Detection of Patients with Diabetes Mellitus using Density-Based Spatial Clustering of Applications with Noize

Serhii Krivtsov[1], Ievgen Meniailov[2] and Kyryl Korobchynskyi[1]

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

Abstract

Machine learning is an effective tool for data-driven medicine. Machine learning methods show high accuracy in the direction of automated diagnostics. Diabetes Mellitus is a significant global problem. Today, more than 400 million people live with this diagnosis. Within the framework of this study, a model for diagnosing patients with suspected Diabetes Mellitus based on the Density-Based Spatial Clustering of Applications with Noize method was developed. An experimental study of the method was carried out on the PIMA Indians Diabetes open dataset. The model shows high accuracy, allowing it to be used in medical institutions for decision support in diagnosing.

Keywords 1

DBSCAN, Diabetes Mellitus, clustering, machine learning

1. Introduction

Diabetes Mellitus is a metabolic disorder of multiple etiologies characterized by chronic hyperglycemia with abnormal carbohydrate, fat, and protein metabolism resulting from impaired insulin secretion and action [1].

Diabetes mellitus is a global public health problem. As of 2022, 422 million people worldwide have diabetes, 6.028% of the planet's total population [2]. At the same time, there is an annual increase in the incidence. According to forecasts of the growth dynamics of the incidence of diabetes, by 2025, the number of patients will increase by two times. Furthermore, by 2030, diabetes will be the world's number 7 cause of death [3].

The main threat posed by Diabetes Mellitus is an early disability and high mortality from concomitant cardiovascular diseases. The main consequences of Diabetes Mellitus are as follows [4]:

- Diabetes is the leading cause of blindness.
- Diabetes is the leading cause of non-traumatic lower limb amputation.
- The risk of stroke, kidney failure, heart attacks, and other cardiovascular diseases increases four times.

The main risk factors for diabetes include [5]:

- Overweight.
- Unbalanced nutrition.
- Hereditary predisposition.
- Physical inactivity.
- Chronic gastritis.
- Cholecystitis.
- Impaired glucose tolerance.

2nd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2022), December 2-4, 2022, Łódź, Poland
EMAIL: krivtsovpro@gmail.com (S. Krivtsov); evgenii.menyailov@gmail.com (I. Meniailov); kirill.korobchinskiy@gmail.com (K. Korobchynskyi)
ORCID: 0000-0001- 5214-0927 (S. Krivtsov); 0000-0002-9440-8378 (I. Meniailov); 0000-0002-3676-6070 (K. Korobchynskyi)

- Age over 40 years.
- Constant stress.

In addition to high social importance, the problem of diabetes is also of economic importance. The fight against diabetes, depending on the country, is spent from 3 to 15% of annual health care budgets [6].

The primary means of combating diabetes are prevention and early diagnosis. Early diagnosis is especially effective in the context of the escalation of Russia's war in Ukraine. Access to facilities and adequate medical care is often difficult in areas with active hostilities. Therefore, physicians' automation tools and decision support systems are of particular relevance.

Over the past few years, the global COVID-19 pandemic has driven the digitization of healthcare worldwide. Information technologies are used to solve such problems as modeling the consequences of epidemics [7], medical diagnostics [8], forecasting infectious diseases [9], assessing resources and consequences of disease outbreaks [10], researching viruses [11], etc.

This study aims to develop a clustering model for patients with suspected Diabetes Mellitus based on the Density-Based Spatial Clustering of Applications with Noize method.

Research is part of a complex, intelligent information system for epidemiological diagnostics, the concept of which is discussed in [12].

2. Materials and Methods

Cluster analysis is the task of grouping a set of objects into subsets so that objects from one cluster are more similar than objects from other clusters according to some criterion. The clustering problem belongs to the class of unsupervised learning problems.

Let X be a set of objects, Y be a set of cluster identifiers. The distance function between objects $\rho(x, x')$ is given on the set X, given a finite training sample of objects $X^m = \{x_1, \ldots, x_m\} \subset X$. It is necessary to divide the sample into clusters, that is, to each object $x_i \in X_m$, assign a label $y_i \in Y$, so that the objects within each cluster are close concerning the metric ρ, and objects from different clusters differed significantly.

In medicine, cluster analysis is used for diagnostics [13], determining the severity of a disease in a patient [14], searching for factors influencing the development of disease [15], and identifying treatment regimens [16].

