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

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

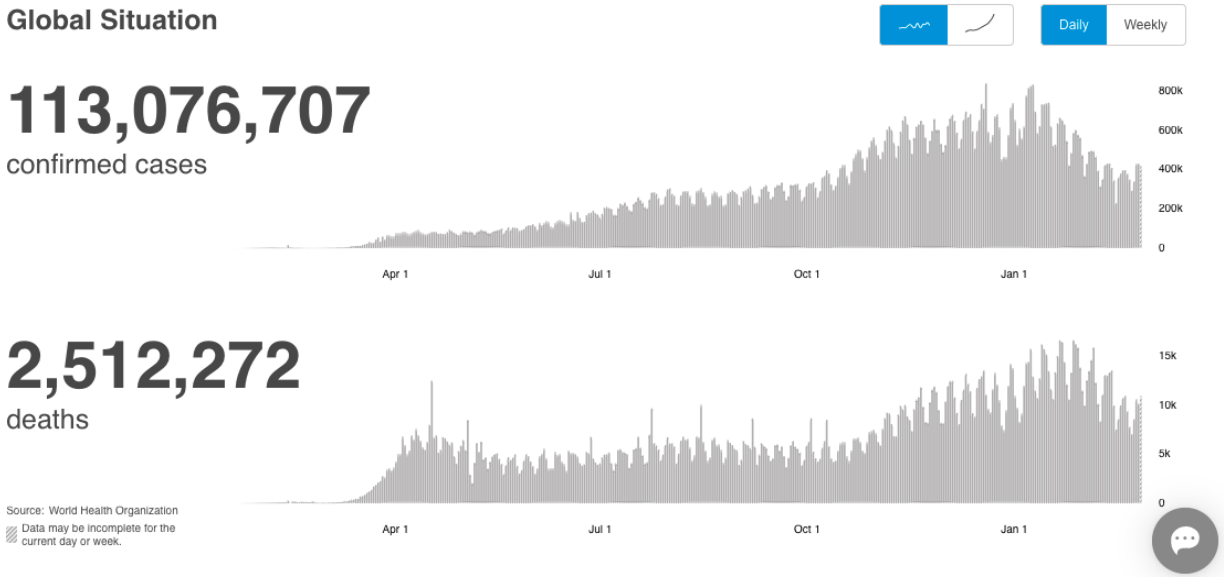
BMED 8813 BHI Presenter: **G-6**

Seonggeon Cho, Rohan Bhukar, Bryce Butler, Zhonghao Dai




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

- COVID-19/Sars-CoV-2 led millions of deaths worldwide
- More than 12,000 mutations reported
- Patients experienced variety of symptoms based on their 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

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

Summary table of literature critique

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-mean 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. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records	Stack of denoising autoencoder	• The method successfully diagnosed disease 1 year time interval with accuracy ranging around 0.84-0.907.	• At least 10 medical record is needed for reasonable prediction. • It only showed 1 year disease prediction. • Predictive power of some of the diseases was low, and it is not guaranteed that health record provided from Dr. Bhavani is enough or suits for predicting patient response to Covid-19.
4. COVID-19 Diagnosis via DenseNet and Optimization of Transfer Learning Setting	Method combining DenseNet and optimization of transfer learning setting (OTLS) strategy	• This method proved to be quite accurate and highlights ways for improvement with further research.	• They did not validate the optimal combination configuration of DA techniques. They did not validate the optimal values of frozen layer
5. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis	SOM algorithm for clustering, Extreme Gradient Boosting	• Was able to demonstrate the use of machine learning models to predict COVID-19 presence using only commonly recorded clinical variables	• Requires large scale clinical data.
6. Affinity network fusion and semi-supervised learning for cancer patient clustering	Affinity Network Fusion; k-Nearest-Neighbor(kNN)-based transformations	• A semi-supervised learning model combining ANF and neural network, which achieved very good results for few-shot learning (e.g., being able to achieve 97% accuracy on test set with training less than 1% of data for classifying patients into correct disease types).	• Only reported experimental results on four cancer types with known disease types. Not known how it will do with multiple types. • Did not how it will work with other types of data sets.

Summary table of literature critique

Title of Paper	Methods / Solutions	Strengths	Weakness
7. 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
8. Machine Learning Models for Analysis of Vital Signs Dynamics: A Case for Sepsis Onset Prediction.	SVM, Feature extraction based on data statistics	• Small number of extracted features allow for comparatively accurate disease onset prediction	• The feature extraction is semi-supervised since it is based on hypothesis that vital sign variability leads to higher onset prediction, not applicable to all disease prediction
9. Unsupervised Machine learning to subtype Sepsis-Associated Acute Kidney Injury.	KNN, K-means clustering	• Phenotyping can be done with relatively small time-series time, dataset from different collection setting can be harmonized into single set for analysis	• Number of clusters extracted is low (K=2) and could not provide more data on clinical characteristics of each group
10. 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
11. Patient Subtyping via Time-Aware LSTM Networks	**Time-Aware LSTM** (Long-Short Term Memory)	• T-LSTM model can handle irregular elapsed times between successive elements of sequential data; • T-LSTM auto-encoder learns powerful representations, uses these to cluster patient populations.	• Since they used synthetic datasets, number of main clusters reported is very low (k=2). • Model yielded only very few features with p-values < 0.05, most of them grouped into 1 cluster.
12. COVID-19 Clinical Phenotypes: Presentation and Temporal Progression of Disease in a Cohort of Hospitalized Adults in Georgia, United States	Gower's dissimilarity matrix, medoids algorithm, Wilcoxon rank-sum tests to evaluate differences	• Novel method to identify clinical phenotypes of COVID, clustering analysis probes complex interactions between patient features and clinical course; • High level of data completeness and quality	• Relatively small sample size cohort (n=305), results are not generalizable to other populations • Interpretation of lab values and vital signs over time was affected by censoring
13. Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study	LSTM (Long-Short Term Memory), **Dynamic Time Warping** (DTW), t-Distributed Stochastic Neighbor Embedding (t-SNE)	• LSTM is used to standardize and densify patient records, while DTW is leveraged to calculate patient similarities from which subtypes are identified; • Better disease subtyping with actionable clinical features	• Approach is completely data-driven without utilising domain knowledge • Interpreting LSTM procedure not straightforward • Study based on only 1 cohort (PPMI).

Summary table of literature critique

Title of Paper	Methods	Strengths	Weakness
14. Learning representations for the early detection of sepsis with deep neural networks	Multilayer perceptrons (MLPs), LSTM (Long-Short Term Memory) networks	• DNN models can improve performance without extraction of features using domain knowledge in raw data; • Using SepLSTM, performance can be improved.	• Model is based upon small number of patients with limited generalizability (n=360 cases) • Patients who were diagnosed with sepsis, but no continuous signs were excluded from analysis.
15. Novel Temperature Trajectory Sub phenotypes in COVID-19	Group – based trajectory modeling (**GBTM**)	• Using phenotypic potential of temperature trajectories novel sub phenotypes in COVID-19 were identified.	• Study was retrospective and conducted at 1 center. • Only temperature trajectories as a clinical factor were considered in the model. Single modality

Dr. SIVA

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

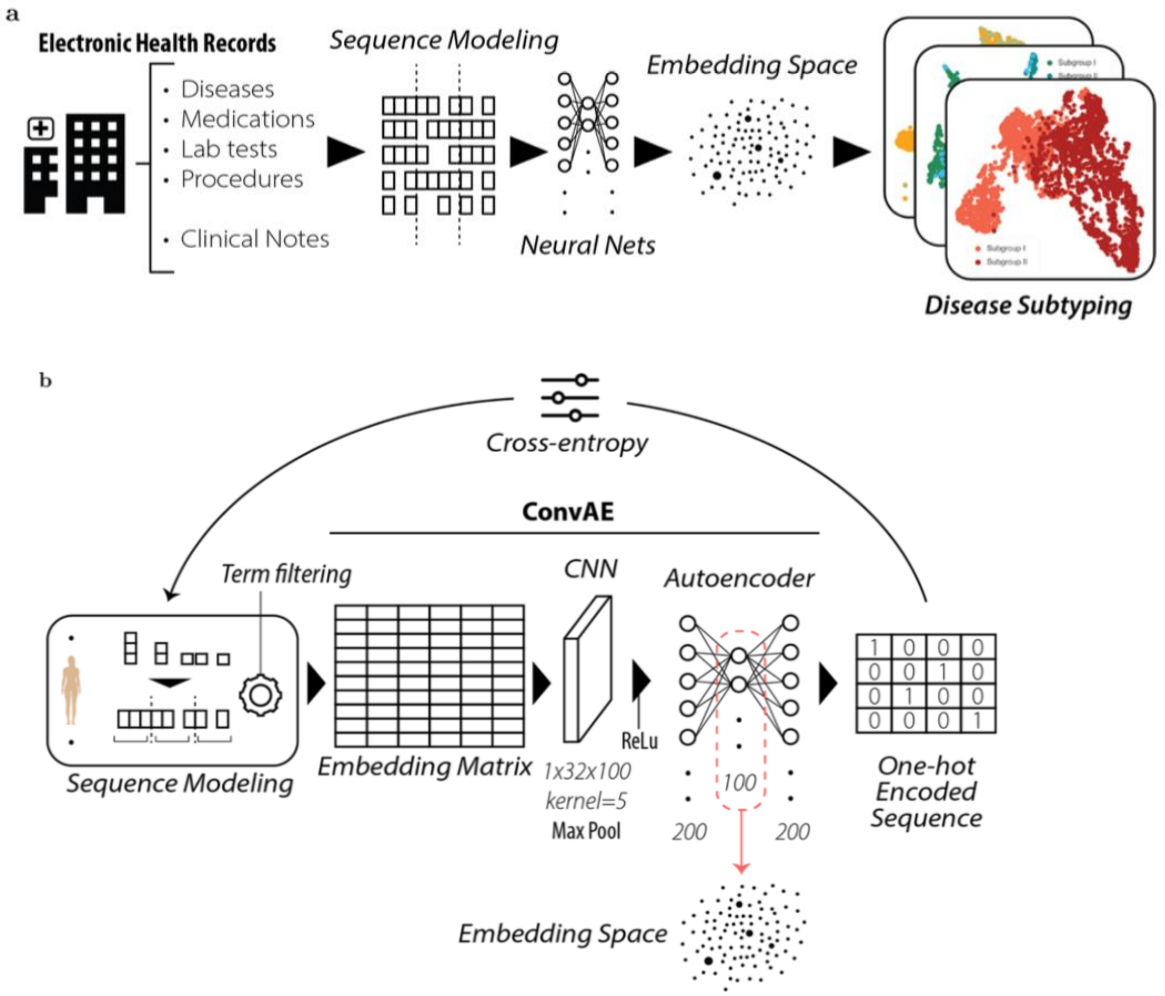
- ***Solutions arising from classical machine learning models*** (gradient boosting, logistic regression, etc.) exist in the case of COVID-19 sub phenotyping ***utilizing only single cohort data***, and ***basic clinical presentation*** factors.
- Either the ***existing methods*** for COVID-19 phenotyping ***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***.

