DOI 10.1007/s10559-023-00578-y
Cybernetics and Systems Analysis, Vol. 59, No. 3, May, 2023

EVALUATING THE USEFULNESS OF BINARY CLASSIFIER BASED ON ENHANCED ROC ANALYSIS

O. A. Zhukovska[1] **and L. S. Fainzilberg**[2]

UDC 004.891.3

Abstract. *The definition of the usefulness of the binary classifier from the point of view of reducing the a priori risk of false classification is formulated. Sufficient conditions are proposed to guarantee the utility of a diagnostic test according to this definition. The obtained conditions improved the traditional ROC analysis by limiting the corresponding region of the ROC curve. The line limiting the region of the guaranteed useful test is shown to coincide with the known iso-performance line corresponding to the a priori risk level. The feasible limits of the ratio of losses from target misses and false alarms were determined, according to which a test with appropriate operational characteristics remains useful for screening a disease with a known prevalence. Based on the obtained results, the authors substantiated the efficiency of the new method of the analysis and interpretation of electrocardiograms, which is based on determining the original diagnostic feature in the phase space and enables detecting persons with a high risk of coronary heart disease in the early stages of the disease.*

Keywords: *binary classifier, ROC curve, diagnostic feature, analysis and interpretation of ECG.*

INTRODUCTION

Binary classifiers are widely used in various fields. Thus, for example, for the prevention of diseases, it is necessary to carry out mass examinations of the population (screening) to identify patients with a high risk of potentially dangerous diseases [1]. In the banking sector, scoring systems [2] are actively used that ensure the detection of unreliable borrowers. The list of such examples can be continued.

Various approaches to evaluating the efficiency of binary classifiers are known [3]. A convenient tool is a method based on the analysis of the so-called operational characteristic curve (Receiver Operating Characteristic curve, ROC) [4, 5]. The traditional ROC analysis is based on the evaluation of two characteristics of a diagnostic test, namely, sensitivity and specificity [6–15], which, in fact, evaluate the probabilities of miss-target and false-alarm errors adopted in the theory of statistical decisions. For the integral evaluation of the efficiency of the binary classifier implementing the diagnostic decision rule, the area under the ROC curve is most often determined [16].

In scientific works [17–21], the development of ROC analysis is proposed for the case when the set of recognized classes contains more than two diagnoses. An improvement of the ROC analysis method for comparing binary classifiers with respect to the expected average losses from wrong decisions is proposed in [22–24].

The purpose of this article is to further improve the ROC analysis method to ensure the selection of a diagnostic test that is guaranteed to reduce the a priori risk of misdiagnosis.

[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine, *zhukovskaya71@gmail.com*. [2]International Scientific and Training Center of Information Technologies and Systems, National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine, Kyiv, Ukraine, and National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," Kyiv, Ukraine, *fainzilberg@gmail.com*. Translated from Kibernetyka ta Systemnyi Analiz, No. 3, May–June, 2023, pp. 95–105. Original article submitted January 3, 2023.

TABLE 1. Results of Testing on a Representative Group of Patients

Current Diagnosis	**The Decision Made Based on the Test Result**	
	A Patient is "Sick"	A Patient is "Healthy"
The "Sick" (class V_1)	The number of true positive results TP	The number of false negative results FN
The "Healthy" (class V_2)	The number of false positive results FP	The number of true negative results TN

TRADITIONAL ROC ANALYSIS

Let us consider the basic details of the traditional ROC analysis [4–15], which are needed in further research. To do this, we let us consider the problem of screening, which comes down to identifying sick people in a large group of examinees where there are also sick (class V_1), and conventionally healthy (class V_2) persons who do not have the analyzed disease.

Let a binary classifier be used to identify sick persons, which, based on the available information, makes one of two decisions, namely, that the patient is either "Sick" (positive test result) or "Healthy" (negative test result). To evaluate the efficiency of such a test, it is tested on a representative group of patients with previously known diagnoses. The test results can be presented in the form of a matrix (Table 1) whose cells indicate the number of relevant observations.

According to the data presented in Table 1, it is easy to evaluate the operational characteristics of the test accepted in medical diagnostics, namely, sensitivity

$$S_E = \frac{TP}{TP + FN}, \tag{1}$$

which determines the proportion of true positive results TP obtained for all sick persons $TP + FN$, and specificity

$$S_P = \frac{TN}{TN + FP}, \tag{2}$$

which determines the proportion of true negative results TN obtained for all healthy individuals $TN + FP$.

The binary classifier with fixed operating characteristics (1) and (2) represent a point in the ROC space with coordinates S_E and $1 - S_P$ [4]. Note that the value

$$1 - S_P = 1 - \frac{TN}{TN + FP} = \frac{TN + FP - TN}{TN + FP} = \frac{FP}{TN + FP}$$

determines the part of false positive results obtained for all healthy individuals $TN + FP$.

The considered ROC space makes it possible to graphically demonstrate the diagnostic value of the test and compare the efficiency of different tests [22].

