Correlation Network and Population Analysis for SLE Patient's Health Monitoring

Ahmed Alabri, Mohammed Alabri

Final Year Project – Fall 2022

B.Sc. in Computer Science

Department of Computer Science

College of Science

Sultan Qaboos University

"A report submitted in partial fulfilment of the requirements for the B.Sc. in Computer Science"

Supervisor: Dr.Hamza Zidoum

Examiner: Dr.Imran Khan

ABSTRACT

Systemic lupus erythematosus (SLE) is a diverse autoimmune disease characterized by mild to life-threatening flares. Severe flares and consequences may necessitate hospitalization, which accounts for most of the direct costs of SLE treatment. Using data from 1160 patients, this study investigates a machine learning algorithm and correlation network to get clusters of SLE patients and analyse these clusters.

In this research, we outline a strategy using correlation networks for analysing and visualizing data from over 1000 SLE patients in the SQU hospital database. We use correlation networks with varied parameters, followed by the Markov clustering algorithm, to evaluate a subset of SLE patients' dataset. “Markov clustering algorithm is fast and efficient algorithm and is designed for undirected and unweighted graphs”. We use the produced sub-network(clusters) to find patterns in this dataset based on: demographic, clinical, laboratory, and therapy information of patients. As a consequence, we built a correlation network of SLE patients using the spearman correlation, and then we used the MCL algorithm to get various clusters, which we subsequently reduced to do some cluster analysis.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: Introduction

1.1 Systemic Lupus Erythematosus

SLE, also known as lupus, is a chronic autoimmune disease that can affect multiple systems in the body, causing a range of symptoms and complications. It is a serious condition that can lead to significant illness and even death [4]. SLE is a chronic autoimmune disease that is known for its complexity and the challenges it poses in terms of diagnosis and treatment. Due to the wide range of symptoms, it can cause and its ability to affect multiple systems in the body, lupus can be difficult to identify and manage effectively. This can make it a frustrating and challenging condition for patients and healthcare providers. The exact cause of SLE is not known, but several risk factors have been identified that are thought to contribute to the development of the disease. These include factors that affect the immune system, such as the production of antibodies and the deposition of immune complexes. This immune system dysregulation leads to organ damage, which is responsible for the wide range of symptoms and the relapsing-remitting nature of the disease. While the specific mechanisms underlying the development of SLE are not fully understood, this immune-mediated process is thought to play a key role in the disease's progression. Only a few drugs are available to manage inflammation and reduce organ damage. Research into the underlying causes of SLE has led to numerous discoveries, increasing the number of drugs available to treat this complex condition [1].

1.2 Research Overview

SLE treatment techniques have advanced significantly in recent years. Nonetheless, despite the improved prognosis, several problems remain in the diagnosis and treatment of SLE. SLE develops gradually and clinically over time; however, a range of illnesses, including viral and hematologic diseases, can resemble SLE characteristics. Database study has shown that individuals with a diagnosis window of fewer than 6 months (between presumptive SLE onset and diagnosis) had lower flare rates and hospitalizations than those with late diagnosis. As a result, SLE is a strong target for ML-based diagnosis approaches that alleviate some of the obstacles associated with early-stage

detection. By employing machine learning-based techniques as clinical support systems for diagnosis [4]. In this study, we provide a correlation network for data visualization and analysis from more than 1000 SLE patients. We use Markov clustering and correlation networks to explore a selection of SLE patients. We use the generated clusters to for monitoring SLE patients and identifying those who need attention. [2].

1.3 Research Objective

The project's goal is to employ the Markov clustering technique after using a correlation network to analyze a dataset of SLE patient data. We use the generated clusters for monitoring and identifying SLE patients who need health care.

ML science is used in many disciplines, even in medical we try to use it int our project to monitor data in easiest way, since it's hard for doctor to compare the hole data to extract pattern. Also, we think it will reduce the harmful of disses since we know how it spread and who. The most important thing is time saving, as we say if doctor try to extract some patterns from patients' data it will take more time instead for that the ML science Is helpful for that.

1.4 Research Methodology

Our planned methodology for this project is as follows:

Figure 1.1: workflow flowchart-diagram

1.4.1 Study systemic lupus erythematosus:

We investigated the etiology, symptoms, and indicators of systemic lupus erythematosus. We also looked at the diagnosis and classification of SLE patients; to study more about this diagnosis to be able to find patterns about the patients.

1.4.2 Data Preparation

The data is now preparing for processing at this step. Data cleaning and data encoding are carried out in two stages here.

1.4.3 Construct the Correlation Network and MCL

The data of 1164 patients are collected as a matrix with 94 features, this called matrix with each row (patient) of the matrix having 94 features, The size of the matrix will be 1164 by 1164 after obtaining a matrix of Pearson's correlation coefficients (correlation matrix) and spearman and will see the different of these features when we apply it to SLE dataset.

In the graph model, each patient is represented as a node. Nodes are only connected by edges if there is a high correlation between the patients they represent. This creates a network of patients, with nodes representing individual patients and edges connecting nodes that are highly correlated. This network allows for the visualization and analysis of the relationships between patients based on their correlations. The previously acquired correlation network is then used to the MCL clustering algorithm in Cytoscape to produce clusters.

1.4.4 Analysing various clusters

In this step, we analyze the clusters that we have obtained and research how to produce better clusters. The number of clusters are obtained by inflation parameter in MCL algorithm, there is no specific relation between inflation parameter and number of clusters, but we do our experiment to the best inflation value which is 1.8 as given in [29]

1.5 Project Plan

Planning must be the first step of each work that the human does it. The project plan includes the main task of what we will do in our project divided

among team members in a detailed timeline as in Figure 1.2A. To do the plan we used the Microsoft Project application. The plan starts by reading about related work and taking a wide background about our project, and then we go to implementation and analyze our models.

GANTT CHART

	Task Mode	Task Name	Duration	Start	Finish	Resource Names
2		Related Work Research papers	6 days	Thu 9/15/22	Thu 9/22/22	Ahmed,Mohamme
3		Machine Learning	2 days	Fri 9/23/22	Sun 9/25/22	Ahmed,Mohamme
4		Correlation Network	2 days	Mon 9/26/22	Tue 9/27/22	Ahmed,Mohamme
5		clusters technique	2 days	Wed 9/28/22	Thu 9/29/22	Ahmed,Mohamme
6		**Progress Report**	**14 days**	**Thu 10/6/22**	**Tue 10/25/22**	
7		First draft	9 days	Thu 10/6/22	Tue 10/18/22	Ahmed,Mohamme
8		Final Draft	5 days	Wed 10/19/22	Tue 10/25/22	Ahmed,Mohamme
9		**implementation**	**49 days**	**Thu 10/6/22**	**Tue 12/13/22**	
10		Data Preperation	3 days	Thu 10/6/22	Sat 10/8/22	Ahmed,Mohamme
11		Correlation network	8 days	Sun 10/9/22	Tue 10/18/22	Ahmed,Mohamme
12		Clusters	18 days	Tue 10/25/22	Thu 11/17/22	Ahmed,Mohamme
13		**Final Report**	**35 days**	**Wed 10/26/22**	**Tue 12/13/22**	
14		First Draft	30 days	Wed 10/26/22	Tue 12/6/22	Ahmed,Mohamme
15		Final draft	5 days	Wed 12/7/22	Tue 12/13/22	Ahmed,Mohamme
16		**submit all deliverable materials of FYP (hard and soft copy)**	4 days	Thu 12/22/22	Tue 12/27/22	Ahmed,Mohamme
17		**Final presentation**	5 days	Thu 12/15/22	Wed 12/21/22	Ahmed,Mohamme
18		**Poster**	6 days	Mon 12/12/22	Sun 12/18/22	Ahmed,Mohamme

Figure 01.2A: GANTT CHART

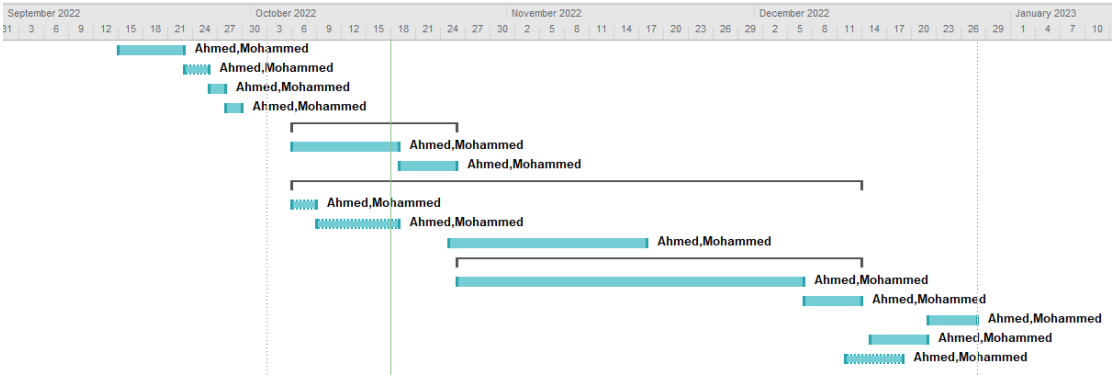

Figure 01.2B: TimeLine of the project

CHAPTER 2: Background

2.1: Correlation Network Model

In recent years, there has been a proliferation of large data sets in the fields of biology and medicine throughout the previous decade. The difficulty is in summarizing, simplifying, and interpreting them. Correlation-based networks can be exploited successfully during this procedure. However, the technique has flaws and necessitates specialized expertise that frequently extends beyond classical biology and comprises a plethora of computer tools and software.

Networks of human, yeast, and plant biological processes have been developed in recent years. Since networks may be seen as graphs. Graphs are a way of representing data as a network of interconnected entities. They consist of a set of vertices, which represent the entities being studied, and a set of edges, which represent the relationships between the entities. Graph theory is a mathematical discipline that studies the properties and structures of graphs, and it has gained popularity in biological research over the past decade. is to determine the degree of linear relationship between all pairs of vectors in the data collection.

2.1.1: Correlation Measures

Correlation measures are statistical techniques used to assess the strength and direction of the linear relationship between two variables.

2.1.1.a Pearson product-Moment

Two variables' monotonic connection between them is measured by correlation. A monotonic connection between two variables is one in which (1) the values of the two variables rise linearly with respect to each other or (2) the values of the two variables decline linearly with respect to each other. When two variables are correlated, their magnitudes might vary in either the same direction or in the opposite way when one changes. One variable's greater value frequently corresponds to the other's higher or lower values and vice versa. A monotonic relationship is a specific example of a linear relationship between two variables [18].

