Available online at www.qu.edu.iq/journalcm

JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS

ISSN:2521-3504(online) ISSN:2074-0204(print)

Graph-Based Community Detection in Hepatitis C Patient Data Using KNN Graph Construction and the Louvain Algorithm

Bashair Mohammed Obaid [a], ***Hayder K. Fatlawi*** [b,*]

[a] *Faculty of Education, Department of Computer Science, University of Kufa, Najaf, Iraq. Email: bashearm.hassan@student.uokufa.edu.iq*

[b] *Center of Information Technology Research and Development, University of Kufa, Najaf, Iraq. Email: hayder.fatlawi@uokufa.edu.iq*

ARTICLE INFO

Article history:
Received: 21 /11/2025
Revised form: 19 /1/2026
Accepted: 25 /01/2026
Available online: 30 /03/2026

Keywords: Hepatitis C virus Data, HCV, Clustering Techniques, Graph_ based, KNN Algorithm, community detection, Louvain Algorithm

ABSTRACT

Analyzing medical data to identify clinically relevant patient groups remains a complex task, particularly for hepatitis C virus patients, as traditional clustering methods often struggle to capture heterogeneous patient relationships and may rely on specific geometric assumptions or be constrained by the number of groups to be found. The proposed framework for detecting patient communities combines local similarity representation and global structural analysis. This framework builds a network of patient similarity using the K-Nearest Neighbors (KNN) algorithm, followed by the detection of patient communities through the Louvain algorithm. The proposed method is implemented on two real hepatitis C virus (HCV) datasets: the Egyptian HCV dataset and the HCV (UCI) dataset. The results of the experiments show effective detection of coherent and stable patient communities. The highest modularity values reach 0.740 and 0.805 for the Egyptian and UCI datasets respectively, at a neighborhood value of K=3. In addition, in this context, low distortion values indicate strong cohesion within the detected communities. Overall, the results confirm that graph-based patient community discovery provides a powerful alternative to traditional clustering techniques for medical data analysis. The proposed framework enables reliable and stable discovery of clinically relevant and interpretable latent patient subgroups.

MSC..

https://doi.org/10.29304/jqcsm.2026.18.12452

1. Introduction

Due to the widespread prevalence of hepatitis C virus (HCV), infection poses a significant challenge, considering the possibility of asymptomatic infection and the potential for serious complications over the long term. These factors raise important questions about the possibility of accurately studying the complex data of HCV patients and identifying clinically significant subgroups of patients. To resolve this issue, both the medical characteristics of the disease and the techniques required for the appropriate analysis of these heterogeneous characteristics must be considered.

[*] Corresponding author: Hayder K. Fatlawi

Email addresses: *hayder.fatlawi@uokufa.edu.iq*

Communicated by 'sub etitor'

1.1 Medical Background of Hepatitis C

Hepatitis C virus (HCV) is a major threat to human health worldwide and can lead to life-threatening complications[1]. According to the World Health Organization (WHO), more than 350,000 people die each year from complications of hepatitis C. More than 58 million people are chronically infected with hepatitis C [2]. In most cases, hepatocellular carcinoma occurs as a result of chronic hepatitis C virus infection [3].

In the Middle East, specifically in Egypt, liver disorders pose a significant health risk [4], [5]. Egypt has the highest prevalence rate in the world. Genotype 4 (G4) is the predominant genotype in Egypt, accounting for more than 90% of isolates [5], and is the most prevalent genotype in the Middle East [6]. Similarly, Egypt and Iraq share genotype 4 [7]. Furthermore, it remains difficult to determine the natural course of infection because hepatitis C virus infection persists for a long time, is usually asymptomatic, and has an indefinite onset [8]. Therefore, detecting hepatitis C using advanced, accurate, and rapid methods is very important for comprehensive screening. This prevents transmission between people and helps doctors initiate appropriate treatment [9].

1.2 Motivation and Technical Background

Given the importance of analysing complex medical data and the urgent need to process it accurately, efficiently and quickly, unsupervised machine learning techniques and graphs are used to detect hidden patterns and similarities between groups of patients with hepatitis C. Health systems must be designed to be consistent, flexible, secure and appropriate [10].

Aggregation plays an important role in medical data analysis due to its unsupervised nature. In this context, dimensionality reduction is an important concept to consider before aggregation, and is often viewed as a component of data preparation. High dimensions cause significant computational complexity. A particular set of mathematical difficulties known as the 'curse of dimensionality' may also arise[11] [12][13]. One of the most important applications of clustering is community detection.Communities can be defined as groups of contracts that have more connections within the group than outside it [14] [15]. in other words, they are groups of nodes and edges that have stronger relationships[16]. The most influential nodes tend to attract other nodes in the network to move closer to them automatically, thus forming a community [17] . A crucial graph mining technique that has been used in numerous real-world health applications is community detection. There are several main algorithms for detecting communities in healthcare applications[18] , such as the Louvain algorithm, the label propagation algorithm (LPA), the Gervais-Neumann algorithm, the PCA-based algorithm, and the graph colouring algorithm[19]. Among these methods, the Louvain algorithm was chosen for its ability to find communities in very large networks, its computational efficiency, and its ability to discover communities without needing to know the number or size of communities before clustering. Despite recent developments, the application of graph-based detection of hepatitis C patient communities using both K-Nearest Neighbors (KNN) and Louvain algorithms is not sufficiently explored.

In this study, a graph-based framework is proposed to discover communities latent in hepatitis C virus patient data. First, a patient similarity network is constructed using the (KNN) algorithm[20] to preserve local relationships between patients. This is followed by the application of the Louvain algorithm to detect communities in the generated graph due to its efficiency and ability to improve modularity in a short time and create a hierarchy for the Louvain network [21] without the need for prior knowledge of the number of communities. The main contribution of this work is the integration between Louvain and KNN that provides discovery of robust, stable, and interpretable patient communities beyond traditional clustering methods. The structure of the paper is organized as follows: in the next section, we review previous work related to our research. The third section explains Proposed Techniques; the fourth section presents Results and Discussion Then the fifth section presents the conclusion.

2. Literature Review

There are many works that have focused on discovering communities in medical data and other types of data. Various methods have been used to discover communities in order to achieve highly efficient and accurate data analysis. In 2022,[22] propose a new algorithm based on Louvain called "NI-Louvain" to discover the community and nodes with the highest influence, taking into account the influence of each node within the group. The proposed algorithm works in three stages. First, the graph to be processed undergoes density reduction, where clusters are calculated from the graph. This clustered graph undergoes a multi-level Louvain method in the second stage.

Communities produced by current algorithms such as edge-betweenness, label propagation, leading eigenvalues, Louvain, fast-greedy, walk-trap, and information map have a lower modularity value than the resulting community. The final stage involves converting the nodes of the resulting community back to the nodes of the original graph. Modality was calculated on multiple databases such as (kite, karate, macaque, us airport, immune, and Enron), where the results indicated the superiority of the proposed algorithm.

According to a study [23] published in 2022, the model was applied to a single-cell RNA sequencing dataset of a rare mouse intestinal cell type (23,630 features and 1,872 samples/cells). Ten cell clusters were obtained with a maximum modulus of 0.8851. The study used "Log Normalize" to normalize the data. The graph was constructed using the nearest neighbor (KNN) algorithm based on Euclidean distance in principal component analysis (PCA) space, which was used to minimize dimensions. Edge weights between any two cells were calculated using the Jacquard similarity coefficient. The first ten principal components were considered the dimensions of the dataset. The KNN graph contained 818 nodes and 20,511 edges. Hierarchical cumulative clustering with Louvain algorithm was used to analyze the single-cell RNA sequencing data.

In study [24],the COBALT model was introduced in 2023, which is a cost-based layered model. To identify phenotypes using a community detection approach, the work maintained the quality of these phenotypes while minimizing the number of features used. Using data from individuals with persistent tinnitus who completed questionnaires, the model was evaluated and the results presented in a multi-layer network structure. Cost is defined as the cost required to acquire features. The algorithm used is the Leiden algorithm, which is a cost-sensitive algorithm. The results were compared between the proposed model and traditional models. Improved prediction quality resulted from the development of interconnected communities. Cost-sensitive questionnaires, i.e., minimizing costs (and ultimately the patient burden) for each community without compromising prediction quality, were selected for the communities. Furthermore, missing data did not significantly affect the partitioning, indicating that communities can be identified even with fewer cases.

