Intelligent Evaluation of the Informative Features of Cardiac Studies Diagnostic Data using Shannon Method

Kseniia Bazilevych, Serhii Krivtsov, Mykola Butkevych

National Aerospace University "Kharkiv Aviation Institute", Chkalow str., 17, Kharkiv, Ukraine

Abstract

The paper is devoted to the important issue of separating more informative data from less informative data for further analysis and use. This determines the relevance of the study. As a result of the study, the methods for assessing the informativeness of signs based on medical data were analyzed. On the basis of Shannon's method, a model for assessing information content has been built and a software package has been implemented. For the experimental study, data from 303 patients and 13 signs were used. The informative value was calculated for various groups of cardiac data. We found that the following signs are the most informative: tala, type of chest pain, colored vessels, angina pectoris, age. The Shannon method is also compared with other methods for assessing the informativeness of features.

Keywords 1

Features informativeness, Shannon method, diagnostics, heart disease, cardiac studies.

1. Introduction

The COVID-19 pandemic caused by the SARS-CoV-2 coronavirus has become a real challenge not only for health systems, but also for the economy around the world [1]. Announced March 11, 2020. It began with the discovery at the end of December 2019 in the city of Wuhan in the Hubei province of central China. There are still no specific antiviral drugs for treatment or prevention against the disease [2]. In severe cases, funds are used to maintain the functions of vital organs. People of all ages are susceptible to infection. Severe forms of the disease are more likely to develop in older people and in people with certain medical conditions, including asthma, diabetes, and heart disease [3].

The coronavirus pandemic has clearly demonstrated that we must act together and give our fight against this crisis the necessary momentum to achieve the Sustainable Development Goals [4]. The COVID-19 pandemic has accelerated the digitalization of all spheres of social activity [5]: education [6], commerce [7], public administration [8], personnel management [9-10], logistics [11], etc. Particular attention should be paid to the many approaches to digitalizing medicine [12]. In this area, information technologies have been developed for insurance [13], decision-making [14], medical diagnostics [15-16], epidemic control systems [17] and morbidity simulation [18]. In this article, we will focus on the diagnostic problem that has arisen sharply in connection with the pandemic. There are not enough people in hospitals [19], and COVID-19 is especially difficult with concomitant diseases [20].

Diseases of the cardiovascular system continue to be the leading cause of death in many countries of the world. Every year 17 million people die from diseases of the cardiovascular system in the world. According to the Centers for Disease Control and Prevention, life expectancy would be 10 years longer in the absence of such a high prevalence of cardiovascular diseases, covering all countries and continents [21]. They lead to long-term disability of the adult population and require colossal economic costs.

International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2021), September 20-21, 2021, Kharkiv, Ukraine
EMAIL: ksenia.bazilevich@gmail.com (K. Bazilevych); krivtsovpro@gmail.com (S. Krivtsov); nikolai.butkevych@gmail.com (M. Butkevych).
ORCID: 0000-0001-5332-9545 (K. Bazilevych); 0000-0001- 5214-0927 (S. Krivtsov); 0000-0001-8189-631x (M. Butkevych).

High-risk groups include people who have had heart attacks and strokes. It is important for patients with repeated heart attacks and high blood pressure to be under medical supervision. High cholesterol in the patient's blood contributes to narrowing of the blood vessels and requires long-term medication. Excess weight, high blood sugar and a sedentary lifestyle have an extremely negative effect on the state of the cardiovascular system, and smoking is one of the most common risk factors. In the development of atherothrombosis, heredity and age play a significant role, and it is noted that in recent years cardiovascular diseases have become significantly "younger" [22]. The growth and occurrence of cardiovascular diseases in young people is associated not only with an incorrect lifestyle, but also with increased neuropsychological stress. The Internet, TV, phones, radio give us such a stream of information that our ancestor cannot cope with in a week. Negative emotions and stress cause an increased amount of adrenaline in the blood, hence fear, anxiety, anxiety, panic, and increased heart rate [23]. The state of the cardiovascular system quickly reacts to changes in mood, and the constant imbalance between physical and neuropsychological stress leads to pathological changes and the development of cardiovascular diseases.

In Ukraine, cardiovascular diseases are the main cause of death among the population [24]. According to this indicator, the country remains one of the world leaders.

According to the ranking data, based on the number of deaths of the population in Ukraine [25], common causes are:

1. Cardiovascular diseases (64.3%)
2. Neoplasm (14.1%)
3. Diseases of the digestive system (4.3%)
4. Neurological disorders (3.1%)
5. Self-harm and interpersonal violence (2.7%)

Nationally, mortality from cardiovascular diseases over the past 29 years has increased by almost 8%: to 449,376 in 2019 and accounts for 64.3% of the total number of deaths, while in 1990 there were 350,605 deaths from cardiovascular diseases, amounted to 56.5% respectively [26].

