Methodology of Disease Risk Assessment

Karina Melnyk, Natalia Borysova and Svetlana Ershova

National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str., 2, Kharkiv, 61002, Ukraine

Abstract

The process of early diagnosis in terms of determining the disease risks is considered. The relevance of the process of medical cards screening as one of the important task of maintaining health of the population of the country is substantiated. An analytical review of existing Data Mining methods for processing of medical data for solving the problem of disease risk assessment is conducted. The choice of Case Based Reasoning approach for searching the relevant medical cards is confirmed. The Methodology of risk disease assessment based on the use of the k-Nearest Neighbors method is proposed. The functional model of the formation of the base of precedents in IDEF0 notation has been developed. A relevant case retrieving model based on the use of the Gower coefficient as a convolution criterion for mixed data has been proposed. Numerical studies of the task have been carried out. The process of early diagnosis using the developed methodology has been improved. Recommendations for assessing the effectiveness of using the proposed methodology in practice based on calculating the values of the confusion matrix have been developed.

Keywords 1

Screening of medical records, early diagnosis of diseases, Case Based Reasoning, k-Nearest Neighbors method, similarity measure for mixed data, Gower coefficient

1. Introduction

Many countries have been seen a decline in economic growth recently. One of the reasons is the labor force shrinking. Labor force or the labor market activity rate is the amount of active persons (occupied labor force and the unemployed) as a percent of the total population [1]. There are many reasons, which have influence on the number of the working population. They are epidemiological situation in the country and in the world, in general, high incidence and death rate, the rejuvenation of many illnesses leading to disability, insufficient funding of the health care system, the lack of mandatory regular medical examinations and programs for the prevention and early diagnosis of serious diseases. For example, Covid-19 has become the reason for the bankruptcy of many enterprises from the sector of small, medium and even large businesses [2]. State policy, social norms, a decrease in the birth rate, and emigration of the population also contribute to reducing the labor force. According to the UN, the population of Ukraine will decrease by 18% over the next 10 years [3]. This problem cannot be solved by raising the retirement age only. Analyzing the aforementioned problems, one can conclude that the way out of the situation can be achieved by different levels of government working together on the problem. It is necessary to make adjustments in a legislative framework of the country: appropriate redeployment of available budget funds, search for additional sources of funding, for example, through charity events and organizations, creating benefits for private clinics that serve retirees. Doctors need the opportunity and funds for the scientific researches: the publication of their medical articles; creating a knowledge base with access for not only medical professionals but also ordinary patients. That is, it is necessary to provide an opportunity to conduct a sufficient number of researches that will increase the effectiveness of medical decisions and reduce

IDDM'2020: 3rd International Conference on Informatics & Data-Driven Medicine, November 19–21, 2020, Växjö, Sweden

EMAIL: karina.v.melnyk@gmail.com (K. Melnyk); borysova.n.v@gmail.com (N. Borysova); svetlana.ershova.2016@gmail.com (S. Ershova)

ORCID: 0000-0001-9642-5414 (K. Melnyk); 0000-0002-8834-2536 (N. Borysova); 0000-0003-3893-117X (S. Ershova)

the risk of wrong medical decisions. Officials need to develop mechanisms to monitor the organization of the health prevention and early diagnosis process.

In general, the procedure of early diagnosis or screening is a method of conducting simple and safe researches of large groups of the population. It can be used in many fields of medicine to determine and assess the likelihood of a disease. Therefore, it is a procedure for the disease risk assessment. To date, in evidence-based medicine, there are proven tools for determining the risk of many diseases, for example, the Framingham scale of 10-year risk can be used for assessing the risks of cardiovascular disease (CVD) [4]; the risk of Alzheimer's disease can be assessed using a specialized software tool brainsalvation, which allows to assess the Risk Reduction Score and take appropriate measures [5]; the Heaf test can be used for assessing the degree of tuberculosis [6] etc. The main source for screening procedures is a database with medical records. Patient information from medical cards can be represented by different types of data [7]. For example, descriptions of complaints and patient examination results are presented in text form. Quantitative data describe the patient's age, weight, and test results, such as leukocytes count. There is also information in the form of scales: the presence or absence of any disease, diagnosis codes, severity of the condition (minor injury, moderate, severe), stage of the disease, stage of treatment, degree of heart failure. Therefore, to solve the problem of disease risk assessment, it is necessary to use such mathematical methods that will process large amounts of medical data and different types of information. At the same time, all mathematical calculations must comply with current health legislation. The WHO has developed set of documents and principles, which form the base for screening studies and disease risk assessment [8, 9]. Medical and technological documents for screening activities for each group of diseases have been identified: unified clinical protocols, local protocols and adapted clinical guidelines. For instance, to determine the risk of CVD, it is necessary to use a clinical protocol for the provision of medical care for CVD disorders [10] and an adapted clinical guideline based on evidence in the prevention of CVD [11]. However, despite the existing research of this task, many questions remain open. The general approach to the processing of data from medical cards for identification the disease risk remains a topical issue.

Thus, the purpose of this work is to develop the methodology of risk disease assessment based on data from medical records.

2. Formal problem statement

According to the current medical and technological documents and the WHO risk assessment guidelines [12], the task of risk disease assessment in general is a classification task. The following risk assessment scale has been chosen in accordance with the work [12]: low risk group, moderate risk, high and very high risk. The mathematical formulation of the task of risk assessment can be represented in the following way.

