Journal of Theoretical and Applied Information Technology
15th April 2019. Vol.97. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# THE APPLICATION OF DATA MINING METHODS FOR THE PROCESS OF DIAGNOSING DISEASES

**SAULE BELGINOVA**[1], **INDIRA UVALIYEVA**[2], **SAMIR RUSTAMOV**[3]

[1-2]D.Serikbayev East Kazakhstan State Technical University, Ust-Kamenogorsk, Kazakhstan

[3]ADA University, Institute of Control Systems, Baku, Azerbaijan

E-mail: [1]sbelginova@gmail.com, [2]iuvalieva@mail.ru, [3]samir.rustamov@gmail.com

## ABSTRACT

Today, the medical field has a large amount of medical information that requires proper processing and further use in the diagnosis and treatment of various diseases. The development of computer technology provides tremendous opportunities for collecting, processing, managing and researching medical information to better understand the complex biological processes of life and help solve the problem of diagnosis and treatment in medical institutions.

Accurate diagnosis and proper treatment provided to patients is one of the main tasks of medical care. Therefore, data mining techniques, which are part of knowledge discovery in databases, are becoming popular tools for medical researchers. The use of these tools makes it possible to identify and use patterns and relationships between numerous variables and to predict specific disease and treatment outcomes.

This paper describes the application of data mining methods for the process of diagnosing diseases. To improve the effectiveness of the methods of data mining for the process of diagnosing diseases, a procedure for assessing the informativeness of heterogeneous diagnostic indicators on the basis of a theoretical information approach has been proposed. The results of diagnosing are described on the basis of revealing the relationship between the supply of magnesium and the risk of somatic diseases with the use of intelligent data analysis.

**Keywords**: *Diagnosis, Information Signs, Intellectual Analysis Of Medical Data, Diagnostically Valuable Signs, Clustering.*

## 1. INTRODUCTION

Improvement of the organization of medical care through the introduction of information technology is one of the priority areas of health development. More and more attention is paid to the use of large data mining technologies to improve the quality of medical care [1]. According to the statistics every third diagnosis that doctors put is incorrect [2].

From the incorrect preliminary diagnosis, not only patients who are treated not from the disease but also medical institutions that incur considerable financial costs suffer, as the Mandatory Medical Insurance Fund finances only the treatment corresponding to the final diagnosis.

The purpose of this work was to check the principle possibility of analytical processing of available data by machine learning methods and determine the accuracy of forecasting, in which the model of machine learning can have practical value.

The field of biomedical Informatics has attracted increasing popularity and attention and has been growing rapidly over the past two decades. Thanks to new advances in medical technology, a huge amount of medical research data is generated every day. Digitizing important medical information such as laboratory reports, patient records, scientific articles, and anatomical images has also led to a large amount of medical data.

Biomedical researchers and practitioners are now faced with the problem of "information overflow". Currently, the data acquisition rate is much higher than the data interpretation rate. These data must be effectively organized and analyzed in order to be useful in medical institutions [3].

New computing and information technologies are needed to manage these numerous medical data repositories and to identify useful knowledge through various models. In particular, in recent years, methods of knowledge management and data mining have been introduced into the field of medical diagnosis.

Intelligent analysis of medical data has great opportunity for identifying and studying hidden patterns in large sets of medical data. These models have been used for clinical diagnosis with great success in recent decades. Data mining technology provides a user-friendly approach to new and hidden patterns in data. However, available baseline health data are widespread, inhomogeneous nature and voluminous.

When working with a large amount of medical data, there is a need to use a system of integrated data mining, which could aggregate and analyze various types of information from different objects of a medical organization. Due to the large volumes of information received, such a system must correctly use existing technologies for working with big data. The analyzed data should be collected in a certain organized form. With the help of information technology, the collected data is integrated into the medical information system. Treatment records for millions of patients can be stored and computerized, and medical facilities will be able to analyze this data. Thus, the processes of diagnosis and therapeutic decision-making are accelerated.

Medical decisions are often made on the basis of doctors' intuition and experience, rather than on the basis of knowledge of the rich data hidden in the database. All this leads to undesirable distortions and errors in the diagnosis of diseases and high medical costs, which affects the quality of patient care. Many researchers in their works prove that the introduction of decision support systems based on computerized records of medical data can reduce the number of medical errors, increase the safety of patients, which is the main task of health care.

The healthcare delivery system is changing in many ways. Technological advances are providing opportunities to optimize patient care. Clinical Information Systems have the potential to address many problems encountered in healthcare, namely, managing large amounts of patient and research data, reduce healthcare costs/errors, increase legibility, and boost the quality of healthcare [4].

Data Mining methods can be used to build decision-making models for procedures such as prediction, diagnosis and treatment planning, which, after testing and evaluation, can be embedded in Clinical Information Systems [5].

There are a number of studies that detail the work of the authors on the use of data mining methods in medicine [6-10]. The article [9] does not only review the literature in this direction, but also considers the problems and the upcoming tasks of applying the data mining methods in medicine. The article [11] discusses the development of data mining methods based on a review of the literature and the classification of articles over the past decades, since this period is important because during this time there has been a proliferation of data mining methods used in the medical industry. A review of previous attempts to develop computer clinical diagnostic tools is made, and problems arising in the development, implementation, evaluation and maintenance of decision support systems for clinical diagnosis are discussed [12].

The authors of [13] suggest that – a) Improvement in the accuracy of Medical Decision Support (MDS) application may be possible by modeling of vague and temporal data, research on inference algorithms, integration of patient information from diverse sources and improvement in gene profiling algorithms; b) MDS research would be facilitated by public release of de-identified medical datasets, and development of open source data-mining tool kits; c) Comparative evaluations of different modeling techniques are required to understand characteristics of the techniques, which can guide developers in choice of technique for a particular medical decision problem; d) Evaluations of MDS applications in clinical setting are necessary to foster physicians' utilization of these decision aids.

When constructing computer medical diagnostic systems, it is important to form an informatively complete space of diagnostic signs, since in medicine for the diagnosis is used very heterogeneous information.

At the same time, the studied subsystems of the organism are, in fact, complex hierarchical systems; therefore, an adequate description of such systems is possible by building a hierarchical system of diagnostic features through the structural identification of diagnostic features.

When structurally identifying diagnostic signs, it is also necessary to select informative diagnostic signs from the proposed set, since the inclusion of non-informative indicators in a computer diagnostic model degrades the quality of a computer diagnosis.

Based on this, today the actual problems are the development of methods to reduce the space of diagnostic signs and the construction of a hierarchical structure of diagnostic signs.

To solve the problem in medicine, the well-known mathematical-statistical methods for reducing the dimension of the space of diagnostic signs are used, namely: cluster, discriminant, factor, dispersion, regression analyzes, multidimensional scaling, principal component method, the method of contrast groups.

With their help, the nature and structure of the relationship of the diagnostic indicators under study are also established.

## 2. REVIEW OF APPLICATION OF DATA MINING METHODS IN MEDICINE

To date, a significant amount of information on public health and habitat has been accumulated in the repositories and database of various health and social development systems. Unfortunately, we have to admit that their processing, which purports to obtain useful knowledge, is insufficient. Huge databases remain unclaimed, while the correct, justified management decisions are essential; analysis and prognosis of public health under the influence of habitat are state functions. Data analysis guidelines, especially adapted to the needs of practitioners, are extremely small, the programs of medical universities do not include training database specialists, so that health solutions are applied at best on the basis of scientific research and expert assessments, often contradictory.

As a way to solve this problem, we offer data mining technologies - Data Mining and Knowledge Discovery in Databases (DM & KDD) [14, 15]. Data Mining is the process of detecting early unknown, non-trivial, practically useful, accessible interpretation of knowledge (regularities) in raw data (row data) necessary for making decisions in various spheres of human activity (G.Pyatetskiy-Shapiro). In the technologies of DM & KDD, various mathematical methods and algorithms are used: classification, clustering, regression, time series prediction, association, consistency [16].

DM & KDD's intelligent tools include neural networks, decision trees, inductive inferences, analogical reasoning methods, fuzzy logical conclusions, genetic algorithms, associative and sequence determination algorithms, selective-action analysis, logical regression, evolution programming, data visualization.

Useful knowledge obtained during Data mining can be represented in the form of regularities, rules, forecasts, relationships between data elements, etc. [17, 18]

In our opinion, Data mining is exactly what should now be actively mastered, adapted to the tasks and opportunities of departments interested in assessing the health of the population and implemented in their routine work. Moreover, they should be used in scientific studies of public health and habitat. We made a number of attempts in this direction.

We used some of the methods of Data mining in our scientific studies [19-23], as well as in methodological developments adapted to the needs of social hygienic monitoring and other specialists performing research in the field of medicine [24-27].

To implement the technology of Data mining, it is necessary to carry out preliminary processing of the initial data. Preliminary processing of diagnostic information in computer diagnostic systems is based on the formalization of the initial characteristics and allocation of space of diagnostic valuable traits. In [28], mathematical methods for evaluating the informativeness of diagnostic features are considered. Traditional methods based on dispersion, regression, correlation analysis are used [29], the information-theoretic approach based on the calculation of conditional probabilities and the amount of information [18, 29], multidimensional statistical analysis, which, as shown in [18, 30-32], is effective only for the complex application of different methods and for a relatively large number of parameters, methods of synthesizing the hierarchical structure of diagnostic features [33].

## 3. RELATED WORK

Data mining in medicine is different from data mining in other areas. First, the data are heterogeneous because they come from different sources, and the methods used must take this heterogeneity into account. Secondly, special ethical, legal and social restrictions apply to private medical information, which is why there may be problems with access to the patient database.

Despite this, data mining methods, which are part of the discovery of knowledge in databases, have become popular tools for researchers who seek to identify relationships between numerous variables and be able to predict the outcome of the disease. It has already been proven that the use of data mining provides advantages in many areas of medicine.

Different types of decision support systems have been developed to help physicians make better or faster decisions. However, most of them focus on the diagnosis [34-46], and on the prediction of the diseases outcome [47-52].

