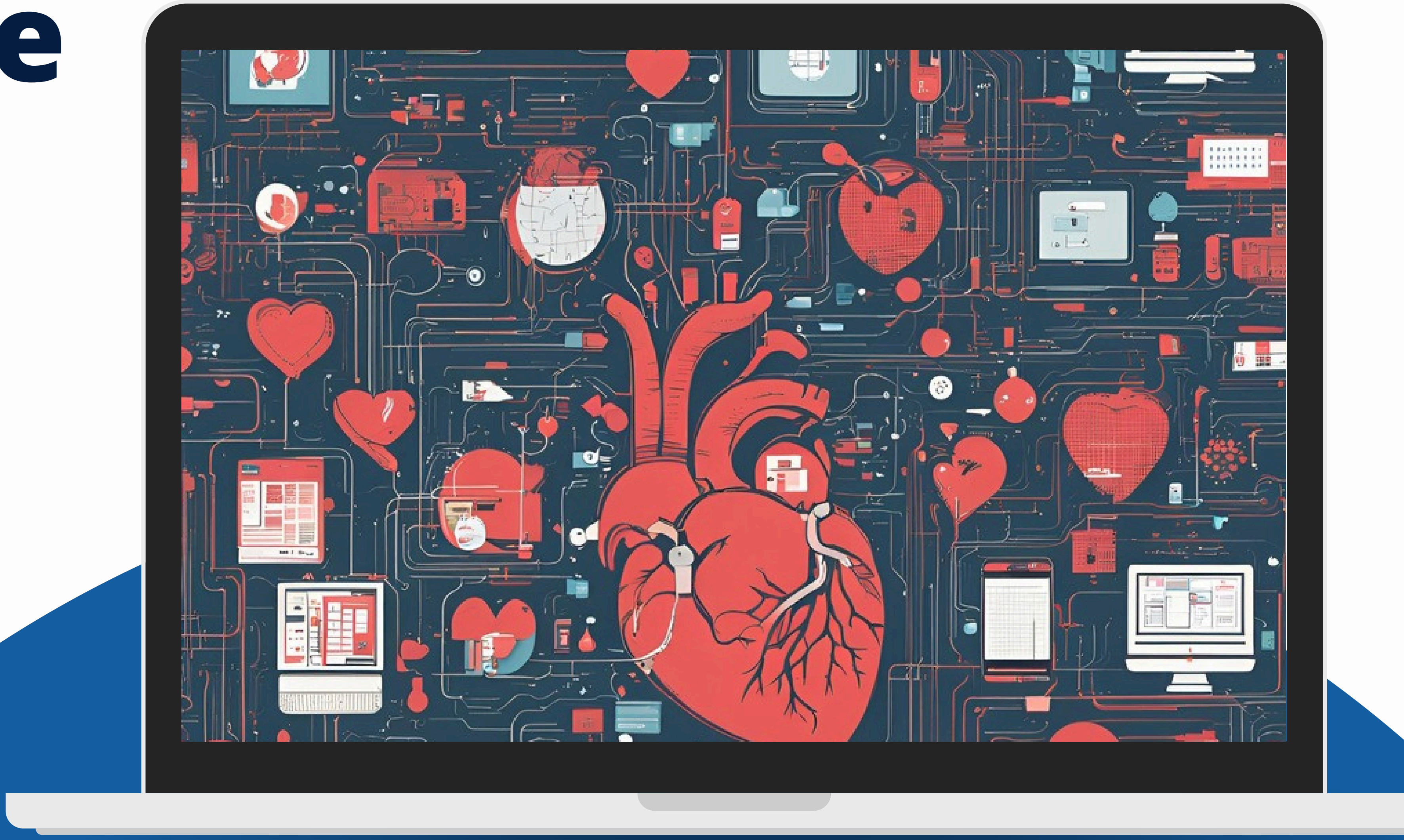
JCT
LEV ACADEMIC CENTER

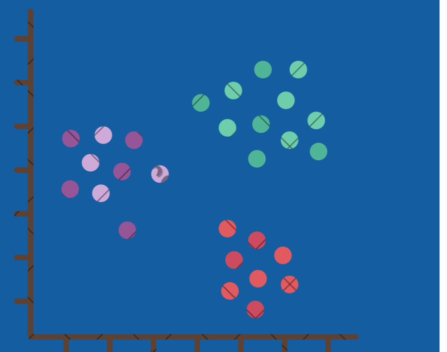
Efficient genetic K-Means clustering for health care knowledge discovery

A. Alsayat and H. El-Sayed, 2016, IEEE



By: Rachelli Adler, Esther Malka Nusbacher

Introduction



- Data mining and machine learning are crucial for healthcare decision-making.
- Clustering helps segment patients for effective treatments. similar patience might have similar treatments.