Within the framework of this study, a model for diagnosing patients with suspected Diabetes Mellitus based on the Density-Based Spatial Clustering of Applications with Noize (DBSCAN) method was developed.

The DBSCAN method consists of the fact that inside each cluster, a typical density of points (objects) is observed, which is noticeably higher than the density outside the cluster, as well as the density in areas with noise is lower than the density of any of the clusters [17]. At the same time, each cluster point's neighborhood of a given radius must contain at least a certain number of points. A threshold value sets this number of points.

Consider a set of points in some space requiring clustering. To perform DBSCAN clustering, points are divided into core points, point density reachable points, and outliers as follows:

- A point p is a core point if at least *minPts* points are at a distance not exceeding ε, the maximum neighborhood radius from p, to it. Such points are reachable from p.
- The point q is directly accessible from p if the point q is at a distance not more significant than ε from the point p, and p must be the main point.
- A point A_q is reachable from p if there is a path $p_1, p_2, \ldots, p_n$ and $p_n = q$, where every point $p_i + 1$ is reachable directly from p_i (all points on the path must be primary except for q).

In this case, all points that are not reachable from the main points are considered outliers. If p is a core point, it forms a cluster along with all points (core or non-core) that are reachable from that point. Each cluster contains at least one central point. Non-core points can also be part of a cluster.

Reachability is not a symmetric relationship because, by definition, no point can be reached from a non-primary point, regardless of distance. Two points, p, and q, are density related if there is a point o such that both p and q are reachable from o. Density connectivity is symmetrical.

Then the cluster satisfies two properties:

- All points in the cluster are pairwise connected in density.
- If a point is a density reachable from some point in the cluster, it also belongs to the cluster.

The DBSCAN algorithm has the following form:

- It is necessary to find points in the ε neighborhood of each point and select the main points with more than minPts neighbors.
- It is necessary to find the connected components of the core points on the graph of neighbors, ignoring all non-core points.
- You must assign each non-principal nearest cluster if the cluster is ε -neighbor. Otherwise, the point is noise.

Advantages of DBSCAN:

- The method does not require specification of the number of clusters in the data.
- The method has the concept of noise and is resistant to outliers.
- The method allows finding arbitrary-shaped clusters.
- Experts can set method parameters if the data is well interpretable.
- The method is insensitive to the order of points in the dataset.

Disadvantages of DBSCAN:

- Edge points that can be reached from more than one cluster may belong to any of those clusters, depending on the order in which the points are viewed. However, such situations rarely occur for most datasets, so they have practically no effect on the final result.
- The quality of DBSCAN depends on the distance measurement. Usually, the Euclidean metric is used for this.
- The method cannot cluster data well with a significant difference in density.

3. Results

The Python programming language was used for the cluster analysis model's software implementation. For the pilot study, we used data from an open dataset of patients with suspected Diabetes Mellitus PIMA Indian Diabetes, collected by the National Institute of Diabetes and Digestive and Kidney Diseases [18]. The dataset contains 768 records with nine attributes. Dataset parameters are presented in Table 1.

Table 1
Medical records description

Attribute	Scale type	Data type	Range
Pregnancies	Metric	Decimal	0…17
PGConcentration	Metric	Integer	0…199
DiastolicBP	Metric	Integer	0…122
TriFoldThick	Metric	Integer	0…99
SerumIns	Metric	Integer	0…846
BMI	Metric	Integer	0…67.1
DPFunction	Metric	Integer	0.08…2.42
Age	Metric	Decimal	21…81
Diabetes	Boolean	0/1	0,1
Attribute	Scale type	Data type	Range

The data distribution is shown in Figure 1.

The data set was divided into objects and objective functions. Figure 2 shows the results of the cluster membership calculation, the clustering quality assessment using the corrected Rand coefficient, and the clustering quality assessment using the normalized mutual information.

The model quality assessment using adjusted Rand index is 0,0028. The model quality assessment using normalized mutual information is 0,0021.

