## *АВТОМАТИЗАЦІЯ ТА ІНТЕЛЕКТУАЛІЗАЦІЯ ПРИЛАДОБУДУВАННЯ*

DOI: 10.20535/1970.67(1).2024.306735
UDC 616.6:004.67

# PROGRAM PROCEDURES FOR TRAINING A RECOGNITION SYSTEM FOR THE DIFFERENTIAL DIAGNOSIS OF PATIENTS BASED ON HETEROGENEOUS SYMPTOM COMPLEXES

*Shulyak O. P., Druzhynin V. V.*
*National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine*
*E-mail: shulyak.alex.47@gmail.com, linagoracaa@gmail.com*

*The machine learning recognition system for the differential diagnosis of patients based on heterogeneous nephrology parameter complexes is being considered, transitioning from instrumental means of examination. Training utilizes empirical statistics of clinical cases in a database with reliable diagnoses. The purpose is to expand the capabilities of information extraction from similar databases for training recognition procedures by enriching this toolkit with new features containing characteristic aspects of the extracted information.*

*The research object is the mathematical and software toolkit for training recognition procedures of patient differential diagnosis based on statistics of reliably diagnosed clinical cases. The subject of the study is the software procedures for forming models of parameter complex incidence during training along scales of their values and the procedures for using these models in diagnostics. Model acquisition is perceived as the main content of the training process in ensuring diagnosis differentiation. A criterion for accepting preferential diagnostic decisions using such models is proposed. To simplify the development of mathematical and software procedures, heterogeneous symptom complexes are normalized and transformed to the [0; 1] scale.*

*The introduction states the significant prevalence in medicine and related fields of databases with medical and biomedical data statistics on parameters and characteristics of human organs and systems in different conditions, their medical interpretation, and their use for various purposes, often associated with patient diagnostics. The problems of their formation and use are outlined on real databases, with one complicating factor in the development of diagnostic hardware-software being the substantial heterogeneity of parameters determined by patient examination instruments.*

***Keywords:*** *patient diagnosis; heterogeneous symptom complexes; parameter normalization; parameter distribution models; decision accumulation criterion.*

**Introduction**

In both theory and practice of medicine and related fields, the prevalence of various open and closed-access databases of medical and biomedical data [1 – 3], diverse in their medical specialization and purpose, has become rooted and continues to progress. The extraction and utilization of information [1 – 3, 4] accumulated in such databases for various purposes, ranging from its study in professional training of specialists [1, 3, 5, 6] to its application in addressing various practical tasks in the field of medicine and related areas [1 – 4, 6 – 15], are gaining increasing relevance and importance. There remains a demand for the development of various software and hardware tools for obtaining necessary information from such databases in different sectors of subject area specialization [1, 3, 5, 16 – 18], including the demand for the development of simple specialized modules in software and hardware implementations [1, 3, 19 – 21].

Each type of toolkit for extracting necessary information from such databases and the corresponding tools that use it to address their issues have their own characteristics, their own emphases, and their effectiveness in extracting and using their available information [1 – 3, 9, 15, 21 – 26, 28] contained in the existing data, as well as their peculiarities in implementing components of accumulated empirical observation experience of objects, processes, and phenomena [1 – 4, 7, 22, 23, 26, 28] are of interest. Perhaps, there is no universal toolkit for such purposes, and each new development can be seen as obtaining data processing tools that complement the existing toolkit and may demonstrate sufficiently high effectiveness in their use, which needs to be verified for its effectiveness [1, 2, 5, 6, 21, 24, 26, 28], and in this sense, the relevance of such research and developments persists.

One of the obstacles to the development of the mentioned simple specialized software and hardware data processing toolkit in the subject area under consideration is the heterogeneity of parameter complexes [1 – 6, 9, 10, 18, 22, 29, 30] collected in databases with descriptions of clinical cases. This complication can be overcome by simple uniform linear data transformations [41, 42] considered in the work.

One of the main reasons for the heterogeneity of the mentioned databases is that they often represent collections of descriptions of clinical cases from medical practice with the results of patient instrumental examinations [1 – 3, 5 – 7, 12, 15] or the results of purposeful statistical studies [1 – 3, 10] related to the analysis of the impact and consequences of professional, climatic, and other conditions on human life processes [1, 2], the analysis of the dynamics of processes and phenomena in the body, the disclosure of relationships between the past, present, and future states of organs and systems at different levels in the body [1, 2, 4, 9, 10, 12 – 14, 19, 22], the identification of influencing factors [1, 7], risk factors [1, 22], chances of favorable outcomes [1, 18], as well as the determination of characteristic regional features [1, 9] related to population health provision, which explains the heterogeneity of parameter complexes in databases.

Such databases contain real factual material of various physical nature, different levels of accuracy and reliability [1, 5, 6, 15, 28]. It is obtained empirically, including the use of software and hardware complexes of various, including medical, purposes and complexities, using unique and widely used means of patient examination, means of studying metabolic processes and products of human life activity, reactions to various influences, as well as tools for studying food products, water, determining environmental parameters, properties of biomedical materials [31 – 38]. Data may be collected during patient observation in the process of their dispensary examination, prevention and treatment, medical examinations, professional selection, surveys, categorization of the examined population by gender, age, working conditions, lifestyle, by risk groups and health level groups, by other characteristics as part of their comprehensive characterization [31 – 38]. This increases the diversity and heterogeneity of information in the obtained similar numerous parameter complexes in databases, the most valuable of which are annotated.

This data heterogeneity about conditions, objects, processes, and phenomena in the subject area, especially in identifying cause-and-effect relationships in the occurrence and development of diseases, is a complicating factor in the study of such data [1 – 3], in the practical application of collected statistics in analysis [1 – 3], and the interpretation of specific clinical cases [1 – 3], in the choice of treatment strategies, tactics, and means for patients [1 – 3], in assessing the effectiveness of treatment in the dynamics of its conduct [1 – 3]. The heterogeneity of parameters and characteristics in databases, the use of quantitative and nominal indicators [1, 5, 6, 39], which differ in their physical nature, ranges of values, characteristics of measurement tools for their obtaining, and other properties, also complicates the development of software and hardware tools for processing collected data, including the development of specialized tools [1 – 3], as well as the use of statistical processing results in the interpretation of collected factual material [1, 5, 6, 31 – 38, 39], and the use of its results in further decision-making support [1, 5, 6, 31 – 39].

The greatest inconvenience, at first glance, may be caused by the heterogeneity of quantitative parameters, the range of values of which may be characterized by several orders of numbers [1, 5, 6, 31 – 39], expressing their value. The need for a joint consideration of data complexes of different in essence nominal and quantitative parameters [1, 5, 6, 39] also poses certain difficulties. However, as previous studies have shown, these difficulties are easy to overcome. Demonstrating this was one of the tasks of this work.

The main focus of this work is dedicated to the heterogeneity of quantitative parameters, for which a database in the field of nephrology [5, 6] was examined in its application to addressing patient diagnostic issues [5, 6]. The question of reconciling nominal parameters, different in sets of their possible values, as well as solving a similar issue for combinations of quantitative and nominal parameters in symptom complexes and other complexes of similar data, can be detailed in a separate study.

The relationship between quantitative and nominal indicators in real databases, as well as the number of interpretation variants of human body conditions, as research has shown, can be extremely diverse and even practically polar, when databases provide statistics for parameters of only one type.

In analyzed open databases, the number of possible interpretation variants of human body conditions turned out to be as follows: [31], [32], [33], [34], [35], [36], [37], [38], where the notation $q$ and $n$ are used for quantitative and nominal parameters, respectively. Initially, the number of interpreted states by the database is indicated here, and in parentheses, the quantitative composition of the indicators whose values characterize these states is disclosed.

The purposes of these databases also vary significantly. The first database provides a dataset at the cellular level on breast cancer, dividing the neoplasm into malignant and benign [31]. The second database contains samples of parameter values in complexes used to predict stages of liver cirrhosis [32]. The third database is intended for informational support in addressing the question of whether a patient with symptoms should be classified into a high-risk group for severe Covid-19 or not. The issue of building an appropriate machine learning model was considered here [33].

The fourth database is built on the classification of fetal health as normal, suspicious, or pathological, with statistics of values of used indicators in each case. The purpose of using the database is diagnostic, providing timely diagnosis in medical practice to avoid child and maternal mortality [34]. The fifth database contains a dataset for the analysis and prediction of heart attacks, considering four types and the distribution of clinical cases into two groups where the occurrence of an attack is present or absent [35]. The sixth database contains sets of 11 clinical signs for predicting heart failure. The interpretation of these sets is considered in two variants [36].

The seventh database provides a dataset of lung cancer patients, who were divided into groups living in areas with high and low levels of air pollution, to identify a range of risk factors for the disease [37]. The eighth database offers a set of indicators for predicting the probability of a stroke in a patient based on such input parameters as gender, age, presence of various diseases, type of occupation, smoking status, and others. High and low stroke risk groups are considered based on statistics of confirmed stroke cases in some patients [38].

The goals of creating databases and the composition of symptom complexes used imply that they are fundamentally diagnostic-oriented and closely aligned in design and usage to address issues of differential diagnosis [31-38], ultimately serving as informational support during the training of diagnostic decision-making procedures [31-38], which is the main focus of this work.

The object of the research and development here are the mathematical and programmatic procedures for training recognition procedures of differential diagnosis systems for patients, which are built on the use of databases with statistics of descriptions of nephrology clinical cases with diverse symptom complexes consisting of eight quantitative parameters and with three diagnoses considered in the database [5, 6].

The subject of the research is the development and use of models for encountering various parameter values for different diagnoses [1, 39] within ranges of their values when forming diagnostic decisions using recognition procedures of differential diagnosis for patients.

The models are formed according to the statistics of symptom complexes in the database. The procedures for making diagnostic decisions are built according to the method of accumulation [1, 5, 21], the decision criterion of which is oriented towards the use of the proposed probability density distribution models along the scales of symptom complex parameters, which differ from histograms [1, 5, 6, 39] and differ in content. The software implementation of the proposed procedures is done in Python [40]. The ability of the models discussed in the work to be used in training recognition decision procedures was tested with an assessment of the sensitivity, specificity, and overall validity of diagnostic decisions [1, 2, 5, 6] based on the statistics of the database used for training.

## I. Key decision of the work and tasks set for the development of software procedures for training a recognition system

The development of software procedures for training is considered in the work concerning the recognition system as part of the diagnostic system for patients based on the results of their examination in the differential diagnosis phase. The situation is examined in which a small list of possible diagnoses is compiled based on patient data and their health status, one of which must be selected using recognition software procedures and available diagnostic experience in similar clinical cases [5, 6].

The experience of such diagnostics in the context of this task is concentrated in a specialized database focused on similar clinical cases [5, 6]. Each such case contains a verified diagnosis in its description, and their descriptions form homogeneous specialized symptom complexes [5, 6]. One part consists of categorical descriptions of patients [5, 6], their overall health status, and individual organs and systems. The second part provides quantitative indicators [5, 6] obtained during patients' instrumental examinations.

Medical practice provides many examples of using diverse symptom complexes characterized by different nature and scales of indicators within specialized health issues [5, 6]. This poses a significant complicating factor in the development and application of recognition procedures in differential diagnosis and in organizing their training. When dealing with different scales of input parameters, specialized multi-input computational devices for forming predominant diagnostic decisions become more complex compared to devices with uniform input scales. Recognition of diagnoses through software procedures becomes more intricate.

The work assumes that the patient's condition undergoing differential diagnosis is characterized by a symptom complex of the same type as in the database, and the recognition system, pre-trained on the database statistics, should propose a diagnosis predominant based on the accumulated experience in the database. The key decision of this work, around which the circle of research and development tasks (Fig. 1) was formed, is the utilization in diagnosing patients of differences in the occurrence of parameter values of symptom complexes on their scales for different diagnoses.

**The main decision of this work is decisive for the formation of a range of tasks for research and development:**

- diagnostics of patients is based on the use of differences in the occurrence of parameter values from the composition of their symptom complexes for various diagnoses;
- this occurrence is characterized by models of distribution of probability densities of such values along operating scales of parameters;
- models are formed on the basis of database statistics with symptom complexes of patients whose diagnoses are verified and reliable;
- the decision regarding the diagnosis of a specific patient can be built on the complex usage of all such models for all parameters and all possible diagnoses.

