Identification of Sub-phenotypes of COVID-19 within Patient Population

Bryce Butler  
*Georgia Institute of Technology*   
Atlanta, U.S.A  
bbutler41@gatech.edu

Rohan Bhukar   
*Georgia Institute of Technology*   
Atlanta, U.S.A  
rbhukar3@gatech.edu

Seonggeon Cho  
*Georgia Institute of Technology*  
Atlanta, U.S.A  
scho392@emory.edu

Zhonghao Dai  
*Georgia Institute of Technology*  
Atlanta, U.S.A  
zdai43@gatech.edu

*Abstract*— COVID-19 over the past year has become the one of the maximum impact pandemics the human population has witnessed in the lifetime. Millions of cases affecting 100s of countries worldwide has brought the world’s growth rates to a slowdown and impacting almost all economies on the global scale. Millions of deaths have happened at an unprecedented scale, with the coronavirus mutating more than 12,000 times. Variety of symptoms based on patient's pre-conditions and type of covid variants exits. Due to this variability, clinical decision making is challenging with the problem still persisting in 2021 as well. Identification of COVID-19 sub-phenotypes could lead to better understanding of the diverse host responses that result in these heterogenous presentations. Current machine learning models of classification are limited to hospitalized patients, with models limited to including patients mostly more than 60 years old. In this article, we are presenting clustering and classification methods to address above issues using deep learning-based methods. Our proposed solution achieves an AUC 0f 0.7287 on evaluation dataset from the COVID-19 DREAM challenge, while the clustering pipeline identifies unique patients' clusters in hospitalized and no-hospitalized patients using the synthetic dataset. This can be used to derive clinical relevance upon tested on real-world datasets and, at the same time provides insights into a new architecture of DNN for patient classification and methods for clustering.

Keywords—Machine learning, COVID-19, autoencoder, K-means clustering, Electronic Health Records

# Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2/ COVID-19) has caused more than thirty million infections and about 574,679 deaths in the United States. With a 7-day case rate per 100,000, 102.3, it only seems that the health care systems will continue to under pressure for the foreseeable future. Patients experienced a variety of symptoms based on their preconditions and type of COVID-19 variants. This variability makes clinical decisions challenging. Understanding the nature of the virus among hospitalized patients is vital to help optimize prognosis and overall clinical management strategies. Identification of COVID-19 sub-phenotypes could lead to a better understanding of the diverse host responses that result in these heterogeneous presentations.

There is a rise in solutions to this problem, utilizing classical machine learning models (such as gradient boosting, logistic regression) and unsupervised learning (like K-means clustering) (Table 1). Most of these solutions exist in COVID-19 sub phenotyping utilizing only single cohort data and basic clinical presentation factors. These methods created have been seen to aim to predict only mortality, or they present novel phenotypes. Unfortunately, researchers use data set derived from only a smaller cohort size, which is the current bottleneck in multi-modal comprehensive research.

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. It would be imperative to use more extensive multi-modal data that measures vital (such as temperature, heart rate, respiratory rate, O2 saturation). Patient clinical presentations also need to be factored in to develop better and novel patient sub-phenotypes and develop more accurate predictive models for clinical use-cases that can handle time-irregularities in the datasets.

We will be using the synthetic data from the COVID-19 DREAM Challenge, data made by the organizers, to give a realistic example of the challenge data while protecting patient privacy. The organizers got the original challenge data from the Enterprise Data Warehouse (EHR collected and stored from multiple large hospitals in the University of Washington Medical System). The data has been converted to the OMOPv5.3 Common Data Model and represents ten years of clinical records (2010-2020) from ~9,500 patients, each with at least one visit record and at least one test COVID-19. These records include medications prescribed, patient conditions, observations such as blood pressure and heart rate, demographic information, procedures, and laboratory measurements. The patients in the dataset, ~800 patients, have tested positive for COVID-19, with all of these patients having at least one visit. More details of the data will be given in the data processing section.