In a 2023 study [25], the hyOPTGB model was proposed which enhanced gradient boosting (GB) using an improved classifier to predict the condition of hepatitis C patients in Egypt. The dataset used in the study contained 1,385 cases and 29 features. The hyOPTGB model outperformed a number of other machine learning models, achieving an accuracy rate of 95.3%. The model used Min-Max normalization pre-processing to reduce the values in the dataset to a fixed range, and features were selected to identify the most relevant ones. The paper also conducts a comparative study between the proposed hyOPTGB model and those used by other researchers who used the same dataset, and the results show that the proposed model achieved the best results.

The model [26] presented in 2023 is based on machine learning methods to predict hepatitis C virus among healthcare workers in Egypt. The research was applied to real data from the National Liver Institute of Menoufia University (Menoufia, Egypt). The collected dataset consisted of 859 patients with 12 different features. To ensure that the proposed framework was robust and reliable, two scenarios were applied: the first without feature selection and the second after feature selection using sequential forward selection (SFS). Furthermore, the selected feature set is evaluated based on the features resulting from SFS. Naive Bayes, Random Forest (RF), K-Nearest Neighbor (KNN), and Logistic Regression algorithms were used as inference algorithms. The results of the experiment indicated that the proposed framework achieved higher accuracy after feature selection using SFS compared to not selecting features. The RF classifier achieved an accuracy of 94.06% with a minimum learning time of 0.54 seconds. Finally, after adjusting the hyperparameter values of the RF classifier, the classification accuracy improved to 94.88% using only four features. In a 2024 study [20], proposed a framework based on stable graph using KNN algorithm and determining the optimal structure k on large-scale single cell data, followed by the use of the (Louvain algorithm to obtain coherent clusters where the optimal accuracy is determined by the Calinski-Hara basz index (CH) index. The study proved its effectiveness by analyzing 15 tissues from the human fetal atlas. Compared to current methods, CDSKNN is highly applicable and robust, and is able to balance complexities across different types of data. More importantly, CDSKNN shows higher operational efficiency on datasets of the order of a million cells, requiring on average only 6.33 minutes to cluster 1.46 million single cells, saving 33.3% to 99% of runtime compared to existing methods.

A study [27] in 2024 propose increasing the efficiency of algorithms used to detect communities in graphs. Using clustering algorithms on this latent representation, these methods begin by partitioning carefully selected nodes, i.e., the partition resulting from embedding the nodes in real numerical space. The study believed that embedding would learn more complex and difficult relationships between nodes or filter out unwanted noise while keeping nodes close to each other and belonging to the same community. The clustering algorithms implemented on this embedding would provide a stable partition that would reduce uncertainty in the early stages of community detection techniques. The proposed method was applied to Facebook data and significantly outperformed basic community detection techniques, such as Louvain and Leiden.

In a 2024 study [28], a network of diseases associated with breast cancer was created based on disease similarity, by creating a similarity matrix and converting it into a neighborhood matrix with a specific threshold. In this research, the community algorithm was applied to discover the community within the giant component of the network. The result is evaluated using modularity and similarity measures. The modularity evaluation using five modularity measures indicates the five best community detection algorithms, which are Leiden, Louvain, RBER Pots, RB Pots, and Walktrap. Similarity measures using Similarity measures with the top three fitness functions (internal edges, normalized density, and size) indicate that the Leiden–Louvain and RBER Pots–RB Pots algorithms are a pair of algorithms with similar results, leaving Walktrap as an algorithm with different community results. Other similarity measurements using the V-measure and visualized with a heatmap indicate that the results of Louvain–Leiden (0.99), RB Pots–Leiden (0.97), and RB Pots-RBER Pots (0.96) are very similar to each other.

In a 2024 study [29],three methods were compared for conducting multiple analysis to discover patterns: first, clustering in natural language processing (NLP); second, graph community detection techniques; and finally, a hybrid approach combining NLP and graphs. As a result, clinical subject matter experts and service systems operations specialists concluded that each of these methods produced clinically similar public service units. These included emergency and acute care services, addiction services, and psychiatric services. In a 2024 study[30], the M-ClustEHR model was introduced a clustering technique that uses multimodal data to create and implement a comprehensive clinical representation of sepsis patients. The method demonstrates excellent stability and routinely outperforms the clustering challenge for both the Angus and Sepsis-3 cohorts. According to the study results, sepsis is a complex medical problem with multiple contributing factors. The results highlight the importance of age and gender in the initial triage after ICU admission and provide compelling evidence of their critical role in the development of sepsis.

In a 2025 study [31], machine learning (ML) was applied to accurately diagnose hepatitis C in patients. Combining two widely used datasets from different sources resulted in a hybrid dataset. A subset of the dataset served as a hold-out set to mimic data from the actual world. K-means for binning and k-modes for categorical clustering in this study were two examples of a multi-dimensional pre-clustering technique. To extract a new feature, the pre-clustering technique was applied. To train a stacked meta-model, this extracted feature column was appended to the original dataset .The model and baseline models were contrasted. With explainable artificial intelligence, the predictions were further developed. XGBoost, K-nearest neighbor, support vector classifier, and random forest (RF) were the models that were employed. RF yielded a baseline score of 94.25%, whereas the meta-model produced a score of 94.82%. Although there have been many studies on community detection based on graphs in social, biological, medical, and other data analysis, current research still lacks the combined use of KNN-based graph construction with pattern-based community detection algorithms, such as Louvain, especially when applied to hepatitis C patient data. Similarly, most research focusing on hepatitis C virus data has aimed to classify and predict disease cases using supervised machine learning. To our knowledge, the discovery of latent patterns in hepatitis C patients remains insufficiently studied, leaving room for our research to shed light on addressing this gap.

3. Proposed Techniques

In our study, we implement two basic stages. The first stage is preprocessing ,it includes cleaning the data, removing missing data, and converting it into digital data with a uniform range to ensure its impartiality and obtain accurate results. At this stage, an KNN algorithm is also applied to create a graph representing the relationships between samples so that they are suitable for entering the next stage, which involves applying the Louvian algorithm for network analysis and detecting patient communities based on the graph. As shown in the figure 1 .

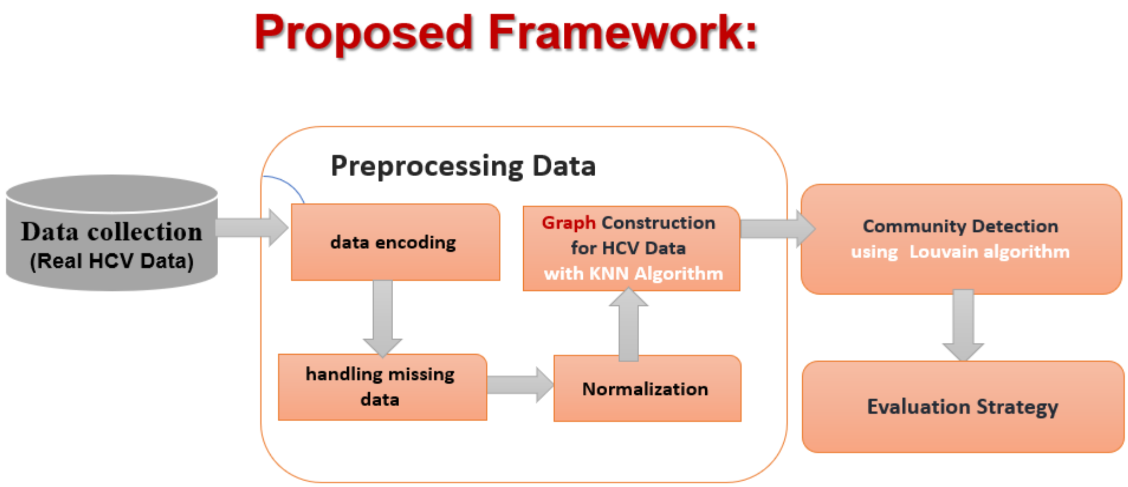


Figure 1. Chart of Proposed Techniques.

