UDC 621.396.4(045)
DOI: 10.18372/2306-1472.87.15719

Kateryna Sazonova[1]
Olena Nosovets[2]
Vitalii Babenko[3]
Olga Averianova[4]

GENERATION OF SYNTHETICAL MEDICAL DATA BY MDR-ANALYSIS

National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»,
37, Peremohy ave., Kyiv, 03056, Ukraine
E-mails: [1]kat2saz@gmail.com; [2]o.nosovets@gmail.com; [3]vbabenko2191@gmail.com;
[4]olgaaveryanova@ukr.net

Abstract

Purpose: *The purpose of this article is to outline an algorithm for generating synthetic medical data in order to augment small samples of data.* ***Methods****: To achieve the research goal, methods such as: correlation analysis (to identify significant variables and the relationships between them), MDR analysis (to build logical chains of relationships between medical data), and regression analysis (to model medical data variables to use this to generate synthetic data) were used.* ***Results****: A database of heart failure patients that is publicly available was used to test the developed algorithm for generating synthetic medical data in action; as a result, statistical relationships between data were found and used to build linear regression models.* ***Discussion****: The proposed algorithm allows, with a few simple, yet important actions, to perform the generation of medical data, which makes it possible to obtain large data sets that can be used to implement machine learning methods in any tasks related to medicine.*

Keywords: data generation, synthetic data, entropy, correlation, communication direction, MDR-analysis

1. Introduction

Nowadays, many activities from different fields, that humans used to do manually are getting automated; the same trend can be observed in medical field. New programs and algorithms, that, for example, help doctors make a diagnosis, are constantly being developed. But there is a problem, when new and improved algorithms require a big amount of data (*Big Data*) for learning (especially when it concerns neural networks), or for testing the results of analysis (decision-making algorithm).

In addition to the necessity of big amounts of data due to the confidentiality of medical records it can be quite difficult to obtain such data. Even if aforesaid data is retrieved, there is no guarantee that there are no missed or false values present, as the human error caused by the doctor who fills out these records, still exists.

This paper addresses the creation of algorithm for generation of synthetic medical data based on already existing data base with the preservation of inter-factor correlations and distribution of every value.

2. Analysis of the latest research and publications

Synthetic data is, in fact, a false data that has the same scheme and statistical properties as the real one. In other words, this data appears so real, that it is impossible for a person to even disprove it [1]. Synthetic data is of great value, for the reason that with a small amount of data it will be difficult to teach the same machine learning algorithms, or with an introduction of the new data there will be a considerable number of errors. For example, the project of *ImageNet* [2], aimed to solve the machine vision problem contains more than 14 million images, that are divided into 22 thousand categories. Due to the amount of material, the objects recognition algorithms are wrong only in 3.75% of cases, for comparison, a human being is wrong in more than 5% of cases [3].

Obtaining synthetic data is not a new problem. The *Gretel* company [4], for example, has created a software that is able to form an artificial array based on already existing database. First of all, the software analyzes the existing information, using data on trips on Uber electric scooters. The software divides trips into categories and marks them, after that the data is anonymized using differential privacy [5].

A similar project was implemented at the University of Illinois, where software engineers have written a Python library, that can generate synthetic data in structured formats (*CSV*, TSV) and partially structured (*JSON*, *Parquet*, *Avro*). [6]. In first case, generative-competitive networks were used [7], in second case – recurrent neural networks [8].

Copyright © 2021 National Aviation University
http://www.nau.edu.ua

However, despite all the advantages of synthetic data, it is often considered less accurate, even if it is generated based on the real data, and it can also lead to machine learning models that generate plausible results, that are impossible to replicate in reality. Nonetheless, real data can always be used for testing of intelligent system algorithms. There is also an argument that such technologies complicate the learning process of the model and increase development costs.

3. Research objective

The aim of this study is to create a synthetic data generation algorithm, based on the existing database, while maintaining inter-factor correlations and distribution of every value. To achieve this, the following tasks were assigned:

1. Evaluate the relationship between variables using correlation, while rejecting the factors, that have no relations at all.
2. Among the remaining factors, determine the direction of the impact using MDR-analysis.
3. Develop mathematical models for the generation of medical data.