Scatterplots of hypothetical data from bivariate normally distributed data are shown in Figure 2.1, with different Pearson correlation coefficients. When "r" equals 0, this indicates that there is no linear relationship between the variables. As the absolute value of "r" increases, the relationship between the variables becomes stronger, and the scatter of the data on a scatter plot will tend to diminish and cluster more closely around a straight line. When "r" approaches -1 or 1, the relationship between the variables will be perfectly linear, and the scatter plot will resemble a straight line.

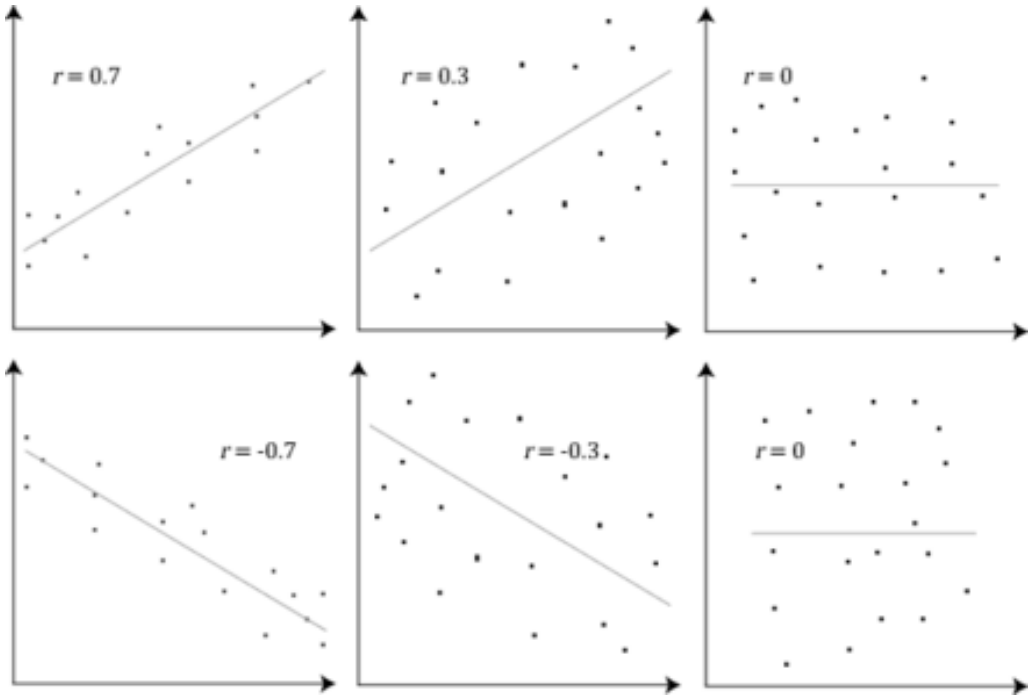

Figure2.1: Example scatterplots of various datasets with various correlation coefficients.

Formula:

$$r = \frac{\Sigma(xi - \bar{x})(yi - \bar{y})}{\sqrt{\Sigma(xi - \bar{x})^2 \Sigma(yi - \bar{y})^2}}$$

2.1.1.b Spearman rank Correlation

The Spearman coefficient expresses the degree and direction of a monotonic association between two continuous or ordinal variables. (rho) or "rs" is a typical abbreviation for a Spearman coefficient. Because ordinal data may be ranked, the Spearman coefficient can be employed. By using rankings, the coefficient measures strictly monotonic relationships between two variables. By rating the data, a strictly monotonic nonlinear connection is transformed into a linear relationship. Furthermore, this trait renders a Spearman coefficient very resistant to outliers [18].

A Spearman coefficient, like the Pearson coefficient, has a range of -1 to +1

Formula:

$$\rho = 1 - \frac{6\Sigma di^2}{n(n^2 - 1)}$$

n = number of observations.

di = difference between the two ranks of each observation.

*ρ = spearman's rank correlation coefficient***.**

2.1.1.c Kendall`s Rank Correlation Coefficients

A statistical technique for determining the link between two variables is called Kendall's Rank Correlation. It is frequently employed when the data is ordinal since it is based on ranks rather than actual numbers. The Kendall tau correlation represents a probability [19].

Formula

Kendall's tau correlation coefficients have the following formulas:

Kendall tau-a (τKen,a) correlation coefficient:

$$\tau Ken, a = \frac{(C - D)}{[\frac{n(n-1)}{2}]}$$

τKen,b (Kendall tau-b correlation coefficient):

$$\tau Ken, b = \frac{C - D}{\sqrt{\left[\left(\frac{n(n-1)}{2-t}\right)\left(\frac{n(n-1)}{2-u}\right)\right]}}$$

τKen,c (Kendall tau-c correlation coefficient):

$$\tau Ken, c = \frac{2(C - D)}{n^2}$$

Interpretation

- If the two ranks agree perfectly and are the same, the coefficient has a value of 1.
- If there is no difference between the two rankings at all, the coefficient is equal to -1.
- Higher values suggest better agreement between the ranks, while the value for all other arrangements varies between -1 and 1.
- Higher values suggest better agreement between the ranks, while the value for all other arrangements varies between -1 and 1.

2.2 Clustering

A cluster is a collection of data elements that are related to one another but not to those in other clusters. It is possible to determine how similar two items are by measuring their distance from one another.

2.2.1 Euclidean Distance

The typical distance between two sites is known as the Euclidean distance. It is the separation between two points as determined by a straight line[20].

$$\sqrt{(\boldsymbol{x1} - \boldsymbol{x2})^2 + (\boldsymbol{y1} - \boldsymbol{y2})^2}$$

Example Using Euclidean Distance Formula

Example: Find the distance between points P(5, 4) and Q(3, 2) Using Euclidean Distance.

Solution:

Given:

P(5, 4) = (x1,y1)

Q(3, 2) = (x2,y2)

Using the Euclidean distance,

$$d = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

$$d = \sqrt{(3 - 5)^2 + (2 - 4)^2}$$

$$d = \sqrt{(-2)^2 + (-2)^2}$$

$$d = \sqrt{8}\, units$$

$Answer$: Euclidean distance between $\mathrm{P}(5, 4)$ and $\mathrm{Q}(3, 2)$ is $\sqrt{8}$ units

2.2.2 Manhattan Distance

When two places are separated by a distance that can only be computed by travelling along grid lines, the distance is known as the Manhattan distance.

$$d = |x1 - x2| + |y1 - y2|$$

Example of Manhattan Distance

Example: Find the distance between points P(5, 4) and Q(3, 2) Using Manhattan Distance.

Solution:

Given:

P(5, 4) = (x1,y1)

Q(3, 2) = (x2,y2)

Using the Manhattan distance,

$$d = |\ 3 - 5\ | + |\ 2 - 4\ | = 2 + 2 = 4$$

Answer: Manhattan distance between P(5, 4) and Q(3, 2) is 4 units

2.2.3 Cosine Index

A metric for comparing two vectors is the cosine index. To figure it out, divide the two vectors' dot products by the sum of their magnitudes. The cosine index is, thus, equal to the sine of the angle formed by the two vectors [20].

$$\theta = arccos\ (A.\ B)\ /(\|A\|\|B\|)$$

Example of Cosine Index

D1 = [1,1,0,1,1,0,0]

D2 = [0,1,1,0,1,1,0]

D1. D2 = 1 * 0 + 1 * 1 + 0 * 1 + 1 * 0 + 1 * 1 + 0 * 1 + 0 * 0 = 2

$||D1|| = \sqrt{(1^2 + 1^2 + 0^2 + 1^2 + 1^2 + 0^2 + 0^2)} = 2$

$||D2|| = \sqrt{(0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 1^2 + 0^2)} = 2$

Similarity (D1 ,D2) = (D1 . D2) / (||D1|| * ||D2||) = 2 / (2 * 2) = 0.5

Cos(θ) = 0.5

θ = arccos(0.5) = 60

2.2.4 Clustering Techniques

Clustering is a data mining and machine learning technique that divides a bunch of data points into smaller groups (called clusters) based on their feature similarity. Consider uploading images to a social media platform. The site may try to group pictures of the same individual to arrange your photos. However, the site has no idea which photographs depict whom, or how many different persons feature in your photo collection. A logical strategy would be to take all of the faces and group them into groups of faces that seem alike. Hopefully, these photographs belong to the same person, and they can be put together for you [3].

2.2.5 Clustering algorithms in machine learning for Healthcare Data

Clustering algorithms can assist find groupings of patients with similar features or outcomes by investigating linkages and patterns in healthcare data.

Clustering algorithms are usually important in predicting diseases since they separate similar patient data based on key criteria [8].

2.2.6 Markov Cluster Algorithm (MCL)

For finding cluster information in graph networks, Markov clustering is a crucial bioinformatics technique. Recently, MCL, which was developed for broad graph clustering, has been used in a variety of bioinformatics applications. MCL has gained popularity as a method for extracting complexes from interaction networks because it is efficient, quick, and frequently more tolerant and resilient to noise [6]. In order for the MCL algorithm to function, the network must first be represented as a matrix, with the nodes' rows and columns serving as their nodes and their entries as their nodes' connections' strength. The algorithm then performs a number of matrix transformations, such as inflation and expansion, to gradually amplify connections between nodes in the same cluster and suppress connections between nodes in different clusters. The network's node clusters are then located using the resulting matrix. The MCL algorithm is a potent tool for discovering clusters in large and complicated datasets due to its mix of scalability, robustness, and adaptability.

Social network analysis is one application of the MCL algorithm. Edges here reflect the connections between people, such friendships, whereas nodes here represent the individuals. The MCL method can be used to locate groups of people who are highly connected to one another, possibly indicating that they are a part of the same community.

CHAPTER 3: Proposed Method

In this chapter, we have made a complete plan to complete our project. Where the proposed plan helps us to work step by step and know the path and the results we will get. First, we filter the SLE patients' data. Then, using the data, we build a correlation network and apply the MCL algorithm. The result will be clusters. We take these clusters, analyze them, compare them, and look for similarities and differences between them.