Figure 1. The main decision in forming the range of tasks for the formation of a range of tasks for research and development in this work

This occurrence is characterized in the work by models of probability density distributions of such values along the working scales of parameters. The models are formed based on the statistics of patient symptom complexes with similar medical issues collected in the database. The description of each clinical case in the database used is provided with a reliable diagnosis [5, 6].

Such a key decision in this work naturally raises and makes one of the central (Fig. 2) questions about the capability, or in other words, the productivity of using sets of probability density distribution models for parameter values of symptom complexes along their working scales to build software procedures for differential diagnosis and their training. In other words, it is a question of the effectiveness of generalizing and using practical diagnostic experience of patients in the form of such models. This led to the need for conducting special research of this nature in the work.

**The main question about the key decision of work:**

question about the ability, efficiency of using complexes of models of probability density distributions for parameter values of symptom complexes along the working range for the construction of software procedures of differential diagnosis and their training;

question about the effectiveness of the generalization and usage in the form of such models of the practical experience of diagnosing patients.

Figure 2. The main question about the key decision of work

In this regard, during the research and development process, the quality of differential diagnosis decisions for patients proposed by recognition software procedures should be verified and confirmed [1, 5, 6]. Consequently, it was envisaged that preliminary tests would be conducted during the research to assess the accuracy of the decisions proposed by the developed software procedures, particularly using the available database statistics [5, 6]. Furthermore, in the process of delineating the tasks outlined in this work, in the development, investigation, and discussion of issues related to the use of recognition software procedures in systems for the differential diagnosis of patients, as well as procedures for their training and quality assessment, such key concepts and corresponding informational objects were used by the software procedures (Fig. 3).

The list of possible diagnoses for differential diagnosis is mandatory here. The range of possible diagnoses should be oriented towards similar patient health issues and be fully and unambiguously defined. It is assumed that the analyzed clinical case should be differentiated based on this list. The list will be used both in dividing clinical cases into groups to obtain characteristics of symptom complexes useful in the training of recognition procedures and in the process of selecting preferred diagnostic decisions based on specific symptom complexes. The medical issues of patients' health and the database discussed in the work are chosen purely to illustrate the content of the research and development, the results of which can be applied to the differential diagnosis of other medical profiles.

The database is considered and used here as actual prior material for machine learning of recognition procedures for differential diagnosis [1, 2, 5, 6, 41, 42], as well as for evaluating the quality of this training during tests of trained decision-making procedures considering the presence of correct diagnoses in it [1, 5, 6].

**Basic concepts and informational objects in the development and training of recognition software procedures**

1. List of possible diagnoses of differential diagnostics of patients with certain problems in their health.
2. A database specialized on similar health problems of patients, which contains their symptom complexes and reliable diagnoses.
3. Similar symptom complexes of patients from the database with a set of categorical and quantitative indicators to characterize their health.
4. Symptom complexes of new admitted patients with related medical problems for their differential diagnosis.
5. List of indicators for evaluating the quality of solutions offered by recognition procedures in the differential diagnosis of patients.
6. Panoramas of the predominant diagnoses on the parameter operating scales of symptom complexes.
7. Constituent criteria of the formation of predominant diagnostic decisions regarding sets of parameters of patient symptom complexes.
8. Diagrams of parameters ranked by diagnostic value symptom complexes of specific patients.

Figure 3. Basic concepts and information objects in the development and training of software recognition procedures in systems of differential diagnosis of patients

Each patient's symptom complex in the database or the symptom complex of a patient undergoing diagnosis is considered as a whole, an indivisible combination of parameter values that together characterize their health issues [1, 5, 6]. The same values are expected to be jointly used in the criteria for accepting preferred diagnostic decisions proposed by the software procedure.

New patient symptom complexes undergoing diagnosis are considered as initial information objects that must be transformed by recognition procedures for differential diagnosis into appropriate diagnoses based on the experience accumulated in the database.

As indicators for assessing the quality of decisions proposed by recognition procedures trained on the available database material, it is advisable to include indicators of sensitivity, specificity, and overall validity of decisions [1, 5, 6]. Along with them, other additional indicators may be considered, the appearance and transparent content of which are related to the logic of data transformation in this work.

Overall, the tasks set in this work for the development of software procedures for training a recognition system have been divided into seven main blocks according to the logic of development, research, and use of recognition procedures in the software of specialized systems for the differential diagnosis of patients, the logic of their training, and the need to assess its quality (Fig. 4).

The block of developing software procedures for data preprocessing should ensure the formation of procedures to standardize heterogeneous symptom complexes from the database into a format convenient for generating diagnostic decisions and training software procedures used for diagnosing patients. The main focus of the block on developing procedures for forming diagnostic decisions is the development of software tools for generating diagnostic decisions based on patient symptom complexes. The block on training recognition procedures ensures the development of software tools for generating and using data necessary for the operation of procedures for forming diagnostic decisions based on patient symptom complexes. Supplementary blocks associated with the development of necessary software tools are also anticipated.

The block for developing software procedures for processing patient data in their differential diagnosis involves developing procedures for processing heterogeneous patient symptom complexes to select the most probable diagnoses based on the available statistics in the database. The block for developing software procedures to investigate the effectiveness of training is designated for developing software tools for assessing the quality of decisions proposed by trained recognition procedures.

The blocks at each level constitute the content of relatively independent stages of development and research. Implementing such stages sequentially, as depicted in the figure, is reasonable. The minimum subset of blocks 1 – 3, 6, 7 is necessary to complete the development cycle. The rest can be realized in further research. The software procedures planned for development in each such block will typically be components, and their content will be elaborated sequentially and in detail later. Nonetheless, at this stage of development, based on the purpose and general content of the designated blocks of tasks, it can be concluded that the main content of training recognition software procedures in decision-making systems for specialized differential diagnosis of patients based on heterogeneous symptom complexes of their clinical cases may involve the following actions (Fig. 5).

It is easy to see that the central task of the developments and research in this case is the coordinated development of two software procedures (Fig. 6). The first procedure is intended to standardize the presentation of symptom-complex parameters in working windows for processing, while the second is for obtaining models that underlie diagnostic decision-making.

The resolution of the stated tasks of developing software procedures in the listed blocks involves the development of corresponding mathematical data processing procedures discussed in the work and their software implementation, primarily oriented towards using the Python programming language toolkit. The general content of such developments and research is further elaborated below.

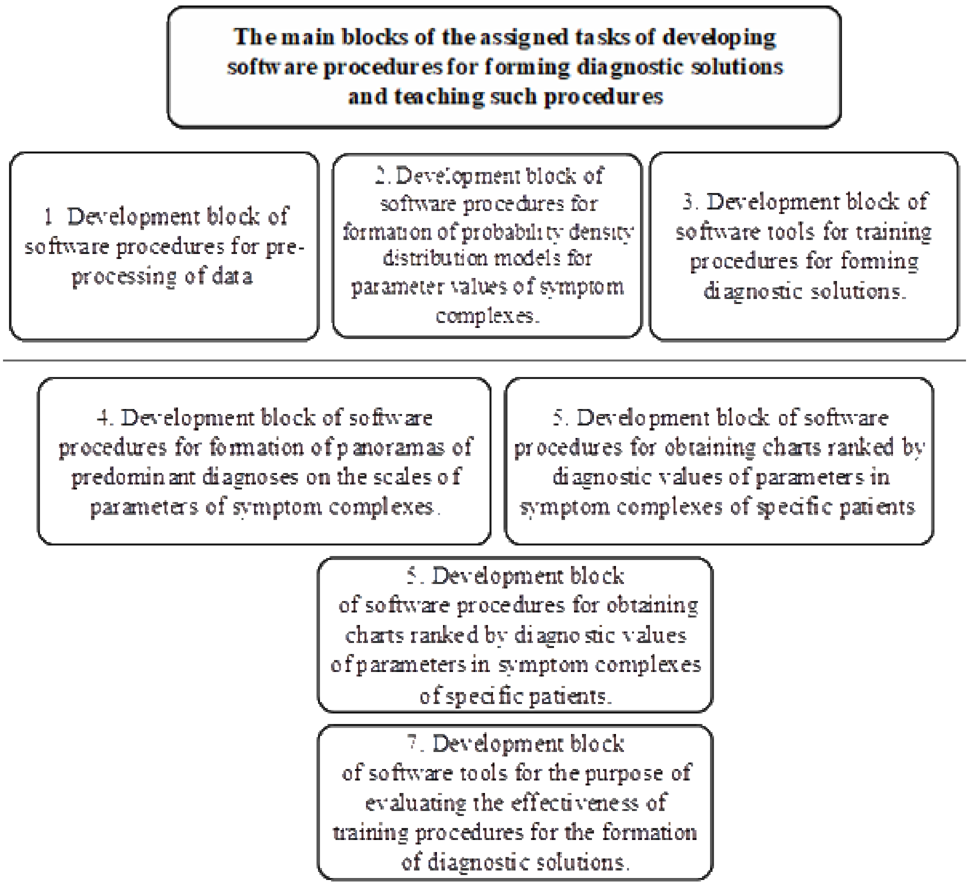


Figure 4. The main blocks of the assigned tasks of developing software procedures for forming diagnostic solutions and teaching such procedures

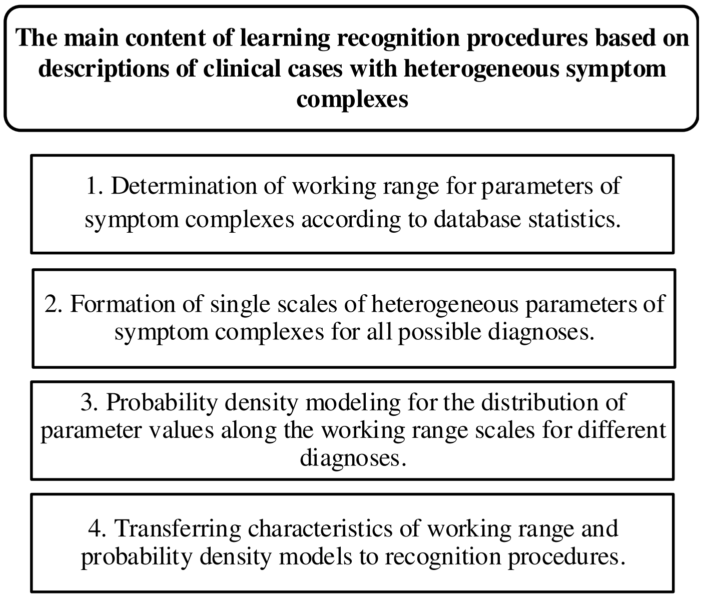


Figure 5. The main content of learning recognition procedures based on descriptions of clinical cases with heterogeneous symptom complexes

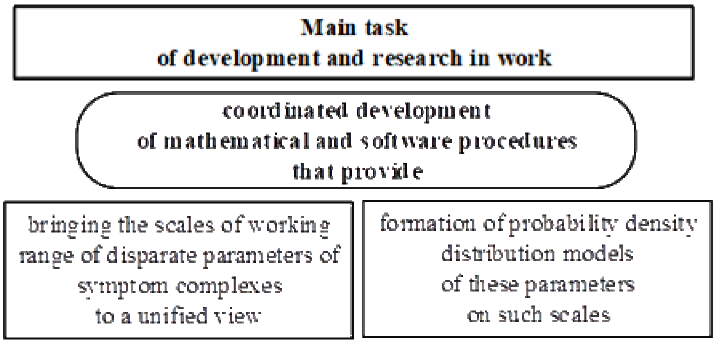


Figure 6. Main tasks of development and research in work

**II. Formalization of the task of developing software procedures for training a recognition system and the general content of mathematical data processing procedures in its resolution**

Adhering to the general logic of developments and research outlined by the list of task blocks for the development of software procedures, we will introduce necessary notations to formalize the formulation of these tasks and elucidate the content of mathematical operations on the data that underlie their resolution.

Pre-processing of symptom complexes in the database. Its essence lies in the general standardization of various parameters of symptom complexes to unified scales, on which the available statistics of the database will be considered and probability density distribution models of different parameter values, used in the criteria for accepting diagnostic decisions, will be constructed. At the beginning of the pre-processing of symptom complexes (SC) in the database, they are sorted by diagnoses into three groups in their initial form [5, 6] and brought into a working format (Table 1). The format of all SCs is the same.

