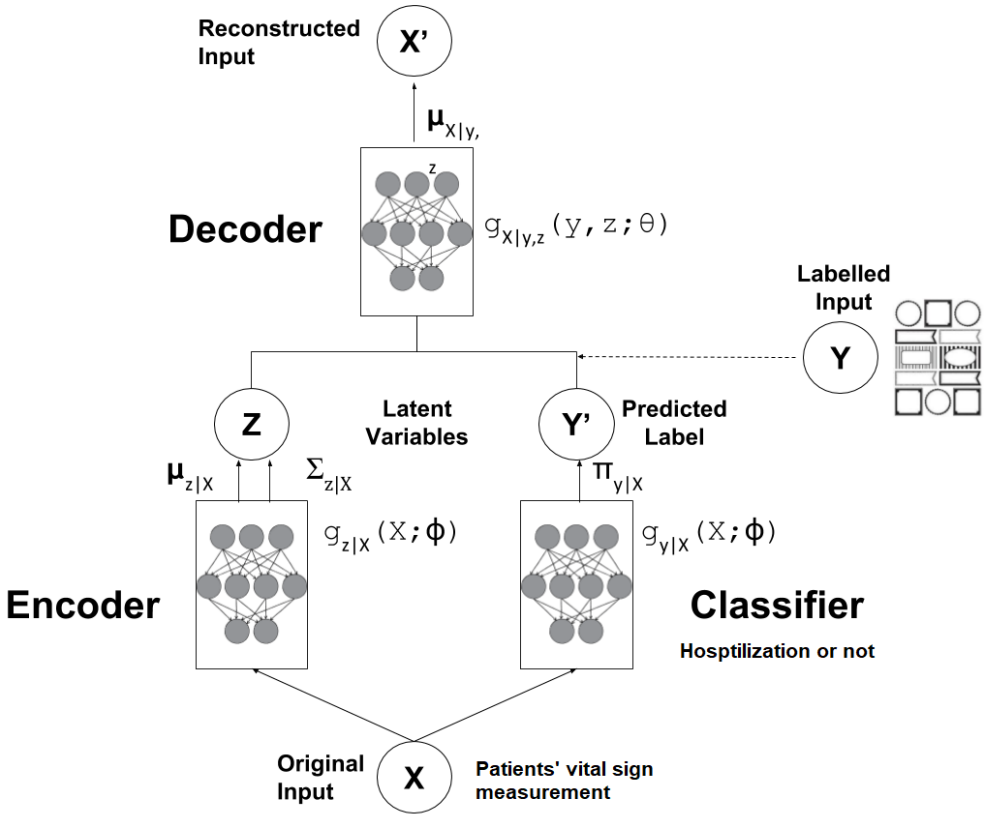
PCA

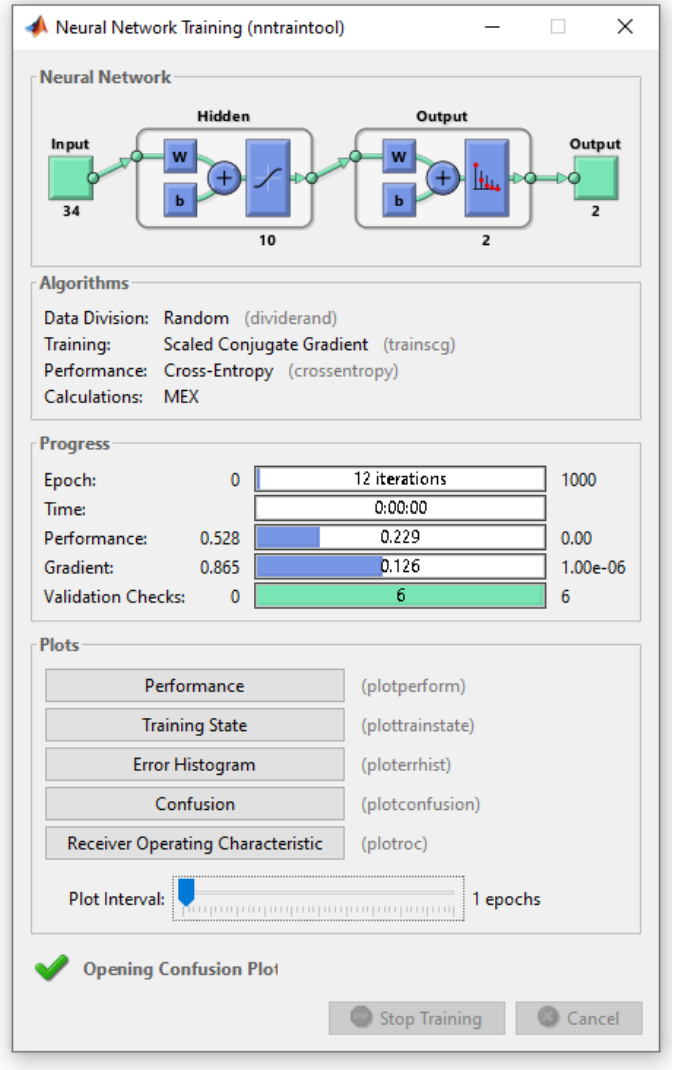




No accurate classification of hospitalization were found

Hospitalization classification based on Autoencoder



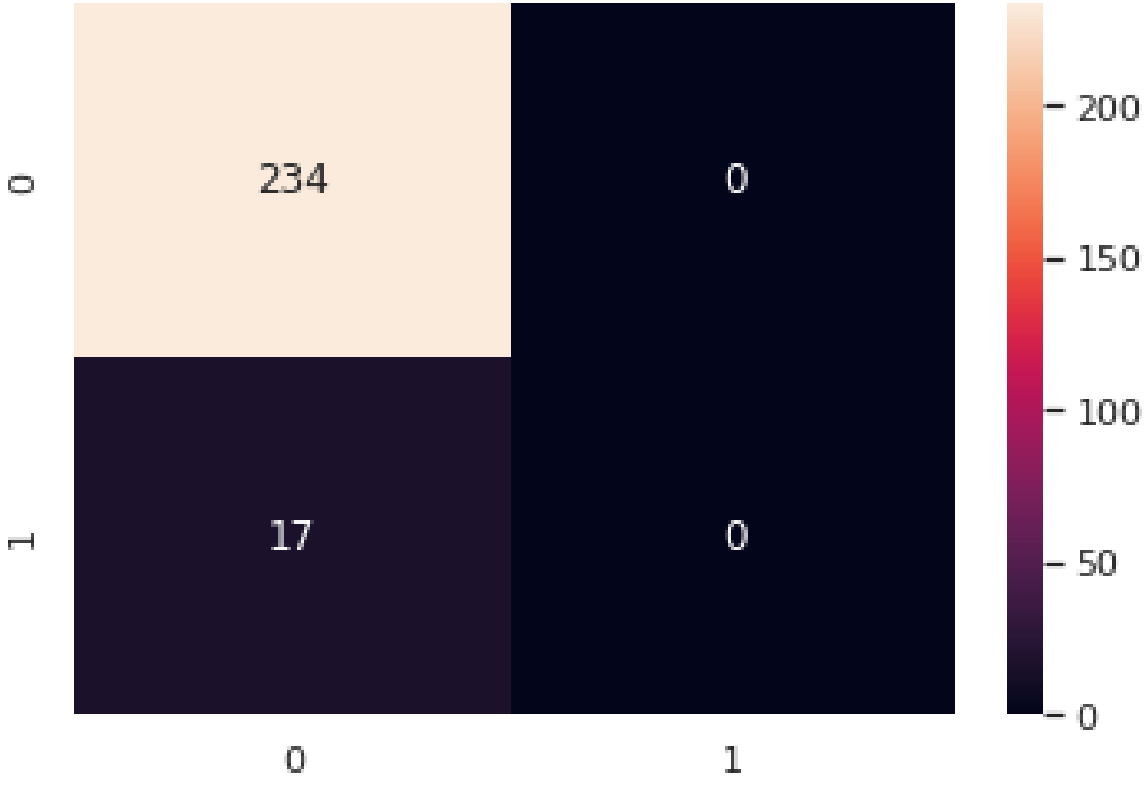


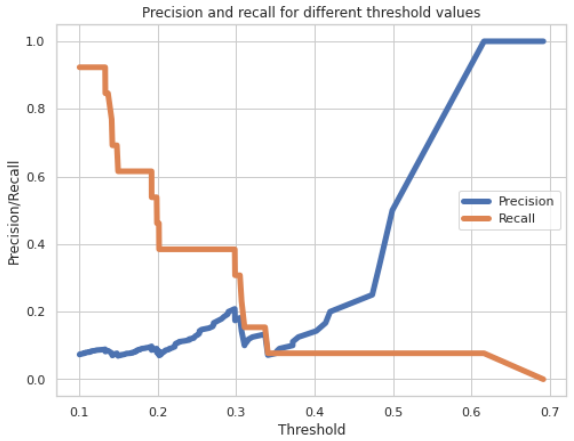
Autoencoder architecture

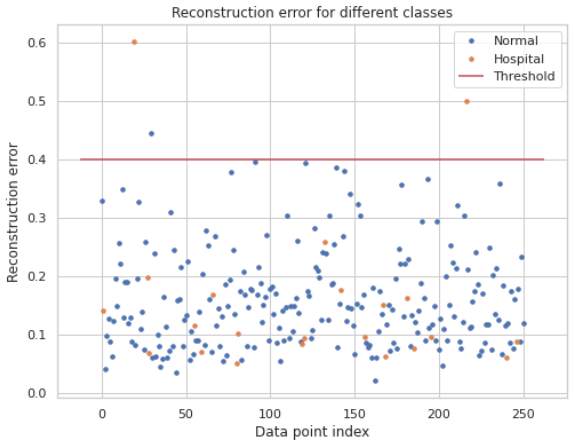
```
[69] autoencoder.summary()
```

```
Model: "model"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 35)]              0
_________________________________________________________________
dense (Dense)                (None, 100)               3600
_________________________________________________________________
dense_1 (Dense)              (None, 50)                5050
_________________________________________________________________
dense_2 (Dense)              (None, 25)                1275