Let $R = \{r_1, \dots r_n\}$ is a set of medical records from some clinic database. Let $G = \{g_1, \dots g_4\}$ is a set of risk assessment groups based on scale from [12]. The task of risk assessment is a mapping of one set to another $f: R \to G$.

To solve this task, it is necessary to:

- develop the methodology for risk assessment;
- choose a classification method for medical records;
- determine a class or appropriate risk group.

3. Literature review

Let's consider the most commonly used Data Mining methods of analysis of medical data for solving the problem of disease risk assessment in terms of the distribution of medical records into separate groups with different risk values.

1. Neural networks.

Using of neural networks in practice for solving the classification task of medical data for linearly separable classes gives good results [13-15]. The data type of patient information is the mixed data, so it can be converted and normalized for using it as input data to the neural network. The neural network adapts every time with new medical data, that is train pattern can be increased. Each group of diseases has its own network model, which takes into account a certain set of input data. The disadvantages of this approach are the large training set with medical information, the problem of choosing the network architecture and method of training.

2. Bayesian networks.

The mathematical apparatus of Bayesian belief network allows to develop a model for assessing the disease risk for each group of diseases separately [16, 17]. The vertices in the directed graph can be the following: bad habits, bad heredity, diseases and their symptoms, test results. The edges in the probability-oriented graph reflect the conditional dependence of one vertex to another. That is, Bayesian networks allow to process mixed data, which is the advantage of this approach. With sufficient statistical information on patients, this approach gives good results in disease risk assessment. Disadvantages include the ambiguity of knowledge about the relationship between the selected input data and the complexity of network creation.

3. Theory of artificial intelligence, method of comparative identification.

The task of disease risk assessment can be solved with the help of the theory of intelligence proposed by school of Yu. P. Shabanov-Kushnarenko, namely the method of comparative identification [18-19]. This approach allows modeling the doctor's reasoning based on the use of finite predicates algebra. In this case, the domain area should be represented as a logical network. The vertices in the network are predicate variables; the edges are the values of these variables. Predicate variables are information from medical records and other sources of medical data. This information can be in the form of qualitative, quantitative and categorical data. The disadvantage of this approach is the complex process of data preprocessing in compiling a risk assessment model, because the expert has to present all possible combinations of all values of predicate variables of the domain area in the form of system of predicate equations.

4. Decision trees.

The domain model is presented in the form of a hierarchical structure or a tree [15, 20, 21]. Each node of the tree is a question related to the results of analyzes or instrumental-computer studies, symptoms of diseases or the presence of complaints. Transitions between nodes are possible answers. To determine the risk class of the disease, it is necessary to answer the questions at the nodes of this tree, starting from its root. The advantage of this method is understandable visualization the doctor's reasoning. A limitation of this method is the difficulty of choosing the next node. There are many various algorithms for this, but they often give too detailed tree structure and lead to errors.

5. Case based reasoning.

Case Based Reasoning (CBR) is an approach that help to make a decision based on the previous cases and experience [15, 22-24]. To assess the disease risk, it is necessary to choose similar precedents that have occurred in the past. Then a similarity measure is calculated between new and all selected cases. The advantages of CBR include the ability of using the experience gained by the system without the constant expert intervention; reducing the decision-making time by using already made decision; usage of heuristics that increase the efficiency of the solution search process. Disadvantages include the following: a large number of precedents can lead to reduced system performance; the problem of choosing the similarity measure or algorithms for determining similar precedents; the impossibility of obtaining solutions if there are no similar precedents or the similarity measure is less, than the specified threshold value.

Comparing Data Mining approaches to determine the disease risk group, this study proposes to choose Case Based Reasoning as a method for classifying medical records.

4. Materials and methods

4.1. Using CBR approach for disease risk assessment

Let's consider the usage of Case-Based Reasoning for resolving the disease risk evaluation task. This approach allows making a decision for new cases using previously solved problems and reusing their solutions. In general, the CBR process for the task of disease risk assessment can be represented in Data Flow Diagram (DFD) notation (Fig. 1). Case base is a repository with medical data about patients and their anamnesis. New case is an external entity, which is considered as current state of some patient with his/her anamnesis. So, input data for CBR-method are appropriate set of features, symptoms and results of analysis. It can be identified according to the particular disease.

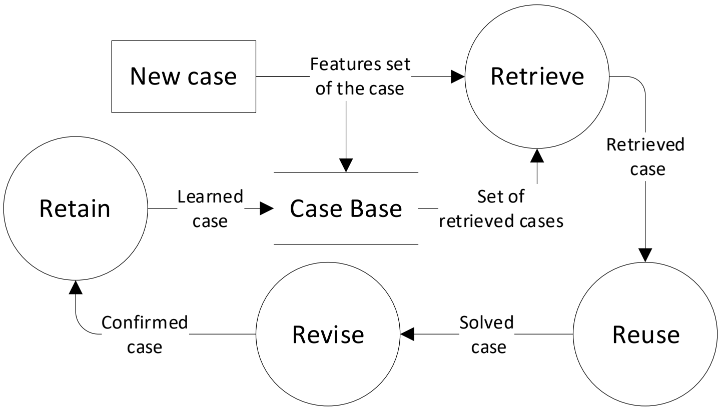

Figure 1: General structure of Case-Based Reasoning cycle

There are four phases of CBR-cycle [24].

- Retrieve. It is a search of relevant precedents. According to chosen similarity measure or comparison method, it is necessary to find similar cases or case from case base datastore.
- Reuse. It is a process of using the retrieved precedent for solving the current problem. If proposed solution is suitable, than the problem is considered as solved.
- Revise. The suggested solution is evaluated for correctness and is adapted if necessary. Result of this stage is the confirmed case.
- Retain. The obtained case stored in the case base datastore as a newly learned case.