There are also integrated systems that link the treatment decision with the diagnosis decision [53]. Thus, the health care system is changing in many ways. Technological advances open up opportunities for optimizing decision-making on disease diagnosis and patient care. Clinical information systems have the potential to solve many of the challenges that arise in the field of health, namely, managing large volumes of patient data and conducting scientific research based on these data.

The paper [4] also describes in detail all the advantages and problems of existing clinical information systems. The main advantages of clinical information systems according to the authors are: direct access to instant updates of the patient's medical records, as well as remote access to patient records; patient-oriented decision-making based on the best clinical data; fast data processing ensures rapid decision-making on the diagnosis of the disease; the ability to develop effective and intuitive software for data processing and bioinformatics tools; more chances to conduct potential research based on real data; improving the quality of data and analysis of patient data by combining them with the doctor's own knowledge. Along with the undeniable advantages of clinical information systems, there are also various problems that need to be solved, such as: the need to develop and integrate subsystems; high initial investment with low expected return; security and confidentiality; the need to create communication between numerous doctors of different specialties;

The research paper [54] illustrates the need for multi-level interdisciplinary integrated intelligent systems in medicine. Such a system will not only reduce costs, speed up the provision of medical services, provide accurate statistical data, allow scientific interdisciplinary and multi-level planning, provide more accurate medical documentation or assist in making critical health care decisions.

Today, a huge amount of data is produced by various medical systems. But they have not yet taken full advantage of the opportunities that these data provide for transformation. The application of methods based on big data from medical institutions can be very useful in the field of health, allowing identifying and extract relevant information and reduce the time spent by biomedical and medical professionals and researchers who are trying to find meaningful patterns and new streams of knowledge [6].

In the article [55] each work is studied on the basis of six medical tasks: screening, diagnosis, treatment, prognosis, monitoring and management. Each task discusses five data mining approaches: classification, regression, clustering, Association, and hybrid.

The authors of the review [56] discuss the main peculiar properties of the data mining in medicine and consider two specific sides of especial interest: methods that can work with temporal data and efforts made to convert the results of molecular medicine into useful data mining models.

Most of the work under consideration focuses on examining data sets for a specific disease such as breast cancer [39, 57, 58, 59] pneumonia [34, 60], blood pressure [48], tuberculosis [61], cardiovascular [36, 42, 47, 62, 63, 64], cancer [50, 51, 65, 66], diabetes [67], cerebrovascular disease [68], autism [43, 69], skin diseases [44], headaches [40], epilepsy [46].

To reduce the dimension of the space of diagnostic features in medicine, well-known mathematical-statistical methods are widely used, namely: cluster, discriminant, factor, dispersion, regression analyzes, multidimensional scaling, principal component method, method of contrasting groups. With their help, the nature and structure of the relationship of the diagnostic indicators under study are also established.

Clustering and classification are two separate phases of data mining that provide a reliable and proven structure from a large set of facts.

The dominant direction of modern medical research is the prediction of diseases and the categorization of diseases. The clustering algorithm divides the data set into several groups in such a way that the similarity within groups is greater than between groups. The main purpose of the classification is to accurately predict the target class for each unknown case in the data.

At the moment, the use of clustering tools for medical data analysis is poorly understood. Cluster analysis in relation to medical data was used to identify cardiac diseases in [70], where cluster analysis is performed on the basis of a search for similar forms of Fourier spectra obtained by simulation of the heart.

The authors of [71] divide asthma into 3 different phenotypes by uncontrolled cluster analysis. In [72] cluster analysis is performed according to the ACT assessment (a test of five questions on the state of asthma control). Each variable is standardized by subtracting the mean and dividing by the standard deviation.

The work [73] focuses on the analysis of clusters of patient records obtained by methods of uncontrolled clustering, and comparison of the effectiveness of classification algorithms on clinical data. Feature selection is a controlled method that attempts to select a subset of predictor attributes based on the information obtained.

Clustering methods are increasingly used in the analysis of high-performance biological data sets. [74] discusses how clustering methods and their possible successors will be used to accelerate the rate of biological discovery in the future.

## 4. THE ANALYSIS OF EXISTING METHODS IN MEDICAL DIAGNOSING

At present, active work is in progress to improve the accuracy of diagnosing diseases. The principal difficulties in the differential diagnosis arise not because of lack of information required, but due to the lack of appropriate methods of structuring it. The advent of personal computers gave rise to their application in medicine for the development of automated methods of recognition of pathologies. It was created by numerous diagnostic systems based on mathematical methods of information processing, carried out extensive work on the creation of software for automated diagnostic systems.

However, computer diagnostics has not widely expanded, which would have to obtain because of the lack of technical support of the health sector and the lack of medical personnel training system. Training of doctors should provide their ability to work on the most modern equipment, making maximum use of the information. The automated diagnostic system must play the role of an active information base with which you can process large amounts of data. Let us consider the diagnostic methods currently used in clinical practice.

1. Traditionally, nosological diagnostic method is used. Subjective factor affects to a large degree on the recognition of the disease process. Preliminary diagnosis may not be accurate and to confirm it is assigned a number of instrumental and laboratory examinations. The method does not work well in an irregular situation (a rare disease or atypical form of the disease course). This method is effective only in the practice of those physicians who are dealing with a fairly limited range of diseases.

2. Syndrome approach to the diagnosis. This is a method of using the minimum amount of medical research and the minimum number of symptoms that are of crucial significance. In fact, this is the same nosological method optimized for its very high diagnostic feasibility with all its shortcomings.

3. Statistical method of information processing. The basic idea is the symptoms of frequency of occurrence analysis helps identify features characteristic of a disease. Symptoms received the “weight” that is, the significance in points. The diagnostic procedure was reduced to the summation of “weights” identified patient’s symptoms for each disease. That diagnosis, which is gaining an increasing number of points, was considered true. Disadvantages: after comparing the significance of clinical signs diagnosed substantially remained undefined.

4. “Anti-syndrome” method. This is a variation of the statistical method. The bottom line: we select a combination of symptoms that have never been encountered in any of the differentiable diseases. If patients show such “anti-syndrome”, it served as the basis for the denial of an appropriate diagnosis. The drawback: it is impossible to give an interpretation of some clinical features statistically significant. As a rule, only allows the comparison of the two alternatives.

5. An algorithmic method of diagnosis. This method is based on a certain amount of accurate knowledge, verified as a result of clinical, laboratory, instrumental and postmortem studies, and in some cases it is based on expert assessments. Working hypothesis: there are signs of the disease and the most general features, designed to consistently refine diagnostic situation. There is a “logical tree” signs. Confirming or denying the existence of the patient's symptoms, it is made a gradual transition to higher stages of the diagnostic procedure. As a result, the choice between alternative diseases takes place. Disadvantages: the algorithm has a rigid schema, when you make any change in unpredictable way, the result of diagnostic procedures also changes.

6. Simulation modeling. This is one of the algorithmic diagnostic methods. The method of simulation modeling enables diagnostic experiment not on a living person, but on a phantom of information, each time returning to the starting position until the desired result. It allows you to display a pathological condition. Disadvantages are the same as that of algorithmic diagnostic method.

7. Expert systems. The peculiarity is a dialogue mode with the operator, when the system can perform a diagnostic assessment of the situation and send requests for missing information. In contrast to the algorithmic method of diagnosis, it is outlined the differences between the database and the mechanisms that operate the data. It allows you to bring the algorithm of its work to the logic of the doctor's actions in the construction of diagnosis. It has a high diagnostic accuracy in cases of detection of rare syndromes. Disadvantages: there is no single unified framework for the development of such systems, the narrow specialization of individual development, the subjective factor, introduced by an expert operator.

8. Intelligent Systems. They represent frame from the set position (slots) that contains declarative and quantitative data, the connection

between them (semantic network) and data application rules and communication (product). There are confirmation mode solutions, including differentiating features or levels of confidence. Association mechanism is used in which one concept is based on similarity, contiguity, or contrast to other. Let's take into account the polymorphic of clinical manifestations and atypical forms of disease. The disadvantage: the difficulty of structuring and formalization of information.

9. The hybrid system. It is a processing the logic knowledge components in conjunction with the computational procedures or mathematical models. It is used either set of mathematical and logical-linguistic models or statistical expert system for differential diagnosis. The integration of software and hardware electrophysiological and biochemical data processing systems significantly enrich the hybrid system. Disadvantage - the same difficulty of structuring and formalization of medical knowledge in the case where it does not apply mathematical data.

10. Neural networks. Artificial neural network is a mathematical model, which is a special case of the discriminant analysis. It is a self-learning system of interconnected and interacting units (artificial neurons). Each unit of such a network is concerned only with the information that it receives from other units, and the information that it sends to other units. Technical training is to find the coefficients of the connections between neurons. In the process of training the neural network is able to identify complex relationships between inputs and outputs, as well as to carry out a generalization. This means that in case of successful learning network will be able to return the correct result on the basis of data that were missing in the training set. When training network offers a variety of diseases and set their attributes. In this case the aggregate of all signs should clearly identify the disease. At the end of the learning network can bring previously unknown to her features, and receive from it a response to their belonging to a specific disease. The network may also signal that the input characteristics do not belong to any of the existing disease condition that is new to the network. The disadvantages of this method include the complexity of establishing the coefficients of connections between neurons, and also that if the signs of insufficient network can correlate them with several diseases.

Modern automatic analyzers allow to obtain a set of numerical indicators, e.g. in hematology. However, all these blood parameters in most cases are not considered. As a rule, hematologists diagnose on the basis of several of them plus history.