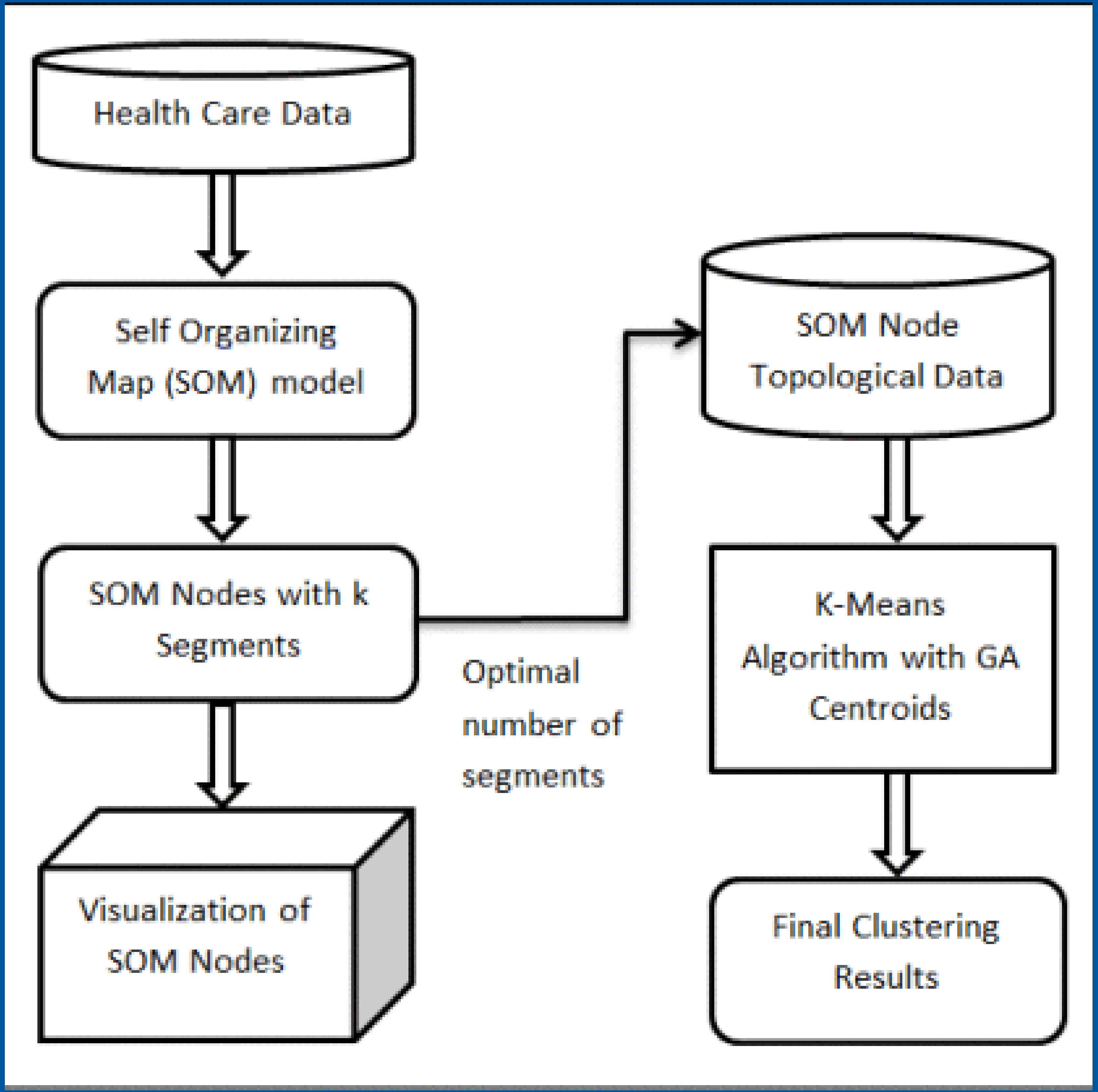
Problem Statement

- How can we improve clustering accuracy in healthcare data?
- Traditional K-Means struggles with selecting the right number of centroids. and may not give the best reults.

Methods

- K-Means Clustering
- Genetic Algorithm
- DBSCAN
- Self Organizing Map (SOM)

