Application of medical data classification methods for a medical decision support system

Ekaterina Yu. Zimina[1][0000-0002-8625-1956], Maxim A. Novopashin[2][0000-0002-8919-4002] and Alexander V. Shmid [1][0000-0002-4672-1458]

[1] National Research University Higher School of Economics, 11, Pokrovsky Boulevard, Moscow, 101000, Russian Federation
[2] EC-leasing Company, 125, Varshavskoe highway, Moscow, 117587, Russian Federation
ezimina@hse.ru

Abstract. Decision support systems (DSS) allow us to help the doctor in making diagnoses to the patient, also medical DSS help to assess the need for a particular examination of the patient. In this article methods of medical data classification are considered, these methods are the part of the medical DSS. The paper includes investigation of data classification methods as hierarchical cluster analysis, k-means analysis and discriminant analysis. The selected methods are implemented using the example of cardiological data. A hypothesis is put forward that it is possible to determine the presence or absence of tuberculosis in a person from cardiological data by using data classification methods. Such indicators as sensitivity and specificity evaluate the effectiveness of the methods. In addition, ROC and AUC are presented. Thus, the DSS will be able to determine a certain degree of probability to assume the presence of tuberculosis in a person. The doctor will decide on the need for additional examinations depending on the values obtained,

Keywords: Decision Support System, Data Analysis, Telemedicine, Classification.

1 Introduction

Currently, the creation of decision support systems (DSS) is relevant, and this direction is also developing in the field of medicine. DSS allow us to help the doctor in making diagnoses to the patient. In addition, with the help of these systems, it is possible to determine the need for various examinations for the patient [1]. The use of the medical DSS for doctors will prevent patients from being sent to expensive additional examinations, which are not always safe [2].

The paper discusses methods of data analysis that will be implemented in the medical DSS in order to help doctors. The paper implements such methods as hierarchical cluster analysis, k-means analysis and discriminant analysis.

The parameters of electrocardiogram (ECG) were used as experimental data. This data is depersonalized.

With implementing the methods of medical DSS a hypothesis is put forward about the possibility of predicting the presence or absence of tuberculosis by ECG parameters. The sample contains a nominal variable (tb), which reflects the presence of diagnosed tuberculosis in a person (tb = 1) or its absence (tb = 0). The experimental data collected ECGs recorded in people with a confirmed form of tuberculosis in the second stage of the disease.

Currently ETU “LETI” under the leadership of Professor Kalinichenko A. N. is doing the similar studies. However, investigations of ETU “LETI” have a direction different from this work. They research the detection of signs of cardiac disease in ECG using machine learning [3].

The sample used in the study is divided into training and test samples, where the mathematical model is created on the training one, and the quality of the obtained model is evaluated on the test one. As a result, a model and an accuracy value of the correct prediction of belonging to the group are obtained for each of the considered methods.

An approach with training of DSS methods based on medical data will make it possible to make an early diagnosis of the patient's health condition. This means that it is possible to assume with a certain degree of probability that a person has signs of tuberculosis or not according to the recorded ECG. If possible signs of the disease are detected, this patient should be sent to get a more detailed examination together with a pulmonologist.

The purpose of this work is to implement classification methods to the medical SPR for early diagnosis to determine the presence or absence of tuberculosis signs. In accordance with this goal, the following tasks were identified: to compile descriptive statistics of the initial experimental data, to investigate and apply methods for classification on experimental data, and to formulate a conclusion.

The performance of the methods is evaluated using sensitivity and specificity indicators. Sensitivity is the percentage of correctly classified "ill" people, and specificity is the percentage of correctly classified "healthy" people [4]. In addition, a ROC is constructed for the results of the methods and the area under the curve (AUC) is calculated. The ROC curve is a tool for assessing diagnostic ability, representing a graph where the sensitivity and specificity values in the range from 0 to 1 are taken as axes [5].