An automated diagnostic system can explore the full information. In this case it would be logical to use a statistical method for processing information, adding to its methods of probability theory. Then the method of probabilistic diagnosis arises, devoid of difficulty of structuring and formalization of medical knowledge, as all indicators are numerical, and they don't have drawbacks of statistical method in the determination of "weights", that are identified patient symptoms as equivalent data. The main point of the method is in the fact that, as in the statistical method, the analysis of the frequency of the values of certain blood parameters specific for this or that disease, calculates the probability of the test conditions for each value of a parameter, and then calculated the probability of a disease for several indicators. Of course, the diagnosis can be made and not with one hundred percent probability, but also in the diagnosis of the doctor cannot exclude the possibility of error. Therefore, firstly, the diagnosis is almost always a probability, and secondly, the increasing number of indicators all diagnosis is more likely to be inclined toward a particular disease.

## 5. LEARNING THE APPLICATION DOMAIN

Medical data for research is collected from a variety of sources, so it may contain various noises, contradictions, missing values, bias, and redundancy. Important information can be stored in unstructured form.

Since there are different data sets for different diseases and the data sets may have different characteristics, the range of low confidence must be defined specifically. In addition, the required data set to verify the structure of the system should consist of clinical measurements, the actual class, the actual treatment performed, and the actual treatment effect [75].

When collecting data for analysis, it is also possible, multi-level integration of health services, in which high efficiency of storage and use of patients input data is achieved [54]. The characteristics of the data collected for analysis may be different: measurable clinical (for example, white blood cell count, neutrophil count, hemoglobin, etc.), observable clinical data (for example, drowsiness, angina, cough, etc.), generally accepted vital signs (for example, respiratory rate, heart rate, oxygen saturation, etc.) [34].

To create a data set, as a rule, existing data stores are used, after extraction of which they need to be normalized.

Discretization and normalization are two data transformation procedures that help to present data and their relationships accurately in a tabular format, making the database easy to understand and efficient to operate. It also reduces data redundancy and increases efficiency. Data elements can be normalized by allocating a unique column number for each possible value. Numeric data fields are sampled by taking values that are within a range defined by the minimum and maximum limits. In such cases, it is possible to divide this range into several sub ranges and assign a unique column number for each sub range, respectively [61].

In [36], preprocessing of data includes recalculating the non-stratified random sampling (R), the synthetic minority over-sampling method (SMOTE), clean data out of range attribute (COR), and remove duplicate (RD).

Thus, there are several reasons for cleaning and preprocessing data, which are described in [9]:

1) The problem of data types. Variables can be of different types: binary, nominal, ordinal, or numeric.

2) The problem of abstraction. A set of variables can express the same concept. This requires aggregating them into one variable, which eliminates redundancy and can prevent errors in the data set.

3) The problem of a temporary nature. Some clinical parameters are recorded in an electronic medical record several times during his stay in a medical facility, and this may lead to inconsistencies in the data set.

4) The problem of missing knowledge and the problem of missing values. The lack of the expected data elements is common, and these elements may be clinically significant for a particular patient or for a specific disease.

Various methods are proposed for solving these problems.

Part of the data is deleted along with deleted rows that have columns with empty or unknown values. For example, in the project [47], data was reduced from 447 patients to 178. This data was translated into a format that WEKA can interpret. Here, the preferred format is an attribute relationship file (.arff), where the type of data loaded can be determined, and then provide the data itself.

It should be noted here that the analysis of the work of many authors on the processing of medical data shows that WEKA is the preferred platform for many researchers [35, 42, 47, 67, 75, 76].

There are also studies where the development of the classifier did not miss a single observation and a single spatial object, which shows the efficiency of using observations with missing values than discarding incomplete observations that cannot be processed by many algorithms [77].

Missing data may have its own information value. In the process of developing an application designed to support a list of medical issues, the researchers examined whether the lack of clinical data could provide useful information when building prediction models. In this study, they experimented with four methods for treating missing values in a clinical data set — two of them explicitly simulate the absence or lack of data. The results showed that in most cases, classifiers trained using explicit methods for processing missing values work better [78].

Knowledge graphs are data structures that efficiently and intuitively encode various entities and relationships between them. In [38], in order to create feature vectors for the machine learning classification method, the system first groups and aggregates all patient data to form feature vectors. Further, feature vectors are encoded as knowledge graphs.

When data sets are imbalanced, standard learning algorithms are faced with the impossibility of finding patterns based on the correct distribution, which determines the number of data points in each class. In other words, these algorithms tend to assign each invisible data point to the most frequent class or dominant class. Therefore, although they can achieve acceptable accuracy, they do not have acceptable performance. For such data sets, a new algorithm is proposed in [57], which initially assumes a number of sub-samples from the majority class with minority sizes. Considering each of the sub-sampling data from the majority class in addition to the minority class data as temporal data, a decision tree, or a multi-layer perceptron is trained on temporal data. Finally, all classifiers work together as an ensemble.

## 6. DEVELOPMENT OF A CONCEPTUAL MODEL OF DIAGNOSTIC OBJECT

In a formalized form, the diagnosis of a patient using a computer system is reduced to the task of determining whether the current state of the organism or its individual subsystem belongs to one of the formalized states from the set of diagnoses $\{D_i\}$. At the same time, the system of diagnostic features $X_i$ adopted in medical practice is analyzed, which, to some extent, reflects the current $j$ -th state of the $i$ -th subsystem of the organism $S_{ij}$ .

In the framework of this study, a conceptual model of Diagnostic object (DO) is proposed, which is based not only on diagnostic features, but on the results of laboratory and clinical diagnostics. To confirm the diagnosis, the results of Biochemical blood test (BBT) were used. Thus, the internal states $S_{ij}$ of the different levels of the hierarchy are projected onto the $X_i$ feature space and the components of the biochemical blood analysis of the $Y_i$ , and on the other hand, the $S_{ij}$ are projected onto the diagnosis system $\{D_i\}$, while the diagnostic task is to determine the dependence $X_i \to \{D_i\}$, as shown in Figure 1.

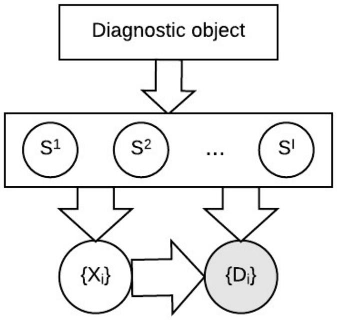
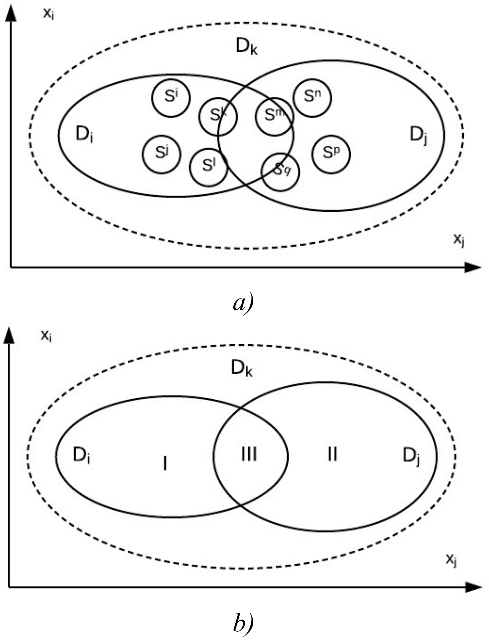


*Figure 1. Scheme of state relations - symptoms - primary diagnoses - components of the BBT - clinical diagnosis*

Let us consider the features of the conceptual model of DO and the nature of the state connections - symptoms - diagnoses.

In Figure 1, the DO display is indicated with the figure arrows (the state of the level subsystems 1 - 9) on the diagnosis space and the symptoms space. It should be noted that some formalized diagnosis states may correspond to the disturbance of homeostasis at several levels of interaction, and even a reliable diagnosis does not always reflect the cause of such disturbance [79]. Based on the diagnostic features, a list of preliminary diagnoses is formed. Further, a list of these preliminary diagnoses forms a number of components BBT, based on the results of which a list of clinical (specifying) diagnoses is formed.

When setting a clinical diagnosis, classes of preliminary diagnoses of different levels of detail and classes of final diagnosis should be distinguished. The classification quality indicators depend on the location of the scattering ellipsoids of objects belonging to the specified classes. Taking into account the features of DO mentioned above, it can be stated that the scattering ellipsoids of the classes of the final diagnosis have a more thin structure consisting of subsets of internal states of the subsystems with different levels of interaction, as shown in Figure 2 (a, b).



*a)*

*b)*

*Figure 2. Illustration of the arrangement of scattering ellipsoids classes in the space of symptoms*

Figure 2, (a) illustrates the hierarchy of internal states of subsystems and diagnoses of different levels of interaction. Diagnoses of one level ( $D_i$ and $D_j$ ) are included in the cluster of diagnosis of a higher level of $D_k$ , and each of them includes a number of internal states, and some states can simultaneously be included in several diagnoses, which lead to the intersection of scattering ellipsoids. In the case of intersecting scattering ellipsoids, in each ellipsoid, it is possible to conditionally select the region of the reliable solution and the region of the unreliable solution. In Figure 2 (b), the regions of the reliable solution for the states $D_i$ and $D_j$ are identified by numbers I and II respectively, and by number 3 - the region of the unreliable decision. Different Decision rule (DR) variants minimize region III by different criteria (mean risk, mini-max, maxi-min, etc. [80, 81]).

The structural diagram of the information support system of the decision-maker (DM) on the basis of the proposed conceptual model of DO is shown in Figure 3.

At the first stage, a set of diagnostic symptoms $X_1$ (of a syndrome) is received and processed, after which a detailed diagnosis (transformation $X_i \to \{D_i\}$) is performed. This diagram depicts a system with the verification of intermediate data, while the $D_M$ confirms or corrects the obtained results (transformations of $D_i \to D_i^V$ and $Y_i \to Y_i^V$).

