DOI https://doi.org/10.15407/csc.2021.02.028
UDC 004.001

L.S. FAINZILBERG, Dr.Sc. (Eng.), Professor, Chief Researcher,
International Research and Training Center for Information Technologies
and Systems of the NAS and MES of Ukraine,
Acad. Glushkova ave., 40, Kyiv, 03187, Ukraine,
fainzilberg@gmail.com

JU.R. DYKACH, Student of Biomedical Engineering Faculty,
the National Technical University of Ukraine
«Igor Sikorsky Kyiv Polytechnic Institute»,
37, Peremohy ave., Kyiv, 03056, Ukraine,
jul.dykach@gmail.com

DEVELOPMENT OF LINGUISTIC APPROACH TO THE PROBLEM OF THE COMPUTER ELECTROCARDIOGRAM'S CLASSIFICATIONS

Six decision rules based on the analysis of the Levenshtein distances and the number of occurrences of characteristic patterns of codewords are considered and investigated. It has been shown that using of the developed decision rules makes it possible to increase the sensitivity and specificity of diagnos-tics even in cases when the ECG does not show traditional electrocardiological signs of myocardial ischemia..

Keywords: *ECG, the Levenshtein distance, occurrence frequency of a substring in a code word, decision rule.*

Introduction

For more than a hundred years, electrocardiography has been widely used in cardiological practice to diagnose diseases of the cardiovascular system. However, it is known that the traditional approach to the analysis and interpretation of the ECG does not always provide the required reliability of diagnostic decisions. So, for example, according to the medical statistics [1] resting ECG, assessed according to generally accepted criteria, remains normal in almost 50% of patients with chronic coronary artery disease (CAD). Therefore, experts are actively exploring new approaches to computerized ECG processing.

One of these new approaches is an intelligent method of ECG processing, called “fasegraphy”. It was developed at the International Research and Training Center for Information Technologies and Systems of the National Academy of Sciences of Ukraine and Ministry of Education and Science of Ukraine [2]. The method is based on the transition from a scalar signal $x(t)$ to a vector signal on the phase plane $x(t)$, $\dot{x}(t)$, where $\dot{x}(t)$ is the rate of change in the electrical activity of the heart, which is determined on the basis of original computational procedures according to the signal $x(t)$ recorded in the standard lead, for example, in the first standard lead (left and right hand).

Large-scale clinical trials have shown that the fasegraphy method provides an increase in the reliability of detecting latent signs of myocardial ischemia even in those cases when the generally accepted electrocardiographic signs of CAD (depression or elevation of the isoelectric line) are absent in all 12 traditional leads. This is achieved through the

use of new diagnostic ECG indicators in the phase space, in particular, the parameter β_T characterizing the symmetry of the repolarization area on the phase plane [3, 4].

Further studies have shown that not only the average value β_T, but also the dynamics of its change from cycle to cycle has diagnostic value. In the simplest case, the variability of the indicator β_T characterizes the mean square deviation of the RMS β_T. For a more subtle analysis of the change in the indicator β_T, it turned out to be useful to calculate the entropy estimates of the signal, in particular, the modified permutation entropy [5].

An effective method for assessing the dynamics of changes in the shape of ECG cycles is based on the use of a linguistic approach to processing cyclic signals. This approach is based on the transition from the observed ECG to a sequence of symbols (word), which uniquely encodes the ECG [6].

The purpose of this article is a further examination of this method.

Basic Method of Linguistic Analysis and ECG Interpretation

Let`s first consider the approach to ECG analysis proposed in [6], which we will need in further studies. Using a microprocessor sensor with finger electrodes, the ECG signal $x(t)$ of the first standard lead is recorded (Fig. 1).

For each i cycle ($i = 1,..., N$) of the digitized signal $x(t)$, using special computational procedures implemented in the fasegraphy method, the durations of the cycles (RR_i-intervals) and the values of the mentioned original indicator $\beta_{T,i}$ are determined.