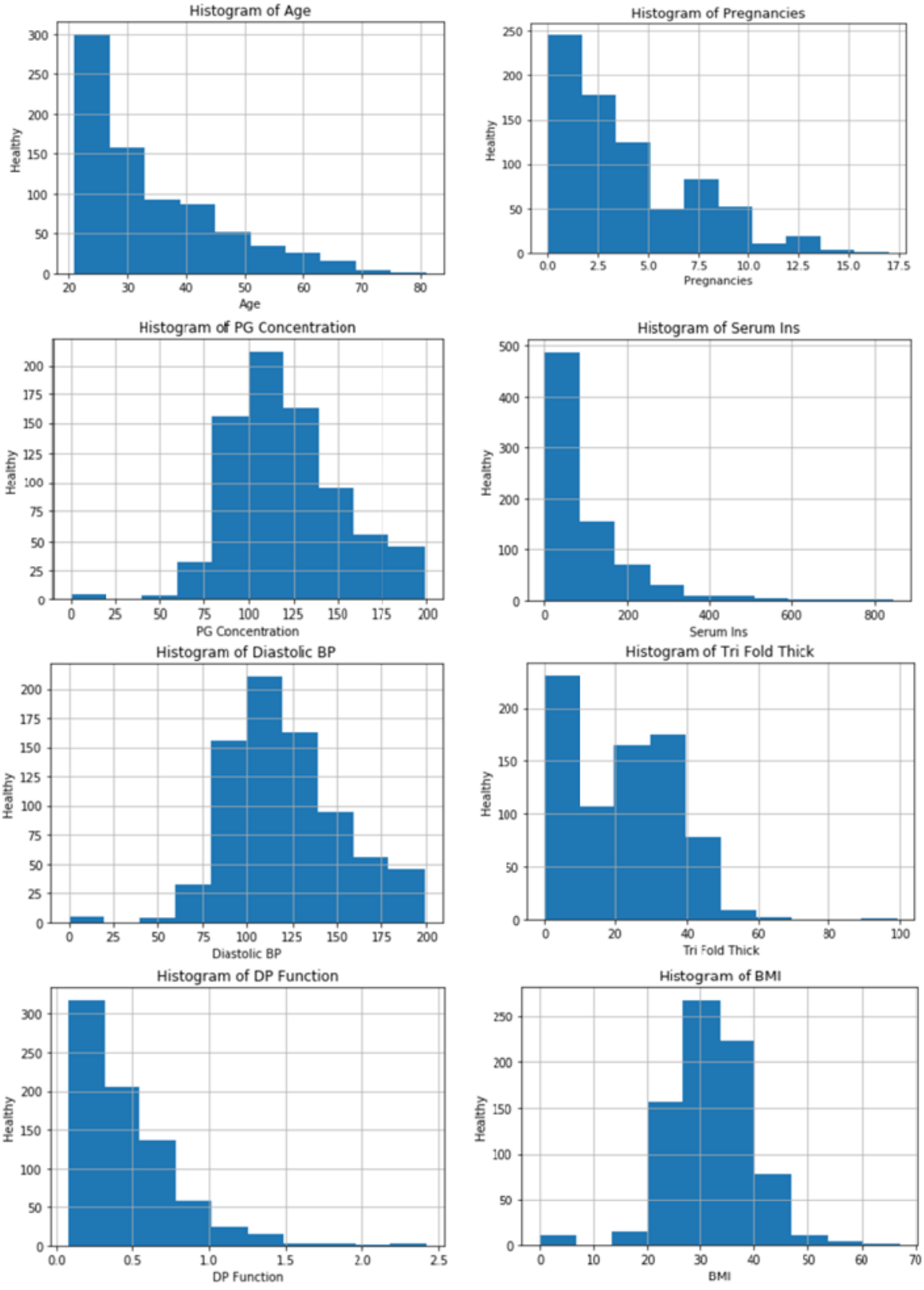

Figure 1: Data distribution

Figure 3 shows the visualization of points belonging to clusters and the display of noise points, i.e. points that took the value "-1".

```
Belonging to the clusters:
[ 0  0  0  0 -1  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1  0  0
  0  1  0  0  0  0  0  0  0  0 -1  0  1  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 -1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0 -1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0 -1 -1  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 -1 -1  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  0  0
  0 -1  0  0  0  0  0  0  0  0  0  0 -1  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
Algorithm quality assessment using adjusted Rand index:
0.00284780698002
Algorithm quality assessment using normalized mutual information:
0.0021023747426503992
```

Figure 2: Results of cluster analysis

The simulation results show sufficient performance to use the model in practical healthcare to support decision-making on the diagnosis of Diabetes Mellitus.