Our primary objective was to identify the principal features/feature sets drive model prediction using K-means clustering, classical classification algorithm, and an autoencoder. Conduct feature extraction on sub-phenotype clusters and extrapolate clinical significance. Our secondary objectives were to understand the significance of our findings then see how we can improve upon them in future works

# Method and Sytem Design

**Data Methodology:**

Within this study, the datasets available upon participation and registration from the COVID-19 DREAM challenge span over the two modalities which are the time independent patient’s demographic data, and time dependent measurement, condition occurrence, device exposure, drug exposure, observation, observation period, procedure occurrence and hospital visits data. The datasets were downloaded from COVID-19 DREAM challenge (<https://www.synapse.org/#!Synapse:syn21978034>). Though the real data in this challenge was collected from multiple large hospitals in University of Washington Medical System including Harborview Medical Center, UW Medical Center, and Northwest Hospital and Medical Center as well as hundreds of regional clinics across Washington. According to organizers, the data has been converted to the OMOPv5.3 Common Data Model and represents 10 years of clinical records (2010-2020) from more than 1000 patients, each of whom has at least one visit record and at least one test for COVID-19. The organizers decided to provide only synthetic data for participants, providing a realistic example of challenge data. The individual datasets size can be described through figure 2A, where column values represent sizes of training and evaluation sets from multiple files. A total of 1251 samples/patient data was part of training set and 536 samples within evaluation set. In terms of quality, the synthetic dataset had high missingness for multiple features of clinical importance, with the measurement data, condition occurrence and person data having satisfied the conditions for imputation as per missing at random (MAR) category.

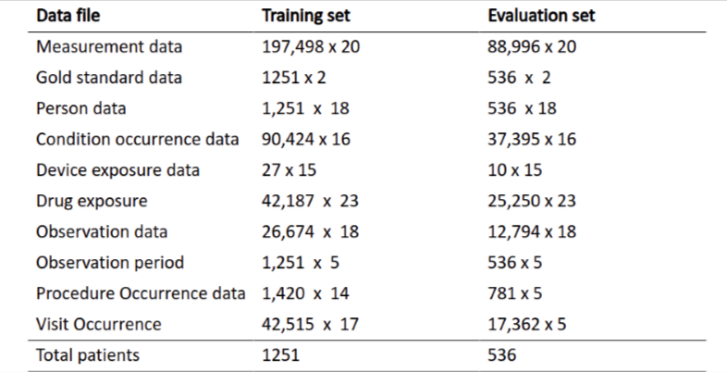


Figure 2A. Distribution and size of data from multiple tables in the competition dataset.

The data was iteratively processed through data preprocessing pipeline for feature engineering, selection and further input to machine learning models for analysis of results, and finally the dataset that had optimal performance from metrics, had features selected from the article for COVID-19 diagnostic aid system []. In total, 38 features of clinical and demographic importance were selected, and data was extracted based on temporal relevance of these features. For the categorical data such as gender, race, etc., we performed one-hot encoding for categorical encoding of these features into numerical values. The training data was split into 80:20 ratio for train and validation sets. Imputation was performed via multiple methods with KNN impute resulting in better applicable to this dataset and was further used downstream pipeline. Z-score normalization of the data was performed for feature scaling before being input to the models. Z-score normalization is a strategy of normalizing the datasets to prevent the problem of outliers and varying range of values of features. The formula for Z-score normalization is: Z = (x – μ) / σ, where μ is the mean value of the feature and σ is the standard deviation of the respective feature. Figure 2B shows the distribution of data before preprocessing and with no extreme values found. Figure 2C, shows data missingness for the selected features whereas figure 2D, shows the data distribution after imputation with KNN method with few features having distribution changed completely such as IL-6 (interleukin) values.

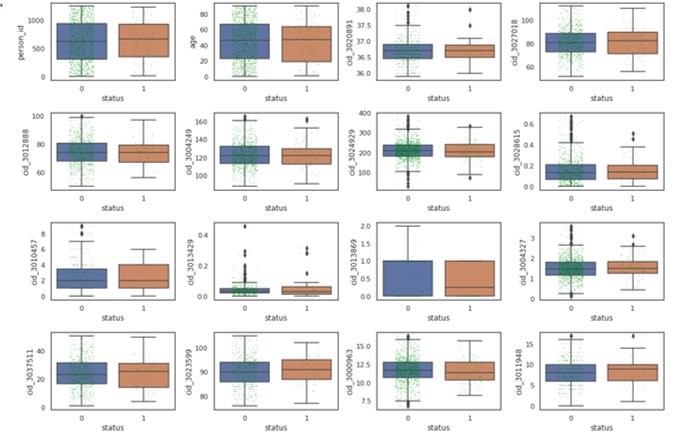


Figure 2B. Boxplot for the selected features to understand data distribution prior to pre-processing.

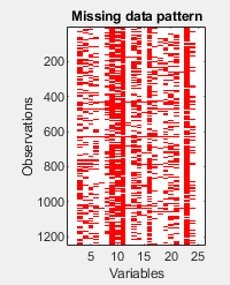


Figure 2C. Data missingness from the training data shown before imputation.