Thus, **aim** of the paper is development of intelligent information system of heart diseases diagnostics. To achieve the aim, we are going to develop model based on Shannon method to evaluate the informative features of cardiac studies.

2. Informative features evaluation

The informativeness of signs is a relative concept. One and the same system of signs can be considered informative for solving some problems and uninformative for others. For example, in medicine, some signs may be significant for the differential diagnosis of diabetes diseases [27], and others for the diagnosis of heart diseases.

In the tasks of medical diagnostics, patients act as objects. Signs characterize the results of examinations, symptoms of diseases and the methods of treatment used. The specifics of modern requirements for data processing in order to discover knowledge are as follows: data are large, heterogeneous (binary, ordinal, quantitative), the results must be specific and understandable. Examples of binary signs are gender, headache, weakness, nausea, etc. An ordinal sign is the severity of the condition (mild, moderate, severe, life-threatening). Quantitative signs are age, pulse, blood pressure, hemoglobin content in the blood, respiratory rate, drug dose, etc. The symptomatic description of the patient is, in fact, a formalized medical history. Having accumulated a sufficient number of precedents, it is possible to solve various problems: to classify the type of disease (differential diagnosis), to determine the most appropriate method of treatment, to predict the duration and outcome of the disease, to assess the risk of complications, and to find syndromes - the most characteristic set of symptoms for a given disease. When studying objects characterized by a large number of factors, it is often important to determine which of these factors most affect the properties of objects of interest to us. In particular, the determination of the informativeness of factors is one of the important stages in the analysis of the object under study.

The block diagram of the method for informative features evaluation is shown in Figure 1.

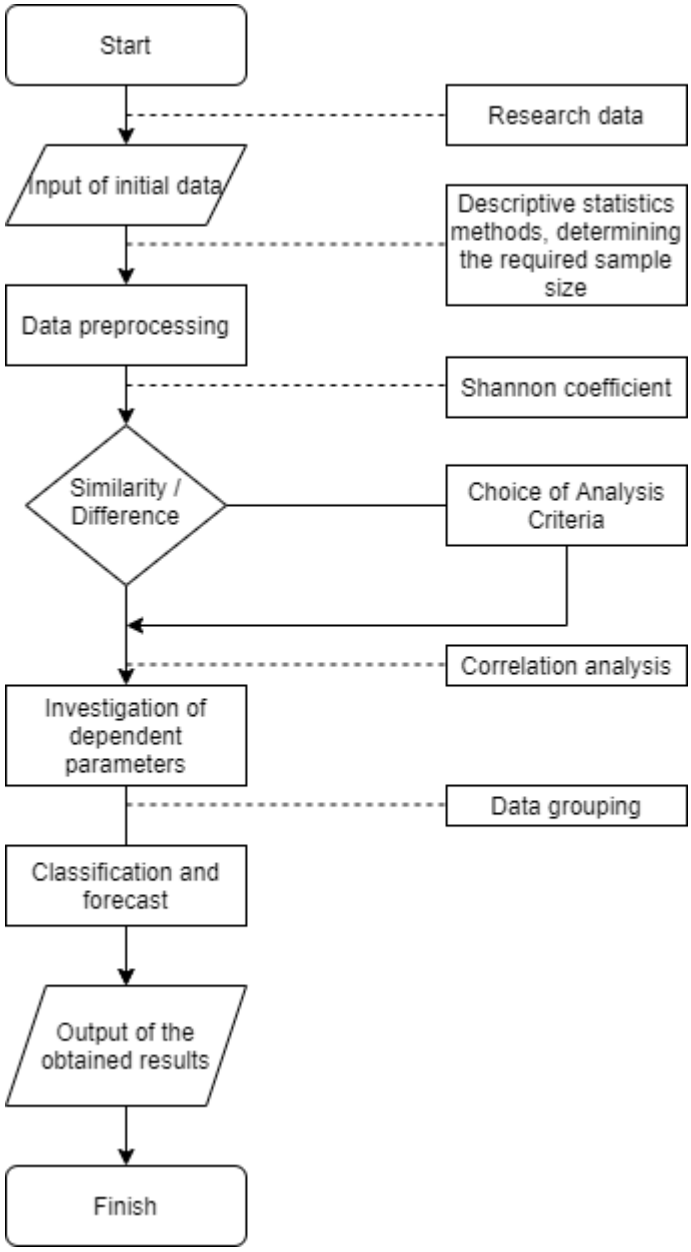

Figure 1: The block diagram of the method for informative features evaluation.