4. Research results

A publicly available medical database, retrieved from Kaggle [9], was used in this study. The database contains 299 patients with heart failure, from whom 13 different indicators were taken (missing data in the database was omitted). These indicators are listed below:

- Patient's age (x_0).
- Anemia (x1) – binary variable, where 0 means absence of anemia, and 1 – its presence.
- Creatine kinase (x2).
- Diabetes (x3) – binary variable (0 – absence, 1 – presence).
- Ejection fraction (x4).
- Hypertension (x5) – binary variable (0 – absence, 1 – presence).
- Number of platelets (x6).
- Creatinine in plasma (x7).
- Sodium in plasma (x8).
- Gender (x9) – binary variable (0 – female, 1 – male).
- Smoking (x10) – binary variable (0 – the patient does not smoke, 1 – the patient smokes).
- Observation period (x11).
- Patient's death (x12).

At the first stage of the study a correlation analysis was performed. Since the data does not have a normal (Gaussian) distribution, Spearman's correlation was used, which can be applied to both quantitative (x_0, x_2, x_4, x_6, x_7, x_8, x_{11}, x_{12}), and binary (x_1, x_3, x_5, x_9, x_{10}). Fig. 1 shows a matrix of correlations between the input indicators.

Коефіцієнт корреляції(r):

1.000	0.088	-0.082	-0.101	0.060	0.093	-0.052	0.159	-0.046	0.065	0.019	-0.224	0.254
0.088	1.000	-0.191	-0.013	0.032	0.038	-0.044	0.052	0.042	-0.095	-0.107	-0.141	0.066
-0.082	-0.191	1.000	-0.010	-0.044	-0.071	0.024	-0.016	0.060	0.080	0.002	-0.009	0.063
-0.101	-0.013	-0.010	1.000	-0.005	-0.013	0.092	-0.047	-0.090	-0.158	-0.147	0.034	-0.002
0.060	0.032	-0.044	-0.005	1.000	0.024	0.072	-0.011	0.176	-0.148	-0.067	0.042	-0.269
0.093	0.038	-0.071	-0.013	0.024	1.000	0.050	-0.005	0.037	-0.105	-0.056	-0.196	0.079
-0.052	-0.044	0.024	0.092	0.072	0.050	1.000	-0.041	0.062	-0.125	0.028	0.011	-0.049
0.159	0.052	-0.016	-0.047	-0.011	-0.005	-0.041	1.000	-0.189	0.007	-0.027	-0.149	0.294
-0.046	0.042	0.060	-0.090	0.176	0.037	0.062	-0.189	1.000	-0.028	0.005	0.088	-0.195
0.065	-0.095	0.080	-0.158	-0.148	-0.105	-0.125	0.007	-0.028	1.000	0.446	-0.016	-0.004
0.019	-0.107	0.002	-0.147	-0.067	-0.056	0.028	-0.027	0.005	0.446	1.000	-0.023	-0.013
-0.224	-0.141	-0.009	0.034	0.042	-0.196	0.011	-0.149	0.088	-0.016	-0.023	1.000	-0.527
0.254	0.066	0.063	-0.002	-0.269	0.079	-0.049	0.294	-0.195	-0.004	-0.013	-0.527	1.000

Fig. 1. Correlation matrix of all indicators in data bases

In order to determine which variables are related, the significance of these correlations was also taken into account (Fig. 2):

Значимість коефіцієнту кореляції(p):

0.000	0.129	0.159	0.081	0.300	0.107	0.367	0.006	0.428	0.259	0.748	0.000	0.000
0.129	0.000	0.001	0.826	0.587	0.511	0.451	0.369	0.471	0.102	0.064	0.014	0.253
0.159	0.001	0.000	0.868	0.448	0.224	0.674	0.778	0.305	0.169	0.967	0.872	0.280
0.081	0.826	0.868	0.000	0.933	0.826	0.112	0.418	0.122	0.006	0.011	0.561	0.973
0.300	0.587	0.448	0.933	0.000	0.674	0.213	0.846	0.002	0.010	0.246	0.472	0.000
0.107	0.511	0.224	0.826	0.674	0.000	0.389	0.932	0.523	0.071	0.337	0.001	0.171
0.367	0.451	0.674	0.112	0.213	0.389	0.000	0.478	0.284	0.031	0.627	0.856	0.397
0.006	0.369	0.778	0.418	0.846	0.932	0.478	0.000	0.001	0.904	0.637	0.010	0.000
0.428	0.471	0.305	0.122	0.002	0.523	0.284	0.001	0.000	0.635	0.934	0.131	0.001
0.259	0.102	0.169	0.006	0.010	0.071	0.031	0.904	0.635	0.000	0.000	0.788	0.941
0.748	0.064	0.967	0.011	0.246	0.337	0.627	0.637	0.934	0.000	0.000	0.694	0.828
0.000	0.014	0.872	0.561	0.472	0.001	0.856	0.010	0.131	0.788	0.694	0.000	0.000
0.000	0.253	0.280	0.973	0.000	0.171	0.397	0.000	0.001	0.941	0.828	0.000	0.000