2 Materials and methods

2.1 Materials

General information of the person and parameters of his cardicycle were taken to study the methods of data classification and subsequent verification of the proposed hypothesis about the possibility of determining the presence or absence of tuberculosis in a person from cardiological data.

A cardiocycle (or cardiac cycle) is a period of blood circulation generated by the cyclic activity of the heart. The measurement unit for this periodicity is one cardiac cycle. The length of the cardiocycle is the period of cardiac contractions [6]. The elements of the cardiocycle are presented below (Fig. 1).

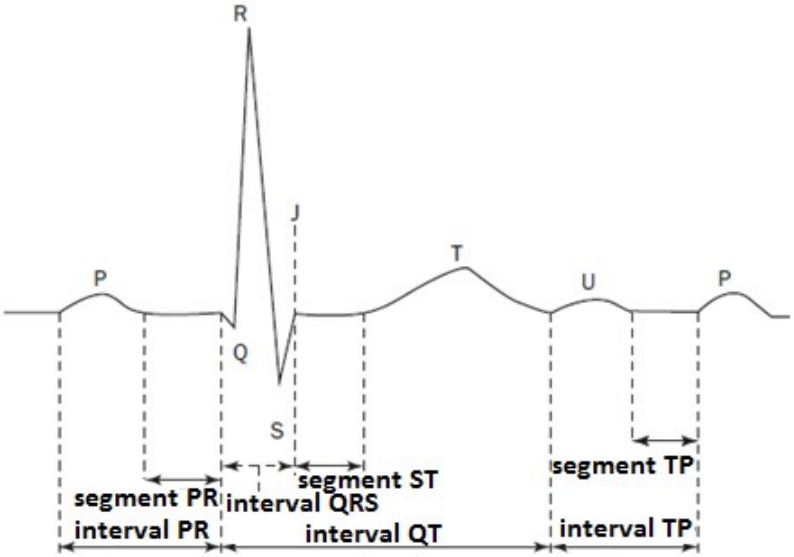

Fig. 1. The elements of the cardiac cycle.

The main elements for the ECG analysis are the start and end time of the elements of the cardiocycle, as well as the PQ interval, the QRS complex, the ST segment, the QT interval and the P wave stands out especially among the elements of the cardiocycle.

The data is presented in the form of a table, where each row corresponds to one ECG. Also in the same row in columns contains non-personally-identifying information about the person. Below is a table with parameters for analysis and explanations to them (Table 1).

The total sample consists of 5928 registered ECGs. Below is a table with general information about people from the data sample (Table 2).

The distribution of the number of people according to their age classification is also presented (Fig. 2). In the sample by age, there is a bias towards people over 18 years old. This is explained by the fact that ECG registration of persons under 18 years old is possible in the presence of a parent and with his permission, so there were few persons of younger groups in the collected data.

Table 1. Data parameters for analysis.

No	The name of parameter	Comments
1	pid	Patient identification number
2	cid	Cardiogram identification number
3	date1	Date of registration of the ECG
4	gender	Gender (1 – M, 0 – W)
5-7	age, weight, height	Age, weight, height
8	cardiostimulator	Presence of pacemaker (1 – yes, 0 – no)
9	smoking	Smoking (1 - yes, 0 – no)
10	Tb	Presence of diagnosed tuberculosis (1 – yes, 0 – no)
11-13	p_a, p_da, p_t	Parameters of P wave
14-15	p_left_slopes, p_right_slopes	The length of the slopes of P wave
16-18	q_a, q_b_t, q_e_t	Parameters of Q wave
19-21	r_a, r_b_t, r_e_t	Parameters of R wave
22-23	r_left_slopes, r_right_slopes	The length of the slopes of R wave
24-27	s_a, s_da, s_b_t, s_e_t	Parameters of S wave
28-30	t_a, t_da, t_t	Parameters of T wave
31-32	t_left_slopes, t_right_slopes	The length of the slopes of T wave
33	interval_pq	The length of interval PQ
34	komplex_qrs	The length of complex QRS
35	segment_st	The length of segment ST
36	interval_qt	The length of interval QT
37	zubets_p	The length of P wave

Where: X_a – the amplitude (height) of the figure X on the ECG, X_da – the amplitude (height) indicator X on the differentiated ECG (ECG is taken first production), X_t – length index of X, X_b_t start time of metric X on ECG, X_e_t – the end time of the index X on ECG.