Subtyping disease with electronic health record

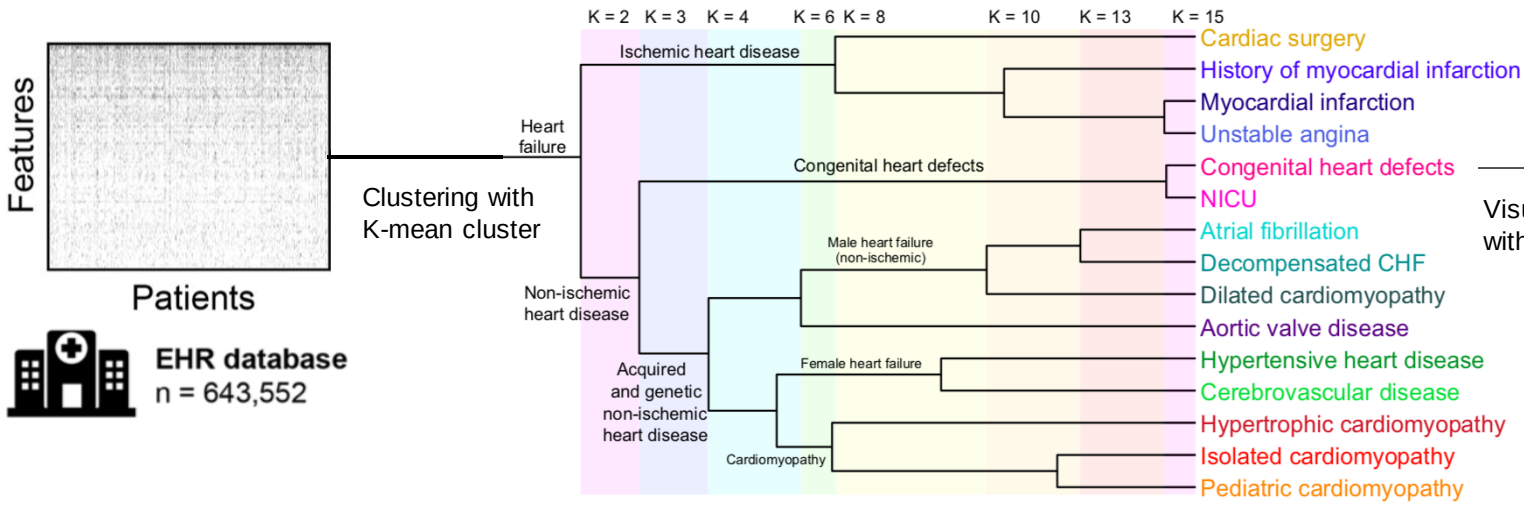
Title	Deep representation learning of electronic health records to unlock patient stratification at scale
Who	Isotta Landi, Benjamin S. Glicksberg, Hao-Chih Lee, Sarah Cherng, Giulia Landi, Matteo Danieletto, Joel T. Dudley, **Cesare Furlanello** and **Riccardo Miotto** (Bruno Kessler Institute, Povo, TN, Italy & Icahn School of Medicine at Mount Sinai, New York, NY, USA)
What	Subtyping within disease-specific population via clustering
When	Published in July 2020 (1980-2016 EHR)
Where	EHR from Mount Sinai Health System
Why	The specific properties of the different subgroups can potentially inform personalized treatments and improve patient care.
How	Convolutional Neural Network, Autoencoder



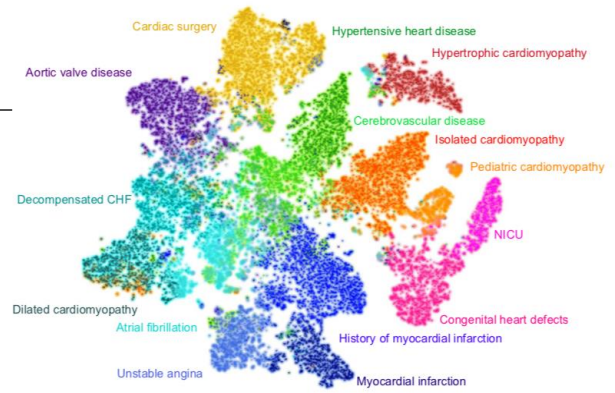
Landi *et al.*, *npj Digital Medicine*, 2020

Subtyping disease with electronic health record

Title	**Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data**
Who	Tasha Nagamine, Brian Gillette, Alexey Pakhomov, John Kahoun, Hannah Mayer, Rolf Burghaus, Jörg Lippert & **Mayur Saxena (**Droice Research, New York, NY, USA **)**
What	Subphenotype heart failure and characterize patients associated with the identified subphenotypes
When	Published in December 2020. (2008-2018 EHR)
Where	EHR from national medical research center located in a major metropolitan center in western Russia
Why	Classification schemes for heart failure help clinicians determine the disease phenotype, select appropriate treatments, and define study populations for randomized controlled trials (RCT) of heart failure interventions.
How	K-mean clustering



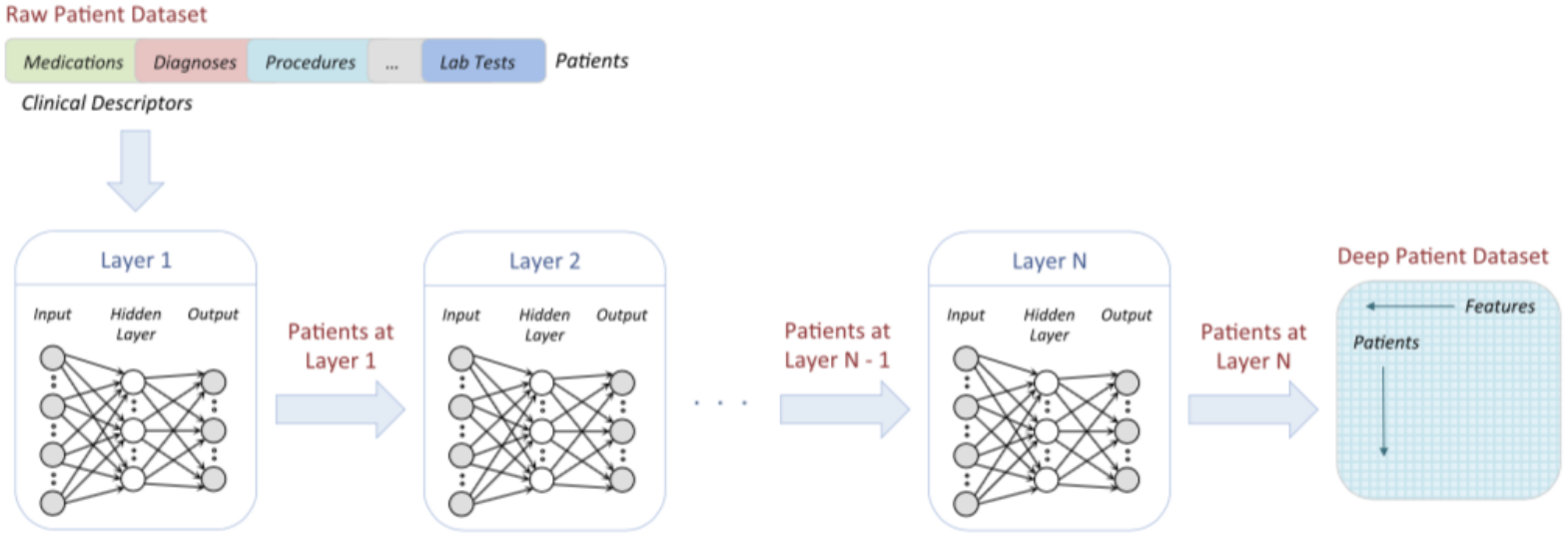
Visualization with t-SNE



Nagamine *et al.*, *Scientific Report*, 2020

Predicting disease outcome with electronic health record

Title	**Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records**
Who	Riccardo Miotto, Li Li, Brian A. Kidd, **Joel T. Dudley** (Icahn School of Medicine at Mount Sinai, New York, NY, USA)
What	Future disease prediction
When	Published in May 2016. (1980-2014 EHR)
Where	EHR from Mount Sinai Health System
Why	Precision medicine is required for predicting health status and preventing diseases and disability.
How	Stack of denoising autoencoder



Miotto *et al.*, *Scientific Report*, 2016

Covid-19 Diagnosis with CT scans

Title	**COVID-19 Diagnosis via DenseNet and Optimization of Transfer Learning Setting**
Who	Yu-Dong Zhang et al. from various schools
What	Presenting the make up of a more accurate method to diagnosis COVID-19
When	The article was published 18 January 2021.
Where	Anywhere where chest CT scan images an be made. (conventionally scanned from the lung tip to the costal diaphragm angle)
Why	A smart diagnosis system via computer vision and artificial intelligence can benefit patients, radiologists, and hospitals as manual diagnosis can be tedious and may have errors due to multiple factors.
How	Using a method combining DenseNet and optimization of transfer learning setting (OTLS) strategy

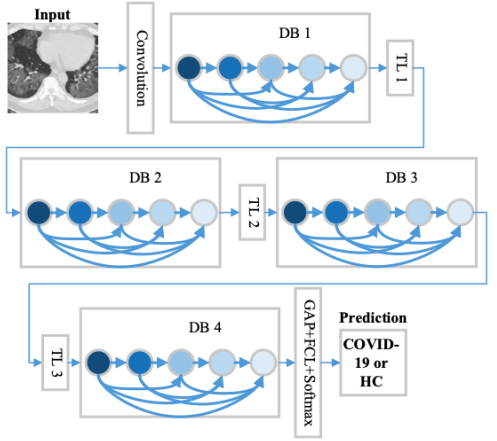
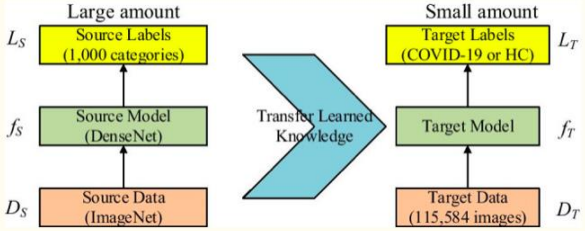


Fig. 6 How DenseNet classify chest CT images (TL transition layer, DB DenseBlock, GAP global average pooling, FCL fully connected layer)



Open in a separate window

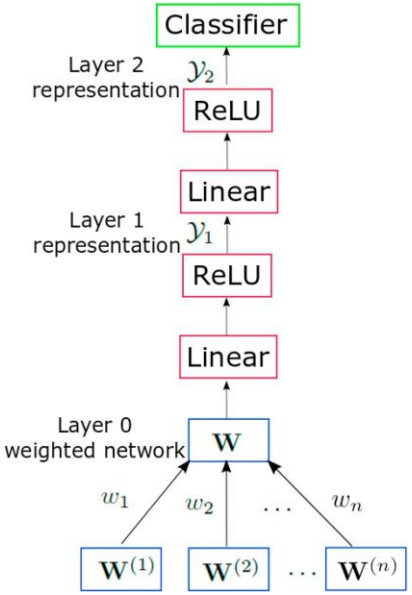
Fig. 4

Idea of transfer learning

Zhang, Y. D., et al. (2021). COVID-19 Diagnosis via DenseNet and Optimization of Transfer Learning Setting. Cognitive computation, 1–17. Advance online publication. https://doi.org/10.1007/s12559-020-09776-8

Patient clustering using ANF approach

Title	Affinity network fusion and semi-supervised learning for cancer patient clustering
Who	Tianle Ma, Aidong Zhang Department of Computer Science and Engineering, University at Buffalo (SUNY)
What	They presented an unsupervised and semi-supervised affinity network fusion that can integrate multi-omics data for cancer patient clustering and subtype discovery
When	The article was published on 26 May 2018, Code and data was pulled from the date range of Sept 2017-Feb 2018
Where	A processed harmonized cancer dataset was downloaded from GDC data portal consisting of 2193 patients during the span of the research. Anywhere where such data can be found
Why	Cancer patients are multifactorial. This results in the need to develop ways to cluster cancer patients of the same cancer type into subgroups; while defining new cancer subtypes with comprehensive molecular signatures associated with distinct clinical features. This helps with proper diagnosis and treatment plans
How	With an Affinity network fusion framework (contains KNN)



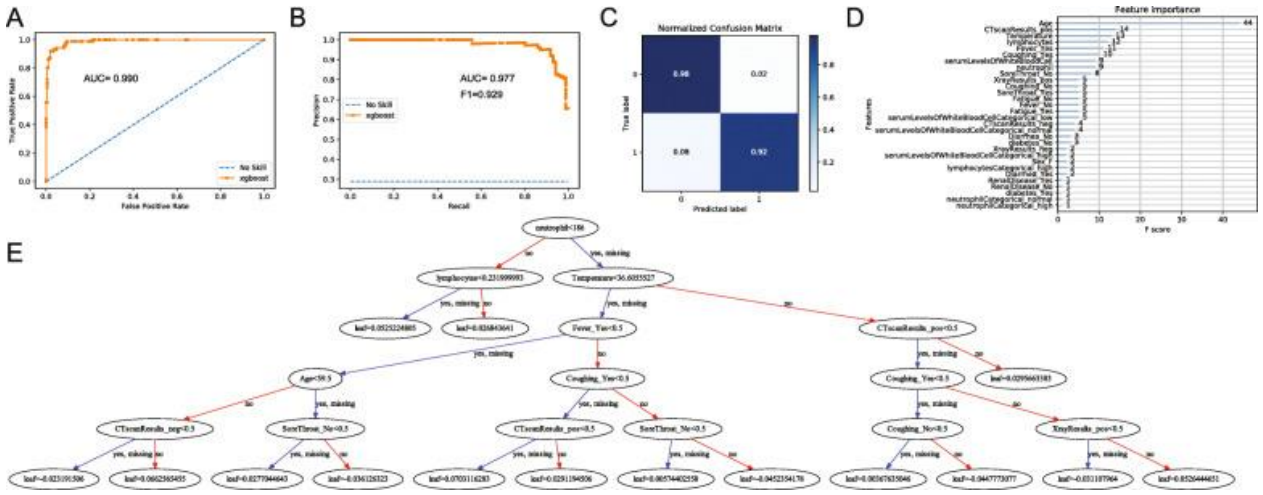
Input : •Patient-feature matrices (n views): $\mathcal{X}^{(v)}, v = 1, 2, \cdots, n$
•Number of clusters: c
•Weight of each view (optional): $\mathbf{w} = (w_1, \cdots, w_n)$
•Other optional parameters
Output: •Patient cluster assignment $\mathcal{A}$
•Fused patient affinity matrix $\mathbf{W}$
•Patient affinity matrices from each view, $\mathbf{W}^{(v)}, v = 1, \cdots, n$