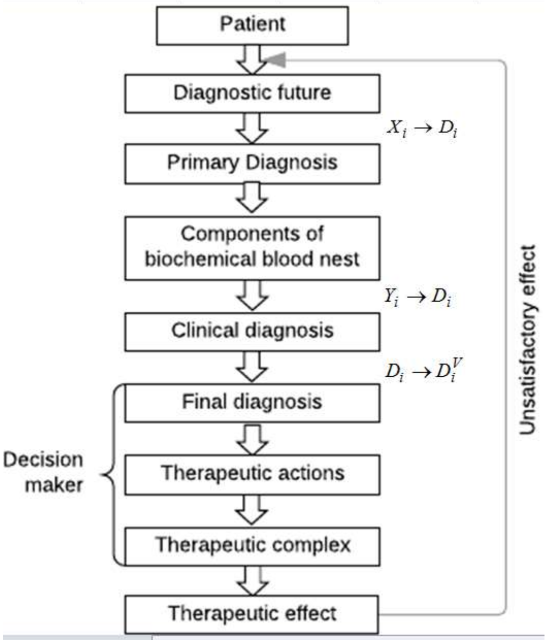


*Figure 3. Informational support of DM*

## 7. AN APPLICATION OF THE PROPOSED MODEL AND ALGORITHM ON THE DIAGNOSIS OF A LIVER FAILURE

A set of liver diseases was selected for the approbation and realization of the experimental study that is proposed by the conceptual model Diagnostic Object. Such choice is reasoned by the fact that today there are about 200 million of people in the world who suffer of liver diseases which is included into a top ten of frequently observed death causes. A more often cases of liver failure are due to viruses and toxic chemicals. The most popular outcome of chronic liver pathologies is cirrhosis [82].

A set of liver diseases $D_i$ includes hepatitis (viral, drug, toxic, ischemic); cirrhosis (alcoholic, biliary, postnecrotic, with hemochromatosis, rare types); Neoplasms of the liver (hepatic cell carcinoma, metastases to the liver, cysts, abscess); Infiltrative lesions of the liver (amyloidosis, glycogenoses, fatty liver, lymphomas, granulomatosis); Functional disorders with jaundice (Gilbert syndrome, pregnant cholestasis, Kriegler-Nayyar syndrome, Dubin-Johnson syndrome); Lesions of intrahepatic bile ducts (occlusion of the bile duct, inflammation of the bile duct); Cardiovascular pathology (congestive liver with heart failure and cardiac cirrhosis, thrombosis of the hepatic veins, arteriovenous fistula).

A set of diagnostic symptoms $X_i$ of a liver disease $D_i$ include: sensation of a crowded abdomen; low levels of working capacity; pain in the right hypochondrium (under ribs) after an alcohol consumption, disturbance of a diet or lifting heavy weights; people quickly gorge; periodic insignificant increase in temperature; bleeding of teeth when cleaning them; periodic nasal bleeding [83].

For the diagnosis of the certain groups of liver diseases BBT is enough. Further, for an example, BBT components were described for the diagnostic search of liver cirrhosis.

During BBT in a step of compensation, insignificant amount of deviations are identified in functional samples of a liver $Y_i$ : hyperproteinemia, a slight increase in bilirubin. In a step of decompensation - expressed disproteinemia, a decrease in the content of protombine and cholisterine, an increase in bilirubin, a moderate increase of aminotransferases activity. Immunological disorders expressed insignificantly. In some patients, the IgA content is clearly increased.

The data of the third stage of diagnostic results search of clinic and biochemical blood analysis reveal symptoms of expressed liver cytolysis and functional insufficiency of hepatocytes: a significant increase in bilirubin levels, a 5-10-fold increase in aminotransferases level, lactate dehydrogenase (LDH or LD) and its 4-5-th fractions, a decrease in amount of cholesterine and prothrombine , dysproteinemia (a sharp increase in the number of y-globulins and a significant decrease in albumins, a change in sediment samples, especially thymolone). A clinical blood test reveals enlargement of erythrocyte sedimentation rate (ESR), a «shift» to the left in the leukocyte formula, there is usually no leukocytosis, hypersplenism is less common, compared to shallow cirrhosis of liver.

In primary biliary cirrhosis, the change in immunological indicators is regularly detected: an increase in IgG and IgM titer (more IgM), severe hyper-y-globulinemia, increased ESR. Very specific for primary biliary cirrhosis is the appearance of antimitochondrial antibodies in high titre.

For the realization of the proposed conceptual model DO, an application was developed in a sphere of Visual Studio 2014. A variety of diagnosis, symptoms, BBT components etc. are collected and stored in a database, designed in MS SQL Server 2014. The architecture of the information system is analogous to [84]. Structure of databases is provided in figure 4.

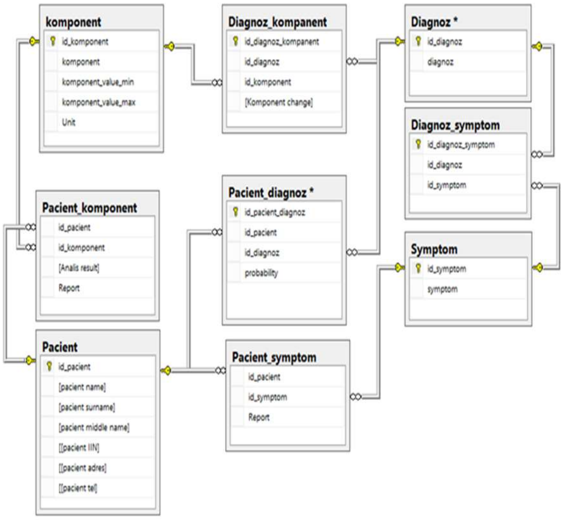


*Figure 4. Structure of databases*

## 8. PRELIMINARY PROCESSING OF MEDICAL DATA BASED ON SELECTION OF INFORMATIVE SIGNS

Mathematical processing of the initial medical data makes it possible to determine the diagnostic value of the indices or their complexes, which in the future helps to construct the optimal survey plan, and significantly reduce the number of necessary for diagnostic studies and ultimately improve the quality of data mining.

To reduce the space of diagnostic signs, methods are used to divide the total number of indicators studied into homogeneous groups, the so-called classification methods, the most acceptable of which are cluster methods.

The task of cluster analysis is to divide the set of objects G into m (m – integer) of clusters (subsets) Q1, Q2,..., Qm based on the data contained in the set X, so that each object $G_J$ belonged to one and only one subset of the partition and that the objects belonging to the same cluster were similar, while the objects belonging to different clusters were heterogeneous.

At the same time, there are quite a lot of cluster analysis methods: the method of complete connections, the maximum local distance method, the centroid method (k-average). [74].

However, when using them, the indispensable condition is the uniformity of indicators, which is almost never observed in medical diagnostics.

This paper presents the result of the algorithm for the structural identification of diagnostic features based on the “defect” algorithm, which performs clustering of diagnostic features with regard to their heterogeneity and internal connections.

The study of blood parameters is necessary to obtain calculations on the basis of which the analysis of the patient's condition and prediction of his condition will be made. Processing of a set of data is made for various target groups on age, sex, pathology information for different clustering parameters.

To improve the effectiveness of the methods of data mining for the process of diagnosing diseases, a procedure for assessing the informativeness of heterogeneous diagnostic indicators on the basis of a theoretical information approach has been proposed.

To determine the value of the obtained results of a patient's examination on the basis of an analysis of simple independent diagnostic features, it is expedient to use the basic provisions of information theory, one of which is the estimation of the amount of information.

Suppose that there is some system of diagnoses D, which consists of n diseases. Based on statistical information and based on medical data, even before the examination of the patient, it is possible to calculate the a priori probabilities of the occurrence of a disease $P(D_i)$. In fact, such a probability will reflect the frequency of occurrence of each diagnosis in the sample being processed.

The uncertainty of the system of possible diagnoses is estimated using entropy or the amount of information (1):

$$H(D) = -\sum_{i=1}^{n} P(D_i) \cdot \log_2 P(D_i) \qquad (1)$$

where $H(D)$ is the measure of uncertainty (entropy) of the diagnosis system; $P(D_i)$ is the a priori probability of diagnosis $D_i$ .

The quantity $H$ is also called the content and is always a positive quantity. In this case, for n possible equiprobable components, its value will be maximal, and formula (1) takes the following form (2):

$$H(D) = -\sum_{i=1}^{n} P(D_i) \cdot \log_2 P(D_i) = -\sum_{i=1}^{n} \frac{1}{n} \cdot \log_2 \frac{1}{n} = \log_2 n \qquad (2)$$

Since entropy reflects the measure of the uncertainty of the system, its magnitude will change when new information enters the system. Such information for diagnoses is the data obtained as a result of a patient's examination. The decrease in entropy occurs by an amount equal to the amount of information entered.

Accordingly, the amount of information entered into the system is defined as the difference between the value of entropy before and after the examination (3):

$$Z_D\left(k_j\right)=H(D)-H\left(\frac{D}{k_j}\right) \tag{3}$$

where $Z_D\left(k_j\right)$ is the amount of information entered into the system after the patient is examined for $k_j$; $H(D)$- initial (primary) entropy of the diagnosis system; $H\left(D/k_j\right)$ is the entropy of the system after the survey, taking into account the sign of $k_{jj}$.

Thus, the value of $Z_D\left(k_j\right)$ characterizes the diagnostic value of the symptom $k_j$ in relation to the diagnosis system $D$ and is based on the amount of information received. The unit of measurement of the diagnostic value of the trait or complex of characteristics is the information, the introduction of which eliminates the uncertainty in $N$ equiprobable diagnoses. The diagnostic value of a simple attribute that takes one of two possible values is determined by the formula (4):

$$Z_{D_i}\left(k_j\right)=\log_2\frac{P\left(k_j/D_i\right)}{P\left(k_j\right)} \tag{4}$$

where $Z_{D_i}\left(k_j\right)$ is the diagnostic weight of $k_j$ for disease $D_i$; $P\left(k_j/D_i\right)$ - frequency of occurrence or a priori probability of presence of a sign at disease $D_i$; $P\left(k_j\right)$ - frequency of occurrence or a priori probability of presence of a sign in all system of possible diagnoses $D$.