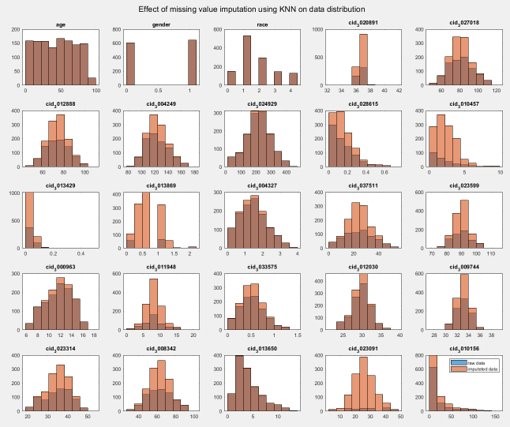


Figure 2D. Data distribution after imputation using KNN method. Some of the features such as IL-6, Basophils ratio, show variation in distribution upon imputation, partially due to very high data missingness.

MRMR and tree-based methods were used for feature importance computation as seen in figure 2E and 2F. The figure 2G shows ML based methods for assessing data imputation at different thresholds of data missingness. Based on this, highest classification accuracy was achieved at 50% missing value threshold and based on this and results from feature importance plots, certain features were dropped from the final input dataset such as IL-6, Basophils ratio, etc. Figure 2H shows the correlation plot for the final selected features, which provides confidence in the dataset considered for the modeling pipelines.

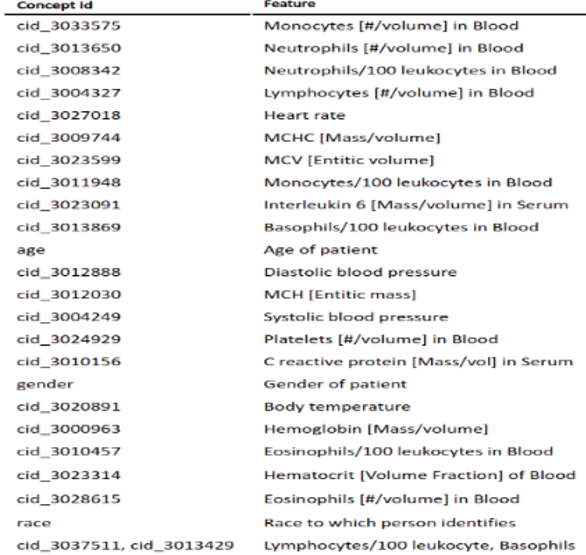


Figure 2E. Feature importance ranks of concept ids using MRMR method.

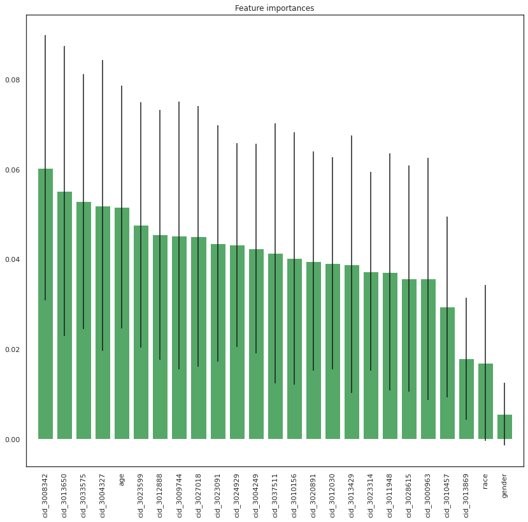


Figure 2F. Feature importance and generated weights based on tree method.

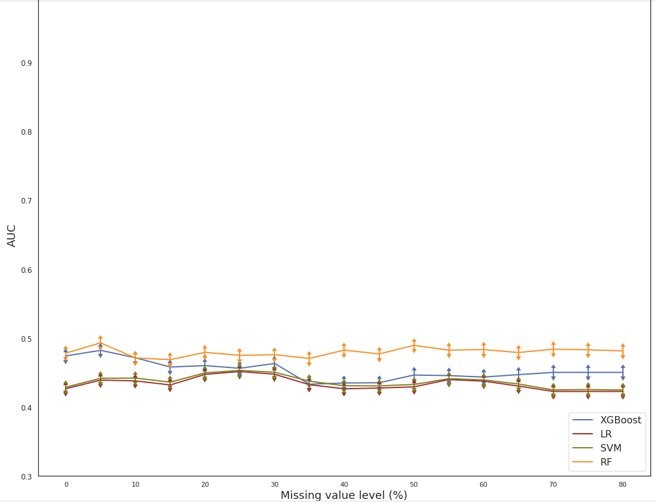


Figure 2G. Upon imputation at different thresholds of missing value, found that highest classification AUCs were achieved at 50% missing value imputation. Additional, column of Body temperature was added manually due small margin of missingness.