3.1 Pipeline

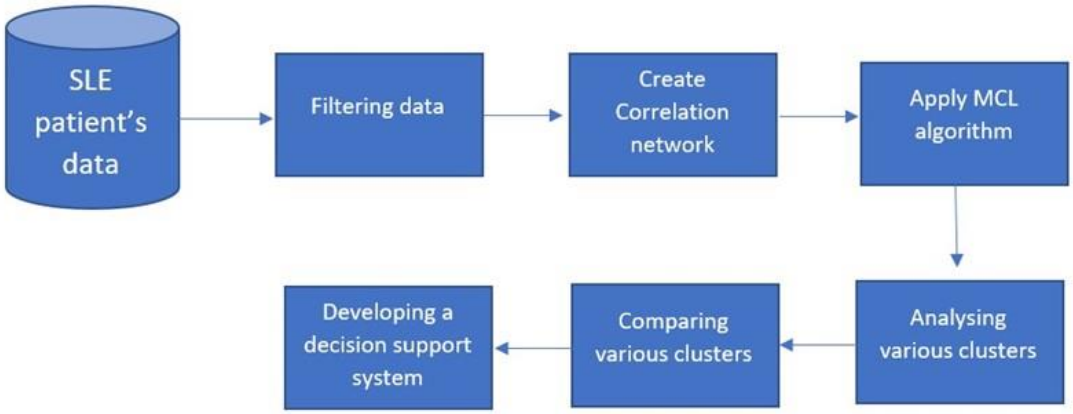

Figure 3.1: Method Architecture pipeline

When filtering data, we remove patients with incomplete information and encode categorical features into numerical ones. Then from this filtered data, we created a correlation network using the spearman correlation coefficient this correlation network represents patients as nodes and the relationship between them as edges. After using the correlation network, we used the MCL algorithm to generate many different clusters, from which we chose the top three for analysis. Finally, we examined clusters based on some key features in cluster analysis. Then we compare clusters to see if there are any similarities or differences.

CHAPTER 4: Literature review

In Oman we have Scarcity in clustering report for systematic lupus erythematosus, but in Arab region we have many, the one of them is (helaly and Mansour, 4.2018) scientific research that published it in the Egyptian journal of hospital medicine.

Researchers in (helaly and Mansour, 4.2018) paper collect 150 Egyptian adult patients form specific hospital. Three different clusters of patients were identified based on some features, he using k-means cluster analysis. The first cluster is the heights cluster with age of disses, patients in cluster 2 is highest with renal and haematological tests, the last cluster 3 is has the most varied characteristics. [31]

Authors in (Fuchsberger and Ali, 2017); (Chetti and Ali, 2020) collected data of 600000 bridges from Federal Highway Administration (FHA). in paper and (Chetti, P. and Ali, H. (2020)) the collected data are for 25 years from 1992 to 2016. in both papers the authors try to inspection the frequencies of bridges health, for that purpose they are used the correlation network and clustering algorithms which we will use it in our project.

Author in (Fuchsberger, A., and Ali, H., 2017) doesn't decide which algorithm they are used to get cluster, but they found 2 cluster with respect to some features deck condition, bridge age and sufficiently.[25]

In other side authors in (Chetti, P. and Ali, H. (2020)) they are used MCL algorithm to get the clustering, they didn't decide the number of clusters they found, but they take hight 5 cluster shave nodes and do different experiment to it.

The main different between the dataset of the last two papers compared to our dataset the we use it is they have sufficient rating (SR). the SR range from 0 to 1000, the bridges with low SR are bad condition.[2]

CHAPTER 5: Dataset Description

The dataset used in this study comes from the Sultan Qaboos University Hospital It contains 1161 rows called samples and 94 variables called features. It includes 1161 Omani SLE patient records that satisfied EULAR's admission criterion. The dataset contains demographic, clinical, and laboratory data [4]. Data are described it in table [1], each feature with its meaning.

To comprehend the general characteristics of the population, we visualized the demographic distribution of our cohort. In figure 5.1 the most of patient age is between 32 and 36 years, And the lowest group of patient age was in from 60 and upward.

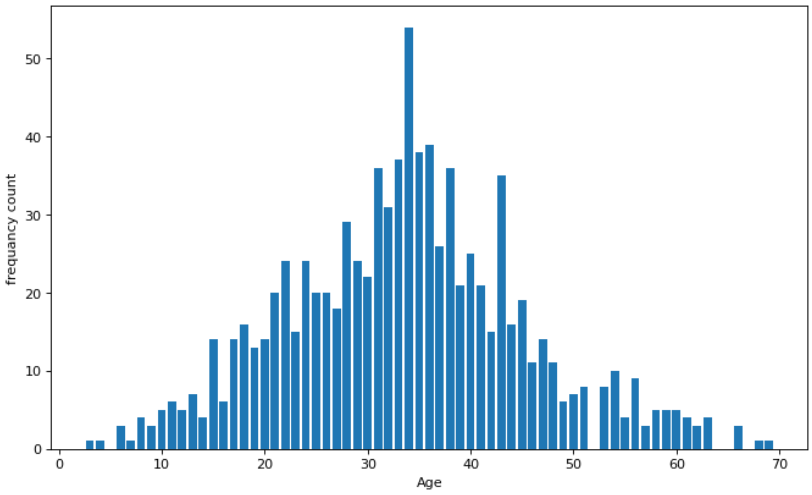

Figure 5.1: The patient Age Histogram

The gender is another similarity distribution; the female is a most patient gender in dataset since represent 88% figure 5.2. Researchers believe that the sex hormone and/or gonadotropin-releasing hormone metabolism in women is to blame for the rise in frequency (GnRH). SLE usually progresses more severely in men, despite their rarity. Tribal distinctions can be seen in how lupus manifests in Omani Gulf Arabs. [22], Figure 5.3 shows the most 10 Tribal that infected by SLE in the dataset, “Al Balushi” is The Most Tribal distributed in our database.

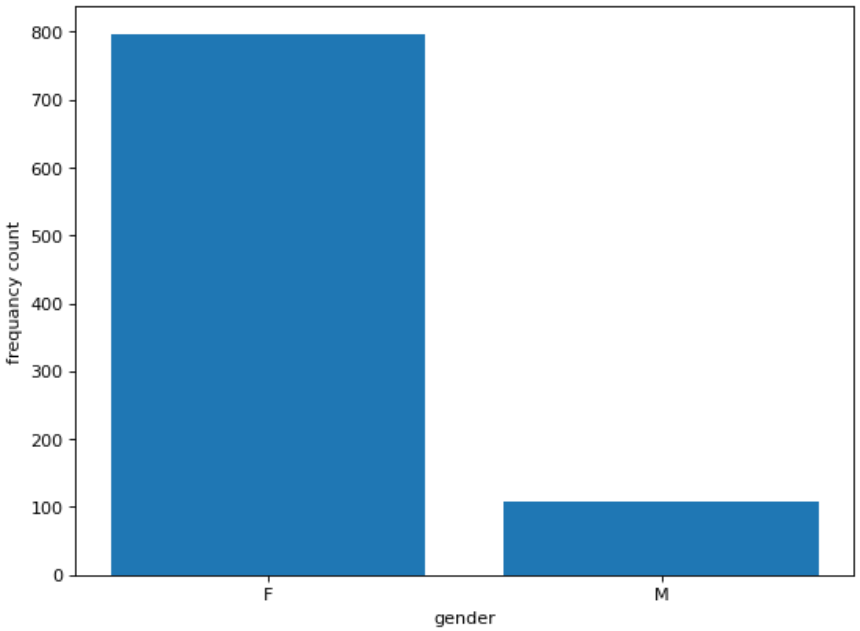

Figure 5.2 Gender Distribution in Dataset

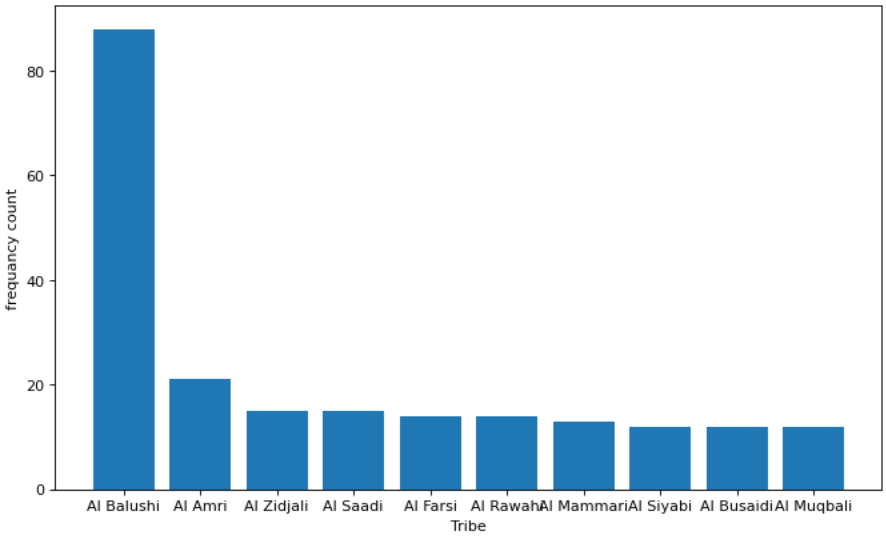

Figure 5.3: The Destitution of Most 10 Tribal in Dataset

5.1 Dataset preparation

There are three processes here: (a) removing irrelevant data, (b) picking an encoding scheme, and (c) data normalization. Data preparation techniques include those associated with processing and analysing raw data to produce quality data. The efforts conducted to improve the data representation are described below [4].

5.1.1 Data Cleaning

Data analysis, Machine learning and Artificial intelligence (AI), these discipline and other disciplines which the Dataset is very relate with the result, so the quality is very important since the data is impact the result. So, the data cleaning is very important process to thesis disciplines, and its mean ensure the dataset is contains no redundant values or inappropriate value.

Analysing the data to find any errors or discrepancies in the database is the first step in the data cleansing process. In other words, this stage is called data auditing, and it's in charge of finding any anomalies in the database. Furthermore, Data analysis will be used to collect metadata about data qualities in order to identify data quality issues. When analysing data, there are two approaches: profiling and mining of data. The study of individual attribute instances is the focus of data profiling. Data mining, on the other hand, focuses on finding certain data patterns in huge databases. The first phase yields an indicator of every probable abnormality that happens within the database [14].