Initial overview of the used database shows that the quantitative parameters of its symptom complexes include: Age – the total number of years lived by the patient at the time of examination; kidney parameters: Length, Width, Thickness, and Thickpair – the length, width, thickness of the patient's kidney, and the thickness of its parenchymal layer, [mm]; dimensionless parameter Index, as well as parameters of blood flow through the kidney: speed Speed, [cm/s] and acceleration Speedup, [cm/s$^2$]. There are two nominal parameters in the symptom complex: Sex – gender and L, R – an indication of whether the kidney is left or right, which was considered in each clinical case. Therefore, the parameters here are diverse in nature.

At the same time, there are also different numerical values in the database not only for different parameters but also for the same parameters for different differentiated diagnoses (healthy), (polycystic), and (hydronephrosis): respectively, Age (21 – 74, 37 – 74, 21 – 68), Length (82 – 135, 98 – 129, 99 – 144); Width (39 – 74, 44 – 89, 41 - 85); Thickness (36 – 72, 37 – 88, 36 – 67); Thickpair (1 – 27, 2 – 11, 6 – 27); Speed (13.3 – 46.6, 2.3 – 43.3, 14.5 – 41.5); Index (0.56 – 0.70, 0.52 – 0.74, 0.69 – 0.80); Speedup (51 – 685, 88 – 656, 98 – 545) – with the provided dimensions. The lengths of the intervals (widths of the ranges) of parameter values are different.

Table 1. Working data format

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Numbers in the groups of each diagnosis $i_a$, $i_b$ or $i_c$ | Diagnosis: healthy, polycystic or hydronephrosis: *A, B* or *C* | Sex | L, R | Length | Width | Thickness | Thiekpar | Speed | Index | Speedup | Age |
| | | | | Quantitative parameters | | | | | | | |

The working format in which the data are presented differs from the original only by some rearrangements of indicators of clinical cases in symptom complexes, but this creates certain conveniences in the development of computational procedures.

In this format, after the patient's clinical case number in his group by diagnosis (which is also his symptom complex number), his reliable diagnosis is indicated, the values of nominal parameters (there are two of them here) are mentioned, and then the values of quantitative indicators are provided. There are eight of them, and they are listed in the table.

Patient cases are further labeled with an index $i$. To this index, an additional auxiliary index $a$, $b$ or $c$ is added according to the patient's diagnosis, which may be established by the task of differential diagnosis. Such a composite index indicates the number of the symptom complex group where the diagnoses are the same. The patient case numbers in the groups are as follows: $i_a$, $i_b$, $i_c$ – symptom complex numbers in groups with corresponding diagnoses, where $i_a = 1, 2, 3, ..., n_a$; $i_b = 1, 2, 3, ..., n_b$; $i_c = 1, 2, 3, ..., n_c$; $n_a$, $n_b$, $n_c$ – the number of symptom complexes with diagnoses $A$, $B$ and $C$.

For the numbering of parameters within the symptom complexes themselves, index $j$ is used; here $j = 1, 2, 3, ..., 12$ – position numbers in the working format, listed in the first row of the table.

Scales of various parameters, which need to be standardized, are determined based on the available statistics of these parameters' values in the symptom complexes of the database. Standardization is done step by step. In this process, any scales for parameters of the same type are chosen uniformly for all diagnoses. Consequently, for any parameter value in the symptom complex, based on the comparison results by the recognition procedure of combined probabilities on a single scale for different diagnoses, a clear preference can be given to one of them based on the highest value of the specified probability.

The parameter values of the symptom complexes in their original form after rearrangement according to the specified format are denoted as elements $x_{ij}$ of the array in the form $x_{i_a j}$, $x_{i_b j}$, $x_{i_c j}$ in groups according to the listed diagnoses.

The boundaries of the working scales (minimum and maximum values $xmi_j$, $xma_j$ of parameters) are determined based on the known statistics of the database, which is considered as a whole regardless of diagnoses. The parameter values on these scales have their natural dimensions.

Further, the parameters are considered within working windows, which are bounded by the limits of the working scales of these parameters based on the available statistics. Therefore, all available statistics of the database fit within these limits, revealing the actual location of the points of the database statistics, and necessary probability distribution models can be constructed for the values of parameters for different diagnoses.

In the initial working windows, which are intermediate, parameter values are denoted and supplemented with additional indices according to diagnoses: $y_{i_a j}$, $y_{i_b j}$, $y_{i_c j}$. The recalibration of parameters in these working windows is performed according to the formula: $y_{ij} = x_{ij} - xmij$. After this, the scales acquire the boundaries of 0 and $\left(xma_j - xmi_j\right)$. The dimensions of the parameters are preserved for now.

Next, transformed working windows are used. They are the previous working windows normalized by the sizes of their scales. The parameter values in them are denoted as $z_{ij}$ with variations $z_{i_a j}$, $z_{i_b j}$, $z_{i_c j}$ according to diagnoses, as in the previous case. The recalibration of parameters in such working windows is carried out according to the formula:

$$z_{ij} = \left(x_{ij} - xmi_j\right) / \left(xma_j - xmi_j\right).$$

Parameters of varying dimensions and ranges of values in the symptom complex are transformed into homogeneous normalized parameters with scales [0;1] within the same working windows. Parameters lose their former dimensions and become dimensionless. The scales of different symptom complex parameters become standardized and have boundaries from 0 to 1 inclusive.

Such parameter transformation is linear. If we consider the key relationship $y = kx + b$, then the provided formula for transformation into a normalized working window takes the following form:

$$z_{ij} = \left[1 \middle/ \left(xma_j - xmi_j\right)\right] * x_{ij} + \left[-xmi_j \middle/ \left(xma_j - xmij\right)\right]$$

what is needed next is to preserve the mutual positions of the database statistics points entirely. The initial location of all points in a certain scale is transferred to such standardized windows.

It is precisely in these windows that statistics of the distribution of parameter values along the specified scales are separately considered for different diagnoses, and necessary probability density models are formed for decision-making.

**Formation of piecewise and smooth models of probability density distributions.** Forming such models for the density of probabilities of different parameter values from the database is done separately by types of parameters. Initially, the values of each parameter encountered in the database are plotted as points on the scales of their working windows. The distribution of points along such scales gives an idea of the probability densities, the models of which are used in making diagnostic decisions.

Placing these points on the scales is done in the usual order of viewing records in the database. After this, the points are renumbered. The numbers $i_a = 1, 2, 3, ..., n_a$; $i_b = 1, 2, 3, ..., n_b$; $i_c = 1, 2, 3, ..., n_c$ in the list of symptom complexes in groups with the same diagnoses are replaced by the numbers of the same points $k_a = 1, 2, 3, ..., n_a$; $k_b = 1, 2, 3, ..., n_b$; $k_c = 1, 2, 3, ..., n_c$ in their hierarchy along the scales in the order of increasing parameter values. The number of points in the groups is preserved, only their order changes. The multiplicity (possible repetitions) of statistics points is then separately considered. Based on the position of these points and their new numbering, a reference stepwise model of the distribution density of probabilities of different values of each parameter along its scale is formed.

The graphical representation of such models is characterized in Figure 7.

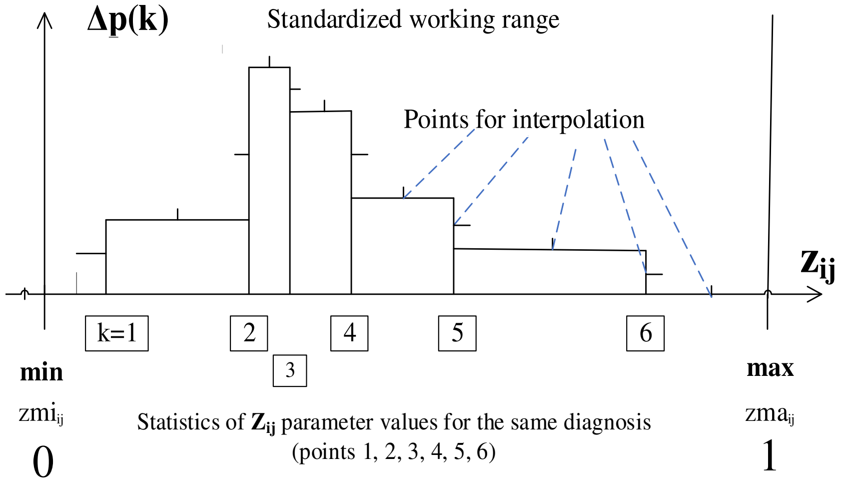


Figure 7. Step model of probability density distribution by parameter values for a specific diagnosis

During modeling, all parameter increments in the available window statistics are considered equal in terms of the share they contribute to the resulting probability. The probability shares $\Delta p_a$, $\Delta p_b$, $\Delta p_c$, contributed by parameter increments for different diagnoses depend on the number of symptom complexes in each group. Given that the total value of such increments under normalization condition equals one [1, 42, 43], it can be expressed as $\Delta p_a = 1/n_a$, $\Delta p_b = 1/n_b$, $\Delta p_c = 1/n_c$.

The increments along the parameter scale in the working window, as shown in the figure, are unevenly distributed. The less frequently the increments occur, the lower the probability density values modeled in this interval. In the reference model under consideration, the probability density is considered constant between adjacent increments. Its magnitude in the first step is dependent on the length of the interval where it is modeled. It is calculated as the reciprocal of this length.

For the $k$ point on the graph, it is considered $\Delta p_1(k) = 1/(z_k - z_{k-1})$. If $k = 1$, then $z(\text{k}-1) = 0$ is taken. On the last interval, up to the right boundary of the window, the value $\Delta p_1(k) = 0$ is selected. If the first interval starts at the window boundary, model formation begins from $k = 2$; $\Delta p_1(2) = 2/(z_2 - z_1)$; and this is taken into account during the normalization of the initial approximation to the probability density model under consideration. This transformation option simplifies the conceptual side of building such models in the first step.

Under the normalization condition for probability density [1, 42, 43], it must be $\sum\limits_k \Delta p_k(\text{k}) = 1$ for any of the diagnoses, and therefore, the relationships $\Delta p_1(k)$ must hold for $\gamma \sum\limits_k \Delta p_1(\text{k}) = \gamma \sum\limits_k \left[1/\left(z_k - z_{k-1}\right)\right] = 1$. Hence, $\gamma = 1/\sum\limits_k \left[1/\left(z_k - z_{k-1}\right)\right]$ and $\Delta p\ (k) = \Delta p_1(k)/\gamma$.

Therefore, for the calculation of $\Delta p\left(k\right)$ and the model density values, $\Delta p_1\left(k\right)$ are initially calculated, then $\gamma$, and then recalculated $\Delta p_1\left(k\right)$ to $\Delta p\left(k\right)$. In this case, the normalization condition is ensured. The magnitude $\gamma$ is a constant needed to maintain the relationships between all $\Delta p_1\left(k\right)$ calculated in the specified manner. Such calculations are performed separately in the groups of symptom complexes for each diagnosis; the coefficient acquires additional indices: $\gamma_a$, $\gamma_b$, $\gamma_c$.

This normalization is important to make the results of the analysis of the frequency of values of all quantitative parameters within the SC composition comparable. Then, for each parameter, obtaining density probability estimates is indeed ensured, creating the possibility of comparing them with each other for different diagnoses and identifying diagnostic advantages for each parameter. These density probabilities and identified advantages are easily taken into account together in complex, component criteria for forming diagnoses based on the totality of values of all parameters in the composition of a specific symptom complex for the clinical case under consideration.

By their form and content, the models of probability density functions used are similar to ordinary histograms of probability distributions over intervals on parameter scales. The separators of intervals in the models under consideration are not points of a uniform grid within the ranges of interest, but rather points of actual parameter values in the available statistics of the database, which is the fundamental difference between these models and histograms.

The discussed reference models can have independent significance in recognition procedures and can be used in forming diagnostic decisions. In this case, to reveal the uncertainty at the breakpoints of intervals, average values between adjacent probability levels were taken. An important feature of such models, arising from the order of their construction, is that within the working ranges of parameter values for any diagnosis, there are no intervals with zero values of probability densities, which could lead to diagnostic uncertainty.