Let's consider the methodology for determining the disease risk in more detailed way. In general, the methodology of any medical business process is the specific procedures or techniques used to identify, select, process, and analyze medical information. The disease risk assessment process is one of the business processes in a healthcare facility. Let us take a look at a functional model of this business processes in IDEF0 notation from the perspective of the CBR-approach (Fig. 2).

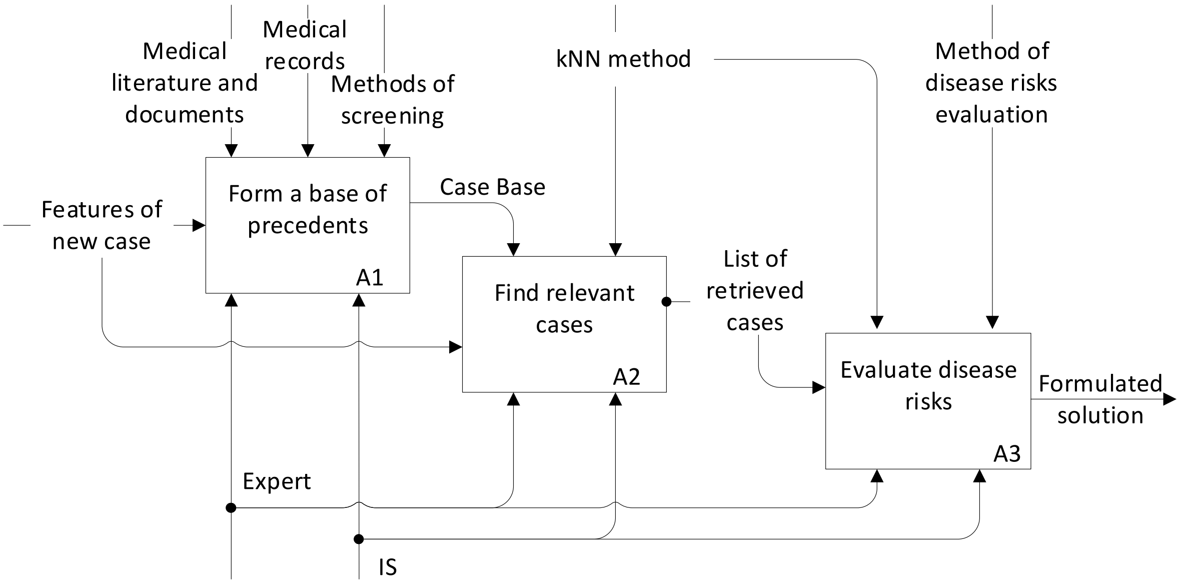

Figure 2: Methodology of disease risk assessment

To eliminate the influence of the human factor, reduce the time spent for making medical decisions and decrease the errors, it is necessary to automate the process of risk assessment. It means the usage of informational technology to process medical data. Disease risk assessment with using CBR-approach consists of several stages. The outcomes of each stage are used as input data for the next one. Therefore, expert and medical informational system (IS) are mechanisms, which should be used together at each stage. Let's take a closer look at the screening process for disease risk assessment.

1. Form a base of precedents. Each disease group is characterized by its own set of input data. Therefore, depending on the conditions of the assessment task, its own base of precedents is formed. It is the basis for medical decision for assessing the risks of a particular disease. At the same time, the formation of the base of precedents must comply with the current medical and technological documents.
2. Find relevant cases. This is the stage for retrieving the precedents. There are many methods for finding relevant cases. For example, Data Mining methods, genetic algorithms, decision trees, cluster analysis. One of the most common methods is the k-th Nearest Neighbors (kNN) method, which allows to calculate a similarity measure of cases [25]. Outcome of this method is a list of relevant k-th precedents. The list is the basis for determination of characteristics of considered precedent.
3. Evaluate disease risks. At this stage, found precedent or list of precedents is reused and revised. The main question of this phase is whether there is a difference between the input dataset (current case) and the found cases. If the difference is insignificant, then the risk group is the same as the most frequently encountered group in the list of relevant cases. If the input data has many differences, then only a preliminary risk of the disease can be determined. Such risk should be refined. For instance, in work [26], it is proposed to use a model for choosing the optimal set of screening actions to form an individual survey plan. The resulting solution is saved to the base of precedents.

Let us consider the process of forming the base of precedents in the proposed methodology in more details. The model for identifying the set of precedents that is the basis for making a decision regarding to the risk group is presented on Fig. 3.

