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

COVID Sub phenotyping Project

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

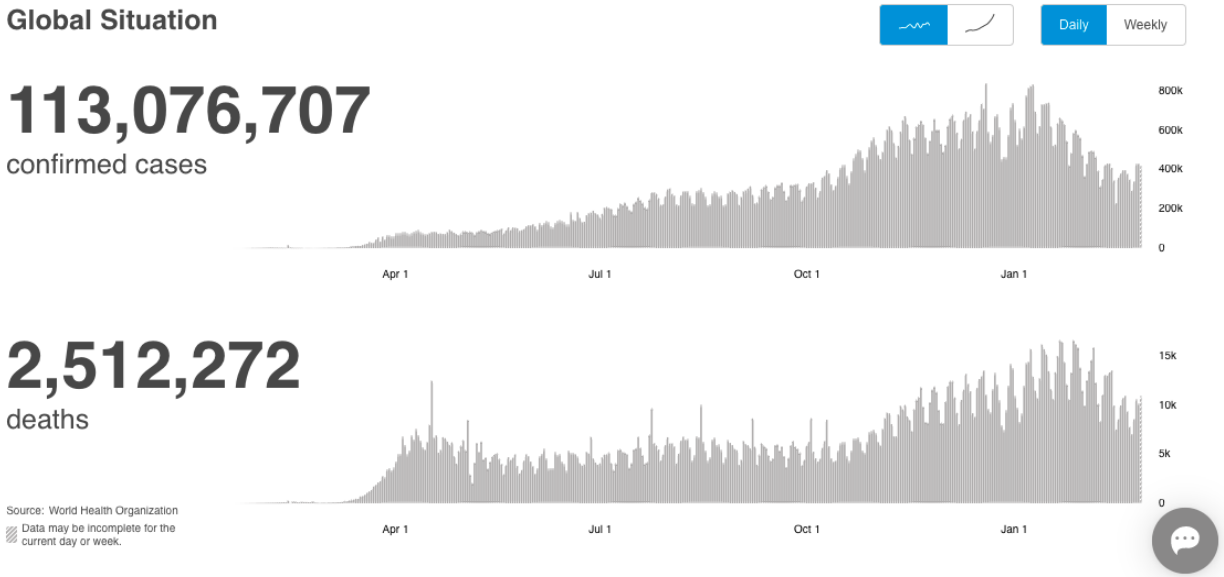
Seonggeon Cho, Rohan Bhukar, Bryce Butler, Zhonghao Dai




May 03, 2021

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



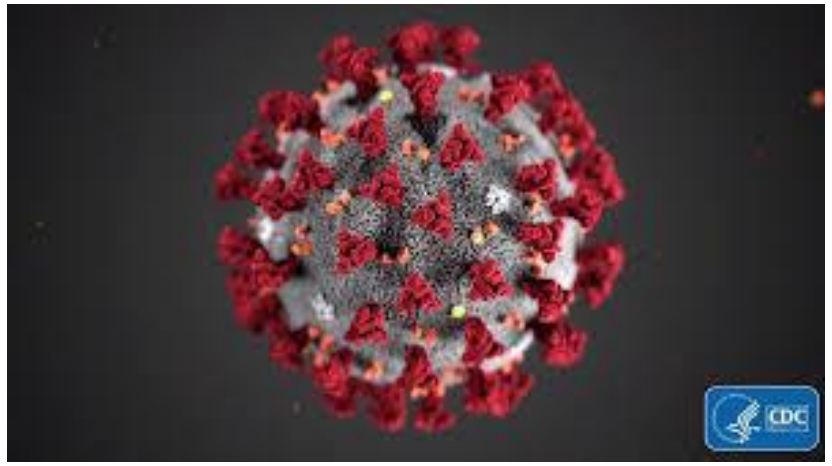
WHO, 2021

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

Current challenges of COVID sub-phenotyping

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



Literature critiques

Title of Paper	Methods / Solutions	Strengths	Weakness
1. Deep representation learning of electronic health records to unlock patient stratification at scale	Convolutional Neural Network, **Autoencoder**	• It showed robust result with sparsely available EHR record	• It used 12 years of EHR record to make meaning subcluster of the disease type.
2. Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data	**K-means clustering**	• Unsupervised methodology of high dimensional subclustering within a single disease type.	• Large sample size is required for training. It used 10 years of EHR record to make meaningful subcluster of the disease type.
3. Phenotyping Clusters of Patient Trajectories suffering from Chronic Complex Disease.	Time Series K means clustering, **Variational Autoencoder**	• Both methods shows promising phenotyping of time-series vital signs data with distinct phenotypic characteristics on	• Phenotype separation are shown to be susceptible to unevenly sampled time-series data and unbalanced class distribution
4. Vital signs assessed in initial clinical encounters predict COVID-19 mortality in an NYC hospital system.	**Multivariate Logistic regression,** **Hyperparameter Tuning**, Extreme Gradient Boosting **Xgboost**	• Immediate, objective measures(age, BMI, heart rate, respiratory rate, O2 saturation rate) collected at time of admit, can be effective predictors of mortality rather than lab-tests with critical lag in response time; • 2-tier analysis. A) identifies critical factors using logistic regression; B) gradient boosting ML uses factors to predict COVID-19 related mortality	• Critical factors, Odds ratio values are derived from demographic data (Race, ethnicity, etc.) comprising of patients from New York area only, cannot be generalized to major ethnic world populations. • Only severe cases of COVID-19 considered, they disproportionately included patients with poor outcomes, limiting the generalizability of study
5. COVID paper	Logistic regression (LR) with lasso, Decision tree, Adaboost algorithm	• Feature importance of clinically relevant measurements and health data • Characterizing suspected & non-suspected pneumonia cases based on top features	• Only a small sample size of 132 patients used for modeling. • Model was developed and validated in a single-center fever clinic

Tasks remain to be explored / What new work can be added to the field

In the *problem statement* of COVID-19 sub phenotyping,

- **Solutions arising from classical machine learning models** (gradient boosting, logistic regression, etc.) exist in the case of COVID-19 sub phenotyping **utilizing only single cohort data**, and **basic clinical presentation factors**.
- Either the **existing methods** for COVID-19 phenotyping **predict only mortality**, or they present more novel phenotypes but **derived from only smaller cohort size**, which is the current bottleneck in multi-modal comprehensive research.

Solutions lacking from the current methods and literature available,

- Utilizing data from larger cohort sizes via multiple systems and developing the models with **data from diverse ethnic populations** as part of demographic factors is lacking from current studies.
- **Not only larger cohort sizes**, but **multi-modal data** which ***measures vital*** (temperature, heart rate, respiratory rate, O2 saturation, etc.) patient clinical presentations, also **needs to be factored in** to develop better and novel patient sub phenotypes, along with developing more accurate predictive models for clinical use-cases which can handle time irregularities in the datasets.

Data Modalities: COVID-19 DREAM CHALLANGE

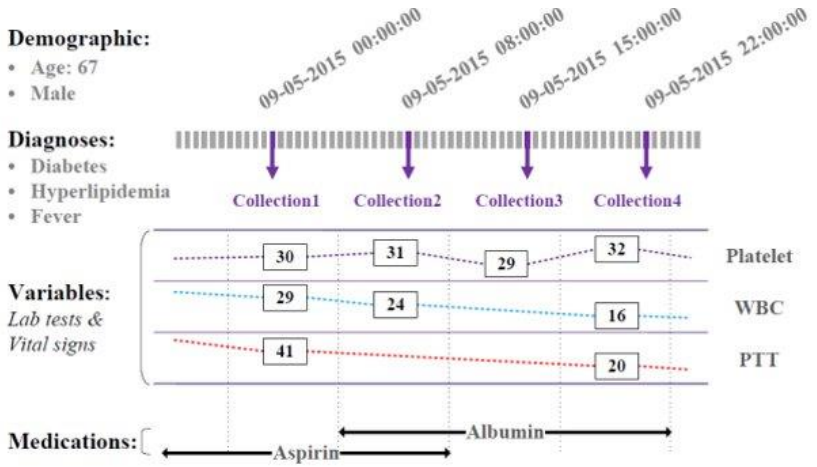
- **Time independent data**
  - Patient's demographic profile
    - Age
    - Gender
    - Race

- **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 Source: 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	

Data preprocessing

- **In total 38 features selected from literature survey (published Covid-19 papers)**
- **Features extracted from the multiple tables and data values selected based on temporal relevance of features**
- **One-hot encoding the categorical data (gender, race, etc.)**
- **Split the training data with 80:20 split ratio, using 5 fold CV**

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)				

Table 1: Candidate features from the model from Covid paper

- **Imputing train, validation and test data using Mean, Median, KNN impute and MICE.**
- **Finally, selected KNN impute for data imputation due to the method better suitable to this dataset**
  - Distribution of original data changed less with KNN impute compared to MICE
- **Z-score normalization of data**
  - $Z = (x - \mu) / \sigma$
  - Dataset further used as input for clustering

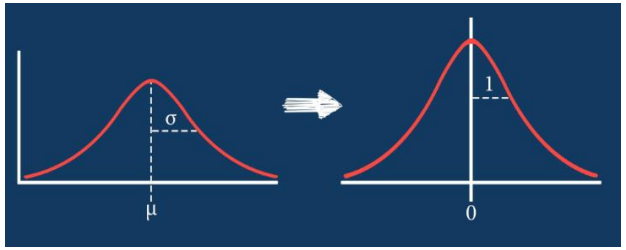


Fig 1: Z-score normalization

- **For Classification,**
  - Selecting the train and validation sets in each CV, imputing and feature scaling the datasets separately

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

Fig 2: Temporally relevant data

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

Data preprocessing

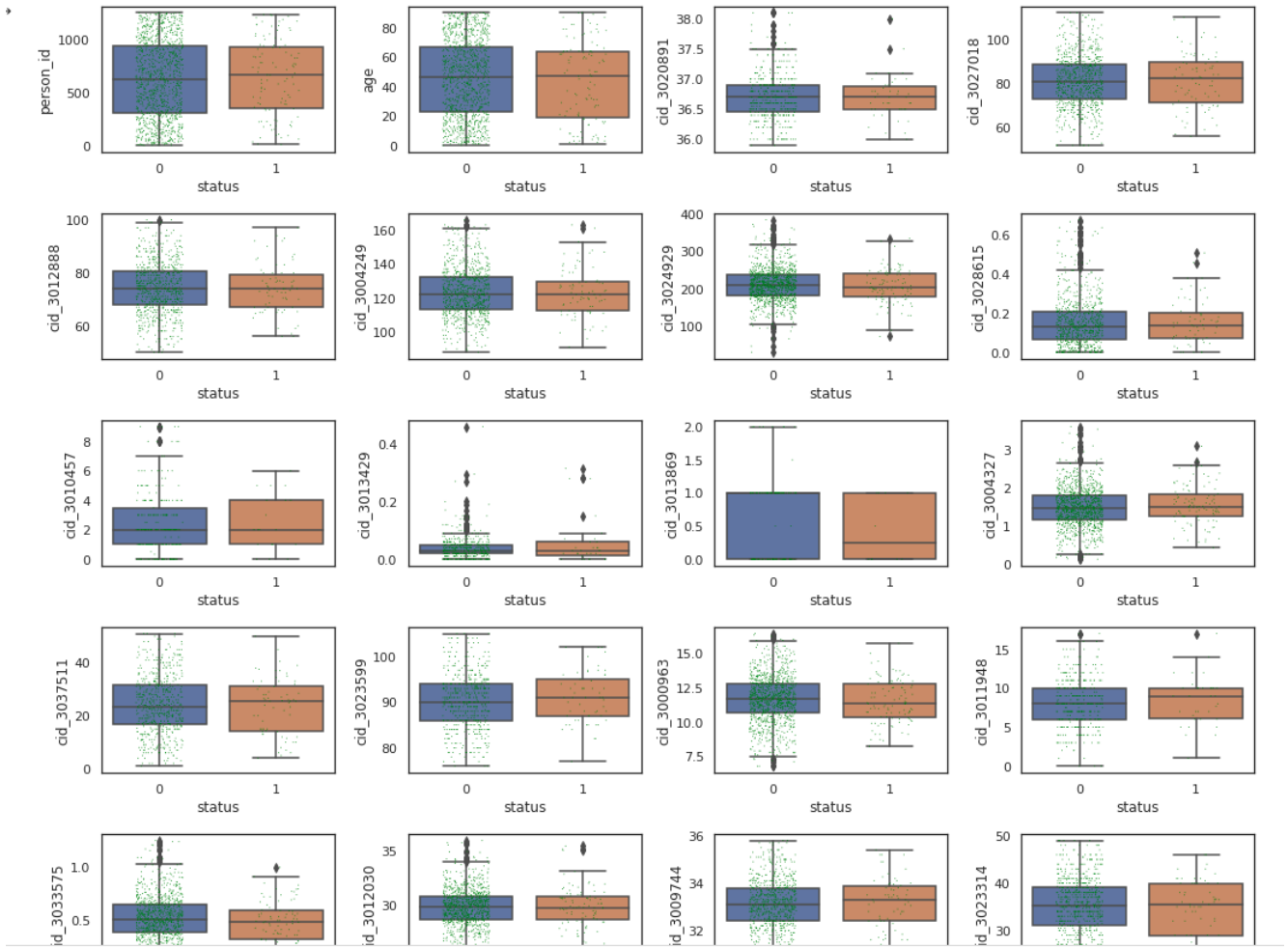


Fig 3: Boxplot for the selected features we try to draw distributions and found no extreme value cases

In this diagram:

- 0 : Non hospitalized patients
- 1 : Hospitalized patients

Data preprocessing

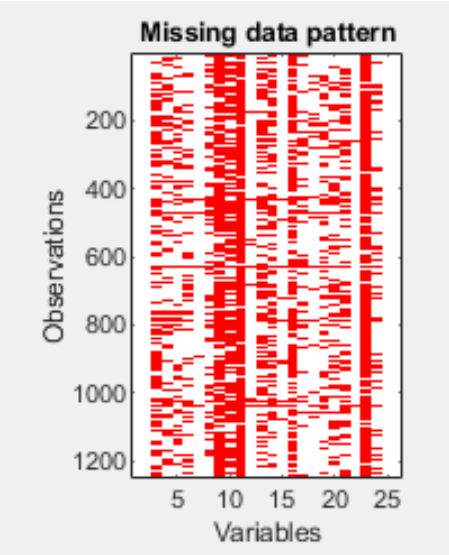


Fig 4: Data missingness

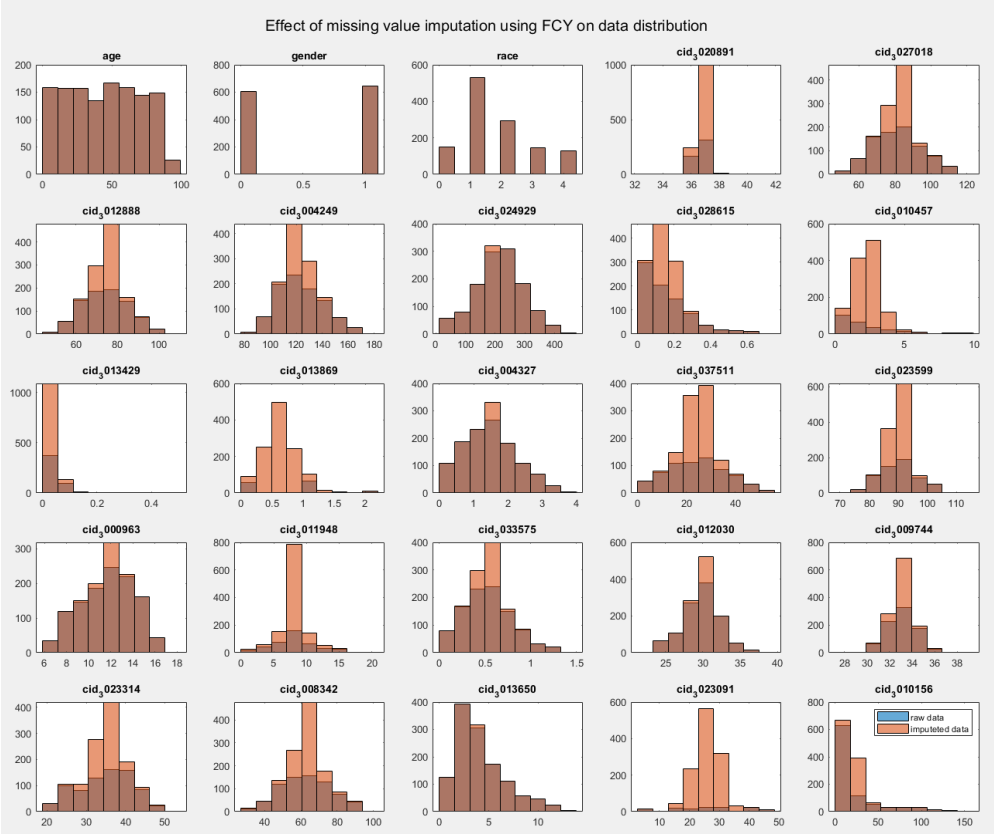


Fig 5a : Data distribution after imputation with MICE method

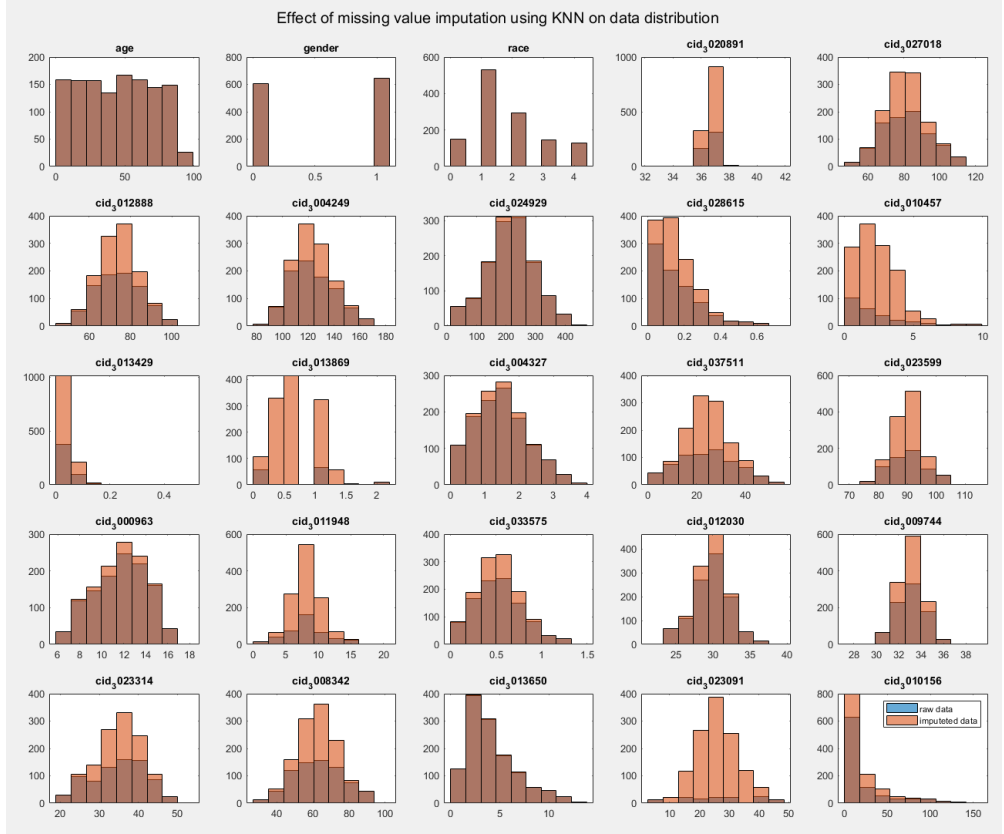


Fig 5b : Data distribution after imputation with KNN method

Feature importance plots

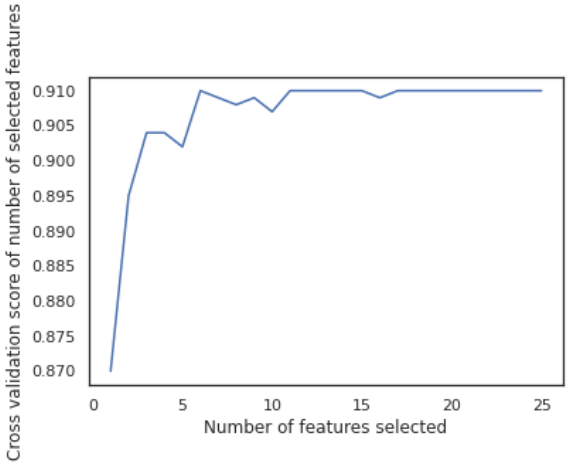


Fig 6 : Number of features selected vs cross validation score

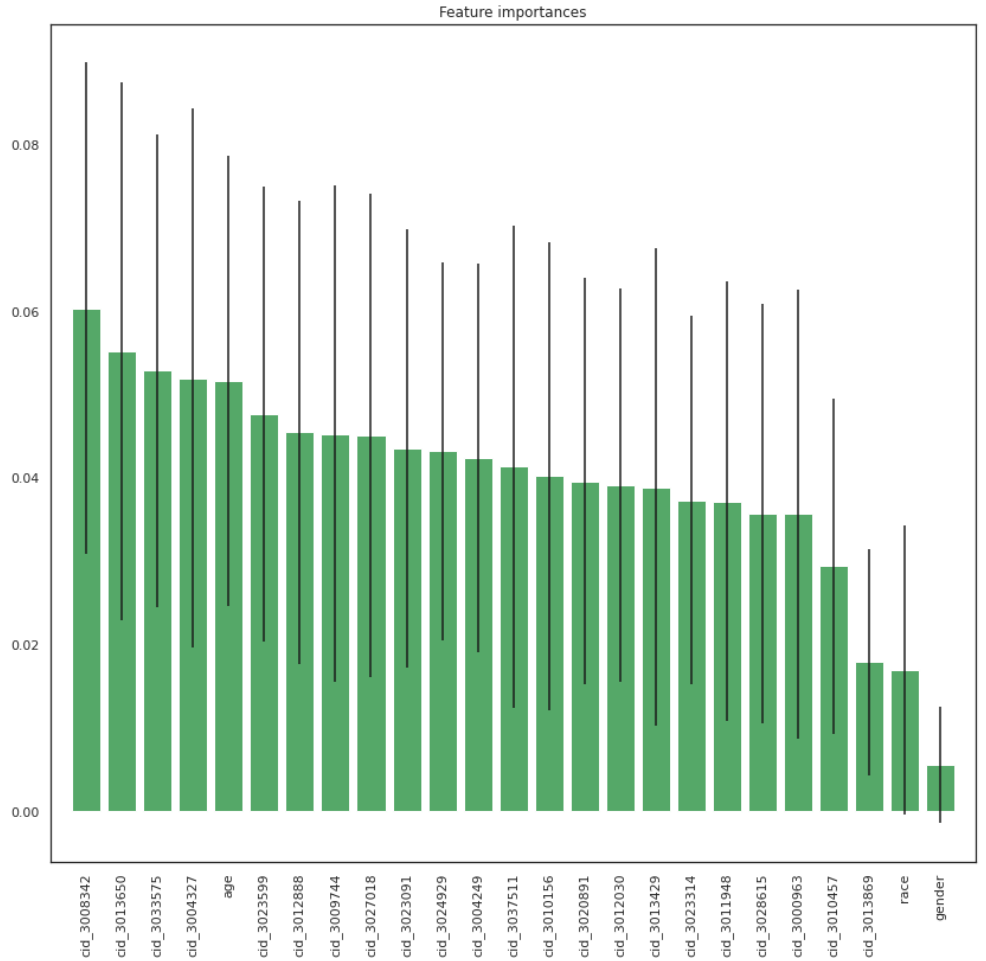


Fig 7 : Feature importance and generated weights based on tree method

Table 1 : Feature importance ranks of concept ids using MRMR method

Concept id	Feature
cid_3033575	Monocytes [#/volume] in Blood
cid_3013650	Neutrophils [#/volume] in Blood
cid_3008342	Neutrophils/100 leukocytes in Blood
cid_3004327	Lymphocytes [#/volume] in Blood
cid_3027018	Heart rate
cid_3009744	MCHC [Mass/volume]
cid_3023599	MCV [Entitic volume]
cid_3011948	Monocytes/100 leukocytes in Blood
cid_3023091	Interleukin 6 [Mass/volume] in Serum
cid_3013869	Basophils/100 leukocytes in Blood
age	Age of patient
cid_3012888	Diastolic blood pressure
cid_3012030	MCH [Entitic mass]
cid_3004249	Systolic blood pressure
cid_3024929	Platelets [#/volume] in Blood
cid_3010156	C reactive protein [Mass/vol] in Serum
gender	Gender of patient
cid_3020891	Body temperature
cid_3000963	Hemoglobin [Mass/volume]
cid_3010457	Eosinophils/100 leukocytes in Blood
cid_3023314	Hematocrit [Volume Fraction] of Blood
cid_3028615	Eosinophils [#/volume] in Blood
race	Race to which person identifies
cid_3037511, cid_3013429	Lymphocytes/100 leukocyte, Basophils

ML methods for feature selection

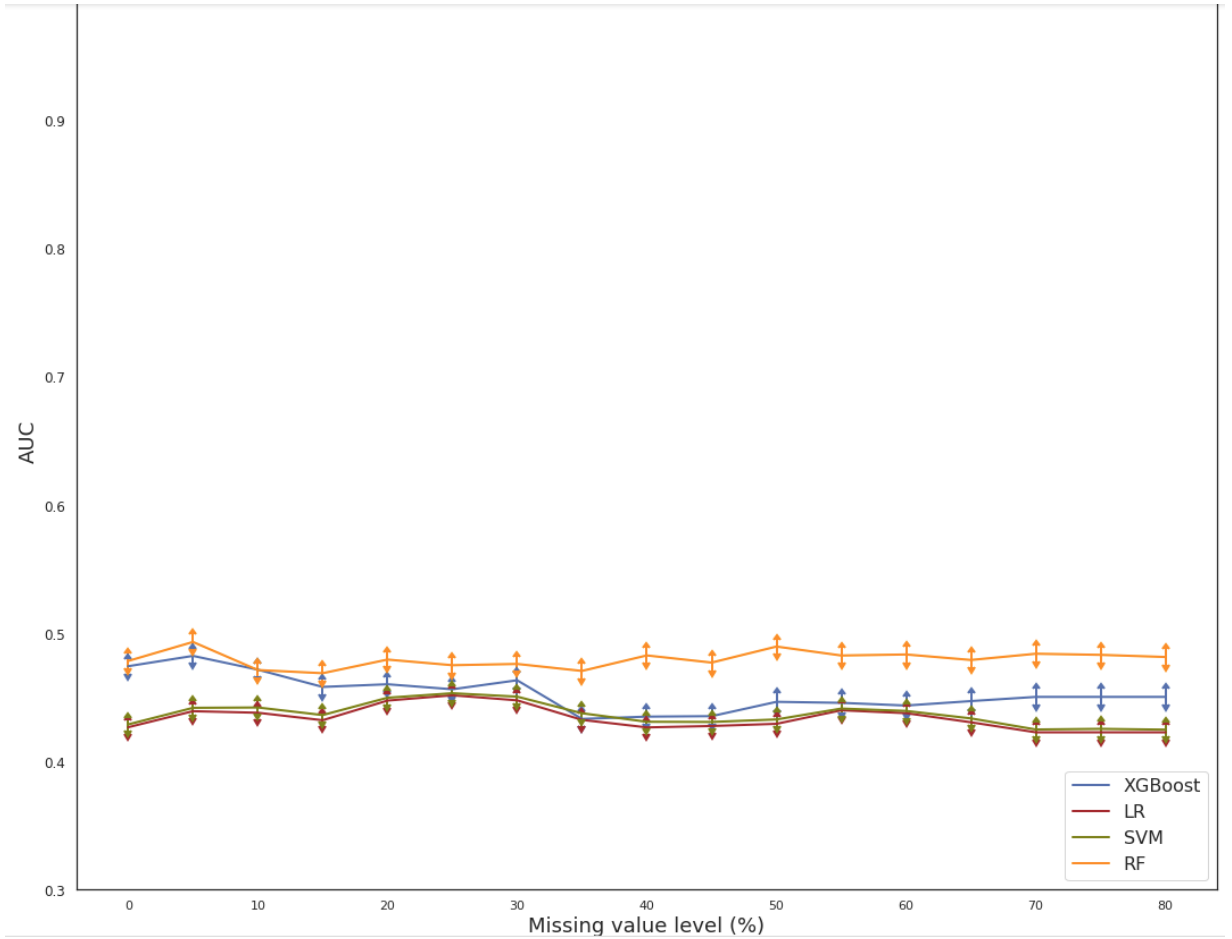


Fig 8: Upon **imputation at different thresholds of missing value**, found that highest classification AUCs were achieved at 50% missing value imputation. The data with <= 50% missing feature values was created and further imputed using KNN. Additional, column of Body_temp was added manually due small margin of missingness.

Feature analysis

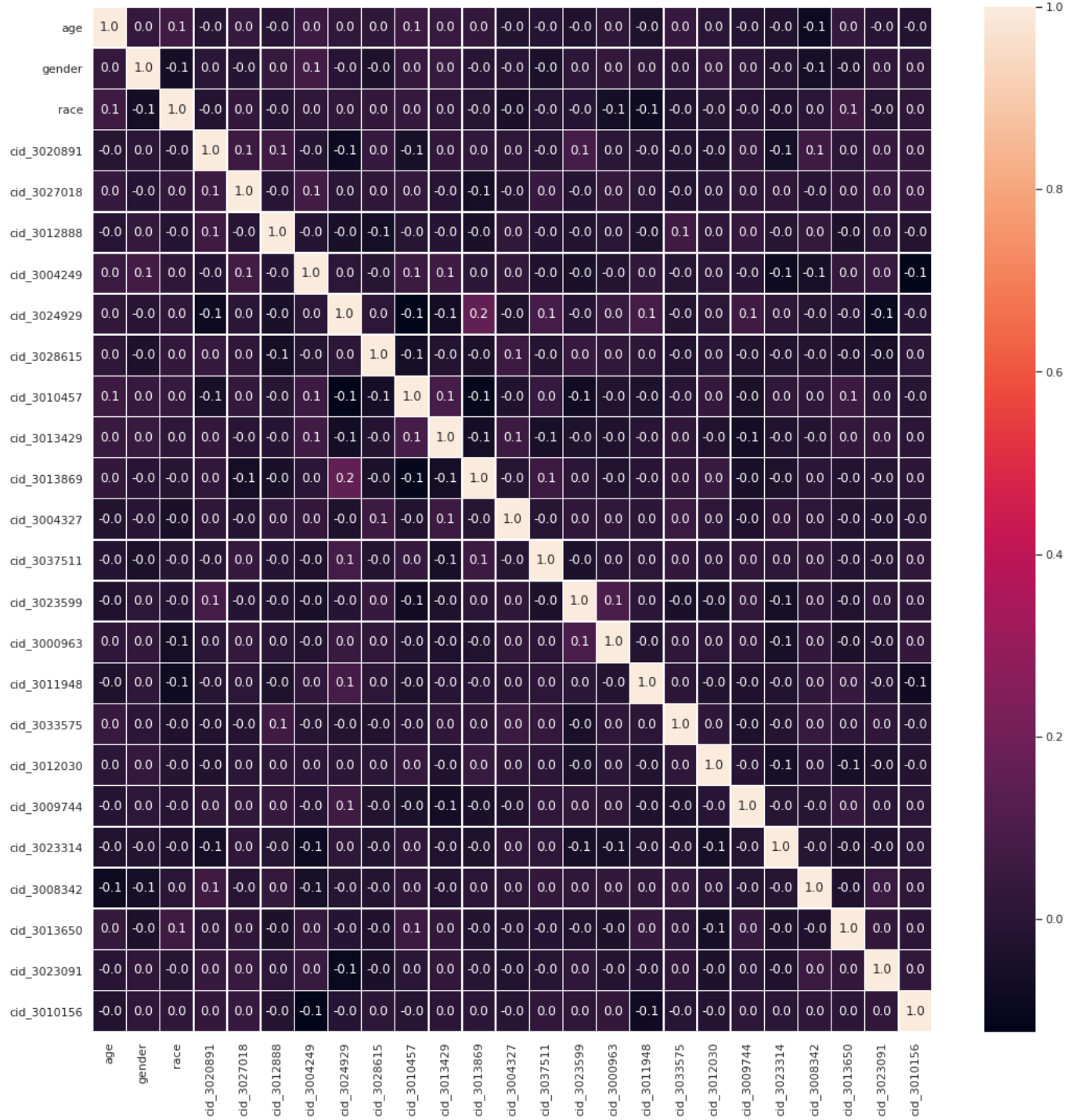


Fig 9: Correlation matrix plot

None of the selected and processed features were found to be correlated.