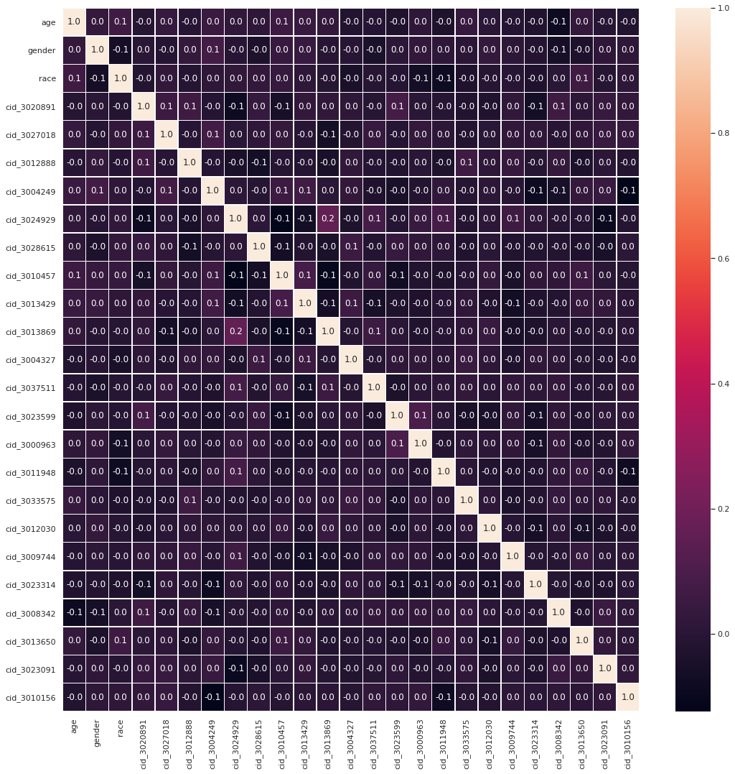


Figure 2H. Correlation matrix plot. None of the selected and processed features were found to be correlated.

**Informatics System Workflow:**

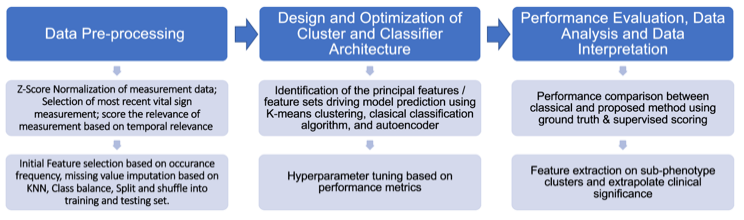


Figure 3. Systematic Workflow

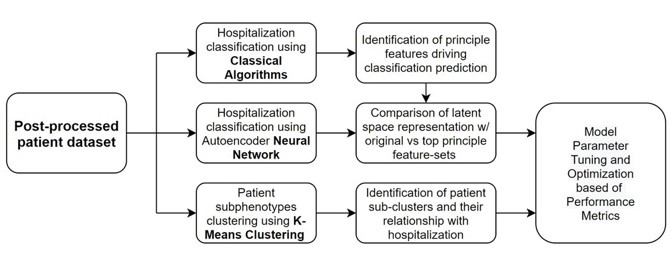


Figure 3A. System workflow of clustering and classification post data pre-processing

The informatics system we implemented is composed of three major phases shown in Figure 3. The first phase is data-processing, as described form above. Followed by the generation of post-processed patient dataset, three parallel pipelines are implemented (Figure 3A.). The first one is performing classification using classical algorithms, including KNN, naïve bayes, SVM which the Matlab has built-in functions. Doing so potentially provides on further identification of principle features driving the prediction, which we can use for the other two pipelines. For the second pipeline, classification is performed based on autoencoder neural network structures with the ability to generate latent space allow us to do comparison between the original and top principal feature sets. The third pipeline is performing patient sub-clustering using K-means method, doing so allow us to identify patient sub-clusters and explore the relationship between the clusters and hospitalization. Upon successfully construction and initial testing on the cluster and classifier architecture, hyperparameter tuning based on performance metrics are used for further performance optimization. For evaluation, performance comparison between classical and proposed method using ground truth & supervised scoring are used, and further feature extraction on sub-phenotype clusters and classification scoring are extrapolated for understanding their clinical significances. Further details on each module of the system will be explained in more detail.