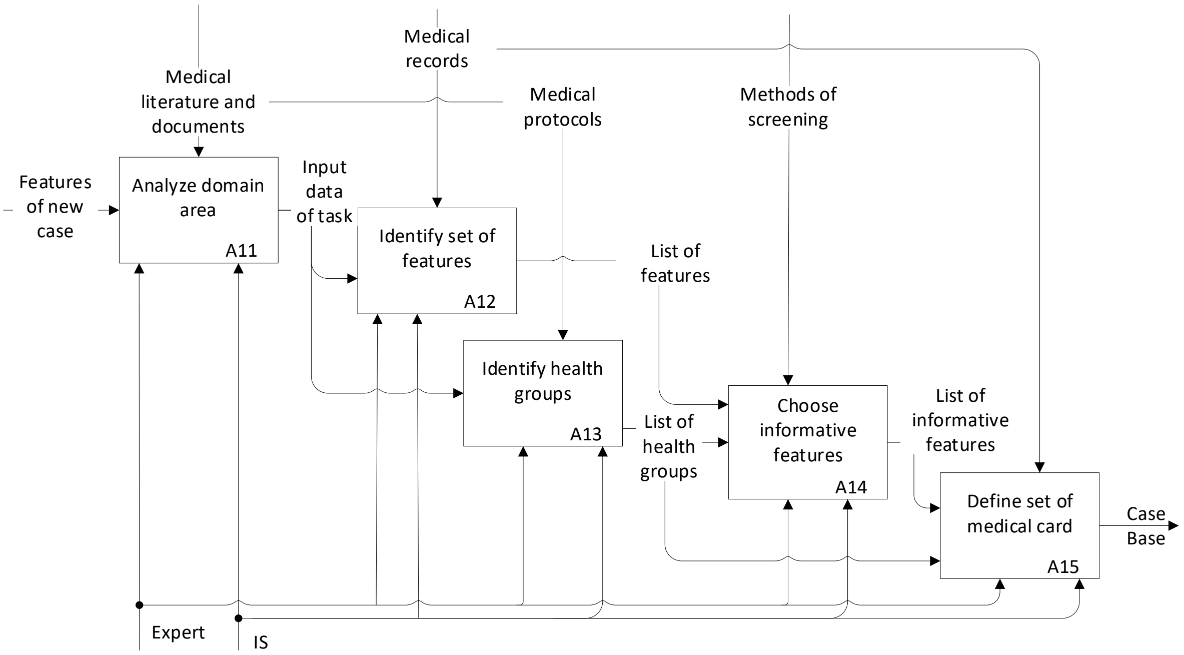

Figure 3: Model of the formation of the base of precedents

Analysis of the domain area allows forming the set of input data. It consists of symptoms of the disease or group of diseases, risk factors, concurrent diseases and the results of various examinations. The next step is to identify the selected medical data. For example, qualitative data should be formalized according to the selected scales; quantitative medical data should correspond to the accepted units of measurement according to WHO. According to the medical protocols and guidelines from WHO, a gradation of health groups is defined for the selected data. If the set of input information is too large, then the efficiency of the IS decreases. Therefore, it is necessary to evaluate and select informative features. This process is called dimensionality reduction [27, 28]. There are many different methods for feature selection or feature screening. The dimensionality reduction usually involves such approaches: filter, wrapper, embedded, and hybrid methods. The filter methods evaluate the relevance of features based on a given function. Spearman's rank correlation coefficient or Information gain can be used as such function. Then, according to the given rule, expert defines which features can be left in the result set. The main idea of the wrapper methods is choosing the selection criterion or given classifier, which is a base for forming the feature sets. After that, the training quality of each set from the precedent base is assessed using given classifier. The next step is changing of the features set. It can be possible with the sequential removal or addition of features according to the convolution criterion changes. Another way is using the algorithm of full search (depth or breadth search), or a genetic algorithm. The embedded methods are similar to the wrapper methods, as the quality of training is also assessed using the selected classifier. The difference is that embedded methods allow the classifier to take into account some feature information during the training. This allows reducing the time of feature selection and reducing the risk of overfitting. The updated set of input data allows to select medical records from the database of records, which provide enough information for decision making.

Thus, the base of precedents is formed, which is then used for searching of relevant cases.

4.2. The model of retrieving the relevant precedents

In this study, it is proposed to conduct the process of retrieving the relevant cases using the kNN-method. The main idea of this method is to select a set of precedents that are similar to the new given precedent. Firstly, the expert should choose the similarity measure. If type of input data is quantitative type, then the similarity measure can be calculated using the following metrics: Euclidean measure, Manhattan distance, the Mahalanobis distance, Chebishev measure and others. For qualitative and categorical information, it is possible to use Hamming distance, Rogers-Tanimoto dissimilarity index, Rao distance etc. But the medical information usually has mixed type, so it is recommended to use Zhuravlyov metric, Gower coefficient, Voronin measure, Mirkin metric [29, 30]. Let's consider the algorithm of the kNN-method for screening the medical records for a selected disease (Fig. 4).

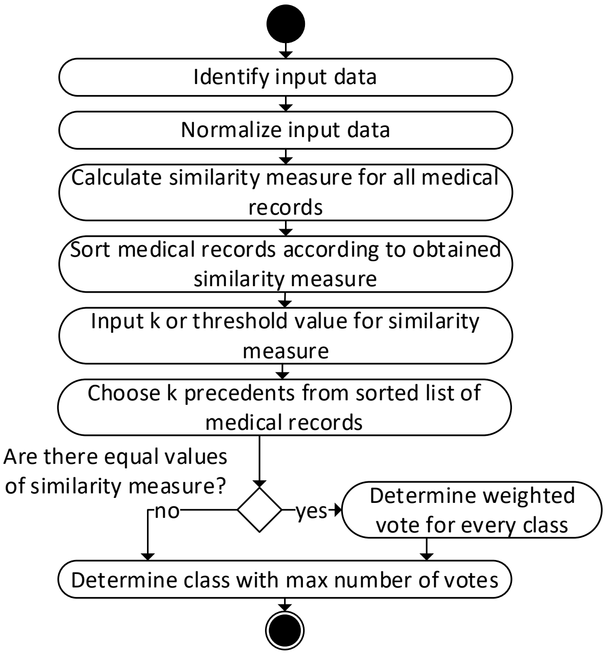
Identify input data
Normalize input data
Calculate similarity measure for all medical records
Sort medical records according to obtained similarity measure
Input k or threshold value for similarity measure
Choose k precedents from sorted list of medical records
Are there equal values of similarity measure?
no
yes
Determine weighted vote for every class
Determine class with max number of votes

Figure 4: The algorithm of kNN - method for determining the risk group

Consider the process of identifying the input data. Let’s denote I as a set of different records of patient information in a medical card, for example, risk factors, diagnoses, symptoms, medical research results. Then $J_i, i \in I$ is the set of records of the i-th type. The example of features from risk factors group is a presence of bad habits, obesity, tobacco smoking, etc.