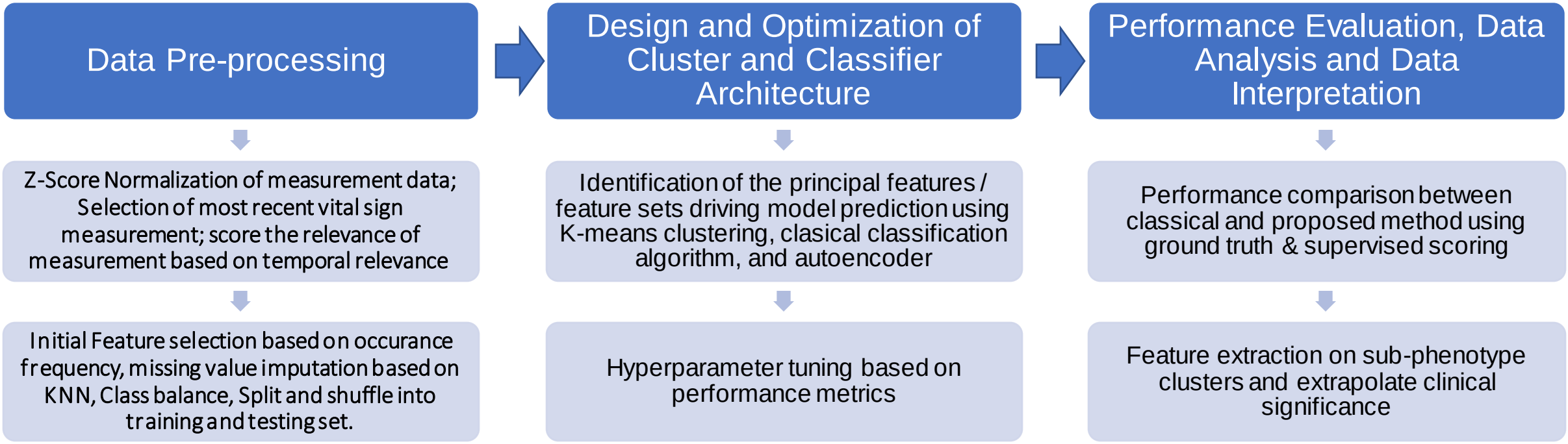
Data format post data pre-processing

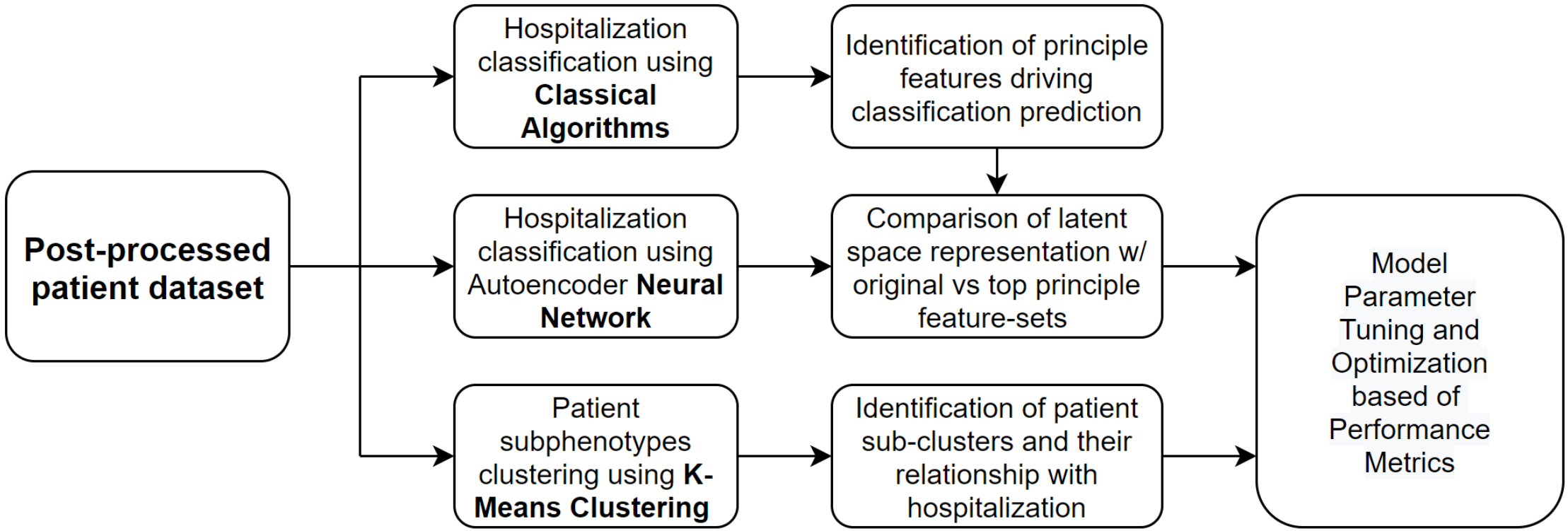
Tknn

214x23 table

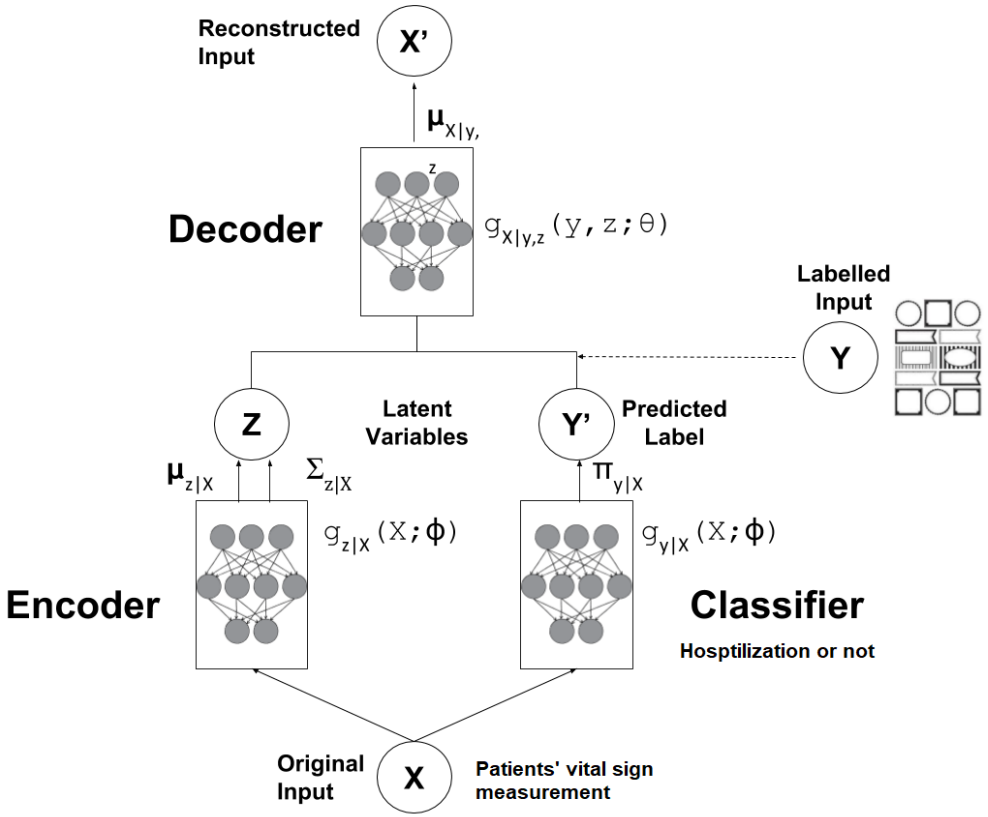
	1 person_id	2 age	3 gender	4 race	5 cid_3020891	6 cid_3027018	7 cid_3012888	8 cid_3004249	9 cid_3024929	10 cid_3028615	11 cid_3013429	12 cid_3004327	13 cid_3037511	14 cid_3023599	15 cid_3000963	16 cid_3033575
1	8	3	0	1	36.6333	92	81.3333	105	372	0.3433	0.0533	1.3300	13.6667	91	11.7000	0.3000
2	12	58	0	1	37.0667	100	67	120	17	0.0900	0.0467	1.6600	30	87.3333	7.6000	0.1000
3	33	63	0	1	36.5000	109	70	139.3333	197	0.1700	0	2.2600	21.3333	85.6667	12.2000	0.4100
4	58	77	1	0	36.6333	75	75	136	277	0.1500	0.0533	1.6900	24.6667	93	12.4000	0.6900
5	70	7	1	1	36.9667	74	85.3333	96	239	0.0733	0.0233	2.8900	8	85	12.5000	0.2267
6	73	58	0	4	37.5000	77	75	133	246	0.2400	0.0400	0.2900	24	85	10.3000	0.2500
7	91	63	0	2	36.2667	84	87	118	316	0.1100	0.0500	0.8900	35	88	8.5000	0.3500
8	96	68	1	4	36.5000	94	66	111	170	0.4300	0.0067	1.4500	38.3333	90.6667	13.4000	0.2600
9	103	8	0	4	36.4000	87.3333	73.3333	140	218	0.3300	0.0167	0.5700	44	92	9.4000	0.1000
10	144	62	0	2	36.5333	82	65	102	236	0.1500	0	0.0500	33	91.3333	11.1000	0.2500
11	152	61	1	2	37	77.6667	65	119	297	0	0.0367	2.0500	38	85	11.8000	0.4100
12	175	64	0	1	36.9000	65	56	135	313	0.1467	0.0267	1.7300	32	103	11.8000	0.8300
13	177	51	0	4	36.5000	60	66	92	254	0.1200	0.0267	2.2500	30	98	14	0.6000
14	184	76	1	0	36.4333	87	83	127	222	0.2600	0.0367	1.2200	19.6667	89	14.5000	0.6700
15	189	57	1	4	36.7000	93	68.3333	159	177	0.2000	0	1.9500	29	104	10.6000	0.5300
16	192	34	1	1	36.7000	70	78	127	391	0.5200	0.0300	1.3500	31	96.3333	14	0.4500
17	220	64	1	1	36.4000	77	75	127	106	0.0700	0.0200	1.6800	18	82	10.8000	0.7400
18	221	28	0	1	36.8000	87	79	127	273	0.0500	0.0400	2.5100	22	87	14.8000	0.9700
19	224	63	1	2	36.6000	70.3333	66.3333	139	350	0.1800	0.0167	2.0800	33	96.3333	12.1000	0.3700
20	226	4	1	1	36.6333	82	74.6667	107	251	0.0900	0.0300	2.5500	18.6667	85	15.5000	0.4500
21	230	32	1	1	36.7000	66	77.3333	100	411	0.1600	0.0200	1.8300	33	88	13.6000	0.1700
22	239	90	1	1	36.3333	92	68	117.6667	229	0.2000	0.0400	1.6800	17	95.3333	15.1000	0.5400
23	267	16	1	1	36.6000	83	75	134.3333	212	0.1300	0.0200	3.1700	20	96	14.3000	0.1900
24	330	74	0	4	37	79	68	121	257	0.0800	0.0500	2.1800	12	88.6667	16	0.7500
25	336	9	0	2	36.3333	70	76	135	269	0.1067	0.0333	1.8200	28	95	12.5000	0.3400
26	345	64	0	4	36.6000	59	84	95	325	0.2367	0.0267	1.7300	45	88	8	0.7100
27	346	82	0	1	36.4333	72.6667	81	128	146	0.0933	0.0733	1.7900	33	87.3333	8.5000	0.6633
28	359	70	1	2	37.3000	74	67	118.3333	209	0.3400	0.0967	1.5100	14.6667	88.6667	12.7000	0.8600
29	385	62	1	1	36.5000	90	60	126	204	0.2200	0.0467	1.5000	29	101	15.7000	0.4200
30	389	47	1	1	36.8000	90	92	122	289	0.0667	0.0233	1.5100	25	93	15.6000	0.2900
31	390	56	0	3	36	98	68.3333	128	179	0.1300	0.0233	0.8800	15	90.6667	11.6000	0.7900
32	400	79	1	1	37.1667	60	78.3333	119	291	0.2700	0.0467	1.9300	28	94	15.7000	0.4700
33	404	38	1	1	37.1000	58	69	135	192	0	0	1.9300	12	93	11.9000	0.4300
34	417	26	0	1	36.5000	103	73	125	104	0.2333	0.0200	1.5000	15.6667	88	11.5000	0.4000
35	440	83	1	0	36.5333	91	81	120	177	0.2733	0.0233	1.8000	16	93	11.9000	0.5267
36	448	14	1	1	36.3000	76.3333	71	122	206	0.3400	0.0167	1.2000	13	94	11.7000	0.2867

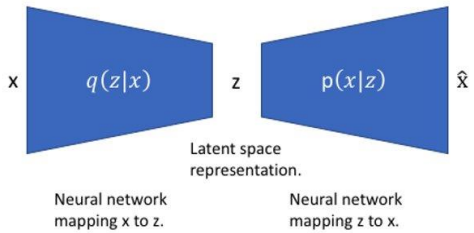
System Workflow




Post-processed patient dataset
Hospitalization classification using **Classical Algorithms**
Hospitalization classification using Autoencoder **Neural Network**
Patient subphenotypes clustering using **K-Means Clustering**
Identification of principle features driving classification prediction
Comparison of latent space representation w/ original vs top principle feature-sets
Identification of patient sub-clusters and their relationship with hospitalization
Model Parameter Tuning and Optimization based of Performance Metrics

Hospitalization classification using Autoencoder





- Compression of patient feature data into latent variables for hospitalization classification while retaining high input integrity.

- Encoder stage: $\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$
- Reconstruction optimization stage:

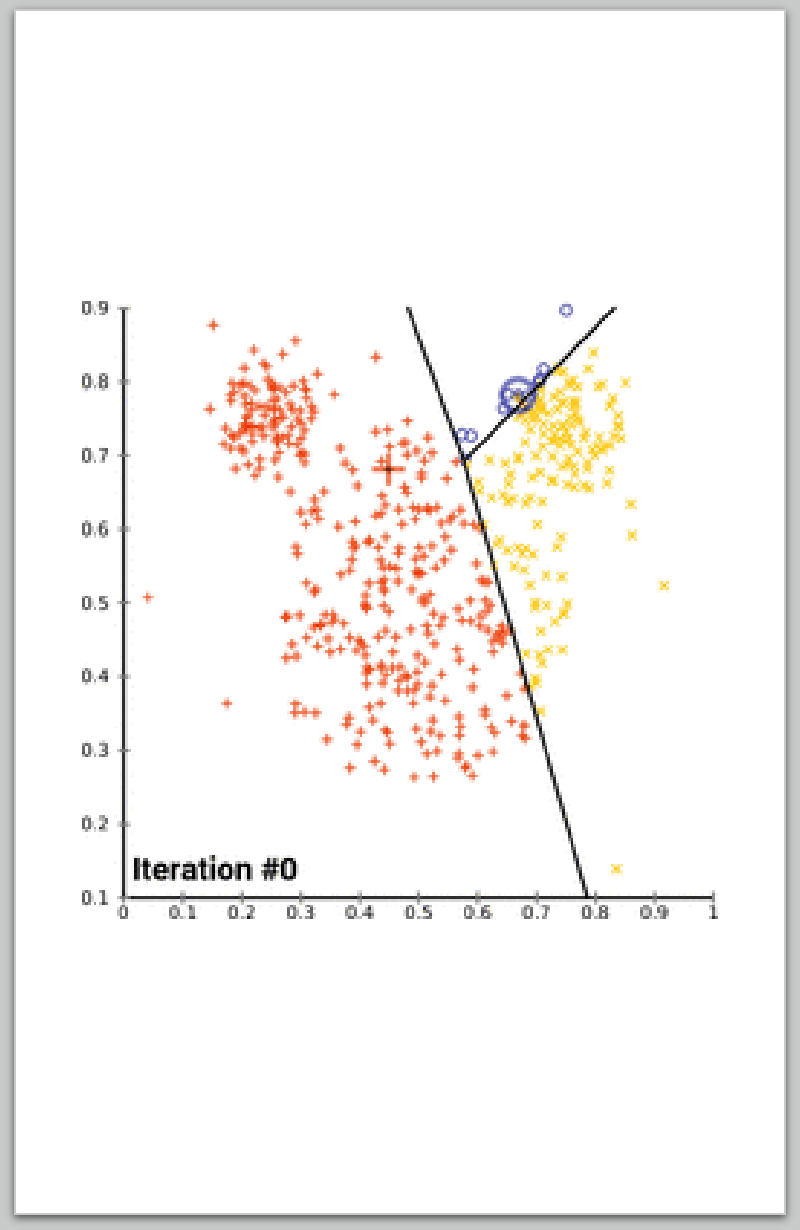
$$\mathcal{L}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2 = \|\mathbf{x} - \sigma'(\mathbf{W}'(\sigma(\mathbf{W}\mathbf{x} + \mathbf{b})) + \mathbf{b}')\|^2$$

- **Why?** 1. Further validation on the effect of dimensionality reduction of patient features on classification accuracy. 2. Latent space extrapolation for further identifications of feature importance. 3. Classifier provides a non-binary output (0-1) that can assist clinical decision more than direct hospitalization classification.