The ideal test A (Fig. 1) is located at the point with coordinates (0, 1). According to the result of this test, all sick persons are assigned to a class V_1, and there are no false alarm errors. It follows that the closer the test with operational characteristics S_E and S_P to the point A is, the more efficient it is.

The test B (see Fig. 1), which is located in the lower left corner, is called "conservative," because with a small percentage of false alarm errors $1 - S_P$, it has low sensitivity S_E. Since the test B has a high specificity value S_P, the decision-making where the patient is "Sick" should be taken with a high degree of confidence. At the same time, the decision that the patient is "Healthy" may be incorrect due to low sensitivity S_E.

The test C (see Fig. 1), which is located in the upper right corner, is called "liberal," because at large sensitivity values S_E, it has a high percentage of false alarm errors $1 - S_P$. Due to high sensitivity S_E, the decision that patient is "Healthy" is more likely to be correct, while the decision that the patient is "Sick" may be wrong due to low specificity S_P.

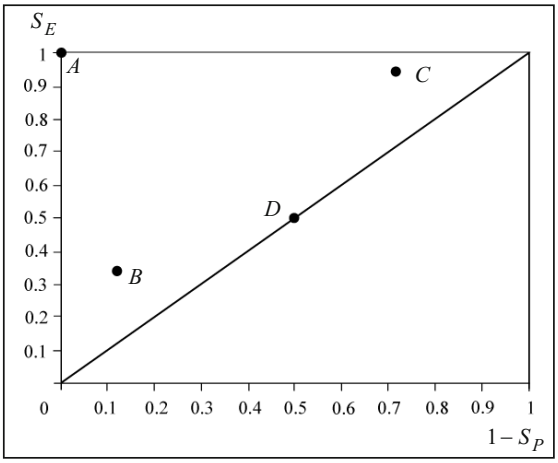

Fig. 1. Binary classifiers in the ROC space.

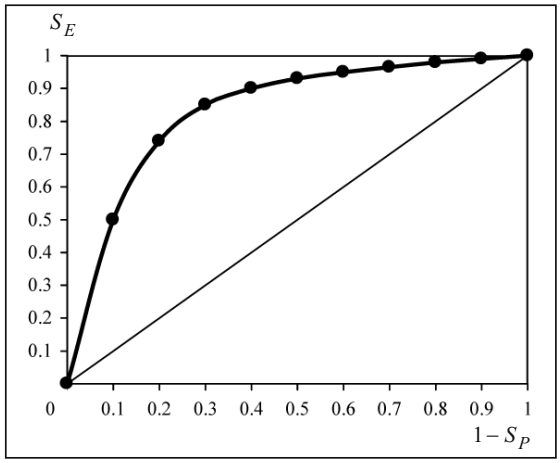

Fig. 2. ROC curve of the test based on decision rule (3).

The test D (see Fig. 1), located on the diagonal $S_E = 1 - S_P$ of the ROC space is useless. Such tests are equivalent to classifiers that use a "random guess" strategy for classes V_1 and V_2.

In medical diagnostics, threshold decision rules are often used, which have the form

$$\begin{aligned} &\text{"Sick"} \quad \text{if } x > x_0, \\ &\text{"Healthy"} \quad \text{if } x \le x_0, \end{aligned} \tag{3}$$

where x is the diagnostic feature, while x_0 is the threshold value of the feature x. In this case, it becomes possible to determine pairs S_E and S_P for different values x_0, which correspond to the sequence of points of the binary classifier in the ROC space.

A simple algorithm for constructing an experimental ROC curve follows from rule (3) as such:

— sort the values of the feature x under consideration in the descending order;

— sequentially decrease (with some step) the threshold x_0 and, using the examination sample of observations, calculate the corresponding pairs of values S_E and S_P;

— display the obtained sequence of pairs S_E and S_P in the ROC space (Fig. 2).

We denote by $p(x|V_1)$ and $p(x|V_2)$ the conditional probability distributions of the feature x in the classes V_1 and V_2. We call the set $X_1 \triangleq \{x : p(x|V_1) \neq 0\}$ the distribution carrier of x for the class V_1, and we call the set $X_2 \triangleq \Delta\{x : p(x|V_2) \neq 0\}$ the distribution carrier of x for the class V_2.

There are four possible cases that characterize the topology of carriers X_1 and X_2 (Fig. 3) and the shape of the ROC curves corresponding to these cases (Fig. 4). In case 1, the sets X_1 and X_2 do not intersect, and a threshold x_0 can be set that completely separates the feature values x for the classes V_1 and V_2. Then, rule (3) determines the ideal test that provides error-free classification of V_1 and V_2. The graph of the corresponding ROC curve (see Fig. 4, curve *1*) runs from the point with coordinates (0, 0) to the point with coordinates (0, 1) and further to the point with coordinates (1, 1).

In case 2, the sets X_1 and X_2 partially intersect, and the corresponding ROC curve is curve *2* (see Fig. 4).