The value of $P\left(k_j/D_i\right)$ is calculated as the ratio of the number of patients with the presence of $k_j$ in $D_i$ disease to the total number of patients with the disease under consideration (5):

$$P\left(\frac{k_j}{D_i}\right)=\frac{\sum\limits_{k_j=1}k_j\left(D_i\right)}{\sum\limits_{D_i=1}D_i} \tag{5}$$

Based on (5), we can conclude that with the same value of the probability of the presence of a characteristic for a particular disease and for the entire system of diagnoses, the diagnostic weight of the trait is zero and the sign does not carry any information content.

The diagnostic weight of the absence of a simple characteristic is determined using the expression obtained from formula (5) by introducing the reciprocal of the probabilities (6):

$$Z_{D_i}\left(k_j\right)=\log_2\frac{1-P\left(k_j/D_i\right)}{1-P\left(k_j\right)} \tag{6}$$

It should be borne in mind that the diagnostic weight of the sign can be either positive or negative, that is, how to reduce, and increase the likelihood of a diagnosis.

The total diagnostic weight of a simple sign for the disease $D_i$ takes into accounts both the presence and absence of the trait, and can be calculated from the expression (7):

$$\begin{aligned}Z_{D_i}\left(k_j\right)&=P\left(k_j/D_i\right)\cdot\log_2\frac{P\left(k_j/D_i\right)}{P\left(k_j\right)}+\\&+\left[1-P\left(k_j/D_i\right)\right]\cdot\log_2\frac{1-P\left(k_j/D_i\right)}{1-P\left(k_j\right)}\end{aligned} \tag{7}$$

Diagnostic value of a simple trait for a disease system (8):

$$Z_D\left(k_j\right)=\sum_{i=1}^{n}P\left(D_i\right)\cdot Z_{D_i}\left(k_j\right) \tag{8}$$

For some simple signs, the diagnostic weights according to (8) assume the limiting values, which indicates the presence of a deterministic connection.

## 9. RESULTS OF INTELLECTUAL DATA ANALYSIS

The results of the analysis of the relationship between the supply of magnesium and the risk of somatic diseases in women aged 18-45 by methods of intellectual data analysis were obtained. For the standard processing of the results of the research, methods of mathematical statistics were used, including calculation of numerical characteristics of random variables, testing of statistical hypotheses using parametric and nonparametric criteria, correlation and variance analysis [85]. A comparison of the predicted and observed frequency of occurrence of the investigated features was carried out using the Chi-square test (Pearson) test, the Wilcoxon-Mann-Whitney T test, and the Student test.

The results of the analysis indicated a pronounced relationship between magnesium levels in the blood (plasma, erythrocytes) and an increased risk of somatic diseases in women of reproductive age.

Using modern methods of data mining allowed not only to establish the risk factors of these diseases, but also to use the received risk factors for finding informative predictors and then sets of logical rules. These rules allow us to identify patients of reproductive age, for whom replenishment of magnesium deficiency is vital.

An analysis of the interactions of the parameters included in the main condensation (Figure 5) showed that lower levels of magnesium in the blood plasma and in erythrocytes were significantly associated with a number of diseases. The results obtained from the analysis of a group of women of reproductive age confirm the earlier general conclusion that magnesium concentrations in plasma less than 0.8 mmol/l correspond to an increased risk of somatic disease. In this case, the average magnesium levels in the blood vary significantly for various diseases.

In Figure 5, the red line shows the thickening boundaries. Each point in the diagram corresponds to one of the studied indicators, including diagnoses by ICD-10, and the distance between a pair of points is the value of the statistical reliability of the interaction between the corresponding parameters.

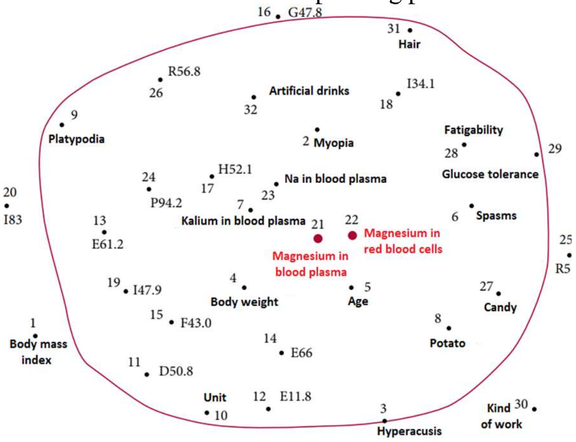


*Figure 5. Condensation of statistically reliable associations between the indicators of the state of somatic health of women of reproductive age*

The next describe the data mining results of age dependence on blood pressure and sugar contained in the blood [86].

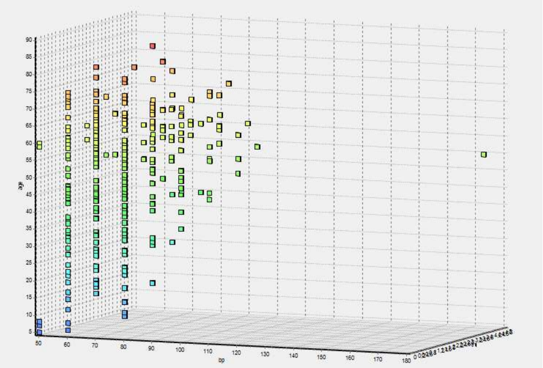


*Figure 6. Age dependence on blood pressure and sugar*

Figure 6 shows the age dependence on blood pressure and sugar contained in the blood.

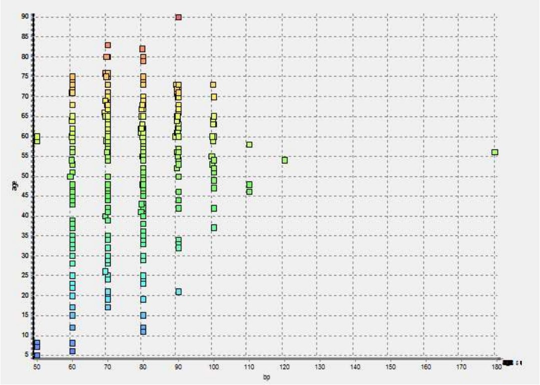


*Figure 7. Age dependence on blood pressure*

In Figure 7, one can observe an increase in blood pressure between the ages of 25 and 75. The peak of blood pressure is 45-60 years. The lowest values of pressure were recorded at the age of up to 5 years.

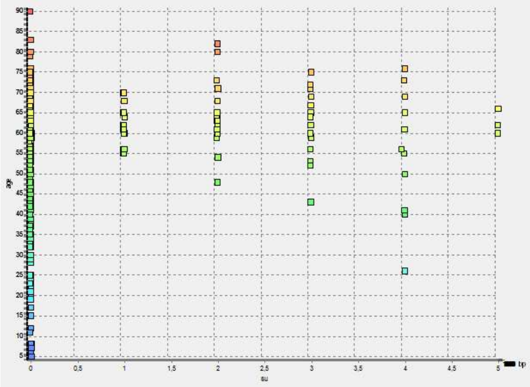


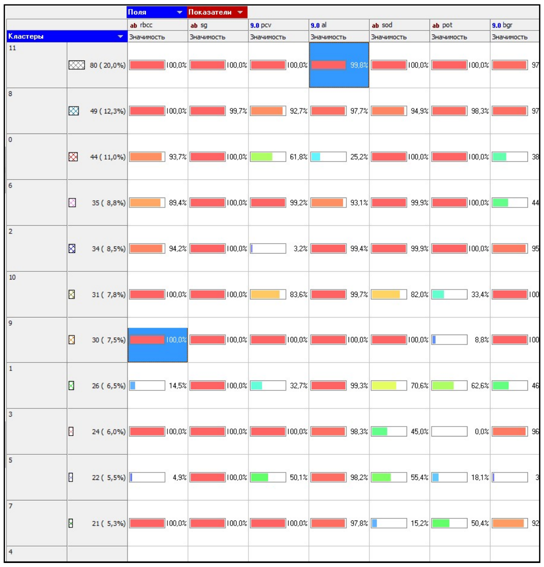
*Figure 8. Age dependence of blood sugar*

Judging from Figure 8, it can be assumed that the average sugar content is generally between 40 and 80 years old, and the minimum values were also found in this age range, from 55 to 77. Based on figure 3, it can be noted that the highest blood sugar is in the range of 60 to 65 years.

Figures 9 show 11 clusters, the clusters are divided according to the number of red blood cells per cm2, which in turn are one of the important signs of anemia.

The largest cluster, cluster 11, which accounts for 20% of all records, in this cluster included people who have all the indicators: the number of red blood cells is rbcc, the specific gravity is sg, the cell volume is pcv, albumin is al, sodium is sod, Potassium - pot, blood glucose - bgr, age - age, sugar - su is within the norms.

