Application of Multidimensional Scaling Model for Hepatitis C Data Dimensionality Reduction

Ievgen Meniailov[a] and Halyna Padalko[b]

[a] *V.N. Karazin Kharkiv National University, Svobody sq, 4, Kharkiv, 61022, Ukraine*
[b] *National Aerospace University "Kharkiv Aviation Institute", Chkalow str., 17, Kharkiv, 61070, Ukraine*

Abstract

Viral hepatitis C is a worldwide disease. About 71 million people worldwide have a chronic form of the disease. Early automated diagnosis of viral hepatitis C is an effective tool for controlling the incidence and providing timely assistance to patients. Using models for automated diagnosis of viral hepatitis C is often tricky due to the overabundance of data used to build the model. Therefore, this study aims to explore a multidimensional scaling model to reduce the dimensionality of patient data with suspected hepatitis C.

Keywords 1

Hepatitis C, dimensionality reduction, multidimensional scaling

1. Introduction

Hepatitis C is an inflammatory liver disease caused by the hepatitis C virus. Hepatitis C virus is a blood-borne virus that is most commonly contracted through contact with blood through unsafe injection practices, transfusion of unscreened blood, injecting drug use, and sexual contact that involves blood [1].

The hepatitis C virus can cause both acute and chronic infections. However, acute infection is usually asymptomatic and, in most cases, does not lead to life-threatening illness. Moreover, about 30% of infected people achieve spontaneous recovery within six months after infection [2]. However, 70% of those infected develop chronic hepatitis C infections. The chronic form is accompanied by an increased risk of liver cirrhosis [3].

The most common ways of transmission of viral hepatitis C are [4]:

- Reuse or insufficient sterilization of medical equipment, especially syringes and needles.
- Transfusion of unscreened blood and blood products.
- Sharing injection equipment while injecting drugs.

Less common transmission mechanisms of the virus are from an infected mother to her child and through sexual contact.

About 71 million people worldwide have a chronic hepatitis C virus [5]. More than 350,000 people die every year from hepatitis C-related liver disease. About 3-4 million people are infected with the hepatitis C virus yearly.

Ukraine belongs to the countries with an average prevalence of hepatitis C [6]. Every year, about 6,000 people become infected with viral hepatitis C. At the same time, a new treatment program launched in 2018 gave access to drugs that are effective against all genotypes of the hepatitis C virus at once.

Primary infection with viral hepatitis C most often occurs asymptomatically. Therefore, the first time after infection, hepatitis is not diagnosed in most infected people. The chronic form of viral

2nd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2022), December 2-4, 2022, Łódź, Poland
EMAIL: evgenii.menyailov@gmail.com (I. Meniailov); galinapadalko95@gmail.com (H. Padalko)
ORCID: 0000-0002-9440-8378 (I. Meniailov); 0000-0001-6014-1065 (H. Padalko)


hepatitis C is also often not diagnosed in the early stages since the disease is asymptomatic until secondary symptoms associated with liver damage are developed.

Diagnosis of infection with viral hepatitis C is carried out in two stages [7]:

1. Serological testing for antibodies to viral hepatitis C.
2. Testing for HCV ribonucleic acid in a patient's blood when an antibody test is positive.

For effective decision-making regarding treating patients, it is necessary to carry out the early diagnosis of the disease. For this, automated tools based on artificial intelligence methods are practical. Such tools are widely used, including for the study of hepatitis [8]. Artificial intelligence systems and tools have shown their effectiveness in the analysis of medical data [9], the analysis of the behavior of viruses [10], the study of various factors affecting the incidence [11], the assessment of resources necessary for the effective operation of healthcare systems [12], and other tasks of data-driven medicine.

However, when building diagnostic models, the problem of information excess arises. To do this, it is necessary to use data dimensionality reduction methods, improving models' accuracy and adequacy.

Therefore, this study aims to build a dimensionality reduction model for these patients with suspected viral hepatitis C based on the multidimensional scaling method.

Research is part of a complex intelligent information system for epidemiological diagnostics, the concept of which is discussed in [13].