_________________________________________________________________
dense_3 (Dense)              (None, 12)                312
_________________________________________________________________
dense_4 (Dense)              (None, 6)                 78
_________________________________________________________________
dense_5 (Dense)              (None, 12)                84
_________________________________________________________________
dense_6 (Dense)              (None, 25)                325
_________________________________________________________________
dense_7 (Dense)              (None, 50)                1300
_________________________________________________________________
dense_8 (Dense)              (None, 100)               5100
_________________________________________________________________
dense_9 (Dense)              (None, 35)                3535
=================================================================
Total params: 20,659
Trainable params: 20,659
Non-trainable params: 0
_________________________________________________________________
```

Autoencoder results

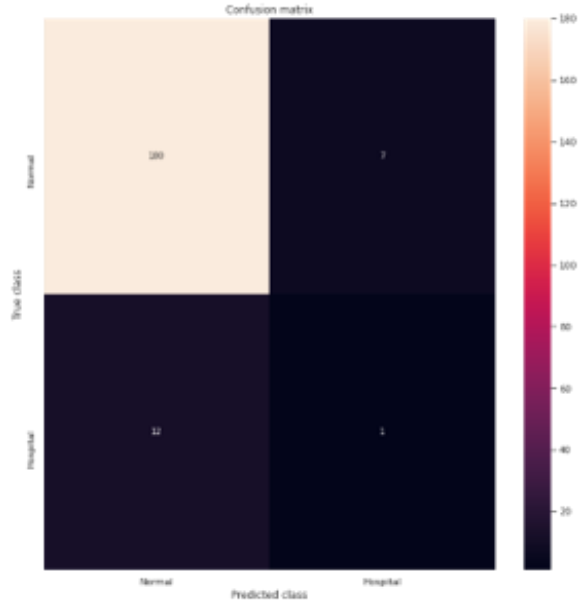






92.4 % accuracy but more negatives wrong

Test set confusion matrix



Patient subpopulation clustering based on K-means clustering

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \ \forall j, 1 \leq j \leq k \right\},$$

Assignment step: This step is step in which the points are assigned to the closest cluster. Distances between every data point and the k centroids are calculated. Based on this calculation the points are assigned

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j$$

Update step: This is the equation used to recalculate the new cluster center, means centroid