Table 2. The number of people.

	All	With tuberculosis	Without tuberculosis	With pacemaker	Without pacemaker	Smoking	No smoking
Men	329	70	259	37	292	121	208
Women	262	66	196	25	237	59	203
All	591	136	455	62	529	180	411

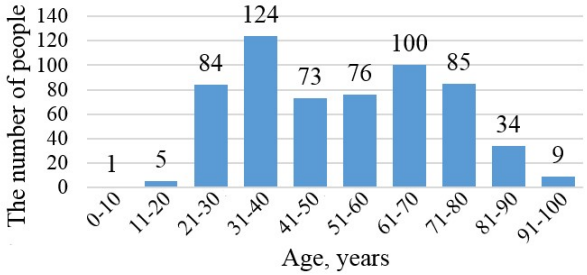

Fig. 2. Distribution of the number of people by age.

The following table includes the main values of the number of ECGs for different groups (Table 3).

Table 3. The number of ECG.

	All	With tuberculosis	Without tuberculosis	With pacemaker	Without pacemaker	Smoking	No smoking
Men	3546	573	2973	776	2770	1945	1601
Women	2382	684	1698	115	2267	972	1410
All	5928	1257	4671	891	5037	2917	3011

It should be noted that not all parameters of the cardiocycle were calculated for all ECGs, so observations with partially uncalculated parameters were automatically discarded methods implementation.

2.2 Methods

This section describes methods of data analysis and provides brief information on them [7].

Cluster analysis is used to separate the original data into groups (clusters) that are amenable to interpretation so that the elements of one group were similar in the parameters, while elements from different groups should differ from each other [8] (Fig. 3).

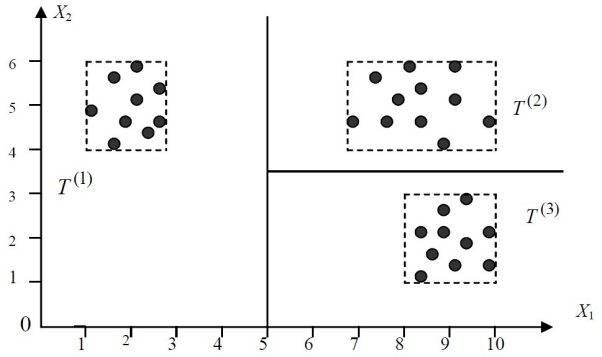

Fig. 3. An example of the points separation in the plane into similar clusters.

Hierarchical cluster analysis is used for relatively small numbers of observations. During the analysis, initially each observation is located in its own cluster, then neighboring clusters are combined in pairs until there are only two clusters left.

K-means analysis allows you to divide an arbitrary data set into a given number of groups so that the objects of the same cluster are close enough to each other, and the objects of different groups do not intersect [9]. In this case, the observation belongs to the cluster to the center of which it is the closest.

First, the center of the class is determined, then all objects within the specified threshold value from the center are grouped.

Discriminant analysis is a method of statistical analysis that allows you to divide data into disjoint groups. This method allows us to identify the variables that affect the separation, as well as their weight coefficients [10]. The result of performing a discriminant analysis is a discriminant function that uses a nominal dependent variable. Discriminant analysis is an alternative to multiple regression analysis.

3 Results

IBM SPSS Statistics 23 software was used in order to implement selected methods. IBM SPSS Statistics is a statistical analysis platform with a set of functions [11].

The use of the hierarchical cluster analysis method did not lead to significant results. The number of observations in 5928 recorded ECGs was too large as a sample for this method.

