

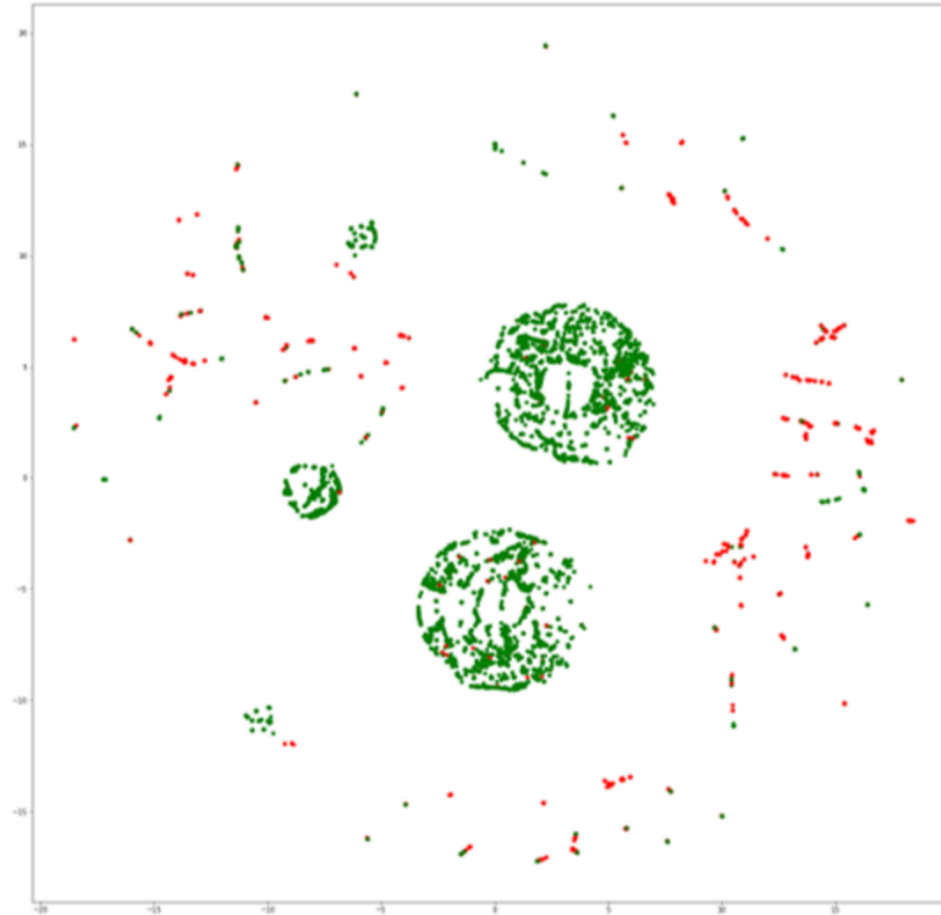
Proposing Tools to Allow Finding "Sub-Types" of Diseases Using Machine Learning Methods