Following that, A set of data processes are described in the transformation workflow for the discovery and elimination of anomalies. To find out about current anomalies, data analysis is specified. The quantity of sources, level of heterogeneity, and "dirtiness" of the data all affect how many transformation stages are required [14].

The third step is the stage of verification. This phase involves evaluating the accuracy and efficiency of the transformation workflow. This stage involves numerous iterations to make sure all errors have been fixed [14].

Following the evaluation and validation of the data, the transformation phases will be taken in order to update warehouse data. To support data quality, it is necessary to document in-depth details on the transformation process. Finally, after removing all errors, Replace the filthy data with the cleansed data [14].

After cleaning schema we get a cleaning dataset without any inconsistent information , inconsistent information like typo Error or missing to input some information , For Typo Error we try to figure out what the person who Enter the information mean : like if we found 1 in columns that that should have Yes we replace the 1 by Yes for example, Another problem is that some columns have they have one choice for the patient who have this column feature ,What we do is replace the empty cell by No choices. After we fix the dataset, we still have some information that is missing and there is no way to fill it, so we use $dropna()$ function in panda library to remove the hole patient who have this missing “information”.

5.1.2 Encoding Scheme

Data might have numerical values or qualitative, i.e., categorical values; one of the key aspects of SLE data is that the majority of the data is categorical; building statistical models on such data to extract useful information often requires a numerical representation of all entries. Encoding is the process of converting data into numerical values. For datasets with low-cardinality features (the number of unique values inside a single feature), the usual strategy is to use an indication encoding method [4].

Indicator encoding is a method that change the qualitative values to integer values based on the order of mapped values. For instance, if a dataset contains a categorical variable with values chosen from the set "Positive", "Negative", and "Not available", label encoding may assign the mapped values from the set "0, 1, 2", respectively. The key disadvantage of this method is that, due to its implicit ranking based on the order of mapped values, it works slightly better for ordinal values. The second method is One-

Hot encoding, one-hot encoding, is a method for encoding categorical data in a machine learning model. In this method, each category is represented as a binary vector, with a "1" in the position corresponding to the category and "0" in all other positions. For example, if a dataset has three categories ("Positive", "Negative", and "Not available"), each category would be represented as a binary vector of length 3. Category "Positive" would be represented as [1, 0, 0], category "Negative" as [0, 1, 0], and category "Not available" as [0, 0, 1]. This method is often used when working with categorical data, as it allows the machine learning model to easily distinguish between the different categories and make predictions based on that information.

Target-based encoders are a type of encoding scheme used in machine learning to represent categorical data. In target-based encoding, the categories are encoded based on their relationship to the target variable. This means that the categories are encoded in a way that reflects how well they predict the target variable. For example, if the target variable is whether or not a customer will purchase a product, the categories might be encoded based on the likelihood that a customer in that category will make a purchase. This can help the machine learning model to more accurately predict the target variable based on the encoded categorical data. Target-based encoders are often used in supervised learning settings, where the relationship between the input data and the target variable is known. [4].

In our data set most of columns have the two choices Yes or No, so we replace Yes by 1and Replace No by 1. The other thing is a gender, also we put 1 for Male and 0 for Female, since the tribal is very important we try to categories the patient with his tribal and we get 369 different tribal.

CHAPTER 6: Implementation and Tools:

For implementation and tools chapter we will describe the tools that we used to in our project and how the we use it to derive our project result.

6.1 Cytoscape

Cytoscape is an open-source software platform for visualizing molecular interaction networks and biological pathways and fusing these networks with annotations, gene expression profiles, and other state data. Although Cytoscape was first created for biological study, it currently serves as a universal platform for sophisticated network analysis and visualization. A fundamental set of capabilities for data integration, analysis, and visualization are offered by the Cytoscape core distribution. Apps with additional functionalities are accessible (formerly called Plugins). There include apps for network and molecular profile analysis, new layouts, extra file format support, scripting, and database connectivity [16].

6.2.1 How Cytoscape works

To download the program, Enter the Cytoscape.org website and thin press on download as figure 6.1, and thin from download folder we find the setup file figure 6.2, double click on it to install the program as shown in figure 6.3.

Figure 6.1: Cytoscape.org

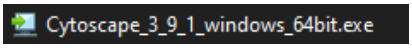

Figure 6.2: the Cytoscape after download

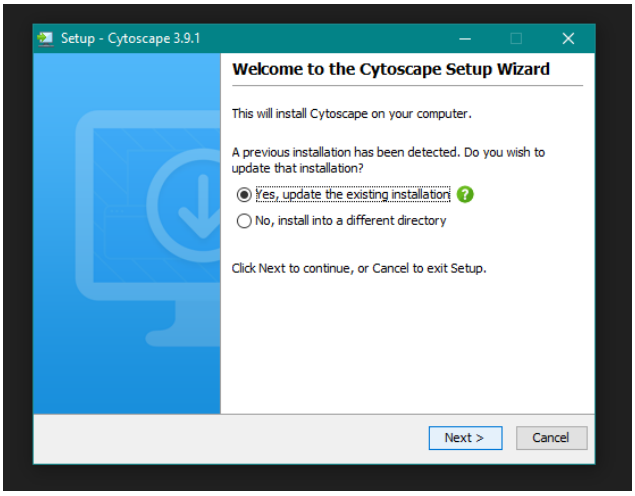

Figure 6.3: install Cytoscape

"Communication in a user-friendly manner. It has a number of data filters that demonstrate its compatibility with other programs." It is a highly interactive tool that allows the user to navigate the network by zooming in and out. It supports 2D representations and is appropriate for large-scale network research involving thousands of nodes and edges. With the assistance of a network manager, several networks may be effortlessly organized [24].

6.2.2 Cytoscape: Network Visualization

Following the import of correlation network, after we create the correlation matrix using python, we upload the matrix in Cytoscape as CSV based file, in Cytoscape we have an apps actually called plugins in Cytoscape that do specific function, in the beginning we start using the MetScape but we didn't arrive to our goal which is a correlation network, since this app is not appropriate for our data. After that we find another plugin which is aMatReader which read the data as a matrix after we got the correlation network, we apply MCL algorithm figure 6.4. We use Clustermaker plugin to use MCL algorithm.

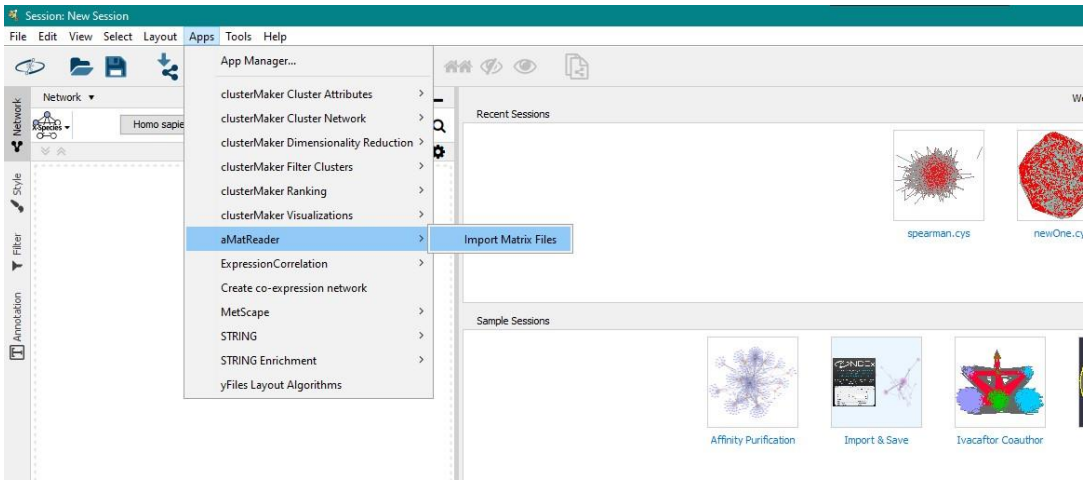

Figure 6.4: The aMatReader app

6.2 Jupiter notebook

The Jupiter Notebook is a web-based platform, it's integrated the code with its output. It's provided the notebook as cells and you can run each cell separately from each cell.

6.2.1 How we use Jupyter

We used Jupyter to clean our data set and to visualise the graphs. in begging we use python languages du to simply used and friendly, in python we have libraries which is a collection of related modules. so, we start to import these libraries: pandas library used for imports data and visualize it as a data frame, matplotlib library which is used to display figures, NumPy library which is a math library.

After we prybar the notebook by import the libraries, we start to clean our data by removing the duplicate features as "Kidney Biopsy LN Class" and try to fix the redundant values for example in some patient we should have later N in the begging but we find he have 1 for like or 0 and this depend to the assigned value in table 3 which describe each number refers to Y or N, after that we decode the values to manipulate with it: since machine learning play with numeric values. we but 0 for no or negative and 1 for yes or positive. the drop data is very important to drop the patients how have missing data, and for that we use $dropna()$ Algorithm form $pandas$ library. in the last of thing, we compute the correlation between patients, each patient with other patients, before we drop the patient how have nan values, we have 1164 patient and after we drop the patient 1054 remains with us. after applying correlation, we get a matrix of 1054*1054 dimension.

After that we save the file of matric correlation and upload it in Cytoscape that describe it above to extract the correlation network and we applying Markov clustering algorithm the gives us the cluster to study it.

CHAPTER 7: Experimental Result

In this chapter we will present our research experiment result. We will provide different experiments with respect to different properties like Demographic features, clinical features and immunology features.