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.
- We plan to do a comparative analysis between the two methods gauging the accuracies/MbN would be beneficiary
- These groupings can be used to provide information for better prognosis of patients



Performance metrics for classification

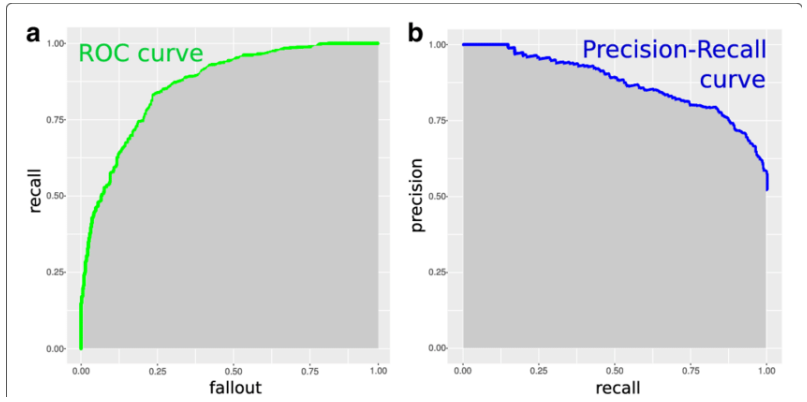
- Confusion matrix
- Specificity, Sensitivity, F1 Score
- Area Under Receiver Operating Characteristic Curve (AUC ROC)
- Area under Precision-Recall Curve (AUC PR)

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

$$F1_{score} = \frac{2 * TP}{2 * TP + FP + FN}$$

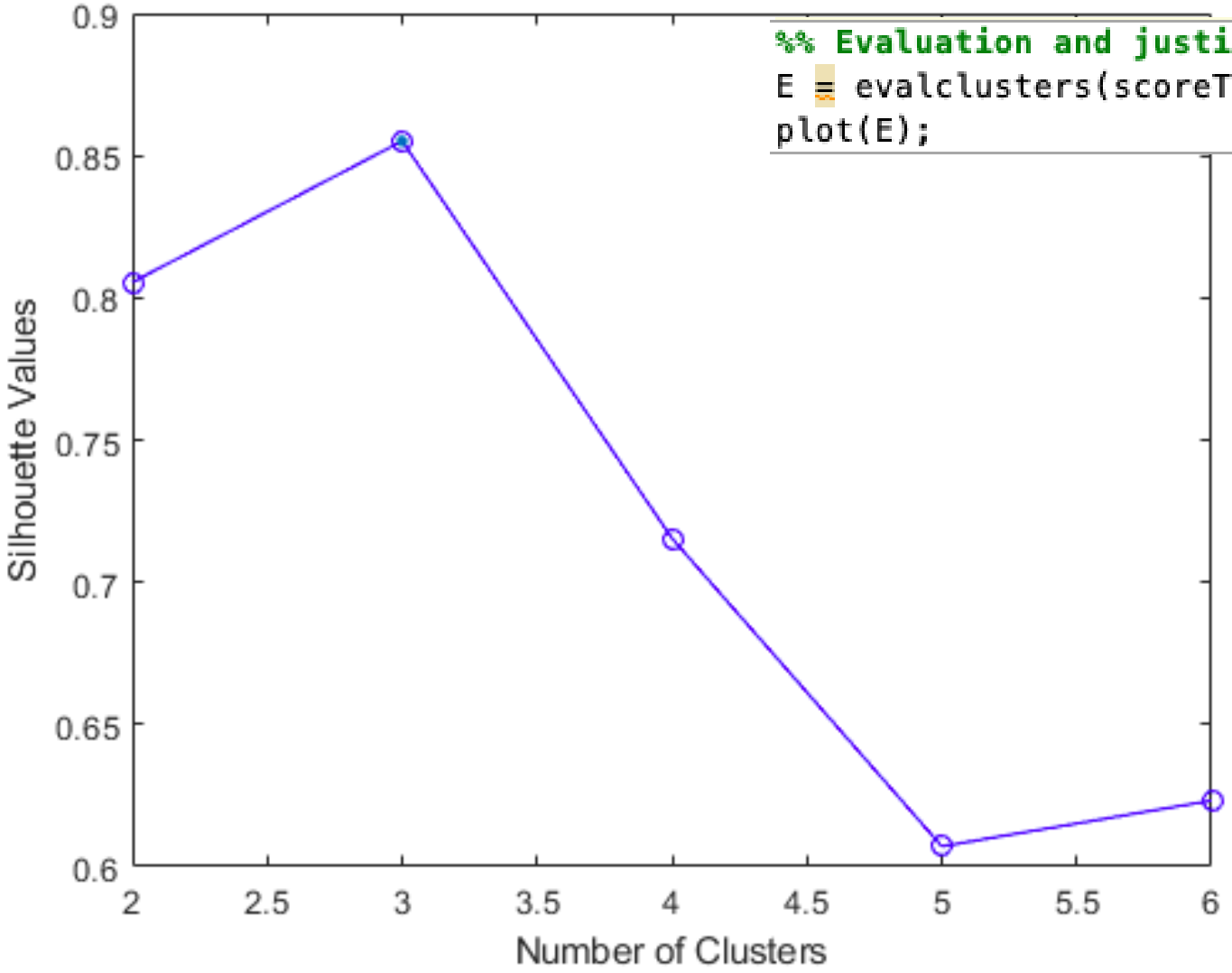


Performance metrics for clustering

The main performance method apart from Evaluating an optimal value of K. This can be done using

- Elbow's method – "elbow" method to help select the optimal number of clusters by fitting the model with a range of values for K. Initially this method is the backbone of clustering algorithm. Where the program is iterated till no data points change clusters, or the sum of distances is minimized
- Silhouette Evaluation - The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that well matched clusters

Example of Silhouette Evaluation

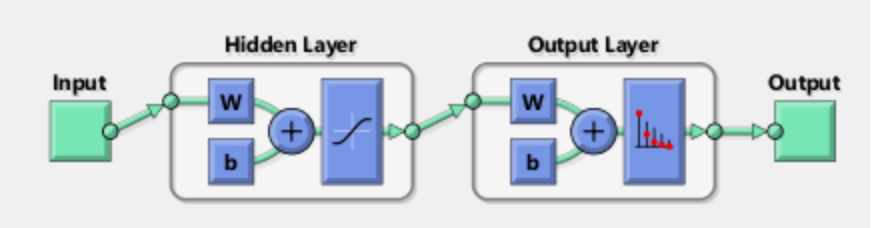



```
%% Evaluation and justification of K silhoutte
E = evalclusters(scoreTrain90,'kmeans','silhouette','klist',[1:12])
plot(E);
```

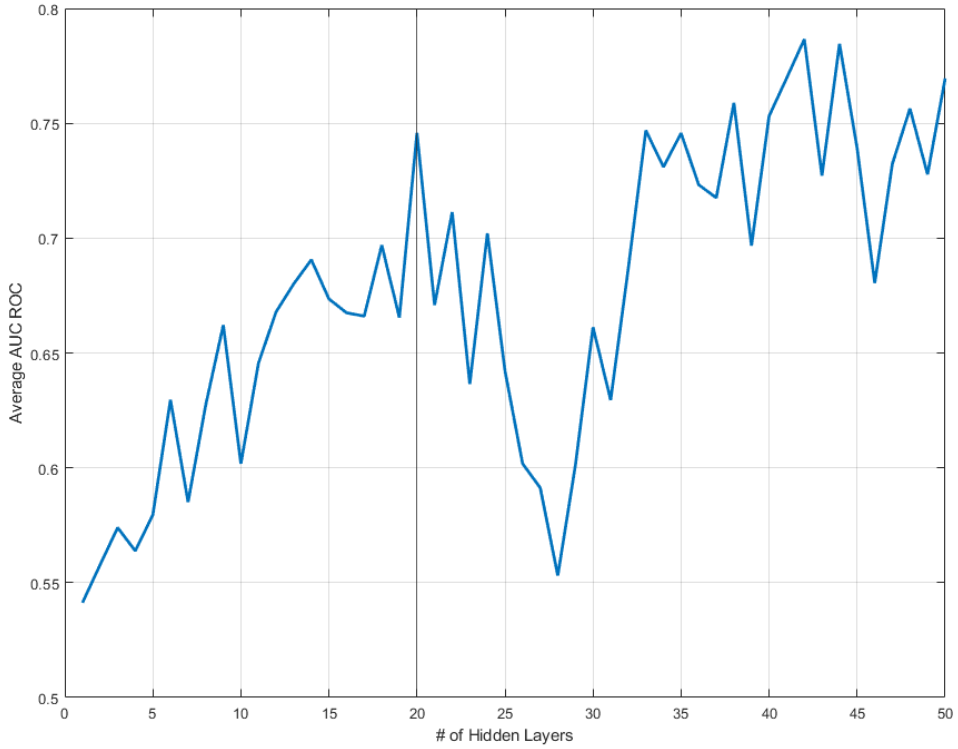
Model Parameter Selection: Autoencoder (architecture 1)

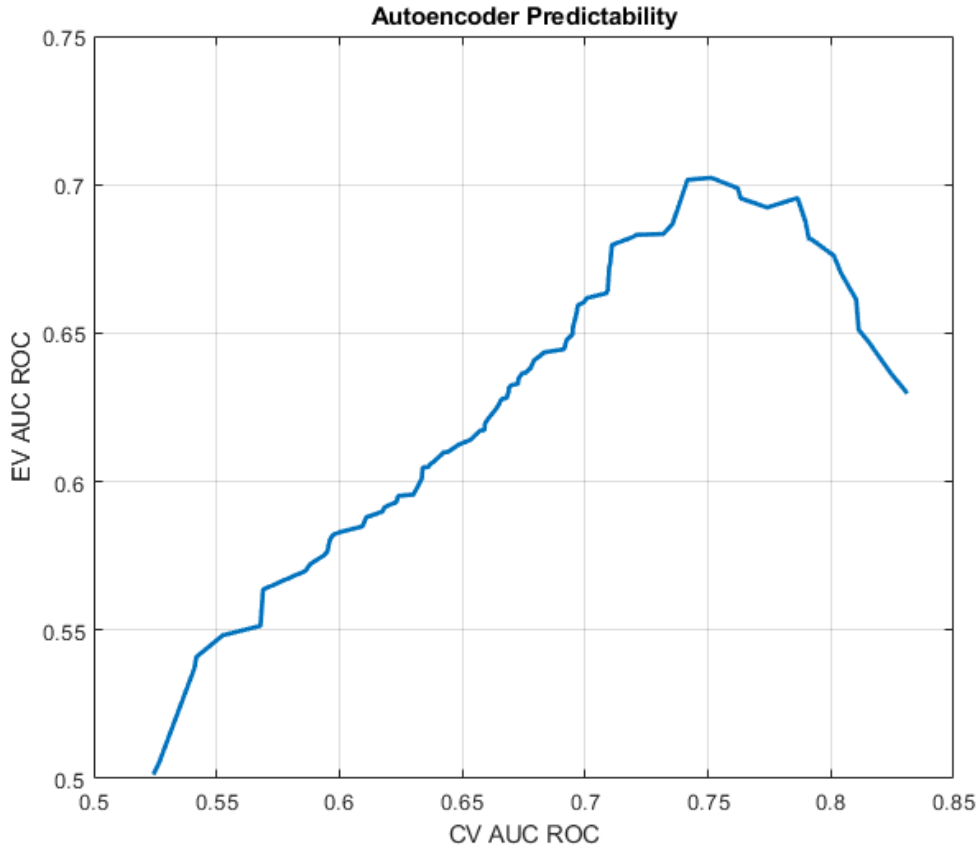
- Numbers of hidden layers (1 – 50) → **20**
- Training functions: **Scaled Conjugate Gradient (SCG)**, damped least-squares (DLS), Resilient Backpropagation, One Step Secant (OSS)
- Training performance function: Mean Square Error, **Cross-Entropy**, Sum Absolute Error, MESREG
- Training epochs: **Cross-Entropy-based**

Parameter Selection Criteria: Hyperparamter grid search using average AUC ROC from generated model as evaluative criteria.



e.g. # hidden layers evaluation; Predictability of CV vs EV





Model Parameter Selection: Autoencoder (architecture 2)

- Numbers of hidden layers = 9
- Training performance function: **Mean Square Error**
- Training epochs: **MSE based**
- **Grid search CV method for model parameter tuning (dynamic range of parameters used)**
  - activation = ['relu', 'tanh', 'sigmoid']
  - learn rate = [1E-0, 1E-1, 1E-2, 1E-3, 1E-4, 1E-5, 1E-6, 1E-7]
  - optimizer = ['sgd', 'adam']
  - batch_size = [16, 32, 64, 128]

Parameter Selection Criteria: Parameter tuning with grid search CV using testing AUC ROC from model as criteria for selection.

Optimum parameter values:

- Activation = 'relu'
- Learn rate = 1E-3
- Optimizer = 'adam'
- Batch size = 32

Modeling Results: Autoencoder (architecture 2)

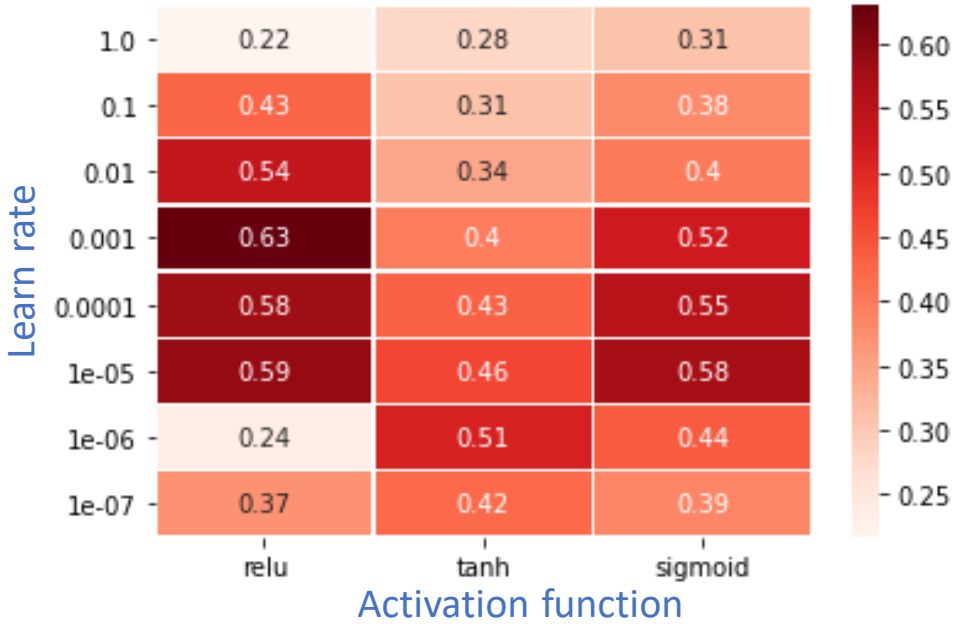


Fig 10 : Plotting range of parameters for 2 parameters amongst 4 used in model tuning.

```
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
_________________________________________________________________
```

Fig 11 : Final model after training.

Parameter for tuning	Range
activation	['relu', 'tanh', 'sigmoid']
learn rate	[1E-0, 1E-1, 1E-2, 1E-3, 1E-4, 1E-5, 1E-6, 1E-7]
optimizer	['sgd', 'adam']
batch_size	[16, 32, 64, 128]

Table 2 : Range of parameters used.

Quantitative Data Analysis Results: Autoencoder (architecture 2)

Cohort	AUC	F-1 score	Precision	Recall	Accuracy
Validation cohort	0.751	0.528	0.495	0.820	0.880
Testing cohort	0.630	0.473	0.384	0.705	0.812

Table 3 : Train dataset is stratified shuffle split with 80:20 ratio into train and validation cohorts, and the test dataset is considered from evaluation data from competition.

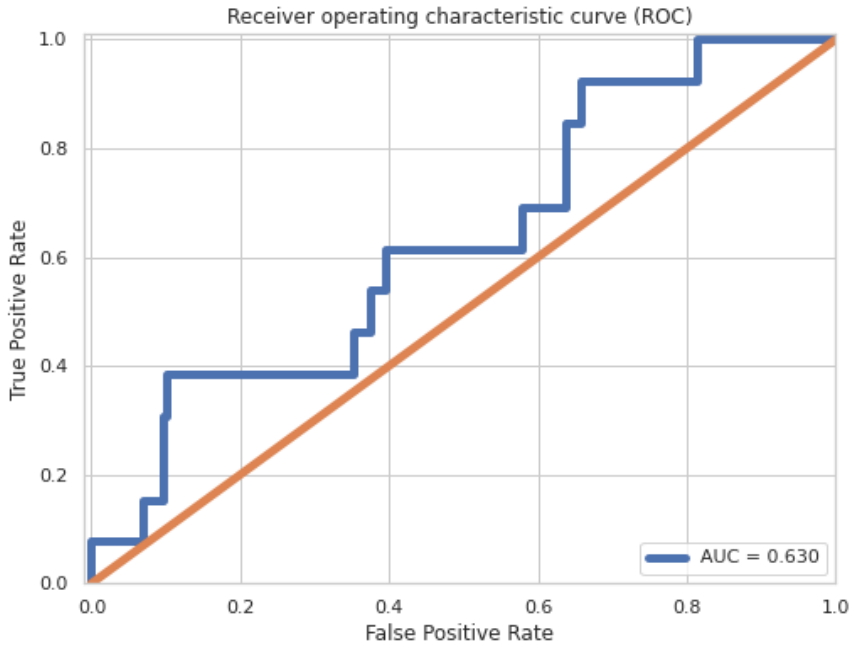


Fig 12 : Final model AUC on evaluation/testing dataset.