2. Materials and Methods

The presence of redundant, non-informative or weakly informative features in the data set can reduce the efficiency of the model. Dimensionality reduction in machine learning is a reduction in the number of features of a dataset [14]. After such a transformation, the model is simplified and the size of the data set is reduced. This reduces the amount of memory required and speeds up the model. The use of this approach is especially important for algorithms that are not scalable, when even a small reduction in the number of entries leads to a significant gain in computational time.

Dimension reduction makes sense to apply when the information necessary for a qualitative solution of the problem is contained in a certain subset of features and it is not necessary to use all of them. This is especially true for correlated traits.

Multidimensional scaling is a technique for visualizing the level of similarity of individual instances of a data set [15]. The method is used to convert information about pairwise distances between a set of n points mapped to an abstract Cartesian space.

Multidimensional scaling exploits the fact that the coordinate matrix X can be obtained by eigenvalue decomposition from

$$B = XX'. \tag{1}$$

Matrix B is computed from proximity matrix D using double centering.

To implement the multidimensional scaling method, you must:

1. To set up square proximity matrix

$$D^2 = \left[d_{ij}^2\right]. \tag{2}$$

2. To apply double centering

$$B = -\frac{1}{2} C D^2 C \tag{3}$$

using centering matrix

$$C = I - \frac{1}{n} J_n, \tag{4}$$

where n is the number of objects, I is the n X n identity matrix, J_n is the n X n matrix.

3. To determine the largest eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$, and the corresponding eigenvectors $e_1, e_2, \ldots, e_m$.

4. Then

$$X = E_m \Lambda_m^{1/2}, \tag{5}$$

where E_m is the matrix of eigenvectors, Λm is the diagonal matrix of eigenvalues.

Within the framework of this study, the SMACOF multidimensional scaling model is considered, the block diagram of which is shown in Figure 1.