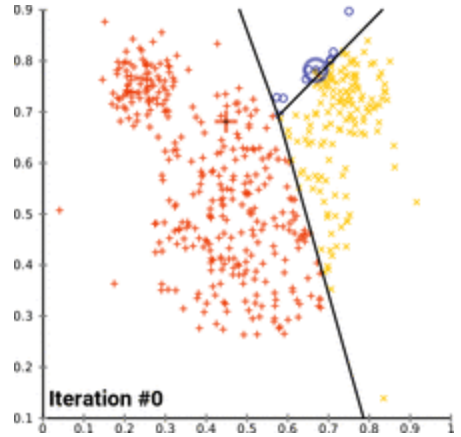


Fig 1. The algorithm iterates between steps the assignment and update steps until the stopping criterion are reached (the maximum number of iterations set by the user, no data points change clusters, or the sum of distances is minimized)

Input: k (the number of clusters),
D (a set of lift ratios)
Output: a set of k clusters
Method:
Arbitrarily choose k objects from D as the initial cluster centers;
Repeat:
1. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
2. Update the cluster means, i.e., calculate the mean value of the objects for each cluster

Until no change;

Finally, evaluate the 2 models' performance with nRMSE, **Silhouette Coefficient, Dunn's Index**

Why K-mean cluster?

In this project we hope to use two methods; learning and adjusting the methods as we get an understanding of each. Then comparing the two at the end; evaluating its performance to the classical methods. During our discussions we initially proposed K-means because of the following.

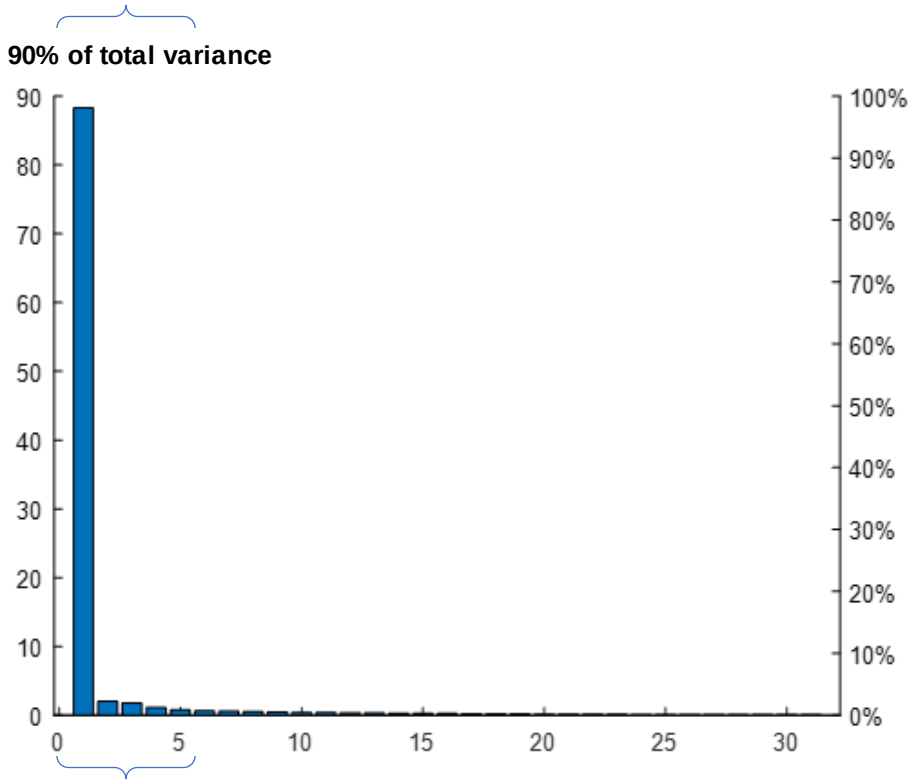
- K means is an unsupervised learning that would help us discover categories that we might not have seen on our own. As we do not have prior information about the grouping found among covid patients.

- Since we plan to transform our time series data to trends etc. It allows us to use K means. K-means is relatively simply to implement and has less computational aspects to it.

- From our wide research, literary critique we have found multiple papers that used K means as a method to find sub phenotypes in diseased patients form HER (Comparable to the data we will be using). Drawing reference to these will aid us in developing the best possible algorithm for our purposes.