We will explain in detail the steps involved in our proposed work, which consists of two main parts:

3.1 Preprocessing

The goal is to convert the data into a format suitable for the community discovery algorithm. This includes:

- *handling missing data* :Missing data indicates that some fields or records in the dataset may not contain data, which can be represented by empty spaces, NaN, null, or other values. Correctly compensating for missing data is extremely important as it affects the model's performance, accuracy, and results. We imputed the missing values by using median method that is robust in the presence of extreme values.
- *Data encoding*: Convert text data into numbers because most algorithms are good at using numerical columns and do not understand text.
- *Normalization* :The goal is to ensure that the different values of the attributes fall within a common range. This is crucial for an algorithm that relies on distances between data points, as KNN algorithm. For example, if the database contains a feature with values ranging from 10,000 to 1,000,000 (such as a test score) and another feature ranging from 0 to 1 (such as gender), the higher values of the attribute (test score) will dominate the distance measurement, and the algorithm will almost completely ignore the influence of the attribute with the lower value. Therefore, normalization was used to prevent bias. The method used in our proposed study is the Min-Max scale. This method converts all values to a specific range, usually between 0 and 1, as in the following equation:

$$X\ norm = (X - X\ min) / (X\ max - X\ min) \quad (1)$$

Where X :the original value, X min : the minimum value in the attribute.

X max : the maximum value in the attribute.

- *K-Nearest Neighbors (KNN) algorithm: Although the K-Nearest Neighbors (KNN) algorithm is typically used for supervised classification and regression, we used it to create a weighted, undirected graph. Each data sample (subject) is represented as a node. The algorithm calculates the Euclidean distance for all node pairs. For each node, it connects each node to its k nearest neighbors to preserve local multi-structures. In this context, edge weights were assigned using the inverse distance function w = 1/(d + eps) to ensure that the influence of the nearest neighbors is stronger in the community detection process using the Louvain clustering-based algorithm.*

K-Nearest Neighbor (KNN) algorithm

```
Inputs: dataset (HCV) (samples* features) (M*N)
K: number nearest neighbors
d:Distance scale (Euclidean)
eps = 1e-9
Outputs
uG: weighted and Undirected graph representing patients and their relationships
begin
Steps:
1- Handle missing values in (HCV) using Median Imputation.
2-Normalize the numerical columns
3-Use the selected scale to calculate the distance between nodes(as Euqlidiean)
4-Calculate edge weight: w = 1 / (d + eps)
5-Represent each patients with(i)
6-Determine the value of (k)
7- Represent neighbors with(j)
8- Add an edge between(i-j) whose proximity was calculated in step 3.
9-Save the final graph
End
```

3.2 Communities Detection by Louvain

In this stage, the Louvain algorithm receives data in the form of a graph created by the Nearest Neighbor algorithm. Initially, the Louvain algorithm considers each node in the graph to be an independent community as shown in the figure 2. (a). The algorithm *optimizes the Modularity* by randomly assigning each node to a specific community from its neighbors. It then calculates the effect of this arrangement on the modularity of the entire network and selects the arrangement that increases this metric. This step is repeated for each node until the increase in modularity value ceases as shown in the figure 2 . (b).

In the second stage, the algorithm *aggregates* the network, considering each community resulting from the first stage as a giant node as shown in the figure 2. (c) . The edge between these nodes carries a weight equal to the number of edges between the communities in the first stage. The algorithm then repeats the process of searching for communities between giant nodes that increase the modularity value. This process is repeated until fewer and more interconnected communities are reached. The algorithm stops when it reaches an aggregation point where the modularity value no longer increases as shown in the figure 2. (d).

In this context, the Louvain algorithm's ability to improve modularity and its efficiency in detecting communities in large and complex networks without the need for prior knowledge of the number of communities was the reason for choosing this algorithm. Similarly, the hierarchical nature of the algorithm allows for the discovery of clinically relevant clusters and does not require significant computational costs, making it a suitable algorithm for detecting communities in high-dimensional medical data.

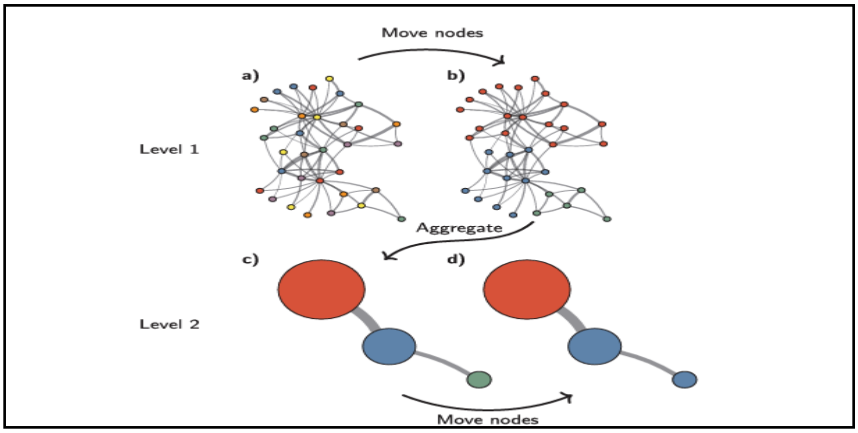


Figure 2. Stages of the Louvain algorithm [32].

Louvain algorthim

```
Input :Graph G =(V,E) from KNN Algorithm
Output :communities of Interconnected nodes
Begin
any node v € V has Private community S(v)={v}
 compute initial Modularity(Q)
Repeat
Step 1 : local optimization of Modularity
Repeat v € V in random arangement
Get S best from ( v € S best) → ΔQ max>0
Step 2: Aggregation of node
Create new graph :G n =( Vn ,En )
Make each community S i node in V n
 Weight En between (S i,S j) = £ Weight E between( S i,Sj) in G
Replace H with H n
until ΔQ <=0
end
```

Based on the above description, the KNN and Louvain algorithms were chosen because of the efficiency of each algorithm individually and their complementary roles within the proposed framework. The KNN algorithm was chosen to create the graph because it preserves local similarities between patients while creating a sparse, stable, and balanced network. This graph structure is highly compatible with the Louvain algorithm, which discovers communities by optimizing of modularity on weighted graphs, allowing locally similar patients connected by KNN to be grouped into globally coherent communities.

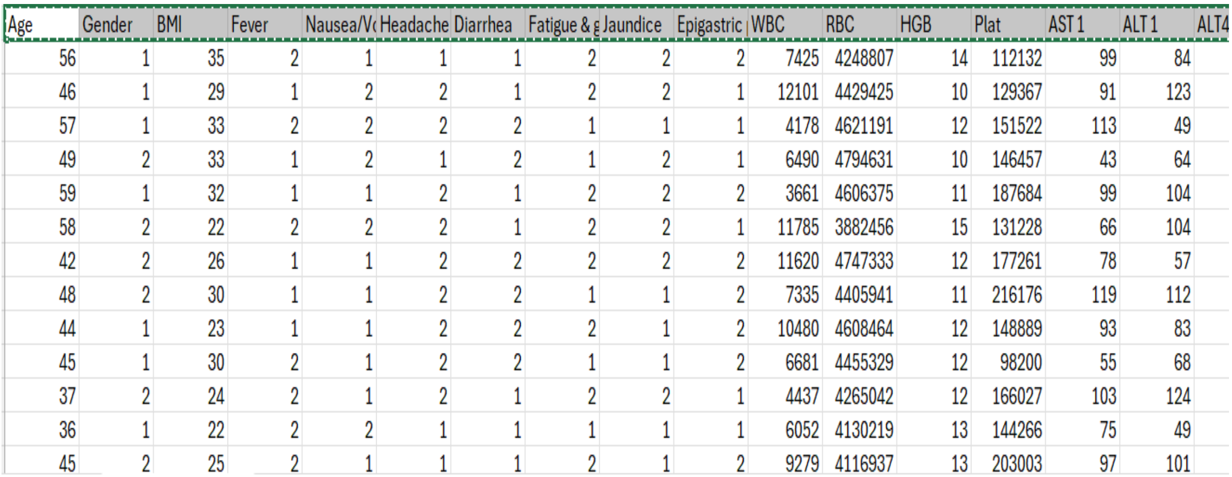
4. Results And Discussion

This section will describe the databases used in the proposed model, along with a preliminary overview of data collection using the K-means algorithm. It will also present the experimental results obtained after applying the proposed steps to the databases, following testing of different K-values in the KNN algorithm. Furthermore, we will highlight the effect of increasing or decreasing the K-value on several performance metrics of hepatitis C virus (HCV) communities detected after applying the Louvain algorithm.

4.1 Dataset Description

The Egyptian HCV database from Kaggle contains 1385 records and 29 attributes, including various characteristics and analysis results such as age, sex, BMI, AST, ALT, etc., as shown in table 1. Before performing the analysis, the label column representing the basic histological stage was excluded to avoid any bias and because the proposed framework is based on unsupervised learning.

Table 1. Sample of Egyptian HCV database.