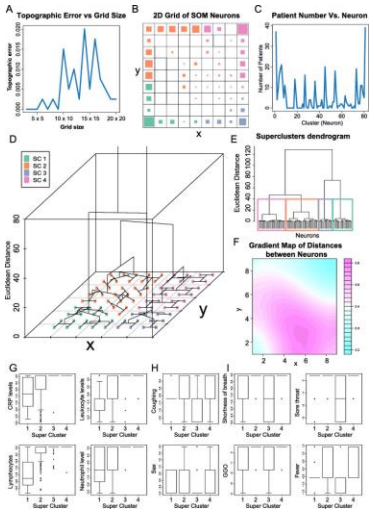
begin
  Feature selection and transformation
  $\mathcal{X}^{(v)} \to \mathbf{X}^{(v)} \in \mathbb{R}^{N \times p_v}, v = 1, 2, \cdots, n$
  Calculate pair-wise distance matrix for each view:
  $\mathbf{\Delta}^{(v)} \in \mathbb{R}_{+}^{N \times N}, v = 1, 2, \cdots, n$
  Calculate kNN affinity matrix for each view:
  $\mathbf{W}^{(v)}, v = 1, 2, \cdots, n$ (Eq. 8 or Eq. 10)
  Calculate fused affinity matrix W (Eq. 6 or Eq. 7)
  Spectral clustering on fused affinity matrix W: $(\mathbf{W}, c) \to \mathcal{A}$
  Return $\mathcal{A}, \mathbf{W}, \mathbf{W}^{(v)}, v = 1, 2, \cdots, n$
end

Ma, T., & Zhang, A. (2018). Affinity network fusion and semi-supervised learning for cancer patient clustering. Methods, 145, 16-24.

Predicting prognosis with available clinical data

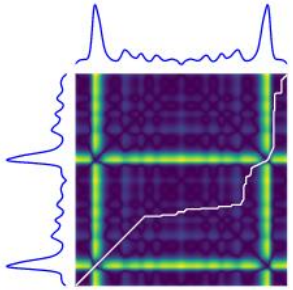
Title	Predicting Prognosis in COVID-19 Patients using Machine Learning and Readily Available Clinical Data
Who	Wei Tse Li et al from various schools
What	They proposed a more accurate diagnosis model of Covid-19; it uses the patient's symptoms and conventional test results, which they analyzed this data through machine learning
When	The paper was published in Sep 29, 2020, with data research ranging from Jan 17, to March 23, 2020
Where	Applicable to places that have patient health records including symptoms and routine test results
Why	More accurate diagnosis models can help overcome the stresses of Covid-19 on health care systems worldwide
How	This is done through machine learning, utilizing tools such as SOM algorithm for clustering, Extreme Gradient Boosting

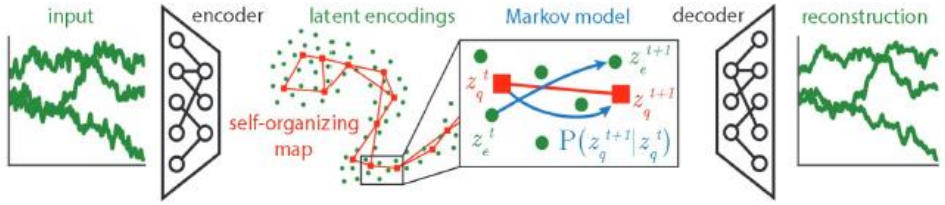




Campbell, Thomas W., et al. "Predicting Prognosis in COVID-19 Patients using Machine Learning and Readily Available Clinical Data." medRxiv (2021).

Phenotyping Clusters of Patient Trajectories using TSK means and VAE



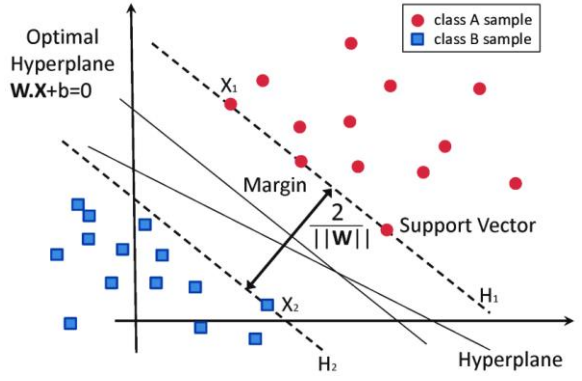


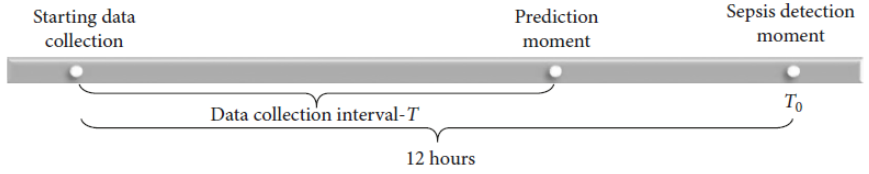
Title	Phenotyping Clusters of Patient Trajectories suffering from Chronic Complex Disease
Who	Henrique Aguiar et al. from Big Data Institute at Oxford University
What	To analyze different clustering methodologies in a hospital dataset of multidimensional, multi-modal vital-sign observations, and evaluated their performance in obtaining well-separated clusters with distinct phenotypic characteristics
When	The article is published in Nov 2020, with datasets range between March 2014 and 31st March 2018. Data contains vital signs observations up to 3 days before a clinical outcome.
Where	Adult patients admitted to four Oxford University Hospitals: the John Radcliffe Hospital, Horton General Hospital, Churchill Hospital, and the Nuffield Orthopaedic Hospital. Applicable to hospitals with time series EHR.
Why	Most clustering methods assume time-invariance of vital-signs and are unable to provide interpretability in clusters that is clinically relevant such event or outcome information.
How	Two types of clustering models were considered: Time Series K means, Variational Autoencoder with proposed loss based on class distribution. Performance of each method was evaluated using supervised scores given true clinical outcome.

Aguiar, Henrique, et al. "Phenotyping Clusters of Patient Trajectories suffering from Chronic Complex Disease." *arXiv preprint arXiv:2011.08356* (2020).

Onset prediction with Vital signs feature extraction

Title	Machine Learning Models for Analysis of Vital Signs Dynamics: A Case for Sepsis Onset Prediction
Who	Eli Bloch et al. from Beilinson Hospital and Tel Aviv University in Israel
What	They extract clinically associated features focusing on variability on patient vital signs and established a ML-based predictive model that can predict the onset of sepsis 4 hours prior to the decision with approximately 80% sensitivity.
When	Article was published in Nov 2019 with patient data collected over the period 2007 – 2014.
Where	electronic medical records (EMRs) of patients admitted to the general intensive care unit (ICU) of the tertiary-level, university-affiliated Rabin Medical Center (RMC), Petah Tikva, Israel
Why	Gaining early insight of sepsis onset can allow timely treatment and lead to better clinical outcome and reduce costs
How	Feature extraction focuses on the hypothesis that unstable patients are more prone to develop sepsis during ICU stay; the features are then dimensionally reduced using SVM with RBF kernel, subsequently ROC classification is used to validate accuracy of prediction.

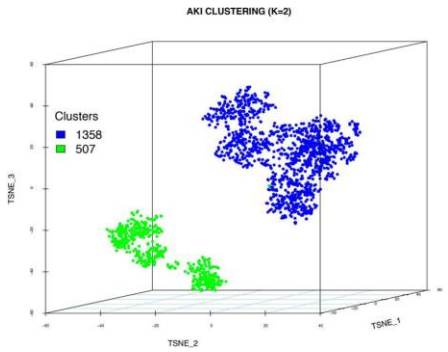


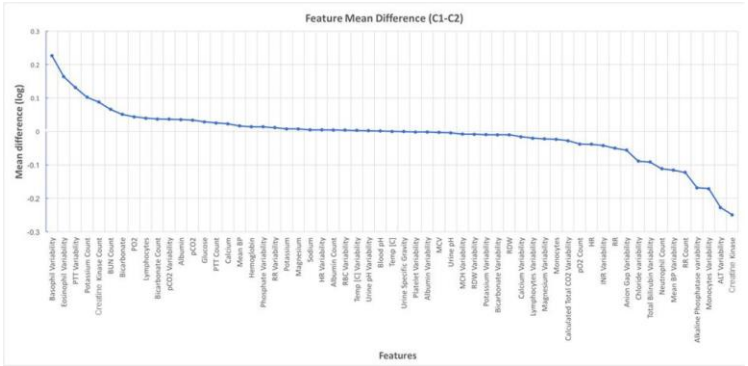


Bloch, Eli, et al. "Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction." *Journal of healthcare engineering* 2019 (2019).

Unsupervised Machine learning to subtype Disease Characteristics

Title	Unsupervised Machine learning to subtype Sepsis-Associated Acute Kidney Injury
Who	Kumardeep Chaudhary et al. From Icahn School of Medicine at Mount Sinai, New York
What	They agnostically identify AKI subphenotypes using machine learning techniques and routinely collected data in electronic health records (EHRs) with finding suggest two distinct clusters of patients from patients within the larger syndrome of sepsis-associated AKI and different mortality rate.
When	The article was published on October 2018 with data collected from patients with sepsis-induced AKI from 2001 to 2012 when they had AKI within 48 hours of ICU admission.
Where	Database contains vital signs of patients from a large, single center tertiary care Beth Israel Deaconess Medical Center in Boston
Why	There has not been any investigation into identifying AKI syndrome comprised of potentially several different subtypes which would potentially results in different clinical characteristics and outcomes
How	Unsupervised clustering using k-means with varying K were performed on feature matrix of combined labs and vitals, then each features were harvested and analyzed for feature importance. The clustering were eventually visualized using T-SNE plot.



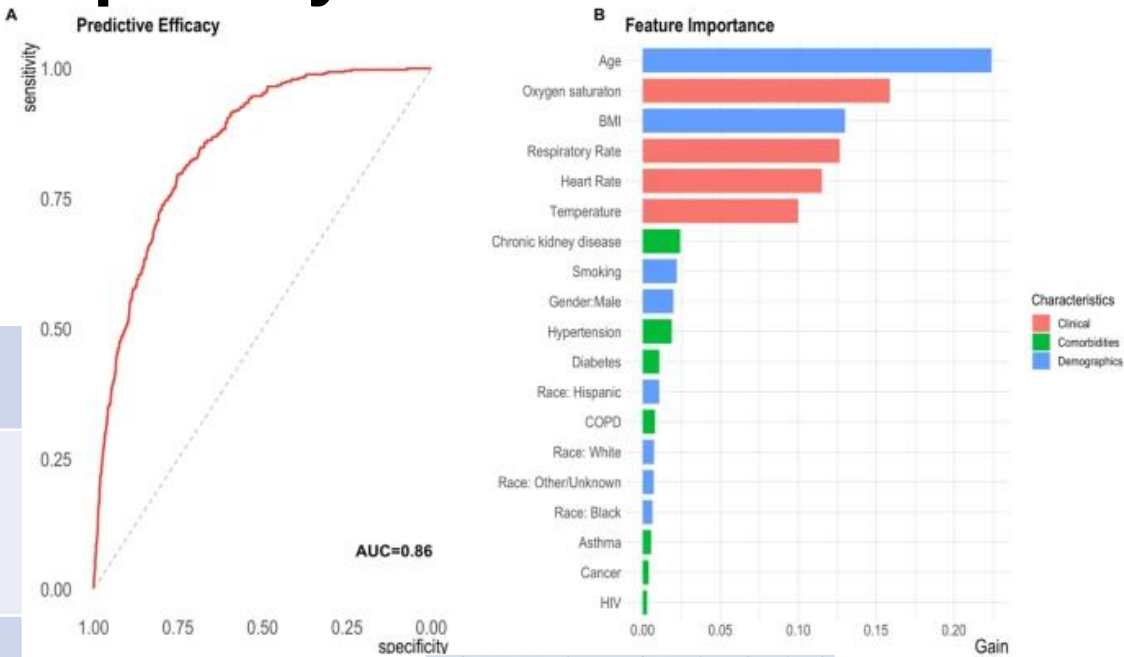


Chaudhary, Kumardeep, et al. "Unsupervised Machine learning to subtype Sepsis-Associated Acute Kidney Injury." *bioRxiv* (2018): 447425.