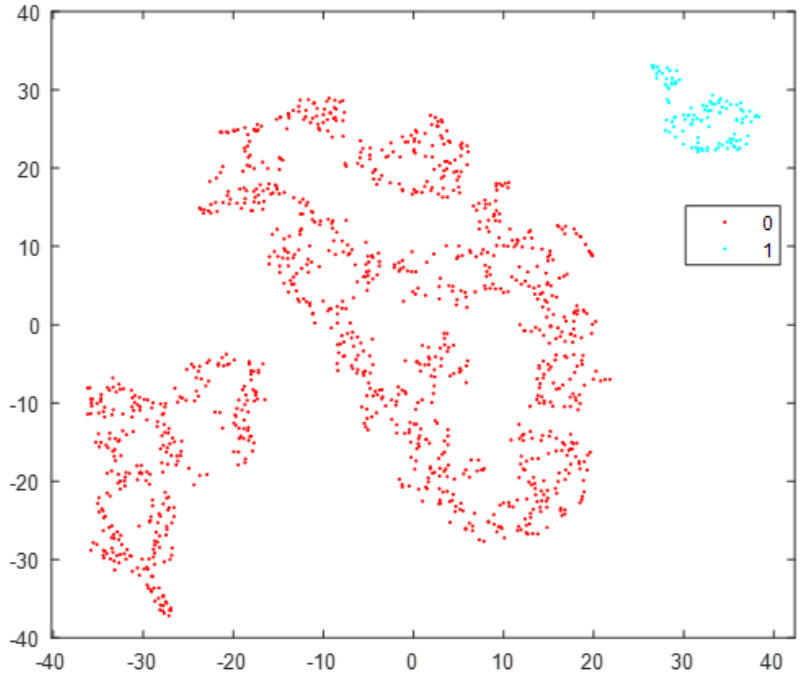
Dimension reduction with PCA

```
13
14      % PCA
15      %X = table2array(sTable);
16 -    [coeff, score, ~, ~, explained] = pca(DFN,'Centered',false);
17 -    hold on
18 -    bar(explained)
19 -    yyaxis right
20 -    h = gca;
21 -    h.YAxis(2).Limits = [0 100];
22 -    h.YAxis(2).Color = h.YAxis(1).Color;
23 -    h.YAxis(2).TickLabel = strcat(h.YAxis(2).TickLabel, '%');
24 -    id = find(cumsum(explained)>90,1);
25 -    scoreTrain90 = score(:,1:id);
```