Further, the number of observations in the sample was reduced to 50% of randomly selected observations. As an assumption, a range was set for the number of classes: there should be 2-3 clusters. This was done because a huge number of clusters are obtained on this sample without this restriction. The values in two or three clusters were chosen based on the following: we need to get two clusters with measurements of people without tuberculosis and with tuberculosis. The possible number of three clusters is taken to compare the results.

The model was built, its variables were saved with the indication of belonging to the cluster. A conjugacy table was constructed for two variables containing

information about the distribution of all variables into 2 and 3 clusters in order to evaluate the performance of this method (Figure 4).

Average Linkage (Between Groups) * Average Linkage (Between Groups)

Quantity

		Average Linkage (Between Groups)		
		1	2	**Total**
Average Linkage (Between Groups)	1	2331	0	2331
	2	0	21	21
	3	0	9	9
Total		2331	30	2361

Fig. 4. A distribution of observations by division into 2 and 3 clusters.

According to the table almost all observations fell into the first cluster, it also observed when the data divided into two clusters and into three clusters. It is also seen that the second and third clusters in both divisions are very small relative to the first cluster. If we compare the decision to divide into two clusters or three, we can conclude that the two-cluster solution is the most stable.

Further, for the sake of clarity of the obtained solution an analysis of the averages was carried out, the part of the resulting picture is presented below (Fig. 5).

Average Linkage (Between Groups)	age	gender	weight	height	cardiosti mulator	smoking	tb	p_a	p_da
1	47,90	,56	76,05	173,29	,07	,45	,24	,006954	,000010
2	44,83	,47	138,13	61,23	,10	,83	,73	,004041	,000008
Всего	47,86	,56	76,84	171,87	,07	,45	,24	,006917	,000010

Fig. 5. The average values when divided into two clusters.

The target variable – the variable of the presence or absence of tuberculosis in this dimension (tb). During the application of hierarchical clustering it was found that the average values of clusters are 0.24 and 0.73, where 0 is “healthy” and 1 is “sick”. You can also pay attention to other parameters, for example, taller people fell into the “healthy” cluster, and almost all smokers fell into the “sick” cluster. It is worth noting the average weight of the subjects in the second cluster – 138 kg, which is quite a lot.

When analyzing this model in detail we conclude that hierarchical clustering is not suitable for working with this sample of medical data.

When implementing the k-means method two clusters are initially set: people without a diagnosis of tuberculosis and people with diagnosed tuberculosis (parameter tb=0 and tb=1, respectively). There are many observations in the medical dataset, so 10 iterations were set for the method to work.

The centers of the two clusters obtained are presented below (Fig. 6).

Final centers of clusters

	Clusters	
	1	2
age	48	45
gender	1	0
weight	76	138
height	173	61
cardiostimulator	0	0
smoking	0	1
tb	0	1
p_a	,0069536787	,0040413144
p_da	,0000102624	,0000080664

Fig. 6. Cluster centers in k–means clustering.

We also obtained an estimate based on Fischer statistics on the significance of the parameter in the differentiation of clusters (Fig. 7).The figure below reveals an example that shows that the target variable tb is significant, as is weight, height, and smoking. The most significant parameters among the parameters of the cardiocycle are the R wave, S wave and the QRS complex.

In addition, the k-means method obtained results is similar to the hierarchical clustering method: outputs data on the number of observations in clusters are 2331 observations in the first cluster and 30 observations in the second cluster. The obtained values coincided with the values for the number of observations when dividing into two clusters during hierarchical clustering. These calculations were obtained by randomly selecting 50 % of all observations.

When using the k-means method with the same parameters on a full sample the following division was obtained by the number of observations in clusters: 4707 and 56 observations, respectively. Thus, the result was obtained that one cluster is dominated by data when clustering into two groups.

Two additional parameters were created using of the k-means clustering method: indicating the number of the membership cluster and the distance to its center.