Further, the dynamics of these indicators is assessed in the process of ECG registration. For this purpose, indicator variables are introduced

$$V_i^{(RR)} = \begin{cases} +1, \text{ if } & RR_i - RR_{i-1} > 0, \\ -1, \text{ if } & RR_{i-1} - R_i > 0. \end{cases} \quad (1)$$

$$V_i^{(\beta)} = \begin{cases} +1, \text{ if } & \beta_{T,i} - \beta_{T,ii-1} > \\ -1, \text{ if } & \beta_{T,i-1} - \beta_{T,i} > \end{cases} \quad (2)$$

where $i = 2,..., N$.

***Fig.* 1.** Microprocessor ECG recorder[1]

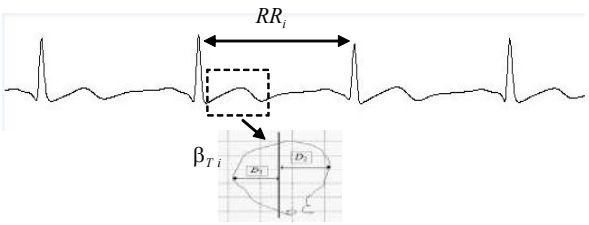

***Fig.* 2**. ECG indicators

Sequences $V_i^{(R)}$ and $V_i^{(\beta)}$ allow you to encode each ECG cycle with one of the alphabet symbols $A = \{a, b, c, d\}$ as follows (Table 1).

As a result, the $N - 1$ — digit word S_k, composed of the symbols a, b, c, d, uniquely encodes the k-th processed ECG (Fig. 3).

The transition from the observed ECG to the code word makes it possible to use the methods of mathematical linguistics to solve the problem of the analysis and interpretation of the ECG. In particular, the proposed method provides for an assessment of the proximity $L(S_\mu, S_\nu)$ between the codewords S_μ, S_ν, of processed ECGs based on the editorial distance $L(S_\mu, S_\nu)$ — the Levenshtein dis-

Table 1. **Principle of ECG cycle coding**

Indicator variable value $V_i^{(RR)}$	+1	+1	-1	-1
Indicator variable value $V_i^{(\beta T)}$	+1	-1	+1	-1
Symbol	***a***	***b***	***c***	***d***

[1] Sensor developed by Solvaig, J.S.C. (Kyiv) https://solvaig.com/fasegraphy

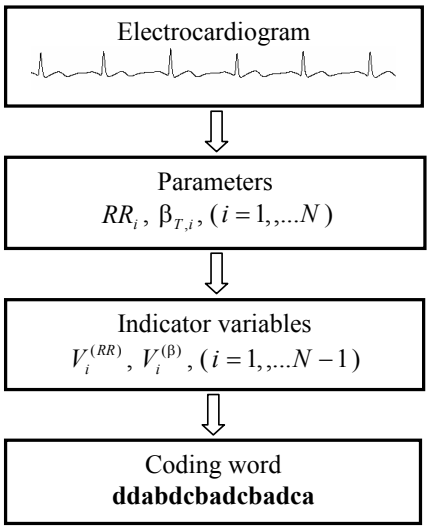

Fig. 3. The principle of forming a code word

tance, which is defined as the minimum number of editing operations (insertion, deletion, and replacement of a symbol) that ensures the transition from word S_μ to word S_ν [9].

To calculate the Levenshtein distance, the Wagner-Fischer algorithm [10], based on the dynamic programming method, is used. Table 2 shows the optimal path to transition from the word,

$$S_\mu = \text{ddabdcbadcbadca} \qquad (3)$$

to the word

$$S_\nu = \text{bacdaaacdadccbb}. \qquad (4)$$

The closeness of these words assesses the Levenshtein distance $L(S_\mu, S_\nu) = 10$.