Let’s assume x_{rij} $(r \in R,\ j \in J_i, i \in I)$ is a value of j-th record of i-th type from r-th medical card. The range of changing of quantitative data in medical records can be very wide and differ from each other by several orders of magnitude. The kNN-method is sensitive to the range of changes of input variables, since the imbalance between the quantitative values of the input medical data set can cause the instability of the algorithm and can worsen the learning outcomes. Therefore, for quantitative data from medical cards, it is desirable to perform normalization or scaling. It means finding $\hat{x}_{rij}$ as the normalized value of j-th record of i-th type from r-th medical card. One of the common variants of normalization is used the standard deviation according to the formula (1):

$$\hat{x}_{rij} = \frac{x_{rij} - \bar{x}}{\delta_x}, \tag{1}$$

where $\bar{x}$ – is an average of j-th record of i-th type of all n-th given medical cards from database, calculated by formula (2); δ_x is the standard deviation of j-th record of i-th type from the average $\bar{x}$, calculated by the formula (3):

$$\bar{x} = \frac{1}{n}\sum\nolimits_{r,j,i} x_{rij}, \tag{2}$$

$$\delta_x = \sqrt{\frac{1}{n}\sum\nolimits_{r,j,i}(x_{rij} - \bar{x})^2}. \tag{3}$$

The minimax normalization is also often used in the process of preparing medical data for the kNN- method:

$$\hat{x}_{rij} = \frac{x_{rij} - \min\limits_{r,j,i}\{x_{rij}\}}{\max\limits_{r,j,i}\{x_{rij}\} - \min\limits_{r,j,i}\{x_{rij}\}}, \tag{4}$$

The next step is to calculate the similarity measure between the medical cards. Let us denote by s_{ij}^{pq} the similarity measure of p-th and q-th medical cards for j-th record of i-th type. The similarity measure for dichotomous and ordinal data is calculated by formula (5), and for quantitative data by formula (6):

$$s_{ij}^{pq} = \begin{cases} 1, x_{pij} = x_{qij} \\ 0, x_{pij} \neq x_{qij} \end{cases}, \tag{5}$$

$$s_{ij}^{pq} = 1 - \left|\hat{x}_{pij} - \hat{x}_{qij}\right|. \tag{6}$$

It is proposed to use the Gower coefficient s^{pq} as a convolution criterion for mixed data. Denote w_{ij} as a weighted coefficient of the importance of the j-th record of i-th type. The coefficient $w_{ij} = 0$, if j-th record of i-th type is absent in the p-th or/and q-th medical cards, in other case it equals to 1.

$$s^{pq} = \frac{\sum_{i,j} w_{ij} s_{ij}^{pq}}{\sum_{i,j} w_{ij}}. \tag{7}$$

The obtained set of values of the convolution criterion s^{pq} is sorted in descending order. Next step is determination of the number of the nearest medical cards, which are base for defining the health risk group. Alternatively, expert can specify a convolution threshold value instead of the number of the nearest neighbors. The top part of the sorted list of medical cards is the set $R_k \subset R$, while $|R_k| = k$, that is, these are k-th precedents, which are the basis for determining the class of the given card. The set R_k consists of 4-th subsets K_m of medical cards according to the number of health risk groups $m = \overline{1,4}$ under guide [12]:

$$\bigcup_{m=\overline{1,4}} K_m = R_k.$$

If the set R_k contains the different values of the Gower coefficient s^{pq} for medical records, then the given case is assigned the class with the highest number of votes, since the similarity measure of each record no longer plays a role in voting. If r_p is the medical card from medical database R and $r_p(g_m)$ is m-th risk group of p-th medical card, where $g_m \in \{1, 2, 3, 4\}$, then the risk group for the current precedent r_p is:

$$r_p(g_m) = \max\left\{ \bigcup_{m=\overline{1,4}} |\{r_l, l \epsilon K_m\}| \right\}. \tag{8}$$

If the set R_k contains the same values of the mixed convolution criterion, then it is necessary to carry out a weighted vote of the relevant cases:

$$r_p(g_m) = \max\left\{ \bigcup_{m=\overline{1,4}} \left(\sum_{l \epsilon K_m} \frac{1}{(s^{pl})^2} \right) \right\}. \tag{9}$$

Thus, the model of retrieving the relevant medical records has been proposed for screening and identifying disease risk groups.

4.3. Experiments

Consider the use of the proposed methodology for determining the disease risk assessment for cardiovascular disease. According to medical protocols and clinical guidelines [10-11, 31], it is necessary to choose a scale for determining risks and to identify risk factors for determining the affiliation of a medical card to a particular health group of cardiovascular disease. Then risks presence of cardiovascular diseases should be identified.

According to medical documents [10-12], the scale of values of classes of CVD risk has four groups: $m = 1$ - low level of risk, $m = 2$ - moderate, $m = 3$ - high, $m = 4$ - very high level.