In case 3, when the sets X_1 and X_2 coincide, but the conditional distributions $p(x|V_1)$ and $p(x|V_2)$ are different, the corresponding ROC curve is curve *3* (see Fig. 4). And finally, in case 4, when not only the sets X_1 and X_2, but also the conditional distributions $p(x|V_1)$ and $p(x|V_2)$ coincide, the diagnostic feature x becomes useless, and the corresponding ROC curve lies on the diagonal $S_E = 1 - S_P$ of the ROC space (see Fig. 4, curve *4*).

Note that two individually useless features, namely, x_1 and x_2, which have the same one-dimensional conditional distributions in classes, i.e., $p(x_i|V_1) \equiv p(x_i|V_2)$, $i = 1, 2$, together can be not only useful, but also provide error-free classification when $X_1^{(2)} \cap X_1^{(2)} = \varnothing$, where $X_m^{(2)} \triangleq \{(x_1, x_2) : p(x_1, x_2|V_m) \neq 0\}$ are carriers of the two-dimensional conditional distributions $p(x_1, x_2|V_m)$, $m = 1, 2$ (Fig. 5).

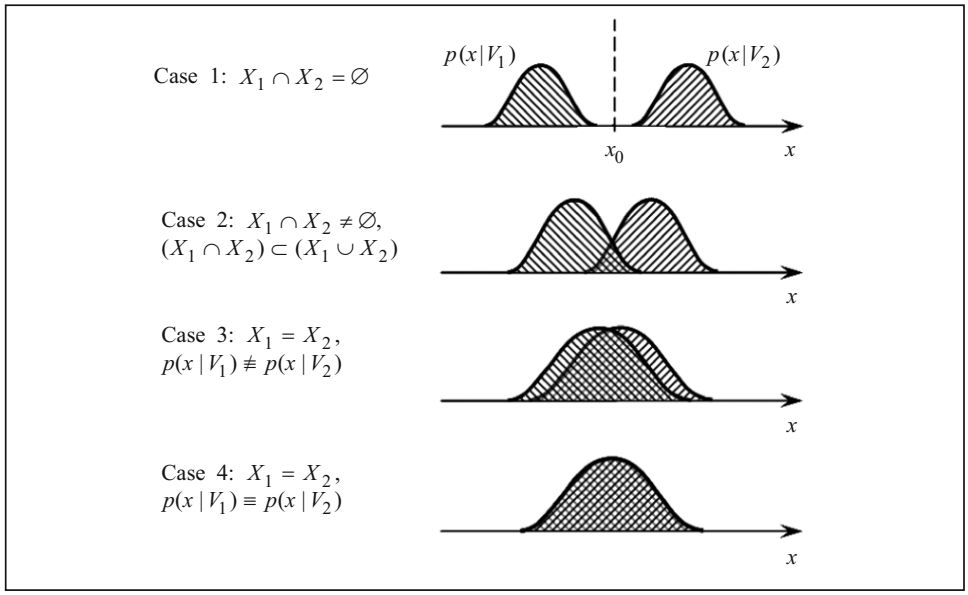

Fig. 3. Topology of the carriers X_1 and X_2 of the classes V_1 and V_2.

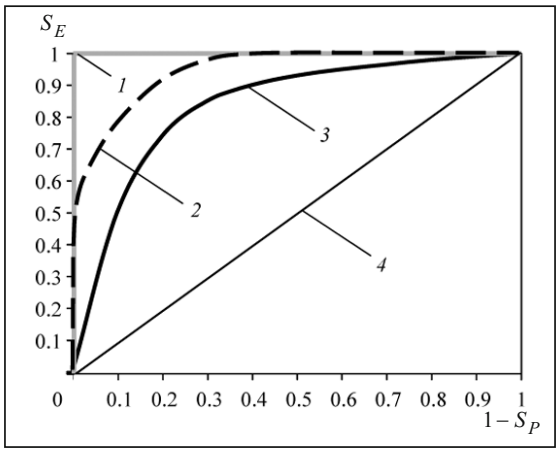

Fig. 4. The relationship of ROC curves with the topology of the carriers of classes V_1 and V_2.

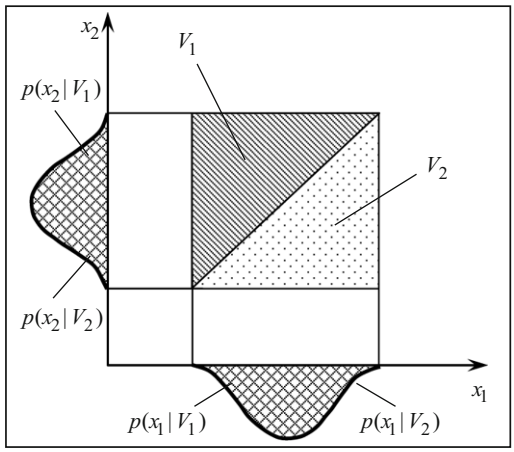

Fig. 5. Errorless classification of V_1 and V_2 based on two separately useless features.