Table 2. **Optimal transition from word to word**

Step	Original word	Operation	Result of editing
1	S_μ = <u>d</u>dabdcbadcbadca	Replacement $d \to b$	S = <u>b</u>dabdcbadcbadca
2	S = b<u>d</u>abdcbadcbadca	Deleting d	S = babdcbadcbadca
3	S = ba<u>b</u>dcbadcbadca	Replacement $b \to c$	S = ba<u>c</u>dcbadcbadca
4	S = bacd<u>c</u>badcbadca	Deleting c	S = bacdbadcbadca
5	S = bacd<u>b</u>adcbadca	Replacement $b \to a$	S = bacd<u>a</u>adcbadca
6	S = bacdaa<u>d</u>cbadca	Replacement $d \to a$	S = bacdaa<u>a</u>cbadca
7	S = bacdaaac<u>b</u>adca	Replacement $b \to d$	S = bacdaaac<u>d</u>adca
8	S = bacdaaacdadc<u>a</u>	Replacement $a \to c$	S = bacdaaacdadc<u>c</u>
9	S = bacdaaacdadcc	Insert b	S = bacdaaacdadcc<u>b</u>
10	S = bacdaaacdadccb	Insert b	S_ν = bacdaaacdadccb<u>b</u>

ECG classification based on the proposed approach assumes:

- construction of class standards based on the Levenshtein distances between pairs of code words of the training set;
- comparison of the code word of the processed ECG with the standards.

The algorithm for constructing standards is as follows. Let, as a result of the experiments, Q_{CAD} electrocardiograms of patients with coronary artery disease (CAD) were recorded, which, in accordance with tables 1, are coded with words $S_q^{(CAD)}$, $q = 1,...$ ..., Q_{CAD}. Let`s determine the Levenshtein distances $L_{\mu\nu}$ $(S_\mu^{(CAD)}, S_\nu^{(CAD)})$ between each pair $S_\mu^{(CAD)}, S_\nu^{(CAD)}$, $\mu = 1,..., Q_{CAD}$, $\nu = 1,..., M_1$ of the indicated words and form a square $Q_{CAD} \times Q_{CAD}$ matrix of distances $L_{\mu\nu}$ $(S_\mu^{(CAD)}, S_\nu^{(CAD)})$, $\mu = 1,..., Q_{CAD}$, $\nu = 1,..., Q_{CAD}$:

$$\Lambda = \begin{pmatrix} L_{11} & L_{12} & \dots & L_{1Q_{CAD}} \\ L_{21} & L_{22} & \dots & L_{2Q_{CAD}} \\ \dots & \dots & \dots & \dots \\ L_{CAD1} & L_{CAD2} & \dots & L_{Q_{CAD}Q_{CAD}} \end{pmatrix}. \qquad (5)$$

Then the CAD patient's reference word will determine the row of the matrix (5), the sum of the elements of which is minimal, i.e.

$$S_0^{(CAD)} = \arg\min_{1 \le \nu \le Q_{CAD}} \sum_{\mu=1}^{Q_{CAD}} L_{\mu\nu}. \qquad (6)$$

The reference word of the healthy group (Healthy) is determined in a similar way by the elements of the Levenshtein distance matrix $L_{\mu\nu}\left(S_\mu^{(Healthy)}, S_\nu^{(Healthy)}\right)$,

$\mu = 1,..., Q_{Healthy}$, $\nu = 1,..., Q_{Healthy}$, built for all pairs of codewords of the control group, i.e.

$$S_0^{(Healthy)} = \arg\min_{1 \le \nu \le Q_{Healthy}} \sum_{\mu=1}^{Q_{Healthy}} L_{\mu\nu}. \quad (7)$$

Reference code words (6), (7) allow to classify the analyzed ECG based on the comparison of the Levenshtein distances between its code word S_t and reference words $S_0^{(CAD)}$ and $S_0^{(Healthy)}$ using the following decision rule:

Decision rule 1:

CAD, if $L(S_t, S_0^{(CAD)}) > L(S_t, S_0^{(Healthy)})$, (8)

Healthy, if $L(S_t, S_0^{(CAD)}) \le L(S_t, S_0^{(Healthy)})$. (9)

The study of the diagnostic capabilities of the proposed method was carried out on the basis of real ECGs registered at the Department of Ischemic Heart Diseases of the V.D. Strazhesko AMS of Ukraine (Kyiv) and four German clinics: Essen University Hospital (Essen), Katholical Hospital "Phillpusstift" (Essen), Heart and Diabetes Center of North Rhein-Weasfalia (Bad-Oeynhausen), German Heart Center (Berlin).