If the data is mixed data with different dimensions, or an indicator is needed that is based on several symptoms or test results, than the input data must be normalized and identified. The following set of informative features with their possible values regarding the presence of cardiovascular disease was proposed in [19]:

- x_1 – sex: x_{11} – woman, x_{12} – man;
- x_2 – age: x_{21} – less, than 40 years, x_{22} – 40-50 years, x_{23} – more, than 50 years;
- x_3 – diabetes: x_{31} – presence, x_{32} – absence, relevant data, x_{33} – absence, irrelevant data, x_{34} – no record;
- x_4 – hypertension: x_{41} – presence, x_{42} – absence, relevant data, x_{43} – absence, irrelevant data, x_{44} – no record;
- x_5 – renal dysfunction: x_{51} – presence, x_{52} – absence, x_{53} – no record;
- x_6 – tachycardia: x_{61} – presence, relevant data, x_{62} – presence, irrelevant data, x_{63} – absence, relevant data, x_{64} – absence, irrelevant data, x_{65} – no record;
- x_7 – inherited heart conditions: x_{71} – presence, x_{72} – absence, x_{73} – no record;
- x_8 – smoking: x_{81} – presence, x_{82} – absence, x_{83} – no record;
- x_9 – alcohol abuse: x_{91} – presence, x_{92} – absence, x_{93} – no record;
- x_{10} – hypodynamia: $x_{10\,1}$ – presence, $x_{10\,2}$ – absence, $x_{10\,3}$ – no record.

Let's consider the medical records of several patients with the unknown risk of cardiovascular disease. Information from medical records is presented as a set of features in Table 1.

Table 1
Input data

New records	Set of informative features of cardiovascular diseases									
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
record 1	x_{11}	x_{22}	x_{33}	x_{41}	x_{53}	x_{61}	x_{72}	x_{81}	x_{93}	$x_{10\,1}$
record 2	x_{11}	x_{21}	x_{34}	x_{43}	x_{52}	x_{63}	x_{71}	x_{82}	x_{91}	$x_{10\,1}$
record 3	x_{12}	x_{22}	x_{31}	x_{41}	x_{52}	x_{65}	x_{73}	x_{82}	x_{93}	$x_{10\,3}$

A base of precedents has been formed for determining the disease risk assessment for cardiovascular disease. The base consist of 197 medical records with confirmed health groups. According to the proposed methodology, the measures of similarity s_{ij}^{pq} were calculated between medical records from the base of precedents and new ones using formula (7). After sorting the obtained values, medical experts proposed to take 14 relevant medical records to determine the risk group. The type of input data is ordinal data (Table 1). Therefore, a large number of obtained data from calculating of the Gower coefficient has the same results. The results are presented in Table 2. In this case, to determine the class, it is necessary to use formula (9) to conduct a weighted vote of the relevant cases.

Table 2
Obtained data

Set of cards	The value of s_{ij}^{pq} with identified and verified number of the class	$r_p(g_m)$				Max of $r_p(g_m)$	Obtained class
		$r_p(1)$	$r_p(2)$	$r_p(3)$	$r_p(4)$		
Card 1	0,6(2); 0,6(2); 0,6(2); 0,6(3); 0,5(1); 0,5(2); 0,5(3); 0,5(2); 0,5(1); 0,5(2); 0,4(4); 0,3(1); 0,3(3); 0,3(2)	19,11	31,44	17,89	6,25	31,44	2
Card 2	0,7(3); 0,6(4); 0,6(3); 0,5(3); 0,5(2); 0,5(3); 0,5(4); 0,4(3); 0,4(2); 0,4(2); 0,4(4); 0,4(2); 0,4(3); 0,2(2)	0,00	47,75	25,32	13,03	47,75	2
Card 3	0,7(4); 0,7(3); 0,7(4); 0,7(4); 0,6(3); 0,6(4); 0,6(2); 0,5(1); 0,5(3); 0,5(4); 0,5(2); 0,3(3); 0,3(3); 0,3(2)	4,00	17,89	31,04	12,90	31,04	3

Analysis of the input and obtained data allows seeing that third card contains many gaps. Nevertheless, the known age and the presence of hypertension are the basis for identifying this card in a group with a high level of cardiovascular disease risk. At the same time, first and second card are filled almost completely, which allows concluding that they belong to the second health group. The obtained results provide information for medical decision-making by expert doctors in the field of cardiology regarding the early diagnosis of patients.

5. Discussion

In order to use the proposed methodology for early diagnosis of patients based on data from medical records in practice, it is necessary to prepare data and check the adequacy of the medical information system. For instance, good way to evaluate the performance of a classifier is to look at the confusion matrix [32]. There are several terms, which are the base for creating of confusion matrix: Precision, Recall, F-measure, Accuracy [32, 33]. The matrix compares the actual target values with those predicted by the kNN-model. If TP is the number of true positive results; TN is the number of true negative results; FP is the number of false positive results; FN is the number of false negative results, then we have the following. Precision is the proportion of objects classified as X that really belong to the class X:

$$Precision = \frac{TP}{TP + FP}. \tag{10}$$

Recall is the proportion of all objects of the class X classified as belonging to the class X:

$$Recall = \frac{TP}{TP + FN}. \tag{11}$$

F-measure is the harmonic mean between Precision and Recall:

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{12}$$

Accuracy is the proportion of right classified objects in the all classified objects:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \qquad (13)$$

It is difficult to compare two sets of input data with low precision and high recall or vice versa. So to make them comparable, it is proposed to use F-measure. It helps to assess Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean.