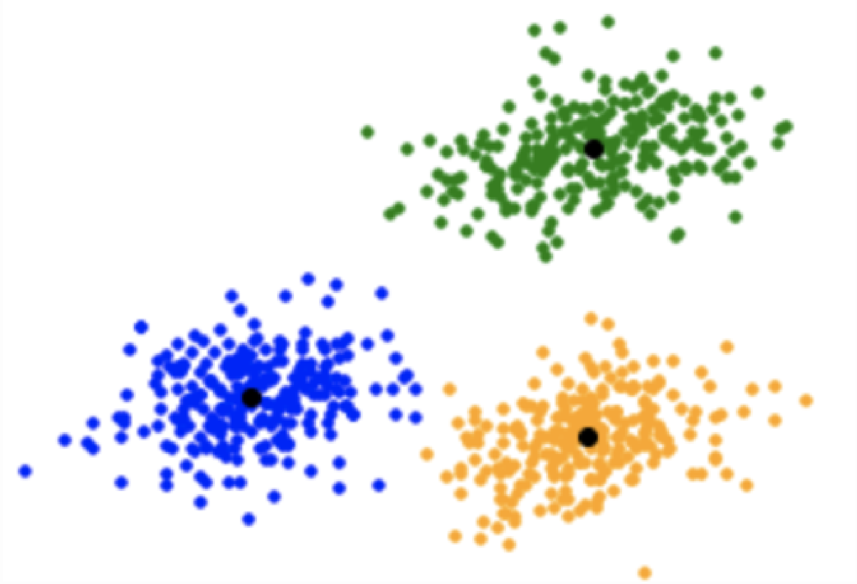
“We propose an **efficient K-Means** clustering algorithm which uses the **SOM** method to discover the optimal segments number in the data as a preprocessing step”


Health Care Data
Self Organizing Map (SOM) model
SOM Nodes with k Segments
Visualization of SOM Nodes
Optimal number of segments
SOM Node Topological Data
K-Means Algorithm with GA Centroids
Final Clustering Results

The Code

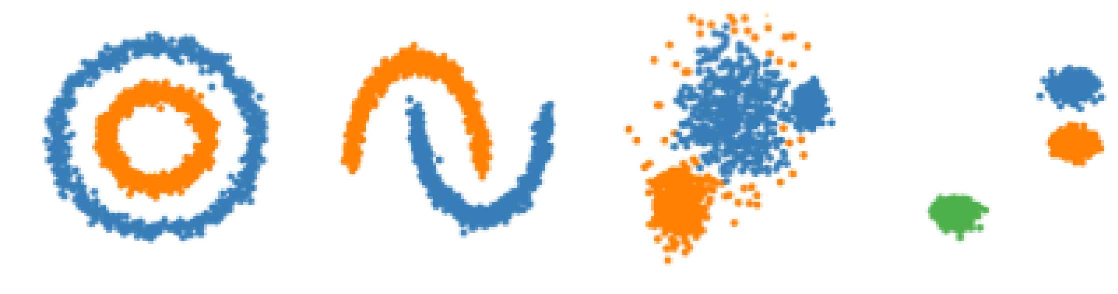
K-Means

- unsupervised
- Groups data into k clusters by assigning points to the nearest centroid and updating centroids repeatedly.



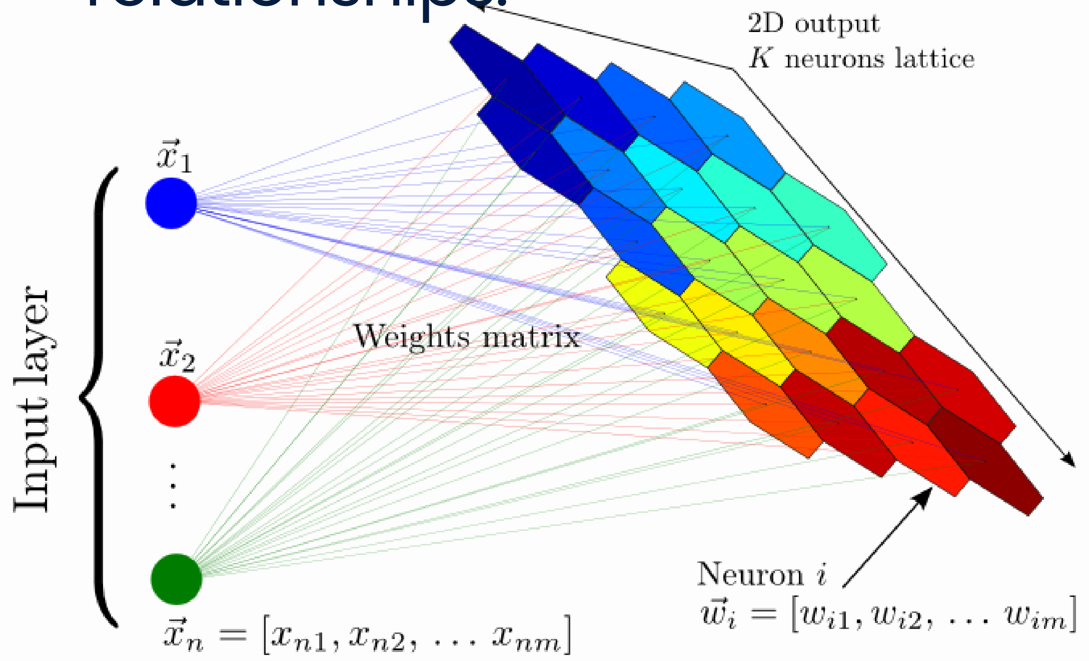
DBSCAN

- unsupervised
- Finds clusters robust to outliers.
- Requires parameters: ε (neighborhood size) and MinPts (minimum points)