4. Conclusions

The task of clustering patients with suspected diseases is relevant for supporting decision-making by doctors when making diagnoses. Automation of diagnostics and the use of machine learning methods for this increase the accuracy of diagnostics and can be used by medical institutions to improve work efficiency. This task is especially relevant in Russia's war in Ukraine because residents of the temporarily occupied territories and territories where active hostilities occur do not have full access to medical care. Therefore, models for diagnosing various diseases can be used by patients in combination with remote consultation by a family doctor.

As part of this study, a clustering model for patients with suspected Diabetes Mellitus was developed based on the DBSCAN method. An open dataset of patients with suspected Diabetes Mellitus PIMA Indian Diabetes, collected by the National Institute of Diabetes and Digestive and Kidney Diseases, was used to verify the model. The developed model showed high performance. The model quality assessment using the adjusted Rand index is 0.0028. The model quality assessment using normalized mutual information is 0.0021. This suggests that the proposed model can effectively be used as a decision support tool for physicians diagnosing Diabetes Mellitus.

In the future, it is planned to apply the model to actual data on patients with suspected Diabetes Mellitus in the Kharkiv region.

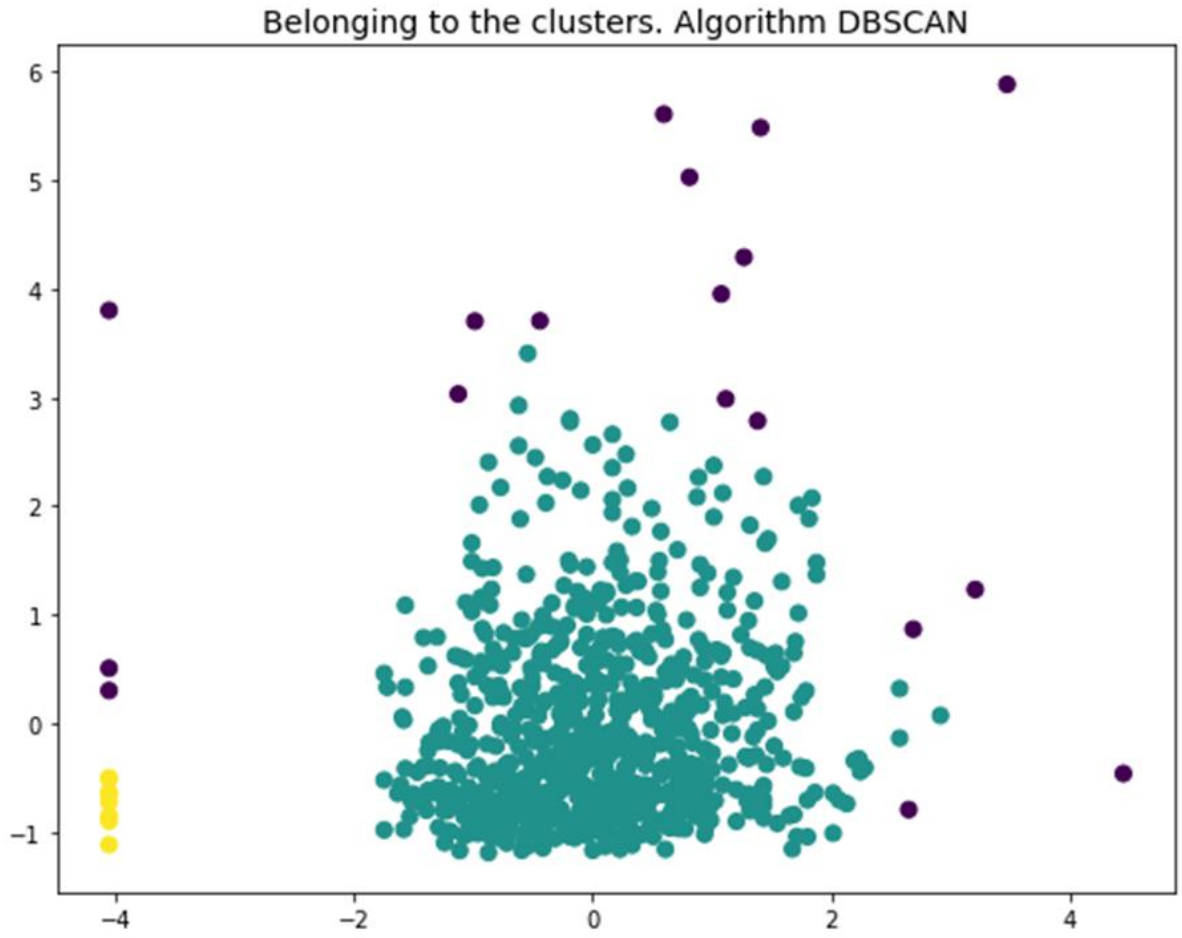

Figure 3: Visualizations of cluster analysis

5. Acknowledgement

The study was funded by the National Research Foundation of Ukraine in the frame-work 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] A.M. Schmidt, Highlighting diabetes mellitus: the epidemic continues. Arteriosclerosis, thrombosis, and vascular biology 38 (1) (2018): e1-e8. doi: 10.1161/ATVBAHA.117.310221

[2] Y. Zheng, S.H. Ley, F.B. Hu, Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nature Reviews. Endocrinology (14, iss. 2, 2018, pp. 88-98. doi: 10.1038/nrendo.2017.151

[3] P. Saeedi, et. al., Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9th edition. Diabetes research and clinical practice 157 (2019): 107843. doi: 10.1016/j.diabres.2019.107843

[4] J.B. Cole, J.C. Florez, Genetics of diabetes mellitus and diabetes complications. Nature Reviews Nephrology 16 (7) (2020): 377-390. doi: 10.1038/s41581-020-0278-5

[5] B. Fletcher, M. Gulanick, C. Lamendola, Risk factors for tupe 2 diabetes mellitus. The Journal of Cardiovascular Nursing 16 (2) (2002): 17-23. doi: 10.1097/00005082-200201000-00003