A series of experiments were carried out with different conditions. During every stage, part of medical cards with known the disease risks was considered as training pattern. The proposed methodology allows determining particular group for records from the training pattern. The results of the experiments are shown in Table 3.

Table 3
Results of classifier verifying

№	TP	FP	TN	FN	Count of cards	$Precision$	$Recall$	$F-measure$	$Accuracy$
1	19	4	13	5	41	0,83	0,79	0,81	0,78
2	10	3	11	3	27	0,77	0,77	0,77	0,78
3	19	5	8	2	34	0,79	0,90	0,84	0,79
4	14	0	9	2	25	1,00	0,88	0,93	0,92
5	27	2	17	1	47	0,93	0,96	0,95	0,94
6	9	3	15	2	29	0,75	0,82	0,78	0,83
7	12	2	33	1	48	0,86	0,92	0,89	0,94

Obtained data from above table proves the medical information system efficiency for determining the disease risks of cardiovascular diseases. Results of the confusion matrix can be the base for changing of the input data. For instance, if F-measure has low value, the number of input features is large and/or the processing time of the data by the informational system is very long, then we can use not only proposed way of screening and selection of informative features, but also the following method. Input features can be grouped by their meaning. For example, the concurrent diseases can be combined when diagnosing a specific disease. Then the numerical value of selected convolution criterion is determined for each group, as proposed in [19]. Such grouping of criteria will allow searching faster for the necessary information in the database with precedents and allow getting the best F-measure values. When training template is being used for checking the correctness of the kNN-algorithm, it is necessary to change the value of k as well, because it determines the cardinality of the set of relevant precedents. If the value of the external proximity criterion is reliably known from the expert, then the number of closest cases can be replaced by the threshold value of the similarity measure between medical records.

Therefore, aforementioned recommendations of changing the input data can help in improving the performance of medical IS. The conducted studies and the obtained results show the feasibility of using the proposed methodology in real conditions.

6. Conclusion

In the course of this study, a methodology for risks disease assessment based on the medical records has been proposed. The following results have been presented:

- the analytical review of early diagnosis methods or methods of disease risk assessment has been conducted, which allowed to choose the CBR-approach to the implementation of the task;
- the formalization of the screening process of medical cards using IDEF0-notation was presented, which allowed to choose several stages of medical data processing and combine them into a common methodology for disease risk assessment;

- the algorithm of kNN-method has been developed for identification of the groups of risks diseases;
- the model of retrieving of relevant precedents has been proposed, which allows to solve the problem;
- the numerical studies have been conducted, which shows the possibility of usage of the proposed methodology in real world settings.

The scientific novelty of the obtained results is the improvement of the process of early diagnosis with the help of the proposed methodology, which allows to process quantitative, qualitative and scaled data simultaneously from medical records for more effective medical decision-making.

7. References

[1] Definition of activity rate. URL: https://www.economicsonline.co.uk/Definitions/Activity_rate.html.

[2] Pandemiya bankrotstva: Kak koronavirus unichtozhayet krupneyshiye v mire kompanii, 2020. URL: https://112.ua/glavnye-novosti/pandemiya-bankrotstva-kak-koronavirus-unichtozhaet-krupneyshie-v-mire-kompanii-541445.html.

[3] Ukraina voshla v top-5 stran po tempam sokrashcheniya naseleniya, 2020. URL: https://delo.ua/econonomyandpoliticsinukraine/ukraina-popala-v-rejting-stran-mira-po-tempam-sokraschenija-nase-338324/.

[4] D'Agostino RB Sr, Vasan RS, Pencina MJ, et al. General Cardiovascular Risk Profile for Use in: Primary Care. The Framingham Heart Study. Circulation. 2008 Jan 22. PubMed ID: 18212285

[5] Prilozheniye pomozhet otsenit' risk razvitiya bolezni Al'tsgeymera, 2016. URL: https://evercare.ru/brainsalvation.

[6] "Heaf Test". Black's Medical Dictionary, 42nd Edition. London: A & C Black. October 2010.

[7] K. V. Mel'nik, and A. Ye. Goloskokov. “Analiz dannykh dlya meditsinskoy informatsionnoy sistemy v lechebno-profilakticheskom uchrezhdenii.” Vestnik NTU KHPI 29 (2012): 60-67.

[8] JMG Wilson, and G. Jungner (1968). "Principles and practice of screening for disease." WHO Chronicle. 22.11 (1968): 281–393.

[9] Anne Andermann, Ingeborg Blancquaert, Sylvie Beauchamp, and Véronique Déry. “Revisiting Wilson and Jungner in the genomic age: a review of screening criteria over the past 40 years.” Bulletin of the World Health Organization 86.4 (2008): 241-320.

[10] Unifikovanyy klinichnyy protokol pervynnoyi, vtorynnoyi (spetsializovanoyi) ta tretynnoyi (vysokospetsializovanoyi) medychnoyi dopomohy. Profilaktyka sertsevosudynnykh zakhvoryuvan′, 2016. URL: https://dec.gov.ua/wp-content/uploads/2019/11/2016_564_ykpmd_pssz.pdf.

[11] Adaptovana klinichna nastanova «Profilaktyka sertsevo-sudynnykh zakhvoryuvan′. Onovlena ta adaptovana klinichna nastanova, zasnovana na dokazakh», 2016. URL: http://www.volyncard.in.ua/files/protocols/2016_564_AKN_PSSZ.pdf.

[12] Rapid Risk Assessment of Acute Public Health Events, 2012. URL: https://apps.who.int/iris/bitstream/handle/10665/70810/WHO_HSE_GAR_ARO_2012.1_eng.pdf?sequence=1.