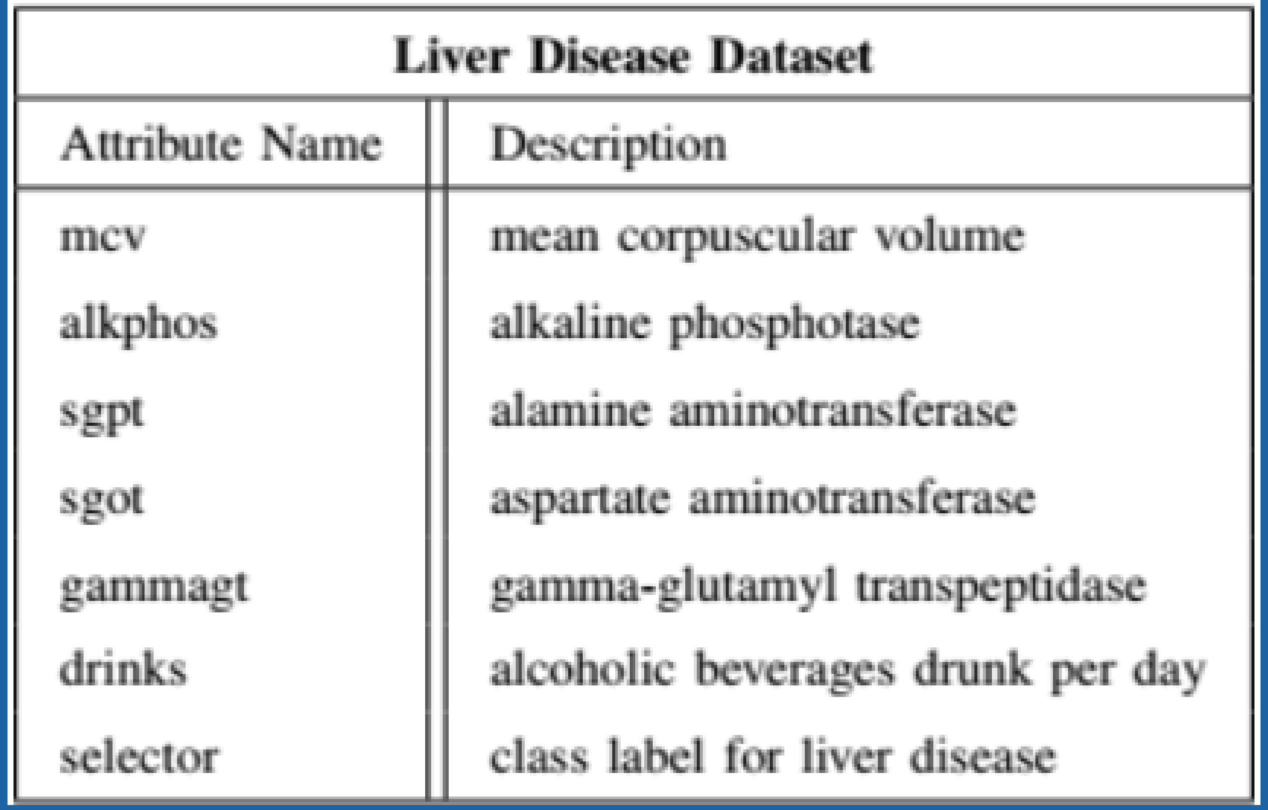
Self Organizing Map

- unsupervised
- Artificial Neural Network
- Maps high-dimensional data onto a lower-dimensional grid while preserving topological relationships.


2D output
K neurons lattice
$\vec{x}_1$
$\vec{x}_2$
Weights matrix
Input layer
Neuron i
$\vec{w}_i = [w_{i1}, w_{i2}, \dots w_{im}]$
$\vec{x}_n = [x_{n1}, x_{n2}, \dots x_{nm}]$