The smallest cluster is cluster 4, in this cluster 4, approximately 1% of records entered. Records with deviations from the norm, according to such indicators as: the volume of blood cells and sugar.



| | | Поля | Показатели | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | ab rbcc | ab sg | 9.0 pcv | 9.0 al | ab sod | ab pot | 9.0 bgr |
| Кластеры | | Значимость | Значимость | Значимость | Значимость | Значимость | Значимость | Значимость |
| 11 | 80 ( 20,0%) | 100,0% | 100,0% | 100,0% | 99,8% | 100,0% | 100,0% | 97 |
| 8 | 49 ( 12,3%) | 100,0% | 99,7% | 92,7% | 97,7% | 94,9% | 98,3% | 97 |
| 0 | 44 ( 11,0%) | 93,7% | 100,0% | 61,8% | 25,2% | 100,0% | 100,0% | 38 |
| 6 | 35 ( 8,8%) | 89,4% | 100,0% | 99,2% | 93,1% | 99,9% | 100,0% | 44 |
| 2 | 34 ( 8,5%) | 94,2% | 100,0% | 3,2% | 99,4% | 99,9% | 100,0% | 95 |
| 10 | 31 ( 7,8%) | 100,0% | 100,0% | 83,6% | 99,7% | 82,0% | 33,4% | 100 |
| 9 | 30 ( 7,5%) | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 8,8% | 100 |
| 1 | 26 ( 6,5%) | 14,5% | 100,0% | 32,7% | 99,3% | 70,6% | 62,6% | 46 |
| 3 | 24 ( 6,0%) | 100,0% | 100,0% | 100,0% | 98,3% | 45,0% | 0,0% | 96 |
| 5 | 22 ( 5,5%) | 4,9% | 100,0% | 50,1% | 98,2% | 55,4% | 18,1% | 3 |
| 7 | 21 ( 5,3%) | 100,0% | 100,0% | 100,0% | 97,8% | 15,2% | 50,4% | 92 |
| 4 | | | | | | | | |

*Figure 9. Clustering*

*Table 1. The number of erythrocytes of each cluster*

| № Cluster | Number of erythrocytes per 1 см2 |
|---|---|
| 0 | 8120 |
| 1 | 7600 |
| 2 | 7790 |
| 3 | 7900 |
| 4 | 6675 |
| 5 | 8390 |
| 6 | 7942 |
| 7 | 7605 |
| 8 | 9020 |
| 9 | 7413 |
| 10 | 8753 |
| 11 | 8961 |

Clusters 1 and 5 should also be selected, which make 6.5 and 5.5% of the total number of records. In these clusters, there are multiple deviations from the norms for such indicators as: the number of red bodies, the volume of cells, the content of such substances as sodium, potassium, glucose.

The result of the stage of clustering the space of diagnostic signs is a multilevel hierarchical structure of diagnostic signs, which more adequately reflects the complex interaction of subsystems of the body.

Evaluation of the informativeness of diagnostic features (intervals) is performed at each level of the hierarchy. After that, the signs are ranked according to their informative value, and a subspace of informative signs is constructed. Alternatives to this stage are the criteria for selecting a system of threshold elements.

At the stage of the hierarchical structure synthesis of decision rules the hierarchical clustering of diagnoses (diagnosable body states) is performed.

In this case, the same methods and clustering algorithms are used as in the synthesis of the hierarchical structure of diagnostic features, and, accordingly, the same alternatives.

Prospects for further research are to verify the adequacy of the application of the developed approach for the synthesis of a diagnostic decision rule by analyzing the reliability of object recognition by the initial set of indicators and by the selected informative subset.

## 10. PROBLEMS OF DATA MINING IN HEALTHCARE

In this section, we will look at a number of challenges faced by both research and data mining practices in healthcare.

a) Heterogeneity of medical data

Currently, centralized medical information databases are limited or non-existent. Much of the potentially important medical information is not stored electronically. Some of the medical information that are stored electronically are scattered across hundreds of small databases in various clinics, hospitals, and laboratories. This data can be in various formats (for example, text, image, and video) and is collected from various sources, such as patient records, physician comments, and laboratory test results.

The main areas of heterogeneity of medical data can be organized according to these headings:

- – Volume and complexity of medical data.
- – Doctor's interpretation.
- – Sensitivity and specificity analysis.
- – Poor mathematical performance.

More and more medical procedures use imaging as a preferred diagnostic tool, as images are easiest for people to understand, and they can provide more information in one picture of the results. Therefore, there is a need to develop methods for efficient development of image data in databases that are more complex than processing in purely numeric databases. This heterogeneity requires high-capacity data storage devices and new tools for analyzing such data.

Another unique feature in the development of medical data is that basic medical data structures are poorly mathematically characterized compared to many fields of the physical sciences. Physicists collect data that they can contribute to formulas, equations, and models that reasonably reflect the relationships between their data. On the other hand, the conceptual structure of medicine consists of word-descriptions and images with a very small number of formal restrictions on vocabulary, composition of images, or permissible relationships between basic concepts [87]

Many of the potential applications of intelligent health information analysis discussed in the previous sections require centralized databases combining different formats of health data from various sources. Although there is an impetus from the government of Kazakhstan to develop such centralized databases, these projects are still in the status of newborns.

The use of constantly growing large volumes of medical data in solving diagnostic, therapeutic, statistical and managerial tasks for the development of e-health in Kazakhstan is relevant. The solution to this problem in the country is to organize a single information space and its technological infrastructure [88].

b) Insufficient communication between information technology and the health community

The introduction of IT in healthcare is especially slow and lags behind other areas. This is due to the complexity of issues such as compatibility, technological rationality, acceptability, management rationality, data security, data quality and standards. Clinical information systems usually provide a wide range of data warehouses, medical reports, clinical decision support systems, etc. As a rule, they are not available in the complex. In addition, current implementations of clinical information systems have a lack of functionality to provide easy access. In general, weak or even missing support for patient information exchange within the health care system can be observed, which impedes immediate access to current and complete patient information [4].

The main problem of the application of data mining methods in health care is the disconnection between information technology and the medical community. At the most basic level, while simple practical questions, such as placing computers in sterile areas or teaching doctors to use different software packages, are often considered trivial in the computer science community, these are actually very complex tasks. For example, one of the main reasons why medical professionals do not use the clinical decision support system is the extreme difficulties that physicians face when interacting with electronic records [89].

c) Legal, ethical and social issues

The composition of the medical data potentially available for data mining is huge. Thousands of terabytes are now generated annually worldwide. However, these data are buried in heterogeneous databases and scattered throughout the medical field without any common format or organization principles. The issue of ownership of patient information is unresolved.

Legal aspects of owning medical data, fear of litigation and privacy issues and other issues that currently limit the widespread use of data mining in the field of health informatics.

The article [82], which considers the features of data mining using medical data, discusses the ethical and legal aspects of using Data Mining for medical data, including data ownership, fear of legal processes, expected benefits and specific administrative issues.

The mathematical understanding of the assessment and hypothesis formation in medical data may be fundamentally different than in other data collection activities. Medicine is primarily aimed at activities for the treatment and care of patients and, only secondarily, as a research resource.

## 11. CONCLUSION

The paper deals with the mathematical apparatus of preliminary processing of medical data based on the selection of informative features. This information-theoretical approach is the most rigorous and formalized and can serve as a basis for constructing more complex methods. The results of the application of intellectual analysis for clustering tasks are described below. Associations were found between the indicators of the state of somatic health of women of reproductive age, as well as the dependence of age on blood pressure and sugar contained in the blood.

An algorithm and methodology for clustering medical data based on the entropy approach has been developed. On the basis of the entropy approach, an algorithm and a method for clustering data that take into account the specifics of the subject area have been developed, namely: a small number of features characterizing the basic essence; the use of nominal scale for measuring features; high variability of features.

The result of their work is the presentation of data in the form of homogeneous groups (clusters), each of which corresponds to the previously specified parameters.

The proposed algorithm allows calculating the dynamic dependence between the obtained clusters and interactions with the medical database.

The work has scientific and practical potential. Clustered data presented in the form of independent statistical entities and possessing a set of additional features are material for further research

From the received results the conclusion about necessity of opening of the analytical centers generalizing the information from different sorts of sources on the basis of Data Mining (probably, it will be Data-centers) was made. In addition, there is a clear need to train specialists to work in such centers, who are able not only to conduct complex types of analysis, but, most importantly, to interpret the results, formulate them in an accessible understanding of the majority of forms and prepare management solutions based on them.

## REFERENCES

[1] Barsegyan, A., Kupriyanov, M., Holod, I., Tess, M., & Yelisarov, S. (2009). *Analysis of data and processes: textbook.* St. Petersburg: BKHV-Petersburg, 2009. (in Russian).

[2] Schiff, G., Hasan, O., Kim, S., Abrams, R, Cosby, K., Lambert, B., … McNutt, R. (2009). Diagnostic Error in MedicineAnalysis of 583 Physician-Reported Errors. *Arch Intern Med.* 2009;169(20):1881–1887. doi:10.1001/archinternmed.2009.333

[3] Chen, H., Fuller, Sh., Friedman, C., & Hersh, W. (2005) Medical Informatics: Knowledge Management and Data Mining in Biomedicine. *Springer Science+Business Media*, New York, USA. ISBN-10: 0-387-24381-X (HB)

[4] Islam, M., Poly, T., & Li, Y-Ch. (2018). Recent Advancement of Clinical Information Systems: Opportunities and Challenges. *Yearbook of medical informatics* 27(01):083-090. DOI: 10.1055/s-0038-1667075

[5] Bellazzi, R., & Zupan, B. (2018). Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics* 77(2):81-97. DOI: 10.1016/j.ijmedinf.2006.11.006

[6] Gozali, E., Rahimi, B., Sadeghi, M., & Safdari R. (2017). Diagnosis of diseases using data mining. DOI: 10.26415/2572-004X-vol1iss4p120-121

[7] Sen, I., & Khandelwal, K. (2018) Data Mining in Healthcare. DOI: 10.13140/RG.2.2.22189.38887

[8] Yoo, I., Alafaireet, P., Marinov, M., & Hua L. (2011) Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *Journal of Medical Systems* 36(4):2431-48. DOI: 10.1007/s10916-011-9710-5

[9] Iavindrasana, J., Depeursinge, A., Cohen, G., & Muller H. (2009) Clinical Data Mining: a Review. *Methods of Information in Medicine*. DOI: 10.1055/s-0038-1638651

[10] Chen. L-Y., & Fawcett, T. (2016) Using Data Mining Strategies in Clinical Decision Making: A Literature Review. *Computers Informatics Nursing* 34(10):1. DOI: 10.1097/CIN.0000000000000282

[11] Jothi, N., Abdul Rashid, N.., & Husain, W. (2015) Data Mining in Healthcare – A Review. *Procedia Computer Science* 72:306-313. DOI: 10.1016/j.procs.2015.12.145

[12] Kala, R., Shukla, A., & Tiwari, R. (2010) Hybrid Intelligent Systems for Medical Diagnosis. *Intelligent Medical technologies and Biomedical Engineering: Tools and Applications*. Chapter: 9. Publisher: IGI Global – 2010

[13] Wagholikar, K., Sundararajan, V., & Deshpande A. (2011) Modeling Paradigms for Medical Diagnostic Decision Support: A Survey and Future Directions. *Journal of Medical Systems* 36(5):3029-49. DOI: 10.1007/s10916-011-9780-4

[14] Kitakaze, M., Asakura, M., Nakano, A., Takashima, S., & Washio, T. (2015). Data Mining as a Powerful Tool for Creating Novel Drugs in Cardiovascular Medicine: The Importance of a "Back-and-Forth Loop" Between Clinical Data and Basic Research. *Cardiovascular Drugs and Therapy* 29(3). DOI: 10.1007/s10557-015-6602-9

[15] Esfandiari, N., Babavalian, M., Moghadam, A., & Tabar, K. (2014) Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications* 41(9):4434–4463. DOI: 10.1016/j.eswa.2014.01.011

[16] Uvalieva, I., & Smailova, S. (2014). Development of decision support system to control the quality of education. *In 2014 IEEE 8th International Conference on Application of*

*Information and Communication Technologies* (AICT) (pp. 1-6). IEEE.