Vital signs assessed in initial clinical encounters predict COVID-19 mortality in an NYC hospital system

Citation	*Rechtman, E., Curtin, P., Navarro, E. et al. Vital signs assessed in initial clinical encounters predict COVID-19 mortality in an NYC hospital system. Sci Rep **10,** 21545 (2020)*
Who	Elza Rechtman, Paul Curtin, **Esmeralda Navarro**, Sharon Nirenberg & **Megan K. Horton** (Icahn School of Medicine at Mount Sinai, New York, NY, USA)
What	Showed that demographic & clinical factors could be combined in a gradient-boosting machine learning model to create an effective predictor of mortality with an AUC of 0.86. Also show that immediate, objective measures collected at clinical presentation, can be effective predictors of mortality.
When	The article was published on December 2020, with data collected from COVID-19 related encounters at all Mount Sinai Health System facilities (n = 53) in New York City till April 2020.
Where	Deidentified data set of all COVID-19 confirmed cases (n=8770) collected at all Mount Sinai Health System Facilities (n=53) in city of New York, with demographic & clinical data extracted from Epic EHR (Verona, WI) databases.
Why	Currently most studies of COVID-19 related health outcomes use demographic statistics offering information only about single risk factors, rather than combined risk. To do a comprehensive risk evaluation based on personal demographic , physical characteristics acquired at 1st encounter to predict COVID-19 mortality.
How	Demographic and clinical data were extracted from EHRs, then a MV logistic regression model was used to estimate association between mortality and factors. To assess utility in predicting COVID mortality, they created a ML model using Extreme Gradient Boosting framework with 10-fold CV.

Fig 1.



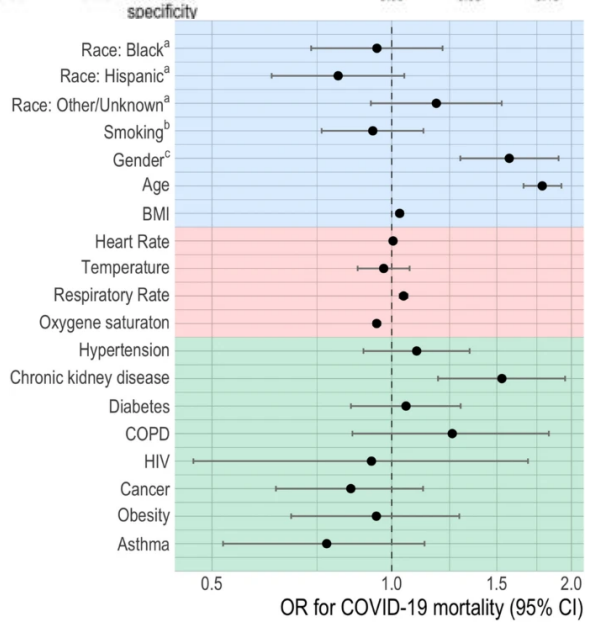


Fig 2.

Patient Subtyping via Time-Aware LSTM Networks

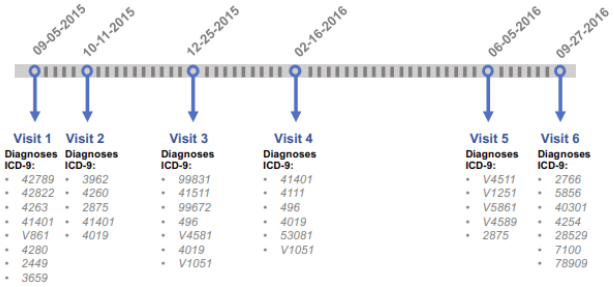


Fig 1.

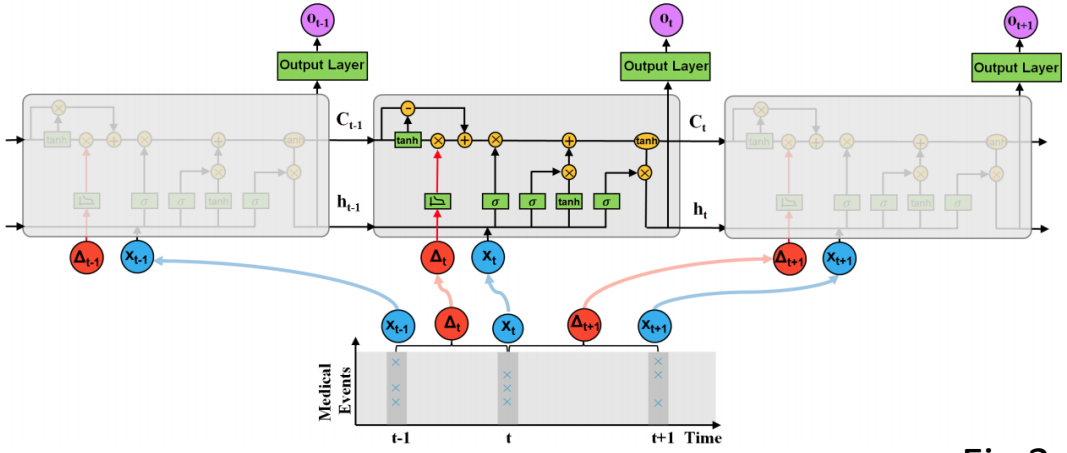


Fig 2.

Methods	Avg. Test AUC	Stdev.
T-LSTM	**0.91**	0.01
MF1-LSTM	0.87	0.02
MF2-LSTM	0.82	0.09
LSTM	0.85	0.02
LR	0.56	0.01

Fig 4.

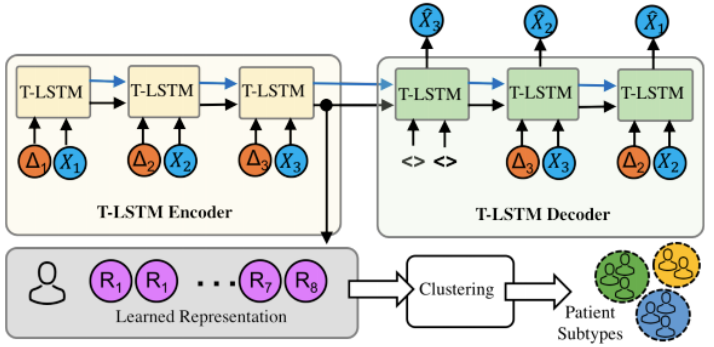


Fig 3.

Citation	*Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient Subtyping via Time-Aware LSTM Networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery, New York, NY, USA, 65–74.*
Who	**Inci M. Baytas** , Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain and Jiayu Zhou. (**Michigan State University, East Lansing, MI** ; Weill Cornell Medical School Cornell University, New York, NY ; IBM T. J. Watson Research Center, New York, NY)
What	Authors propose a patient subtyping model that leverages the proposed novel T-LSTM (Time-Aware LSTM) in an auto-encoder to learn a powerful single representation for sequential records of patients, which are then used to cluster patients into clinical subtypes.
When	The article was published on August 2017, with data collected from 2 synthetic publicly available EHRs data and 1 real world PPMI database with patient data collected over 2010-2017.
Where	Apart from the 2 synthetic datasets, the real-world Parkinson's Progression Markers Initiative (PPMI) database consists of EHRs which is a longitudinal dataset with unstructured elapsed time, from more than 1,400 individuals at 33 clinical sites in 11 countries.
Why	Subtyping from complex patient data is challenging because of the information heterogeneity and temporal dynamics. LSTM units are designed to handle data with constant elapsed times between consecutive elements of a sequence, but given that time lapse between successive elements in patient records can vary from days to months, the design of traditional LSTM may lead to suboptimal performance.
How	A novel LSTM architecture (T-LSTM) is proposed to handle time irregularities in sequences. T-LSTM has forget, input, output gates of standard LSTM, but the memory cell is adjusted in a way that longer the elapsed time, smaller the effect of previous memory to the current output. For this purpose, elapsed time is transformed into a weight using a time decay function. T-LSTM learns a neural network that performs a decomposition of cell memory into short & long-term memories. The short-term memory is discounted by the decaying weight before combining it with the long-term counterpart.

COVID-19 Clinical Phenotypes: Presentation and Temporal Progression of Disease in a Cohort of Hospitalized Adults in Georgia, United States

Citation	*Juliana F da Silva et al. COVID-19 Clinical Phenotypes: Presentation and Temporal Progression of Disease in a Cohort of Hospitalized Adults in Georgia, United States, Open Forum Infectious Diseases, Volume 8, Issue 1, January 2021.*
Who	**Juliana F da Silva** et al. - From Centers for Disease Control and Prevention (**CDC**), Atlanta, Georgia, USA
What	They showed 1 phenotype of 6 identified was characterized by high mortality (49%), older age, male sex, elevated inflammatory markers, high prevalence of cardiovascular disease, and shock. This can be used to inform tailored clinical management for COVID-19
When	The article was published on December 2020, with longitudinal clinical data of 305 hospitalized COVID patients during March – April 2020.
Where	Longitudinal clinical data from EMRs on 305 hospitalized patients with laboratory-confirmed SARS-CoV-2 infection in the US state of Georgia, collected from 7 hospitals were in metropolitan Atlanta, and 1 in southern Georgia, reviewed by CDC, Georgia DPH and 3 hospital networks.
Why	Though advanced age, comorbid conditions, & inflammatory markers have been identified as independent risk factors for severe disease, how these risk factors interact and their relationship to severe pulmonary, extrapulmonary complications remains poorly understood.
How	Authors examined laboratory & vital sign trends, by mortality status & length of stay. For identifying clinical phenotypes, calculated Gower's dissimilarity matrix between each patient's clinical characteristics & clustered similar patients using the partitioning around medoids algorithm.