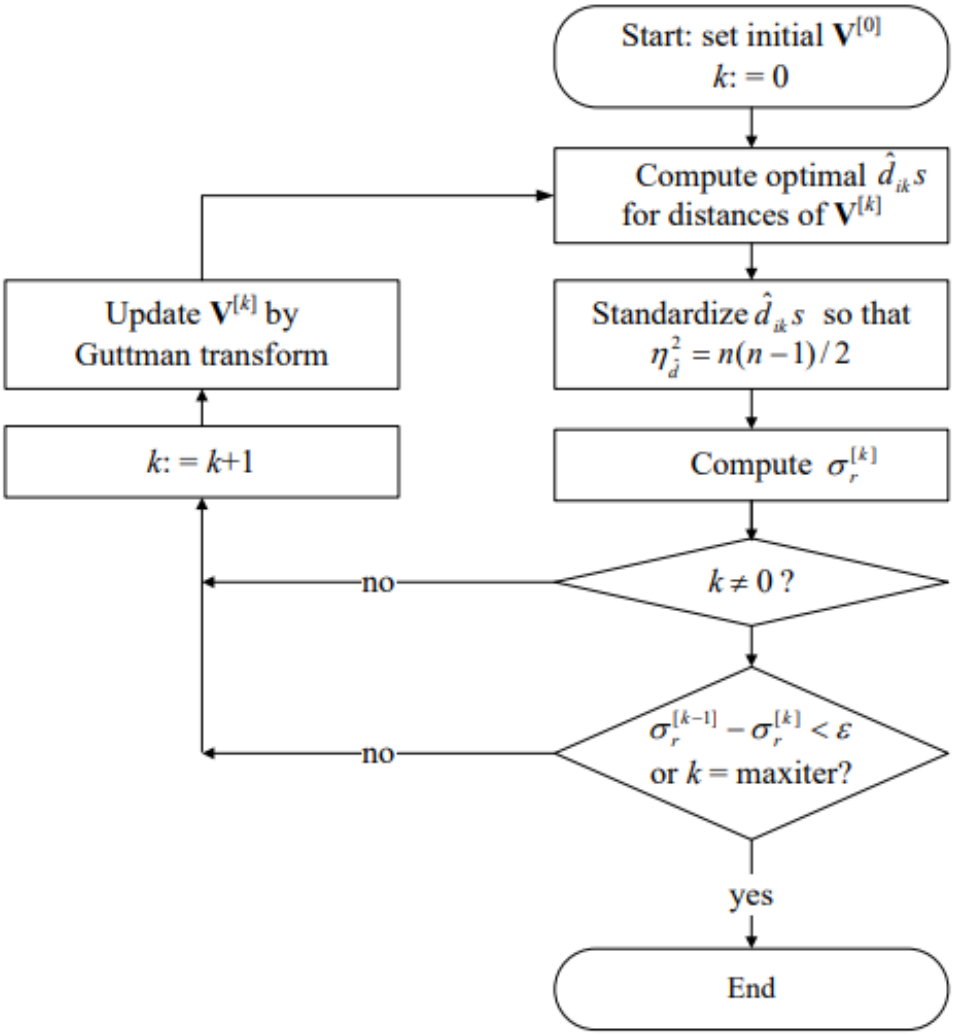

Figure 1: SMACOF multidimensional scaling method

Euclidean Distance [16] and Manhattan Distance [17] were used as model performance metrics.

The Euclidean distance can be calculated from the Cartesian coordinates of points using the Pythagorean theorem. For observations a and b computed in multiple dimensions, the Euclidean distance is:

$$E = \sqrt{\sum_i (a_i - b_i)^2}. \tag{6}$$

Even if scaling, normalization, or dimension weighting is used, the distance measure will still be the result. Therefore, Euclidean distance is a good default distance measure to use if it makes sense to combine dimensions.

According to the Manhattan distance, the distance between two points is equal to the sum of the modules of the differences in their coordinates:

$$M = \sum_{i=1}^{N} |a_i - b_i|. \quad (7)$$

The Manhattan distance depends on the inversion of the coordinate system, but does not depend on its mapping to the coordinate axis or shift.

3. Results

For the experimental investigation, a data set of patients with suspected hepatitis C was used [18]. After preprocessing, this set includes 152 patients, the list of attributes of which is presented in Table 1.

Table 1
Attributes of the dataset

Name	Type of the scale	Range
Class	Nominal	Die/Live
Age	Metric	10..80
Sex	Nominal	Male/Female
Steroid	Nominal	No/Yes
Antivirals	Nominal	No/Yes
Fatigue	Nominal	No/Yes
Malaise	Nominal	No/Yes
Anorexia	Nominal	No/Yes
Liver Big	Nominal	No/Yes
Spleen Palpable	Nominal	No/Yes
Spiders	Nominal	No/Yes
Ascites	Nominal	No/Yes
Varices	Nominal	No/Yes
Bilirubin	Metric	0,39..4,00
Alk Phosphate	Metric	33..250
Sgot	Metric	13.500
Albumin	Metric	2,1..6,0
Protime	Metric	10..90
Histology	Nominal	No/Yes

Visualization of some parameters is presented in Figure 2.

The software implementation for data dimensionality reduction by multidimensional scaling was considered in the Python programming language in the Anaconda programming environment.

The import of the data is presented in Figure 3.

The Manhattan distance matrix is presented in Table 2.

The Euclidean distance matrix is presented in Table 3.

Visualization of the obtained results is presented in Figures 4-5.