Presented by : Mor Zukin

Supervised by Prof. Ran Gilad-Bachrach

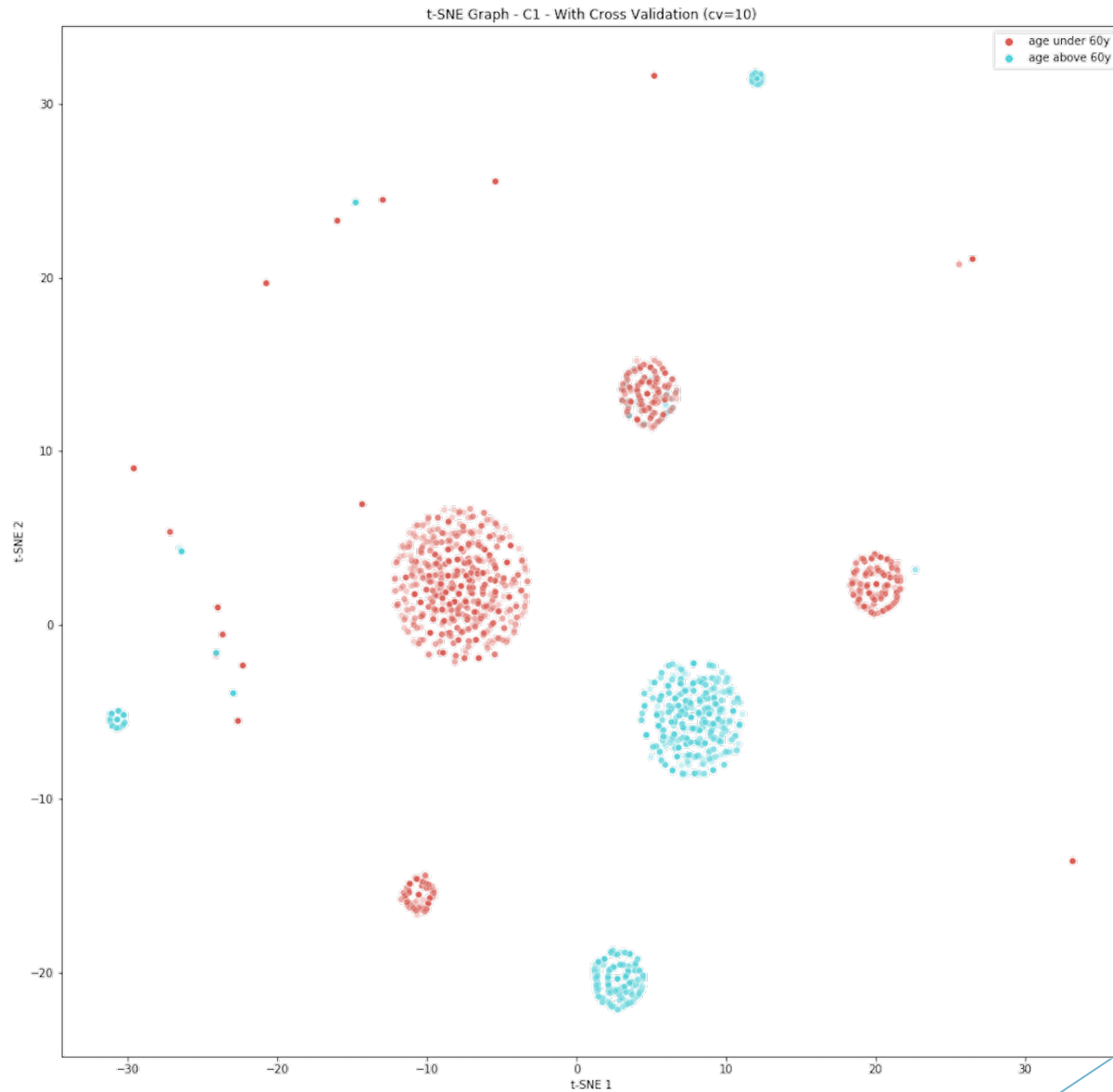
Biomedical Engineering Department
Faculty of Engineering
Tel-Aviv University, ISRAEL
April 2020

Corona Time



תרשים מס' 6: דיאגרמת פיזור של בדיקות חיוביות לקורונה (אדום) ושליליות לקורונה (ירוק).
הערכים של כל ציר עבור כל בדיקה נקבעו באמצעות הפעלת טכניקת T-SNE להפחתת מימד
- המאפיינים הקליניים (סימפטומים, קטגוריית גיל, מגדר, אינדיקציה לביצוע הבדיקה)

Corona Time



Introduction

Background

- ▶ Some diseases may have sub-types that react differently to treatment
- ▶ In this work we implement an existing method, using machine learning (ML) tools, for the purpose of finding sub-types of diseases

Background

- ▶ ML algorithms are often categorized as supervised learning or unsupervised learning

	Supervised learning	Un-supervised learning
Data used for training	Labeled data	Un-labeled data
Purpose	classification	clustering