Fig. 2. Significance of correlation coefficients

In statistics, relationships between variable are considered significant if the significance of correlation is less than 0.05. Fig. 3 shows the following variables:

Між якими змінними є звязок (p < 0.05):
x0: x7(p=0.006)
x1: x2(p=0.001)
x2: x1(p=0.001)
x3: x9(p=0.006) x10(p=0.011)
x4: x8(p=0.002) x9(p=0.010)
x5:
x6: x9(p=0.031)
x7: x0(p=0.006) x8(p=0.001)
x8: x4(p=0.002) x7(p=0.001)
x9: x3(p=0.006) x4(p=0.010) x6(p=0.031) x10(p=0.000)
x10: x3(p=0.011) x9(p=0.000)
x11:
x12: x0(p=0.000) x4(p=0.000) x7(p=0.000) x8(p=0.001)

Fig. 3. Statistical relationships between variables

Hence, the obtained results allowed us to conclude that hypertension (x_5) and observation period (x_{11}) do not correlate with other factors, so their synthetic values will be generated based on data of their distribution and standard deviation in a range from minimum to maximum values.

It was decided to use analysis via Multifactor dimensionality reduction (MDR) for the remaining variables. Using this approach, the direction and severity of the variables by using entropy indicators. For implementation «Multifactor Dimensionality Reduction 3.0.2» software application was used, that outputs the result of analysis in the form of circle graph or dendrogram.

At first, using this application, a circle graph of 10 chosen variable was built (Fig. 4).

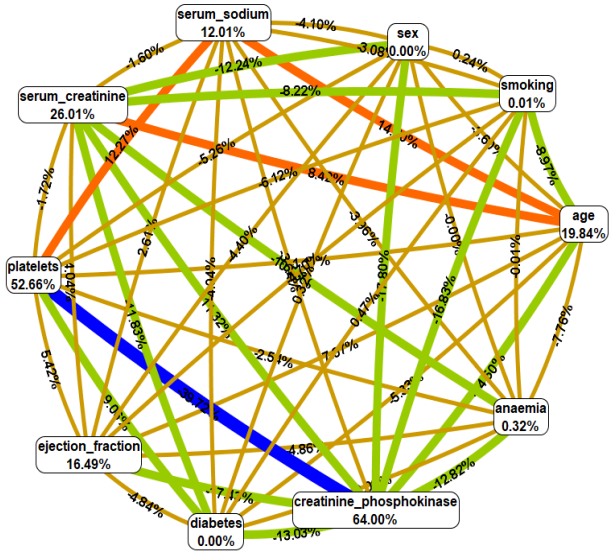

Fig. 4. Circle graph of all variables

When analyzing the graph show above, it is necessary to describe the steps of its construction in detail (the used software allows to perform this). Fig. 5. shows a graph in the shape of an «arrow» from age to sodium in plasma (direct orientation).

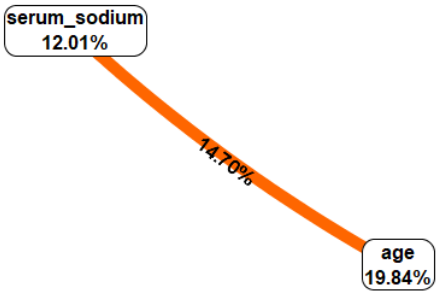

Fig. 5. The start of a general graph

The next step is to add an arrow from sodium in the plasma to the number of platelets (Fig. 6).