The symptomatic descriptions of patients are formalized case histories [28]. Having accumulated the required number of use cases in the database, you can solve various problems:

- classification of types of diseases;
- differential diagnostics;
- determination of effective methods of treatment;
- prediction of the outcome of the disease;
- prediction of the duration of the disease;
- risk assessment of complications;
- identification of syndromes that are most typical for this disease.

Speaking about the tasks of medicine, the following features can be distinguished:

- Quantitative features are features measured in a certain numerical scale.
- Qualitative features are features used to express terms and concepts that do not have numerical values, which are measured in ordinal scales.
- Nominal features are features measured in a naming scale (e.g. blood group). When analyzing such features, each mark of the nominal scale is converted to a boolean scale.

It is also possible to single out various methods for assessing the informativeness of signs: energy and information.

The energy approach is based on the fact that the information content is assessed by the value of the attribute. The signs are sorted by values, and those whose values are greater are considered the most informative. For example, according to the amplitude-time analyzes of the electrocardiogram, the amplitude of the R waves is considered the most informative signs among the amplitudes. But, such approaches to assessing the information content may turn out to be poorly suitable for object recognition. If some features are large in absolute values, but are almost the same for objects of different classes, then by the values of these features it is difficult to assign objects to some classes. Conversely, if the features are relatively small in magnitude, but differ greatly for objects of different classes, then objects can be easily classified by their values.

The method for determining the informativeness is selected depending on the purpose of the study, the number of studied classes and medical data (coding methods, the number of gradations, the sample size, etc.)

Therefore, information methods are more suitable for classification in medical diagnostics, according to which information of signs is considered as reliable differences between classes of images in spaces of signs. If, when classifying objects, they need to be attributed to one of two classes, then the differences in the probability distributions of features constructed from samples of two compared classes can act as such a reliable difference.

3. Shannon method application

Shannon's method suggests evaluating information content as a weighted average amount of information per different grades of a feature [29]. In information theory, information is understood as the value of the eliminated entropy.

$$I(x_i) = 1 + \sum_{i=1}^{G}(P_i \sum_{k=1}^{K} P_{i,k} \log_K P_{i,k}), \quad (1)$$

where G is the number of gradations of the feature;

K is quantity of classes;

P_i is the probability of the i-th gradation of the feature

$$P_i = \frac{\sum_{k=1}^{K} m_{i,k}}{N}, \quad (2)$$

where $m_{i,k}$ is the frequency of occurrence of the i-th grade in the K-th class,

N is the total number of observations;

$P_{i,k}$ is probability of occurrence of the i-th gradation of a feature in the K-th class.

$$P_{i,k} = \frac{m_{i,k}}{\sum_{k=1}^{K} m_{i,k}}. \quad (3)$$

Shannon's method gives an estimate of the informativeness as a normalized value, which varies from 0 to 1. Therefore, the informativeness of a feature determined by Shannon's method can be said in absolute terms: closer to 1 for high; closer to 0 for low.

The block diagram of the Shannon method for informative features evaluation is shown in Figure 2.