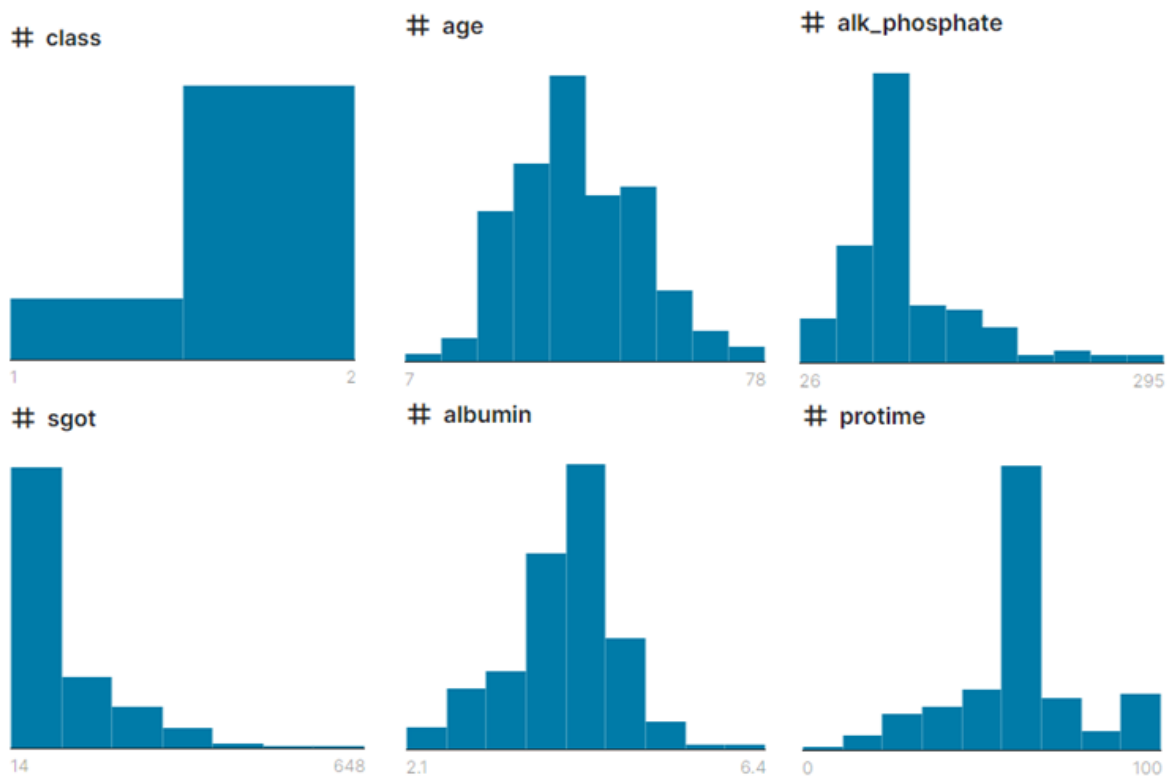

Figure 2: Visualization of the data.

	Class	Age	Sex	Steroid	...	SGOT	Albumin	Protine	Histology
0	Live	30	Female	No	...	18	4.0	0	No
1	Live	50	Male	No	...	42	3.5	0	No
2	Live	78	Male	Yes	...	32	4.0	0	No
3	Live	31	Male		...	52	4.0	80	No
4	Live	34	Male	Yes	...	200	4.0	0	No
..	...	...	...	...	...	...	...	...	...
147	Die	46	Male	Yes	...	242	3.3	50	Yes
148	Live	44	Male	Yes	...	142	4.3	0	Yes
149	Live	61	Male	No	...	20	4.1	0	Yes
150	Live	53	Female	No	...	19	4.1	48	Yes
151	Die	43	Male	Yes	...	19	3.1	42	Yes

[152 rows x 20 columns]

Figure 3: Import of the data.

Table 3
Manhattan MDS

[[0	68	62	...	48	56	53]
[68	0	52	...	66	80	65]
[62	52	0	...	38	66	67]
...	...	...	...	...	...	...
[48	66	38	...	0	42	65]
[56	80	66	...	42	0	41]
[53	65	67	...	65	41	0]]

Table 4
Euclidean MDS

[[0	35.4964787	40.86563348	…	31.04834939	29.3257566	21.56385865]
[35.4964787	0	27.4226184	…	36.02776707	36.46916506	27.51363298]
[40.86563348	27.4226184	0	…	17.94435844	27.89265136	31.95309062]
…	…	…	…	…	…	…
[31.04834939	36.02776707	17.94435844	…	0	20.24845673	27.54995463]
[29.3257566	36.46916506	27.89265136	…	20.24845673	0	17.63519209]
[21.56385865	27.51363298	31.95309062	…	27.54995463	17.63519209	0]]