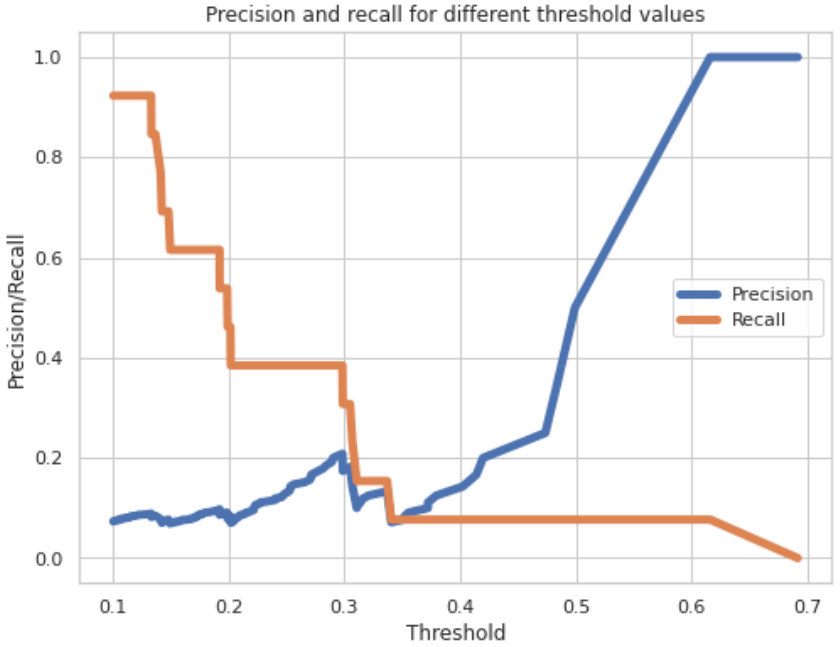
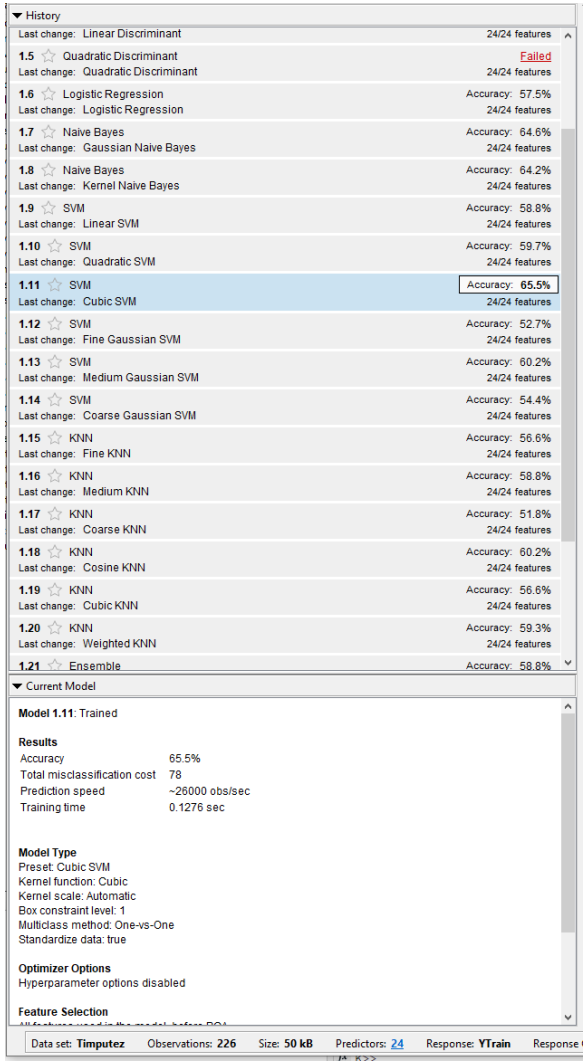
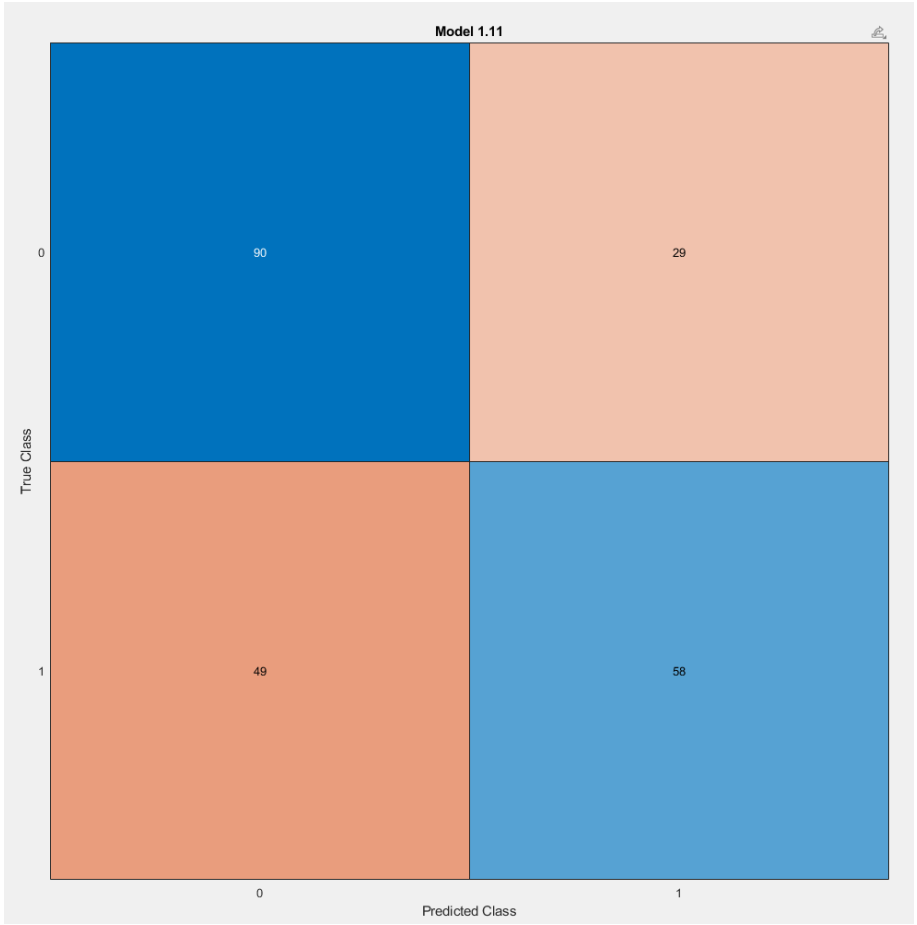
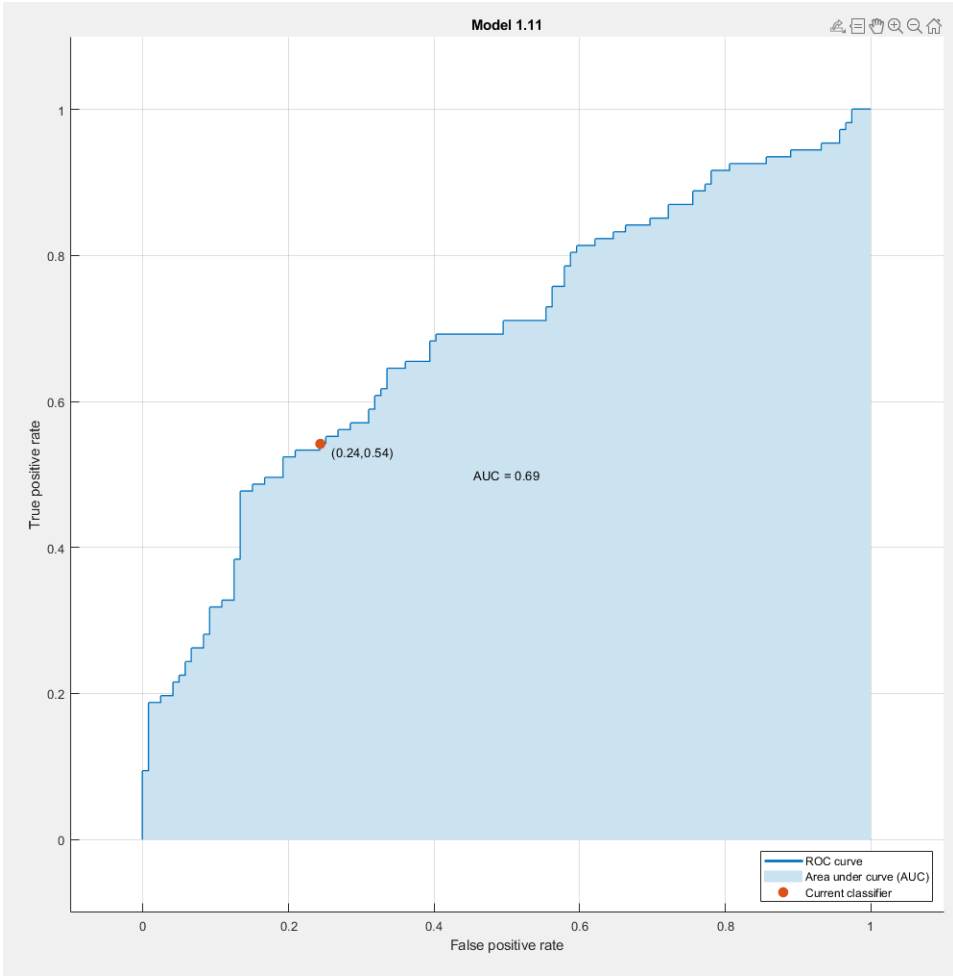


Fig 13 : AUC PR for model on evaluation data set is 0.124

Classical Method Result w/ clinical feature selection

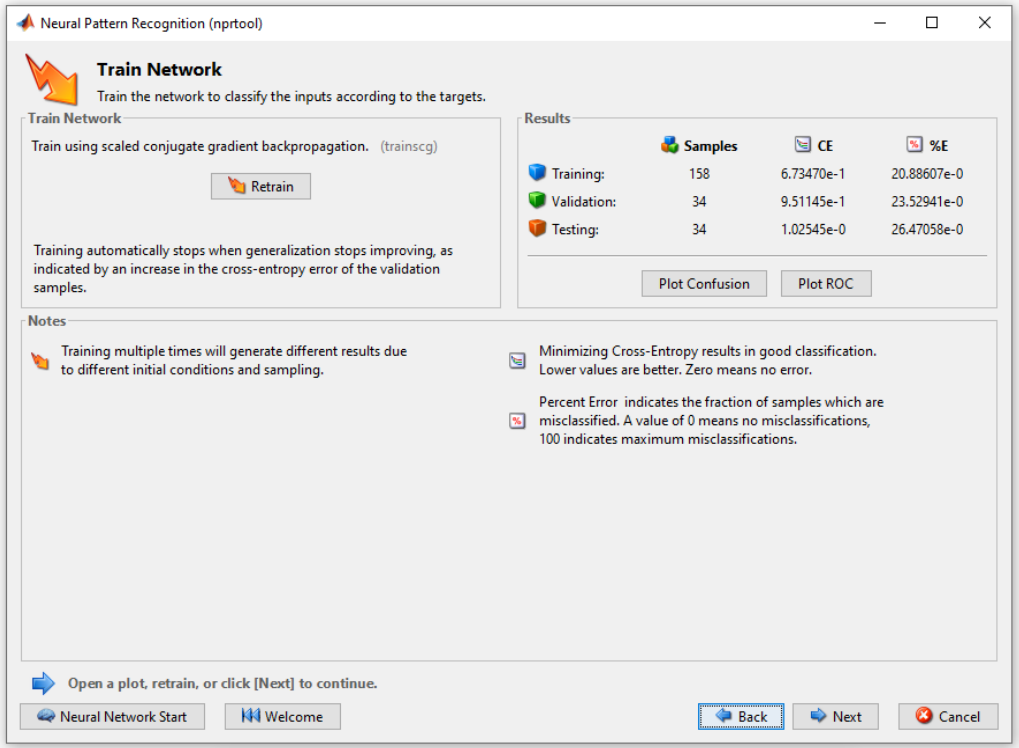


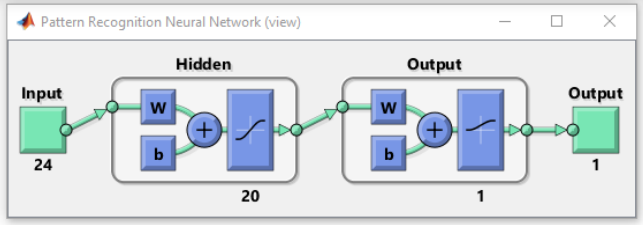


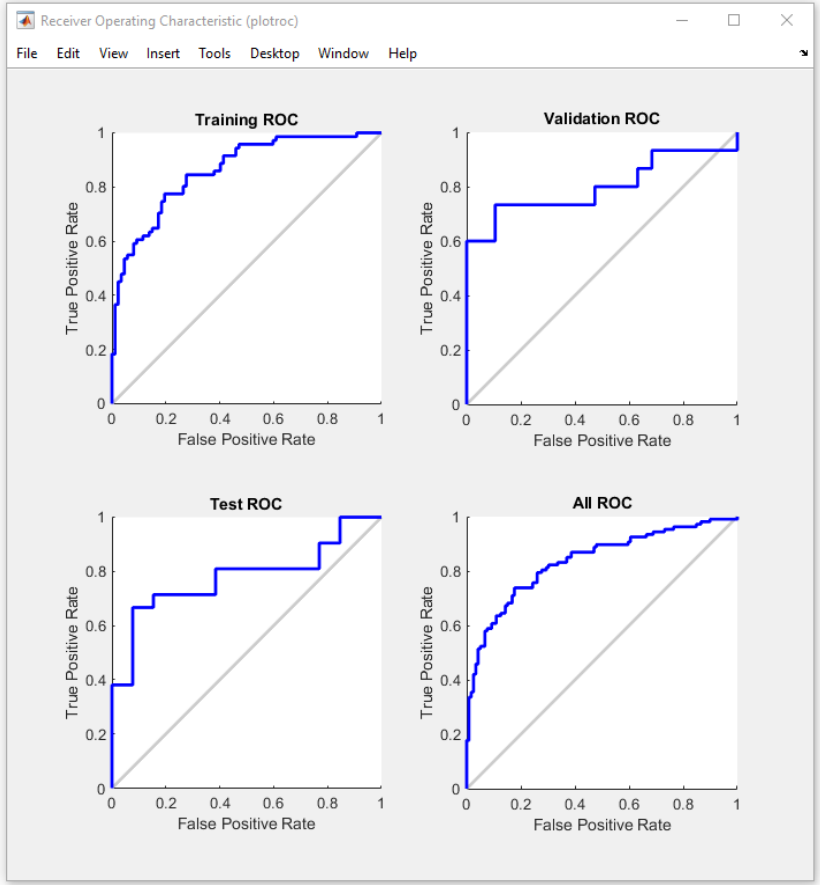


Improved classification accuracy with modified features

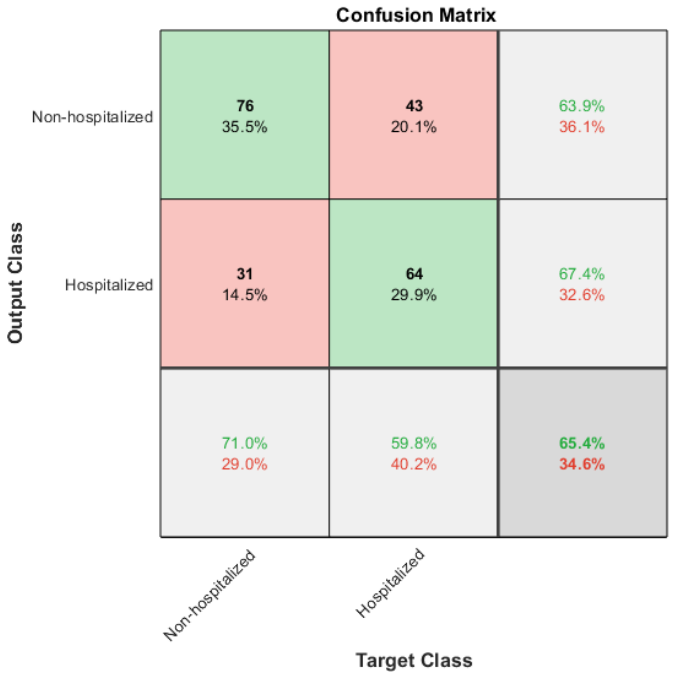
Autoencoder Classication Result on Training Set