On the contrary, the combination of probability density models for all possible differentiated diagnoses in a common working window does not lead to the situation where all densities simultaneously have zero values at certain parameter values in this window. At least for one of the diagnoses, the probability density is not zero, which allows determining the diagnosis with the dominant value of this density for any parameter value on the scale of the common working window and reaching a clear predominant decision when choosing it.

On the other hand, the discussed step models can be used as a basis for their transformation into a smoother form with the calculation that in such a modified form they will be used in diagnostics similarly to step models. Such modifications can be obtained by applying known interpolation and approximation procedures [1, 42, 43] to the reference step models. The use of interpolation procedures [1, 42, 43] was considered in the work.

All points of the reference model, which should be covered by interpolation in the parameter value window, are indicated in the illustrative example in the figure with dashes. These are points at the discontinuities, the positions of which were discussed, and points at the midpoints between these discontinuities. The peculiarity of choosing and applying standard interpolation procedures to obtain smooth models lies in the fact that the graphs of functions obtained with their help cannot go into the area of negative values, since these functions are models of probability density.

Such models create the possibility of developing and using coordinated criteria and procedures for forming diagnostic decisions in the differential diagnosis of patients.

Criterion and procedure for making diagnostic decisions on probability density models. Having a system of smooth models of probability density distribution of all parameters of symptom complexes for different diagnoses, it is possible to select a predominant diagnosis for each value separately, based on the diagnosis for which the probability density is maximum. In principle, for making diagnostic decisions, methods [1, 5, 21] of accumulation and voting can be implemented (Fig. 8).

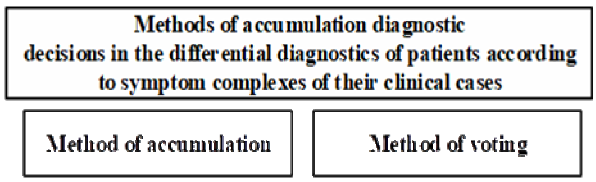


Figure 8. Methods of accumulation diagnostic decisions in the differential diagnostics of patients according to symptom complexes of their clinical cases

The voting method is not considered in this work. The content of the decision-making procedure on probability density models by accumulation method is depicted in Figure 9.

Among the diagnoses, the one for which the sum of probability density model values for all parameters within the analyzed symptom complex of the patient for different diagnoses is the highest is selected.

For the implementation of this method discussed in the work, the model values of probability densities $p_{ja}(z_{ij})$, $p_{jb}(z_{ij})$, $p_{jc}(z_{ij})$ for parameter values they have within the patient's analyzed symptom complex for different diagnoses are summed up after being transformed into the format of normalized dimensionless homogeneous indicators in normalized working windows on identical scales $\left[0;1\right]$:

$r_a = \sum_j p_{jb}(z_{ij})$, $r_b = \sum_j p_{jb}(z_{ij})$, $r_c = \sum_j p_{jb}(z_{ij})$.

Among these values, the maximum sum indicating the diagnosis is selected: $Ds = \arg\max_{a,b,c}\left(r_a\ ,\ r_b\ ,\ r_c\right)$.

The software implementation is subsequently considered for the main mathematical data processing procedures developed in the work.

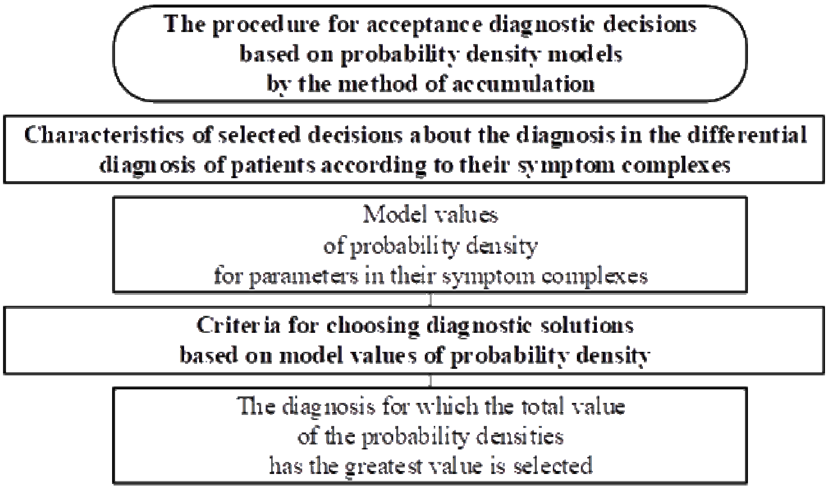


Figure 9. Content of the accumulation procedure in acceptance diagnostic decisions

**III. Software Implementation of Recognition Mathematical Procedure and Training Procedure for Use in Patient Differential Diagnosis Systems**

The construction of software procedures is considered for mathematical procedures transforming clinical case symptom complexes into normalized indicator complexes within working windows, for forming smooth distribution models of probability density along normalized parameter scales in these windows for different diagnoses, and for implementing the criterion for forming predominant diagnostic decisions based on patient symptom complexes using accumulation method.

The software implementation of these procedures is primarily discussed in the order in which their mathematical content was revealed in the previous section. The developed software toolkit is characterized by flowcharts and necessary descriptions. The terminology and notation from previous sections of the work are retained.

The software implementation of mathematical procedures is carried out by recording into software procedures consisting of descriptions of data processing operations (descriptions of data and actions on them) that reveal their mathematical content, using the Python programming language.

**The software toolkit for transforming heterogeneous symptom complexes of the database into complexes of homogeneous indicators is organized.** The data processing organization to address this issue is characterized by the flowchart of the software procedure in Figure 10.

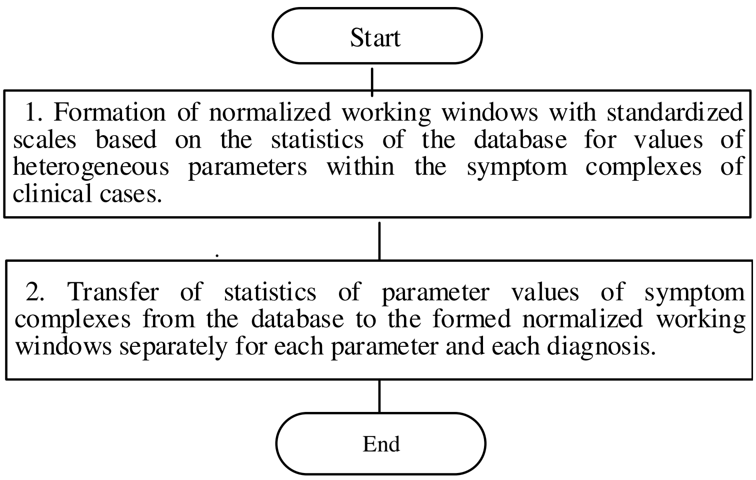


Figure 10. Block diagram of the software procedure for transforming heterogeneous symptom complexes of the database into complexes of homogeneous indicators

First, the symptom complexes are sorted by diagnoses and by the arrangement of parameters within them. Then, heterogeneous parameters with different scales of values and dimensions are gradually transformed into complexes of dimensionless homogeneous indicators, which describe clinical cases with their numerical values on convenient standardized scales from zero to one. Overall, procedures for defining and transforming the boundaries and sizes of parameter scales with their normalization and transferring the original parameter values to these scales with a transition to dimensionless quantities are used in this software module.

As a result of such transformations, the description of a clinical case in the form of a symptom complex of heterogeneous parameters from the database or the description of a clinical case of a patient who has returned can be represented as a complex of normalized dimensionless homogeneous parameters with values on identical scales [0;1].

The boundaries and sizes of the scales of the working windows, determined during the execution of this procedure, should be stored as part of the training data of the recognition system. The boundaries of the initial working windows limit the ranges of acceptable values of symptom complex parameters, for which the recognition procedures are designed to work. Numerical characteristics of intermediate working scales are necessary for transforming the parameters of symptom complexes of new patients arriving for diagnosis. Their symptom complexes must undergo the same transformations as the symptom complexes of the database used to train the diagnostic decision-making procedures.

The discussed software procedure ensures the transformation of the original database into a new form, in which symptom complexes consist of sets of dimensionless homogeneous indicators, which are convenient for building models used in recognition procedures of differential diagnosis.

Each complex of homogeneous indicators obtained in this way can be easily represented graphically by a bar chart of its components' values, to which a characterization of its form [24 – 27] can be applied. Such a characteristic can be used as one of the characteristics of a clinical case for its diagnosis.

In a multi-dimensional space with identical scales [0;1], the same sequence of numbers can be considered as coordinates of a point representing the entire such symptom complex [1, 24 – 27, 41, 42].

**The software toolkit for building smooth models of probability density distributions along the scales of parameters of symptom complexes** repeats the sequence of mathematical operations discussed for this purpose. The block diagram of the software procedure is presented in Figure 11.

This procedure is a component. First, the normalization of scales and parameter values is carried out. Then, the toolkit for forming basic stepwise models is used. Then follows the procedure for transforming them into smooth models.

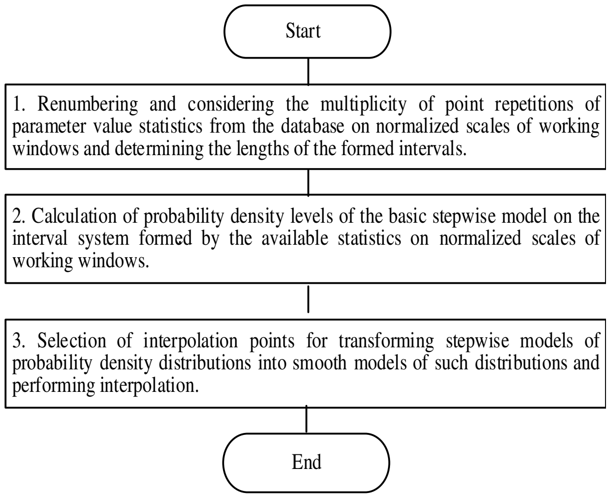


Figure 11. Block diagram of the procedure for constructing smooth models of probability density distributions along the scales of symptom-complex parameters

Stepwise models of probability density distributions are formed separately for all types of parameters of symptom-complexes and for all possible differentiated diagnoses using the same program procedure. The result of its application in processing symptom-complexes from the database consists of three sets of probability density models for all parameters. Each set corresponds to one of the possible diagnoses in the differential diagnosis of clinical cases.

In the cycle of forming stepwise models from the database, first, the entire set of parameter values for the selected diagnosis is listed in a separate array. From these values and the points representing them on the numerical axis, a sequence of separators of the parameter value range into intervals is compiled, within which probabilities can be considered constant. For this purpose, all listed parameter values are ranked in ascending order from the minimum to the maximum values using a standard procedure. They become the boundaries, the extreme points of the parameter value intervals on its scale in the available statistics of the database for the diagnosis under consideration.

In general, such statistics only occupy a certain part of the [0;1] scale of the normalized working window, without exceeding its boundaries. To the left and right of such a segment in the [0;1] working window, there may be intervals not filled with points. However, these intervals are also subsequently filled with probability density values.

In the process of ranking the parameter value statistics, there may be repetitions, and this must be taken into account in the probability density model. Therefore, on the one hand, an additional row of parameter values is formed where there are no such repetitions. At the same time, these repetitions are separately counted; their quantity is recorded in a separate synchronous row of repetition multiplicities.

If there are no repetitions of a certain value, then its multiplicity is one. With each repetition, another unit is added to it. This is done until the repetitions of this parameter value stop, i.e., until its new successive value becomes greater than the previous one.

In addition, the calculation of points remaining after excluding repetitions is also conducted separately. This numerical indicator is necessary as a parameter in the loop of the stepwise modeling procedure. If there were no points 0 and 1 in each group formed for a particular parameter, they are added with a multiplicity of zero. When any of them already existed with a certain multiplicity, it remains with the same multiplicity here.

Overall, performing such operations leads to the formation of a two-dimensional array. The first row of this array enumerates all others, the second one contains all parameter values from the database for the considered diagnosis, the third row is filled with all different parameter values in ascending order, and the fourth row represents the multiplicity of repeats for each of the different parameter values.