The Data

We used two datasets on **liver disease** and **heart disease.**

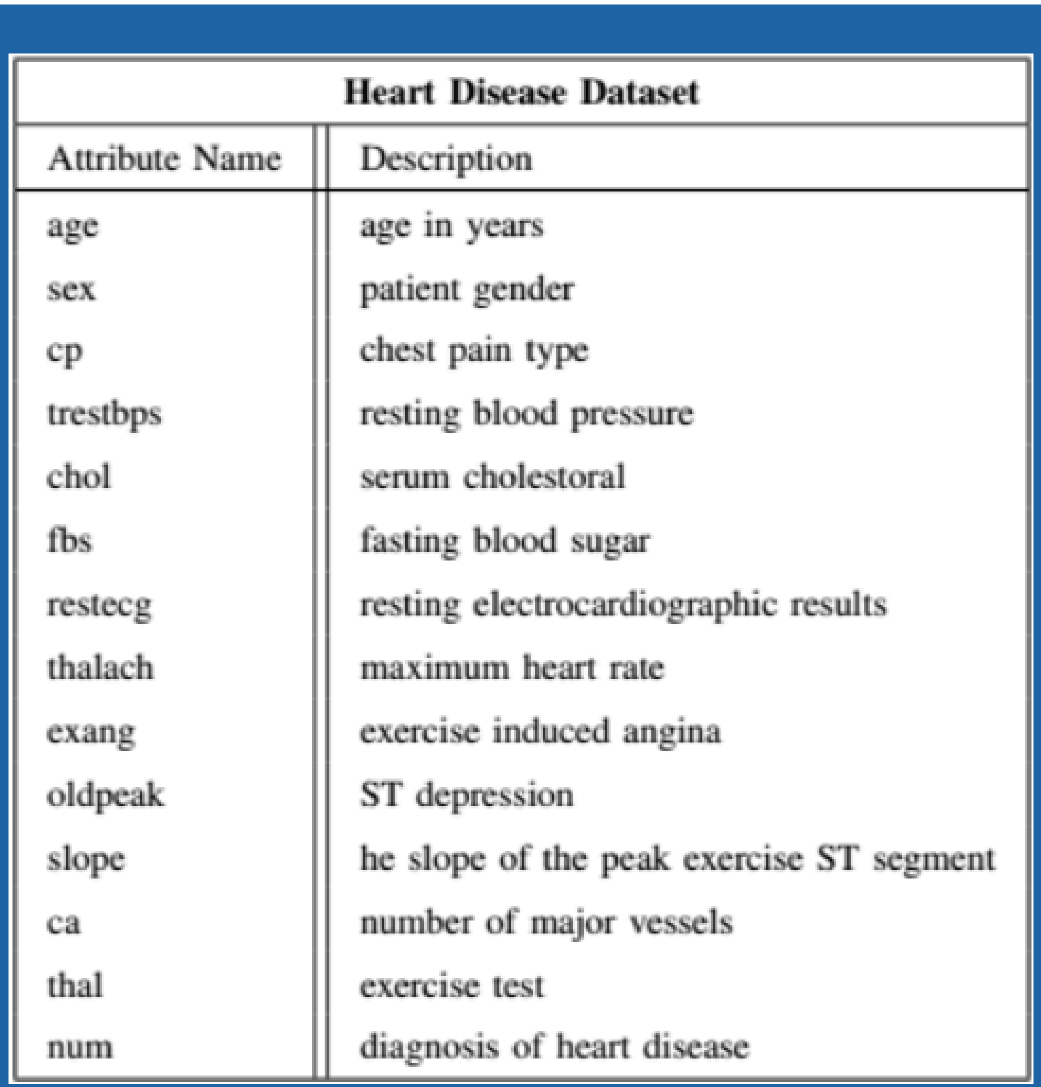


Liver Disease Dataset

Attribute Name	Description
mcv	mean corpuscular volume
alkphos	alkaline phosphotase
sgpt	alamine aminotransferase
sgot	aspartate aminotransferase
gammagt	gamma-glutamyl transpeptidase
drinks	alcoholic beverages drunk per day
selector	class label for liver disease

- 345 patients
- 7 blood test variables related to liver disease (e.g., alcohol-related)
- Class label: drinks > 2