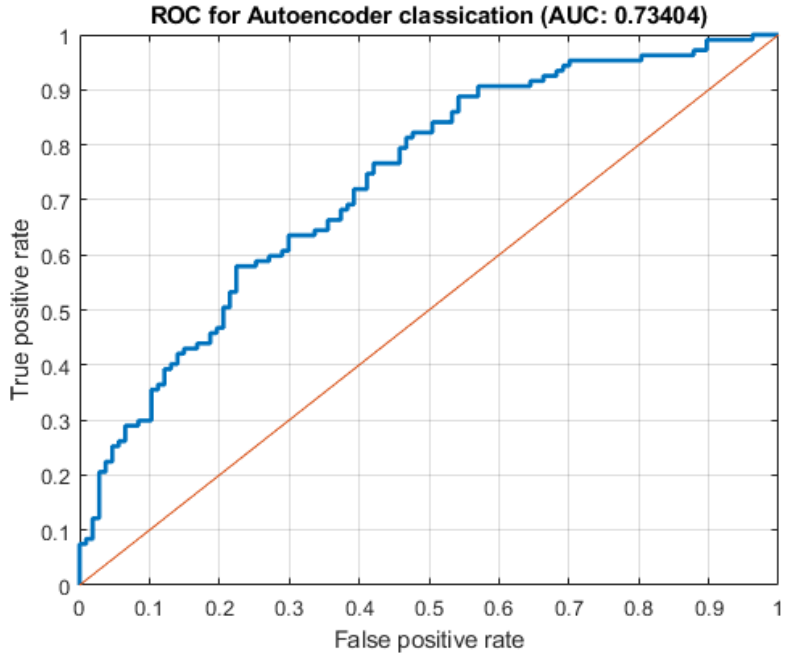


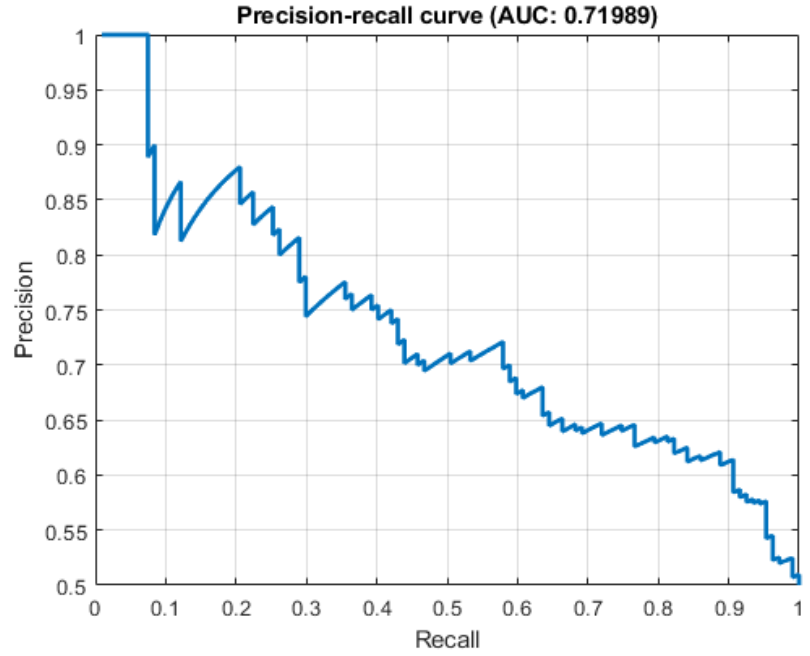




Autoencoder Classication Result on Training Set

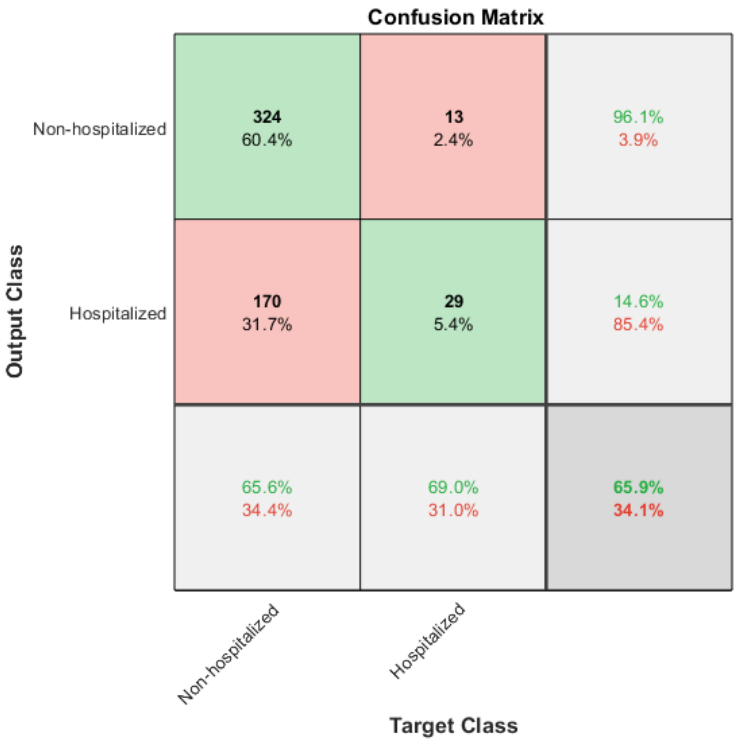


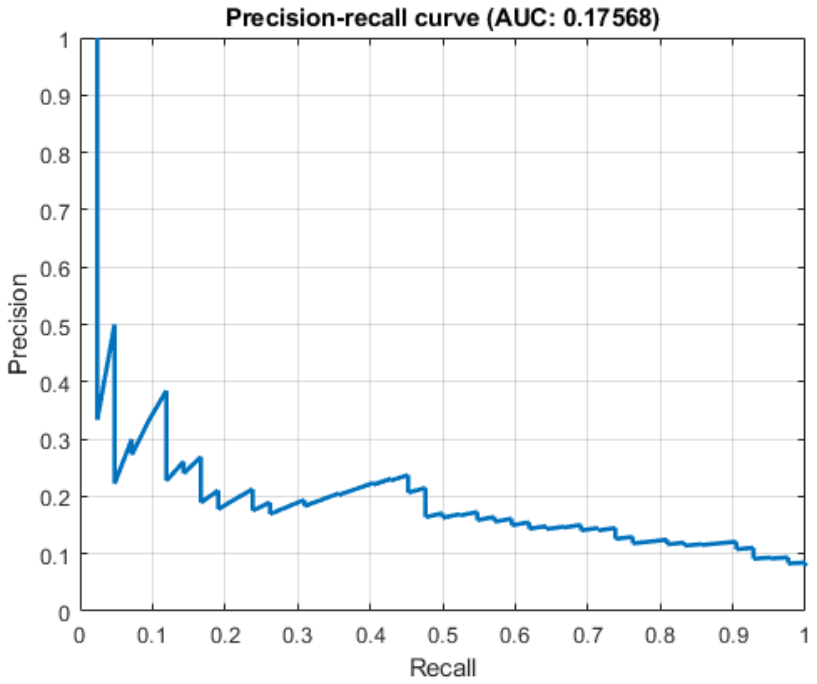


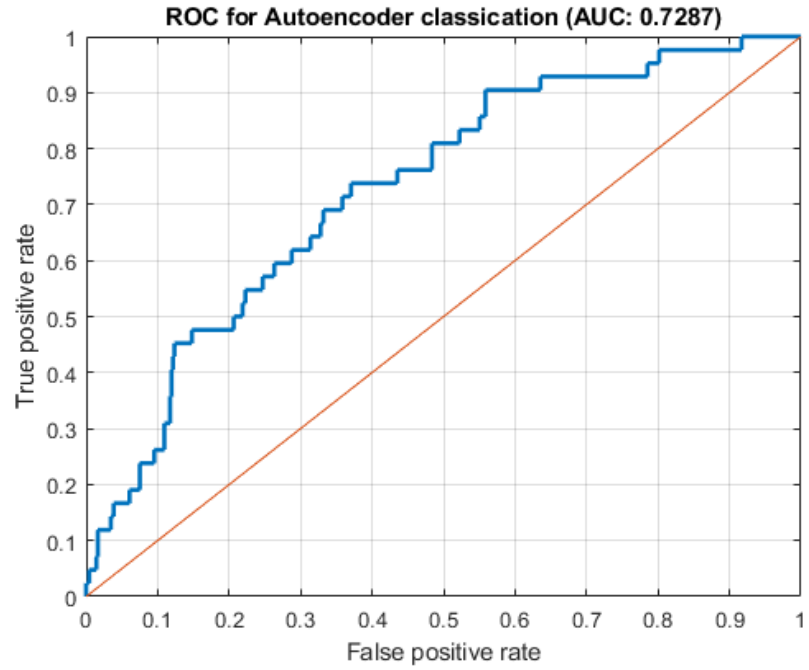


Specificity: 64 / 64 + 43 = 0.6;
Sensitivity: 76 / 76 + 31 = 0.71;
F1 score: 2 * 64 / 2 * 64 + 31 + 43 = 0.63

Autoencoder Classication Result on EV set

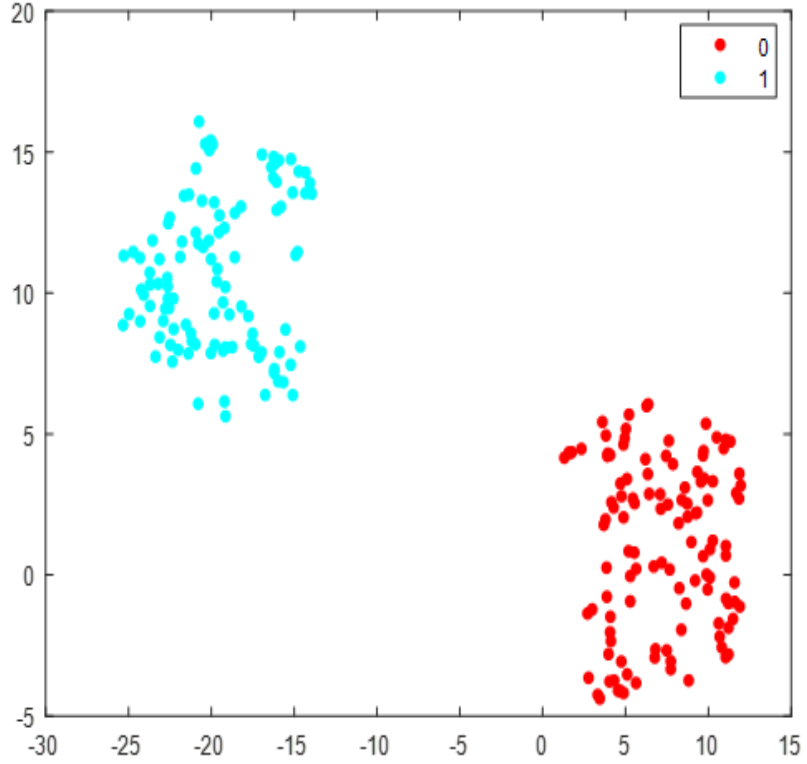




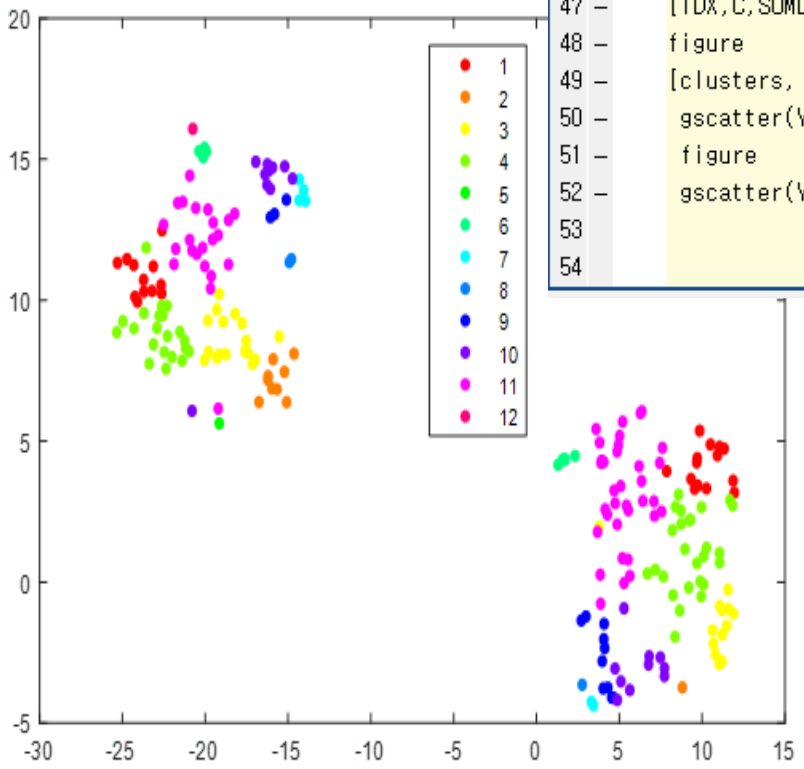


Specificity: 29 / 29 + 13 = 0.69;
Sensitivity: 324 / 324 + 170 = 0.656;
F1 score: 2 * 29 / 2 * 29 + 13 + 170 = 0.24

Clustering with K-mean without feature selection



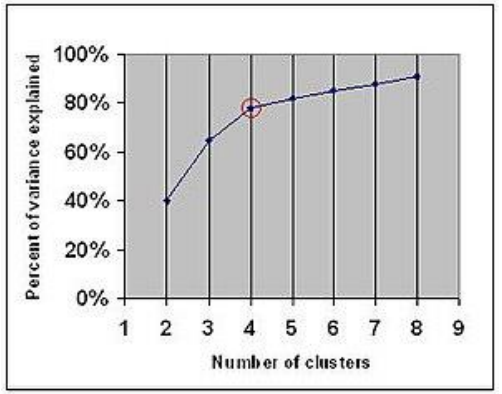
hospitalized vs Non hospitalized



Clustering with measurement data

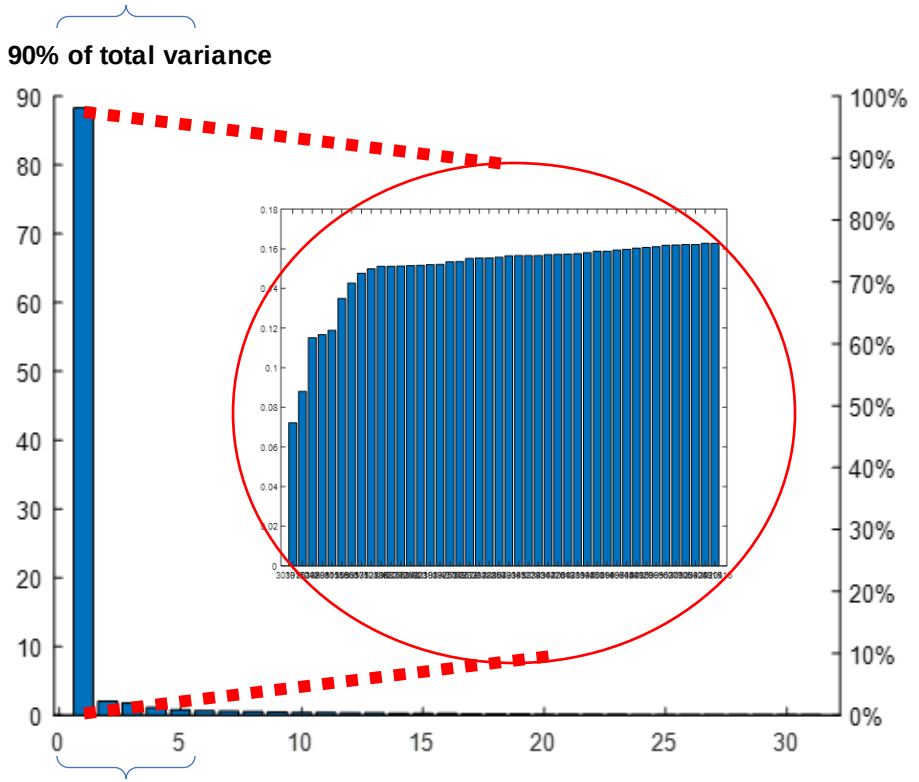
```
%% K-mean clustering and t-SNE plot
 Y = tsne(scoreTrain90wC);
[IDX,C,SUMD,K]=kmeans_opt(scoreTrain90);
figure
[clusters, centroid] = kmeans(scoreTrain90 , 12);
 gscatter(Y(:,1),Y(:,2),clusters)
 figure
 gscatter(Y(:,1),Y(:,2),GS)
```