Next a graphical illustration of the results of this method was constructed: the grouping variable is the cluster number, the differentiating variable is the distance to the cluster center. The figure below, as well as the line on the cluster, shows the median value (Fig. 8).

	Cluster		Error			
	The average square	St. sv.	The average square	St. sv.	F	Value
age	278,089	1	277,172	2359	1,003	,317
gender	,279	1	,246	2359	1,133	,287
weight	114171,290	1	382,013	2359	298,868	,000
height	371908,687	1	107,397	2359	3462,948	,000
cardiostimulator	,035	1	,062	2359	,566	,452
smoking	4,440	1	,246	2359	18,053	,000
tb	7,315	1	,181	2359	40,447	,000
p_a	,000	1	,000	2359	1,085	,298
p_da	,000	1	,000	2359	,020	,887
p_t	,001	1	,001	2359	,557	,455
p_left_slopes	,000	1	,000	2359	,538	,463
p_right_slopes	,000	1	,000	2359	1,300	,254
q_a	,000	1	,000	2359	,254	,614
q_b_t	,000	1	,000	2359	,837	,360
q_e_t	,000	1	,000	2359	,010	,918
r_a	,042	1	,002	2359	17,760	,000
r_left_slopes	,000	1	,000	2359	17,571	,000
r_right_slopes	,000	1	,000	2359	15,852	,000
r_b_t	,001	1	,000	2359	5,096	,024
r_e_t	,000	1	,000	2359	,392	,532
s_a	,001	1	,000	2359	2,352	,125
s_da	,000	1	,000	2359	7,366	,007
s_b_t	,001	1	,000	2359	5,910	,015
s_e_t	,001	1	,000	2359	3,275	,070
t_a	,000	1	,001	2359	,287	,592
t_da	,000	1	,000	2359	,054	,817
t_t	,006	1	,001	2359	5,108	,024
t_left_slopes	,000	1	,000	2359	,268	,604
t_right_slopes	,000	1	,000	2359	,355	,551
interval_pq	,001	1	,001	2359	1,025	,312
komplex_qrs	,002	1	,000	2359	6,241	,013
segment_st	,011	1	,001	2359	8,614	,003
interval_qt	,007	1	,002	2359	3,239	,072

Fig. 7. Values of Fisher statistics.

Based on the results of the analysis it can be concluded that the parameters of the R wave were the most significant parameters in clustering by this method. Below is the spread of the R wave indicator depending on the presence or absence of tuberculosis (Fig. 9).

The figure shows that the variations of this indicator differ depending on the presence or absence of tuberculosis, but visually almost half of the values of the indicator are the same both in the presence of tuberculosis and in its absence. According to the results of the clustering analysis the k-means indicator is the most

significant when divided into groups. This leads to the conclusion that the model is not sufficiently accurate using the k-means method.

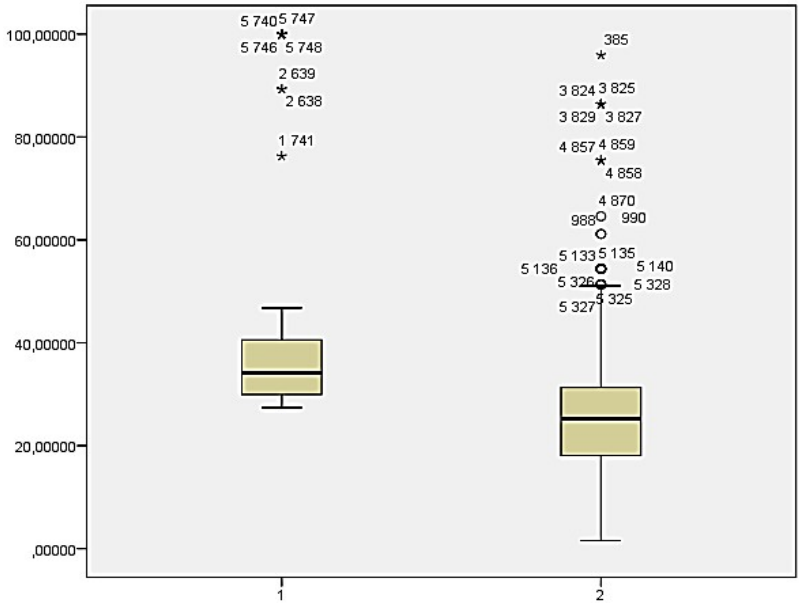

Fig. 8. The vertical axis - distance from observation to cluster center, the horizontal axis is the cluster number.