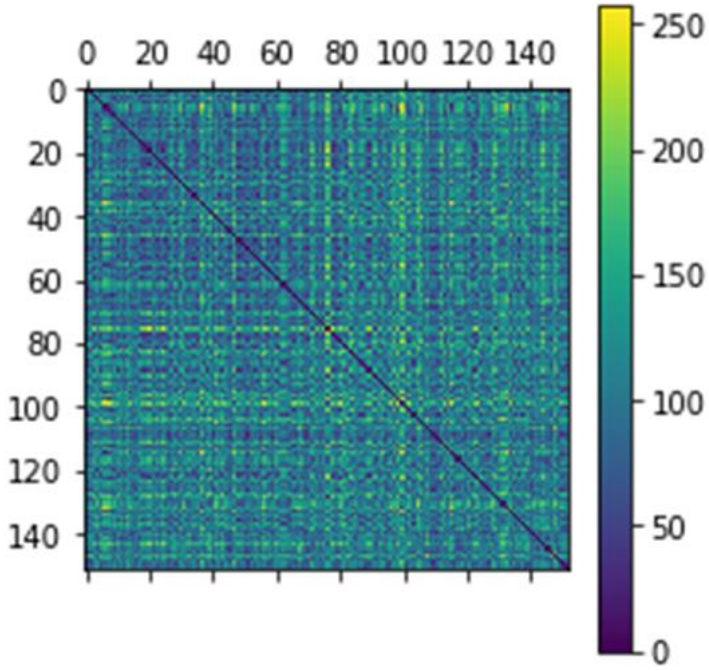

Figure 4: Manhattan MDS.

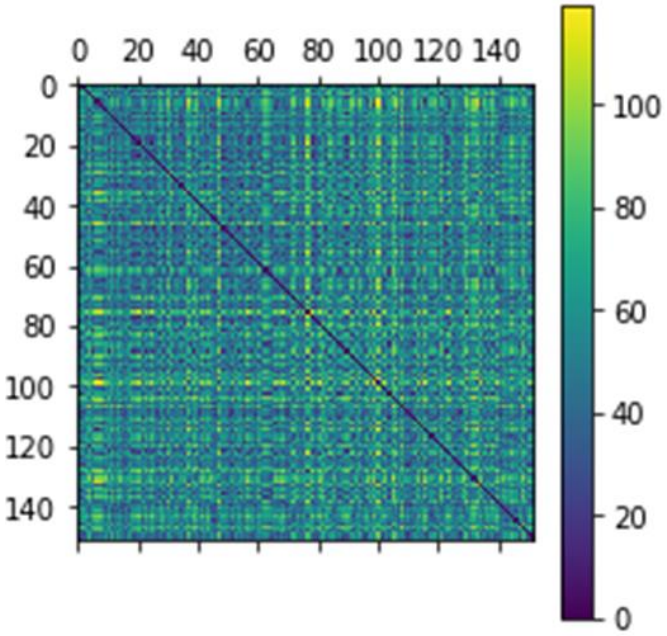

Figure 5: Euclidean MDS.

Stress value according to Manhattan distance MDS is 0.1259161019273947.

Stress value according to Euclidean distance MDS is 0.08271906360960252.

As we can see, the stress factor for multidimensional scaling based on Euclid distance is smaller, which suggests that the new dataset based on Manhattan distance has more errors.

The new data sets contain information about 152 patients, namely the x,y,z coordinates of the points depicting each patient. New data set according to Manhattan distance MDS is presented in Table 5.

Table 5

New database (Manhattan MDS)

[[20.2938038	-44.1092983	30.5774195]
[-13.5033604	-50.4097332	-22.7673883]
[-0.378301973	-70.1431781	13.8080757]
[58.6541139	38.7255543	34.2156564]
[-41.0441253	66.7508857	40.7634540]
[62.2854220	-11.8633235	-3.05123159]
[-21.1234832	-45.1411934	92.7766274]
[-1.89588739	-23.8666618	100.310135]
[-3.90607280	7.89892377	61.2980443]
[-31.0997297	48.2789317	46.6093780]
[72.3965905	-2.33119261	7.89822863]
[-7.67626985	86.3445669	-2.95718356]
[17.3602805	8.92972452	-4.93874964]
[28.2649986	94.8415580	-0.361476353]
[-42.5757988	-0.803932683	67.4518015]
[-19.9791194	34.4316643	-4.16562869]
[-24.2660481	-47.4501230	-4.58395112]
[...	...	...]]