It was proved in [25] that such an interesting case is fundamentally possible only when there is a conditional statistical dependence between the features x_1 and x_2 in both classes at once, i.e.,

$$p(x_1, x_2|V_m) \not\equiv p(x_1|V_m)\,p(x_2|V_m),\ \ m = 1, 2.$$

To compare the efficiency of diagnostic tests, it is convenient to use an integral characteristic, which is usually considered to be the area AUC (Area Under Curve) calculated under the ROC curve [16]. The value AUC determines the average sensitivity $\overline{S}_E$ of the test for possible values of specificity $0 \le S_P \le 1$ or the average specificity $\overline{S}_P$ of the test for possible values of sensitivity $0 \le S_E \le 1$, and $AUC = 1$ for the ideal test and $AUC = 0.5$ for a useless test. Therefore, it is believed that the closer the value AUC to unity, the more efficient the test.

It is clear that for the operational characteristics of potentially useful tests that lie above the diagonal of the ROC space, the following condition is satisfied:

$$S_E > 1 - S_P. \tag{4}$$

ROC ANALYSIS USING STATISTICAL DECISION THEORY

Let us now consider some refinements of the traditional ROC analysis, which make it possible to evaluate the efficiency of the diagnostic test in terms of expected losses (average risk).

Let the a priori probability $P(V_1)$ of the appearance of a sick person in a group of examined persons be known according to medical statistics and characterize the prevalence of a particular disease. Accordingly, the value $P(V_2) = 1 - P(V_1)$ characterizes the a priori probability of the presence of a person who does not have the considered disease in the group of examined persons.

We characterize the possible errors of the diagnostic test by values L_{12} and L_{21}, which determine the losses from the error of missing the target (the "Sick" patient from the class V_1 was not detected) and the false alarm (the "Healthy" patient is classified as a sick one of the class V_1).

Then, the expected losses from false diagnosis are determined by the relation

$$R = P(V_1)(1 - S_E)L_{12} + [1 - P(V_1)](1 - S_P)L_{21}. \quad (5)$$

Condition (5) makes it possible to calculate the average risk of false diagnosis (aposterior risk) based on the known $P(V_1)$, L_{12}, and L_{21} for a test that has operating characteristics S_E and S_P. Here, a natural question arises on whether such a test is always going to be useful.

To answer this question, first of all, it is necessary to formulate the definition of the usefulness of a diagnostic test. According to the authors of this article, such a definition can be the condition that a test provides lower average losses than those that would be observed without the use of the diagnostic test [26]. In other words, the concept of usefulness can be logically formulated as follows.

Definition 1. A diagnostic test is useful if a strict inequality

$$R < R_0 \quad (6)$$

holds, i.e., aposterior risk R is less than a priori risk R_0 implemented by a decision-making strategy based only on the known values $P(V_1)$, L_{12}, and L_{21}.

Let us present some explanations. Obviously, if the binary classifier implements a Bayesian decision-making strategy that minimizes the aposterior risk, then either condition (6) or the condition $R = R_0$ is satisfied. In other words, a formally constructed diagnostic algorithm cannot be "harmful." However, in practice, the information needed to implement the Bayesian strategy is often missing, and the diagnostic test is not constructed in a formal way.

In such cases, important for practical use, at fixed $P(V_1)$, L_{12}, and L_{21}, and known S_E and S_P, instead of condition (6), a seemingly "paradoxical" condition $R > R_0$ can be satisfied.

Let us illustrate this possibility with a model example. Suppose a diagnostic test is used to identify "Sick" patients in a group of 10,000 people. This test has sufficiently high operational characteristics $S_E = 0.9$ and $S_P = 0.9$. Let the prevalence of the disease be $P(V_1) = 0.02$, and let the losses from target miss error and false alarm be assumed to be the same, i.e., $L_{12} = L_{21} = 1$.

The expected test results are presented in Fig. 6. If a diagnostic test is not used, then the optimal strategy should be as follows: to recognize all the examined patients as "Healthy" ones (otherwise the losses R_0 are greater!). Then, the average a priori losses per examined person, due to the fact that 200 "Sick" patients are not detected, are

$$R_0 = (L_{12} \cdot 200)/10000 = (1 \cdot 200)/10000 = 0.02. \quad (7)$$

In the case of using a diagnostic test with a sensitivity of $S_E = 0.9$ and a specificity of $S_P = 0.9$, the average losses due to false positive diagnostic results of 980 "Healthy" patients and non-detection of 20 "Sick" patients are

$$R = (L_{21} \cdot 980 + L_{12} \cdot 20)/10000 = (1 \cdot 980 + 1 \cdot 20)/10000 = 0.1. \quad (8)$$

It follows from relations (7) and (8) that $R > R_0$, and such a test cannot be considered as useful, because for the given losses from target miss and false alarm errors, it only increases the expected losses.