from UCI Machine Learning Repository



Heart Disease Dataset

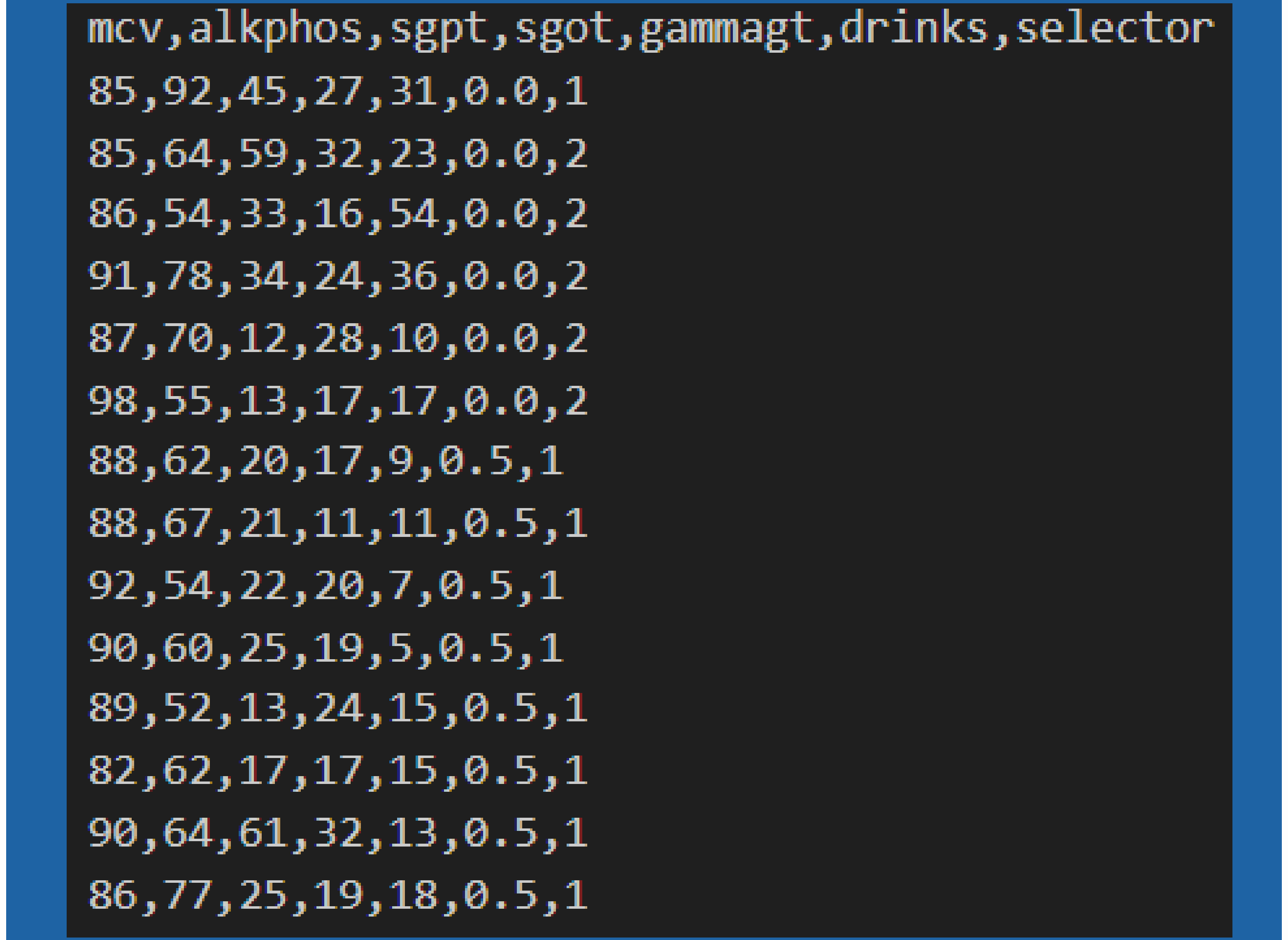
Attribute Name	Description
age	age in years
sex	patient gender
cp	chest pain type
trestbps	resting blood pressure
chol	serum cholestoral
fbs	fasting blood sugar
restecg	resting electrocardiographic results
thalach	maximum heart rate
exang	exercise induced angina
oldpeak	ST depression
slope	he slope of the peak exercise ST segment
ca	number of major vessels
thal	exercise test
num	diagnosis of heart disease

- 303 patients
- 14 variables related to heart disease diagnosis
- Class label: target = 1