- ▶ The ML problem we are dealing with in this project is closely related to an unsupervised clustering problem

Our Problem

- ▶ Different disease sub-types are likely to appear in different ratios across populations
- ▶ One of the guidelines is that there is a chance that the odds of acquiring a certain sub-disease may be influenced by measurable properties
- ▶ Premise - division into populations is sufficient to expose the sub types

Methods

Method

- ▶ Our goal is to identify the different clusters by building a tree of classifiers
- ▶ Each node in the tree is a classifier trained to separate between two populations
- ▶ In each classifier, different symptoms will be used in order to classify each patient as population A or population B
- ▶ Each leaf at the final tree represents a cluster (sub - type)

Method

$$P(A | X) > \setminus < P(B | X) \quad \text{st.} \quad X = \{\text{symptoms}\}_{i=1}^m$$



$$\frac{P(X | A) \cdot P(A)}{P(X)} > \setminus < \frac{P(X | B) \cdot P(B)}{P(X)}$$

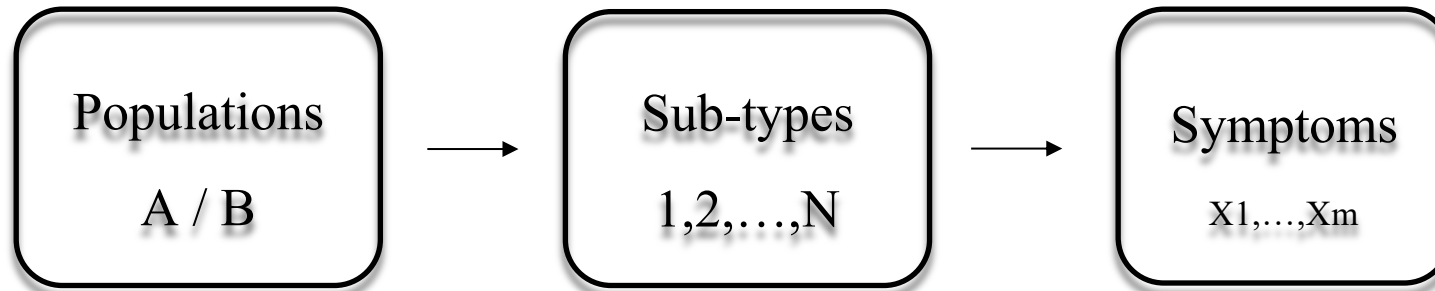


$$P(X | A) > \setminus < P(X | B)$$

Method

$$P(X | A) = P(A) \cdot P(sub - type_i | A) \cdot P(X | sub - type_i)$$

$$P(X | B) = P(B) \cdot P(sub - type_i | B) \cdot P(X | sub - type_i)$$



Method

$$P(X | A) = P(A) \cdot P(sub - type_i | A) \cdot P(X | sub - type_i)$$

$$P(X | B) = P(B) \cdot P(sub - type_i | B) \cdot P(X | sub - type_i)$$



$$P(X | A) \propto P(sub - type_i | A)$$

$$P(X | B) \propto P(sub - type_i | B)$$



$$P(sub - type_i | A) > \setminus < P(sub - type_i | B)$$

Method

- ▶ Given symptoms and two known population, the network can be used to compute the probabilities of the presence of various sub-types

$$P(A \text{ or } B \mid X) \rightarrow P(\text{sub-type}_i \mid A \text{ or } B)$$