[17] Kaisler, S., Armour, F., & Espinosa, J. (2016). Introduction to the Big Data and Analytics: Concepts, Techniques, Methods, and Applications Minitrack. *Conference: 49th Hawaii International Conference on System Sciences* (HICSS). DOI: 10.1109/HICSS.2016.135

[18] Witten, I., Frank, E., Hall, M., & Pal, C. *Data Mining: Practical machine learning tools and techniques*. – Morgan Kaufmann, 2016

[19] Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: literature review and challenges. *International Journal of Distributed Sensor Networks*, 11(8), 431047.

[20] Uvalieva, I., Chettykbayev, R., Utegenova, A., & Toibayeva, S. (2015). Mathematical basis and information system software for educational institutions ranking. *In 2015 9th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 486-490). IEEE.

[21] Holzinger, A., Dehmer, M., & Jurisica, I. (2014). Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. *BMC Bioinformatics* 15(Suppl 6):I1. DOI: 10.1186/1471-2105-15-S6-I1

[22] Dyczkowski, K. (2018). Intelligent Medical Decision Support System Based on Imperfect Information. The Case of Ovarian Tumor Diagnosis. *Studies in Computational Intelligence*, 735. DOI:10.1007/978-3-319-67005-8.

[23] Abdullah, F., Manan, S., Ahmad, A., Wafa, Sh., Razif, Sh., Zulaily, N., …, Ahmed, A. (2017) Data Mining Techniques for Classification of Childhood Obesity Among Year 6 School Children. Conference: *International Conference on Soft Computing and Data Mining*. DOI:10.1007/978-3-319-51281-5_47.

[24] Deo, R. Machine Learning in Medicine. (2015). *Circulation* 132(20):1920-1930. DOI: 10.1161/CIRCULATIONAHA.115.001593

[25] Jung W., Lee, T., Lee, I-S., Kim, S., Jang, H., Kim, S-Y., …, Chae, Y.(2015). Spatial Patterns of the Indications of Acupoints Using Data Mining in Classic Medical Text: A Possible Visualization of the Meridian System. *Evidence-based Complementary and Alternative Medicine* 2015(2):457071. DOI: 10.1155/2015/457071

[26] Futoma, J., Sendak, M., Cameron, B., & Heller, K. (2016). Predicting Disease Progression with a Model for Multivariate Longitudinal Clinical Data. *Proceedings of the 1st Machine Learning for Healthcare Conference*, PMLR 56:42-54.

[27] Weiss, J., Natarajan, S., Peissig, P., McCarty, C., & Page, D. (2012). Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records. *AI Magazine*. DOI: 10.1609/aimag.v33i4.2438

[28] Jaarsma, T., Boshuizen, H., Jarodzka, H., Nap, M., Verboon, P., & Merrienboer, J. (2016). Tracks to a Medical Diagnosis: Expertise Differences in Visual Problem Solving. *Applied Cognitive Psychology* 30(3):n/a-n/a. DOI: 10.1002/acp.3201

[29] Schwaederle, M., Zhao, M., Lee, J., Eggermont, A., Schilsky, R., Mendelsohn, J., …, Kurzrock, R. (2015). Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials. *Journal of Clinical Oncology* 33(32). DOI: 10.1200/JCO.2015.61.5997

[30] Vergara J. R., Estévez P. A. A review of feature selection methods based on mutual information //Neural computing and applications. – 2014. – Т. 24. – №. 1. – С. 175-186.

[31] Schabenberger, O., & Gotway C. (2017). Statistical methods for spatial data analysis. *Stochastic Environmental Research and Risk Assessment*. DOI: 10.1201/9781315275086

[32] Heera, S., & Deshpande, P. (2015). Data Analysis using Multidimensional Modeling, Statistical Analysis and Data Mining on Agriculture Parameters. *Procedia Computer Science* 54:431-439. DOI: 10.1016/j.procs.2015.06.050

[33] Kim, H., Keifer, C., Rodriguez-Seijas, C., Eaton, N., Gadow, K., (2017). Structural hierarchy of autism spectrum disorder symptoms: An integrative framework. *Journal of Child Psychology and Psychiatry* 59(1). DOI: 10.1111/jcpp.12698

[34] Naydenova, E., Tsanas, A., Howie, S., Casals-Pascual, C., De Vos, M. (2016) The power of data mining in diagnosis of childhood pneumonia. *Journal of The Royal Society Interface* 13(120):20160266. DOI: 10.1098/rsif.2016.0266

[35] Nalluri, M., Kannan, K., Manisha M., & Roy, D. (2017) Hybrid Disease Diagnosis Using Multiobjective Optimization with Evolutionary Parameter Optimization. *Journal of Healthcare Engineering* 2017(1):1-27. DOI: 10.1155/2017/5907264

[36] Wiharto, W., Kusnanto, H., & Herianto, H. (2016). Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm. *Healthcare Informatics Research* 22(1):30. DOI: 10.4258/hir.2016.22.1.30

[37] Alizadeh, B., Safdari, R., Zolnoori, M., & Bashiri A. (2015). Developing an Intelligent System for Diagnosis of Asthma Based on Artificial Neural Network. *ACTA Inform Med* 23(4):220. DOI: 10.5455/aim.2015.23.220-223

[38] Vandewiele, G., De Backere, F., Lannoye, K., Berghe, M., Janssens, O., Hoecke, S.., …, Turck, F. (2018). A decision support system to follow up and diagnose primary headache patients using semantically enriched data. *Medical Informatics and Decision Making* 18(1). DOI: 10.1186/s12911-018-0679-6

[39] Selvathi, D., & AarthyPoornila, A. (2017). Performance analysis of various classifiers on deep learning network for breast cancer detection. *International Conference on Signal Processing and Communication* (ICSPC). DOI: 10.1109/CSPC.2017.8305869

[40] Keight, R., Aljaaf, A., Al-Jumeily, D., Hussain, A., Ozge A., & Mallucci C. (2017). An Intelligent Systems Approach to Primary Headache Diagnosis. *International Conference on Intelligent Computing*. DOI: 10.1007/978-3-319-63312-1_6

[41] Sepandi, M., Rahimikazerooni, S., & Rezaianzadeh, A. (2018). Assessing Breast Cancer Risk with an Artificial Neural Network. Asian Paciffic *Journal of Cancer Prevention*, Vol 19. DOI: 10.22034/APJCP.2018.19.4.1017

[42] Shafique, U., Majeed, F., Qaiser, H., & Mustafa, I. (2015). Data Mining in Healthcare for Heart Diseases. *International Journal of Innovation and Applied Studies*. ISSN 2028-9324 Vol. 10 No. 4 Mar. 2015, pp. 1312-1322

[43] Nermend, K., Ghani, W., Hussain, A., & Shihab A. (2018). Classification and Monitoring of Autism using SVM and VMCM. *Journal of Theoretical and Applied Information Technology* 96(14) ISSN: 1992-8645

[44] Afifi Sh., Gholamhosseini, H., & Sinha, R. (2018). Dynamic hardware system for cascade SVM classification of melanoma. Neural Computing and Applications. DOI: 10.1007/s00521-018-3656-1

[45] Huang, M-W., Chen, Ch-W., Lin, W-Ch., Ke, Sh-W., & Tsai, Ch-F. (2017). SVM and SVM ensembles in breast cancer prediction. *PLoS ONE* 12(1):e0161501. DOI: 10.1371/journal.pone.0161501

[46] Jirka, J., Prauzek, M., Krejcar, O., & Kuca, K. (2018) Automatic epilepsy detection using fractal dimensions segmentation and GP–SVM classification. *Neuropsychiatric Disease and Treatment* V 14:2439-2449. DOI: 10.2147/NDT.S167841

[47] Pereira, J., Peixoto, H., Machado, J., & Abelha A. (2017). A Data Mining Approach for Cardiovascular Diagnosis. *Open Computer Science* - 2017; 7:36–40. DOI: 10.1515/comp-2017-0007

[48] Kaur, G., Arora, A., & Jain, V. (2017). Comparative Analysis of Hybrid Models for Prediction of BP Reactivity to Crossed Legs. *Journal of Healthcare Engineering* Volume 2017, Article ID 2187904, 13 pages. doi: 10.1155/2017/2187904

[49] Jatav, Sh., & Sharma, V. (2018). An Algorithm for Predictive Data Mining Approach in Medical Diagnosis. *International Journal of Computer Science & Information Technology* (IJCSIT) Vol 10, No 1, February 2018. DOI: 10.5121/ijcsit.2018.10102

[50] Kourou, K., Exarchos, T., Exarchos, K., Karamouzis, M., & Fotiadis, D. (2014). Machine learning applications in cancer prognosis and prediction. ing applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 13(C). DOI: 10.1016/j.csbj.2014.11.005

[51] Rosado, P., Lequerica-Fernandez, P., Villallain, L, Pena, I., Sanchez-Lasheras, F., & De Vicente, J. (2013). Survival model in oral squamous cell carcinoma based on clinicopathological parameters, molecular markers and support vector machines. *Expert Systems with Applications* 40(12):4770–4776. DOI: 10.1016/j.eswa.2013.02.032

[52] De Canete, J., Gonzalez-Perez S., & Ramos-Diaz, J. (2011). Artificial neural networks for closed loop control of in silico and ad hoc type 1 diabetes. *Computer methods and programs in biomedicine* 106(1):55-66. DOI: 10.1016/j.cmpb.2011.11.006

[53] Hamouda, S., Hamouda, A., & Wahed, M. (2017). Intelligent System for Predicting, Diagnosis and Treatment of Breast Cancer.