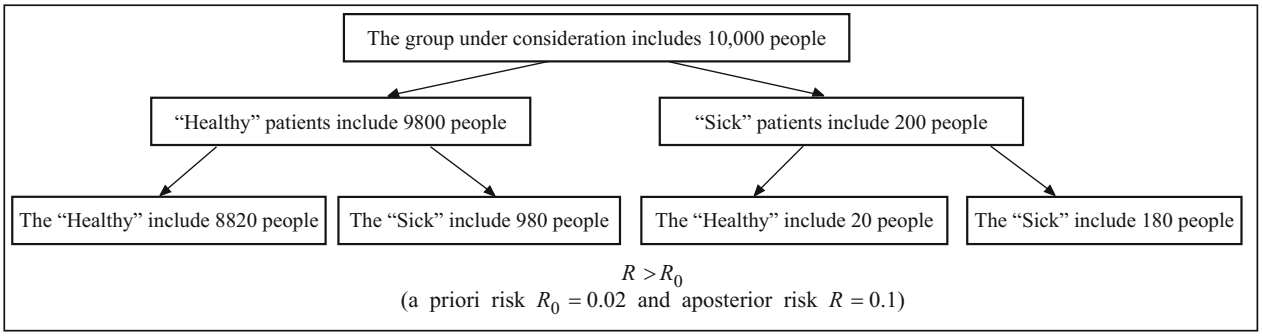

Fig. 6. Expected test results.

Despite the "paradoxical" fact that $R > R_0$, the obtained result is fully justified. It is not difficult to show that if we base the decision that the patient is "Sick" on the considered values $P(V_1)$, L_{12}, L_{21}, S_E, and S_P, it does not justify the minimization of the average risk. In other words, we must always make the decision for each examinee that the patient is "Healthy," which coincides with a priori decisions.

Let us show that (4) is only a necessary but not a sufficient condition that guarantees the fulfillment of strict inequality (6).

Statement 1. A diagnostic test is guaranteed to be useful in the sense of Definition 1 if there is a relationship between its sensitivity S_E and specificity S_P expressed by the relations

$$S_E > m(1-S_P) \text{ if } m \geq 1 \tag{9}$$

or

$$S_E > 1-m+m(1-S_P) \text{ if } m<1, \tag{10}$$

where

$$m = \frac{[1-P(V_1)]}{\omega P(V_1)} \tag{11}$$

is a value that depends on the a priori probability $P(V_1)$ and the ratio between errors of missing target and false alarms

$$\omega = \frac{L_{12}}{L_{21}}. \tag{12}$$

Proof. If you do not use a diagnostic test and make decisions based only on values $P(V_1)$, L_{12}, and L_{21}, then the a priori strategy is reduced to one of the two following possible options:

— classify all examinees as "Sick" V_1, and then the a priori risk has the form

$$R_0^+ = L_{21}[1-P(V_1)]; \tag{13}$$

— make the decision that each examinee is "Healthy," and then the a priori risk has the form

$$R_0^- = L_{12}P(V_1). \tag{14}$$

From (13) and (14), taking into account (11), it follows that the minimum a priori risk, which appears in the right-hand side of inequality (6), can be written as follows:

$$R_0 = \begin{cases} L_{21}[1-P(V_1)] & \text{if } m<1, \\ L_{12}P(V_1) & \text{if } m \geq 1. \end{cases} \tag{15}$$

As a result of substituting expressions (5) and (15) into inequality (6), after elementary transformations, we obtain relations (9) and (10). The statement is proved.

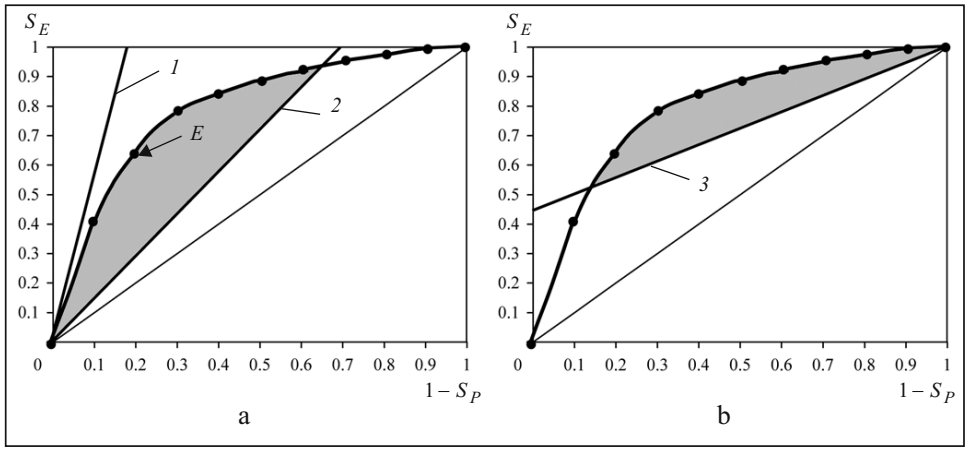

Fig. 7. The illustration of the idea of improved ROC analysis for cases $m \geq 1$ (a) and $m < 1$ (b).

As can be seen from expression (11), to check conditions (9) and (10), it is sufficient to have only information about the loss ratio (12), and such information is more accessible when solving applied problems than information about the absolute values of losses from the error of missing the target L_{12} and false alarm L_{21}.

Conditions (9) and (10) make it possible to reasonably evaluate the usefulness of new diagnostic tests. For this purpose, it is advisable to improve the traditional ROC analysis by limiting the corresponding fragment of the ROC curve (Fig. 7).