New data set according to Euclidean distance MDS is presented in Table 6.

Table 6

New database (Euclidean MDS)

[[14.34082277	-14.72113167	28.28945039]
[29.68727476	-14.70348599	-0.38039605]
[20.71660585	-29.27847018	12.35350779]
[-21.83629431	23.7830797	20.52382905]
[-48.48090021	-8.33583728	-21.57916453]
[15.01731202	20.31933842	20.97493686]
[-13.09174086	-30.41483387	41.96942425]
[-18.83816817	-20.50830841	47.18319403]
[-27.88528687	-14.41177379	20.41489463]
[-42.18887234	-9.08665496	-6.30699151]
[4.20155951	25.09237554	24.13960161]
[-35.86184375	13.86963771	-31.11551236]
[-1.8425913	5.84892742	2.00522366]
[-31.54290666	29.65109918	-13.55392365]
[-30.75246417	-23.0910951	11.85288018]
[-13.16247211	-23.0910951	11.85288018]
[...	...	...]]

Figure 6 is a graphical representation of the new dataset in 3D space obtained by multidimensional scaling using Manhattan distance.

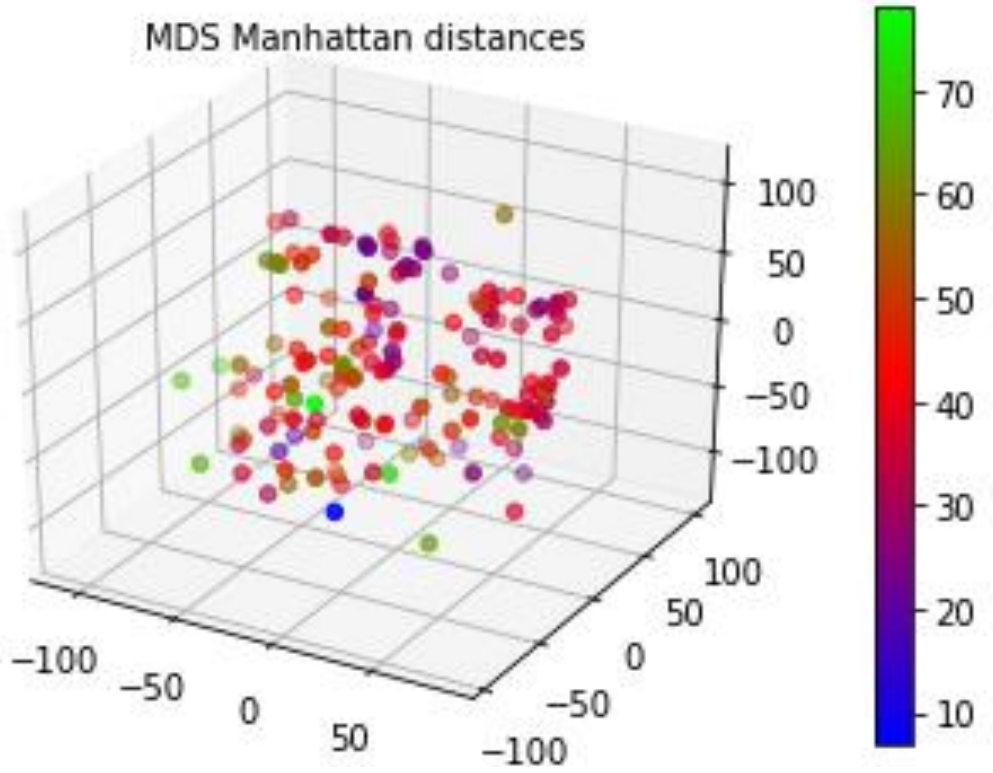

Figure 6: Visualization of the data (Manhattan MDS).

Figure 7 shows a visual representation of the new data set obtained using MDS based on Euclidean distance.