Age	Gender	BMI	Fever	Nausea/V(	Headache	Diarrhea	Fatigue & g	Jaundice	Epigastric	WBC	RBC	HGB	Plat	AST 1	ALT 1	ALT4
56	1	35	2	1	1	1	2	2	2	7425	4248807	14	112132	99	84	
46	1	29	1	2	2	1	2	2	1	12101	4429425	10	129367	91	123	
57	1	33	2	2	2	2	1	1	1	4178	4621191	12	151522	113	49	
49	2	33	1	2	1	2	1	2	1	6490	4794631	10	146457	43	64	
59	1	32	1	1	2	1	2	2	2	3661	4606375	11	187684	99	104	
58	2	22	2	2	2	1	2	2	1	11785	3882456	15	131228	66	104	
42	2	26	1	1	2	2	2	2	2	11620	4747333	12	177261	78	57	
48	2	30	1	1	2	2	1	1	2	7335	4405941	11	216176	119	112	
44	1	23	1	1	2	2	2	1	2	10480	4608464	12	148889	93	83	
45	1	30	2	1	2	2	1	1	2	6681	4455329	12	98200	55	68	
37	2	24	2	1	2	1	2	2	1	4437	4265042	12	166027	103	124	
36	1	22	2	2	1	1	1	1	1	6052	4130219	13	144266	75	49	
45	2	25	2	1	1	1	2	1	2	9279	4116937	13	203003	97	101	

In the table2 , we list the names of the features with a description of each one and the data type, in order to provide a comprehensive overview of the Egyptian HCV database used.

Table 2 : Description of Egyptian HCV dataset.

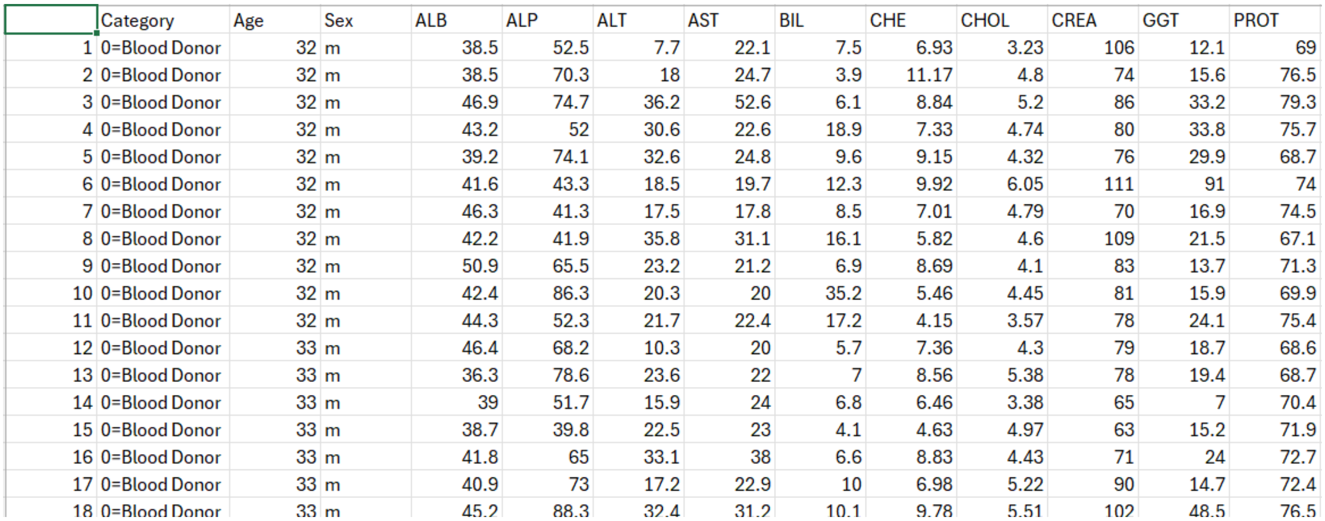
features	Description	Type of data
Age	Patient's age	Numerical
Gender	Patient's Gender	Binary
(BMI)	Body Mass Index (BMI)	Numerical
Fever	Elevated body temperature	Binary
Nausea/Vomiting	Presence nausea or Vomiting	Binary
Headache	Presence headache	Binary
Diarrhea	Presence diarrhea	Binary
Fatigue & generalized bone ache	Fatigue and generalized bone ache in body	Binary

Jaundice	Yellowing of the skin and whites of the eyes	Binary
Epigastric pain	Pain in the lower rib cage (upper abdomen)	Binary
(WBC)	White Blood Cells	Numerical
(RBC)	Red Blood Cells	Numerical
(HGB)	Hemoglobin	Numerical
Plat	Platelets	Numerical
AST 1	Liver enzyme level: (start)	Numerical
ALT4	ALT After 4 weeks	Numerical
ALT 12	ALT After 12 weeks	Numerical
ALT 24	ALT After 24 weeks	Numerical
ALT 36	ALT After 36weeks	Numerical
ALT 48	ALT After 48 weeks	Numerical
ALT after 24 w	ALT after After 24 week of treatment	Numerical
RNA Base	Baseline viral load (BML)	Numerical
RNA 4	Viral load in the blood after 4 weeks.	Numerical
RNA 12	Viral load of HCV in the blood after 12 weeks.	Numerical
RNA EOT	Viral load of HCV at end of treatment	Numerical
RNA EF	Viral load of HCV at the end of follow-up.	Numerical
Baseline histological Grading	Grade of liver inflammation. Baseline histological grading	ordinal
Baseline histological staging	(Class Label) Liver cirrhosis stage	ordinal

4.1.1 Description of HCV (UCI)

The study used the HCV dataset from the UCI Machine Learning Repository. The dataset consists of 615 instances representing a mix of healthy donors and HCV patients as shown in Table 3. The database contains 14 attributes. To increase the accuracy of similarity calculations and community detection, prior to analysis, record identifiers were removed from the dataset because they do not convey analytical or clinical information. Similarly, category label was excluded because the proposed framework does not rely on predefined category classifications but rather on unsupervised analysis.

Table 3. Sample of HCV (UCI).



	Category	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT
1	0=Blood Donor	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106	12.1	69
2	0=Blood Donor	32	m	38.5	70.3	18	24.7	3.9	11.17	4.8	74	15.6	76.5
3	0=Blood Donor	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.2	86	33.2	79.3
4	0=Blood Donor	32	m	43.2	52	30.6	22.6	18.9	7.33	4.74	80	33.8	75.7
5	0=Blood Donor	32	m	39.2	74.1	32.6	24.8	9.6	9.15	4.32	76	29.9	68.7
6	0=Blood Donor	32	m	41.6	43.3	18.5	19.7	12.3	9.92	6.05	111	91	74
7	0=Blood Donor	32	m	46.3	41.3	17.5	17.8	8.5	7.01	4.79	70	16.9	74.5
8	0=Blood Donor	32	m	42.2	41.9	35.8	31.1	16.1	5.82	4.6	109	21.5	67.1
9	0=Blood Donor	32	m	50.9	65.5	23.2	21.2	6.9	8.69	4.1	83	13.7	71.3
10	0=Blood Donor	32	m	42.4	86.3	20.3	20	35.2	5.46	4.45	81	15.9	69.9
11	0=Blood Donor	32	m	44.3	52.3	21.7	22.4	17.2	4.15	3.57	78	24.1	75.4
12	0=Blood Donor	33	m	46.4	68.2	10.3	20	5.7	7.36	4.3	79	18.7	68.6
13	0=Blood Donor	33	m	36.3	78.6	23.6	22	7	8.56	5.38	78	19.4	68.7
14	0=Blood Donor	33	m	39	51.7	15.9	24	6.8	6.46	3.38	65	7	70.4
15	0=Blood Donor	33	m	38.7	39.8	22.5	23	4.1	4.63	4.97	63	15.2	71.9
16	0=Blood Donor	33	m	41.8	65	33.1	38	6.6	8.83	4.43	71	24	72.7
17	0=Blood Donor	33	m	40.9	73	17.2	22.9	10	6.98	5.22	90	14.7	72.4
18	0=Blood Donor	33	m	45.2	88.3	32.4	31.2	10.1	9.78	5.51	102	48.5	76.5

we list the names of the features with a description of each one and the data type, in order to provide a comprehensive overview of HCV(UCI) dataset used, as shown in the table 4.