**Method for Autoencoder**

For machine learning algorithm on the classification of patient hospitalization, we implemented our framework using a stack of autoencoders that are independently trained layer by layer but same structure and functionalities. Briefly, an autoencoder takes an input  and first transform it via an encoder to a hidden representation through a deterministic mapping of: 

Parameterized by , where s (.) in a non-linear transformation functions that act as activation function to the layer, **W** is a weight coefficient matrix and **b** is a bias vector. The latent representation y is then mapped back through the decoder stage to a reconstructed vector **Z** with same dimensions as input x, through:  with transformed weight coefficient matrix and bias vector. The goal of autoencoder is that the code **y** is a distributed representation that captures the main factors of variation in the data, thus the major clinical features. When training the model, the algorithm searches the parameters that minimize the difference between x and z. Where θ and θ′ are optimized over the training dataset to minimize the average reconstruction error: 

Autoencoder is implemented in our system for hospitalization classification due to its ability to further validate on the effect of dimensionality reduction of patient features on classification accuracy, where both features from initial missingness based feature selection and features selected from literature survey are implemented to the autoencoder architecture to perform classification. Also, the generated latent space create room for further feature extrapolation that may not be possible with conventional PCA approaches. The classifier attributed by the autoencoder also provides a non-binary output (0-1) that can assist clinical decision more than direct hospitalization classification.

For the selection of model parameters, several elements are considered, including # of hidden layers (1-50), training function (i.e. Scaled Conjugate Gradient (SCG), damped least-squares (DLS), Resilient Backpropagation, One Step Secant (OSS)), performance function (i.e. Mean Square Error, Cross-Entropy, Sum Absolute Error, MESREG), and training epochs (1-20, or performance function optimized). The performance analysis for the classification result used to evaluate the performance of the autoencoder designs include common metrics: Confusion matrix, accuracy, specificity, sensitivity, F-1 score, area under receiver operating characteristic curve (AUC ROC), and area under precision-recall curve (AUC PR). For optimizing parameter selection of the autoencoder, only AUC ROC value is selected. Briefly, the autoencoder model are built with the grid selected parameters, from which 10 models are built and evaluate with AUC ROC to receive an average AUC ROC score for comparative purposes. For the final selection of MATLAB-based autoencoder architecture, we adopted number of hidden layers to be 20, training function to be SCG, performance function to be Cross-Entropy and epochs to be cross-entropy based evaluation.

Two distinct architectures of deep neural networks (DNNs) were designed and further optimized based on input parameters, with the method described above with 20 hidden layers resulted in optimal performance on the evaluation dataset. The autoencoder DNN with 9 layer architecture resulted in final AUC score of 0.630, whereas the 20 layer autoencoder DNN with parameters tuning and optimization resulted in higher AUC score of 0.7287, and was considered for the classification results analysis and interpretation.

**Method for k-means clustering**

K-means clustering method was employed to find unique clusters from COVID-19 patients. k-means clustering is a method that partition data into k cluster in which each observation belongs to the cluster with the nearest mean. The method attempts to minimize within-cluster variances. The value k was optimized with Elbow method, which plot explained variation as a function of the number of clusters. We selected k value where it explains more than 80% of variance. Silhouette plot was generated to evaluate performance of k-means clustering with k selected from Elbow method.

We 2D visualized the resultant clusters using t-Distributed Stochastic Neighbor Embedding (t-SNE) method. In order to better visualize hospitalized and non-hospitalized patients, hospitalization status was added into the dataset to make clear boundary. While finding clusters from k-means clustering method, we did not include the hospitalization status but plotted on the same t-SNE plot to locate unique clusters found only from hospitalized patients.

There were three datasets used for k-means clustering method. First one is a dataset right after data preprocessing without any feature selection. Second dataset includes features selected from principal component analysis (PCA). PC with variance that explains 90% above were selected. Total 8 PC were selected. PC1 explained more than 85% of variance. The eigenvalues for PC1 were sorted to see which variables account for PC1. Last dataset includes features selected from feature importance plot based on tree method.

# Result

**Result for k-means clustering**

A picture containing graphical user interface

Description automatically generatedA picture containing graphical user interface

Description automatically generated

Figure 4. Unique clusters were identified from K-means clustering. For figure (A), (B), and (C), left figures show non-hospitalized (0) and hospitalized patients (1). Right figures show clusters with k-means clustering method drawn on same t-SNE plot. (A) There were 12 clusters identified from the dataset without any feature selection. All the clusters shared both hospitalized and non-hospitalized patients. (B) There were 13 clusters identified from the dataset with PCA feature selection. Cluster 4 and 13 only contained hospitalized patients. (C) There were 5 clusters identified from the dataset with features selected by feature importance plot generated with tree method. (D) Performance of clustering was indicated by Silhouette plot for each dataset.