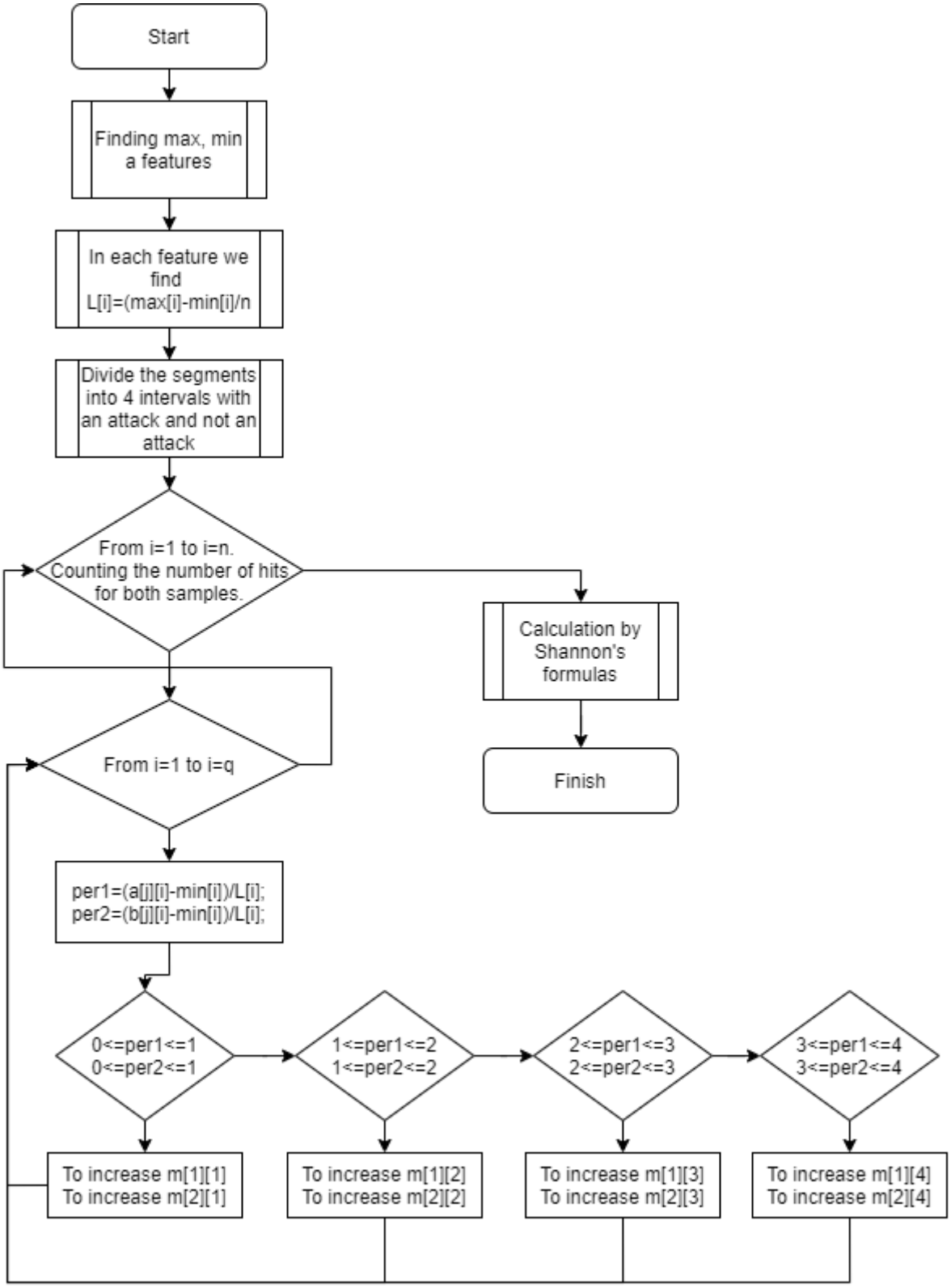

Figure 2: The block diagram of the Shannon method for informative features evaluation.

4. Results

The input data is a dataset of information on the diagnostic data of patients based on cardiac studies, their age, gender, type of chest pain, cholesterol level, etc., a complete list of parameters in Table 1.

Table 1
Parameters list

Name of parameter	Description	Data type
Name	ID	Count
Age	Years	Count
Sex	Sex	String
Chest pain type	Pain type	String
Blood pressure	Scores	Count
Cholesterol	Scores	Count
Fasting blood sugar < 120	+/-	0/1
Resting ECG	Normal/Hyper	String
Maximum heart rate	Scores	Count
Angina	+/-	0/1
Peak	Scores	Float
Slope	Flat/Up/Down	1/2/3
Colored Vessels	0/1/2	0/1/2
Thal	Normal/Rev/Fix	String

Before software implementation of an information system, it is necessary to design it. For this, the IDEF0 and DFD methodologies were used.

The model is based on the concepts of an external entity, process, data storage (storage) and data flow.

An external entity is a material object or individual acting as sources or receivers of information, for example, customers, personnel, suppliers, bank customers, and the like.

Process is converting input data streams to output in accordance with a certain algorithm. Each process in the system has its own number and is associated with the executor who performs this transformation. As in the case of functional diagrams, physical transformation can be carried out by computers, manually or by special devices. At the upper levels of the hierarchy, when the processes have not yet been defined, instead of the concept of “process”, the concepts of “system” and “subsystem” are used, which respectively denote the system as a whole or its functionally complete part.

A data warehouse is an abstract device for storing information. The type of device and methods of placement, removal and storage for such a device are not detailed. Physically, it can be a database, a file, a table in RAM, a card file on paper, and the like.

Data flow is the process of transferring some information from a source to a receiver. Physically, the process of transferring information can occur through cables under the control of a program or software system, or manually with the participation of devices or people outside the designed system.

The functional model of the system is presented in Figure 3.

Decomposition of the system is presented in Figure 4.

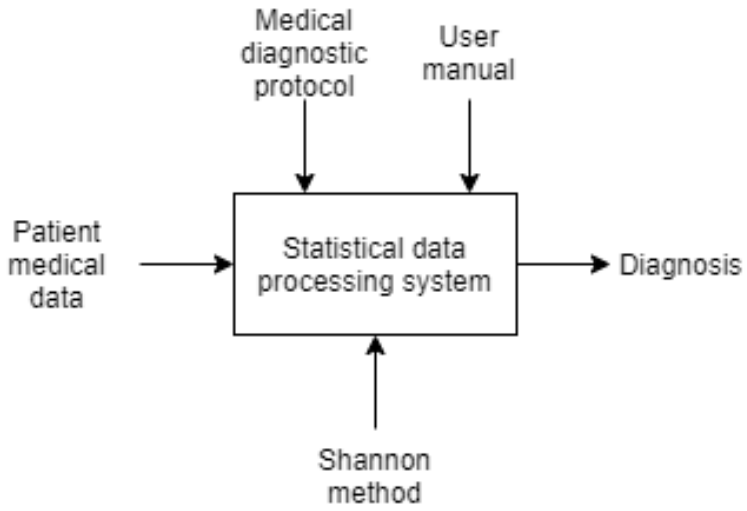

Figure 3: The functional model of the information system.