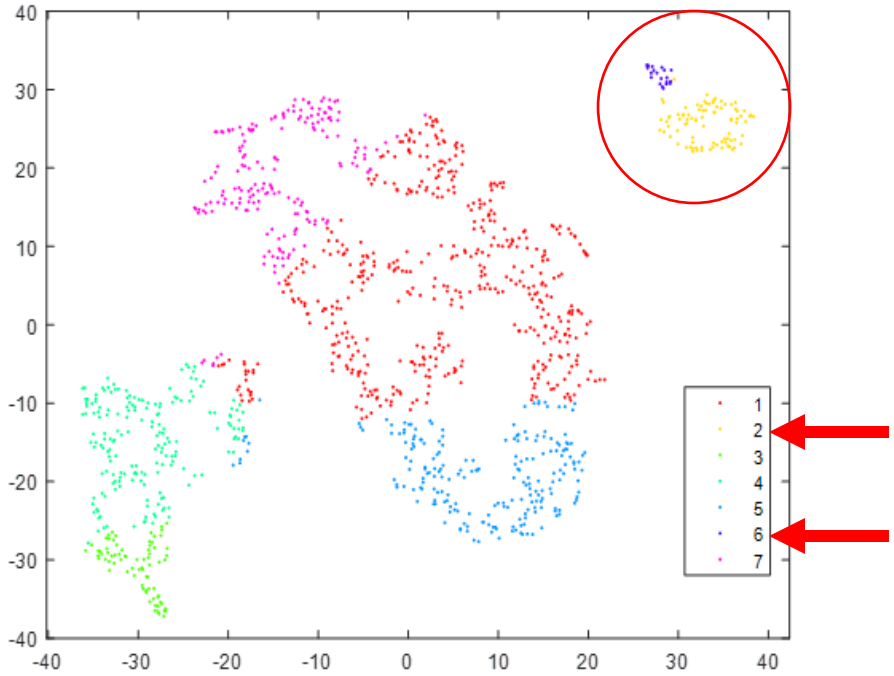


Clustering with K-mean

```
29      %% K-mean clustering and t-SNE plot
30 -     Y = tsne(scoreTrain90);
31 -    [IDX,C,SUMD,K]=kmeans_opt(scoreTrain90);
32 -    figure
33 -    [clusters, centroid] = kmeans(scoreTrain90, K)
34 -     gscatter(Y(:,1),Y(:,2),clusters)
35 -     figure
36 -     gscatter(Y(:,1),Y(:,2),CLH)
37
```