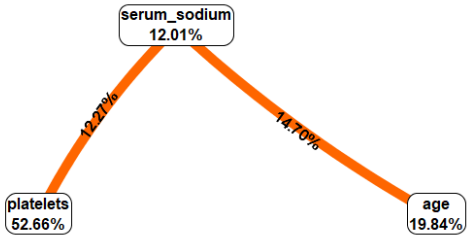

Fig. 6. The second step in constructing a general graph

After that, an arrow from age to creatinine in plasma was added. (Fig. 7).

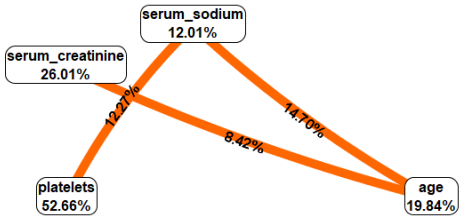

Fig. 7. The third step in constructing a general graph

In the fourth stage, an arrow from age to ejection fraction has appeared. (Fig. 8).

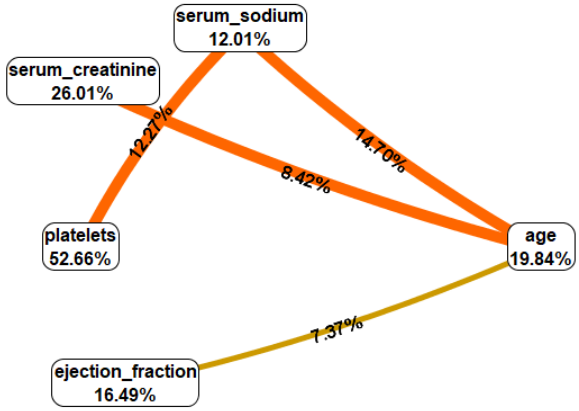

Fig. 8. The fourth step in constructing a general graph

In contrast to the previous arrows shown above, the new arrow in Fig. 8. has a different, brown color, which shows a pairwise relationship between variables.

In the fifth step, a link between smoking and diabetes has appeared (Fig. 9). There are two things to note from the figure: first of all, the link between smoking and diabetes appeared on its own, meaning that they are not related to other variables that were present in the previous steps; and secondly, as a general graph is obtained, the relationships can also be formed between variables that have already been presented.

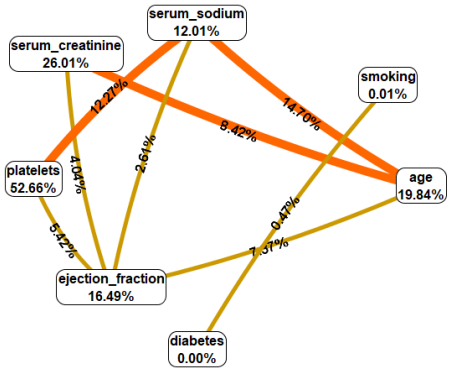

Fig. 9. The fifth step in constructing a general graph

In the sixth step, a new variable was added: gender due to the connection with diabetes. (Fig. 10).

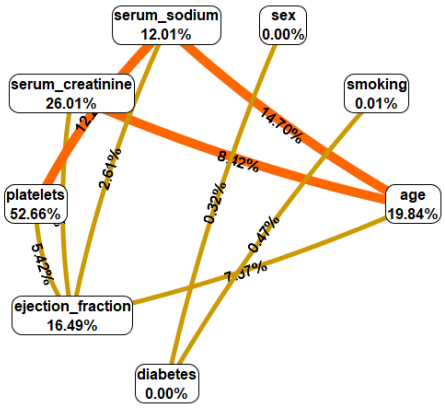

Fig. 10. The sixth step in constructing a general graph

And in the last step, two new variables appear: creatinine kinase from creatinine in plasma and anemia from creatinine kinase (Fig. 11).

After performing these operations, it is possible to display communication chains:

- **Age (x_0)** → sodium in plasma (x_8) → number of platelets (x_6).
- **Age (x_0)** → creatinine in plasma (x_7) → creatine kinase (x_2) → anemia (x_1).
- **Smoking (x_{10})** → diabetes (x_3) → sex (x_9) → ejection fraction (x_4).