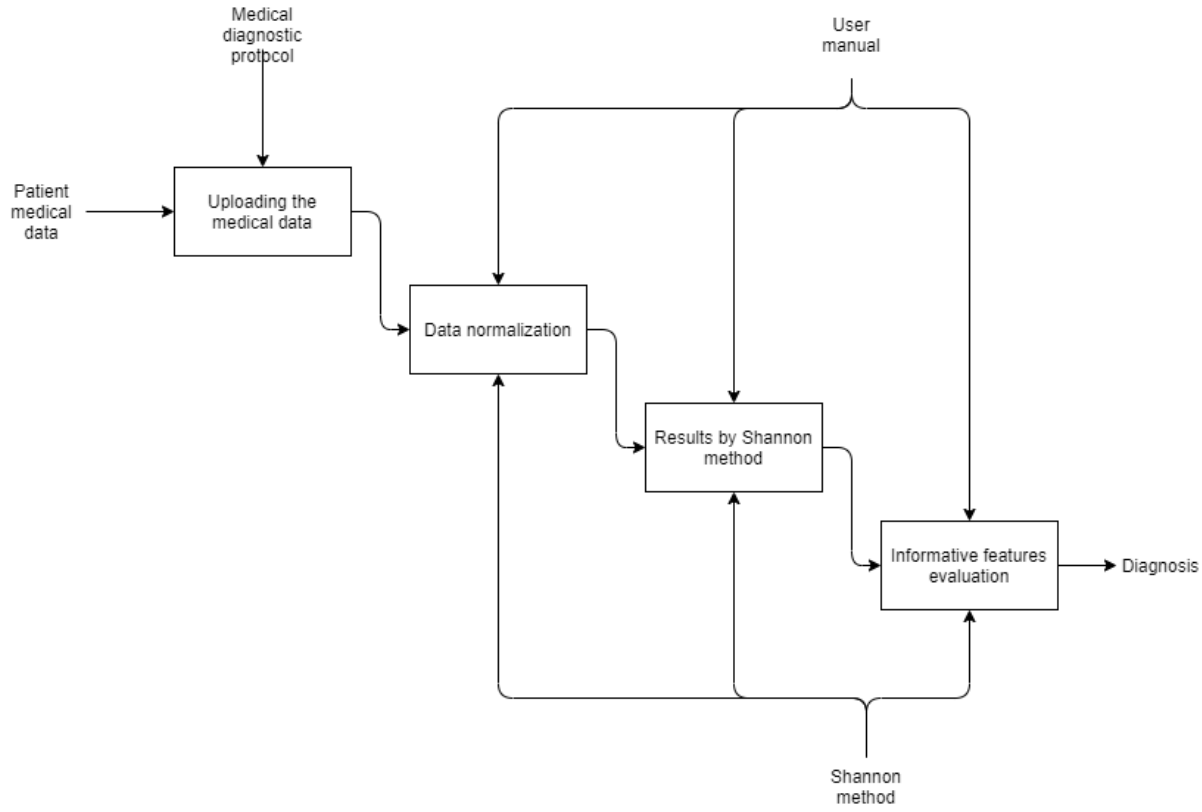

Figure 4: Decomposition of the system.

In total, for example, data from 303 patients and 13 features was taken (their age, gender, type of chest pain, cholesterol level, ECG, blood pressure, maximum pressure, blood sugar level, type and presence of tonsillitis, colored vessels, etc.)

For software implementation, the C# programming language was used in the Microsoft Visual Studio environment. To start the software package, you need to upload the data presented in the *.csv file (Figure 5).