hospotalized vs nonhospitalized



Clustering with measurement data

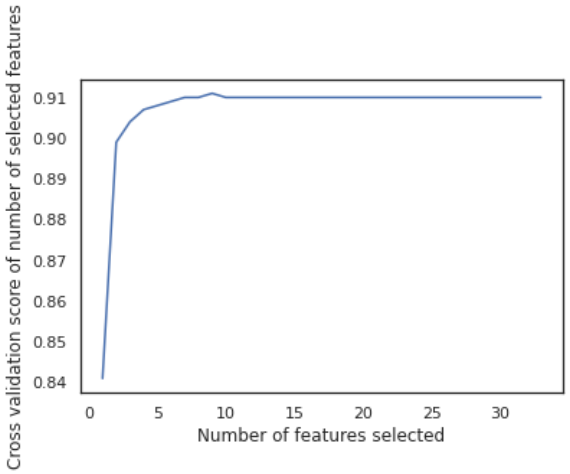
K-means interpretation and improvement

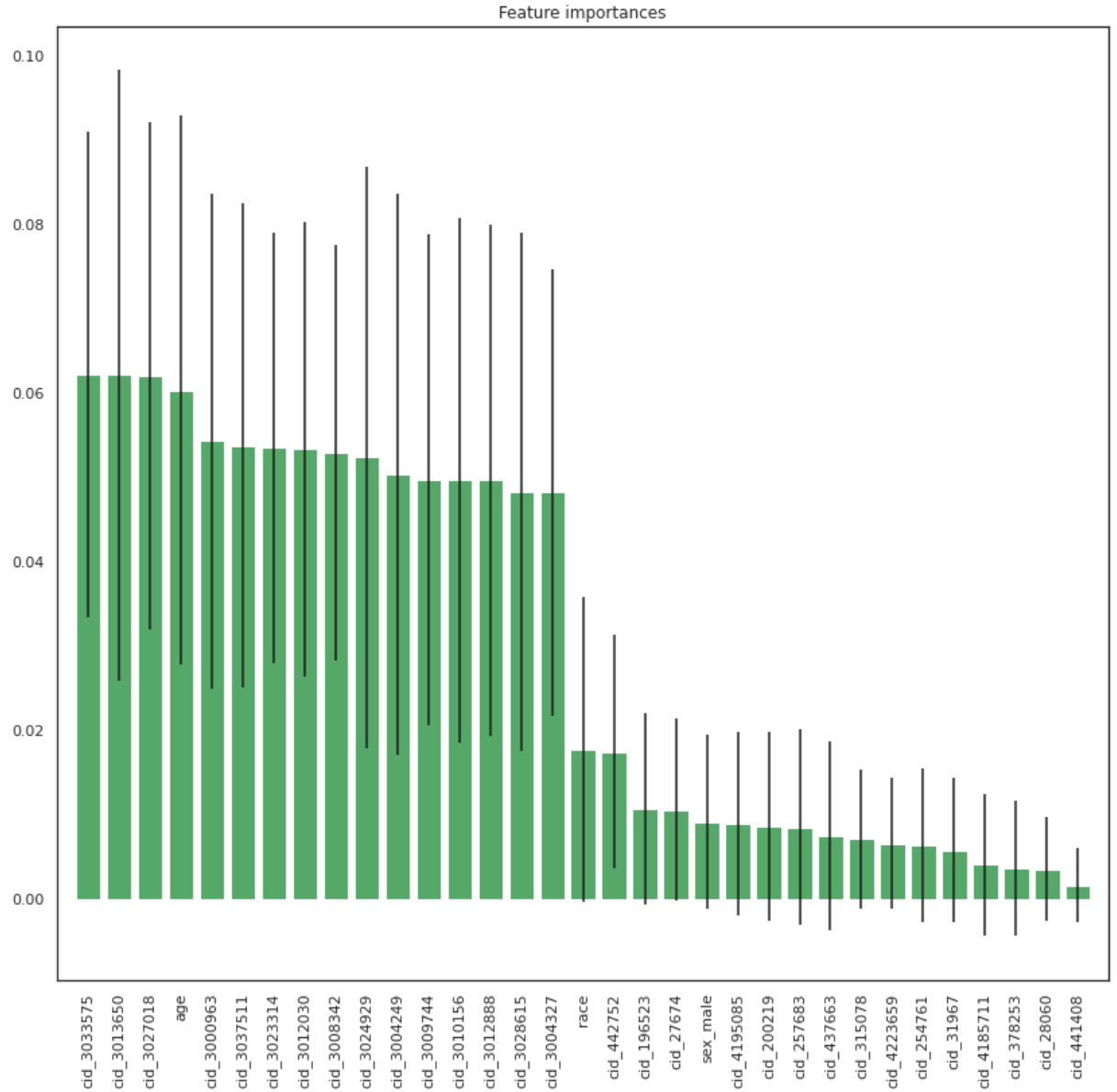
```
%% Data Tuning (normalization)
DF = readmatrix("mydata.csv");
DFR = DF(2:end,:);
[B,TF] = rmoutliers(DFR, 'movmean', 10);
DFN = normc(B); %data that is normalized and prepared for clustering this value will be inputed into the
%SilhouetteEvaluation is an object consisting of sample data, clustering data,
%and silhouette criterion values used to evaluate the optimal number of data clusters.
%Create a silhouette criterion clustering evaluation object using evalclusters.
 E = evalclusters(DFN,'kmeans','silhouette','klist',[1:10]);
%because the previous clustering was done with K=7 we decided to make the
%values of k evaluated 1 through 10, to leave a margin.
```

```
>> E = evalclusters(DFN,'kmeans','silhouette','klist',[1:10])

E =

  SilhouetteEvaluation with properties:

    NumObservations: 1251
         InspectedK: [1 2 3 4 5 6 7 8 9 10]
    CriterionValues: [NaN 0.2007 0.2044 0.1354 0.1033 0.0955 0.0926 0.0910 0.0895 0.0832]
           OptimalK: 3
```

```
%% K-mean clustering and t-SNE plot
Y = tsne(scoreTrain95 );
[IDX,C,SUMD,K]=kmeans_opt(scoreTrain95);
figure
idx = kmeans(scoreTrain95 ,E.OptimalK); %rendition of previous clustering method including the optimal value of K
gscatter(Y(:,1),Y(:,2),idx)
```

Feature importance plot: using random forest classifier

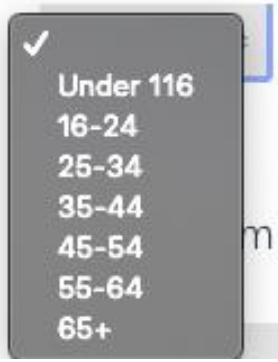




Graphical user interface(GUI)