Straight line *1* (Fig. 7a) corresponds to the case where $P(V_1) = 0.15$ and $\omega = L_{12}L_{21}^{-1} = 1$. Therefore, the test is useless, because straight line *1* does not intersect the ROC curve. However, if we assume that the loss from the error of missing the target is four times higher than the loss from a false alarm, i.e., $L_{12}L_{21}^{-1} = 4$, then the corresponding straight line *2* already crosses the ROC curve. Thus, a test that has sensitivity $S_E = 62.5\%$ and specificity $S_P = 80\%$ is guaranteed to be useful according to (6) since under such conditions the point E, which corresponds to the given operational characteristics, is located on the admissible part of the ROC curve. In Fig. 7b, where $m < 1$, limiting straight line *3* begins from the values $S_P = 0$ and $S_E = 1$. And if it crosses the ROC curve, then there are values S_P and S_E on its corresponding fragment, for which the test is useful.

Note that for $m = 1$, sufficient conditions (9) and (10) coincide with necessary condition (5). In this case, the straight line limiting the fragment of the ROC curve of the guaranteed useful test lies on the diagonal of the ROC space.

It is interesting to compare the obtained conditions with the results given in [22] where the so-called iso-performance lines are considered. It follows from expression (5) that the equation of the isoline with expected losses $R = R^*$ has the form of a linear dependence in coordinates S_E, $1 - S_P$

$$S_E = m(1 - S_P) + 1 - \frac{R^*}{L_{12}P(V_1)}. \tag{16}$$

Equation (16) coincides with the equations of straight lines (9) and (10) if the condition

$$R^* = \begin{cases} L_{21}P(V_1) & \text{if } m \geq 1, \\ L_{12}[1 - P(V_1)] & \text{if } m < 1 \end{cases}$$

is satisfied, whose right-hand side coincides with the right-hand side of (15). That is, the area of useful tests is limited by the iso-performance line

$$S_E = m(1 - S_P) + 1 - \frac{R_0}{L_{12}P(V_1)}, \tag{17}$$

which corresponds to the a priori risk level R_0 and is requiered to pass through the point with coordinates (0, 0) if $m \geq 1$, or through the point with coordinates (1, 1) if $m < 1$.

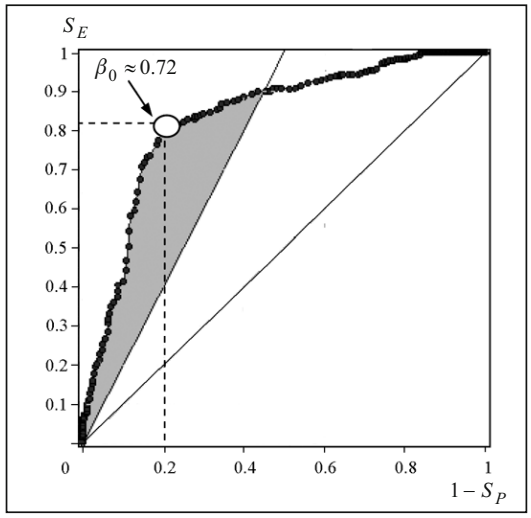

Fig. 8. The result of evaluating the usefulness diagnostic rule (19).

It is clear that the following cases are possible:

— isoline (17) crosses the ROC curve and then $R < R_0$;

— isoline (17) is tangent to the ROC curve and then $R = R_0$;

— isoline (17) does not cross the ROC curve and then $R > R_0$.

Conditions (9) and (10) make it possible to determine the feasible interval of the loss ratio $\omega = L_{12}L_{21}^{-1}$, according to which the test with sensitivity S_E and specificity S_P remains useful for screening a disease with prevalence $P(V_1)$. Such an interval is determined by the formula

$$\frac{1-P(V_1)}{P(V_1)}\frac{1-S_P}{S_E} \le \omega \le \frac{1-P(V_1)}{P(V_1)}\frac{S_P}{1-S_E}. \tag{18}$$

PRACTICAL RESULTS

The obtained results made it possible to substantiate the efficiency of the method of analysis and interpretation of electrocardiograms, which is innovative in cardiology, namely, the phaseography method [27]. The method is based on the transition from a scalar signal $x(t)$, which contains information about the electrical activity of the heart, to the trajectory on the phase plane with coordinates $x(t)$ and $\dot{x}(t)$, where $\dot{x}(t)$ is the rate of change of the signal [28].

Dues to this transition, it was possible to determine a new diagnostic feature β_T, which is the symmetry of the repolarization area of the electrocardiogram (ECG). To test the diagnostic value β_T, statistical processing of clinical data was carried out, which resulted in 441 ECG recordings of pre-verified patients with the coronary heart disease (CHD) and 387 ECG recordings of healthy individuals from the control group.

The result of data processing showed that the average values β_T have statistically significant differences in groups ($p < 0.01$) and are 0.956 ± 0.43 and 0.665 ± 0.12 in the CHD group and the control group, respectively.