*International Journal of Biomedical Data Mining* 06(02). DOI: 10.4172/2090-4924.1000128

[54] Sagi, A., Sabo, A., Kuljić, B., & Szakall T. (2015). Multilevel Interdisciplinary Intelligent Systems in Medicine. Proceedings of the *International Symposium on Intelligent Systems and Informatics*. DOI: 10.1109/SISY.2015.7325356

[55] Esfandiari, N., Babavalian, M., Moghadam, A-M., & Tabar, V. (2014). Knowledge discovery in medicine: Current issue and future trend. discovery in medicine: Current issue and future trend. *Expert Systems with Applications* 41(9):4434–4463. DOI: 10.1016/j.eswa.2014.01.011

[56] Belazzi, R., Ferazzi, F., & Sacchi L. (2011). Predictive data mining in clinical medicine: A focus on selected methods and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(5):416-430. DOI: 10.1002/widm.23

[57] Parvin, H.,, Minaei-Bidgoli, B., & Alinejad, H. (2013). A New Imbalanced Learning and Dictions Tree Method for Breast Cancer Diagnosis. *Journal of Bionanoscience* 7(6). DOI: 10.1166/jbns.2013.1162

[58] Cichosz, P., Jagodzi´nskijagodzi´nski, D., Matysiewicz, M., & Oleszkiewicz, W. (2016). Novelty Detection for Breast Cancer Image Classification. Proceedings of the Conference: *XXXVIII-th IEEE-SPIE Joint Symposium on Photonics, Web Engineering, Electronics for Astronomy and High Energy Physics ExperimentsAt*: Wilga, PolandVolume: 10031. DOI: 10.1117/12.2249183

[59] Bochare, A., Gangopadhyay, A., Yesha, Y., & Rishe, N. (2014). Integrating domain knowledge in supervised machine learning to assess the risk of breast cancer. *International Journal of Medical Engineering and Informatics* 6(2):87 – 99. DOI: 10.1504/IJMEI.2014.060245

[60] Iakovidis, D. & Smailis, C. (2012). A semantic model for multimodal data mining in healthcare information systems. *Studies in health technology and informatics*. DOI: 10.3233/978-1-61499-101-4-574

[61] Asha, T., Natarajan, S. & Murthy, K. (2012). Data Mining Techniques in the Diagnosis of Tuberculosis. In book: *Understanding Tuberculosis - Global Experiences and Innovative Approaches to the Diagnosis*. DOI: 10.5772/30504

[62] Cabral, G., & De Oliveira, A. (2014). One-class Classification for heart disease diagnosis. *Proceedings of the Conference: 2014 IEEE International Conference on Systems, Man and Cybernetics – SMC*. DOI: 10.1109/SMC.2014.6974311

[63] Masethe, H., & Masethe, M. (2014). Prediction of Heart Dieases using Classification Algorithms. *Proceedings of the World Congress on Engineering and Computer Science* 2014 Vol II. WCECS 2014, 22-24 October, 2014, San Francisco, USA. ISBN: 978-988-19253-7-4

[64] Srinivas, K., Kavihta, R., & Govardhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering* Vol. 02, No. 02, 2010, 250-255. ISSN : 0975-3397

[65] Nahar, J., Ali, Sh., Imam T., Tickle, K., Chen, P. (2016). Brain Cancer Diagnosis-Association Rule Based Computational Intelligence Approach. Proceedings of the Conference: *International Conference on Computer and Information Technology (CIT*). DOI: 10.1109/CIT.2016.106

[66] Chen, Y-Ch, Ke, W-Ch., & Chiu, H-W. (2014). Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Computers in Biology and Medicine* 48C(1):1-7. DOI: 10.1016/j.compbiomed.2014.02.006

[67] Rahman, R., & Afroz, F. (2013). Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis. *Journal of Software Engineering and Applications* 06(03):85-97. DOI: 10.4236/jsea.2013.63013

[68] Yeh, D-Y., Cheng, Ch-H., & Chen, Y-W. (2011). A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications* 38(7):8970-8977. DOI: 10.1016/j.eswa.2011.01.114

[69] Kosmicki, J., Sochat, V., Duda, M. & Wall, D. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational Psychiatry* 5(2):e514. DOI: 10.1038/tp.2015.7

[70] Zimina, E. Cluster analysis of cardiac data. (2018). *Statistics and Economics.* V. 15. No 2. DOI:10.21686/2500-3925-2018-2-30-37

[71] Mason, P., Frigo, A., Scarpa, M., Maestrelli, P., & Guarnieri, G. (2018). Cluster Analysis Of Occupational Asthma Due to Isocyanates.

*Journal of Allergy and Clinical Immunology* 142(6). DOI: 10.1016/j.jaci.2018.08.018

[72] Seino, Y., Hasegawa, T., Koya, T., Sakagami, T., Mashima, I., Shimizu, N., ..., Kikuchi, T. (2018). A Cluster Analysis of Bronchial Asthma Patients with Depressive Symptoms. *Internal Medicine* 57(14). DOI: 10.2169/internalmedicine.9073-17

[73] Jacob, Sh., & Ramani, R. (2012). Evolving Efficient Clustering and Classification Patterns in Lymphography data through Data Mining Techniques. International Journal on Soft Computing (IJSC) Vol.3, No.3, DOI: 10.5121/ijsc.2012.3309

[74] Jamous, B., Fa, R., & Nandi, A. (2015). Book. Integrative Cluster Analysis in Bioinformatics. *John Wiley*ISBN: 978-1-118-90653-8. DOI: 10.1002/9781118906545

[75] Li, Y., Chou, Ch-A., & Wang, H. (2015). A Machine Learning Based Intelligent System for Diagnosis and Treatment. Proceedings of the *Industrial and Systems Engineering Research Conference*. pp. 1646-1653

[76] Jacob, Sh., & Ramani, R. (2012) Evolving Efficient Clustering and Classification Patterns in Lymphography Data Through Data Mining Techniques. *International Journal on Soft Computing* (IJSC) Vol.3, No.3. DOI: 10.5121/ijsc.2012.3309

[77] Bilska-Wolak A., & Floyd C. (2004). Tolerance to missing data using a likelihood ratio based classifier for computer-aided classification of breast cancer. *Physics in Medicine and Biology* 49(18):4219-37. DOI: 10.1088/0031-9155/49/18/003

[78] Lin, J-H., & Haug, P. (2008). Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. *Journal of Biomedical Informatics* 41(1):1-14. DOI: 10.1016/j.jbi.2007.06.001

[79] Androuchko, V., & Kelly, C.(2006). Intelligent medical diagnostic system. HEALTHCOM 2006 8th *International Conference on e-Health Networking, Applications and Services*. DOI: 10.1109/HEALTH.2006.246448

[80] Belginova, S., Uvaliyeva, I., & Ismukhamedova, A. (2018). Decision support system for diagnosing anemia. In 2018 4th *International Conference on Computer and Technology Applications* (ICCTA) (pp. 211-215). IEEE.. doi: 10.1109/CATA.2018.8398684

[81] Sivasankar, E., & Rajesh, R., Knowledge discovery in medical datasets using a Fuzzy Logic rule based classifier. *International Conference on Electronic Computer Technology (ICECT)*. - Kuala Lumpur, 2010. - P. 208-213.

[82] Poonja Z., Brisebois A., Van Zanten, S.. Patients With Cirrhosis and Denied Liver Transplants Rarely Receive Adequate Palliative Care or Appropriate Management. *Clinical Gastroenterology and Hepatology*, Vol. 12, Issue 4, April 2014. – P. 692–698.

[83] Zhang, Zh., & Wang, F-Sh.. Stem cell therapies for liver failure and cirrhosis. *Journal of Hepatology*, Vol. 59, Issue 1, July 2013. – P. 183–185.

[84] Uvalieva, I., Garifullina, Z., Utegenova, A., Toibayeva, S., & Issin, B. (2015). Development of intelligent system to support management decision-making in education. In 2015 6th *International Conference on Modeling, Simulation, and Applied Optimization* (ICMSAO) (pp. 1-7). IEEE. doi: 10.1109/ICMSAO.2015.7152249.

[85] Shalev-Shwartz S, Ben-David S. Understanding Machine Learning: From Theory to Algorithms. New York: Cambridge University Press; 2014. 410 p. DOI: 10.1017/CBO9781107298019

[86] Indira, U., Belginova, S., & Ismukhamedova, A. (2018). Informational and Analytical System to Diagnose Anemia. *In Proceedings of the Fourth International Conference on Engineering & MIS 2018* (p. 18). ACM. https://doi.org/10.1145/3234698.3234716

[87] Cios, K., Moore, G. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine* 26(1-2):1-24. DOI: 10.1016/S0933-3657(02)00049-0

[88] Kalimoldayev M., Belginova S., Uvaliyeva I., Ismukhamedova A. (in press) IT infrastructure of e-Health of the Republic of Kazakhstan. In: Shokin Y., Shaimardanov Zh., Zhumagulov B. (eds) Mathematical Modeling of Technological Processes. CITech 2018. *Communications in Computer and Information Science* (unpublished)

[89] Gray, G. Challenges of building clinical data analysis solutions. *Journal of Critical Care* 19(4):264-70 DOI: 10.1016/j.jcrc.2004.08.008