This array is complemented by two more rows. In the fifth row, for each point serving as a separator of adjacent intervals, the value of the constant probability density on the interval extending from this point immediately to the right to the next point is recorded. The sixth row provides the probability density values selected for each point of transition between density levels at the intervals' junctions.

In calculating the constant probability densities on the specified intervals, intermediate values of this density are first calculated as quantities inversely proportional to the length of their interval. If the multiplicity of the start point of the interval is greater than one, this quantity is multiplied by the multiplicity coefficient value. The calculated intermediate probability density values are then normalized by the sum of the areas of rectangles formed accordingly, taking into account the end intervals of the working windows. As a result, stepwise characteristics of probability density distribution are obtained on complete normalized parameter scales.

To use any of the stepwise models prepared in this way to determine the model probability density value at a given parameter value, it is sufficient to organize a review of all points in their sequence without repetitions and stop at the last point where the parameter value is still not greater than the specified one. If the parameter value of interest coincides with the parameter value at the interval separator point, then the probability density is read from the sixth row of the formed array, otherwise, it is taken from the fifth row for the selected point.

Further, the stepwise models of probability density distributions for each parameter and diagnosis are transformed into smooth models by interpolating selected points in the working window using a standard procedure.

The choice of such points boils down to forming an array where the points of the intervals' junctions alternate with the midpoints of these intervals. The characteristic density values are taken at the midpoint points, and the average values are taken at the junctions of their gradients. Procedures suitable for interpolation are those that do not lead to obtaining negative values in the primary characteristics formed because negative values cannot exist in probability density distribution models along parameter value scales. Further normalization of the primary continuous distribution dependency values for any of the parameters is performed by the areas between the graphs of these dependencies and the horizontal axes of the working scales in the normalized working windows [0;1].

The obtained stepwise and continuous models of probability density distributions are one of the main results of training recognition procedures based on the experience of medical practice, which are used for forming preferred differential diagnostic decisions for patients' symptom-complexes. They are one of the main components of information support, prior information about the varieties of these complexes, on the basis of which criteria and procedures for making plausible diagnostic decisions and possible confidence level assessments are built.

**The program procedure for forming preferred diagnoses by accumulating probabilities on model distributions for parameter values of clinical cases' symptom-complexes**. This program procedure is a Python language tool implementation of mathematical procedures for forming preferred diagnoses using probability density distribution models along parameter value scales based on the implementation of the criterion for accumulating preferred diagnostic decisions in differential diagnosis, as described in Chapter II. The block diagram of this program procedure is provided in Figure 12.

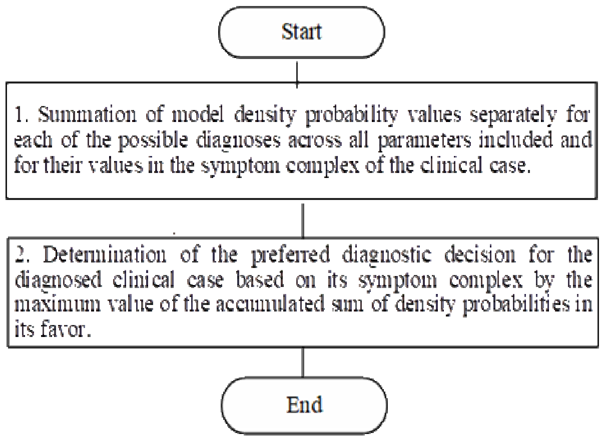


Figure 12. Block diagram of the program procedure for forming preferred diagnostic decisions based on symptom complexes of clinical cases of patients on probability density models for parameter values by accumulation

Each proposed diagnostic decision may be accompanied by numerical confidence values, which take into account the relationship between the selected diagnosis advantage and the sum of advantages characterizing their overall resource across all possible differentiated diagnoses.

The procedures use sets of probability density distribution models for values of all parameters of symptom complexes for all possible differentiated diagnoses. Models are considered in normalized parameter working windows. Clinical cases of patient symptom complexes are input into the procedures, and recommended diagnoses are output.

The procedure envisages a cyclic review of probability density models for all diagnoses and all parameters. During this process, model density probabilities are read and transferred to a separate array for each diagnosis for those parameter values in the symptom complex defined during patient examination. The decision-making criterion by accumulation is guided by the values of accumulated sums of probability densities for each diagnosis across the full set of parameters in the symptom complex. Priority is given to the diagnosis for which such a sum has the highest value. The numerical confidence indicator for choosing the proposed diagnosis can be calculated as the ratio of the sum of probabilities in favor of the accepted decision to the total sum of similar sums for all possible diagnoses.

Program procedures for assessing the effectiveness of training diagnostic decision-making procedures are discussed in the next section.

## IV. Testing of the software toolkit for training recognition procedures of differential diagnosis in patients with heterogeneous symptom complexes of their clinical cases with an assessment of its effectiveness

These tests can be regarded primarily as a kind of readiness check for the developed software procedures before their practical use, as a verification of the tuning of these procedures, the correctness of their compilation to confirm their functionality as software tools. At the same time, in the process of these tests, the main question is probably the verification of the ability of the implemented concept of forming diagnostic decisions for clinical cases of patients based on the use of specially constructed models of probability density distributions for parameter values, heterogeneous in the initial form of symptom complexes, and developed criteria for such a case decision-making. The ability to use such models in terms of their content and format of presentation, as well as the way they are obtained from symptom complexes in the database containing verified reliable diagnoses, should be verified.

Initially, in the testing process, it is necessary to go through the data processing stage proposed in the work, which provides for the transformation of heterogeneous patient symptom complexes into complexes of homogeneous parameters. Then it makes sense to check the implementation of training recognition procedures with obtaining the mentioned probability density distribution models at this stage, using which diagnostic decisions will be formed. Then it makes sense to organize a stage of testing the proposed recognition procedures on the statistics of symptom complexes with reliable diagnoses, with an assessment of the level of accuracy (overall validity) of the proposed decisions.

Such were the goals and tasks of the tests in their most general form, the solution of which was carried out step by step according to the presented plan.

Testing the software procedure for transforming heterogeneous symptom complexes of the database into complexes of homogeneous parameters. The necessary transformation of informational objects by this software procedure is illustrated in Figure 13.

Heterogeneous symptom complexes from the database describing clinical cases of patients were to be transformed by this procedure into sequences of normalized homogeneous parameters, the values of which fall within the range [0;1]. In the testing process of this procedure from the database, a symptom complex of a specific patient's clinical case was taken into account in terms of quantitative parameters, the values of which are provided in row 3 of Table 2. In row 4, the results of transformations performed by the tested procedure are recorded. It is characteristic of the processing results that the parameter values transformed by the procedure do not exceed the interval [0;1]. Furthermore, they coincide with the results of calculations performed manually on a calculator for verification, which was the expected outcome.

For the numbered sequence of normalized parameter values of any of the symptom complexes, a panorama can be compiled, which resembles the graphs of discrete sequential samples of a certain signal over time (Figure 14).

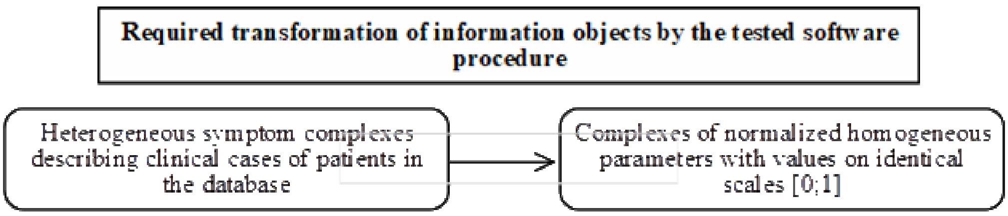


Figure 13. Required transformation of information objects by the tested software procedure

Table 2. Heterogeneous numerical parameters of the symptom complex and the result of their transformation to a unified scale [0;1]

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| Length | Width | Thickness | Thiekpar | Speed | Index | Speedup | Age |
| 125 mm | 66 mm | 51 mm | 18 mm | 26,2 sm/s | 0,667 | 275 sm/s$^2$ | 69 years |
| 0,69 | 0,54 | 0,29 | 0,65 | 0,54 | 0,53 | 0,35 | 0,91 |

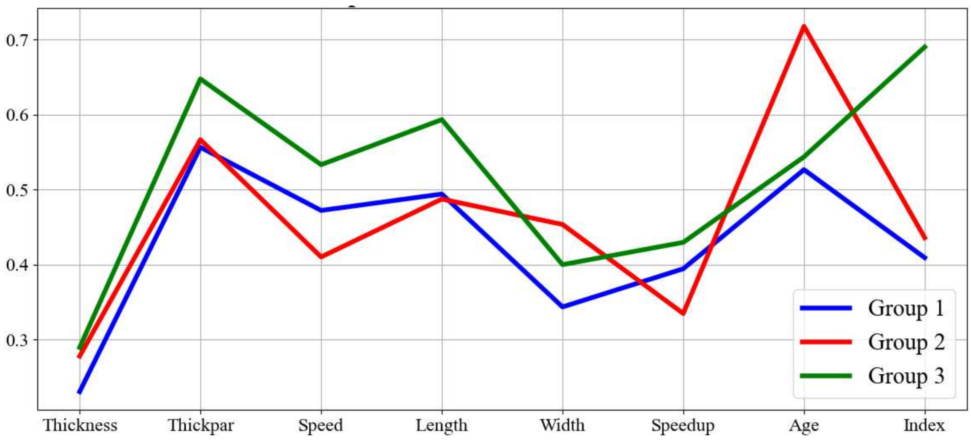

Figure 14. Panoramic presentation format of average normalized values of the symptom complex as a process for three different diagnoses

The independent variable here is the parameter number in the symptom complex. Depending on this number, the normalized parameter value is given. The scale for all parameters is the same.

The symptom complex in the format of such panoramas can be considered as a process. In this case, it acquires an additional characteristic - the characteristic of its form [24 – 27], which is undoubtedly significant for choosing a diagnosis for a clinical case.

Using such a characteristic, one can find its characteristic appearance (as shown in the figure) for each of the possible diagnoses. To form a diagnostic decision, you can compare the shape of the patient's transformed symptom complex with its characteristic appearance, with such prior information about these complexes for different diagnoses, for example, using the scalar product operation [24 – 27], the value of which expresses the level of similarity of the shape of the compared processes and allows you to search for a preferred solution.

Overall, based on the results of the tests, this software procedure performs the necessary mathematical operations in processing the input data. The panoramic presentation format of the results of such processing of symptom complexes provides certain conveniences in conducting (if necessary) additional visual data analysis to address issues of differential diagnosis of clinical cases of patients.

**Testing of the software tool for training procedures for diagnosing diagnoses based on symptom complexes of clinical cases**. By the content of the work, training software procedures should form and provide the recognition procedure of differential diagnosis of patients with the necessary prior information about these symptom complexes for different diagnoses. Based on the use of this information, the operation of selecting preferred diagnostic decisions for analyzed clinical cases is built.

The primary information about the parameter values of symptom complexes for different diagnoses in the available statistics of the database regarding reliably diagnosed clinical cases is transformed at the stage of training recognition procedures (Figure 15) into statistical characteristics of the occurrence of different values of symptom complex parameters, which is revealed by model sets of probability density distributions for the values of each of the parameters for each of the possible diagnoses.

Therefore, the entire purpose of training diagnostic recognition procedures in this work lies in forming models of probability density distributions for parameter values of symptom complexes under different diagnoses, obtaining such prior information about the types of these complexes that are recognizable. The relationship between the values of these probability density distributions taken from these models for the parameters of specific symptom complexes processed after training is necessary for the criteria for selecting preferred diagnoses embedded in the software modules of the recognition procedure.

Therefore, the entire purpose of training diagnostic recognition procedures in this work lies in forming models of probability density distributions for parameter values of symptom complexes under different diagnoses, obtaining such prior information about the types of these complexes that are recognizable. The relationship between the values of these probability density distributions taken from these models for the parameters of specific symptom complexes processed after training is necessary for the criteria for selecting preferred diagnoses embedded in the software modules of the recognition procedure.

The testing of training software procedures involved a preliminary assessment of the correctness of the obtained results in checking the possibility of obtaining complete sets of necessary models as a result of processing the available database statistics and in visually comparing these models with histograms of parameter value distributions obtained from the same statistics.