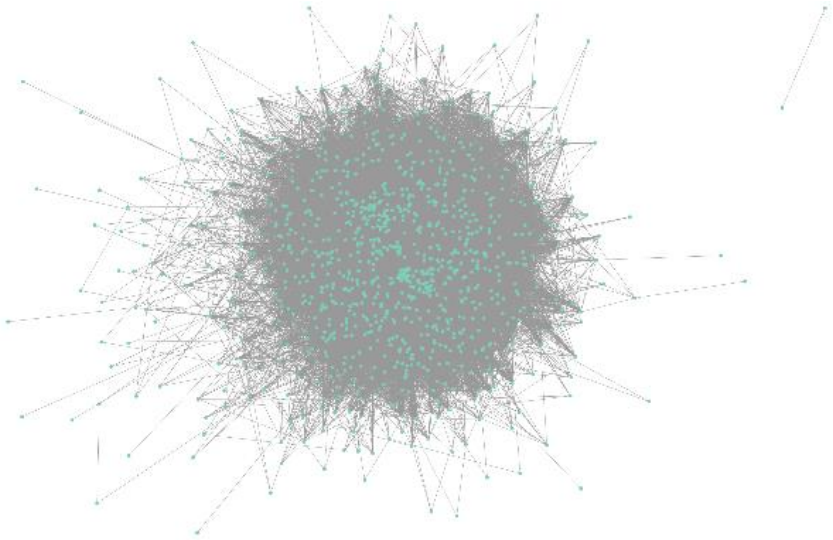

Figure 7.1: Correlation network (correlation p>0.9) with 977 nodes, and 106234 edges (average degree = 217.914, and 2 connected component)

7.1 Network properties of top 5 clusters

The correlation network with correlation coefficient p>0.9 is presented with 977 nodes and 106234 edges and 2 connected component (number of subgraphs) (90) are presented in Figure (7.1). To extract this figure, we begin start with expression correlation plugin in cystoscope, but since it doesn't appropriate for our data, we used aMatReader plugin, which is read a correlation matrix. We do the correlation matrix in python using spearman correlation measure and then we upload it to cystoscope. early we used a Pearson correlation measurement and then we change to spearman correlation duo to data; when we apply a Pearson correlation measurement to our data, the most of correlation coefficient between samples greater than 0.9, which is mean we don't get verity in correlation coefficient cambarid to spearman correlation measurement, also as the correlation coefficients r-value are the determining elements in CN construction. and we think this is due to Rank of patient in each feature in spearman measurement.

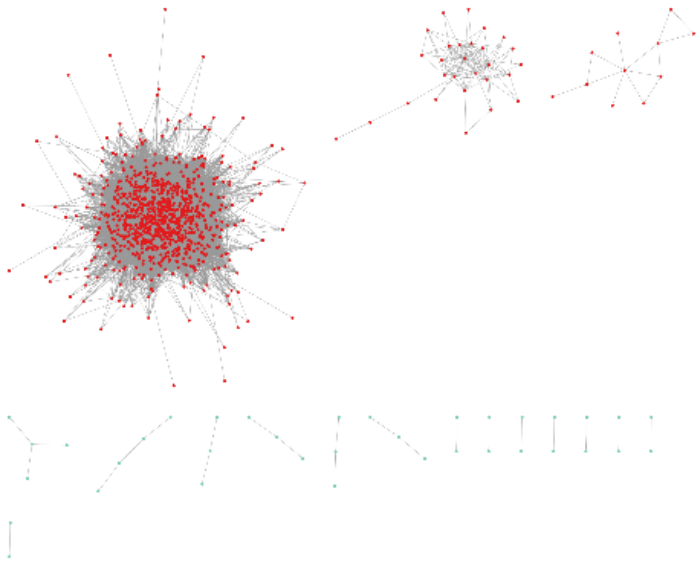

Figure 7.2: Top 3 clusters (red coloured clusters) produced by MCL algorithm

Figure (7.2) presents the top 3 cluster (red coloured cluster) generated by MCL algorithm. Table (comparison between three clusters) shows us the statistics of these 3 clusters, first cluster have the heights number of nodes, with 902 nodes and 105298 edges. The smallest cluster have only 11 nodes and 14 edges. the average degree of the clusters refers to the "total amount of incoming flow" [27]/ the number of edges. "The average clustering coefficient (cl) is calculated it for each node as the number of links between its neighbours over the number of possible links" [28]. the cluster has higher clustering coefficient from the top 3 clusters is first one, these top 3 clusters are in table 1, will be considered for further analysis. the network density in table 1, refers to "description of the potential number of edges present in the sub-network compared to the possible number of edges in the sub-networks" [2]. The heights cluster with density network for top 3 clusters describes in table 1, is second cluster.

Table 1: Network Statistics of top 3 cluster produced by the MCL Algorithm

	# Node	# Edge	Avg. degree	Network density	clustering coefficient
1	902	105298	233.5	0.259	0.675
2	28	105	7.5	0.278	0.658
3	11	14	2.545	0.255	0.498

7.2. Population analysis of clusters with respect to Demographic features

Specific population characteristics are referred to as demographics. In this section we will see a set of charts that show clusters with respect to Demographic features.

Figure 7.3 shows the classified ages of SLE patients in each cluster, in cluster 1, the most affected age group is middle-aged adults with 443 cases, followed by young adults with 268 cases, old adults with 123 cases, and children with 67 cases, and babies come in last with one case. In cluster 2, we notice that there is a change in the ages of most people with the disease, as the young adult became the most infected group with 11 cases, followed by middle-aged adults with 10 cases, old adults with 6 cases, and children with one case. With 8 cases, young adults make up the majority of cluster 3's infected individuals, in stark contrast to the other age groups; children come in second with two cases, and old adults come in last with one case.

Figure 7.4 show the classified age at diagnosis of SLE patients in each cluster, in cluster 1, the most affected age group is young adults with 429 cases, followed by children with 229 cases, middle-aged adults with 129 cases, and old adults with 46 cases, babies come in last with 8 cases. In cluster 2, we notice that there is a change in the age of diagnosis of most people with the disease, as young adults and children became the most infected groups with 9 cases, followed by middle-aged adults with 8 cases, old adults with 2 cases, and babies with no case. With 6 cases, children make up the majority of cluster 3's infected individuals, in stark contrast to the other age groups; young people come in second with 4 cases, and middle-aged adults come in last with one case.

Figure 7.5 show number of SLE patients according to the duration of the disease in each cluster, in cluster 1, the disease duration of more than 10 years is the most prevalent with 347 cases, followed by the period between 0–5 years with 303 cases, and finally the period between 5–10 years with 252 cases. Cluster 2 is slightly different from Cluster 1, as the duration of the disease ranges from 0–5 years and is the most prevalent, and this indicates

that in this cluster the duration of the disease is less than in Cluster 1, where the duration of the disease was recorded between 0–5 years, about 13 cases, followed by a longer duration from 10 years with 9 cases, and then the period between 5 and 10 years with 6 cases. In Cluster 3, there is an equal number of cases of disease that lasted from 5–10 years and 0–5 years, with two cases, while there are seven patients whose disease lasted longer than 10 years.

Figure 7.6 show number of SLE patients according to the gender in each cluster, in cluster 1 The number of females is about eight times the number of males, in cluster 2 The number of females is about four times the number of males, in cluster 3 The number of males is greater than the number of females, we notice that Cluster 3 has the highest percentage of males compared to females, followed by Clusters 2 and 3.

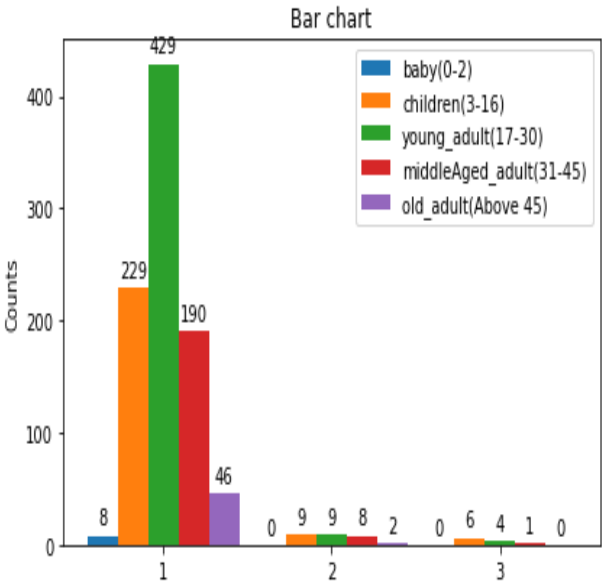

Figure 7.3: Age

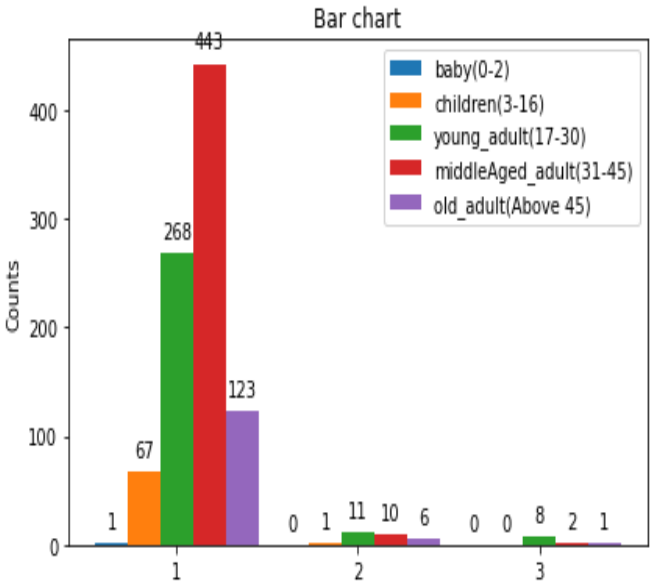

Figure 7.4: Age at diagnosis

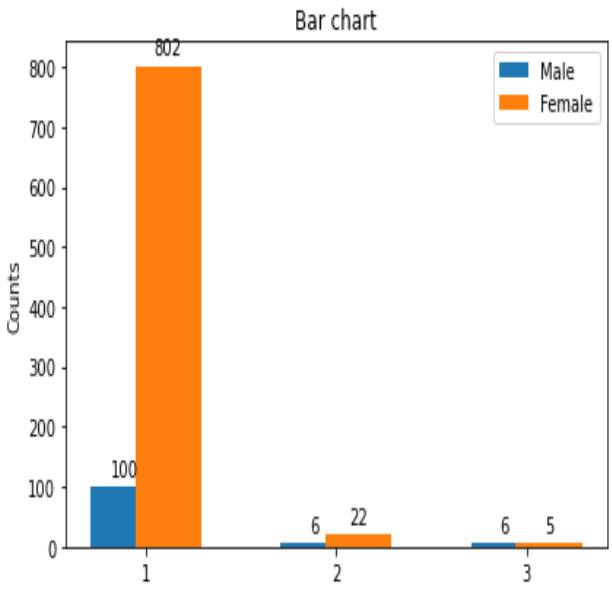

Figure 7.6: Gender

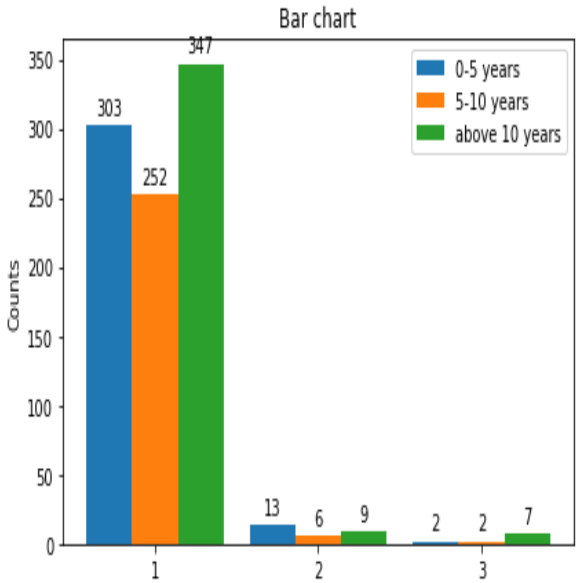

Figure 7.5: disease duration

7.3 Population analysis of clusters with respect to Clinical Features

In this section, we will do a population analysis of clusters with respect to some important clinical features (Discoid Rash, Malar Rash, Alopecia, Nephritis).