K-means clustering with datasets with PCA feature selection and features selected from feature importance plot generated unique clusters that were found only from hospitalized patients (Fig. 4B and 4C). Both methods had steep drop of Silhouette value indicating the datasets are not suited for clustering. However, performance was better for features selected from feature importance plot compared to PCA feature selection (Fig. 4D).

**Result for Classification using autoencoder**

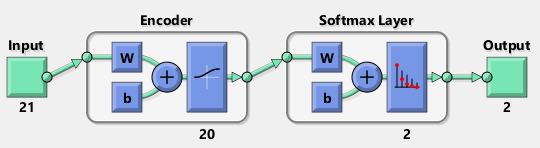


Figure 5A. Finalized autoencoder model.

The finalized autoencoder architecture used for hospitalization classification is demonstrate in Figure 5A. Subsequent feeding of the model into the Neural Pattern Recognition Tool from MATLAB indicate high training, validation, and test ROC across the training stage demonstrated by Figure 5B. Reaching an AUC ROC of 0.734 and AU PR of 0.720. The sensitivity and specificity of the classification are 0.6 and 0.71 respectively, meaning that this classification algorithm would provide 30% to 40% false negative and/or false positive rate.

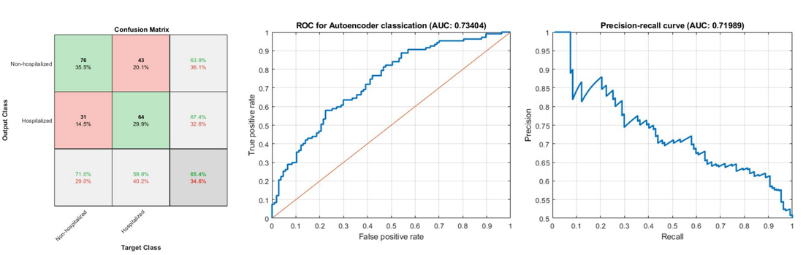


Figure 5B. Classification result of autoencoder on Training set

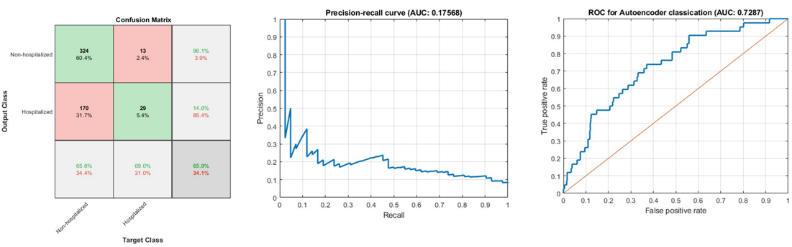


Figure 5B. Classification result on autoencoder on External Validation set

Moving onto the external validation set shown in Figure 5C., which has high imbalance of class size, the classification model still retains a similar AUC ROC of 0.7287 and a drastically low AUC PR of 0.17568, which is resultant from misclassifying non-hospitalized patient into hospitalized class. In terms of clinical decision, such misclassification wouldn’t affect non-hospitalized patient at an individual standpoint since being classified as requiring hospitalization when one does not need to be hospitalized is less severe than vice versa, yet at in terms of pandemic standpoint such misclassification would result in over-saturated number of patient hospitalization that would impair the healthcare system. Therefore, such misclassification error should also be taken to great consideration in the cases of COVID-19 particularly. For specificity and sensitivity on the EV set, the model receives a score of 0.69 and 0.656 respectively, which are also similar with that of the training set, indicating that our classification model wasn’t affected by data overfitting as similar scores are retained on EV set whose data are not fed into the model during construction. To further demonstrate the predictability of our model, we re-iterate the autoencoder architecture using training set as input and feed the resultant model with external validation set for generate AUC ROC on both set, the resultant CV-EV AUC ROC plot (Figure 5C.) indicated that while the model has a dynamic range of AUC ROC, a positive correlationis observed between the CV and EV set, demonstrating certain degree predictability of our classification model.