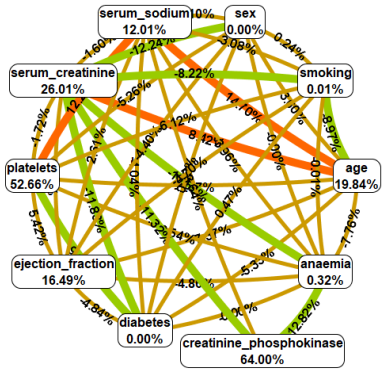

Fig. 11. The sixth step in constructing a general graph

These chains are confirmed by correlation analysis (Fig. 1-3), so the results can be considered variable. This makes it possible to perform the main task of the study, namely – the generation of synthetic data.

Regression analysis is the most suitable for the implementation of this function. It allows you to build a mathematical model of the variable using another variable or set of variables. The following are the models that were built after MDR analysis:

$$\begin{cases} x_8 = -0.017x_0 + 137.667 \\ x_6 = -407.483x_0 + 1333.558x_8 + 105730.652 \\ x_7 = 0.014x_0 + 0.553 \\ x_2 = -6.608x_0 - 3.674x_7 + 992.007 \\ x_1 = 0.003x_0 + 0.019x_7 - 0.00009x_2 + 0.285 \\ x_3 = -0.155x_{10} + 0.468 \\ x_9 = 0.442x_{10} - 0.091x_3 + 0.545 \\ x_4 = -0.107x_{10} - 0.702x_3 - 3.74x_9 + 40.839 \end{cases} \quad (1)$$

The models were obtained using «IBM SPSS Statistics 23» statistical software. The determination coefficient (R^2) of these models ranged from 0.5 to 0.6, which are acceptable values, but far from ideal. This is not surprising, since the models were built using linear regression, which has a disadvantage of undertraining.

5. Conclusions

As a result of the study, a synthetic data generation algorithm was created. The following approaches were used for implementation: correlation analysis, MDR analysis and linear regression. Data of patients with heart failure, which is available in the public domain, was used to test the developed approach.

Correlation analysis has helped to identify those medical symptoms that have nothing to do with others. With the help of MDR analysis, logical chains of connection between medical data were obtained, which gives an understanding of which variables need to be used to model and generate data. After obtaining regression models with a sufficient determination coefficient on average, this thesis was confirmed. Therefore, the synthetic medical data generation algorithm was performed successfully.

References

[1] Patki N. The Synthetic Data Vault / N. Patki, R. Wedge, K. Veeramachaneni // IEEE International Conference on Data Science and Advanced Analytics (DSAA). – 2016. – Available at: https://bit.ly/3uU1IWU.

[2] Towards Fairer Datasets: Filtering and Balancing teh Distribution of the People Subtree in the ImageNet Hierarchy / [K. Yang, K. Qinami, L. Fei-Fei та ін.] // Conference on Fairness, Accountabiility and Transparency. – 2020. – Available at: https://doi.org/10.1145/3351095.3372833.

[3] Dodge S. A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions / S. Dodge, L. Karam. – 2017. – Available at: https://arxiv.org/pdf/1705.02498.pdf.

[4] Watson A. Using generative, differentially-private models to build privacy-enhancing, synthetic datasets from real data. / Alexander Watson. – 2020. – Available at: https://medium.com/gretel-ai/using-generative-differentially-private-models-to-build-privacy-enhancing-synthetic-datasets-c0633856184.

[5] Privacy: Theory meets Practice on the Map / [A. Machanavajjhala, D. Kifer, J. Abowd et.al.]. – 2018. – Available at: https://bit.ly/33RpdnC.

[6] Walters A. Why You Don't Necessarily Need Data for Data Science / Austin Walters // Capital One Tech. – 2018. – Available at: https://bit.ly/2SZm4Qz.

[7] Pouget-Abadie J. Generative Adversarial Networks / J. Pouget-Abadie, M. Mirza, B. Xu. – 2014. – Available at: https://arxiv.org/abs/1406.2661.

[8] Fernández S. An application of recurrent neural networks to discriminative keyword spotting / S. Fernández, A. Graves, J. Schmidhuber // ICANN'07: Proceedings of the 17th international conference on Artificial neural networks. – 2007. – Available at: https://dl.acm.org/doi/10.5555/1778066.1778092.