Samples of obtained step models for the diagnoses "healthy", "polycystic", "hydronephrosis" with corresponding histograms for the parameters Thiekpar and Index obtained in the tests are presented as examples in Figures 16 and 17.

During the visual inspection of the graphs, their alignment reveals correspondence between the models and histograms in the locations of point concentrations on the normalized working window scales, as required. An example of aligned models of probability density distributions for the same parameter values across different diagnoses from the set of full model complements in the information provision of diagnostic recognition procedures is shown in Fig. 18. An example of the alignment of a smooth and a stepwise model, from which it was constructed, for the Index parameter from the set of symptom complexes for the "healthy" and "hydronephrosis" diagnoses is provided in Fig. 19.

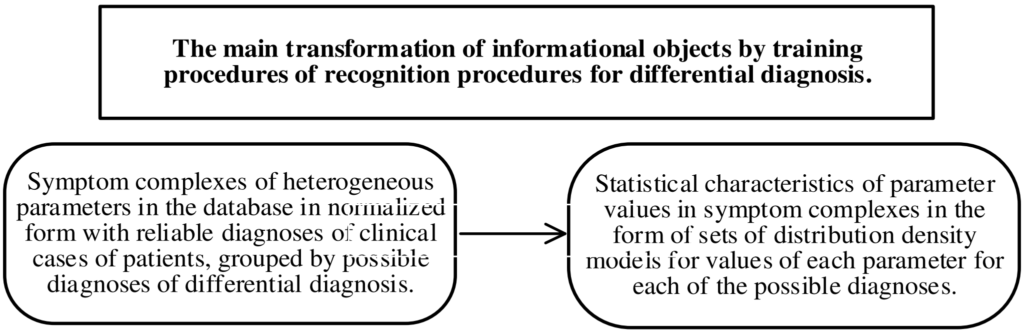


Figure 15. The main transformation of informational objects by training procedures of recognition procedures for differential diagnosis.

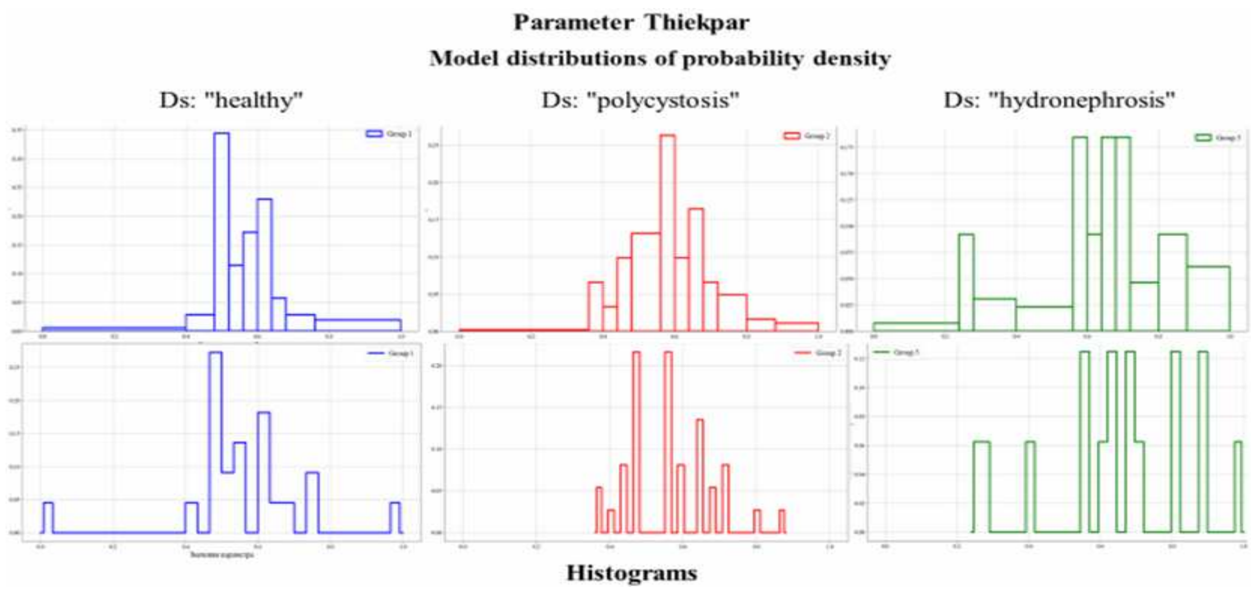


Figure 16. Samples of model probability density distributions and histograms of the Thiekpar parameter for different diagnoses obtained in the trials

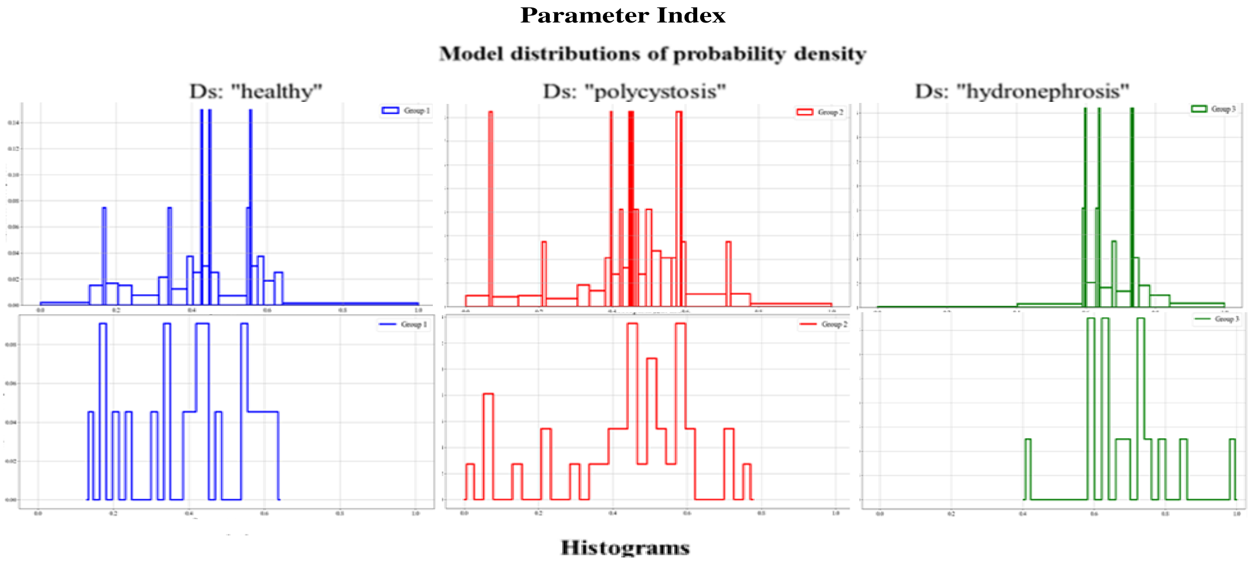


Figure 17. Model distributions of probability densities and histograms for values of the Index parameter based on the results of tests of the software procedure for learning the recognition procedures of differential diagnosis

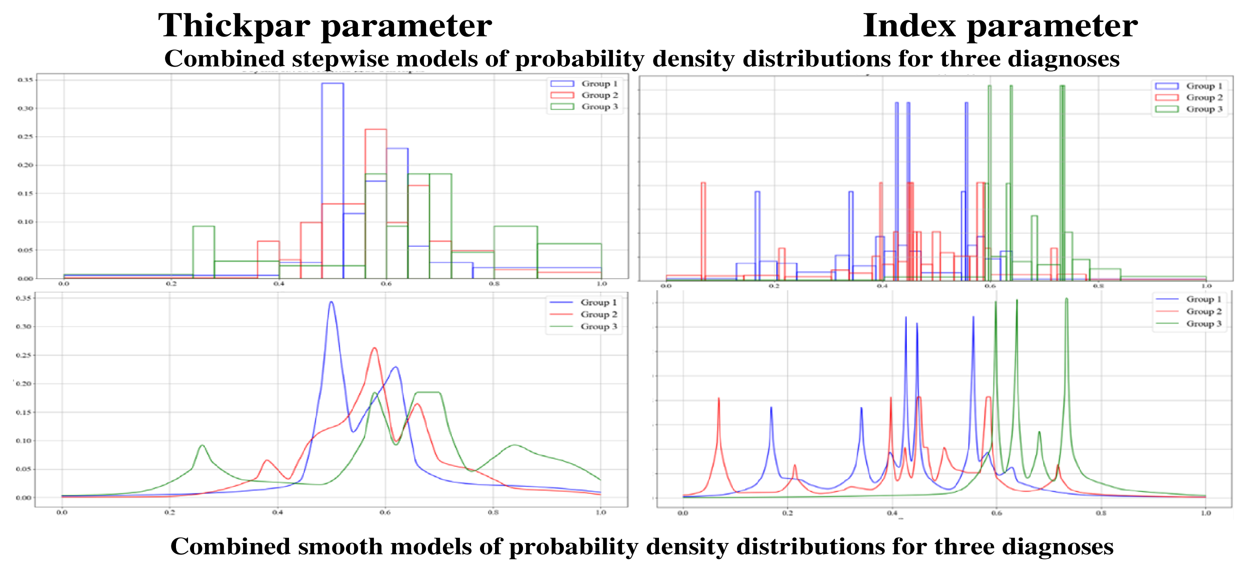


Figure 18. Combined stepwise and smooth distributions of probability densities for three diagnoses as results of the tests

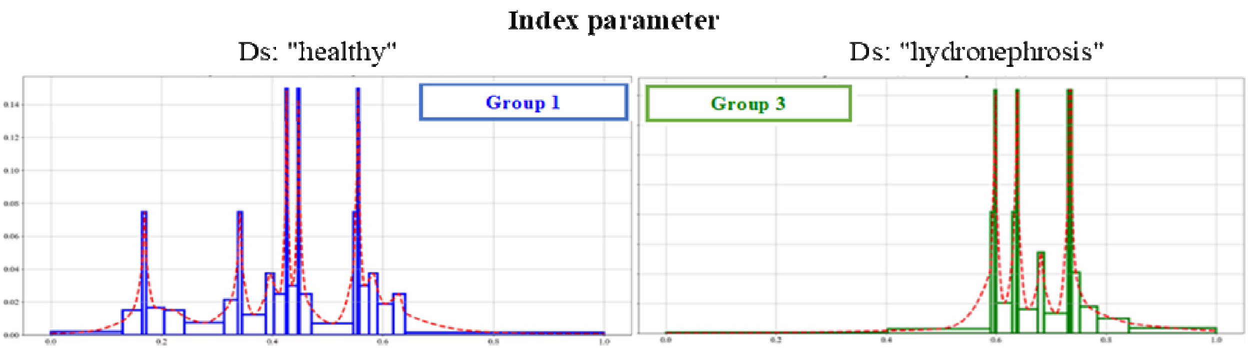


Figure 19. Combining step and smooth models of the distribution of probability densities of the Index parameter

So, overall, the tested procedures are prepared as software tools for use. The models obtained through the discussed procedures are the main result and the main component of the information support for the operation of recognition procedures in selecting their preferred diagnostic decisions in the differential diagnosis for patients with symptom complexes describing their clinical cases.

**Testing of recognition procedures, trained on reliably diagnosed symptom complexes of the database, with an assessment of the validity of the proposed decisions, was conducted.** The goal of the tests was to confirm the hypothesis of using the proposed models of probability density distributions for the values of symptom complex parameters in forming diagnostic decisions based on the criteria discussed in the work.

For this purpose, statistical tests of software recognition procedures were carried out to form their preferred diagnostic decisions by accumulating on reliably diagnosed symptom complexes from the available database used in the necessary stage of their training. Testing in each case was carried out in two stages. First, statistics of decisions proposed by the investigated recognition procedures were collected. Then the correctness indicators of these decisions were evaluated. The task of the first stage of testing was to collect statistics of decisions proposed by the investigated recognition procedures. At the second stage, the level of correctness of these decisions was evaluated, taking into account the correct diagnoses in the database.