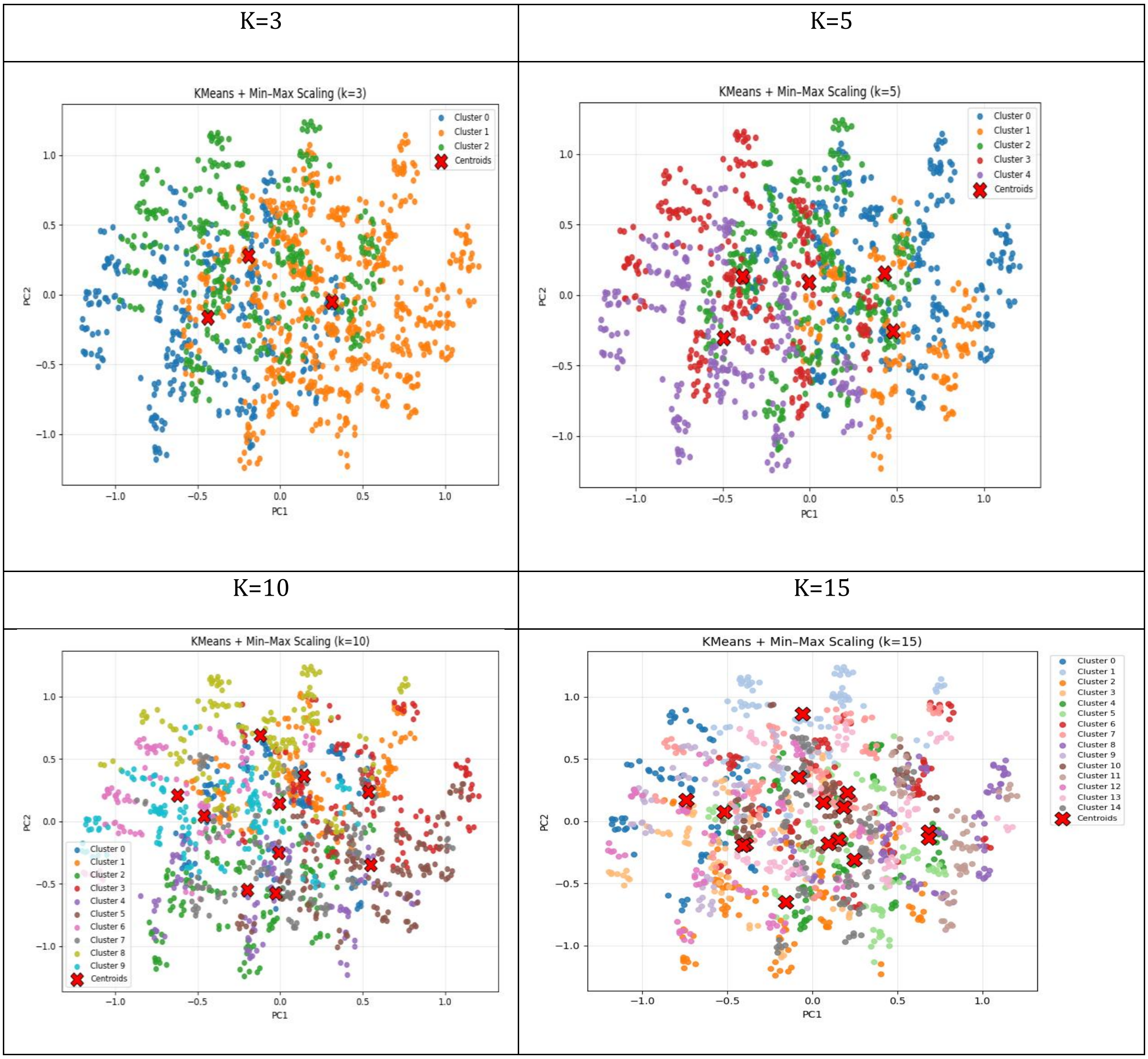
Table 4 : Description feature of HCV (UCI) Dataset.

features	Description	Type of data
unnamed	Sequence	Numerical
Category	Target (values: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis')	Categorical
Age	Number of years of the patient's age	Numerical
Sex	Male/Female	binary
ALB	(albumin): The main protein produced by the liver. A low level may indicate liver damage.	Numerical
ALP	(alkaline phosphatase): An enzyme found in the liver and bones. A high level may indicate liver disease or blockage of the bile ducts.	Numerical
ALT	(Alanine transaminase)It is found mainly in the liver	Numerical
AST	(aspartate aminotransferase): Another enzyme found in the liver, heart, and muscles. Elevated levels may indicate liver damage.	Numerical
BIL	(bilirubin): A yellow substance produced when red blood cells break down. It is processed by the liver, and high levels may lead to yellowing of the skin and eyes (jaundice).	Numerical
CHE	(Cholinesterase): One of the enzymes produced by the liver.	Numerical
CHOL	(Cholesterol): The level of fat in the blood.	Numerical
CREA	(Creatinine): Used specifically to assess kidney function.	Numerical

GGT	(gamma-glutamyl transferase): One of the liver enzymes.	Numerical
PROT	The total amount of protein present in blood serum	Numerical

4.2 Initial Cluster Analysis

Before applying community detection using the Louvain algorithm, the nature of the data was explored, and a preliminary view of the clustering and homogeneity of the data points was obtained using one of the unsupervised clustering algorithms, the K-means algorithm, which requires the value of K to be specified from the outset. The algorithm was applied using different values for K (3,5,10,15) on (Egyptian HCV dataset) as shown in the figure 3 .


K=3

K=5

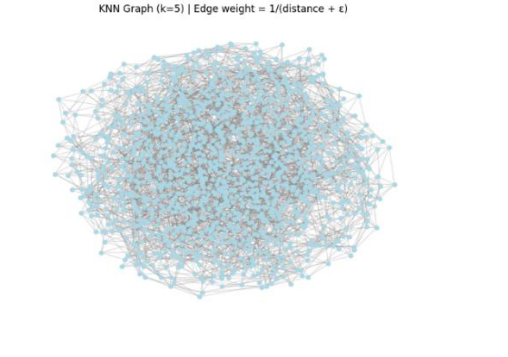
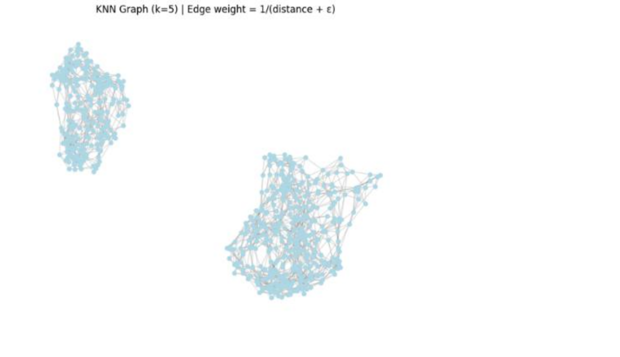
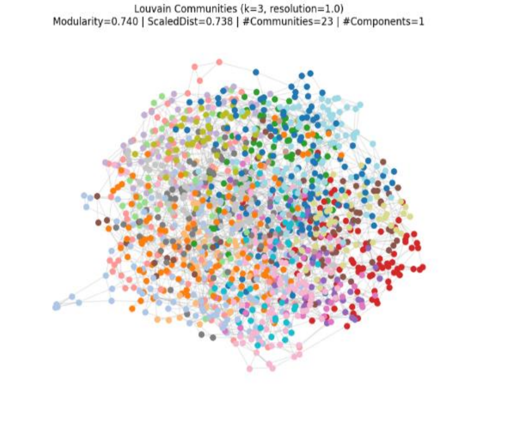
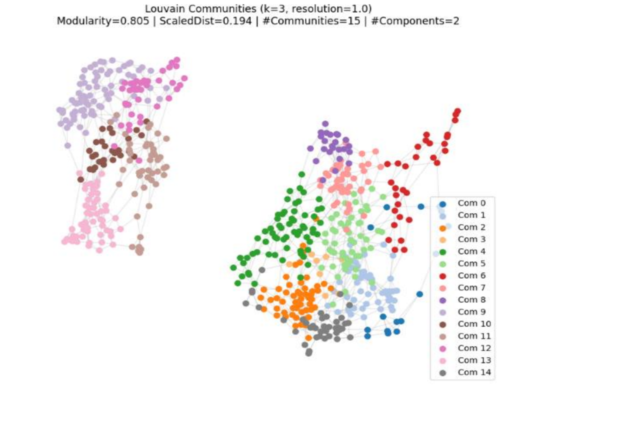
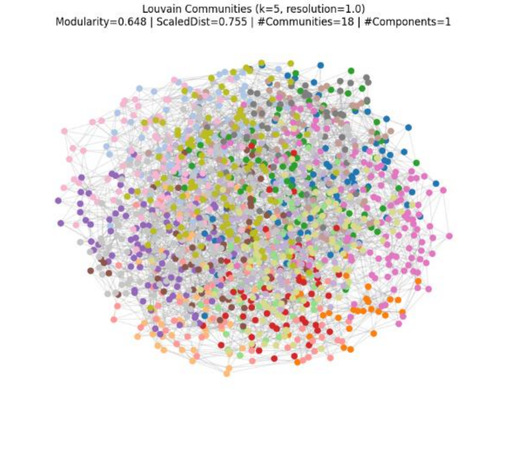
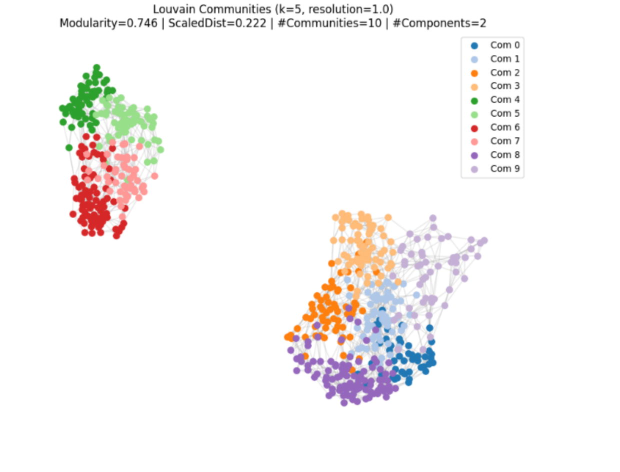
K=10