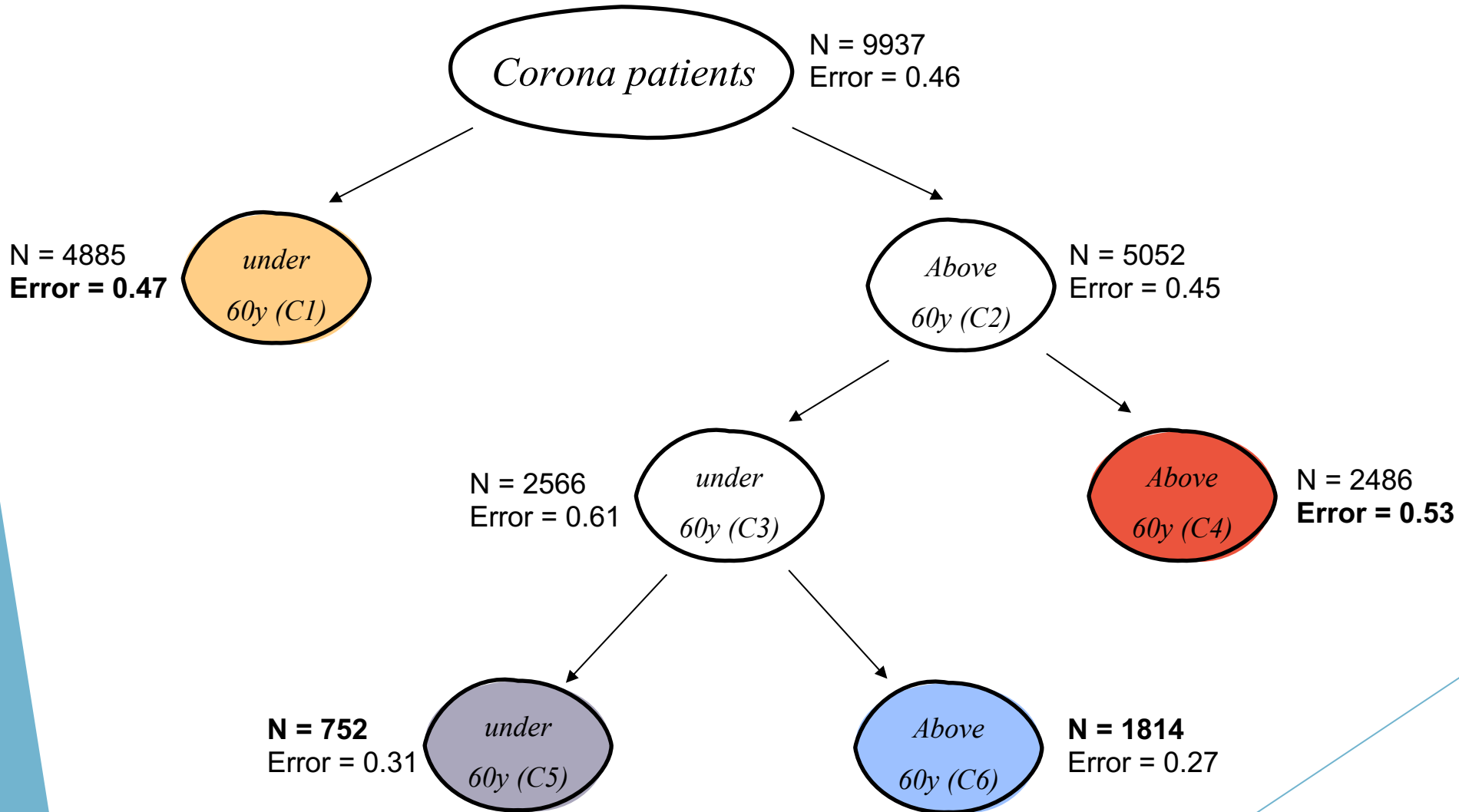
We are dealing with un-supervised problems but using tools of supervised learning