[6] C. Bommer, et. al., The global economic burden of diabetes in adults aged 20-79 years: a cost-of-illness study. The Lancet. Diabetes & Endocrinology 5 (6) (2017): 423-430. doi: 10.1016/S2213-8587(17)30097-9
[7] D. Chumachenko, V. Dobriak, M. Mazorchuck, I. Meniailov, K. Bazilevych, On agent-based approach to influenza and acute respiratory virus infection simulation. 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET 2018 – Proceedings (2018): 192-195. doi: 10.1109/TCSET.2018.8336184
[8] A.S. Nechyporenko, et. al., Implementation and analysis of uncertainty of measurement results for lower walls of maxillary and frontal sinuses, 2020 IEEE 40th International Conference on Electronics and Nanotechnology, ELNANO 2020 – Proceedings (2020): 460-463. doi: 10.1109/ELNANO50318.2020.9088916
[9] D. Chumachenko, I. Meniailov, K. Bazilevych, Y. Kuznetsova, T. Chumachenko, Development of an intelligent agent-based model of the epidemic process of syphilis. International Scientific and Technical Conference on Computer Sciences and Information Technologies 1 (2019): 42-45. doi: 10.1109/STC-CSIT.2019.8929749
[10] 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.
[11] D. Chumachenko, K. Chumachenko, S. Yakovlev, Intelligent simulation of network worm propagation using the code red as an example. Telecommunications and Radio Engineering 78 (5) (2019): 443-464. doi: 10.1615/TELECOMRADENG.V78.I5.60
[12] S. Yakovlev, et. al., The concept of developing a decision support system for the epidemic morbidity control. CEUR Workshop Proceedings 2753 (2020): 265-274.
[13] Wartelle, A., et. al., Clustering of a health dataset using diagnosis co-occurrences. Applied Sciences 11 (5) (2021): 2373. doi: 10.3390/app11052373
[14] M. Liao, Y. Li, F. Kianifard, S. Arcona, Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. BMC Nephrology 17 (2016): 25. doi: 10.1186/s12882-016-0238-2
[15] O. Skitsan, I. Meniailov, K. Bazilevych, H. Padalko, Evaluation of the informative features of cardiac studies diagnostic data using the Kullback method. CEUR Workshop Proceedings 2917 (2021): 186-195.
[16] S. Windgassen, R. Moss-Morris, K. Goldsmith, T. Chalder, The importance of cluster analysis for enchancing clinical practice: an example from irritable bowel syndrome. Journal of Mental Health 27 (2) (2018): 94-96. doi: 10.1080/09638237.2018.1437615
[17] M. Ester, H.P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (1996): 226-231.
[18] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, R.S. Johannes, Using the ADAP learning algorithm to forecast the onset of Diabetes Mellitus. Proceedings of the Symposium on Computer Applications and Medical Care (1988): 261-265.