[13] T. V. Ovchinkina, V. V. Mitin, and A. A. Kuz'min. “Primeneniye gibridnykh neyronnykh setey v prognosticheskikh modelyakh otsenki funktsional'nogo sostoyaniya serdechno-sosudistoy sistemy.” Sovremennyye problemy nauki i obrazovaniya 5 (2013). URL: http://www.science-education.ru/ru/article/view?id=10551.

[14] A. B. Kraskovskiy, A. V. Nosov, and O. V. Shatalova. “Gomeostaticheskiye modeli vliyaniya psikhoemotsional'noy napryazhennosti na risk psikhosomaticheskikh zabolevaniy.” Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskiye nauki, 110.9 (2010): 17-21. URL: https://cyberleninka.ru/article/n/gomeostaticheskie-modeli-vliyaniya-psihoemotsionalnoy-napryazhennosti-na-risk-psihosomaticheskih-zabolevaniy/viewer.

[15] K. V. Mel'nik, and S. Í. Êrshova. “Problemy i osnovnyye podkhody k resheniyu zadachi meditsinskoy diagnostiki.” Sistemi obrobki ínformatsíї 2.92 (2011): 244-248.

[16] G. A. Dmitriyev, and Al'-Fakikh Ali Salekh Ali. “Sistema diagnostiki i otsenki riska osteoporoticheskogo pereloma na osnove intellektual'nogo analiza dannykh.” Programmnyye produkty i sistemy 3.115 (2016): 208-212. DOI:10.15827/0236-235X.115.208-212. URL: https://cyberleninka.ru/article/n/sistema-diagnostiki-i-otsenki-riska-osteoporoticheskogo-pereloma-na-osnove-intellektualnogo-analiza-dannyh.

[17] K. V. Mel'nik, and V. N. Glushko. “Primeneniye apparata Bayyesovykh setey pri obrabotke dannykh iz meditsinskikh kartochek.” Science and Education a New Dimension: Natural and Technical Sciences. I(2).15 (2013): 126-129.

[18] M. F. Bondarenko, YU. P. Shabanov-Kushnarenko Teoriya intellekta. – Khar'kov: SMIT, 2007. – 576 s.

[19] K. V. Melʹnyk. “Modelyuvannya protsesu intelektualʹnoyi obrobky medychnykh danykh.” Systemy obrobky informatsiyi 4.150 (2017): 237-244.

[20] Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov. “Data mining in healthcare and biomedicine: a survey of the literature.” Journal of medical systems 36.4 (2012): 2431–2448.

[21] O. S. Zharkova, K. A. Sharopin, A. S. Seidova, Berestneva Ye.V., and Osadchaya I.A. “Postroyeniye sistem podderzhki prinyatiya resheniy v meditsine na osnove derev'yev resheniy.” Sovremennyye naukoyemkiye tekhnologii 6-1 (2016): 33-37. URL: http://www.top-technologies.ru/ru/article/view?id=35973.

[22] Nabanita Choudhury, and Shahin Ara Begum. “A Survey on Case-based Reasoning in Medicine” International Journal of Advanced Computer Science and Applications 7.8 (2016): 136-144. URL: https://thesai.org/Downloads/Volume7No8/Paper_20-A_Survey_on_Case_based_Reasoning_in_Medicine.pdf.

[23] R. Schmidt, and L. Gierl. “Case-based Reasoning for Medical Knowledge-based Systems.” Studies in health technology and informatics 77 (2000): 720-7255. DOI: 10.3233/978-1-60750-921-9-720. URL: https://folk.idi.ntnu.no/agnar/publications/medical-cbrws.pdf.

[24] A. Aamodt, and E. Plaza. “Case-based reasoning: foundational issues, methodological variations, and system approaches.” AI Communications 7.1 (1994): 39-59.

[25] Hugo Bowne-Anderson, Preprocessing in Data Science (Part 1): Centering, Scaling, and KNN, 2016. URL: https://www.datacamp.com/community/tutorials/preprocessing-in-data-science-part-1-centering-scaling-and-knn.

[26] K. V. Mel'nik, and A. Ye. Goloskokov. “Zadacha planirovaniya skriningovykh meropriyatiy.” Problemy informatsionnykh tekhnologiy 14 (2013): 60-68.

[27] Rinu Gour, Dimensionality Reduction in Machine Learning, 2019. URL: https://medium.com/@rinu.gour123/dimensionality-reduction-in-machine-learning-dad03dd46a9e.

[28] Judy T Raj, A beginner’s guide to dimensionality reduction in Machine Learning, 2019. URL: https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e.

[29] Brian S. Everitt, Sabine Landau, and Daniel Stahl. Cluster Analysis. Wiley, 2011.

[30] K. V. Mel'nik. “Primeneniye algoritma kollaborativnoy fil'tratsii dlya obrabotki meditsinskikh dannykh.” Vestnik NTU "KHPI" 2.1111 (2015): 193-198.

[31] Sertsevo-sudynni zakhvoryuvannya. Klasyfikatsiya, standarty diahnostyky ta likuvannya, 2007. URL: http://ukrcardio.org/wp-content/uploads/2015/10/Recomendations-UAKSSZ.pdf.

[32] K. M. Ting, Confusion Matrix. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA., 2011. https://doi.org/10.1007/978-0-387-30164-8_157 https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_157.

[33] Everything you Should Know about Confusion Matrix for Machine Learning, 2020. URL: https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/.