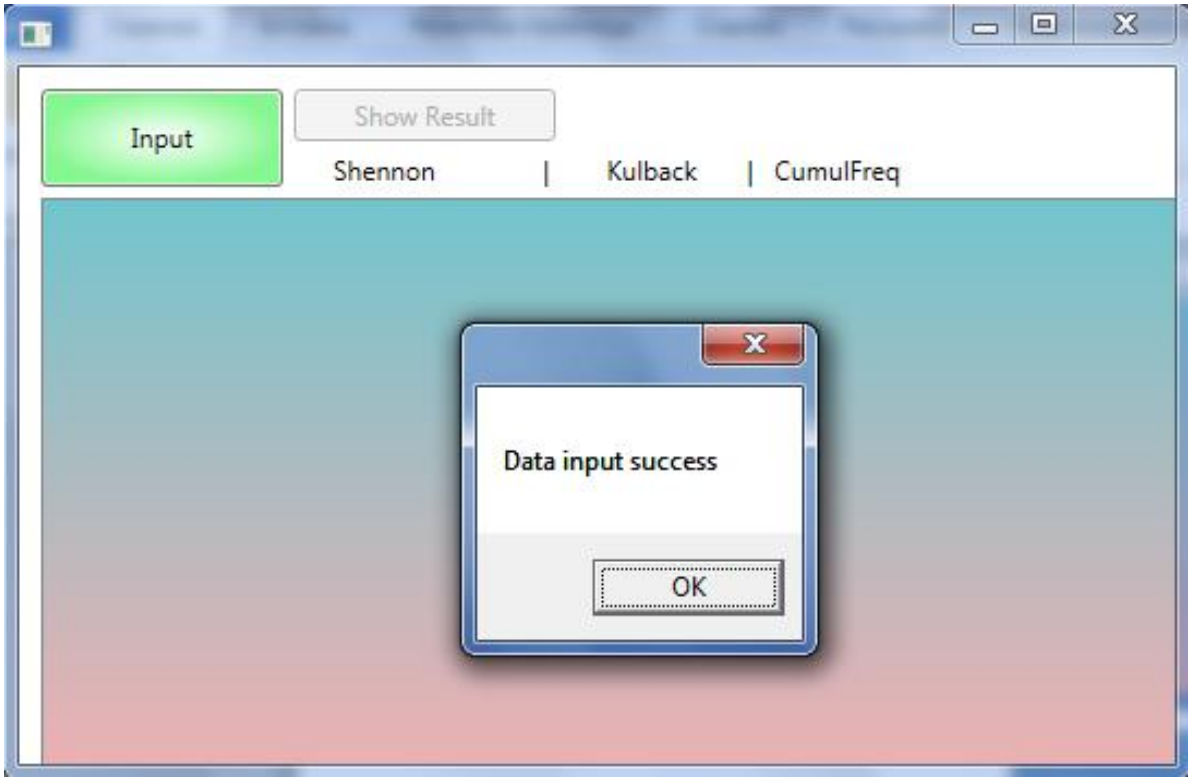

Figure 5: Setting the initial data for program use.

The data is divided into two classes A – "Healthy" and B – "Sick".

The results of the calculation by the Shannon method for assessing the informativeness of the attribute m = "Patient's age" is shown in Figure 6.