To evaluate the usefulness of the diagnostic rule

$$\begin{aligned} &\text{CHD} \quad \text{if} \ \ \beta_T > \beta_0, \\ &\text{Normal} \quad \text{if} \ \ \beta_T \le \beta_0, \end{aligned} \tag{19}$$

based on the comparison of β_T with the threshold value β_0, a developed software application was used, in which an improved ROC analysis was implemented. The experimental ROC curve is shown in Fig. 8. The range of values S_E and S_P, which under the conditions $P(V_1) = 0.1$ and $\omega = 5$ satisfy relations (9) and (10), is highlighted in gray.

It is established that decision-making according to rule (19) provides sensitivity $S_E = 81\%$ and specificity $S_P = 78\%$ if $\beta_0 \approx 0.72$ and is the optimal value of the threshold corresponding to the point farthest from the diagonal.

According to (18), if $S_E = 81\%$ and $S_P = 78\%$, the diagnostic rule (19) is guaranteed to be useful for screening CHD with prevalence $P(V_1) = 0.1$ in a fairly wide range of loss ratio values $2.44 \le \omega \le 37$.

Note that only patients for whom traditional 12-lead ECG analysis did not reveal any abnormalities were included in the CHD group. At the same time, the proposed diagnostic rule (19) confirmed the sufficiently high indicators on such a "complex" clinical material. Therefore, we consider it quite acceptable for the CHD screening.

CONCLUSIONS

It is proved that conditions (9) and (10) guarantee the usefulness of the diagnostic test in the sense of reducing the a priori risk. The obtained conditions made it possible to improve the traditional ROC analysis due to the limitation of the corresponding fragment of the ROC curve. It is shown that the straight line that limits the fragment of the ROC curve of the guaranteed useful test coincides with the iso-performance line corresponding to the level of a priori risk. The feasible limits of the ratio of losses from the error of missing a target and a false alarm, according to which the test with sensitivity S_E and specificity S_P remains useful for screening a disease with a known prevalence $P(V_1)$, are determined.

The results of the study have a practical application for confirming the efficiency of the innovative method of phasography in cardiology, which makes it possible to detect persons with hidden initial signs of the coronary heart disease during mass preventive examinations.

REFERENCES

1. D. Maxim, R. Niebo, and M. J. Utel, "Screening tests: A review with examples," Inhal. Toxicol., Vol. 26, Iss. 13, 811–828 (2014). http://doi.org/10.3109/08958378.2014.955932.
2. O. Zhukovska, "Decision-making model on potential borrower lending for independent experts group," in: Proc. IEEE 3rd Intern. Conf. on System Analysis & Intelligent Computing (SAIC) (Kyiv, Ukraine, Oct. 4–7, 2022), IEEE (2022), pp. 118–121. http://doi.org/10.1109/SAIC57818.2022.9923015.
3. C. Dendek and J. Mańdziuk, "Improving performance of a binary classifier by training set selection," in: V. Kurková, R. Neruda, and J. Koutník (eds.), Artificial Neural Networks — ICANN 2008. Lecture Notes in Computer Science, Vol. 5163, Springer, Berlin–Heidelberg (2008), pp. 128–135. https://doi.org/10.1007/978-3-540-87536-9_14.
4. C. E. Metz, "Fundamental ROC analysis," in: R. L. Van Metter, J. Beutel, and H. L. Kundel (eds.), Handbook of Medical Imaging, Vol. 1, Physics and Psychophysics, Ch. 15, SPIE Press, Bellingham (2000), pp. 751–769. https://doi.org/10.1117/3.832716.ch15.
5. T. Fawcett, "Using rule sets to maximize ROC performance," in: Proc. IEEE Intern. Conf. on Data Mining (ICDM-2001), IEEE Computer Society (2001), pp. 131–138. https://doi.org/10.1109/ICDM.2001.989510.
6. P. Flach and S. Wu, "Repairing concavities in ROC curves," in: Proc. 2003 UK Workshop on Computational Intelligence, University of Bristol (2003), pp. 38–44.
7. P. Sonego, A. Kocsor, and S. Pongor, "ROC analysis: Applications to the classification of biological sequences and 3D structures," Brief. Bioinform., Vol. 9, Iss. 3, 198–209 (2008). https://doi.org/10.1093/bib/bbm064.
8. K. Feng, H. Hong, K. Nang, and J. Wang, "Decision making with machine learning and ROC curves," arXiv:1905.02810v1 [stat.ME] 5 May (2019). https://doi.org/10.48550/arXiv.1905.02810.
9. J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves" in: Proc. 23rd Intern. Conf. on Machine Learning (ICML'06), Association for Computing Machinery, New York (2006), pp. 233–240. https://doi.org/10.1145/1143844.1143874.
10. W. B. van den Hout, "The area under an ROC curve with limited information," Med. Decis. Making, Vol. 23, Iss. 2, 160–166 (2003). https://doi.org/10.1177/0272989X03251246.