6 clusters

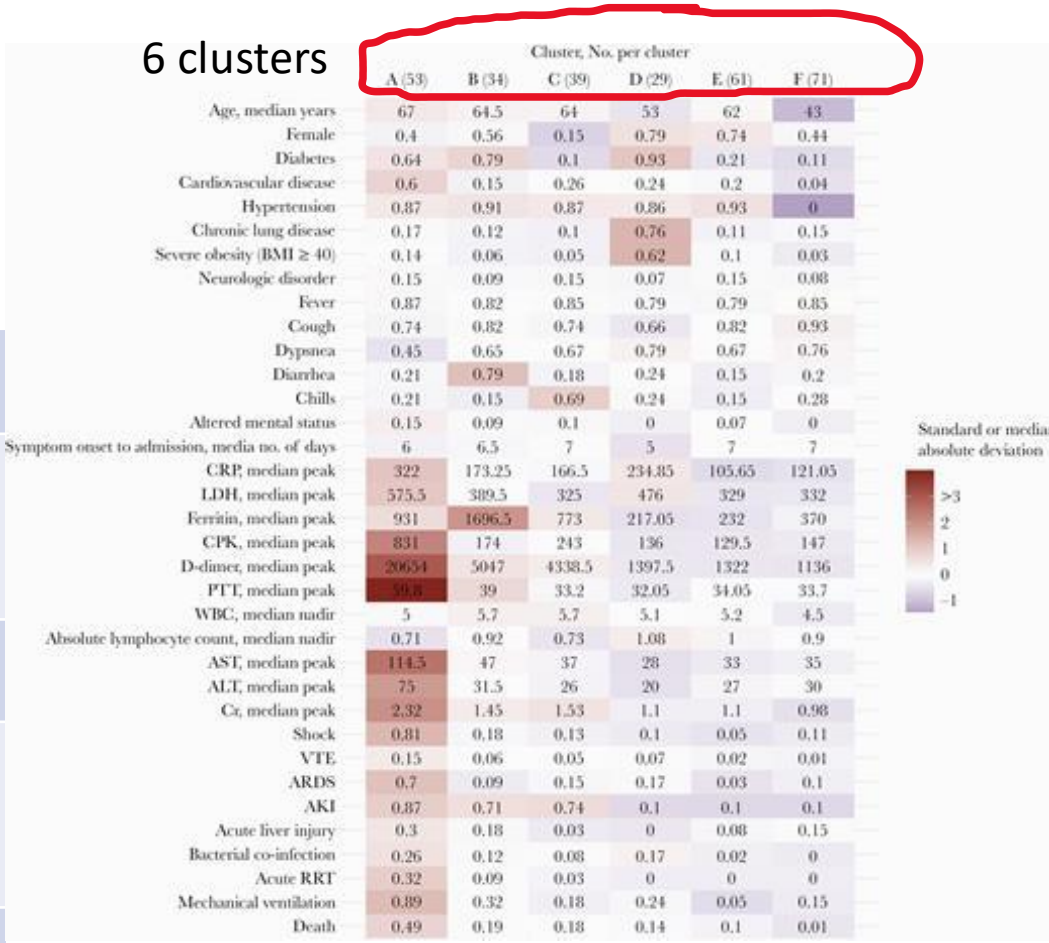


Fig 1.

Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study

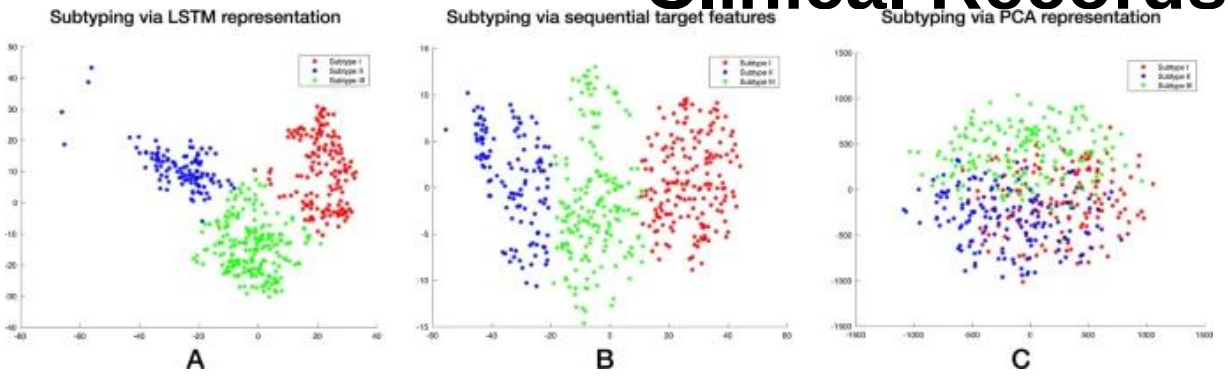


Fig 2.

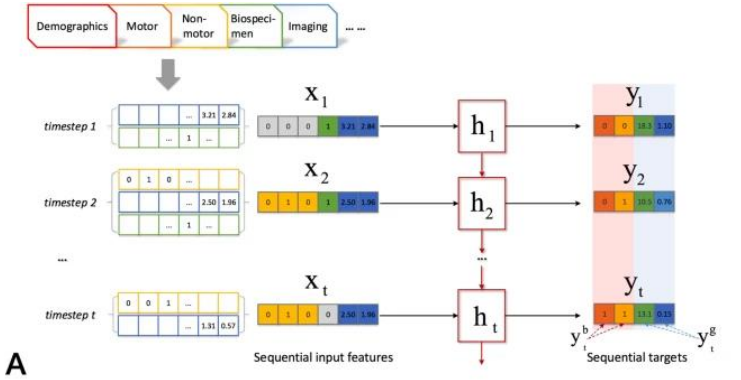


Fig 1.

Citation	*Zhang, X., Chou, J., Liang, J. et al. Data-Driven Subtyping of Parkinson's Disease Using Longitudinal Clinical Records: A Cohort Study. Sci Rep **9,** 797 (2019).*
Who	Xi Zhang, Jingyuan Chou, **Jian Liang**, Cao Xiao,Yize Zhao, Harini Sarva, Claire Henchcliffe & Fei Wang (from Weill Cornell Medical College, Cornell University, New York, USA ; Tsinghua University, Beijing, China ; IBM Research, Cambridge, USA)
What	466 patients with idiopathic PD were investigated and three subtypes were identified. Subtypes suggest that when comprehensive clinical, biomarker data are used in deep learning algorithm,disease progression rates do not necessarily associate with baseline severities, & progression rate of non-motor symptoms is not necessarily correlated with the progression rate of motor symptoms
When	The article was published on January 2019, with PPMI (Parkinson's Progression Markers Initiative) data contained archives of enrolled subjects from June 1, 2010, to June 1, 2016.
Where	The EHRs data was downloaded from PPMI database on June 21, 2016. The de-identified data contained archives of enrolled subjects from June 1, 2010 to June 1 2016, from more than 1,400 individuals at 33 clinical sites in 11 countries.
Why	Conventional approaches for PD subtyping typically focus on 1 specific aspect (e.g., motor or cognition) of the patient characteristics. Therefore, to develop more comprehensive approaches that can consider different aspects of patient characteristics during the subtyping process using deep learning did not exist.
How	Comprehensive clinical information was extracted from PPMI database. A deep learning algo, Long-Short Term Memory (LSTM), was used to represent each patient as a multi-dimensional time series for subtype identification.

Learning representations for the early detection of sepsis with deep neural networks

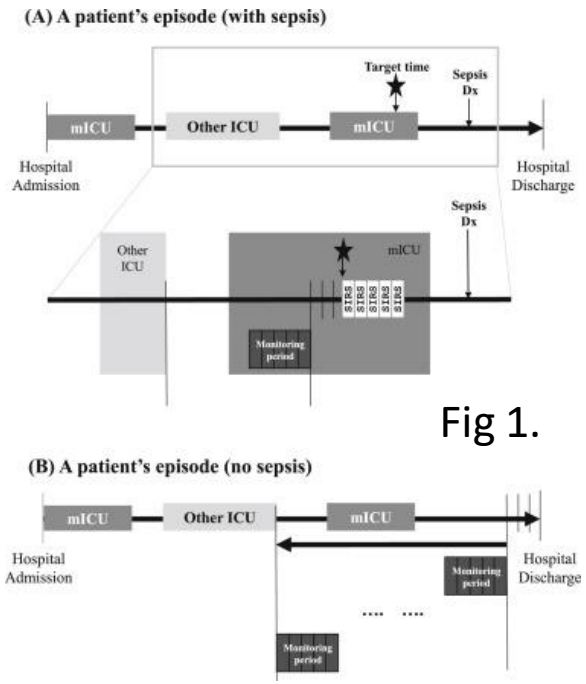


Fig 1.

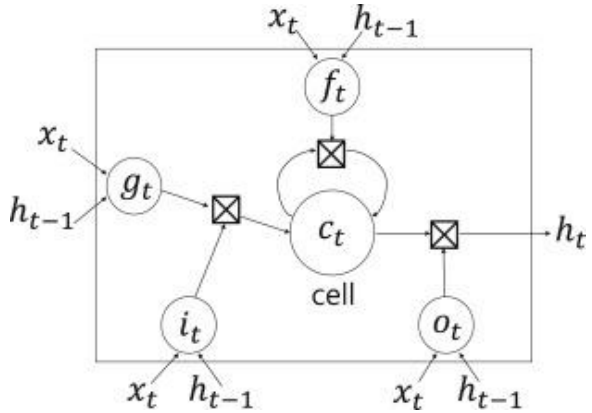


Fig 2.

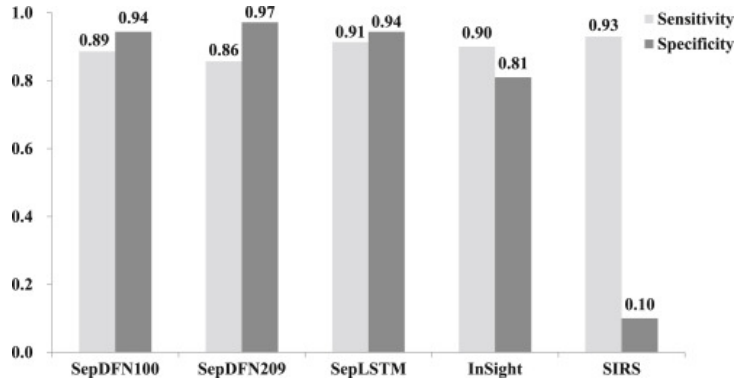


Fig 3.

Citation	*H.J. Kam, H.Y. Kim. Learning representations for the early detection of sepsis with deep neural networks. Comput. Biol. Med., 89 (2017), pp. 248-255*
Who	Hye JinKam, **Ha YoungKim** (from Asan Institute for Life Sciences, South Korea ; **Ajou University, South Korea**)
What	They developed sepsis detection models with deep-feedforward networks and LSTM model, with AUC of proposed models improved to up to 0.929. Improved performance of proposed models show capability of feature extraction, & LSTM has superior capability for sequential patterns.
When	The article was published on August 2017, with patient data from MIMIC-II database comprising of 400,000 patients enrolled in ICU from 2001 to 2012.
Where	Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II, version 3) database, an open-source, clinical database encompassing approximately 400,000 patients enrolled in the ICU at the Beth Israel Deaconess Medical Center in Boston.
Why	Whether it is possible to replace the feature extraction steps in the existing regression and machine learning methods, which is an important factor in performance. Also, to judge whether newly applied models such as RNNs or LSTMs can reflect signal volatility in the raw data that Calvert et al. tried to extract with temporal features.
How	Study group selection was adhered to the InSight model. 4 conditions satifying SIRS model was chosen as target definition. Out of 460, 9 variables were taken after data-preprocessing, for modeling they experimented with 2 DNN architectures : Multilayer perceptrons (MLPs), LSTM (Long-Short Term Memory) networks

Novel Temperature Trajectory Sub phenotypes in COVID-19

Dr. Bhavani SV

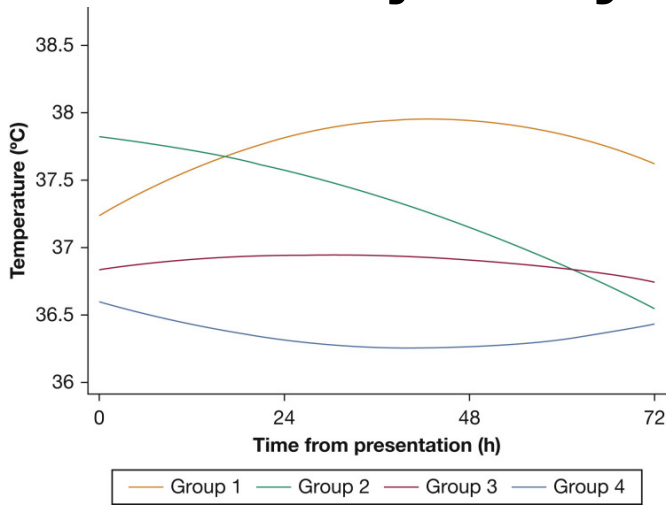


Fig 1.