The clinical material consisted of 100 ECG records of patients with coronary artery disease (CAD), the diagnosis of which was previously established based on the results of coronary angiography, and 100 ECG records of healthy volunteers included in the control group. It is important to note that the training set included only those ECGs in which traditional electrocardiographic signs of coronary artery disease (flat or negative wave T, depression or isoelectric line elevation) were absent. In other words, from the point of view of traditional cardiology, all ECGs, including those of verified patients, would be classified as healthy (Healthy).

Table 3 shows a fragment of the database of code words built for the ECG of the training set.

Table 3. **ECG training set code words**

CAD	Healthy
abdcbbdddaabdaacdcab	adccaddadcbacddaaddbbcadcc
daadddaccdaaddabaddaa	caddadcbbcdaddaddcbaccbcdd
aadbaadcabcdaabddcaddca	ddababddabadabacbcabddaadb
dddcaadcabaddcacddcbdc	cadcddabdbdabdcaaabab dabdc
abddcabddaadddaaddcaab	bdcbaadcabcdaadcdcbbacddccb
dcaabdaaddcadcaabdcab	aadcabdaabadbcddaabccbccddda
adddaabddcabbdcbaddcad	bbadabacbdcddacdacddbadccda
cbcdcabdcabddcaadcaa	addcacddadbabdabcdbadabdab
dcabdcaabdcaaddcaaabc	bddbcdaadcdcdaabadcdbcbcba
ddcbaaadbccdcccbabcdcbdccda	abdcdcbcccdcacbdbbcdcaddcab
cdaabadcaabddaabdacadccd	dcaabbcdcacdddcddaabbbcbcad
cbaadcdabbacbcdcdabbabdc	ddcbddcbddacddaadddcaaddacd
dcccdcaaddadcdacddacdcb	cabcabdabccdcadbcaaddaabdab
dabcbdadbcdbdcdcdaddcba	bcdcaadcbbcacbdcbddcadbcaab
bdcadacddababdcbadccbcd	dcbdbcadbddadcddaabddcadddca
aadcaabdccbcdcabbaabcdd	dcdccbabaaadcdabcddadbddcbc
bdcacddaabbdacdcdcacbdcc	dcaaddcaddcdaccbabddabdabdc
daaabcabadbaddcdadcaadc	abddcabdcadddcadbdacdcbaaab
cddbcdaadcdbadcacddacdaad	dadcabcadabbdadcbdaadabdaadcbcb
dabcbcddaadcbcdcaccbdcbad	aadcadaadcbcddcbddcbabdadcbdcad

Using the training set according to formulas (6), (7), reference code words of the indicated classes were determined:

$S_0^{(CAD)}$ = adcbdadcadabdabcadabdadcbdab,

$S_0^{(Healthy)}$ = cbcdcdabdcabddcaadcaa.

For illustration, we present the results of the ECG assessment of a verified sick patient (male, 69 years old) whose codogram had the form

$S_t^{(1)}$ = adcabdadcadabdaddabdaadabdbdda

and a representative of the control group — a 54-year-old man whose codogram looked like,

$S_t^{(2)}$ = bdcbbcdcabcdcabcdcbaa.

It is easy to verify that $L(S_t^{(1)}, S_0^{(CAD)}) = 13$ and $L(S_t^{(1)}, S_0^{(Healthy)}) = 15$, i.e.

$$L(S_t^{(1)}, S_0^{(CAD)}) < L(S_t^{(1)}, S_0^{(Healthy)})$$

and in accordance with rule (8), the subject was assigned to the CAD group.