Then the discriminant method was implemented and investigated. The sample was first divided into training and test samples: 60% and 40%, respectively [12-13]. For implementation the method of forced inclusion of variables was used and grouping was performed by the variable of the presence or absence of tuberculosis tb.

A table "Group statistics" was obtained with an indication of the average values of each parameter, its standard deviation by group. The inequality of the mean and standard deviation does not prove that these variables are distinctive features of the selected clusters.

The figure below shows the calculated values of the variables. Parameters whose values in the table are greater than 0.05 can be excluded later for analysis purposes.

From the figure below it can be seen that there are parameters that are insignificant when divided into groups, for example, p_da, t_da and others. Thus, they can be excluded when composing the equation (Fig. 10).

The coefficients of the canonical discriminant function were also obtained to create the equation (Fig. 11).

The accuracy of the division into clusters is determined by the distance between the average values of the discriminant function in the studied clusters. The greater the distance, the better the groups are separated. The values of the centroids of the groups are as follows: -0.376 and 1.242.

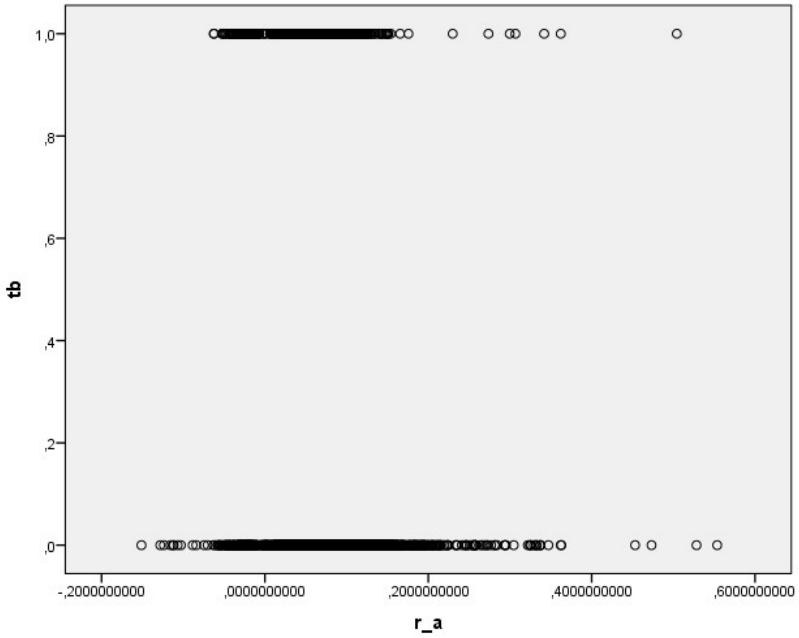

Fig. 9. The scatter plot of figure R wave and tb.

You can determine the quality of the model based on the results of the classification at the following table (Table 4).

Table 4. The results of classification discriminant analysis.

		Tb	Predicted group membership		Total
			0	1	
Selected observations	Quantity	0	1826	389	2215
		1	200	500	700
	%	0	82.4	17.6	100
		1	28.6	71.4	100
Unselected observations	Quantity	0	1195	243	1438
		1	117	293	410
	%	0	83.1	16.9	100
		1	28.5	71.5	100

In the training sample the sensitivity is 71.4% and the specificity is 82.4%. In the test sample the sensitivity is 71.5% and the specificity is 83.1%. This shows good accuracy of this model.