Method - Corona dataset

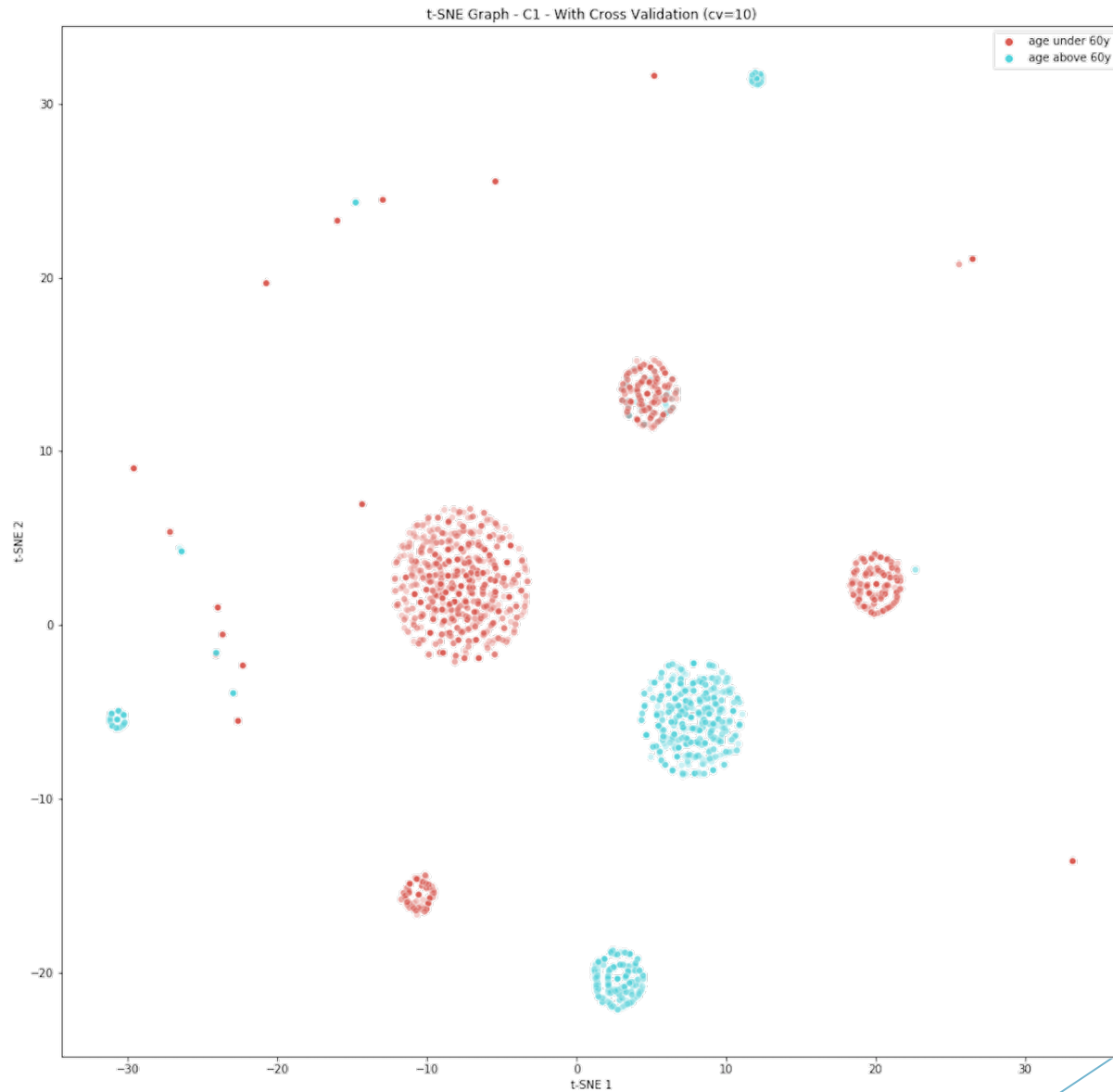
- ▶ We take Corona patients with symptoms like fever, sore throat, shortness of breath etc
- ▶ The patients are divided to two population by age -
 - ▶ Population A - patients over the age of 60 years old
 - ▶ Population B - patients under the age of 60 years old
- ▶ Training the model to create the classifiers tree