[9] Heart Failure Prediction Available at: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

К.М. Сазонова[1], О.К. Носовець[2], В.О. Бабенко[3], О.А. Аверьянова[4]

Генерація синтетичних медичних даних за допомогою MDR-аналізу

Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", пр. Перемоги, 37, Київ, Україна, 03056
E-mails: [1]kat2saz@gmail.com; [2]o.nosovets@gmail.com; [3]vbabenko2191@gmail.com; [4]olgaaveryanova@ukr.net

Мета: Метою даної статті є викладення алгоритму генерації синтетичних медичних даних для того, щоб доповнити маленькі вибірки даних. **Методи**: Для досягнення мети дослідження були використані такі методи, як: кореляційний аналіз (для виявлення значимих змінних та взаємозв'язків між ними), MDR-аналіз (для побудови логічних ланцюгів зв'язку між медичними даними) та регресійний аналіз (для моделювання змінних медичних даних, щоб використати це для генерації синтетичних даних). **Результати**: Була використана база даних пацієнтів з серцевою недостатністю, яка доступна у відкритому доступі, щоб перевірити розроблений алгоритм генерації синтетичних медичних даних у дій; в результаті були знайдені статистичні взаємозв'язки між даними, які використовувались для побудови моделей лінійної регресії. **Обговорення**: Запропонований алгоритм дозволяє за допомогою декількох простих, але в той час важливих дій виконати генерацію медичних даних, що дає можливість отримати великі масиви даних, які можна використовувати для реалізації методів машинного навчання у будь-яких задачах пов'язаних з медициною.

Ключові слова: генерація даних, синтетичні дані, ентропія, кореляція, направленість зв'язку, MDR-аналіз

Е.М. Сазонова[1], Е.К. Носовец[2], В.О. Бабенко[3], О.А. Аверьянова[4]

Генерация синтетических медицинских данных с помощью MDR-анализа

Национальный технический университет Украины "Киевский политехнический институт имени Игоря Сикорского", пр. Победы, 37, Киев, Украина, 03056
E-mails: [1]kat2saz@gmail.com; [2]o.nosovets@gmail.com; [3]vbabenko2191@gmail.com; [4]olgaaveryanova@ukr.net

Цель: Целью данной статьи является изложение алгоритма генерации синтетических медицинских данных для того, чтобы дополнить маленькие выборки данных. **Методы**: для достижения цели исследования были использованы такие методы, как: корреляционный анализ (для выявления

значимых переменных и взаимосвязей между ними), MDR-анализ (для построения логических цепей связи между медицинскими данными) и регрессионный анализ (для моделирования переменных медицинских данных, чтобы использовать это для генерации синтетических данных). **Результаты**: Была использована база данных пациентов с сердечной недостаточностью, которая доступна в открытом доступе, чтобы проверить разработанный алгоритм генерации синтетических медицинских данных в действии; в результате были найдены статистические взаимосвязи между данными, которые использовались для построения моделей линейной регрессии. **Обсуждение**: Предложенный алгоритм позволяет с помощью нескольких простых, но в то же время важных действий выполнить генерацию медицинских данных, что дает возможность получить большие массивы данных, которые можно использовать для реализации методов машинного обучения в любых задачах, связанных с медициной.

Ключевые слова: генерация данных, синтетические данные, энтропия, корреляция, направленность связи, MDR-анализ

Kateryna Sazonova. Student.
Department of Biomedical Cybernetics, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».
Research area: information technologies in medicine, computer science, data science, deep learning.
Publications: 0.
E-mail: kat2saz@gmail.com

Olena Nosovets. PhD of Technical Science.
Department of Biomedical Cybernetics, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».
Education: National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute» (2015).
Research area: information technologies in medicine, computer science, data science, deep learning.
Publications: 100.
E-mail: o.nosovets@gmail.com

Vitalii Babenko. Master of Science.
Department of Biomedical Cybernetics, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».
Education: National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute» (2021).
Research area: information technologies in medicine, computer science, data science, deep learning.
Publications: 20.
E-mail: vbabenko2191@gmail.com

Olga Averianova. Senior Lecturer
Department of Biomedical Cybernetics, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».
Research area: information technologies in medicine, computer science, data science, deep learning, system analysis, information system design, IT management
Publications: 5.
E-mail: olgaaveryanova@ukr.net