In addition, a ROC curve was constructed, the area under the curve of which was 0.853 (Fig. 12).

Criteria for equality of group averages

	Wilkes lambda	F	St. sv. 1	St. sv. 2	Value
age	,997	7,442	1	2822	,006
gender	,987	38,083	1	2822	,000
weight	,921	241,124	1	2822	,000
height	,966	98,803	1	2822	,000
cardiostimulator	,980	57,919	1	2822	,000
smoking	,944	167,686	1	2822	,000
p_a	,990	29,447	1	2822	,000
p_da	1,000	,693	1	2822	,405
p_t	,993	19,690	1	2822	,000
p_left_slopes	,997	9,430	1	2822	,002
p_right_slopes	,989	30,454	1	2822	,000
q_a	,999	2,869	1	2822	,090
q_b_t	,998	6,710	1	2822	,010
q_e_t	,996	11,784	1	2822	,001
r_a	,938	186,799	1	2822	,000
r_left_slopes	,942	174,480	1	2822	,000
r_right_slopes	,937	189,736	1	2822	,000
r_b_t	,993	20,350	1	2822	,000
r_e_t	,993	20,936	1	2822	,000
s_a	,988	33,700	1	2822	,000
s_da	,994	16,531	1	2822	,000
s_b_t	,987	36,569	1	2822	,000
s_e_t	,998	5,603	1	2822	,018
t_a	,991	24,402	1	2822	,000
t_da	1,000	,347	1	2822	,556

Fig. 10. Evaluation of the significance of the parameter in the distribution into groups.

Using the table with the coordinates of the curve points the threshold value for the final discriminant equation 0.4511434 was selected. At this threshold the sensitivity is 76.4% and the specificity is 76.5%.

The threshold value was selected from the points of the coordinate ROC. The sensitivity and specificity values were selected so that the sum of sensitivity and specificity was the maximum.

Coefficients of the canonical discriminant function

	Function 1
age	,001
gender	,513
weight	-,033
height	-,032
cardiostimulator	-2,337
smoking	1,303
p_a	-38,099
p_da	739,055
p_t	1,385
p_left_slopes	1608,659
p_right_slopes	624,581
q_a	9,906
q_b_t	-18,702
q_e_t	20,070
r_a	-2,008
r_left_slopes	54,939
r_right_slopes	79,316
r_b_t	-8,807
r_e_t	17,468
s_a	2,800
s_da	47,839
s_b_t	35,761
s_e_t	-31,883
t_a	14,026
t_da	542,249
t_t	-11,077
t_left_slopes	-157,506
t_right_slopes	620,090
interval_qt	4,877
zubets_p	-2,906
Constant	7,247

Fig. 11. Coefficients of the discriminant function.

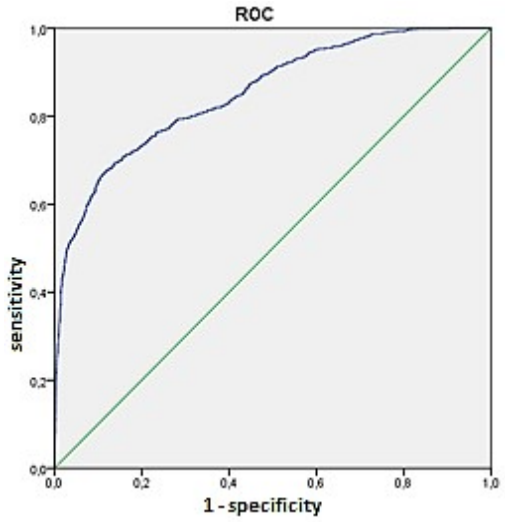

Fig. 12. ROC in discriminant analysis.

4 Discussion

Three methods were implemented: hierarchical cluster analysis, k-means analysis, and discriminant analysis.