11. M. S. Pepe, G. Longton, and H. Janes, “Estimation and comparison of receiver operating characteristic curves,” Stata J., Vol. 9, No. 1, 1–16 (2009). https://doi.org/10.1177/1536867X0900900101.
12. T. A. Alonzo and M. S. Pepe, “Distribution-free ROC analysis using binary regression techniques,” Biostatistics, Vol. 3, Iss. 3, 421–432 (2002). https://doi.org/10.1093/biostatistics/3.3.421.
13. F. Provost and T. Fawcett, “Robust classification for imprecise environments,” Machine Learning, Vol. 42, No. 3, 203–231 (2001). https://doi.org/10.1023/A:1007601015854.
14. M. Majnik and Z. Bosnić,“ROC analysis of classifiers in machine learning: A survey,” Intell. Data Anal., Vol. 17, No. 3, 531–558 (2013). https://doi.org/10.3233/IDA-130592.
15. N. Hu, “Using receiver operating characteristic (ROC) analysis to evaluate information-based decision-making,” in: M. Khosrow-Pour, D.B.A. (ed.), Advanced Methodologies and Technologies in Business Operations and Management, IGI Global, Hershey, PA (2019), pp. 764–776. https://doi.org/10.4018/978-1-5225-7362-3.ch057.
16. D. J. Hand and R. J. Till, “A simple generalization of the area under the ROC curve to multiple class classification problems,” Machine Learning, Vol. 45, No. 2, 171–186 (2001). https://doi.org/10.1023/A:1010920819831.
17. C. Ferri, J. Hernández-Orallo, and M. A. Salido, “Volume under the ROC surface for multi-class problems,” in: N. Lavrač, D. Gamberger, H. Blockeel, and L. Todorovski (eds.), Machine Learning: ECML 2003. ECML 2003; Lecture Notes in Computer Science, Vol. 2837, Springer, Berlin–Heidelberg (2003), pp. 108–120. https://doi.org/10.1007/978-3-540-39857-8_12.
18. D. C. Edwards, C. E. Metz, and M. A. Kupinski, “Ideal observers and optimal ROC hypersurfaces in N-class classification,” IEEE Trans. Med. Imaging, Vol. 23, No. 7, 891–895 (2004). https://doi.org/10.1109/TMI.2004.828358.
19. D. C. Edwards, C. E. Metz, and R. M. Nishikawa, “The hypervolume under the ROC hypersurface of “Near-Guessing” and “Near-Perfect” observers in N-class classification tasks,” IEEE Trans. Med. Imaging, Vol. 24, No. 3, 293–299 (2005). https://doi.org/10.1109/tmi.2004.841227.
20. X. He and E. C. Fry, “An optimal three-class linear observer derived from decision theory,” IEEE Trans. Med. Imaging, Vol. 26, No. 1, 77–83 (2007). https://doi.org/10.1109/TMI.2006.885335.
21. B. Sahiner, H.-P. Chan, and L. M. Hadjiiski, “Performance analysis of 3-class classifiers: Properties of the 3D ROC surface and the normalized volume under the surface,” IEEE Trans. Med. Imaging, Vol. 27, No. 2, 215–227 (2008). https://doi.org/10.1109/TMI.2007.905822.
22. T. Fawcett, “ROC graphs with instance-varying costs,” Pattern Recognit. Lett., Vol. 27, Iss. 8, 882–891 (2006). https://doi.org/10.1016/j.patrec.2005.10.012.
23. R. Meekins, S. Adams, P. A. Beling, K. Farinholt, N. Hipwell, A. Chaudhry, S. Polter, and Q. Dong, “Cost-sensitive classifier selection when there is additional cost information,” Proc. Mach. Learn. Res., Vol. 88, 17–30 (2018).
24. R. C. Holte and C. Drummond, “Cost-sensitive classifier evaluation using cost curves,” in: T. Washio, E. Suzuki, K. M. Ting, and A. Inokuchi (eds.), Advances in Knowledge Discovery and Data Mining, PAKDD 2008; Lecture Notes in Computer Science, Vol. 5012, Springer, Berlin–Heidelberg (2008), pp. 26–29. https://doi.org/ 10.1007/978-3-540-68125-0_4.
25. L. S. Fainzilberg, “Plausible but groundless premises when constructing diagnostic models,” J. Autom. Inform. Sci., Vol. 52, Iss. 5, 38–50 (2020). https://doi.org/10.1615/JAutomatInfScien.v52.i5.40.
26. L. S. Fainzilberg, “Conditions of utility of diagnostic tests from the point of view of the statistical theory of decision making,” J. Autom. Inform. Sci., Vol. 35, Iss. 4, 63–73 (2003). https://doi.org/10.1615/JAutomatInfScien.v35.i4.30.
27. L. S. Fainzilberg, “New opportunities of phasegraphy in medical practice,” Sci. Innov., Vol. 1, Iss. 3, 37–50 (2017). https://doi.org/10.15407/scine13.03.037.
28. L. S. Fainzilberg, “New approaches to the analysis and interpretation of the shape of cyclic signals,” Cybern. Syst. Analysis, Vol. 56, No. 4, 665–674 (2020). https://doi.org/10.1007/ s10559-020-00283-0.