K=15

figure 3. Results of applying the k-mean algorithm using different K values

4.3 Results of the proposed model.

In this section, we review the results of applying the Louvain algorithm used to detect patient communities when applied to two separate databases to determine the algorithm's success in distinguishing communities and its clarity when applied to different databases. As shown in figure 6, the algorithm was tested on graphs generated at four different values of k (number of nearest neighbors).We also review the results of the KNN algorithm at K=5 for each database, which represents an input for the community detection algorithm.

Model	Data set	
	HCV (Egyptian)	HCV (UCI)
KNN Algorithm(K=5)		
KNN (K=3) +Louvian		
KNN (K=5) +Louvian		

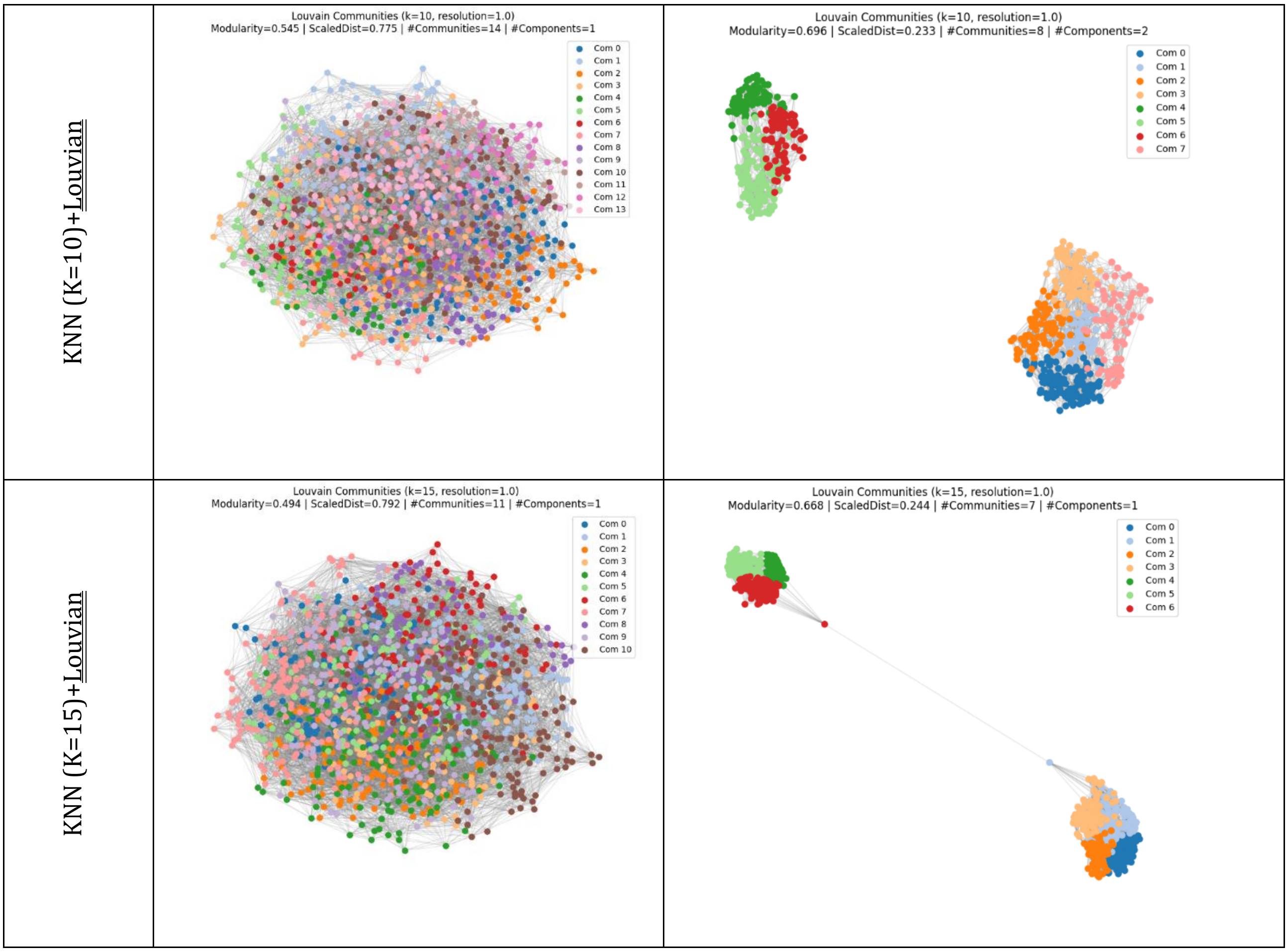


Figure 4. Results of applying KNN and Louvain algorithm using different values of neighborhood K.

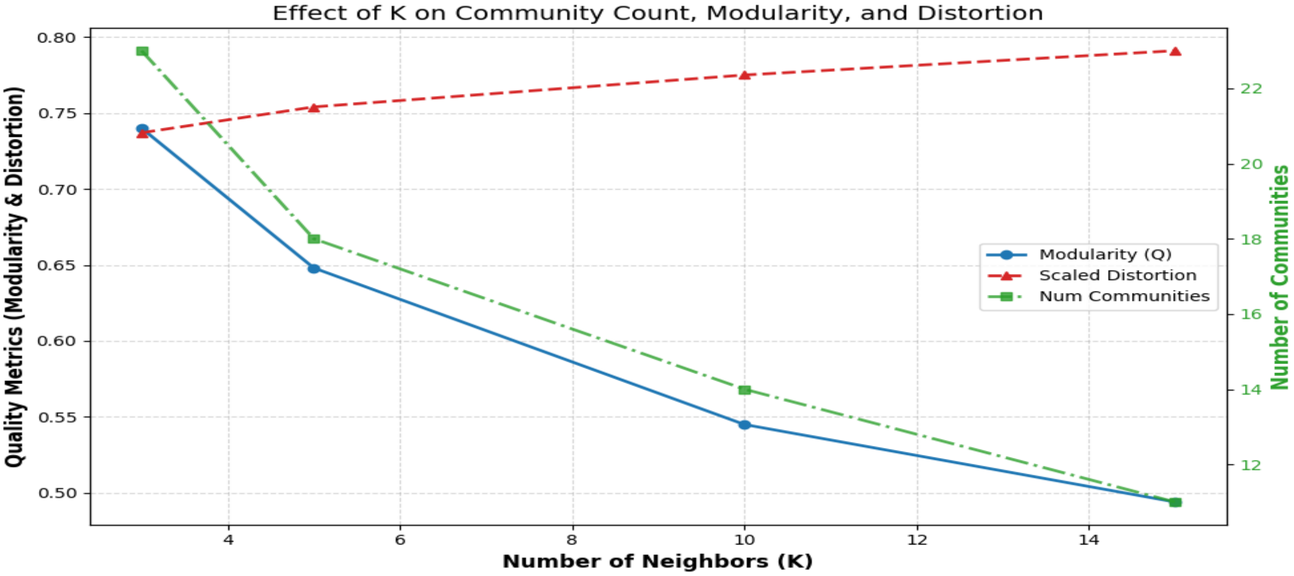


Figure 5: Effect of K in KNN on the amount of modularity, measured distortion, and number of communities detected in Louvain on the Egyptian HCV dataset.