Figure 7.7 count how many SLE patients have a discoid rash and how many do not have a discoid rash in each cluster, we see from all clusters that most SLE patients do not have a discoid rash.

Figure 7.8 count how many SLE patients have a malar rash and how many do not have a malar rash in each cluster, we see from all clusters that most SLE patients do not have a malar rash.

Figure 7.9 count how many SLE patients have alopecia and how many do not have alopecia in each cluster, we see from all clusters that most SLE patients do not have an alopecia rash.

Figure 7.10 count how many SLE patients have nephritis and how many do not have nephritis in each cluster. We see that in clusters 1 and 2, the number of non-nephritis patients with SLE is slightly higher than the number of patients with nephritis, while in cluster 3, the number of people with nephritis among SLE patients is much higher than the number of those without nephritis.

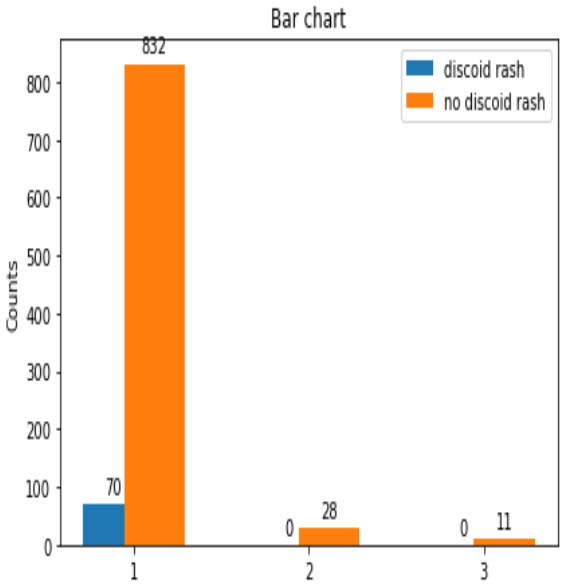

Figure 7.7: Discoid rush

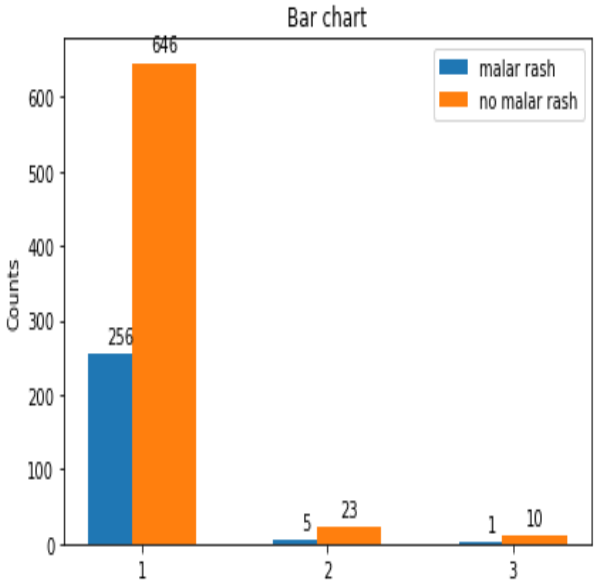

Figure 7.8: Malar rush

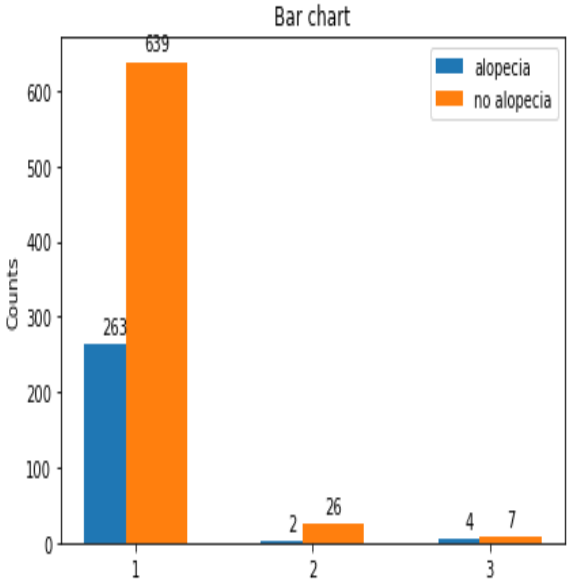

Figure 7.9: alopecia

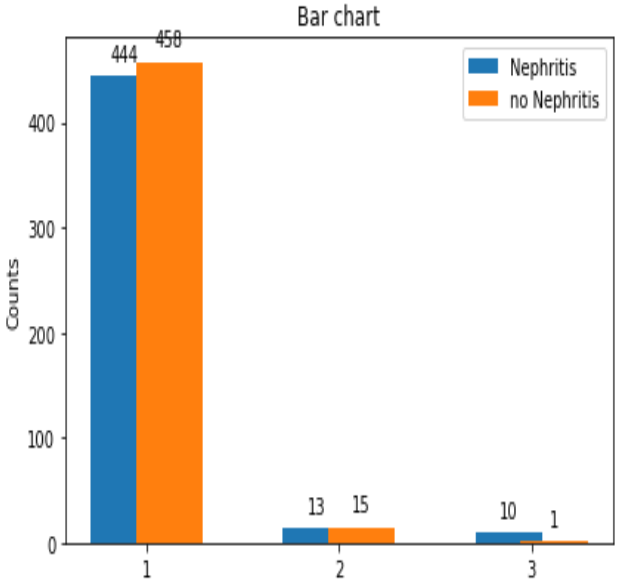

Figure 7.10: Nephritis

7.4 population analysis of clusters with respect to Immunology Features

The immunology features that are considered for population analysis are: Direct coombs test, Anti-Nuclear Antibody test, Anti-dsDNA test, Anti phospholipid antibody, low complement, Anti-sm. these sex different immunology features are compared with respect to their Mode values in each cluster in top 3 clusters as figure 7.11 display, we used the Mode value instead of mean because of our dataset values is discreate not continues. from figure we find that the clusters are different in Direct coombs test but are same in other tests.

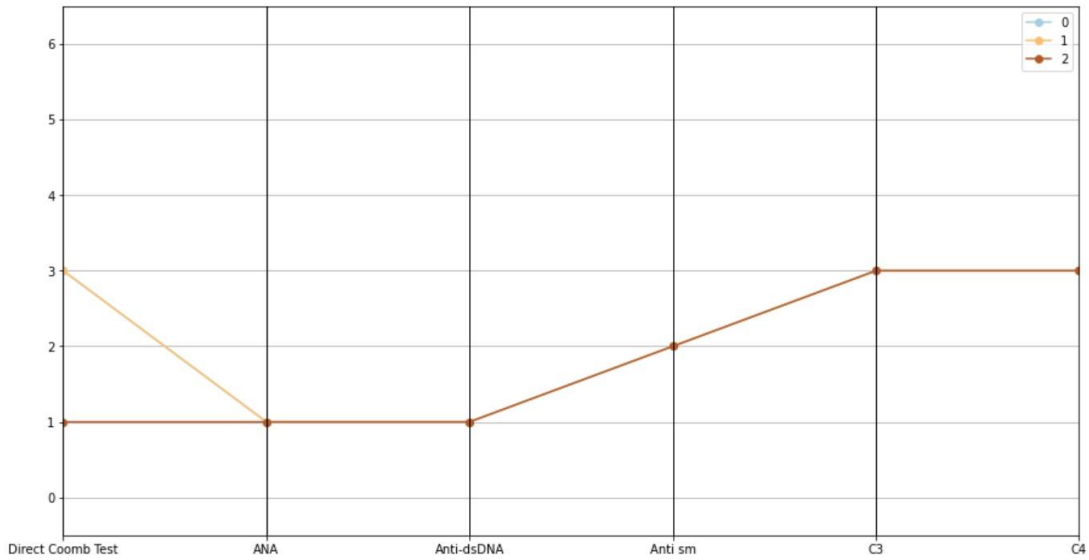

Figure 7.11: compaction for top 3 clusters

Figure 7.12 through figure 7.17 demonstrate the enrichment of each cluster`s immunity rating values. Figure 7.13 demonstrate that the cluster 1 and are highly enriched with not available test in direct coomb's test. Figure 7.12 and 7.17 shows the testing rating of ANA and Anti-dsDNA, respectively. From these two figures we can see that the clusters are highly enriched with positive rating compared with other result, hence these two testing play significant of this disease. From figure 7.17 we see the 3 clusters are highly enriched with negative result from Anti-sm test. figure 7.14 and 7.16 demonstrate the complement testing c3 and c4, respectively. From these two figures we can see that the clusters 2 and 3 are with same numbers of patients in each test complement which is different to cluster 1. In public all clusters enriched with not available result in both complement testing.