ELBOW Method

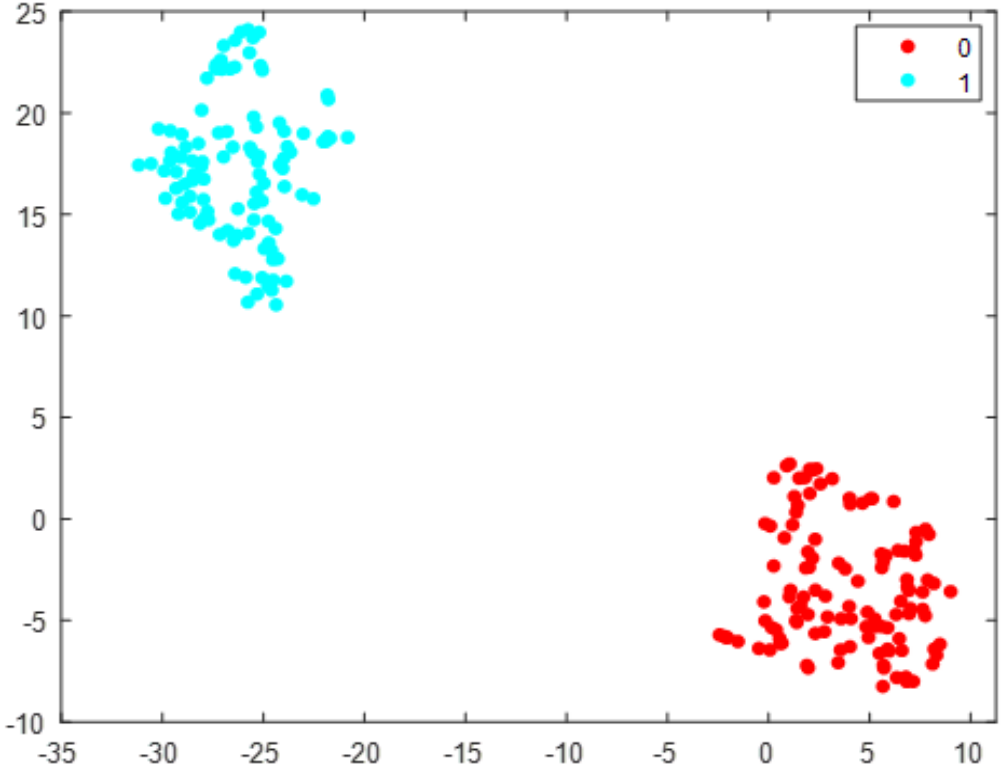


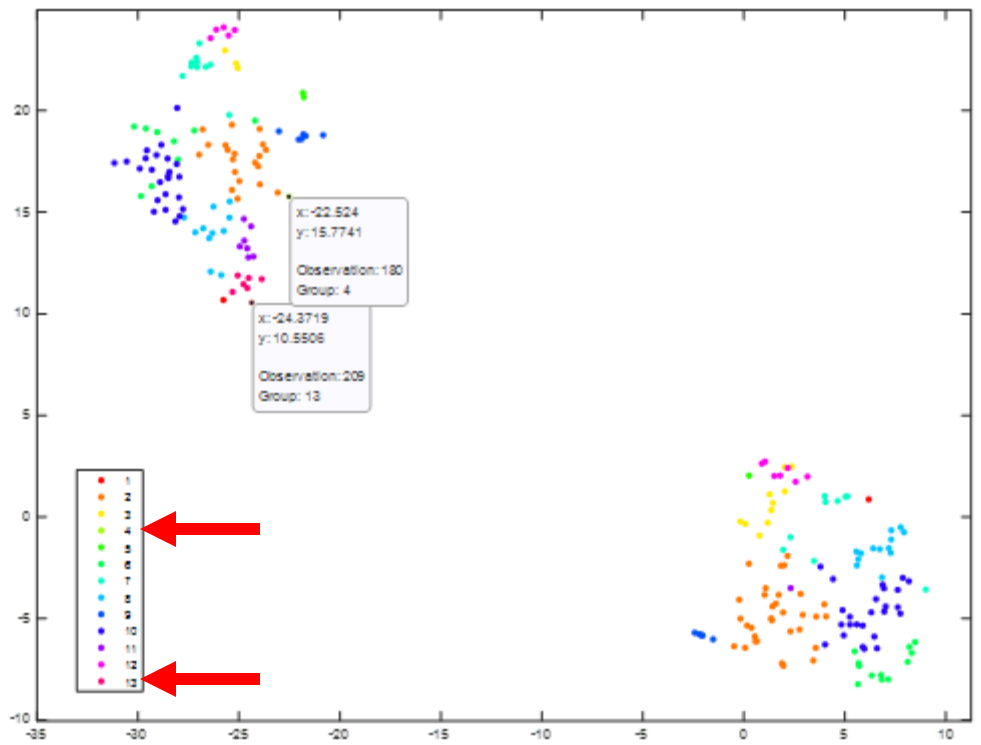
Dimension reduction with PCA

```
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);
```

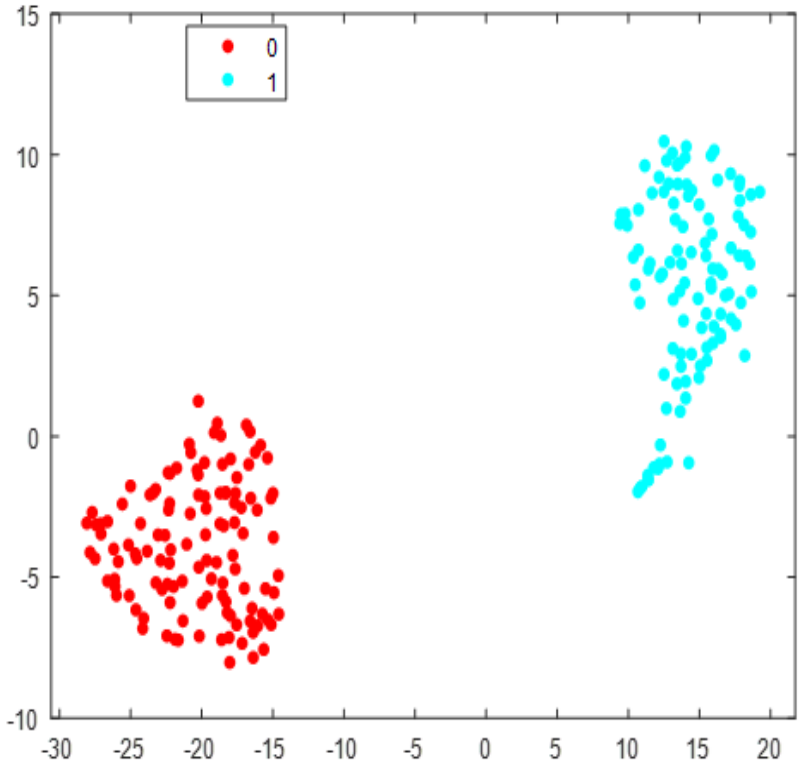


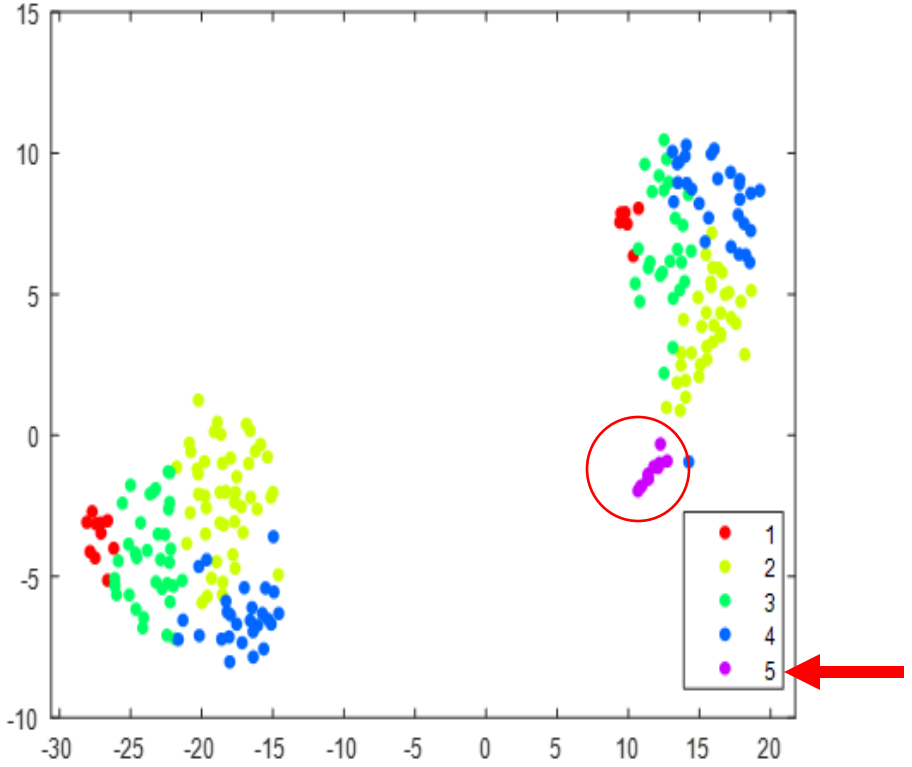
Clustering with feature selection with PCA





Clustering with feature selection from feature importance plot

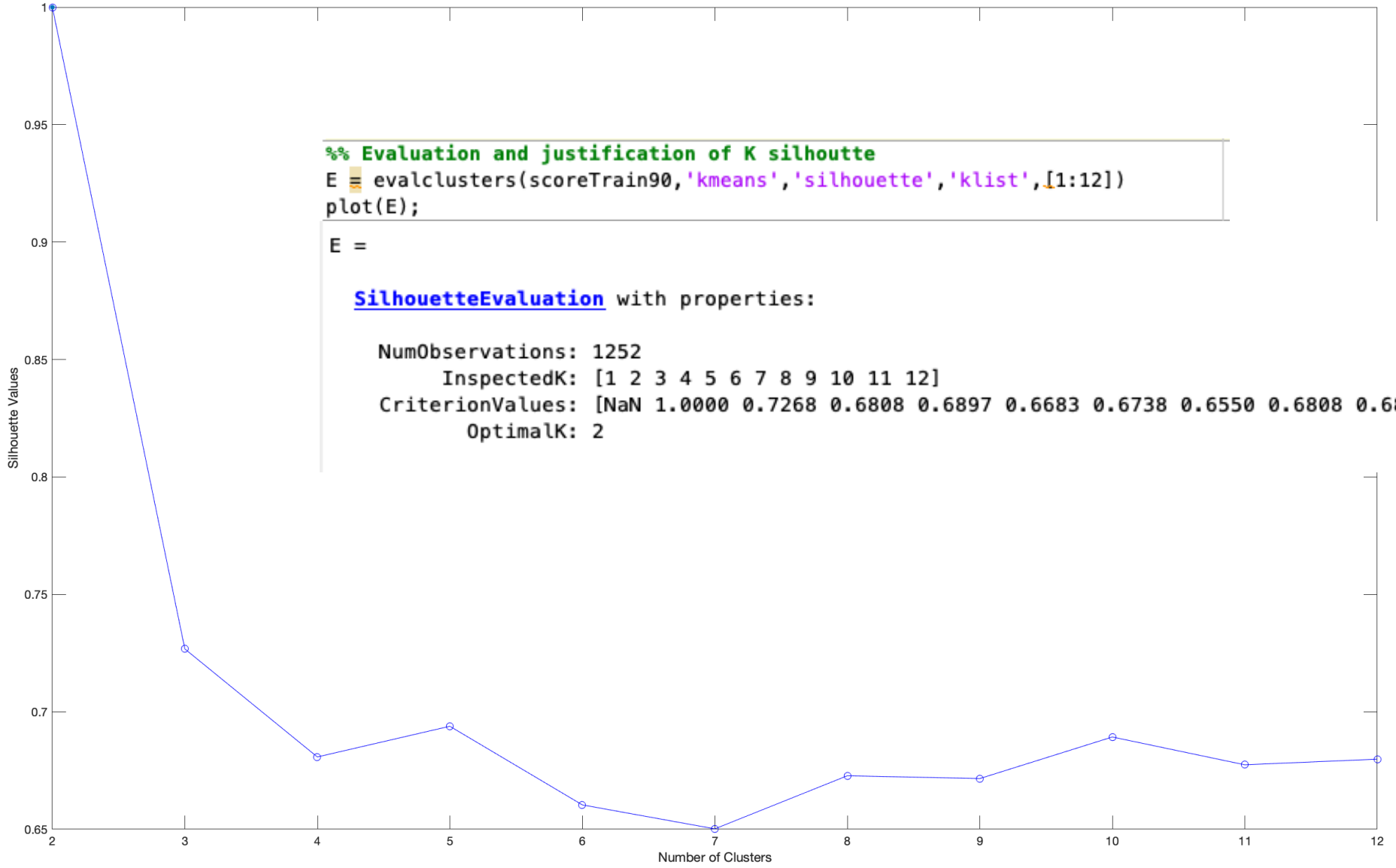




Evaluation of clustering feature selection with PCA

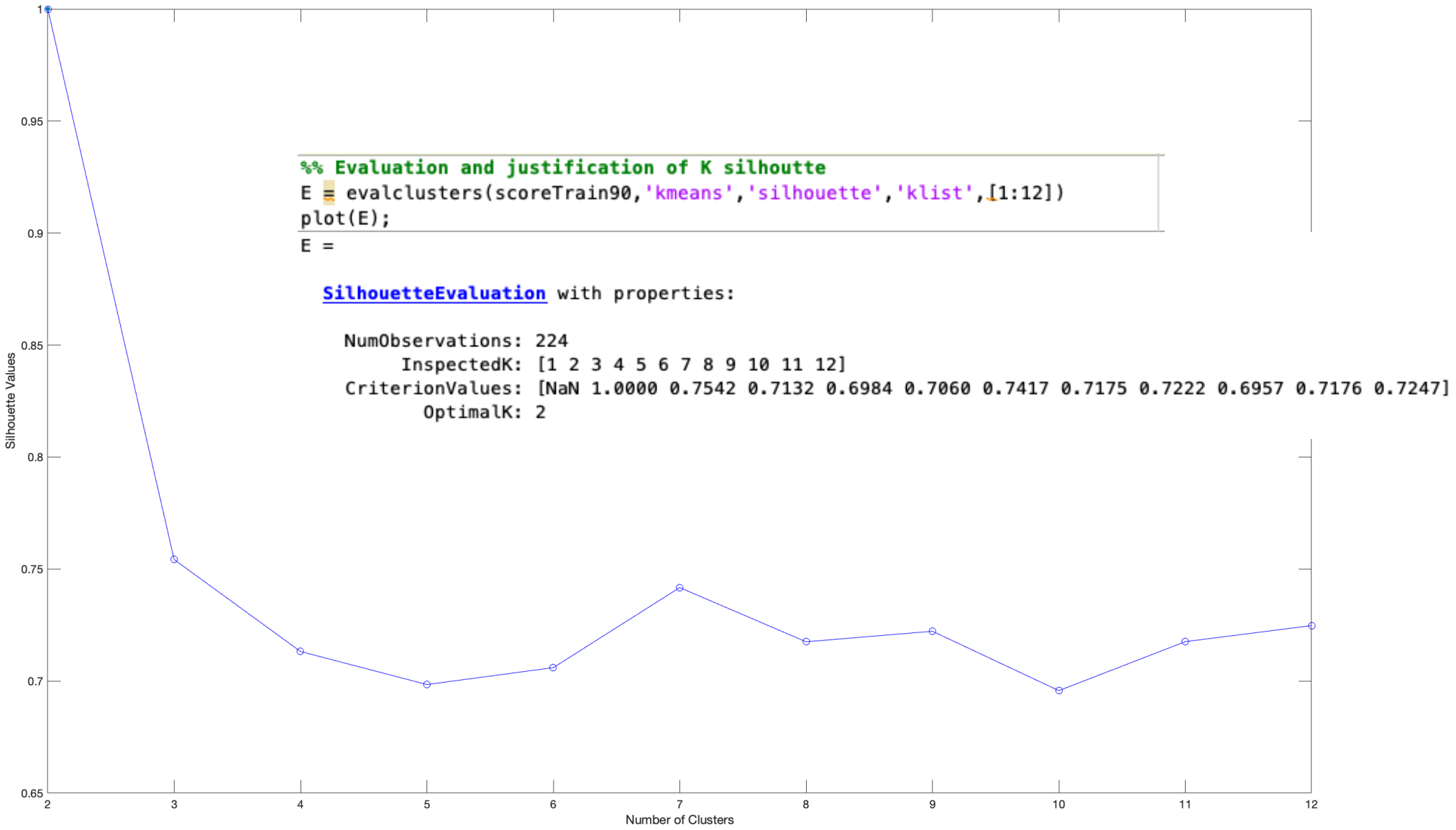
```
%% Evaluation and justification of K silhoutte
E = evalclusters(scoreTrain90,'kmeans','silhouette','klist',[1:12])
plot(E);
```

```
E =

  SilhouetteEvaluation with properties:

    NumObservations: 1252
         InspectedK: [1 2 3 4 5 6 7 8 9 10 11 12]
    CriterionValues: [NaN 1.0000 0.7268 0.6808 0.6897 0.6683 0.6738 0.6550 0.6808 0.6884 0.6857 0.6817]
           OptimalK: 2
```



Evaluation of clustering
feature selection from feature importnace plot

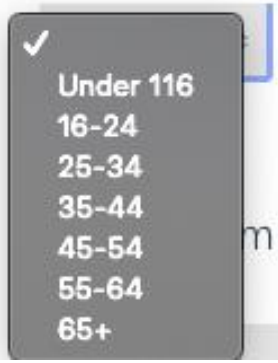
```
%% Evaluation and justification of K silhoutte
E = evalclusters(scoreTrain90,'kmeans','silhouette','klist',[1:12])
plot(E);
E =

  SilhouetteEvaluation with properties:

    NumObservations: 224
         InspectedK: [1 2 3 4 5 6 7 8 9 10 11 12]
    CriterionValues: [NaN 1.0000 0.7542 0.7132 0.6984 0.7060 0.7417 0.7175 0.7222 0.6957 0.7176 0.7247]
           OptimalK: 2
```