To analyze performance metrics for the graph-based Louvain algorithm using KNN across different neighborhood sizes (K) on HCV Egyptian dataset the results are summarized in the table 5.

Table 5: effect of KNN (K) on community detection metrics using Louvain on the Egyptian HCV dataset.

K	Resolution	Num _Edges	Num_ Components	Num_ Communities	Modularity	Scaled_ Distortion
3	1.0	2884	1	23	0.740030	0.737750
5	1.0	4633	1	18	0.648353	0.754526
10	1.0	8983	1	14	0.545123	0.775354
15	1.0	13238	1	11	0.494343	0.791624

The effect of different K-values in the KNN algorithm on graph quality, cluster detection using the Louvain algorithm, and overall clustering behavior, based on real-world hepatitis C virus (HCV) data in Egypt, is examined. Table 5 summarizes cluster detection performance metrics for different K-values, including the number of edges, the number of components, the number of clusters, modularity, and measured distortion. Similarly, Figure 6 illustrates the effect of different K-values on modularity, measured distortion, and the number of detected clusters. Lower K-values result in higher modularity and a greater number of clusters. In particular, the value K = 3 achieves the highest level of modularityand the lowest level of distortion, indicating that it has achieved a remarkable balance between similarities among patients locally and the separation of communities globally with lower K values. Conversely, as K values increase, the number of connections increases, leading to an increase in graph density and thus introducing false similarities between dissimilar patients, resulting in the merging of some communities, a decrease in modularity, and an increase in distortion. This behaviour illustrates the trade-off between cluster accuracy and graph connectivity, highlighting the importance of choosing an appropriate value for K when analyzing patient health data.

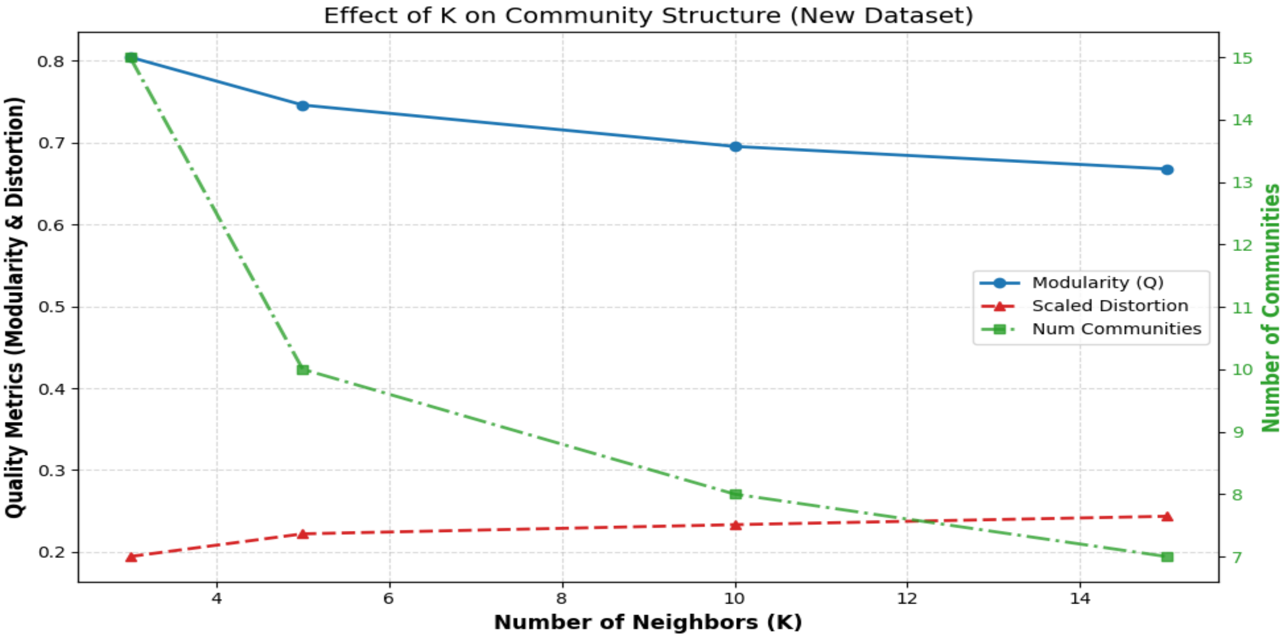


Figure 6: Effect of K in KNN on the amount of modularity, measured distortion, and number of communities detected in Louvain on HCV(UCI) dataset.

Table 6: Impact of KNN neighborhood size (K) on community detection metrics using Louvain on HCV (UCI) dataset.

k	Resoluti n	Num _ Edges	Num _ Components	Num_ Communities	Modularity	Scaled_ Distortion
3	1.0	1367	2	15	0.804597	0.194450
5	1.0	2246	2	10	0.746066	0.222178
10	1.0	4448	2	8	0.695526	0.233374
15	1.0	6642	1	7	0.667985	0.243628

For further evaluation of the generalizability of the proposed framework, it was applied to another dataset with different characteristics. The purpose of this experiment is to verify the amount of interaction between the graph constructed by the KNN algorithm and the communities discovered by the Louvain algorithm. Table 6 shows the performance metrics corresponding to the second dataset across different K values. Similar to the Egyptian HCV dataset, the graphs become more connected, and the number of communities detected decreases as K increases, corresponding to a decrease in modularity and an increase in distortion. Conversely, decreasing K values again lead to higher modularity and an increase in the number of communities with higher accuracy and less distortion.

Figure 6 visually illustrates these observations by showing the relationship between K, modularity, measured distortion, and the number of communities. This consistency in the behavior of the proposed framework indicates that it does not depend on a specific data set, but rather on the choice of basic characteristics in constructing the graph using KNN.

In general, when observing the behavior of the proposed framework in both databases, a clear pattern emerges regarding the role of the K factor in the KNN algorithm's community discovery performance. Low K values effectively preserve local connections, enabling the Louvain algorithm to more accurately identify discrete and cohesive communities. Conversely, as K values increase, graph density increases, weakening community boundaries and thus reducing their regularity. Similarly, the observed increase in measured distortion with increasing K values supports this interpretation. High distortion indicates that local structural similarities deteriorate progressively with increasing neighborhood size, resulting in less significant community divisions.

Regularity and distortion together constitute two complementary measures of the accuracy and stability of these communities. In contrast, traditional clustering methods such as K-means operate directly in the feature space and require a prior assumption about the number of communities, while the proposed graph-based approach requires no assumption about the shape or number of communities. The community structure is automatically determined by the Louvain algorithm based on the network structure, making it highly suitable for discovering unknown communities in patient data networks, thus providing stable and interpretable detection.

4.4 Comparative Evaluation at the Optimal Neighborhood Size (k = 3)

To avoid unnecessary complexity and ensure clarity of interpretation, the comparative analysis focuses on the number of communities detected at a neighborhood value (k = 3) that exhibited the best performance metrics for the proposed model when replicated across the two datasets. Table (7,8) presents the results of a comparison between the proposed method and the (k-mean, k-mean+spline) algorithm using common evaluation metrics.

Table 7: Comparative Evaluation on the Egyptian HCV dataset.

Method	Num. Communities	Silhouette ↑	Davies-Bouldin ↓	Calinski-Harabasz ↑	Stability(ARI) ↑	Stability(NMI)↑
Proposed (KNN+Louvain)	23	0.025	3.076	22,01	**0.536**	**0.758**
K-mean	23	0,060	2.613	29.04	0.294	0.556
Spline+ K-mean	23	0.025	3.203	18.57	0.175	0.395

Table 8: Comparative Evaluation on HCV (UCI) dataset

Method	Num. Communities	Silhouette ↑	Davies-Bouldin ↓	Calinski-Harabasz ↑	Stability (ARI) ↑	Stability (NMI) ↑
Proposed (KNN+Louvain)	15	0.100	1.923	177.54	**0.741**	**0.850**
K-mean	15	0,170	1.371	222,98	0.730	0.811
Spline+ K-mean	15	0.126	1,712	60.16	0,627	0.768

Looking at Tables 7 and 8, it is clear that geometric metrics tend to be higher in both data sets when using the K-means algorithm. This is due to the design of the k-means algorithm to optimize geometric and spherical clustering, reflecting its compatibility with geometric assumptions, but not demonstrating its robustness and reproducibility. Examples of such metrics include Silhouette, Davies–Bouldin, and Calinski–Harabasz. In contrast, the proposed model demonstrated its strength and robustness by achieving the highest stability metrics (ARI and NMI) for both databases.