EMORY UNIVERSITY SCHOOL OF MEDICINE

Citation	*Bhavani SV, Huang ES, Verhoef PA, Churpek MM. Novel Temperature Trajectory Sub phenotypes in COVID-19. Chest. 2020;158(6):2436-2439. doi:10.1016/j.chest.2020.07.027*
Who	**Sivasubramanium V. Bhavani**, Elbert S. Huang, Philip A. Verhoef, and Matthew M. Churpek (from University of Chicago, Chicago, IL ; Kaiser Permanente, Honolulu, HI; University of Wisconsin, Madison, WI ; **University of Chicago Medical Center, IL**)
What	They report the use of longitudinal temperature measurements to identify novel sub phenotypes in COVID-19 illness. High mortality rate that was seen in group 1 and the organ dysfunction that was seen in group 4 suggest that both sub phenotypes have a dysregulated response to COVID-19.
When	This study was published on July 2020, comprising of all adult patients admitted to University of Chicago Medicine between March 1 and June 24, 2020, who tested positive for SARS-CoV-2 within 72 hours.
Where	Temperature measurements that had been taken in the first 72 hours of hospitalization at the University of Chicago Medicine for COVID +ve patients.
Why	Authors previously published a novel method of identifying sub phenotypes in hospitalized patients with all-cause infection with the use of longitudinal body temperature measurements, and with this they hypothesized that using a similar approach that is specific to patients with COVID-19 would identify sub phenotypes with unique clinical characteristics and inflammatory and coagulation abnormalities.
How	They included all adult patients who were admitted to University of Chicago Medicine between March 1 and June 24, 2020, who tested positive and excluded patients who had been tested for SARS-CoV-2, > 3 days after admission. They included temperature measurements that had been taken in the first 72 hours of hospitalization in the group-based trajectory modeling (GBTM) algorithm, a finite mixture model used to identify clusters of patients following similar trajectories of a variable of interest.

Expanding the literature critique in more details : ends

https://med.emory.edu/directory/profile/?u=SBHAVA2

Expected Data Modalities

- COVID-19 patients' vital signs: **Time series data**
  - Body temperature
  - Blood pressure
  - Heart rate
  - Respiratory rate
- Demographic background: **Time independent data**
  - Age
  - Gender
  - Race
  - Medical History
- Clinical Outcome: **Ground Truth**
  - Mortality/Morbidity
- Potential Clinical Data: **Time dependent data**
  - Time of drug administration
  - White Blood Cell & Neutrophil Count

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

Fig 1.

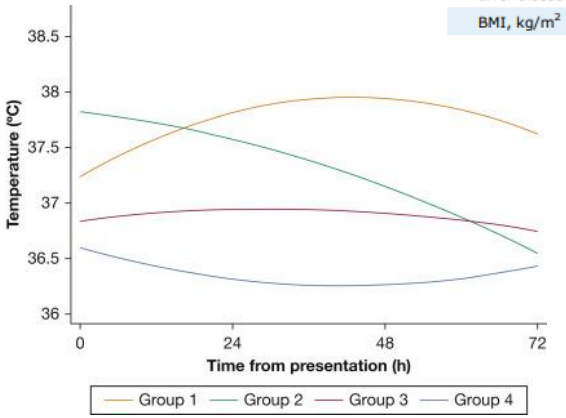


Fig 2.

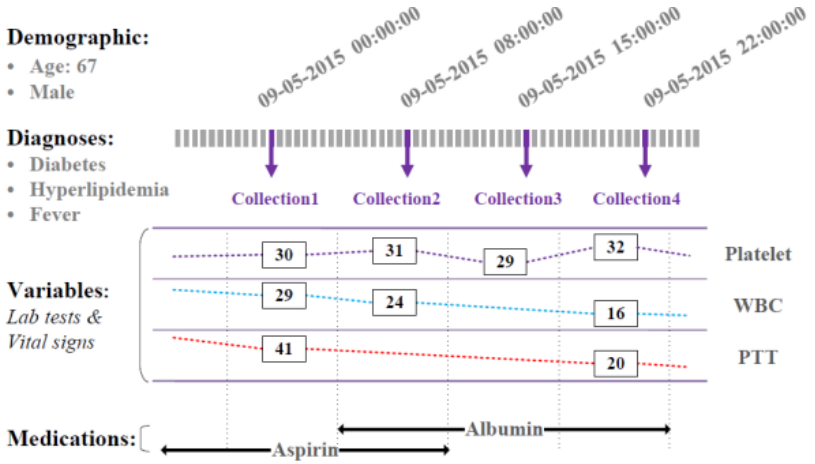
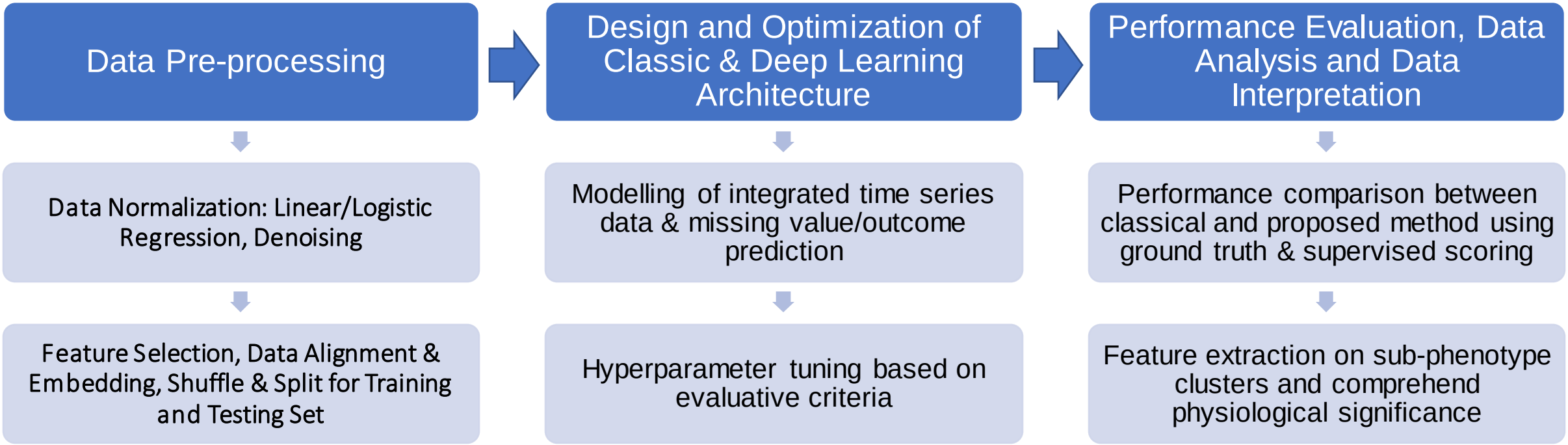


Fig 3.

Proposed Workflow



Deep Learning methodology: LSTM Auto-Encoder for Time Series Data

Auto-encoder: Compresses input data into a latent space representation composed of much small dimension but can still be reconstructed back to match the input; aimed to learn a representation/encoding of input data similar to dimensionality reduction.

Long Short-Term Memory (LSTM) Model: Network designed to support sequences of input data.

- Autoencoder structure were applied to initial section of time series data to encode, decode and evaluate performance of the model.
- Once the model achieves good decoding performance, the decoder part of the model is removed, leaving only the encoder model for encoding the time series data into series of fixed vectors of latent representation that can be used for supervised clustering (i.e. K-means clustering)

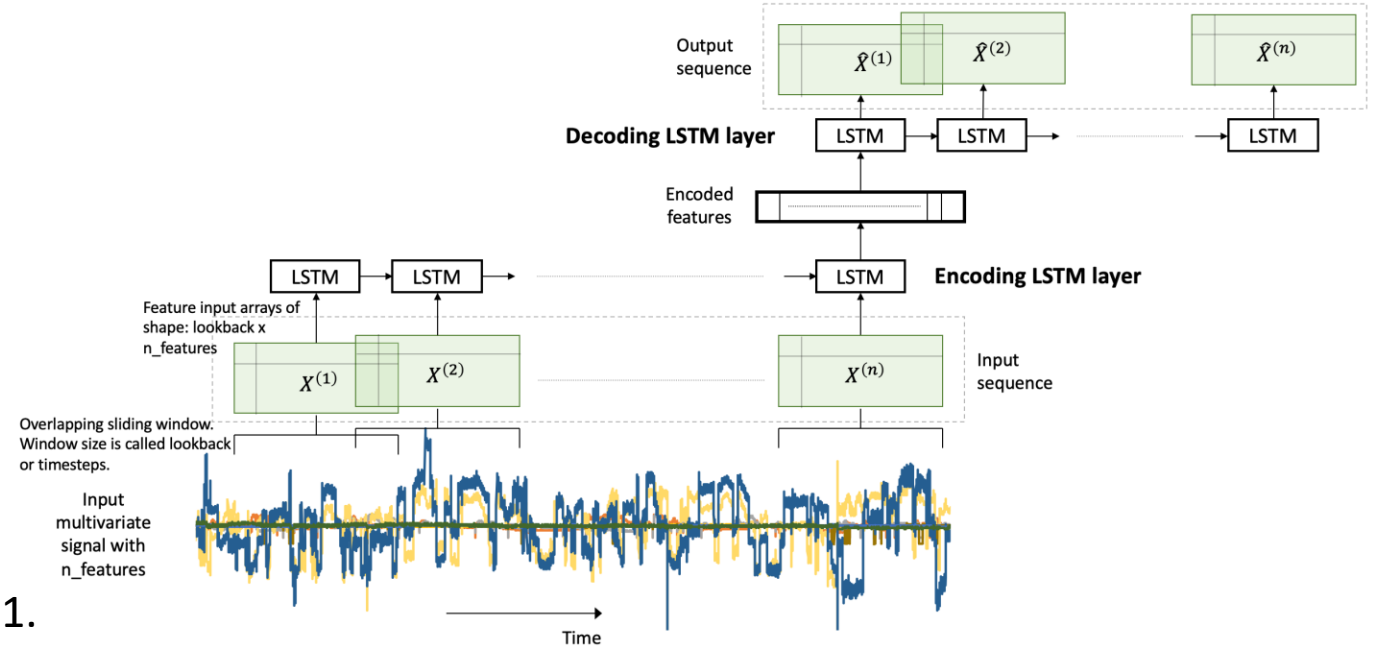


Fig 1.

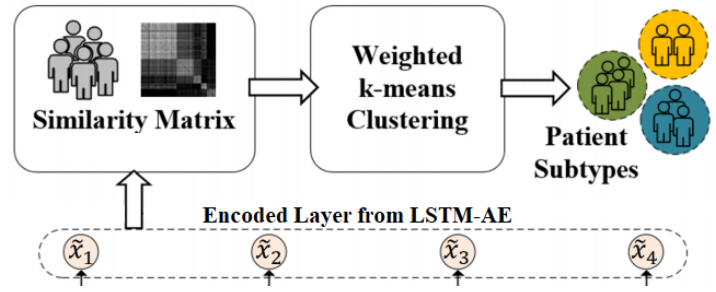


Fig 2.

Mathematics behind proposed Time-aware LSTM autoencoder

Data Imputation

$$c_t^i = \begin{cases} 1 & \text{if } x_t^i \text{ is observed in input data,} \\ 0 & \text{else,} \end{cases}$$ - eq 1

$$a_t^i = \begin{cases} 1 & \text{if } c_t^i = 0 \ \& \ x_t^i \text{ is observed in ground truth,} \\ 0 & \text{else,} \end{cases}$$ - eq 2

Here, two masking matrix C & A

$$\Delta_t^{(l),i} = \begin{cases} \Delta_t & \text{if } c_{t-1}^i = 1, \\ \Delta_{t-1}^{(l),i} + \Delta_t & \text{else} \end{cases}$$ - eq 3

$$\Delta_t^{(n),i} = \begin{cases} \Delta_{t+1} & \text{if } c_{t+1}^i = 1, \\ \Delta_{t+1}^{(n),i} + \Delta_{t+1} & \text{else} \end{cases}$$ - eq 4

Above, 3 time gap vector and matrices,
Δ ∈ R |X | , Δ(l) ∈ R |X | × Kx ,
Δ(n) ∈ R |X | × Kx

$$x_t^{(l),i} = \begin{cases} x_{t-1}^i & \text{if } c_{t-1}^i = 1, \\ x_{t-1}^{(l),i} & \text{else} \end{cases}$$ - eq 5

$$x_t^{(n),i} = \begin{cases} x_{t+1}^i & \text{if } c_{t+1}^i = 1, \\ x_{t+1}^{(n),i} & \text{else} \end{cases}$$ - eq 6

2 neighboring value matrices
X (l) and X (n).

to incorporate longitudinal information.