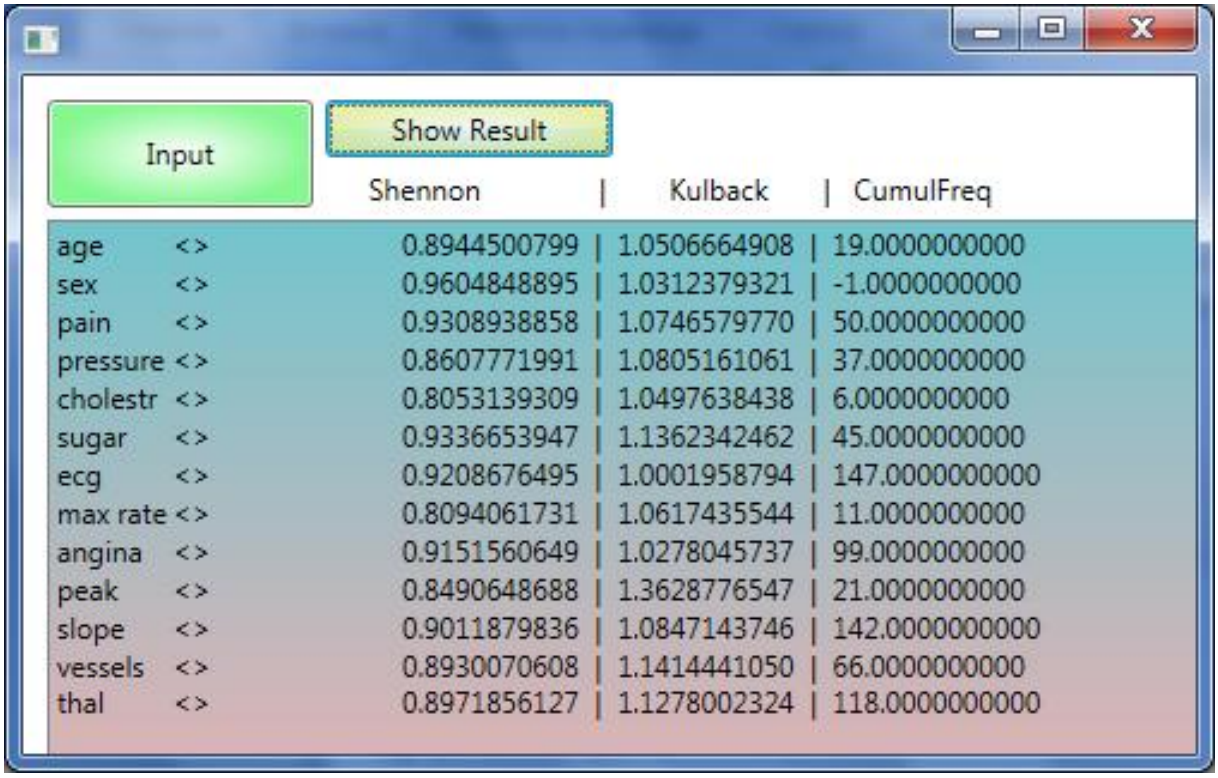

Figure 6: Result of the information system calculations.

The numerical results are shown in Table 2.

Table 2
Informative features by Shannon method

Name of parameter	Informativeness
Age	0.8944500799
Sex	0.9604848895
Chest pain type	0.9308938858
Blood pressure	0.8607771991
Cholesterol	0.8053139309
Fasting blood sugar < 120	0.9336653947
Resting ECG	0.9208676495
Maximum heart rate	0.8094061731
Angina	0.9151560649
Peak	0.8490648688
Slope	0.9011879836
Colored Vessels	0.8930070608
Thal	0.8971856127

Shannon method gives an estimate of the informativeness of the investigated feature in the form of a value, takes values from 0 to 1. In this case, it is believed that the closer I (x) to 1, the higher the informativeness of the feature, on the contrary, the closer I (x) to 0, the lower the informative value of x.

5. Conclusions

As a result of the study, methods for assessing the informativeness of signs for medical data were analyzed. The Shannon method was chosen as the most appropriate method for medical data. On the basis of the Shannon method, a model for assessing the information content was built and a software package was implemented. For the experimental study, data from 303 patients and 13 features were used. The information content was calculated for various groups of cardiac data. We got that the following signs are the most informative: thal, chest pain type, colored vessels, angina, age. The Shannon method is used to determine the informativeness of a feature that is involved in the recognition of two classes of objects. Also, comparisons of the Shannon method with other methods (Kullback and Cumulative frequency method) for assessing the informativeness of features are made.

6. 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” [30].

7. References

[1] E.R. Fox, Budgeting in the time of COVID-19, American Journal of Health-System Pharmacy: official journal of the American Society of Health-System Pharmacists 77 (15) (2020) 1174-1175. doi: 10.1093/ajhp/zxaa185.
[2] M. Gavriatopoulou M, et. al., Emerging treatment strategies for COVID-19 infection, Clinical and Experimental Medicine 21 (2) (2021) 167-179. doi: 10.1007/s10238-020-00671-y.
[3] H. Ejaz, et. al., COVID-19 and comorbidities: Deleterious impact on infected patients, Journal of Infection and Public Health 13 (12) (2020) 1833-1839. doi: 10.1016/j.jiph.2020.07.014.
[4] K. Heggen, T.J. Sandset, E. Engebretsen, COVID-19 and sustainable development goals, Bulletin of World Health Organization 98 (10) (2020) 646. doi: 10.2471/BLT.20.263533.