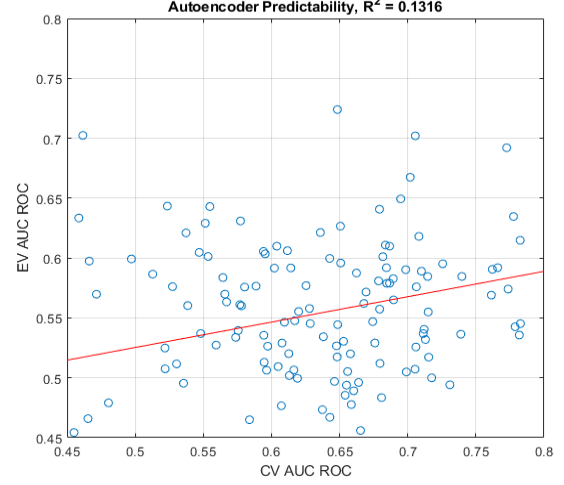


Figure 5C. Predictability of autoencoder based on AUC ROC of CV versus EV.

# Discussion

**Interpretation of K-means clustering**

From the unique clusters identified from PCA feature selection and features selected from feature importance plot, we have found same hospitalized patients. We compared which features are overlapping between two datasets. Three major groups were identified. First group includes features related to immune cell number, which are Monocytes, neutrophils, and lymphocytes. Second group includes RBC features, which are platelets, hemoglobin, MCHC, MCH, and hematocrit. Last group includes features related to myocyte health, which are heart rate, systolic blood pressure, and diastolic blood pressure.

Previous study indicates monocytes, neutrophils, and lymphocytes as key determinant of COVID-19 disease presentation and severity [6]. Recent study also indicates close correlation of RBC and sever or hospitalized COVID-19 patients [7]. It is also well-known that COVID-19 has effect on cardiovascular health [8].

**Interpretation of autoencoder classification results**

Based on the classification results from the 20-layer autoencoder DNN, which achieved an AUC of 0.7287 on the evaluation dataset, we found few interesting results. The results were separated into hospitalized and non-hospitalized patients from a total of 536 patients from the evaluation dataset. The median age of patients was 45 years, almost balanced number of male and female samples. The vital signs were found to be very higher in the hospitalized patients with 88.66 (n/minute) heart rate compared to 82 in non-hospitalized. The systolic and diastolic pressures were slightly higher as 122.83 and 75.83 (mmHg) in the hospitalized patients, and highest body temperature of 38.3 (degree Celsius) in that category. We also, found that the most common blood routine values were lower at 11.3 (g/L) of Hemoglobin, 201.5 (x 10^9 /L) platelet count, a higher lymphocyte count of 1.73 (x 10^9 /L). This also, provides insights into the various measurement data available from clinical records that can be used to identify whether the patient needs to hospitalized upon the first visit after COVID-19 RT-PCR positive test, and enable clinical decision making for prioritizing patients. General numerical value of likelihood of hospitalization, which can assist clinical decision more than binary output, can developed on real world datasets. Figure 6A, shows the final features which were ranked based on their feature importance from the MRMR (minimum Redundance Maximum Relevance) method.

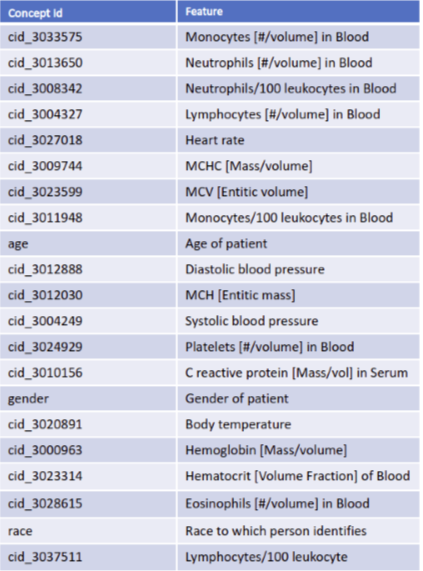


Figure 6A. Top features selected for making classification predictions of hospitalized vs non-hospitalized patients.

**Comparative discussion and weakness**

The DNN method developed provides a novel way to utilize selected reduced clinical features and achieve an AUC of 0.7287 and AUPR of 0.1757 in the evaluation dataset. Though the highest scores available on the current ranking board provide an AUC of 0.8103 and AUPR of 0.2154, which the competition organizers have publicly provided. Though the competition is still ongoing, the authors have not released the methods developed yet for doing strict comparison with those methods on this dataset. Though, the current understanding suggests that features from the complete dataset including all 10 tables needs to be understood evaluatively and further methods can be developed to incorporate more clinical features of importance based on further analysis and guidance from clinical experts, utilizing the knowledge based component into DNN methods development. 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.