Similarly, for the second subject we have $L(S_t^{(2)}, S_0^{(CAD)}) = 14$ and $L(S_t^{(1)}, S_0^{(Healthy)}) = 8$, i.e.,

$$L(S_t^{(1)}, S_0^{(CAD)}) > L(S_t^{(1)}, S_0^{(Healthy)})$$

and in accordance with rule (9) the subject was assigned to the healthy group.

Clinical studies have shown that, despite the presence of traditional electrocardiographic signs on the ECG of patients with coronary artery disease (CAD), decision rule 1 provides sensitivity $S_E = 72\%$ and specificity $S_P = 79\%$.

Fig. 4 presents estimates of the conditional distributions of Levenshtein distances with respect to the reference codograms of the sick $P(L(S_t, S_0^{(CAD)}))$ and the Healthy $P(L(S_t, S_0^{(Healthy)}))$.

Checking the hypothesis about the homogeneity of conditional distributions $P(L(S_t, S_0^{(CAD)}))$ and $P(L(S_t, S_0^{(Healthy)}))$ according to the Kolmogorov-Smirnov criterion showed, that with high statistical significance ($p < 0{,}001$) the hypothesis of equality of distributions must be rejected. A similar fact was confirmed by the Man-Whitney test for independent samples.

Statistically significant difference in conditional distributions $P(L(S_t, S_0^{(CAD)}))$ and $P(L(S_t, S_0^{(Healthy)}))$ allows to hypothesize that, the Levenshtein distance is not only a useful diagnostic feature, but also beneficial in combination with other diagnostic features [11]. Therefore, the next stage of our research was aimed at finding additional diagnostic signs that would increase the sensitivity and specificity of the decision rule (8), (9).

Extension of the Basic Method

According to [12], the probability of the appearance of symbols in code words carries valuable information in the linguistic analysis of physiological signals. Based on this idea, lets introduce into consideration three types of patterns, which are substrings of code words of the training set:

- $\pi_1 = x,\ x \in \{a, b, c, d\}$ — one-character pattern;
- $\pi_2 = xy,\ x, y \in \{a, b, c, d\}$ — two-character pattern;
- $\pi_3 = xyz,\ x, y, z \in \{a, b, c, d\}$ — three-character pattern.

Based on the data of the training sample, we calculate the average frequencies $P^{(Healthy)}(\pi_1)$ and $P^{(CAD)}(\pi_1)$ of the appearance of single-symbol patterns in the ECG code words of healthy and sick groups:

$$P^{(Healthy)}(\pi_1) = \frac{1}{Q_{Healthy}} \sum_{i=1}^{Q_{Healthy}} \frac{G_i(\pi_1)}{W_i}, \qquad (10)$$

$$P^{(CAD)}(\pi_2) = \frac{1}{Q_{CADA}} \sum_{i=1}^{Q_{CAD}} \frac{G_i(\pi_1)}{W_i}, \qquad (11)$$

where $G(\pi_1)$ — the number of occurrences of a two-character pattern $\pi_1 = x,\ x \in \{a, b, c, d\}$ in the i-th codeword, W_1 — the total number of characters in the i-th codeword, $Q_{Healthy}$ и Q_{CAD} — number of ECGs of healthy and CAD in the training set.

The results of evaluating the frequency $P^{(Healthy)}(\pi_1)$ and $P^{(CAD)}(\pi_1)$ of occurrence of single-character patterns in groups are summarized in Table 4. To assess the statistical significance (*p-value* of deviations in the mean frequency $P^{(Healthy)}(\pi_1)$ and $P^{(CAD)}(\pi_1)$ was used the Student's test, since checking by the Kolmogorov-Smirnov test confirmed the normal distribution of the appearance of patterns π_1 in the codewords.