from the Cleveland Clinic Foundation

Preprocessing The Data

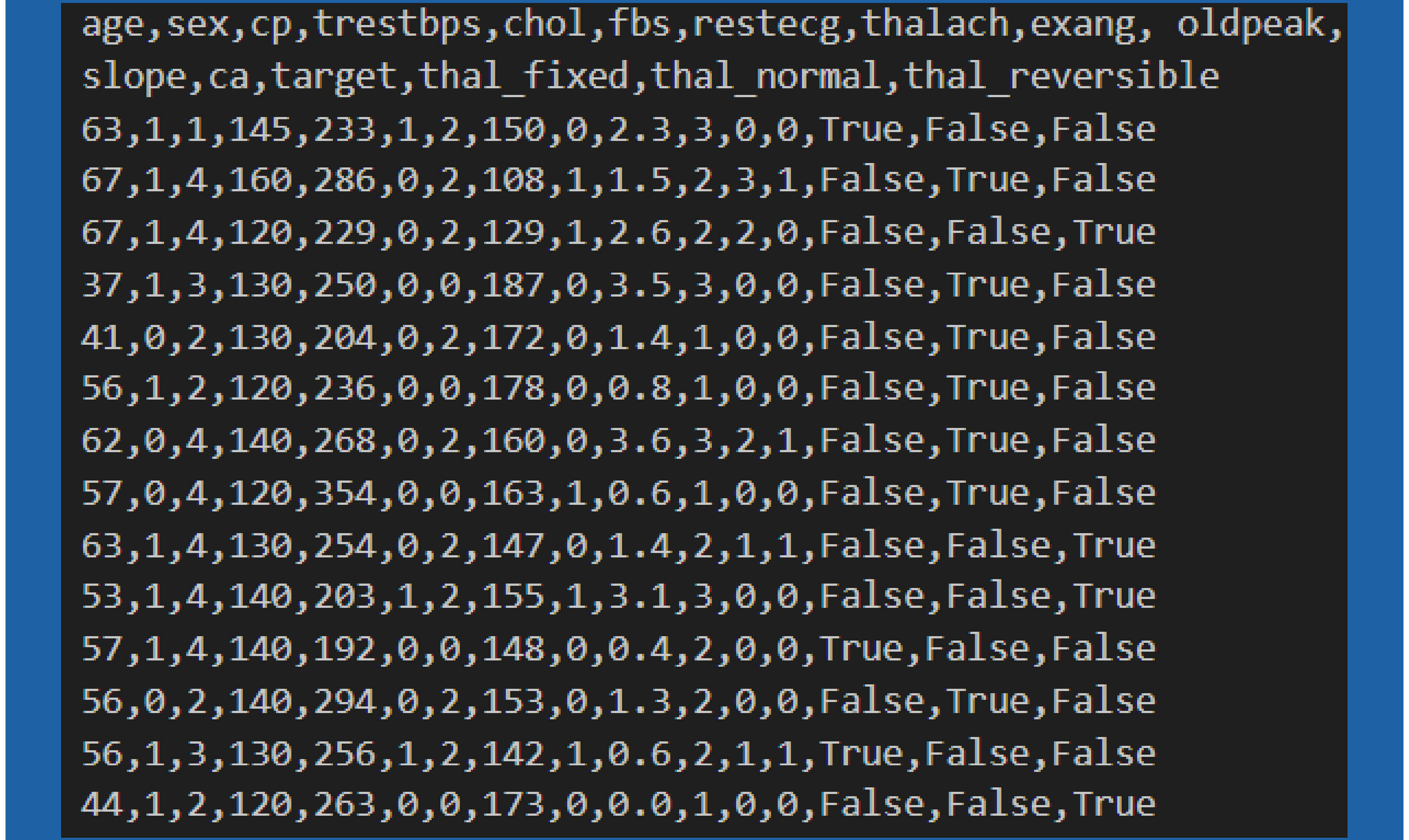
Liver Disease Dataset



```
mcv,alkphos,sgpt,sgot,gammagt,drinks,selector
85,92,45,27,31,0.0,1
85,64,59,32,23,0.0,2
86,54,33,16,54,0.0,2
91,78,34,24,36,0.0,2
87,70,12,28,10,0.0,2
98,55,13,17,17,0.0,2
88,62,20,17,9,0.5,1
88,67,21,11,11,0.5,1
92,54,22,20,7,0.5,1
90,60,25,19,5,0.5,1
89,52,13,24,15,0.5,1
82,62,17,17,15,0.5,1
90,64,61,32,13,0.5,1
86,77,25,19,18,0.5,1
```

- ✓ No preprocessing required
- ✓ No train/test split needed

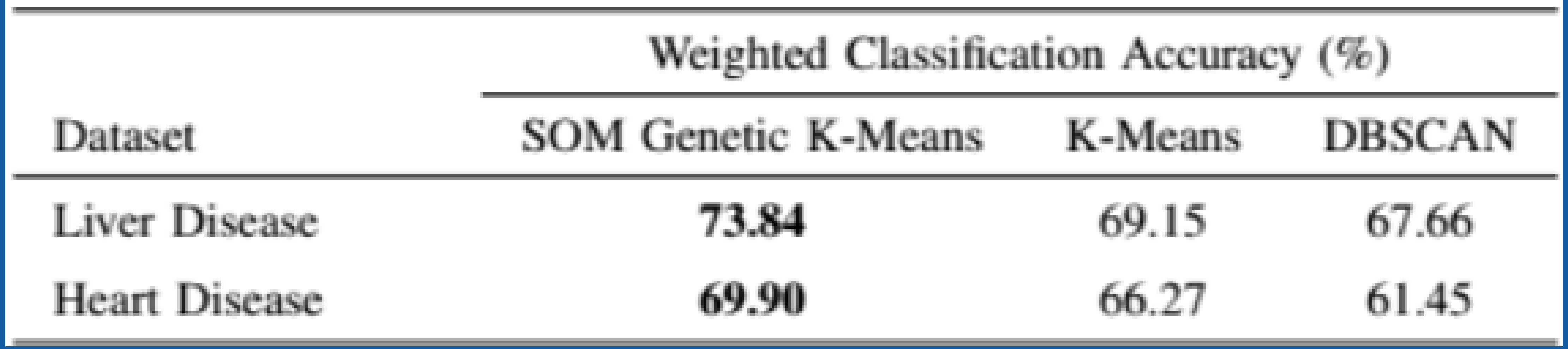
Heart Disease Dataset



```
age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang, oldpeak,
slope,ca,target,thal_fixed,thal_normal,thal_reversible
63,1,1,145,233,1,2,150,0,2.3,3,0,0,True,False,False
67,1,4,160,286,0,2,108,1,1.5,2,3,1,False,True,False
67,1,4,120,229,0,2,129,1,2.6,2,2,0,False,False,True
37,1,3,130,250,0,0,187,0,3.5,3,0,0,False,True,False
41,0,2,130,204,0,2,172,0,1.4,1,0,0,False,True,False
56,1,2,120,236,0,0,178,0,0.8,1,0,0,False,True,False
62,0,4,140,268,0,2,160,0,3.6,3,2,1,False,True,False
57,0,4,120,354,0,0,163,1,0.6,1,0,0,False,True,False
63,1,4,130,254,0,2,147,0,1.4,2,1,1,False,False,True
53,1,4,140,203,1,2,155,1,3.1,3,0,0,False,False,True
57,1,4,140,192,0,0,148,0,0.4,2,0,0,True,False,False
56,0,2,140,294,0,2,153,0,1.3,2,0,0,False,True,False
56,1,3,130,256,1,2,142,1,0.6,2,1,1,True,False,False
44,1,2,120,263,0,0,173,0,0.0,1,0,0,False,False,True
```