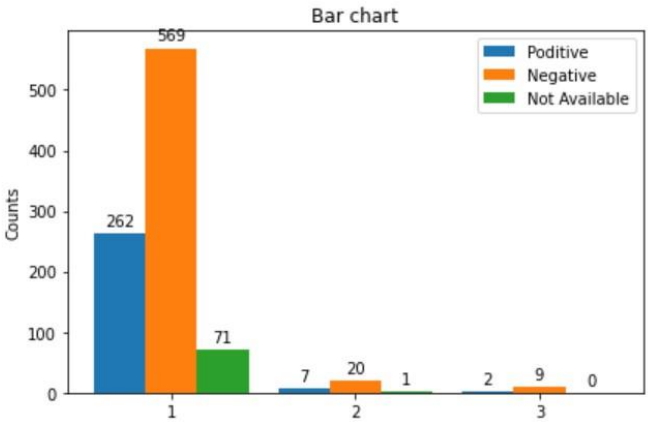

Figure 7.12: Anti sm

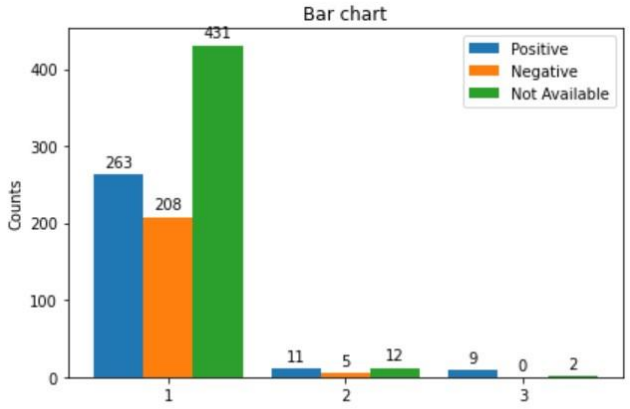

Figure 7.13: Direct Coomb Test

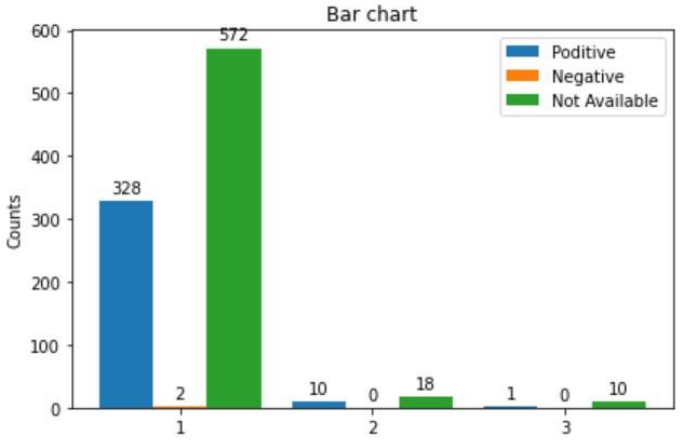

Figure 7.14: C3

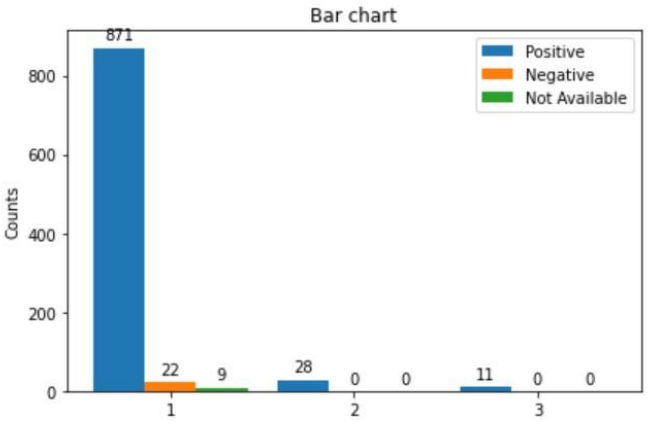

Figure 7.15: ANA

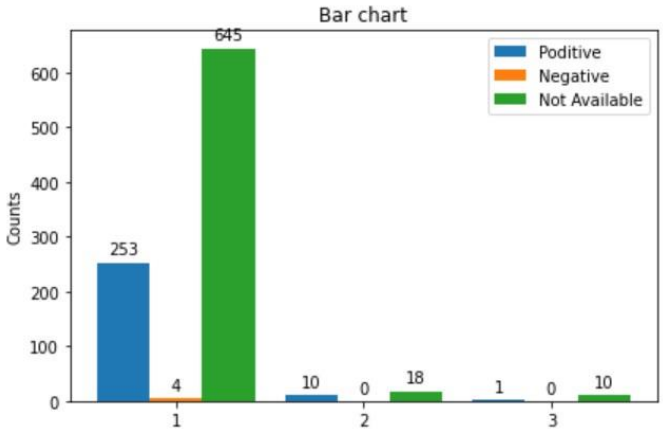

Figure 7.16: C4

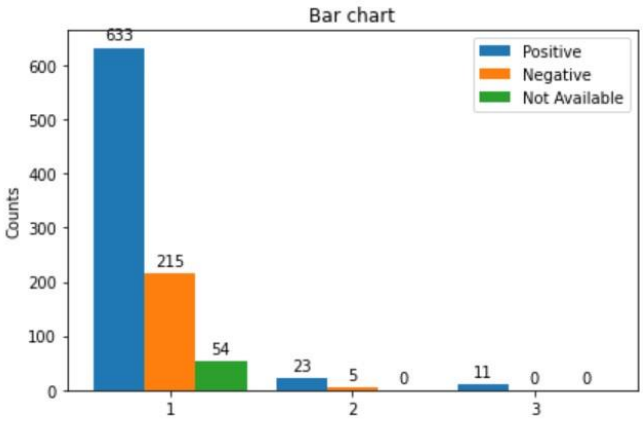

Figure 7.17: Anti-dsDNA

7.5 population analysis of clusters with respect to Tribes

Figure 7.18 through 7.20 shows the most tribes frequent in top 3 clusters ,in cluster 1 Al -balushi then Al-amri is the most two tribs, in cluster two also Al-balushi is one of most frequent tribs with equal frequent with Al-jabri and Al-riyami, but in third cluster the Al-Saadi is most frequent tribe. From Figure 5.3 we can see that the tribes Al-bulshi, Al-amri, Al-Saadi are one of most 10 tribes in Dataset.

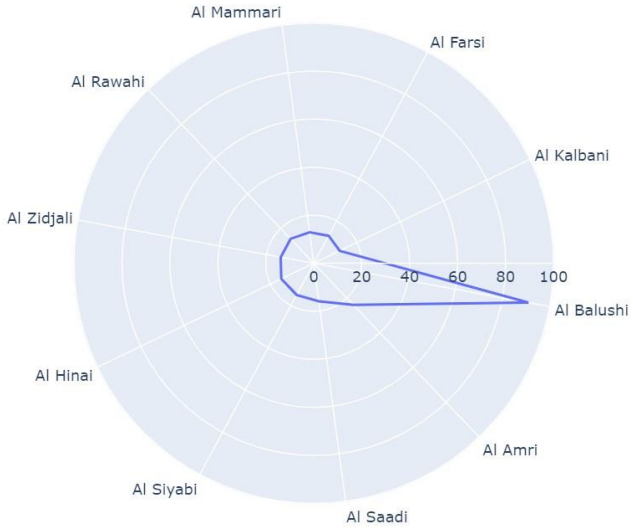

Figure 7.18: Cluster 1 tribes

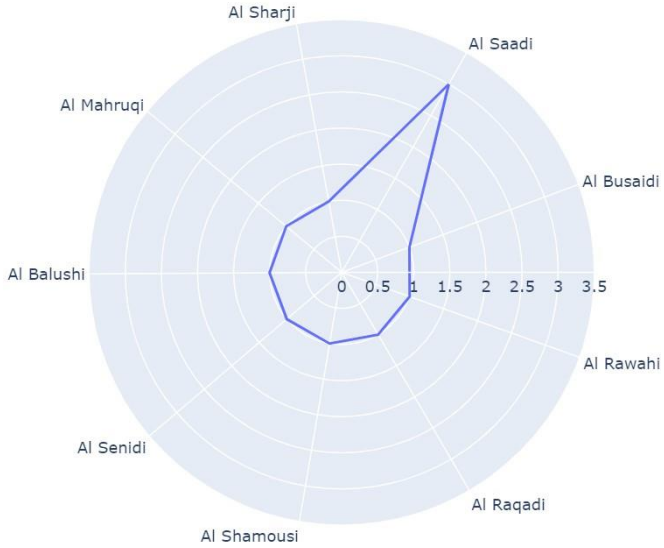

Figure 7.19: Cluster 2 tribes

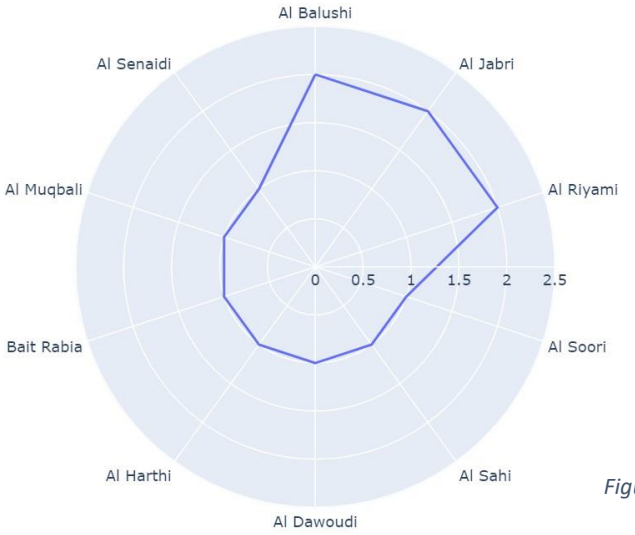

Figure 7.20: Cluster 3 tribes

CHAPTER 8: Conclusion

In this study, we have seen the systematic lupus erythematosus which is an autoimmune disease that fights body tissues to render the body unable to perform basic functions or death, to solve this problem we apply pre-processing methods to clean patient dataset and then we encode it; to visualize correlation network. Next, we apply the MCL cluster algorithm to the visualized correlation network; to cluster the patients into groups. Different clusters we have enriched with respect to different parameters.

We take the 3 most clusters that have nodes to do different experimental, with respect to trebles we find that the first cluster and the third cluster are enriched of patient whose tribes are Al-balushi, from this cluster we can conclude that the hereditary may a cause of spreading disease.

In the last there is many goals or benefit that we achieved from this project, for example but not limited:

1- We find many patterns from this study like the impact of tribes in infection of SLE diseases, also the disease is spread between females.

2- Learn how to arrive to scientific papers, and read many of them.

3- Manipulate real data and sensitive data

8.1: limitations:

Correlation network and population analysis has some limitations that can be mentioned here. the one of the limitations is that the model is taken for limited dataset, which is good to find some patterns, but not enough to release general pattern that appropriate for all countries.