The k-means clustering method successfully identified unique clusters and features that represent hospitalized patients. However, the synthetic data from COVID-19 DREAM Challenge was designed for classification problem and may not be suited for clustering or sub-phenotyping, which probably is a reason for our Silhouette plot not being able to find optimal cluster number for clustering. In addition, combined measurement and categorical/frequency dataset is not suitable for K-means clustering method.

**Conclusion and Future work**

Based on the classification and clustering results, we found some interesting results which have been briefly discussed prior in previous sections. While working on this project, we realized that working on synthetic datasets helps in developing and prototyping new methods but eventually deriving meaning from these datasets for relevance to clinical decision making, should be avoided due to synthetic nature of the datasets. We were able to achieve high scores of AUC on evaluation dataset, given low clinical features incorporated. We are able to compute general numerical values of likelihood of hospitalization, which can assist clinical decision more than binary output, though the current values are of synthetic nature. We will continue working on improvements of our methods, to implement better hyperparameter tuning techniques (i.e. Bayesian optimization). Also, scoring the quality of measurement data based on temporal relevance might be another factor to look into.

For clustering, we are going to use latent space from autoencoder for better clustering, apply k-mode for mixed dataset which include both categorical and measurement data, and lastly implement UMAP for better visualization quality.

##### Acknowledgment

The group acknowledges Dr. May D Wang, Professor, Kavli Fellow, GCC Distinguished Cancer Scholar at Biomedical Engineering department of Georgia Institute of Technology for her consistent mentoring, evaluation and project support. The PhD student mentors from the BMED 8813 course at Biomedical department, Felipe Oliveira Giuste and Wenqi Shi for their efforts into organizing team, sessions for weekly meetups and critical evaluation of results based on weekly developments helped in further development of this project.

##### References

1. Landi, I., Glicksberg, B. S., Lee, H. C., Cherng, S., Landi, G., Danieletto, M., ... & Miotto, R. (2020). Deep representation learning of electronic health records to unlock patient stratification at scale. NPJ digital medicine, 3(1), 1-11.
2. Nagamine, T., Gillette, B., Pakhomov, A., Kahoun, J., Mayer, H., Burghaus, R., ... & Saxena, M. (2020). Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. Scientific reports, 10(1), 1-13.
3. Aguiar, H., Santos, M., Watkinson, P., & Zhu, T. (2020). Phenotyping Clusters of Patient Trajectories suffering from Chronic Complex Disease. arXiv preprint arXiv:2011.08356
4. da Silva, J. F., Hernandez-Romieu, A. C., Browning, S. D., Bruce, B. B., Natarajan, P., Morris, S. B., ... & Wong, K. K. (2020, December). COVID-19 clinical phenotypes: presentation and temporal progression of disease in a cohort of hospitalized adults in Georgia, United States. In Open Forum Infectious Diseases.
5. Renoux, C., Fort, R., Nader, E., Boisson, C., Joly, P., Stauffer, E., ... & Connes, P. Impact of COVID-19 on red blood cell rheology. British journal of haematology.
6. Nishiga, M., Wang, D. W., Han, Y., Lewis, D. B., & Wu, J. C. (2020). COVID-19 and cardiovascular disease: from basic mechanisms to clinical perspectives. Nature Reviews Cardiology, 17(9), 543-558.
7. Brodin, P. (2021). Immune determinants of COVID-19 disease presentation and severity. Nature Medicine, 27(1), 28-33.
8. Feng, C., Huang, Z., Wang, L., Chen, X., Zhai, Y., Zhu, F., ... & Li, T. (2020). A novel triage tool of artificial intelligence assisted diagnosis aid system for suspected COVID-19 pneumonia in fever clinics. MedRxiv.

Table 1. Summary of strength/weakness of existing methods/solutions

|  |  |  |  |
| --- | --- | --- | --- |
| **Title of Paper** | **Methods / Solutions** | **Strengths** | **Weakness** |
| Deep representation learning of electronic health records to unlock patient stratification at scale [1] | 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. |
| Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data [2] | **K-means clustering** | Unsupervised methodology of high dimensional sub clustering 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. |
| Phenotyping Clusters of Patient Trajectories suffering from Chronic Complex Disease. [3] | 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 |
| Vital signs assessed in initial clinical encounters predict COVID-19 mortality in an NYC hospital system. [4] | **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 |
| A Novel Triage Tool of Artificial Intelligence Assisted Diagnosis Aid System for Suspected COVID-19 pneumonia In Fever Clinics [5] | 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 |