1. Patient demographic

2. Patient Age



4. Heart rate

5. If you have other data you can place them in a matrix form as shown in preview

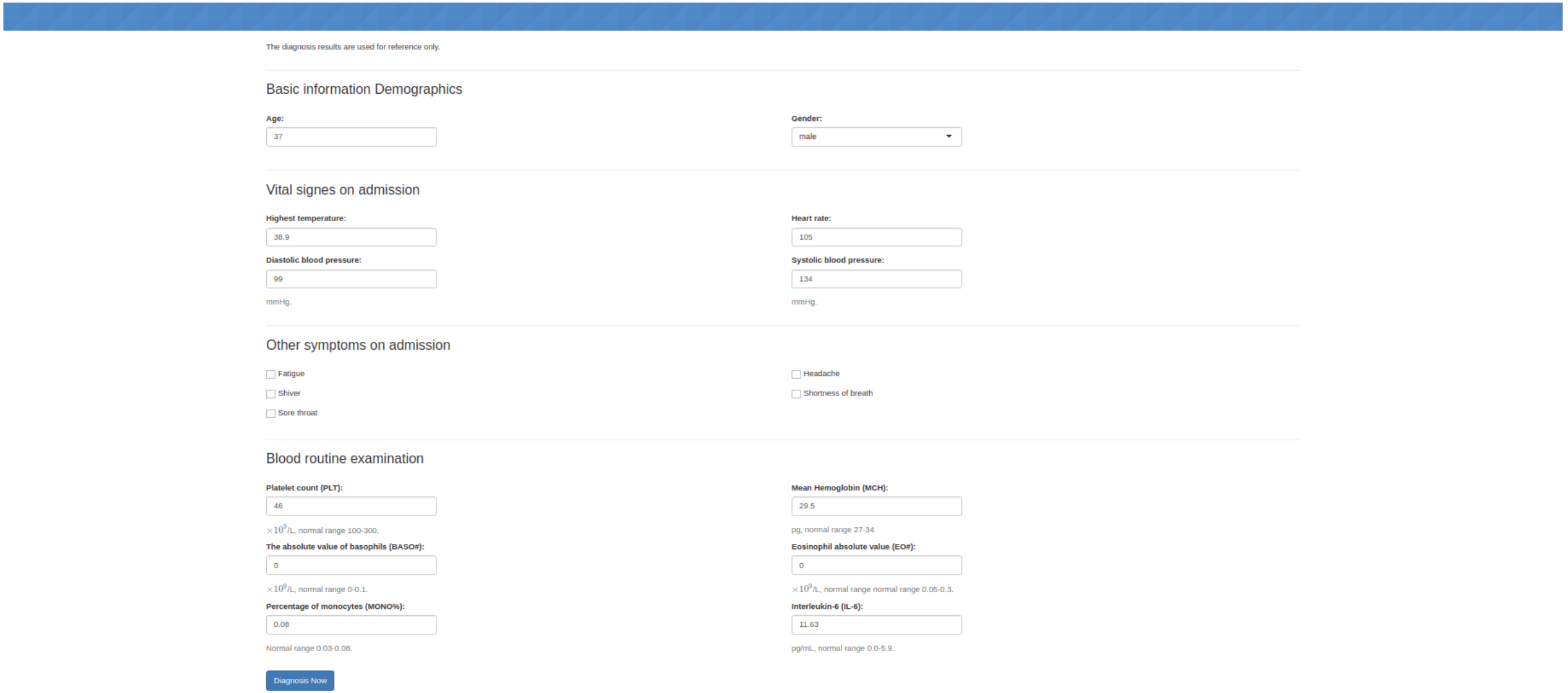
3. Maximum body temperature

6. If you familiar with the process and already have a matrix of the patient's EHR upload here

CSV file upload

Choose File No file chosen

Sample GUI


The diagnosis results are used for reference only.
Basic information Demographics
Age:
37
Gender:
male
Vital signes on admission
Highest temperature:
38.9
Heart rate:
105
Diastolic blood pressure:
99
Systolic blood pressure:
134
mmHg.
mmHg.
Other symptoms on admission
Fatigue
Headache
Shiver
Shortness of breath
Sore throat
Blood routine examination
Platelet count (PLT):
46
Mean Hemoglobin (MCH):
29.5
×10⁹/L, normal range 100-300.
pg, normal range 27-34
The absolute value of basophils (BASO#):
0
Eosinophil absolute value (EO#):
0
×10⁹/L, normal range 0-0.1.
×10⁹/L, normal range normal range 0.05-0.3.
Percentage of monocytes (MONO%):
0.08
Interleukin-6 (IL-6):
11.63
Normal range 0.03-0.08.
pg/mL, normal range 0.0-5.9.
Diagnosis Now

Plan of Action

- Explore more on psyciological representations of selected measurement features for additional feature engineering.
- Perform further tuning on hyperparameters associated with K-means clustering and Autoencoder for better classification accurary or more distinct cluster formation.
- Extract feature importance information from clustering and classification results upon reaching desired predictability and model the top-scoring parameters' sensitivity.
- Non-binarize the classfication of patient hospitaliztion into a 0-1 spectrum representing the probability of hospitlization.