Analysis of the hierarchical cluster method showed that this method is not suitable for analyzing large datasets. Even with the usage of reduction in the number of observations in the sample it was not possible to obtain acceptable results.

Analysis of the k-means method showed that this method can be used for classification problems into two clusters "sick" and "healthy", but the accuracy of this method did not show high results. The parameters that were identified by the method as the most significant have not significant differences in the spread between "healthy" and "sick". For more accurate operation of this method it is necessary to filter out the least significant parameters and continue a more detailed study.

The discriminant method analysis allowed us to obtain a discriminant equation with sensitivity and specificity values of 76.4% and 76.5% respectively. Based on the selected sensitivity and specificity values, a threshold was selected for working with the discriminant equation.

In order to implement the best method in terms of sensitivity and specificity in medical DSS it should be tested on a larger sample size. Also for better accuracy in predicting the probability of a person having second-stage tuberculosis, cross-validation should be performed.

5 Conclusion

In this paper several classification methods for working in the medical DSS were investigated. The idea of creating a medical DSS is as follows: according to the ECG parameters the trained methods determine the degree of probability of the presence of tuberculosis of the second type in the examined person, whose open symptoms are practically not observed. Thus, the system will help in the early stages of the disease to determine the presence of it on the ECG.

Medical data were used as experimental data in order to train DSS methods. Medical data includes parameters calculated from an electrocardiogram, as well as general impersonal parameters about its owner (height, weight, age, etc.). The experimental sample consisted of more than five thousand electrocardiograms. Also a hypothesis was put forward and tested about the possibility of determining the presence or absence of signs of tuberculosis in a person by the parameters of an electrocardiogram. This hypothesis was confirmed.

Three methods were implemented: hierarchical cluster analysis, k-means analysis, and discriminant analysis. The program for statistical data processing IBM SPSS Statistics was used to carry out the work.

Of the methods considered in this paper the most suitable for the classification problem with a nominal target variable on the example of the study of medical experimental data was the method of discriminant analysis. This method is similar to regression analysis, which will be studied further, as well as other classification methods that are not included in this work. In the future the model should be refined to obtain a higher accuracy of the medical DSS.

References

1. Sim, I., et al.: Clinical decision support systems for the practice of evidence-based medicine. Journal of the American Medical Informatics Association, 8.6, 527-534 (2001).
2. Golub, J., et al.: Delayed tuberculosis diagnosis and tuberculosis transmission. The international journal of tuberculosis and lung disease, 10.1, 24-30 (2006).
3. Nemirko, A., Manilo, L., Kalinichenko, A.: Intellectual analysis of biomedical signals. Biotekhnosfera Journal 2(20), 30-37 (2012).
4. Parikh, R., et al.: Understanding and using sensitivity, specificity and predictive values. Indian journal of ophthalmology, 56.1, 45-50 (2008).
5. Fawcett, T.: An introduction to ROC analysis. Pattern recognition letters, 27.8, 861-874 (2006).
6. Lu, B. I. N., et al.: Coronary artery motion during the cardiac cycle and optimal ECG triggering for coronary artery imaging. Investigative radiology, 36.5, 250-256 (2001).
7. Ott, R. Lyman, Longnecker, M. T.: An introduction to statistical methods and data analysis. 7th edn., Brooks/Cole, Boston, United States (2015).
8. Kaufman, L., Rousseeuw P. J.: Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York, United States (2009).
9. Kanungo, T., et al.: An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24.7, 881-892 (2002).
10. McLachlan, G. J.: Discriminant analysis and statistical pattern recognition. John Wiley & Sons, New York, United States (2004).
11. Field A.: Discovering statistics using IBM SPSS statistics. 4th edn., SAGE Publications Ltd., London (2013).
12. Breiman L.: Bagging predictors. Machine Learning, 24, 123-140 (1996).
13. Fletcher, G., Ades, P., Kligfaild, P.: Exercise standards for testing and training: a scientific statement from the American Heart Association. Circulation Journal, 128, 873-934 (2013).