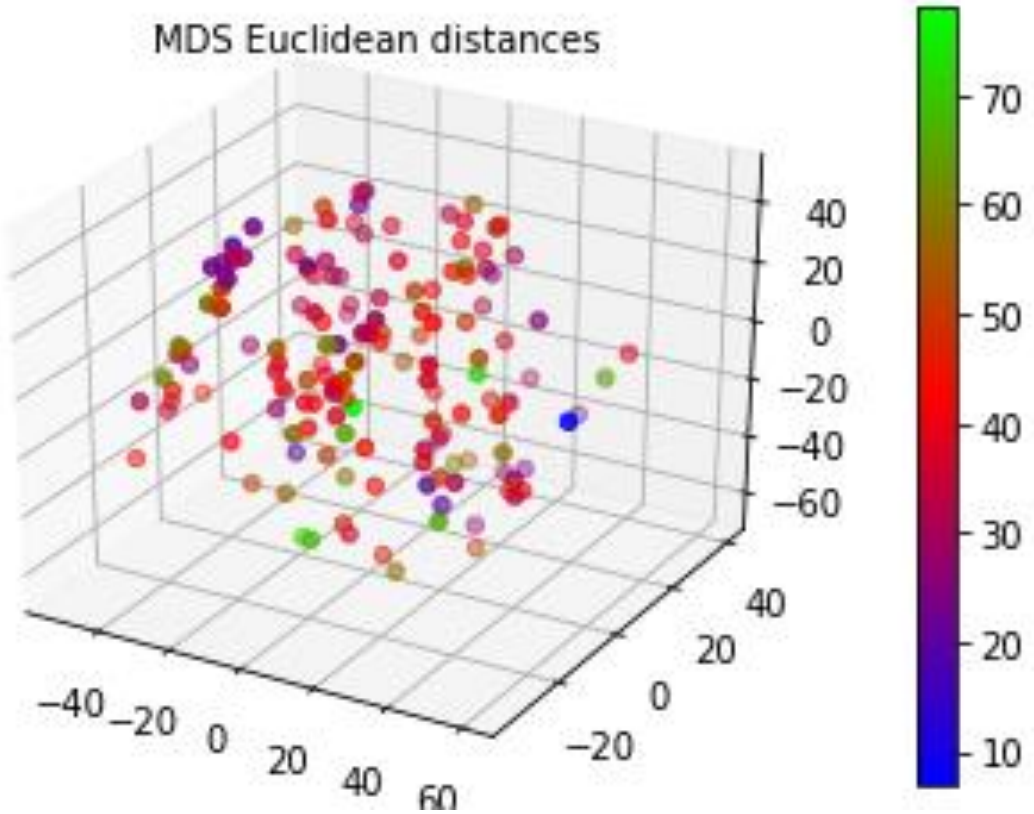

Figure 7: Visualization of the data (Euclidean MDS).

In general, the results obtained are a graphical representation of the dissimilarity of patients among themselves. The new data do not contain the original signs, but they include, so to speak, a comparison of patients with each other, which makes it possible to judge their relationship with each other. This can be explained as follows: the closer the points are to each other, the more similar their initial parameters were. If the point is isolated, this can be explained by the fact that the indicators of the patient displayed by this point differ from the parameters of other patients. Therefore, this patient is unlike others.

4. Conclusions

The task of data dimensionality reduction is relevant in the context of applying machine learning models to real data. Such methods are of particular relevance when analyzing medical data to support physicians' decision making.

Hepatitis C is a common and dangerous disease throughout the world. Therefore, within the framework of this study, a data dimensionality reduction model based on the multidimensional scaling method was developed for these patients with suspected hepatitis C.

As a result, the number of attributes has been reduced from 20 to 3. The model shows high performance, as evidenced by the stress value according to the Manhattan distance, which is 0.126, and the stress value according to the Euclidean distance, which is 0.083.

5. Acknowledgements

The study was funded by the National Research Foundation of Ukraine in the framework of the research project 2020.02/0404 on the topic “Development of intelligent technologies for assessing the epidemic situation to support decision-making within the population biosafety management”.