This indicates the success and reliability of the proposed method when used with heterogeneous clinical medical data to obtain stable, meaningful, and interpretable clusters, with the model being reproducible across multiple databases. Conversely, the (Spline + K-means) method performed worse than either the proposed model or the k-means algorithm for both datasets. This performance suggests that smoothing based on cubic functions is unsuitable for this type of non-time medical data, resulting in poor cluster separation.

5. Conclusions

This proposed study aims to uncover hidden clinical patterns in hepatitis C virus (HCV) data using graph-based clustering techniques. The framework applies data preprocessing, followed by the creation of a similarity graph using the nearest neighbors (KNN) algorithm, and then the Louvain algorithm to identify clinically relevant communities. The model was evaluated on two sets of HCV data, and the results showed its effectiveness in detecting homogeneous and overlapping relationships, achieving the highest patterning values at a neighbor size of K=3, reaching 0.740 for the first data set and 0.805 for the second. Furthermore, the low distortion values indicate strong internal coherence within the detected communities. Stability analysis also confirms the robustness and reliability of the proposed framework, as the highest stability metrics were observed for the number of detected communities at the optimal neighbor size (k=3) on both data sets.

The proposed framework demonstrates that data analysis using unsupervised, graph-based machine learning techniques can reveal latent relationships useful for identifying similarities between groups in the data. In future work, this framework can be expanded by applying larger and more diverse medical datasets from different sources. For example, strategies such as incorporating the Leiden algorithm can improve community stability and internal cohesion. Additionally, unsupervised deep learning techniques, including graphical neural networks (GNNs) and variable graphical automatic coders (VGAEs), can also be applied to enhance community discovery.

Reference

[1] M. A. A. Mohammed M. El Behery, Ahmed I. Elghwab, Ashraf A. Tabll, Elsherbeny H. Elsayed, “Review Article: Egyptian Efforts to Control Hepatitis C Virus in Egypt .,” 2025.

[2] N. E. Izere Salomon, Sibomana Olivier, “Advancing Hepatitis C Elimination in Africa: Insights from Egypt.,” 2024.

[3] M. B. Mojca Matičič, “Towards eliminating hepatitis C as a public health threat: different speeds, different needs,” 2024.

[4] H. H. A. Elbahrawy, Ashraf, Marwa K. Ibrahim, Ahmed Eliwa, Ali Madian Mohamed Alboraie, “Current situation of viral hepatitis in Egypt,” 2021.

[5] I. Gomaa, Asmaa, Mohamed Gomaa, Naglaa Allam, “Hepatitis C Elimination in Egypt: Story of Success.,” 2024. [Online]. Available: https://doi.org/10.3390/pathogens13080681

[6] W. K. Al-hamoudi and Department, “Management of Hepatitis C Genotype 4 in the Liver Transplant Setting,” 2016. doi: 10.4103/1319-3767.182453.

[7] S. Mahmud, Z. Al-Kanaani, H. Chemaitelly, K. Chaabna, S. P. Kouyoumjian, and L. J. Abu-Raddad, “Hepatitis C virus genotypes in the Middle East and North Africa: Distribution, diversity, and patterns,” 2018. doi: 10.1002/jmv.24921.

[8] H. A. A. Khudhair* and K. R. H. , Ali A. H. Albakaa, “Detecting the Prevalence of Hepatitis C Virus among Iraqi People.,” 2023.

[9] S. D. Warkad, K. Song, and D. Pal, “Developments in the HCV Screening Technologies Based on the Detection of Antigens and Antibodies,” 2019.

[10] G. S. Vincenzo Moscato, “Community detection over feature-rich information networks: An eHealth case study,” *ELSEVIER*, 2022, [Online]. Available: https://doi.org/10.1016/j.is.2022.102092

[11] C. Krantsevich, “Digital medicine and the curse of dimensionality,” pp. 1–8, 2021, doi: 10.1038/s41746-021-00521-5.

[12] L. G. Buch, Amanda M., Conor Liston, “Simple and Scalable Algorithms for Cluster-Aware Precision Medicine,” 2023.

[13] C. Simpson *et al.*, “Lifting the curse from high-dimensional data : automated projection pursuit clustering for a variety of biological data modalities,” pp. 1–20, 2025, doi: 10.1093/gigascience/giaf052.

[14] J. B. L. RYAN A. ROSSI, DI JIN, SUNGCHUL KIM, NESREEN K. AHMED, DANAI KOUTRA, “On Proximity and Structural Role-based Embeddings in Networks: Misconceptions, Techniques, and Applications.,” 2020.

[15] L. Brahim, M. Mouad, C. Chihab-eddine, and I. Ali, “A SURVEY ON COMMUNITY DETECTION : APPLICATIONS , ALGORITHMS , AND CHALLENGES,” vol. 102, no. 12, pp. 4923–4945, 2024.

[16] I. K. Isa Inuwa-Dutse, Mark Liptrott, “A multilevel clustering technique for community detection.,” 2021.

[17] W. Zhao, J. Luo, T. Fan, and Y. X. Ren, Yan, “Analyzing and visualizing scientific research collaboration network with core node evaluation and community detection based on network embedding,” 2021, [Online]. Available: https://doi.org/10.1016/j.patrec.2021.01.007

[18] 2 MEHRDAD ROSTAMI 1, MOURAD OUSSALAH 1, 2, (Senior Member, IEEE), KAMAL BERAHMAND 3, AND VAHID FARRAHI 1, “Community Detection Algorithms in Healthcare Applications: A Systematic Review.,” 2023, *IEEE*.

[19] J.-K. H. Goudet, Olivier, B´ eatrice Duval, “Population-based Gradient Descent Weight Learning for Graph Coloring Problems,” 2020.

[20] J. Ren, X. Lyu, J. Guo, X. Shi, Y. Zhou, and Q. Li, “CDSKNN XMBD : a novel clustering framework for large - scale single - cell data based on a stable graph structure,” *J. Transl. Med.*, pp. 1–11, 2024, doi: 10.1186/s12967-024-05009-w.

[21] V. D. Blondel, J. Guillaume, and E. Lefebvre, “Fast unfolding of communities in large networks,” pp. 1–12, 2008.

[22] R. G. Dipika Singh, “NI-Louvain: A novel algorithm to detect overlapping communities with influence analysis.pdf,” 2022, *Elsevier*.

[23] S. Seth, S. Mallik, T. Bhadra, and Z. Zhao, “Dimensionality Reduction and Louvain Agglomerative Hierarchical Clustering for Cluster-Speci fi ed Frequent Biomarker Discovery in Single-Cell Sequencing Data,” vol. 13, no. February, pp. 1–17, 2022, doi: 10.3389/fgene.2022.828479.

[24] W. Makelaardij, “A cost-based multi-layer network approach for the discovery of patient phenotypes Clara,” 2023.

[25] A. M. Elshewey *et al.*, “Optimizing HCV Disease Prediction in Egypt : The hyOPTGB Framework,” 2023.

[26] H. M. Farghaly, M. Y. Shams, and T. A. El-hafeez, “Hepatitis C Virus prediction based on machine learning framework : a real-world case study in Egypt,” *Knowl. Inf. Syst.*, vol. 65, no. 6, pp. 2595–2617, 2023, doi: 10.1007/s10115-023-01851-4.

[27] B. Pankratz and P. Prałat, “Performance of community detection algorithms supported by node embeddings,” 2024.

[28] A. A. Permana and R. M. Yaputra, “Analyzing breast cancer comorbidities : a network approach using community detection algorithms,” *Appl. Netw. Sci.*, 2024, doi: 10.1007/s41109-024-00644-0.

[29] J. Bambi *et al.*, “Approaches to Extracting Patterns of Service Utilization for Patients with Complex Conditions : Graph Community Detection

vs . Natural Language Processing Clustering,” no. 1, pp. 1884–1900, 2024.

[30] M. Bampa, I. Miliou, B. Jovanovic, and P. Papapetrou, “M-ClustEHR : A multimodal clustering approach for electronic health records,” vol. 154, no. May, 2024.

[31] A. Sharma, T. Khade, and S. M. Satapathy, “A cross dataset meta-model for hepatitis C detection using multi- dimensional pre-clustering,” pp. 1–17, 2025.

[32] N. J. van E. Traag, V. A., L. Waltman, “From Louvain to Leiden: guaranteeing well-connected communities,” 2019.