Multi-modal Embedding

$$e_j'^v = sin(\frac{v * j}{V * k})$$ - eq 7

$$e_{k+j}'^v = cos(\frac{v * j}{V * k}),$$ - eq 8

embed variables & sub-ranges into a vector e^('v) ∈ R^2k

$$e^{iv} = [e^i; e'^v]W_{iv} + b_{iv},$$ - eq 9

fully connected layer is followed to map concatenation vector into a new value embedding vector e^(iv) ∈ R^k

$$e_j^\delta = sin(\frac{\delta * j}{T_m * k})$$ - eq 10

$$e_{k+j}^\delta = cos(\frac{\delta * j}{T_m * k}),$$ - eq 11

Given a time gap δ, the time embedding layer outputs a vector e^(δ) ∈ R^2k

Changchang Yin et al. 2020

BiLSTM Architecture

Given, multi-modal embedding vectors e^t

$$\hat{e}_t = e_t W_e + e_t^\Delta W_\Delta + b_e$$

$$\overrightarrow{h}_1, \overrightarrow{h}_2, ..., \overrightarrow{h}_{|X|} = \overrightarrow{LSTM}(\hat{e}_1, \hat{e}_2, ..., \hat{e}_{|X|})$$

$$\overleftarrow{h}_1, \overleftarrow{h}_2, ..., \overleftarrow{h}_{|X|} = \overleftarrow{LSTM}(\hat{e}_1, \hat{e}_2, ..., \hat{e}_{|X|})$$

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \quad \text{for } t = 1, 2, ..., |X|,$$

- eq 12

the time gap embedding vector e^(Δ) t is also an input to auto-encoder

→LSTM and ←LSTM are forward and backward directional

Temporal Similarity with DTW

The distance between $S^{(i)}$ and $S^{(j)}$ is :

$$Dist_p(S^{(i)}, S^{(j)}) = \frac{Dist(S_1^{(i)}, S_1^{(j)})}{\max(|S^{(i)}|, |S^{(j)}|)}$$

$$Dist(S_k^{(i)}, S_l^{(j)}) = dist(s_k^{(i)}, s_l^{(j)}) + \min \begin{cases} Dist(S_{k+1}^{(i)}, S_l^{(j)}) \\ Dist(S_k^{(i)}, S_{l+1}^{(j)}) \\ Dist(S_{k+1}^{(i)}, S_{l+1}^{(j)}), \end{cases}$$ - eq 13

where $dist(s_k^{(i)}, s_l^{(j)})$ is defined with Euclidean distance:

$$dist(s_k^{(i)}, s_l^{(j)}) = ||s_k^{(i)} - s_l^{(j)}||_2$$

The boundary condition is as follows:

$$Dist(S_k^{(i)}, S_{|S^{(j)}|}^{(j)}) = \sum_{m=k}^{|S^{(i)}|} dist(s_m^{(i)}, s_{|S^{(j)}|}^{(j)})$$

$$Dist(S_{|S^{(i)}|}^{(i)}, S_l^{(j)}) = \sum_{m=l}^{|S^{(j)}|} dist(s_{|S^{(i)}|}^{(i)}, s_m^{(j)})$$

Here, S is an imputed matrix for each patient after replacing missing values

Weighted K-means Clustering

$$Dist_g(S^{(i)}, G_k) = \frac{\sum_{j \in G_k} Dist_p(S^{(i)}, S^{(j)}) * w_j}{\sum_{j \in G_k} w_j}$$

$$w_j = (1 + exp(\sum_{l \in G_k} \frac{Dist_p(S^{(j)}, S^{(l)})}{|G_k|}))^{-1} \quad j \in G_k$$

- eq 14

Finally, evaluate the models' performance with nRMSE

Architecture of proposed Solution, how it works, why we chose this ?

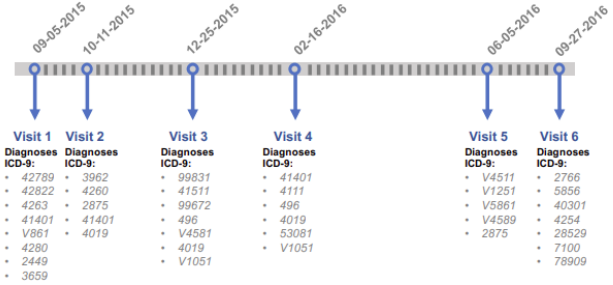


Fig 2. Longitudinal medical records data need some kind of imputation strategy and a layer/step to handle such irregular data

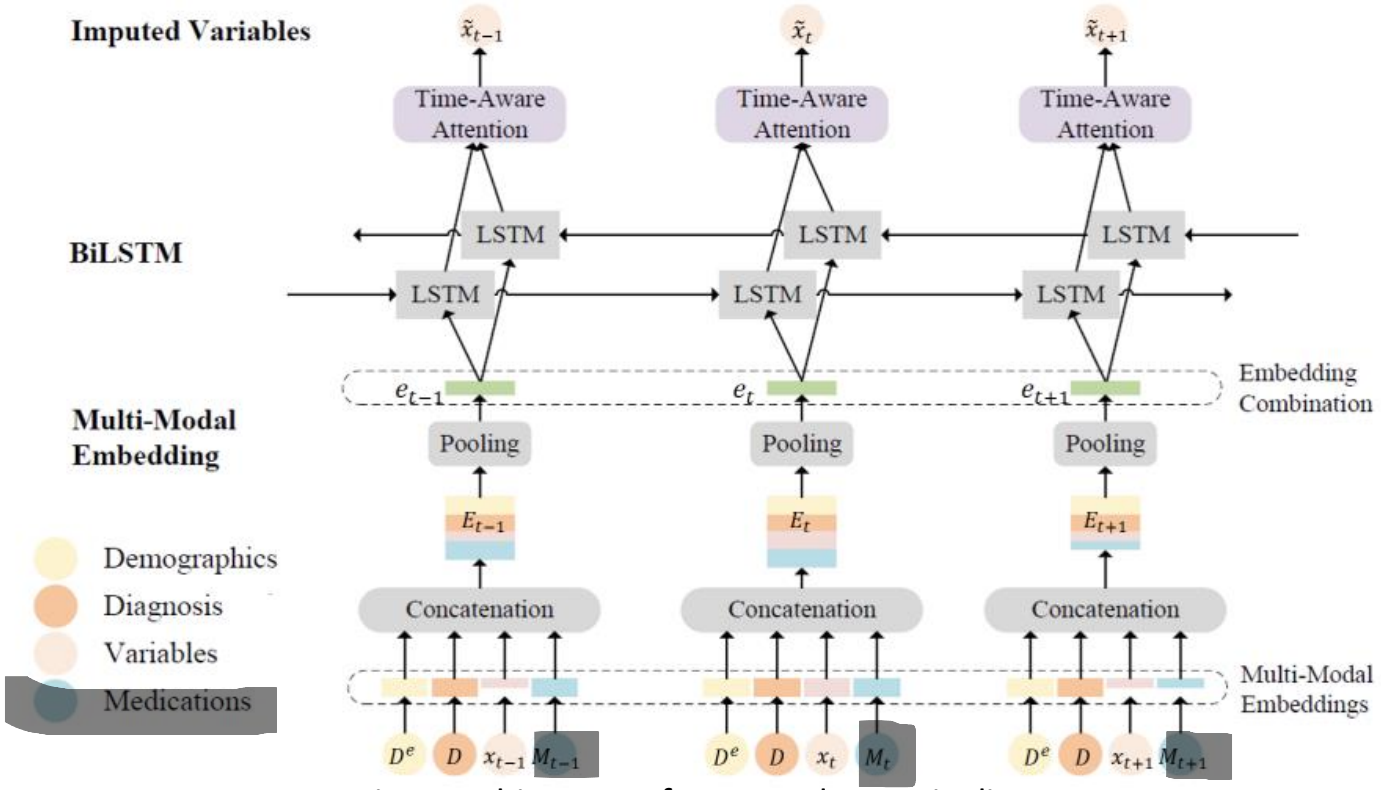


Fig 1. Architecture of proposed DNN pipeline.

DNN architecture will take multi-modal data as inputs, including demographics, diagnoses, 5 variables (i.e., vital signs), and maybe medications. Any patient may have varying numbers of missing values at different times. Input dimensions are expected to vary across collected data. By concatenating the embeddings of the inputs, we will obtain a matrix E(t) containing multi-modal information. After that a Pooling layer is to output a fixed-size vector e(t) , which will be sent to Bi-LSTM. Towards top, a time-aware attention module is proposed to be used to handle the longitudinal information and then impute the missing values.

how it works ?

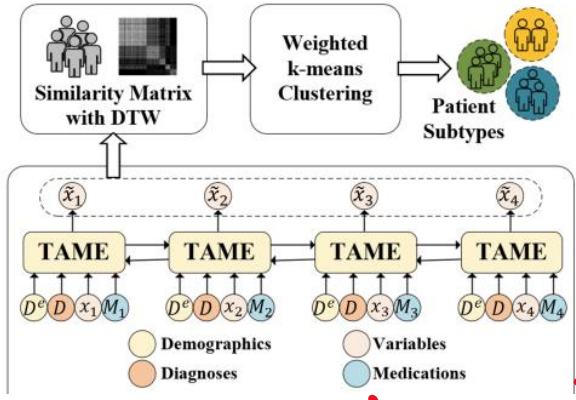
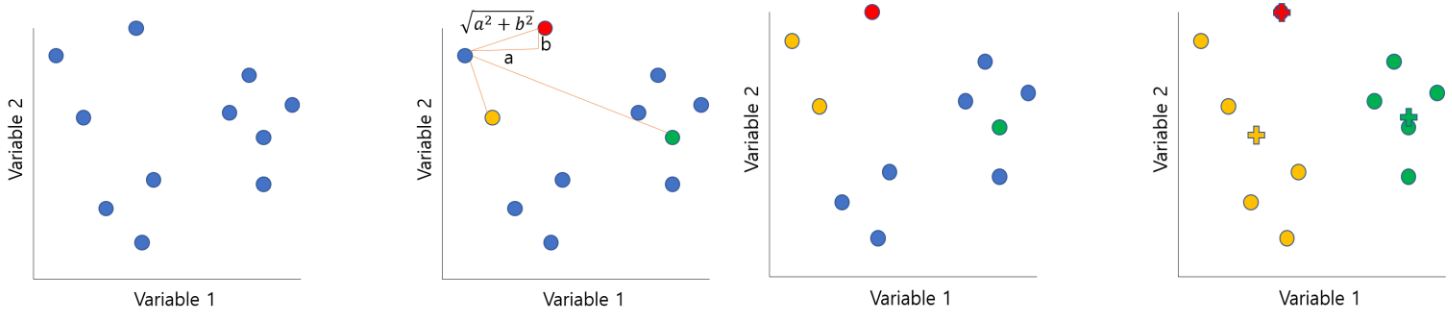


Fig 3. : Solution proposal to Cluster patients with time-aware LSTM autoencoder, DTW and weighed k-means.

why we chose this ?

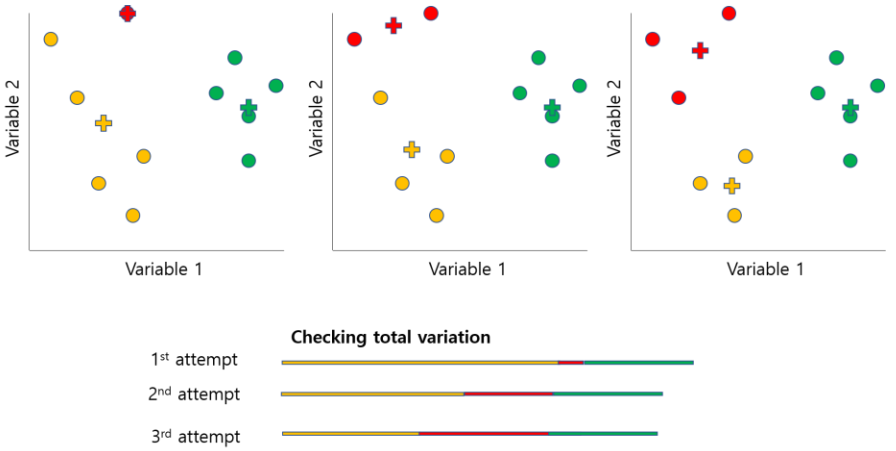
- Existing classical ML models factor in either basic clinical factors or focus on only 1 modality
- Handling longitudinal EHRs data with time irregularities isn't feasible with classical ML
- Time-Aware LSTM (T-LSTM) have been proven to handle irregular time intervals in longitudinal patient records.
- Patient subtyping is posed as an unsupervised clustering problem since we do not have any prior information about the groups inside the patient cohort, and so AUTO-ENCODERS

How K-mean cluster works?