The second stage involved counting the number of correct and incorrect decisions separately by their types according to a known scheme and calculating indicators of sensitivity, specificity, and overall validity [1, 5, 6]. Sensitivity of recognition procedures [1, 5, 6] is understood as their sensitivity to the presence of a specific health problem, as the ability to detect it in case of guaranteed presence. It requires non-omission, detection by the recognition procedure of each type of medical problem listed in the list of possible diagnoses. The procedure must be sensitive to the problem for each of its specific types, capable of detecting it based on the manifestations of the disease in the values of parameters in the clinical cases of symptom complexes.

Specificity [1, 5, 6] assesses the correctness level of all diagnostic decisions with the same assigned diagnosis. It is the percentage of correct diagnoses among all those that turned out to be the same. This indicator, like sensitivity, is related to the diagnosis of a specific type, although it has a completely different meaning and should be evaluated for each of the possible diagnoses separately.

The overall level of correctness of decisions proposed by the recognition procedure is a general characteristic of their quality (validity of decisions) [1, 5, 6], that is, the percentage or proportion of correct decisions among all decisions accepted by the recognition procedure during its testing on the available statistics.

The general format of working tables for calculating sensitivity, specificity, and overall validity of decisions of recognition procedures used in the work [1, 5, 6] has the following form (Table 3).

In the cells of the working field of the table, the number of correct and incorrect decisions of each type is provided. The calculated values of sensitivity and specificity are placed, respectively, in the last cells of the columns and rows of this matrix. The validity value (overall correctness level of decisions) is entered into the last (bottom) diagonal element of the matrix.

Here, the first indices are the numbers of the actual states of patients' organisms and their corresponding diagnoses (1 – "healthy", 2 – "polycystic", 3 – "hydronephrosis"), and the second indices are the selected diagnosis numbers from the same list by recognition procedures. The letter "n" denotes the number of cases, the content of which is revealed by the values associated with it indices.

The formulas used in the calculations have the following form [1, 5, 6]:

- for sensitivity:

$$sens\ 1 = n_{11} \Big/ \left(n_{11} + n_{12} + n_{13}\right);$$

$$sens\ 2 = n_{22} \Big/ \left(n_{21} + n_{22} + n_{23}\right);$$

$$sens\ 3 = n_{33} \Big/ \left(n_{31} + n_{32} + n_{33}\right);$$

- for specifity:

$$spec1 = n_{11} \Big/ \left(n_{11} + n_{21} + n_{31}\right);$$

$$spec2 = n_{22} \Big/ \left(n_{12} + n_{22} + n_{32}\right);$$

$$spec3 = n_{33} \Big/ \left(n_{13} + n_{23} + n_{33}\right);$$

- for overall validity:

$$vd = \left(n_{11}+n_{22}+n_{33}\right) \Big/ \left(n_{11}+n_{22}+n_{33}+n_{12}+n_{13}+n_{21}+n_{23}+n_{31}+n_{32}\right).$$

The validity of these formulas stems from the meaning of the analyzed indicators being evaluated.

Table 3. Format of working tables for calculating sensitivity, specificity, and overall validity of decisions of recognition procedures

| The recognition procedure has determined that the clinical case type can be one of the following | The type of clinical case | | | Specificity |
|---|---|---|---|---|
| | ***1*** «Healthy» | ***2*** «Polycystic» | ***3*** «Hydronephrosis» | |
| ***1*** | $n_{11}$ | $n_{21}$ | $n_{31}$ | spec 1 |
| ***2*** | $n_{12}$ | $n_{22}$ | $n_{32}$ | spec 2 |
| ***3*** | $n_{13}$ | $n_{23}$ | $n_{33}$ | spec 3 |
| Sensitivity | sens 1 | sens 2 | sens 3 | validity |

Table 4 – Table of decisions made by the recognition procedure by accumulation method during its testing process

| The recognition procedure has determined that the clinical case type can be one of the following | The type of clinical case | | | Specificity |
|---|---|---|---|---|
| | 1 «Healthy» | 2 «Polycystic» | 3 «Hydronephrosis» | |
| 1 | 20 | 7 | 1 | 0,7143 |
| 2 | 0 | 22 | 0 | 1 |
| 3 | 2 | 8 | 15 | 0,6 |
| Sensitivity | 0,9091 | 0,5946 | 0,9375 | 0,76 |

The table of the quantities of different decisions made by recognition procedures in terms of their correctness, accepted by them during the testing process, and the values of the sensitivity, specificity, and overall validity of these decisions accumulated by the method are presented below (Table 4).

Overall, the obtained results, as indicated by their analysis, are positive, which is what was needed. The level of validity of the decisions indicates that the decision-making process is not chaotic but guided by the models and criteria for correct decisions provided in the database.

**Conclusions**

The main results of the conducted research, developments, and their performance evaluation can be summarized as follows:

1. Overall, a fairly simple method of learning and a corresponding criterion for making preferred decisions have been developed for differential diagnostic recognition procedures of patients based on diverse symptom complexes composed of parameters obtained through instrumental medical examination tools to characterize their clinical cases for health problem diagnosis.
2. For informational support of diagnostic decision-making, mathematical and software procedures for forming sets of convenient models of probability density distributions for different parameter values have been developed. Based on these models and considering the specific parameter values in the descriptions of patients' clinical cases, preferred diagnoses will be formed.

The construction of models is planned for all parameters within the symptom complexes being diagnosed, for all possible diagnoses. The proposed criterion, based on the use of these models, which is implemented in mathematical and software procedures for diagnostic decision-making through accumulation.

3. To simplify the process of developing software procedures for training recognition procedures for the differential diagnosis of patients, whose clinical cases are described by diverse symptom complexes of magnitudes, and also to simplify the content of the recognition procedures themselves and the construction of specialized software and hardware tools for their implementation, if necessary, it is proposed to conduct preliminary processing of the available prior training data. During this processing, diverse symptom complexes of the database with reliable diagnoses are transformed into complexes of homogeneous quantitative indicators with values on uniform scales [0;1] in uniform working windows. For each type of parameter in the symptom complexes, a separate window is provided, which is common for all possible diagnoses in the database statistics.

Thus, the basis of these training procedures is the use of medical practice experience represented by real statistics of reliably diagnosed clinical cases. Uniform computational procedures convenient for their implementation are used to transform various heterogeneous parameters. These transformations are linear and therefore simple. At the same time, they do not disrupt the initial arrangement of parameter values on their scales in the prior data for different diagnoses. This information is preserved but is expressed more compactly and conveniently for further use. The values of constants necessary for constructing working windows and transforming values of heterogeneous parameters into new normalized scales are selected according to the real statistics of the database used in training.

4. As characteristics of symptom complexes used in the considered problem by recognition procedures as signs in determining preferred decisions, it is proposed to use descriptions of various values of each parameter for each of the possible diagnoses in the form of specially constructed models of probability density distributions along normalized scales of parameters.

These models are formed during the statistical analysis of the occurrence of parameter values in symptom complexes in the database of prior data and are harmonized with them. Such models are further used as the outcome of learning, as a form of its presentation for use in diagnostic decision-making, for which corresponding computational procedures have been developed, implemented in the Python programming language. Unlike traditionally used histograms, where empty intervals on the parameter scales may occur synchronously for all diagnoses, leading to uncertainty in the choice of diagnostic decisions, the proposed models do not allow such situations due to the established order of construction, which can be considered as their advantage.

5. The peculiarity of probability density models formed in the work lies in the fact that with their use, a

preferred decision for a specific patient's symptom complex can be separately proposed for each parameter, although the final diagnostic decision must be made based on the entirety of all parameters for the clinical case. This peculiarity stems from the fact that for any of the parameters, the models provide probability density values for each of the possible diagnoses. These densities can be compared to each other, and the most probable diagnostic decisions can be chosen.

The presence of such a feature facilitated the development of a corresponding modification of the decision accumulation method used in making diagnostic decisions, which utilizes the discussed models. To form the final decision in such a case, it is sufficient to collect probabilities in favor of each diagnosis across all parameters (or their part), find the largest such sum, and it indicates the preferred decision for the clinical case, which is the implementation of the decision accumulation method here. Moreover, based on the use of this feature of the models, indicators of confidence level for each of the diagnostic decisions proposed by such recognition procedures were constructed and proposed.

6. The procedure for calculating the value of the confidence level indicator for the diagnostic decision based on the symptom complex of a specific patient, built using probability density models, is characterized by its simplicity. The calculation of such an indicator value for the diagnosis of each specific clinical case involves computing the fraction of the model probability density attributed to the selected diagnosis, from the total sum of such densities for all possible diagnoses.
7. Testing of recognition procedures that underwent training, during which traditional indicators of sensitivity, specificity, and overall validity were evaluated on clinical cases from the database used in training, confirmed the ability of the general concept of using the proposed probability density models both at the training stage of such procedures and directly in providing necessary information for decision criteria regarding symptom complexes of specific clinical cases.

With the help of these fidelity indicators, the quality of training and the performance of recognition procedures for differential diagnosis, which utilize the results of this training in the form of models of probability density distributions built on reliable statistics, were comprehensively evaluated.

The completed research and developments, while having their own significance, have also laid the groundwork and prerequisites for expanding the scope of their application and enriching its content. There are several directions for further work in these areas:

1. Multidimensional Analysis: By normalizing the parameters of symptom complexes and considering them in multidimensional spaces, various mathematical concepts such as norms, metrics, and spatial orientation can be employed to obtain characteristics of specific symptom complexes. This facilitates determining the characteristic features of symptom complexes for different diagnoses, constructing decision criteria, conducting clustering of symptom complexes based on reliable statistical data, and more.

2. Panoramic View of Normalized Symptom Complexes: Viewing normalized symptom complexes as processes opens up opportunities to treat them as signals, utilizing diverse tools for their processing, including recognition and classification. Characteristics of symptom complexes as signals can be obtained using procedures such as decomposition of different systems into basic functions, including decomposition in frequency, time, and other domains. Characteristics of process forms and scalar products can be used to compare processes based on their form, which can be useful in solving issues of differential diagnosis of clinical cases considering the form of the characterizing symptom complexes.

3. Transformation of Basic Stepwise Models: Instead of using interpolation and approximation procedures, the transformation of basic stepwise models of probability density distributions along parameter scales into smooth models can be considered. A procedure involving the review of the original model with a sliding interval, which collects the probabilities densities within its bounds, can be used to smooth sharp transitions in the model. Naturally, normalization of the obtained dependencies will be necessary to transition to probability density models.

4. Obtaining Characteristics of Parameter Informativeness: It's evident that characteristics of parameter informativeness within symptom complexes can be obtained, along with the possibility of forming bases where differences between symptom complexes are perceived most distinctly and compactly, facilitating the resolution of many other useful questions.

**References**

[1] О. П. Мінцер, Ю. В. Вороненко, В. В. Власов, *Інформаційні технології в охороні здоров'я і практичній медицині*: У 10 кн. Кн. 5. *Оброблення клінічних і експериментальних даних у медицині*: навч. посіб. Київ, Україна: Вища шк., 2003.

[2] S. Uddin, A. Khan, M. Hossain et al. "Comparing different supervised machine learning algorithms for disease prediction", *BMC Med Inform Decis Mak*, 19, 281, 2019. DOI: 10.1186/s12911-019-1004-8.

[3] Z. Yu, K. Wang, Z. Wan, et al. "Popular deep learning algorithms for disease prediction: a review", *Cluster Comput*, 26(2), 1231–1251, 2023. DOI: 10.1007/s10586-022-03707-y.