Results

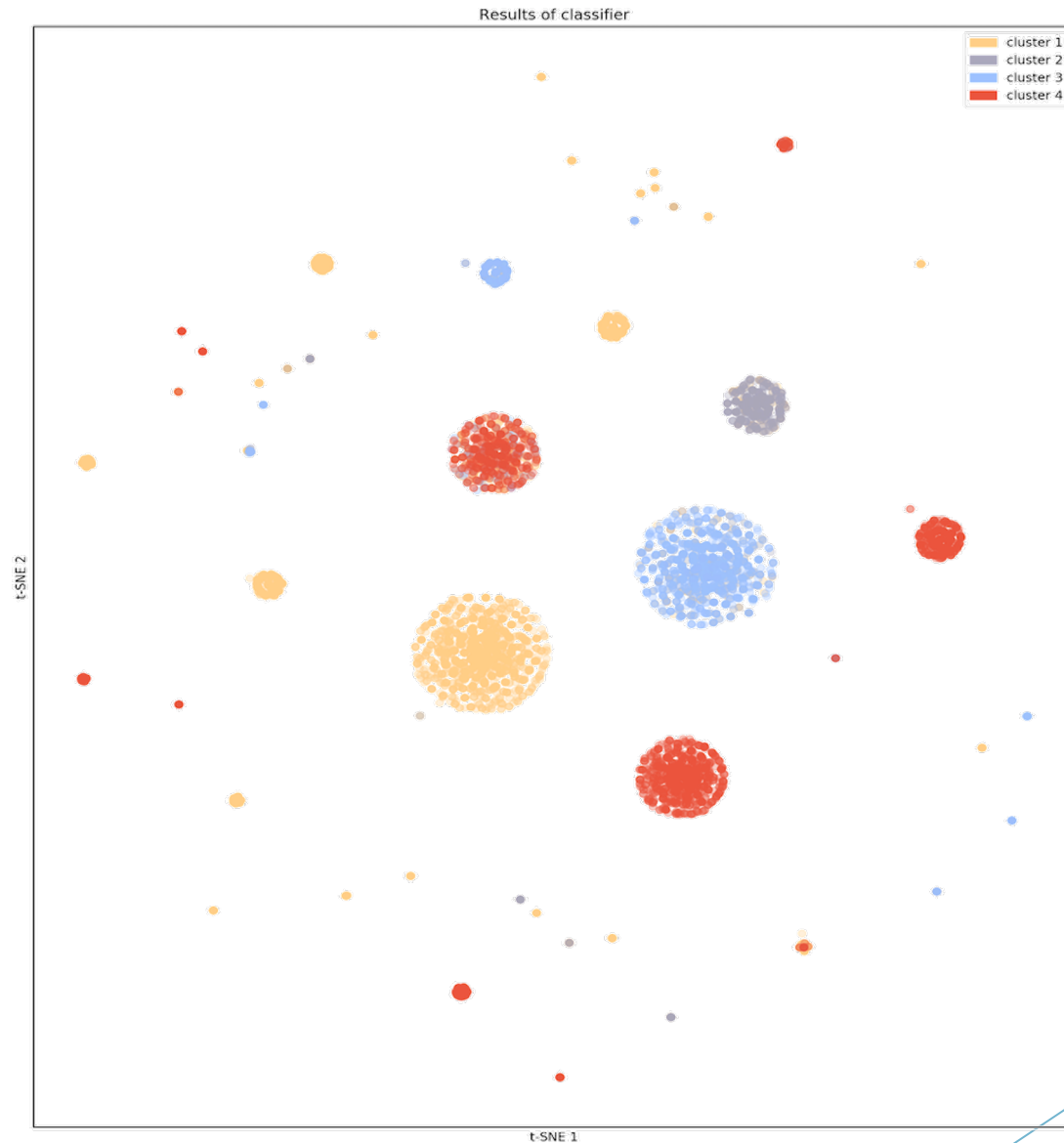
Results - Corona dataset



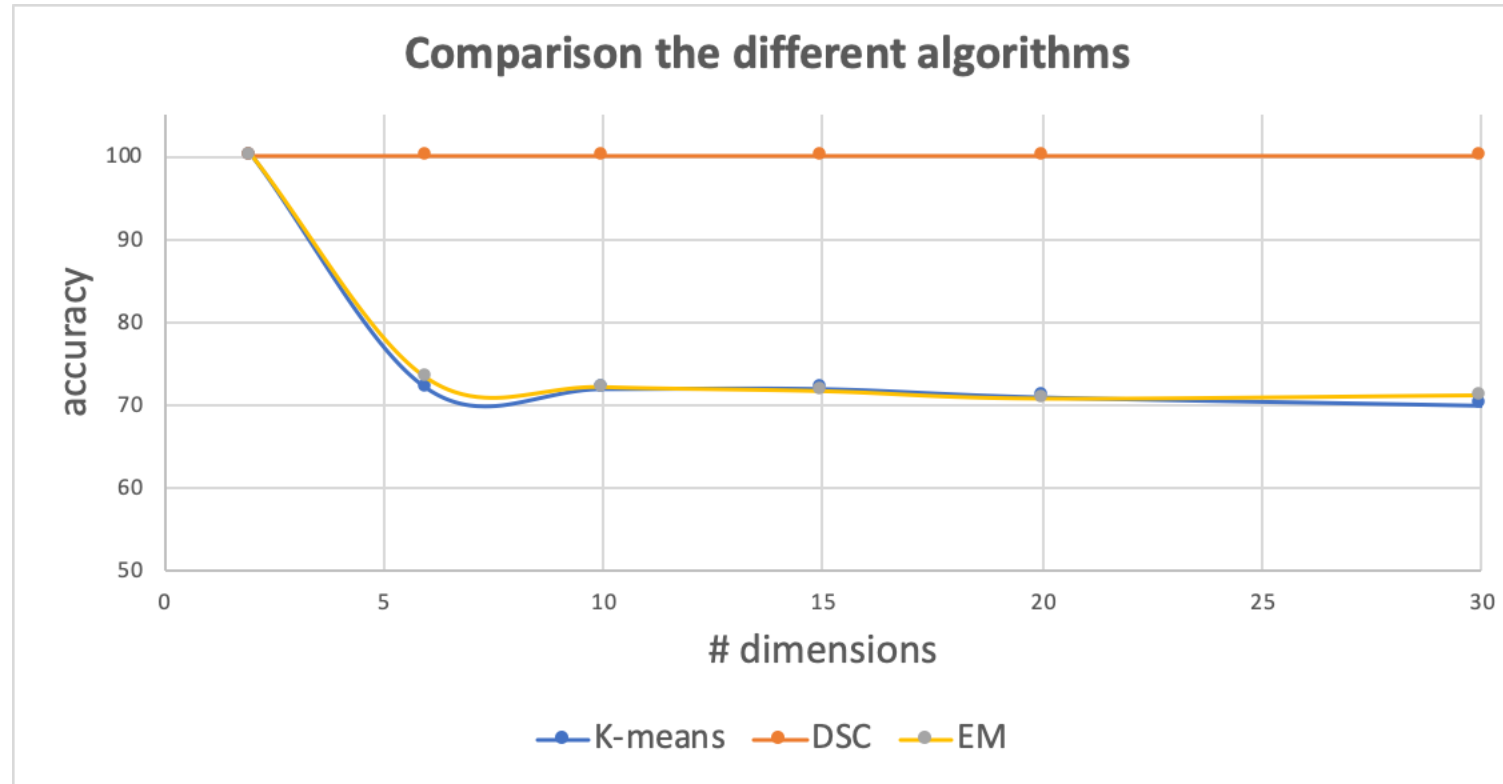
Results - Corona dataset



Results - Corona dataset



Validation of the performance



- There is no need to know in advance the amount of clusters (sub-types)

Conclusions

Conclusions

- ▶ The results indicate that there may be clusters of people who response differently to a disease
- ▶ Our work can provide a good start point for further research into the characteristics of the disease
- ▶ This algorithm can be applied to different data sets of patients to better recover the underlying structure of a disease

Questions ?