The difficulty we encountered while working on the project,

The first difficulty we encountered was that the machine learning requires the stronger device and faster, and also, it’s required the graphic card to show the correlation networks clearly.

The other difficulty is taking high credited course with graduation project, which is causing some problem in time management, since other course is require also projects, and have many assignments.

Appendix:

Table 2: Eular/Acr weighted classification criteria

Feature Type	Feature Name		Description
Demographic Data	age		The patient Age.
	Age at diagnosis		Age of patient at which they started symptoms.
	Gender		Male or Female.
	Tribe		a social division in a traditional society.
	Region		Area where the patient love.
Clinical Manifestations	Discoid Rash		chronic inflammatory condition that is limited to the skin and is caused by an autoimmune disease.
	Malar Rash		Rash over the cheeks
	Alopecia		Hair loss
	Nephritis		Inflammation of the kidney
Immunology Criteria	Direct Coombs test		Detect antibodies that act against red cell
	Anti-Nuclear Antibody test		Detect antinuclear antibodies in blood
	Anti-dsDNA Test		Extension to ANA test
	Anti-phospholipid antibody		Detect phospholipid antibody
	low complement	C3	Detect level of complements proteins C3 and C4
		C4	
	Anti-sm		Detect the presence of antibodies against smooth muscle

Table 3: SLE Prevalence in Oman Data Collection Sheet

Study Serial Number	
Patient National ID	
Patient MRN	
Hospital Name	1- RH 2-SQUH 3- AFH. 4- Salalah 5- Al Buraimi. 6- Nizwa
Age	
Sex F/M	1- Male. 2-Female
Tribe	
Region	1- Muscat. 2- Al-Batinah. 3- Al-Dhakhilia. 4- Al Sharqiyah. 5- Al Buraimi 6- Al Wasta. 7- Dhofar. 8- Musandam 9- Al-Dhahira
Disease Duration	
Age at diagnosis	
Is there an Overlaping syndrome?	1- Yes. 2- No
If yes to the above question, what is the other Overlapping Disease?	1- Rheumatoid Arthritis. 2- Scteroderma 3- Myositis. 4- Sjogren's Syndrome 5- Mixed Connective tissue diseases. 6- Others
Arthritis	1- Yes. 2- No
Discoid Rash	1- Yes. 2- No
Malar Rash	1- Yes. 2- No
Photosensitive Rash	1- Yes. 2- No
Alopecia	1- Yes. 2- No
Mucosal ulcers	1- Yes. 2- No
Hemolytic anemia	1- Yes. 2- No
Thrombocytopenia	1- Yes. 2- No
Leucopenia	1- Yes. 2- No
Direct Coomb's test	1 -Positive. 2-Negative 3- Not available
Lymphadenopathy	1- Yes. 2- No
Cardiac	1- No 2- Pericardial effusion/ Pericarditis 3- Myocarditis. 4-Conduction defects. 5- valvular disease. 6- coronary artery disease. 7- Peripheral Vascular Disease 8- Others
Nephritis	1- Yes. 2- No
Kidney Biopsy	1- Available. 2- Not available
If kidney biopsy report available what is the LN Class?	1- Class I/II. 2- Class III. 3- Class IV 4-Class V. 5- Class VI. 6- Difficult to classify the biopsy sample.
If kidney biopsy was repeated, what was the reason?	1- To assess flare up. 2- To assess response to Rx. 3- For second opinion
What was the LN Class of the repeated kidney biopsy?	1- Class I/II. 2- Class III. 3- Class IV 4-Class V. 5- Class VI 6- Difficult to classify the biopsy sample
Neuropeychiatdc	1- No 2- Seizure 3- Psychosis. 4- Myelitis. 5- Stroke. 6- Neuropathy 7- Depression 8- Cognitive impairment 9- Vasculitis. 10- Others
Pulmonary	1- No 2- Pleural Effusion 3- ILD. 4- Pulmonary embolism 5- Pneumonitis. 6- Pulmonary HTN 7- Others

GI Involvement	1- No. 2- ischemic Colitis. 3- Mesenteric insufficiency. 4- Chronic peritonitis 5- Spleen/Liver infarction 6- Others
Eye Involvement	1-No. 2- Retinopathy. 3- Keratoconjunctivitis sicca. 4- Optic Atrophy. 5- Episcleritis/Scleritis 6- Others
ANA	1 -Positive. 2-Negative 3- Not available
AntI-daDNA	1 -Positive. 2-Negative 3- Not available
Anti sm	1 -Positive. 2-Negative 3- Not available
Anti Ra/S5A	1 -Positive. 2-Negative 3- Not available
Anti La/SGB	1 -Positive. 2-Negative 3- Not available
ACL IgM	1 -Positive. 2-Negative 3- Not available
ACL IgG	1 -Positive. 2-Negative 3- Not available
B2-glycoproteIn I	1 -Positive. 2-Negative 3- Not available
Lupus anticoagulant	1 -Positive. 2-Negative 3- Not available
C3	1-Normal. 2-High. 3- Low. 4- Not available
C4	1-Normal. 2-High. 3- Low. 4- Not available
TSH	1-Normal. 2-High. 3- Low 4- Not available
Mortality	1- Alive. 2- Died. 3- Unknown
If the patient died, what was the cause of death?	1- Spesis. 2- CV involvement 3- Lung involvement. 4- Renal involvement. 5- CNS Involvement. 6- Others

CHAPTER 9: Reference

[1] Cunha, J. S., & Gilek-Seibert, K. (2016). Systemic lupus erythematosus: A review of the clinical approach to diagnosis and update on current targeted therapies. Rhode Island medical journal, 99(12), 23.

[2] Chetti, P. and Ali, H. (2020). Estimating the Inspection Frequencies of Civil Infrastructures using Correlation Networks and Population Analysis.

[3] Müller Andreas C., Guido, S. (2018). Introduction to machine learning with python: A guide for data scientists. O'Reilly Media, Inc.

[4] Al Shareedah, A., 2022. Interpretable Approach for Predicting Systemic Lupus Erythematosus in Oman-based Cohort.. 2022.

[5] Siadati, S. (2018). What is unsupervised learning. Research Gate.

[6] Bustamam, A., Burrage, K., A. Hamilton, N. (2012, June). Fast Parallel Markov Clustering in Bioinformatics Using Massively Parallel Computing on GPU with CUDA and ELLPACK-R Sparse Format.

[7] Russell, R. (2018). Machine learning step-by-step guide to implement machine learning algorithms with python.

[8] Munusamy, A.; Sridharan, D. (2020). Analysis of Clustering Algorithms in Machine Learning for Healthcare Data. Springer.

[9] Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A comparison study on similarity and dissimilarity measures in clustering continuous data. PloS one, 10(12), e0144059.

[10] Jiang, J.-Y., Cheng, W.-H., Chiou, Y.-S., & Lee, S.-J. (2011). A similarity measure for text processing. 2011 International Conference on Machine Learning and Cybernetics. https://doi.org/10.1109/icmlc.2011.6016998

[11] Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., & Herawan, T. (2014, June). Big data clustering: a review. In International conference on computational science and its applications (pp. 707-720). Springer, Cham.

[12] Mohebi, A., Aghabozorgi, S., Ying Wah, T., Herawan, T., & Yahyapour, R. (2015). Iterative Big Data Clustering Algorithms: A Review. Software: Practice and Experience, 46(1), 107–129. https://doi.org/10.1002/spe.2341

[13] Young Lee, G., Alzamil, L., Doskenov, B., Termehchy, A. (2021, September). A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance.

[14] Ridzuan, F., & Wan Zainon, W. M. N. (2019). A Review on Data Cleansing Methods for Big Data.

[15] Chetti, P., & Ali, H. (2019). Analyzing the structural health of civil infrastructures using correlation networks and population analysis. In Proceedings of the eighth international conference on data analytics (pp. 12-19).

[16] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504.

[17] Bastian, M., Heymann, S., & Jacomy, M. (2009, March). Gephi: an open-source software for exploring and manipulating networks. In Proceedings of the international AAAI conference on web and social media (Vol. 3, No. 1, pp. 361-362).

[18] Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. Anesthesia & Analgesia, 126(5), 1763-1768.

[19] Bolboaca, S. D., & Jäntschi, L. (2006). Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. Leonardo Journal of Sciences, 5(9), 179-200.

[20] Shraddha Pandit, Suchita Gupta, "A Comparative Study on Distance Measuring Approaches for Clustering". International Journal of Research in Computer Science, 2 (1): pp. 29-31, December 2011.doi:10.7815/ijorcs.21.2011.011

[21] Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in Machine Learning: A new perspective. Neurocomputing, 300, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

[22] Al Maini, M. H., El-Ageb, e., Al-Shukaily, A. K., Al-Wahaibi, S., & Richens, E. R. (2002). Tribal and geographical variations of lupus in the Sultanate of Oman: a hospital-based study. Rheumatology international, 21(4), 141–145. https://doi.org/10.1007/s00296-001-0151-1

[23] Yacoub Wasef, S. Z. (2004). Gender differences in systemic lupus erythematosus. Gender Medicine, 1(1), 12–17. https://doi.org/10.1016/s1550-8579(04)80006-8

[24] Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., & Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. Bioinformatics, 26(18), 2347-2348.

[25] Fuchsberger, A., & Ali, H. (2017, January). A correlation network model for structural health monitoring and analyzing safety issues in civil infrastructures. In Proceedings of the 50th hawaii international conference on system sciences.

[26] Neufeldt V. & Guralnik D. B. (1996). *Webster's new world college dictionary* (Third). Macmillan.

[27] Dongen, S. V. (2000). Graph clustering by flow simulation. PhD thesis, University of Utrecht.

[28] Durán, C., Muscoloni, A., & Cannistraci, C. V. (2021). Geometrical

inspired pre-weighting enhances Markov clustering community detection in complex networks. *Applied Network Science*, *6*(1), 1-16.

[29] Brohee, S., & Van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, *7*(1), 1-19.

[30] Aringer, M., Costenbader, K., Daikh, D., Brinks, R., Mosca, M., Ramsey-Goldman, R., ... & Johnson, S. R. (2019). 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. *Arthritis & rheumatology*, *71*(9), 1400-1412.

[31] Helaly, M., & Mansour, M. (2018). Clinical Features Clusters in Systemic Lupus Erythematosus. The Egyptian Journal of Hospital Medicine, 71(5), 3136-3141.