6. References

[1] T.L. Applegate, E. Fajardo, J.A. Sacks, Hepatitis C virus diagnosis and the holy grail, Infectious disease clinics of North America 32 (2) (2018): 425-445. doi: 10.1016/j.idc.2018.02.010

[2] A. Lombardi, M.U. Mondelli, Hepatitis C: is eradication possible?, Liver International: official journal of the International Association for the Study of the Liver 39 (3) (2019): 416-426. doi: 10.1111/liv.14011

[3] D.E. Bailey Jr, D.M. Zucker, Supportive interventions during treatment of chronic Hepatitis C: a review of the literature, Gastroenterology nursing: the official journal of the Society of Gastroenterology Nurses and Associates 43 (5) (2020): E172-E183. doi: 10.1097/SGA.0000000000000488

[4] A. Mane, et al., Phylogenetic analysis of spread of Hepatitis C virus identified during HIV outbreak investigation, Unnao, India, Emerging Infectious Diseases 28 (4) (2022): 725-733. doi: 10.3201/eid2804.211845

[5] V. Solitano, M.C.P. Torres, N. Pugliese, A. Aghemo, Management and treatment of Hepatitis C: are there still unsolved problems and unique populations?, Viruses 13 (6) (2021): 1048. doi: 10.3390/v13061048

[6] M. Benade, et al., Impact of direct-acting antiviral treatment of Hepatitis C on the quality of life of adults in Ukraine, BMC Infectious Diseases 22 (1) (2022): 650. doi: 10.1186/s12879-022-07615-9

[7] J. Grebely, T.L. Applegate, P. Cunningham, J.J. Feld, Hepatitis C point-of-care diagnostics: in search of a single visit diagnosis, Expert Review of molecular diagnostics 17 (12) (2017): 1109-1115. doi: 10.1080/14737159.2017.1400385

[8] D. Chumachenko, On intelligent multiagent approach to viral Hepatitis B epidemic processes simulation, Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing (2018): 415-419. doi: 10.1109/DSMP.2018.8478602
[9] I. Izonin, R. Tkachenko, N. Shakhovska, N. Lotoshynska, The additive input-doubling method based on the SVR with nonlinear kernels: small data approach, Symmetry 13 (4) (2021): 612. doi: 10.3390/sym13040612
[10] D. Chumachenko, et. al., Investigation of statistical machine learning models for COVID-19 epidemic process simulation: random forest, k-nearest neighbors, gradient boosting, Computation 10 (6) (2022): 86. doi: 10.3390/computation10060086
[11] N. Davidich, et. al., Monitoring of urban freight flows distribution considering the human factor, Sustainable Cities and Society 75 (2021): 103168. doi: 10.1016/j.scs.2021.103168
[12] N. Dotsenko, et al., Project-oriented management of adaptive teams' formation resources in multi-project environment, CEUR Workshop Proceedings 2353 (2019): 911-923.
[13] S. Yakovlev, et al., The concept of developing a decision support system for the epidemic morbidity control, CEUR Workshop Proceedings 2753 (2020): 265-274.
[14] J.T. Vogelstein, et. al., Supervised dimensionality reduction for big data, Nature Communications 12 (2021) 2872. doi: 10.1038/s41467-021-23102-2
[15] J. Tzeng, H.H.S. Lu, W.H. Li, Multidimensional scaling for large genomic data sets, BMC Bioinformatics 9 (2008) 179. doi: 10.1186/1471-2105-9-179
[16] A. Ultsch, J. Lotsch, Euclidean distance-optimized data transformation for cluster analysis in biomedical data (EDOtrans), BMC Bioinformatics 23 (2022) 233. doi: 10.1186/s12859-022-04769-w
[17] R. Shahid, S. Bertazzon, M.L. Knudtson, W.A. Ghali, Comparison of distance measures in spatial analytical modeling for health service planning, BMC Health Services Research 9 (2009) 200. doi: 10.1186/1472-6963-9-200
[18] G. Cestnik, I. Konenenko, I. Bratko, Assistant-86: a knowledge-elicitation tool for sophisticated users, Progress in Machine Learning (1987): 31-45.