- ✓ **preprocessing required:** One-hot encoding and dummy variables
- ✓ No train/test split needed

Articles results

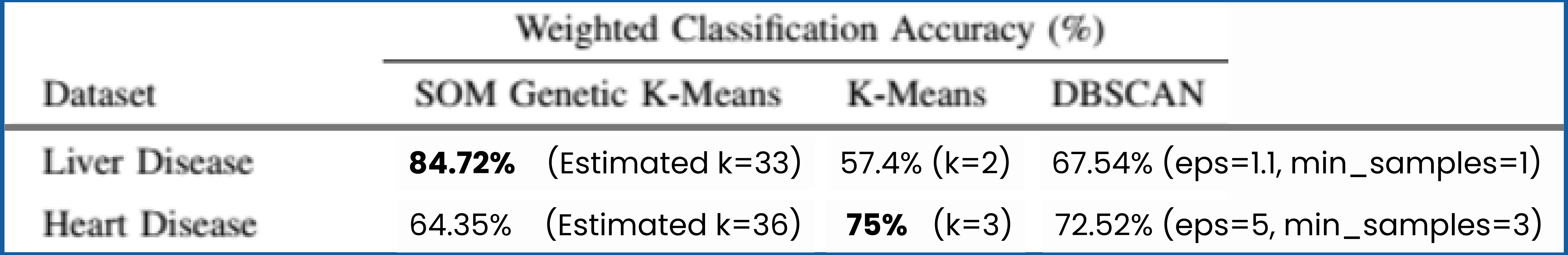


	Weighted Classification Accuracy (%)		
Dataset	SOM Genetic K-Means	K-Means	DBSCAN
Liver Disease	**73.84**	69.15	67.66
Heart Disease	**69.90**	66.27	61.45

Key Findings

SOM–Genetic K-Means achieved higher accuracy, leading to more accurate treatment suggestions.

Our Results



	Weighted Classification Accuracy (%)		
Dataset	SOM Genetic K-Means	K-Means	DBSCAN
Liver Disease	**84.72%** (Estimated k=33)	57.4% (k=2)	67.54% (eps=1.1, min_samples=1)
Heart Disease	64.35% (Estimated k=36)	**75%** (k=3)	72.52% (eps=5, min_samples=3)

Key Findings

Unlike the article, SOM Genetic k-means improved clustering for the liver dataset, but not for the heart dataset.

Parameter Optimization

Clustering performance may be improve by tuning key parameters such as learning rate and grid size (SOM), k (K-Means), and eps/min_samples (DBSCAN).

Project Breakdown

✓	**Dataset Preparation**	• Find and download the datasets. • Clean data- Use one-hot encoding for categorical features.
✓	**Baseline Clustering**	• Run K-Means clustering- Choose appropriate k values. • Run DBSCAN- Tune eps and min_samples.
✓	**Advanced Clustering**	• Implement Self-Organizing Maps (SOM)- Train SOM on the dataset. • Apply Genetic K-Means- Use SOM output as input for clustering.
✗	**Evaluate performances**	• Compare K-Means, DBSCAN and som’s results. • Generate visualizations- using graphs and plots.
✗	**Present conclusions**	• Summarize findings and write conclusions

Bibliography

01

THE ARTICLE:

A. Alsayat and H. El-Sayed, "Efficient genetic K-Means clustering for health care knowledge discovery," 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), Towson, MD, USA, 2016, pp. 45-52, doi: 10.1109/SERA.2016.7516127.
https://ieeexplore.ieee.org/document/7516127

02

DATASET 1 : liver disease

UCI Machine Learning Repository

Discover datasets around the world!

ics.uci.edu

03

DATASET 2: heart disease

Cleveland Clinic Foundation Heart Disease

Kaggle is the world's largest data science community with powerful tools and resources to...

kaggle.com

04

Reference Article:
Richard Forsyth and Roy Rada. 1986. Machine learning: applications in expert systems and information retrieval. Halsted Press, USA.
https://dl.acm.org/doi/abs/10.5555/6736

THANK YOU!