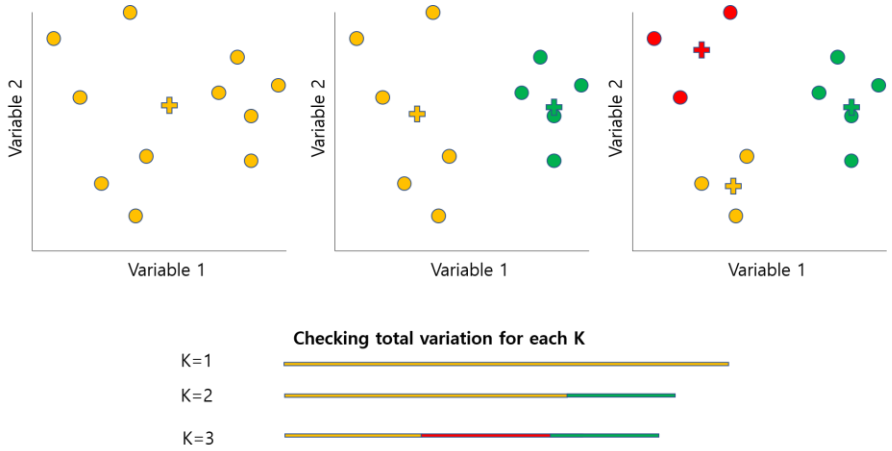


Step 1: Select number of cluster (K) and find distance of each point to the cluster randomly selected (Euclidian distance)

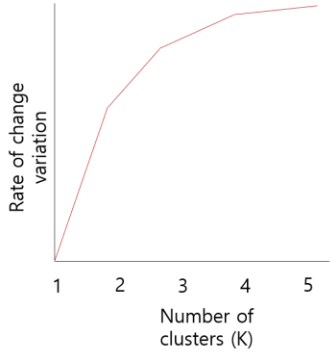
Step 2: Assign each point to the cluster based on minimum distance. Find mean and variance of each cluster



Step 3: Repeat the procedure until we get smallest total variation



Step 4: Repeat step 1-3 with different K until we get smallest total variation



Step 5: Select best (K) for clustering

Mathematics behind K-mean cluster

Given a set of observations, $(x_1, x_2, x_3, \ldots, x_n)$, where each observation is a d-dimensional real vector, k-mean clustering aims to partition n observations into k($\leq n$) set $S = \{S_1, S_2, S_3, \ldots, S_k\}$ so as to minimize the within-cluster sum of square (variance).

Goal: $$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

Where μ_i is the mean of points in cluster S_i

Algorithm of K-mean cluster

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

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

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;

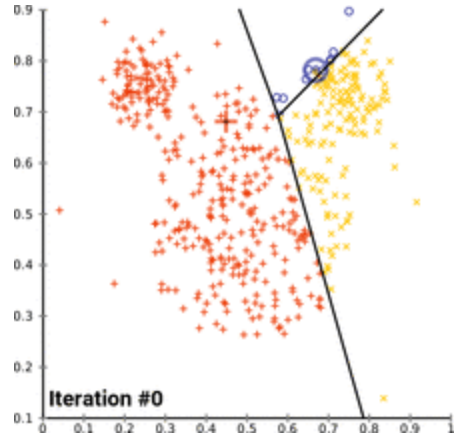


Fig 1. The algorithm iterates between steps the assignment and update steps until the stopping criterion are reached (the maximum number of iterations set by the user, no data points change clusters, or the sum of distances is minimized)

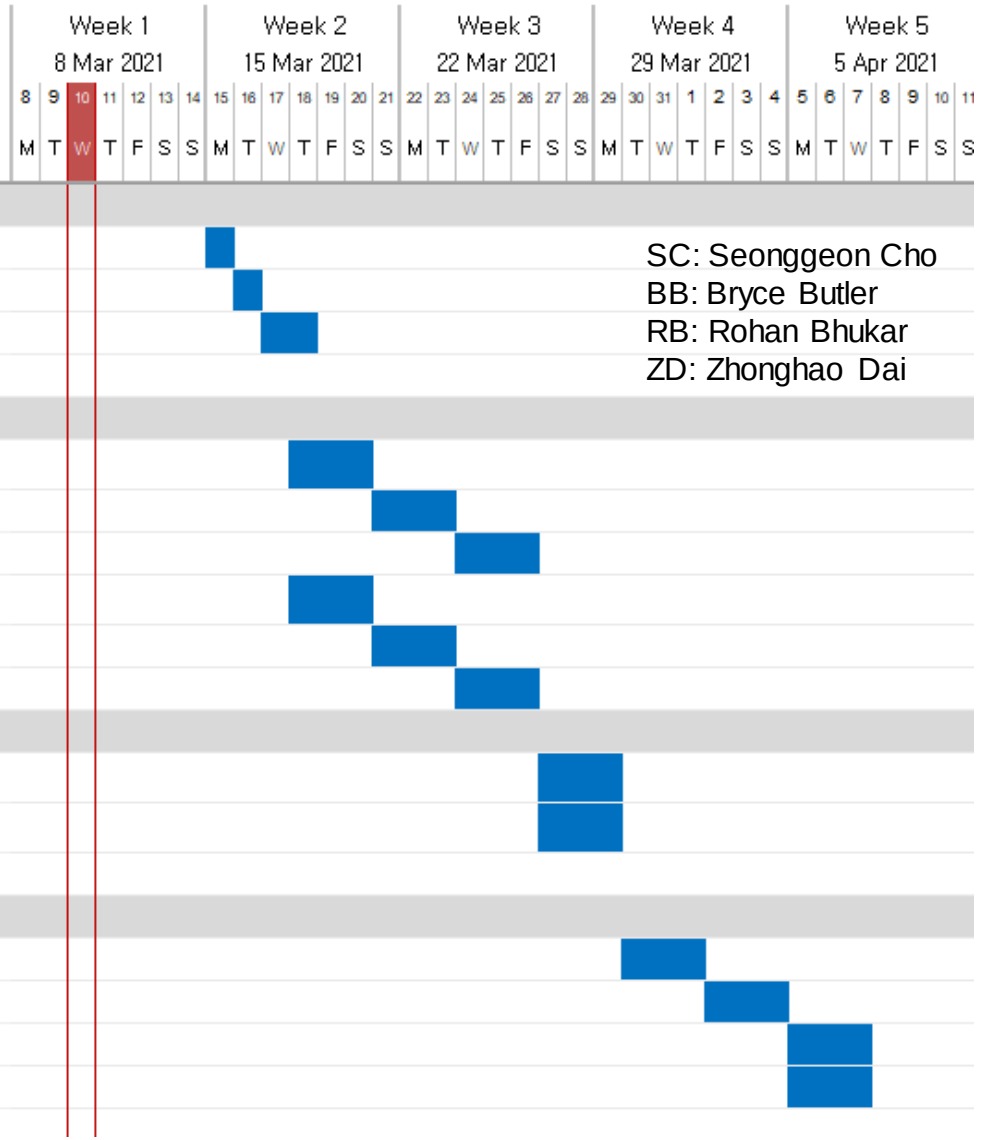
Why K-mean cluster?

- Relatively simple to implement
- Scales to large data sets and adapts to new examples
- Results are less susceptible to outliers in data and influence of chosen distance measured
- This method is less computational
- We were able to find multiple papers to use as a framework in building our algorithm

Project timeline and individual tasks

Project Start Date: 3-10-2021 (Wednesday) | Display Week: 1
Project Lead: Seonggeon Cho, Rohan Bhukar, Bryce Butler, Zhonghao Dai

WBS	TASK	LEAD	START	END	DAYS	% DONE	WORK DAYS
1	**Data Pre-processing**			-			-
1,1	Data normalization	All	Mon 3-15-21	Mon 3-15-21	1	0%	1
1,2	Feature selection	All	Tue 3-16-21	Tue 3-16-21	1	0%	1
1,3	Data alighment & Embedding	All	Wed 3-17-21	Thu 3-18-21	2	0%	2
1,4				-		0%	-
2	**Design Model and Training**			-			-
2,1	Model K-mean clustering (Data aggregated over entire timeline)	SC, BB	Thu 3-18-21	Sat 3-20-21	3	0%	2
2,2	Training K-mean clustering	SC, BB	Sun 3-21-21	Tue 3-23-21	3	0%	2
2,3	K-mean clustering model optimization	SC, BB	Wed 3-24-21	Fri 3-26-21	3	0%	
2,4	Model LSTM auto-encoder (Integrated time series data)	ZD, BR	Thu 3-18-21	Sat 3-20-21	3	0%	2
2,5	Training LSTM auto-encoder	ZD, BR	Sun 3-21-21	Tue 3-23-21	3	0%	2
2,6	LSTM auto-encoder model optimization	ZD, BR	Wed 3-24-21	Fri 3-26-21	3	0%	3
3	**Model Comparison**			-			-
3,1	Visualization and evaluation of K-mean clustering	SC, BB	Sat 3-27-21	Mon 3-29-21	3	0%	1
3,2	Visualization and evaluation of LSTM auto-encoder	ZD, BR	Sat 3-27-21	Mon 3-29-21	3	0%	1
3,3	[Task]			-		0%	-
4	**Interpretation**			-			-
4,1	Common findings from each methods	All	Tue 3-30-21	Thu 4-01-21	3	0%	3
4,2	Significance of our works	All	Fri 4-02-21	Sun 4-04-21	3	0%	1
4,3	Evaluate limitation of our works	All	Mon 4-05-21	Wed 4-07-21	3	0%	3
4,4	Possible Future work	All	Mon 4-05-21	Wed 4-07-21	3	0%	3
4,5				-		0%	-



Reference articles

1. Nature. (2020). "The coronavirus is mutating — does it matter?." Retrieved from https://www.nature.com/articles/d41586-020-02544-6
2. World Health Organization. (2021). "WHO Coronavirus (COVID-19) Dashboard." Retrieved from https://covid19.who.int/
3. Landi, I., Glicksberg, B.S., Lee, HC. *et al.* Deep representation learning of electronic health records to unlock patient stratification at scale. *npj Digit. Med.* **3,** 96 (2020).
4. Nagamine, T., Gillette, B., Pakhomov, A. *et al.* Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. *Sci*
5. Miotto, R., Li, L., Kidd, B. *et al.* Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* **6,** 26094 (2016).
6. Changchang Yin, Ruoqi Liu, Dongdong Zhang, and Ping Zhang. 2020. Identifying Sepsis Subphenotypes via Time-Aware Multi-Modal Auto-Encoder. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)
7. Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient Subtyping via Time-Aware LSTM Networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery, New York, NY, USA, 65–74.
8. Zhang, Y. D., et al. (2021). COVID-19 Diagnosis via DenseNet and Optimization of Transfer Learning Setting. Cognitive computation, 1–17. Advance online publication. https://doi.org/10.1007/s12559-020-09776-8

Reference articles cont.

1. Ma, T., & Zhang, A. (2018). Affinity network fusion and semi-supervised learning for cancer patient clustering. Methods, 145, 16-24.
2. Campbell, Thomas W., et al. "Predicting Prognosis in COVID-19 Patients using Machine Learning and Readily Available Clinical Data." medRxiv (2021).
3. Aguiar, Henrique, et al. "Phenotyping Clusters of Patient Trajectories suffering from Chronic Complex Disease." *arXiv preprint arXiv:2011.08356* (2020).
4. Bloch, Eli, et al. "Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction." *Journal of healthcare engineering* 2019 (2019).
5. Bloch, Eli, et al. "Machine learning models for analysis of vital signs dynamics: a case for sepsis onset prediction." *Journal of healthcare engineering* 2019 (2019).
6. Aguiar, Henrique, et al. "Phenotyping Clusters of Patient Trajectories suffering from Chronic Complex Disease." *arXiv preprint arXiv:2011.08356* (2020).
7. Chaudhary, Kumardeep, et al. "Unsupervised Machine learning to subtype Sepsis-Associated Acute Kidney Injury." *bioRxiv* (2018): 447425.