[5] A. Abd-Alrazaq, et. al., Artificial Intelligence in the Fight Against COVID-19: Scoping Review, Journal of Medical Internet Research 22 (12) (2020) e20756. doi: 10.2196/20756.
[6] D. Chumachenko, V. Balitskii, T. Chumachenko, V. Makarova, M. Railian, Intelligent expert system of knowledge examination of medical staff regarding infections associated with the provision of medical care, CEUR Workshop Proceedings 2386 (2019) 321-330.
[7] P. Piletskiy, et. al., Development and Analysis of Intelligent Recommendation System Using Machine Learning Approach, Advances in Intelligent Systems and Computing 1113 (2020) 186-197. doi: 10.1007/978-3-030-37618-5_17.
[8] 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.
[9] N. Dotsenko, et. al. Modeling of the processes of stakeholder involvement in command management in a multi-project environment, Proceedings of 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies 1 (2018) 29-33. doi: 10.1109/STC-CSIT.2018.8526613
[10] N. Dotsenko, et. al. Project-oriented management of adaptive teams' formation resources in multi-project environment, CEUR Workshop Proceedings 2353 (2019) 911-920.
[11] M. Bielecki, et. al., Air travel and COVID-19 prevention in the pandemic and peri-pandemic period: A narrative review, Travel Medicine and Infectious Disease 39 (2021) 101915. doi: 10.1016/j.tmaid.2020.101915.
[12] S.C. Mathews, et. al., Digital health: a path to validation, NPJ Digital Medicine 2 (2019) 38. doi: 10.1038/s41746-019-0111-3.
[13] K. Bazilevych, et al. Stochastic modelling of cash flow for personal insurance fund using the cloud data storage, International Journal of Computing 17 (3) (2018) 153-162. doi: 10.47839/ijc.17.3.1035
[14] D. Chumachenko, et. al. On Intelligent Decision Making in Multiagent Systems in Conditions of Uncertainty, Proceedings of 2019 11th International Scientific and Practical Conference on Electronics and Information Technologies (2019) 150-154. doi: 10.1109/ELIT.2019.8892307
[15] M. Mazorchuck, et. al. Web-Application Development for Tasks of Prediction in Medical Domain, 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT) (2018) 5-8. doi: 10.1109/STC-CSIT.2018.8526684
[16] O. Skitsan, et. al., Evaluation of the informative features of cardiac studies diagnostic data using the Kullback method, CEUR Workshop Proceedings 2917 (2021) 186-195.
[17] D. Chumachenko, et. al. On-Line Data Processing, Simulation and Forecasting of the Coronavirus Disease (COVID-19) Propagation in Ukraine Based on Machine Learning Approach, Communications in Computer and Information Science 1158 (2020) 372-382. doi: 10.1007/978-3-030-61656-4_25
[18] Yu. Polyvianna, et. al. Computer Aided System of Time Series Analysis Methods for Forecasting the Epidemics Outbreaks, 2019 15th International Conference on the Experience of Designing and Application of CAD Systems (2019) pp. 7.1-7.4. doi: 10.1109/CADSM.2019.8779344
[19] J. Wosik, et. al., Telehealth transformation: COVID-19 and the rise of virtual care, Journal of American Medical Informatics Association 27 (6) (2020) 957-962. doi: 10.1093/jamia/ocaa067
[20] M.S. Gold, et. al., COVID-19 and comorbidities: a systematic review and meta-analysis, Postgraduate Medicine 132 (8) (2020) 749-755. doi: 10.1080/00325481.2020.1786964.
[21] C.Y. Cheng, C.Y. Hsu, T.C. Wang, Y.C. Jeng, W.H. Yang, The risk of cardiac mortality in patients with status epilepticus: A 10-year study using data from the Centers for Disease Control and Prevention (CDC), Epilepsy and Behaviour 117 (2021) 107901. doi: 10.1016/j.yebeh.2021.107901
[22] R.D. Bagnall, E.S. Singer, J. Tfelt-Hansen, Sudden Cardiac Death in the Young, Heart, Lung and Circulation 29 (4) (2020) 498-504. doi: 10.1016/j.hlc.2019.11.007.
[23] A. Tajbakhsh, et. al., COVID-19 and cardiac injury: clinical manifestations, biomarkers, mechanisms, diagnosis, treatment, and follow up, Expert Review of Anti-Infective Therapy 19 (3) (2021) 345-357. doi: 10.1080/14787210.2020.1822737
[24] O. Makar, G. Siabrenko, Influence of physical activity on cardiovascular system and prevention of cardiovascular diseases (review), Georgian Medical News 285 (2018) 69-74.

[25] R.O. Moiseienko, N.G. Gojda, O.O. Dudina, N.M. Bodnaruk, Development of perinatal medicine in Ukraine in the context of international approaches, Wiadomosci Lekarskie 74 (3) 2 (2021) 761-766.
[26] J. Luck, J.W. Peabody, L.M. DeMaria, C.S. Alvarado, R. Menon, Patient and provider perspectives on quality and health system effectiveness in a transition economy: evidence from Ukraine, Social Science and Medicine 114 (2014) 57-65. doi: 10.1016/j.socscimed.2014.05.034.
[27] T. Dudkina, et. al., Classification and prediction of diabetes disease using decision tree method, CEUR Workshop Proceedings 2824 (2021) 163–172
[28] D. Chumachenko, O. Sokolov, S. Yakovlev, Fuzzy recurrent mappings in multiagent simulation of population dynamics, International Journal of Computing 19 (2) (2020) 290-297. doi: 10.47839/ijc.19.2.1773.
[29] Qing Tian, T. Arbel, J. J. Clark, Shannon information based adaptive sampling for action recognition, 2016 23rd International Conference on Pattern Recognition (ICPR) (2016) 967-972, doi: 10.1109/ICPR.2016.7899761.
[30] S. Yakovlev, et. al., The concept of developing a decision support system for the epidemic morbidity control, CEUR Workshop Proceedings 2753 (2020) 265–274.