[4] Hung C-Y, Chen W-C, Lai P-T, Lin C-H, Lee C-C. "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database", in *Engineering in Medicine and Biology Society* (EMBC), 2017 39th Annual International Conference of the IEEE, vol. 1; 2017.

p. 3110–3. IEEE. DOI: 10.1109/EMBC.2017.8037515.
[5] О. П. Шуляк, *Основи будови телемедичних систем: Загальні теоретичні і прикладні питання телемедицини і будови телемедичних систем*. навч. посібн. Київ, Україна: КПІ ім. Ігоря Сікорського, 2022.
[6] Телемедичні системи: Комунікаційне та інформаційне забезпечення телемедичних процедур. Лабораторний практикум (ділові ігри) та розрахункова робота до нього [Електронний ресурс]: навч. посіб. для студ. спец. 153; уклад.: О. П. Шуляк. Київ: КПІ ім. Ігоря Сікорського, 2020.
[7] S. Ekız, P. Erdoğmuş, "Comparative study of heart disease classification", in *Proceedings of the 2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*; Istanbul, Turkey. 20–21 April 2017; New York, NY, USA: IEEE; 2017. pp. 1–4. DOI:10.1109/EBBT.2017.7956761.
[8] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, K. Bashir, "Improving heart disease prediction using feature selection approaches", in *Proceedings of the 2019 16th International Bhurban Conference on Applied Sciences and Technology* (IBCAST); Islamabad, Pakistan. 8–12 January 2019; New York, NY, USA: IEEE; 2019. pp. 619–623. DOI: 10.1109/IBCAST.2019.8667106
[9] M. S. Borah, B. P. Bhuyan, M. S. Pathak, P. Bhattacharya, "Machine learning in predicting hemoglobin variants", *Int J Mach Learn Comput,* 8(2):140–3, 2018. DOI: 10.18178/ijmlc.2018.8.2.677
[10] M. Chen, Y. Hao, K. Hwang, Lu Wang, Lin Wang "Disease prediction by machine learning over big data from healthcare communities", *IEEE Access,* 5:8869–79, 2017.
DOI: 10.1109/ACCESS.2017.2694446.
[11] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, X. Wei, "Predicting the risk of heart failure with ehr sequential data modeling", *IEEE Access*, 6:9256–61, 2018. DOI:10.1109/ACCESS.2017.2789324.
[12] D. Sisodia, D. S. Sisodia, "Prediction of diabetes using classification algorithms", *Procedia Comput Sci.,* vol. 132, pp. 1578–85, 2018.
DOI:10.1016/J.PROCS.2018.05.122.
[13] T. Marikani, K. Shyamala, "Prediction of heart disease using supervised learning algorithms", *Int J Comput Appl.*, 165(5):41–4, 2017.
DOI: 10.1016/j.compbiomed.2021.104672.
[14] H. Mansoor, I. Y. Elgendy, R. Segal, A. A. Bavry, J. Bian, "Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: a machine learning approach", *Heart Lung.*, 46(6):405–11, 2017. DOI: 10.1016/j.hrtlng.2017.09.003.
[15] O. P. Shuliak, P. Hénaff, A. D. Shachykov, D. R. Kulakhmetov, R. K. Haponenko, "Ranking the functional states of a group of individuals by the activity indicators of regulatory systems evaluated using electrocardiography data," *Bull. Kyiv Polytech. Inst. Ser. Instrum. Mak.,* is. 57(1), pp. 84-96, 2019. DOI: 10.20535/1970.57(1).2019.172029. (8)
[16] Shulyak A. P., Shachykov A. D. "Decomposition of Biomedical Signals on Mutually Orthogonal Components in the Diagnosis of Diseases", in *2014 IEEE XXXIV International Scientific Conference Electronics and Nanotechnology* (ELNANO), April 15-18, Kyiv, Ukraine, 2014, pp. 291-294. DOI: 10.1109/ELNANO.2014.6873914
[17] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, I. Chouvarda, "Machine learning and data mining methods in diabetes research", *Comput Struct Biotechnol J.* 15:104–16, 2017. DOI: 10.1016/j.csbj.2016.12.005
[18] C. M. Lynch, et al. "Prediction of lung cancer patient survival via supervised machine learning classification techniques", *Int J Med Inform.*, 108:1–8, 2017. DOI: 10.1016/j.ijmedinf.2017.09.013.
[19] Choi E., Bahadori M. T., Schuetz A., Stewart W. F., Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR Workshop Conf Proc.* 2016 Aug;56:301-318. Epub 2016 Dec 10. PMID: 28286600; PMCID: PMC5341604.
[20] T. Erol, A. Mendi, D. Dogan, "The digital twin revolution in healthcare", in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT),* pp. 1–7, 2020. DOI:10.1109/ISMSIT50672.2020.9255249.
[21] A. Shulyak, T. Saurova, A. Shachykov, V. Sirotenko and V. Lahutin, "Modifications of recognition procedures of biomedical signals using shape characteristics for their description and in decision criteria," in *2020 IEEE 40th International Conference on Electronics and Nanotechnology* (ELNANO), Kyiv, Ukraine, 2020, pp. 520-525, DOI: 10.1109/ELNANO50318.2020.9088862.
[22] T. Puyalnithi, V. M. Viswanatham, "Preliminary cardiac disease risk prediction based on medical and behavioural data set using supervised machine learning techniques", *Indian J Sci Technol.*, 9(31):1–5, 2016.
DOI:10.17485/ijst/2016/v9i31/96740.
[23] A. B. Nassif, O. Mahdi, Q. Nasir, M. A. Talib, M. Azzeh, "Machine learning classifications of coronary artery disease", in *Proceedings of the 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing* (iSAI-NLP); Pattaya, Thailand. 15–18 November 2018; New York, NY, USA: IEEE; 2018. pp. 1–6. https://doi.org/10.48550/arXiv.1812.02828
[24] A. P. Shulyak, A. D. Shachykov, "Criteria and Procedures for Estimating the Informativity and Feature Selection in Biomedical Signals for their Recognition", *«Вісник НТУУ КПІ» Серія – Радіотехніка. Радіоапаратобудування*. Вип. 66, с. 79 -86, 2016.
WoS DOI: 10.20535/RADAP.2016.66.79-86.
[25] A. P. Shulyak, A. D. Shachykov, "Analysis of the biomedical signals' structure in the problem of

recognition", in *2014 IEEE XXXIV International Scientific Conference Electronics and Nanotechnology (ELNANO)*, April 15-18, 2014, Kyiv, Ukraine. p. 281-285.
DOI: 10.1109/ELNANO.2014.6873982

[26] A. P. Shulyak, A. D. Shachykov, "About the Impact of Informative Features Selection in the Mutually Orthogonal Decompositions of Biomedical Signals for their Recognition", in *Electronics and Nanotechnology (ELNANO), 2016 IEEE 36th International Conference*, 19-21 April 2016. pp. 228-231.
DOI: 10.1109/ELNANO.2016.7493054. (4)

[27] В. В. Лагутін, В. І. Сиротенко, А. Д. Шачиков, О. П. Шуляк, "Кластеризація медико-біологічних сигналів в розпізнавальних системах, що навчаються з учителем, "*Мікросистеми, Електроніка та Акустика*", т. 24, вип. 6, с. 38-52, 2019.
DOI: 10.20535/2523-4455.2019.24.6. (15)

[28] В. В. Лагутін, В. І. Сиротенко, А. Д. Шачиков, О. П. Шуляк, "Вибір меж для шкал медико-біологічних сигналів в алгоритмах їх розпізнавання", «*Мікросистеми, Електроніка та Акустика*», т. 25, вип. 1, с. 11 – 19, 2020.
DOI: 10.20535/2523-4455.mea.197291. (8).

[29] D. Li, C. Zheng, J. Zhao, Y. Liu, "Diagnosis of heart failure from imbalance datasets using multi-level classification", *Biomed. Signal Process. Control*., 81:104538, 2023.
DOI: 10.1016/j.bspc.2022.104538.

[30] L. Hussain, et al. "Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies", *Cancer Biomarkers*, 21(2):393–413, 2018. DOI: 10.3233/CBM-170643.

[31] "Body Fat Prediction Dataset". [Electronic resource]. Available: https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset.

[32] "Breast Cancer Wisconsin (Diagnostic) Data Set". [Electronic resource]. Available: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.

[33] "Cirrhosis Prediction Dataset". [Electronic resource]. Available: https://www.kaggle.com/datasets/fedesoriano/cirrhosis-prediction-dataset

[34] "COVID-19 Dataset". [Electronic resource]. Available: https://www.kaggle.com/datasets/meirnizri/covid19-dataset?resource=download

[35] Fetal Health Classification". [Electronic resource]. Available: https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification /code

[36] "Heart Attack Analysis & Prediction Dataset". [Electronic resource]. Available: https://www.google.com/search?q=Heart+Attack+Analysis+%26+Prediction+Dataset&oq=Heart+Attack+Analysis+%26+Prediction+Dataset&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIICAEQABgWGB7SAQgzMDc4ajBqN6gCALACAA&sourceid=chrome&ie=UTF-8.

[38] "Heart Failure Prediction Dataset". [Electronic resource]. Available: https://www.google.com/search?q=Heart+Failure+Prediction+Dataset&oq=Heart+Failure+Prediction+Dataset&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIMCAEQABgUGIcCGIAEMgcIAhAAGIAEMggIAxAAGBYYHjIICAQQABgWGB4yCAgFEAAYFhgeMggIBhAAGBYYHjIICAcQABgWGB4yCggIEAAYChgWGB4yCAgJEAAYFhge0gEIMzgxN2owajeoAgCwAgA&sourceid=chrome&ie=UTF-8.

[39] І. А. Голованова, І. В. Бєлікова, Н. О. Ляхова, *Основи медичної статистики*: навч. посіб. для аспірантів та клінічних ординаторів. Полтава, Україна, 2017.

[40] Python 3.8.3 [Electronic resource]. Available: https://www.python.org/downloads/release/python-383/

[41] О. В. Чалий, Я. В. Цехмістер, К. О. Чалий, *Математична обробка медико-біологічних даних. Навчальний посібник з модуля № I дисципліни "Медична і біологічна фізика"*. Київ, Україна: Вид-во НВП "Інтерсервіс", 2011.

[42] Основи молекулярної біології та біоінформатики: комп`ютерний практикум [Електронний ресурс]: навч. посіб. для студ. спеціальності 122 «Комп`ютерні науки та інформаційні технології спеціалізації «Інформаційні технології в біології та медицині» / С. В. Кисляк, Є. А. Настенко; КПІ ім. Ігоря Сікорського. – Електронні текстові данні (1 файл, 2957 Кбайт). – Київ: КПІ ім. Ігоря Сікорського, 2018. – 95 с.

УДК 616.6:004.67

**О. П. Шуляк, В. В. Дружинін**
*Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна*

## ПРОГРАМНІ ПРОЦЕДУРИ НАВЧАННЯ РОЗПІЗНАВАЛЬНОЇ СИСТЕМИ ДЛЯ ДИФЕРЕНЦІЙНОЇ ДІАГНОСТИКИ ПАЦІЄНТІВ ЗА РІЗНОРІДНИМИ СИМПТОМОКОМПЛЕКСАМИ

Розглядається система розпізнавання машинного навчання для диференціальної діагностики пацієнтів на основі гетерогенних комплексів нефрологічних параметрів, перехідна від інструментальних засобів обстеження. Під

час навчання використовується емпірична статистика клінічних випадків у базі даних із надійними діагнозами.
Мета полягає в тому, щоб розширити можливості вилучення інформації з аналогічних баз даних для навчання процедурам розпізнавання шляхом збагачення цього інструментарію новими функціями, що містять характерні аспекти витягнутої інформації.
Об'єктом дослідження є математичний та програмний інструментарій для навчання процедур розпізнавання диференціальної діагностики пацієнтів на основі статистики достовірно діагностованих клінічних випадків.
Предметом дослідження є програмні процедури формування моделей падіння комплексу параметрів під час навчання за шкалами їх значень та процедури використання цих моделей у діагностиці. Освоєння моделі сприймається як основний зміст навчального процесу в забезпеченні диференціації діагнозу. Запропоновано критерій прийняття преференційних діагностичних рішень з використанням таких моделей.
Для спрощення розробки математичних і програмних процедур різнорідні симптомокомплекси нормалізуються і перетворюються на [0; 1] масштаб.
У вступі констатується значна поширеність у медицині та суміжних галузях баз даних зі статистичними даними медико-біологічних параметрів і характеристик органів і систем людини в різних станах, їх медичної інтерпретації та використання в різних цілях, часто пов'язаних з діагностикою пацієнтів.
Проблеми їх формування та використання окреслено на реальних базах даних, причому одним з ускладнюючих факторів у розробці діагностичного апаратно-програмного забезпечення є значна неоднорідність параметрів, що визначаються приладами обстеження пацієнтів.

**Ключові слова**: діагностика пацієнтів; гетерогенні симптомокомплекси; нормалізація параметрів; моделі розподілу параметрів; критерій накопичення рішень.

*Надійшла до редакції*
*06 березня 2024 року*

*Рецензовано*
*10 квітня 2024 року*