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

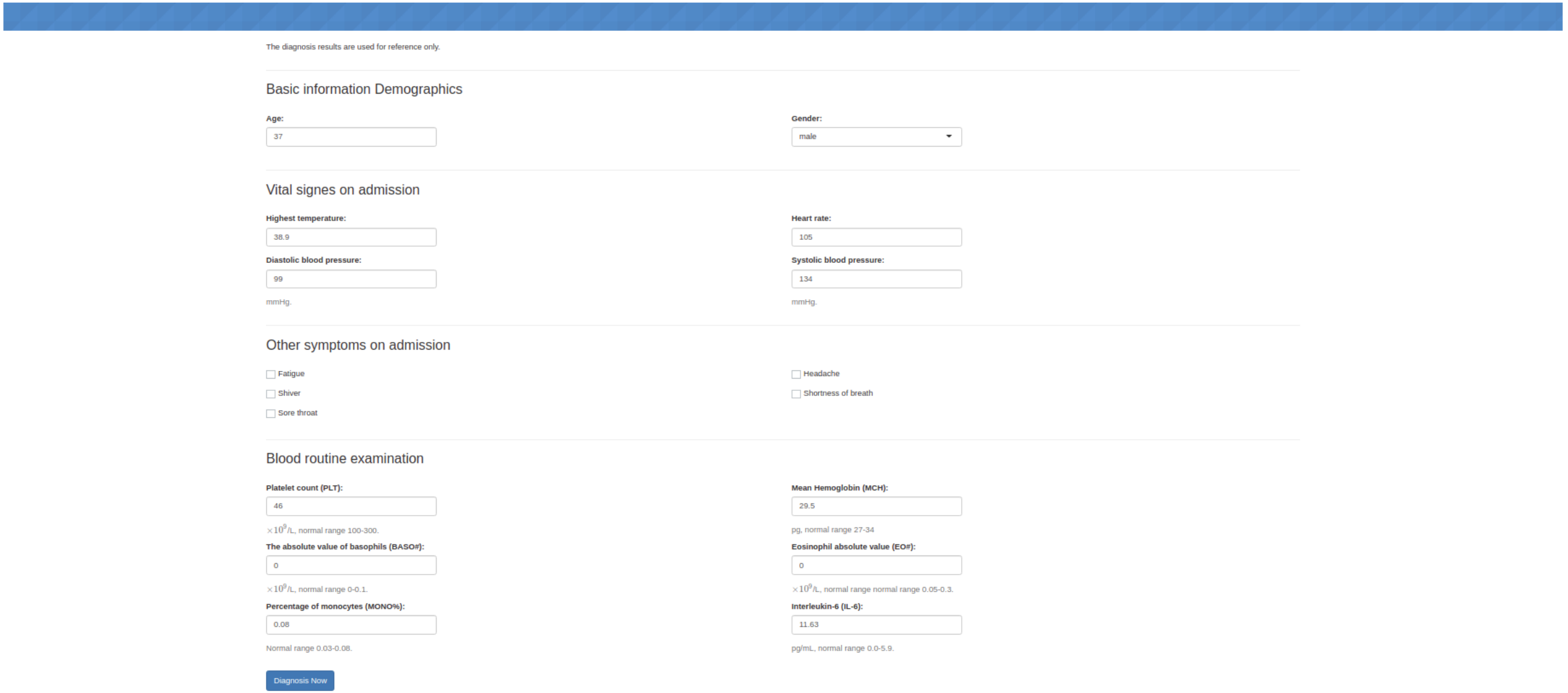
3. Maximum body temperature

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

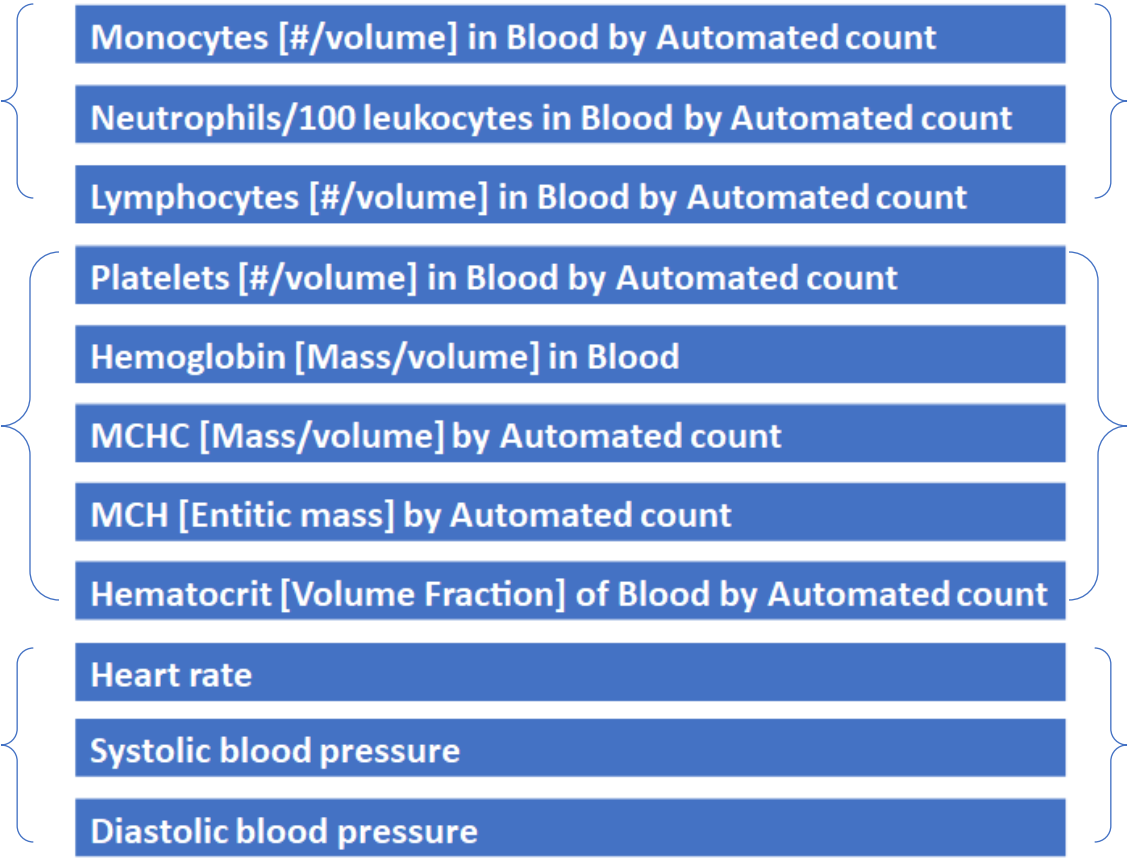
Sample GUI



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

K-mean clustering result interpretation

Overlapping features between PCA and feature importance plot



Monocytes, neutrophiles, and lymphocytes as key determinants of COVID-19 disease presentation and severity Brodin, (2021)

Close correlation of RBC and severe/hospitalized COVID-19 patients Renoux et al., (2021)

Effect of COVID-19 on Cardiovascular health Nishiga *et al.*, (2020)

K-mean clustering Summary and Conclusion

1. **Significance**
   - **Identified unique clusters and features that explain hospitalized patients**

2. **Weakness:**
   - **The synthetic data from COVID-19 DREAM Challenge was designed for classification problem and may not be suited for clustering/subphenotyping**
   - **Combined measurement and categorical/frequency dataset is not suitable for K-mean clustering method**

3. **Future work**
   - **Use latent space from autoencoder**
   - **Apply k-mode for mixed dataset**
   - **Implement UMAP for better visualization quality**

Classification task result interpretation

Table 4: Top features selected for making classification predictions of hospitalized vs non-hospitalized patients

Concept id	Feature
cid_3033575	Monocytes [#/volume] in Blood
cid_3013650	Neutrophils [#/volume] in Blood
cid_3008342	Neutrophils/100 leukocytes in Blood
cid_3004327	Lymphocytes [#/volume] in Blood
cid_3027018	Heart rate
cid_3009744	MCHC [Mass/volume]
cid_3023599	MCV [Entitic volume]
cid_3011948	Monocytes/100 leukocytes in Blood
age	Age of patient
cid_3012888	Diastolic blood pressure
cid_3012030	MCH [Entitic mass]
cid_3004249	Systolic blood pressure
cid_3024929	Platelets [#/volume] in Blood
cid_3010156	C reactive protein [Mass/vol] in Serum
gender	Gender of patient
cid_3020891	Body temperature
cid_3000963	Hemoglobin [Mass/volume]
cid_3023314	Hematocrit [Volume Fraction] of Blood
cid_3028615	Eosinophils [#/volume] in Blood
race	Race to which person identifies
cid_3037511	Lymphocytes/100 leukocyte
cid_3013429	Basophils count

- Monocytes, neutrophiles, and lymphocytes as key determinants of COVID-19 disease presentation and severity *Brodin et al. (2021)*
- Close correlation of RBC and severe or hospitalized COVID-19 patients. *Renoux et al. (2021)*
- Effect of COVID-19 on Cardiovascular health. *Nishiga et al., (2020)*
- Blood routine values at the time of admission readily available, play a critical role in deciding patient critical situation. *Feng et al. 2020*

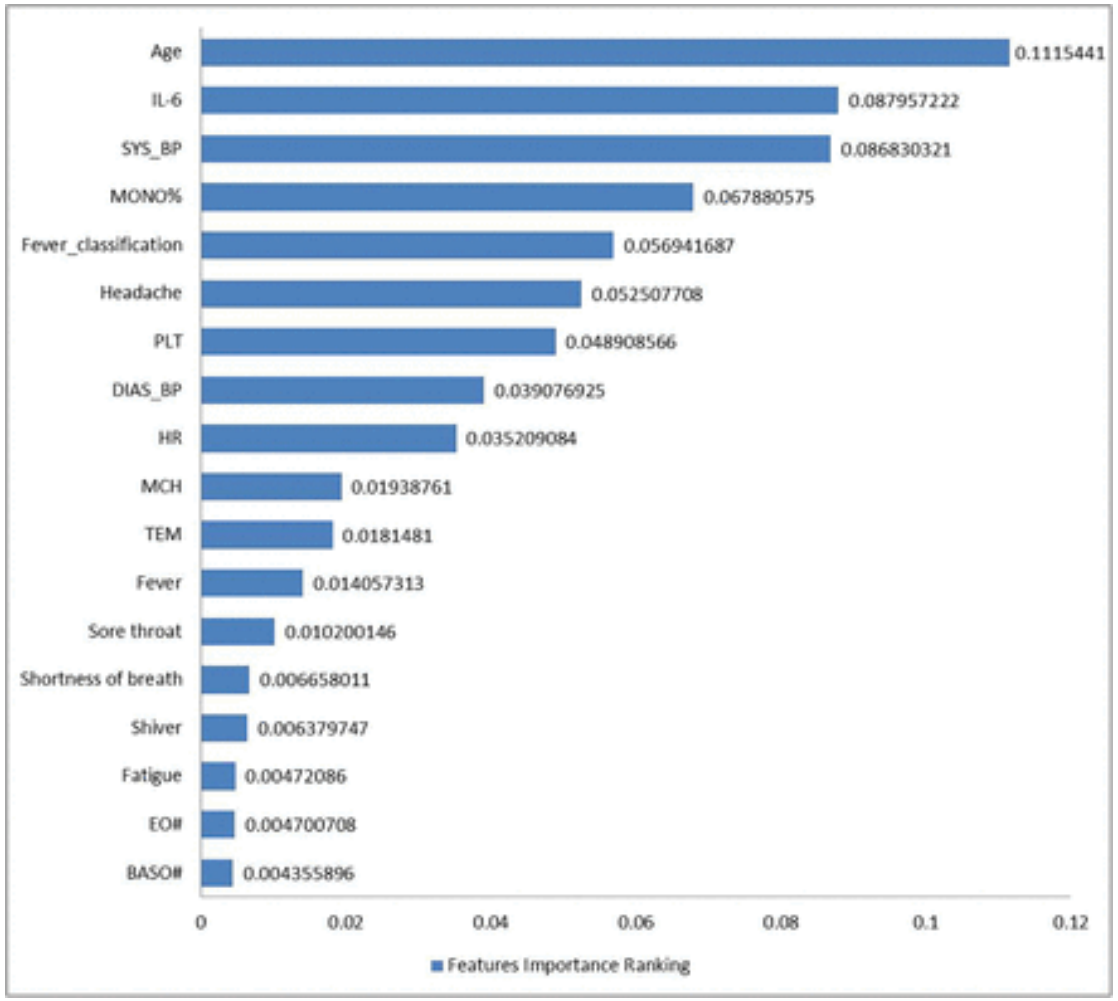


Fig 15: Top features from the Covid-19 article for Diagnosis aid system

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

	All patients	Non hospitalized	Hospitalized
Cohort (n)	536	500	36
Age (in years) (median)	45	46	36
Gender (n%)			
Male	264	244	20
Female	272	256	16
Vital signs			
Heart rate n/min (median)	82.66	82	88.66
Diastolic BP mmHg (median)	73	72.5	75.83
Systolic BP mmHg (median)	117	116	122.83
Body Temperature deg. C (highest)	37.36	37.36	38.3
Blood routine values			
Hemoglobin (g/L)	11.6	11.6	11.3
Hematocrit	34.66	34.66	33.5
Platelet count (x 10^9 / L)	217	220	201.5
Lymphocyte count (x 10^9 / L)	1.395	1.35	1.73
Lymphocyte ratio (%)	22.16	22.16	22.5
Neutrophil count (x 10^9 / L)	3.555	3.59	2.745

Classification Result Comparison

Rank	Submission Id	Created On	Submitter	description	AUPR	AUROC	Train Dataset Version	Infer Dataset version	Repository	Digest	ranked_features
1	9710459	02/15/2021 7:48 AM	Home Sweet Home	hospitalization baseline prediction	0.2154	0.8103	01-26-2021	01-26-2021	docker.synapse.org/syn22242573/co...	sha256:f59bc18f34be5e6dc402945...	age,gender,race
2	9711406	03/26/2021 3:16 PM	@Amhar	hospitalization baseline prediction	0.2046	0.8025	01-26-2021	01-26-2021	docker.synapse.org/syn23763328/co...	sha256:ddacec32dcdb20891c41b11...	age,gender,race
3	9711200	03/09/2021 11:49 AM	sucovid	COVID diagnosis baseline prediction	0.2025	0.7532	01-26-2021	01-26-2021	docker.synapse.org/syn22156808/q2...	sha256:6ecce80c88c939f8df38288e...	age,ventilator
3	9711217	03/12/2021 1:50 PM	@egearikan	COVID diagnosis baseline prediction	0.2025	0.7532	01-26-2021	01-26-2021	docker.synapse.org/syn22842873/co...	sha256:f76a9cff23b60670e449c624...	age,ventilator
3	9711221	03/13/2021 2:34 AM	@egealpay	COVID diagnosis baseline prediction	0.2025	0.7532	01-26-2021	01-26-2021	docker.synapse.org/syn22156808/q2...	sha256:1fd76c54c824d809ad6e50c...	age,ventilator
3	9711223	03/13/2021 3:01 AM	@ealpy	COVID diagnosis baseline prediction	0.2025	0.7532	01-26-2021	01-26-2021	docker.synapse.org/syn22333713/q2...	sha256:1fd76c54c824d809ad6e50c...	age,ventilator
3	9711245	03/16/2021 4:35 PM	@semayilmazer	COVID diagnosis baseline prediction	0.2025	0.7532	01-26-2021	01-26-2021	docker.synapse.org/syn24829156/aw...	sha256:bd08e75d55c74b52e0d88f6...	age,ventilator
8	9711247	03/16/2021 5:53 PM	@alperbingol	COVID diagnosis baseline prediction	0.1887	0.7666	01-26-2021	01-26-2021	docker.synapse.org/syn24829191/aw...	sha256:3c759dd12d828edf1520783...	age,ventilator
9	9710542	02/16/2021 9:35 PM	@ivanbrugere	COVID diagnosis baseline prediction	0.1833	0.7878	01-26-2021	01-26-2021	docker.synapse.org/syn20833371/co...	sha256:d134a9ef21cd81313ed4c0c...	all concept ids
10	9710539	02/16/2021 8:53 PM	Bryson and Yao Team	COVID diagnosis baseline prediction	0.1585	0.6367	01-26-2021	01-26-2021	docker.synapse.org/syn22248060/co...	sha256:d6bea121564e8bcbc1edd73...	cough,fever,loss of smell,sore throat,r...
11	9710308	02/09/2021 8:48 PM	ArkansasAICampus20	hospitalization baseline prediction	0.1572	0.7759	01-26-2021	01-26-2021	docker.synapse.org/syn22351720/aic...	sha256:7ab91d5e93da10b5acdfc97f...	age,gender,race, measurement, condi...
12	9711342	03/20/2021 10:19 PM	Social Distancer Team	Covid baseline prediction	0.1425	0.6081	01-26-2021	01-26-2021	docker.synapse.org/syn22089460/q1...	sha256:39fae11544b7d092f89cc00...	see features.json file
13	9711111	03/01/2021 9:44 PM	QTeam	hospitalization baseline prediction	0.1179	0.7029	01-26-2021	01-26-2021	docker.synapse.org/syn24262913/ba...	sha256:5ce379feacc46a6121fab454...	age,gender,race
13	9710244	02/06/2021 12:23 AM	@thomas.yu	COVID diagnosis baseline prediction	0.1179	0.7029	01-26-2021	01-26-2021	docker.synapse.org/syn21849256/co...	sha256:f042a09c1bc5cb6b6372621...	cough,body temperature,hematocrit,...

Our Best Classification result: AUROC: 0.7287; AUPR: 0.1757
Sigh Still need more improvement...

Autoencoder classification Summary and Conclusion

1. Significance
   - Retain relatively high accuracy of hospitalization prediction given low clinical measurement features
   - General numerical value of likelihood of hospitalization, which can assist clinical decision more than binary output.

2. Weakness:
   - The synthetic data from COVID-19 DREAM Challenge contains measurement data across a decade of time, from which many measurement data would not be indicative for hospitalization prediction

3. Future work
   - Implement better hyperparameter tuning techniques (i.e. Bayesian optimization)
   - Score the quality of measurement data based on temporal relevance