Table 4 shows that the one-character pattern $\pi_1 = d$ occurs in the codewords of healthy patients and patients with CAD with different probabilities. This fact made it possible to formulate the second decisive rule:

Decision rule 2:

CAD, if

$$\left|\frac{G_t(d)}{W_t} - P^{(CAD)}(d)\right| \le \left|\frac{G_t(d)}{W_t} - P^{(Healthy)}(d)\right|, \quad (12)$$

Healthy, if

$$\left|\frac{G_t(d)}{W_t} - P^{(CAD)}(d)\right| > \left|\frac{G_t(d)}{W_t} - P^{(Healthy)}(d)\right|, \quad (13)$$

Else uncertain.

where $G_t(d)$ — number of occurrences of a character $x = d$ into the analyzed codeword of length W_t, a $P^{(CAD)}(d) \approx 0{,}311$ and $P^{(Healthy)}(d) \approx 0{,}286$ — estimating the probabilities of a one-character pattern $\pi_1 = d$ in the codewords of the corresponding groups.

However, the single-character pattern $\pi = d$ provides low reliability of the decisions made: the decision rule (12), (13) provided sensitivity $S_E = 59\%$ and specificity $S_p = 58\%$.

Therefore, the diagnostic value of the decision rule that was investigated, which allows to make a decision with the Levenshtein distances together with an estimate of the number of occurrences. $G_t(d)$ pattern $\pi = d$ into the analyzed codeword. Thus, it somewhat improves the decision rule 1.

Decision rule 3:

CAD, if $L(S_t, S_0^{(CAD)}) \le L(S_t, S_0^{(Healthy)})$ **AND**

$$\left|\frac{G_t(\pi_1)}{W_t} - P^{(CAD)}(d)\right| \le \left|\frac{G_t(\pi_1)}{W_t} - P^{(Healthy)}(d)\right|, \quad (14)$$

Healthy, if $L(S_t, S_0^{(CAD)}) > L(S_t, S_0^{(Healthy)})$ **AND**

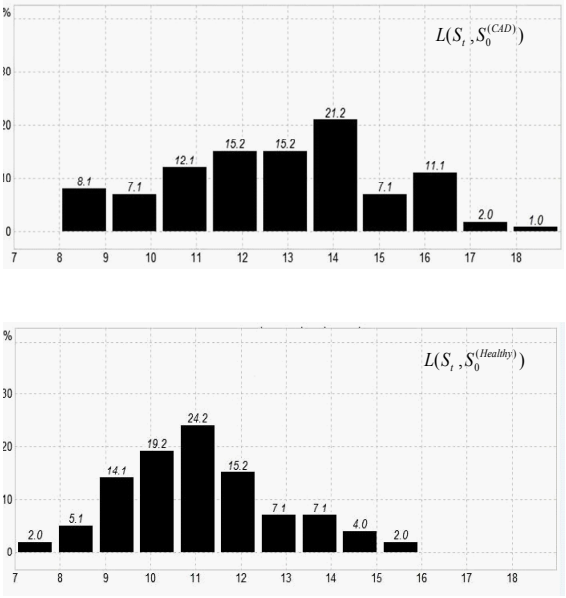

Fig. 6. Conditional distributions of the Levenshtein distances to the reference codograms CAD and Healthy

$$\left|\frac{G_t(\pi_1)}{W_t} - P^{(CAD)}(d)\right| \le \left|\frac{G_t(\pi_1)}{W_t} - P^{(Healthy)}(d)\right|, \quad (15)$$

Else uncertain.

The rule (14), (15) provided decision with sensitivity $S_E = 77{,}2\%$ and specificity $S_p = 86{,}2\%$.

Consider the possibility of improving the decision rule by analyzing two-symbol patterns $\pi_2 = xy$, $x, y \in \{a, b, c, d\}$. Table 5 presents the results of assessing the probability of the appearance of such patterns in the words of the training sample.

Since the variety of two-pattern π_2 characters is greater than that of single-pattern π_1, characters, the frequency of occurrence π_2 is low. Nevertheless, as can be seen from Table 5, the two-character pattern $\pi_2 = ab$ is almost twice as common in the

Table 4. **Estimating the frequencies of occurrence of one-symbol patterns**

Pattern	$P^{(Healthy)}(\pi_1)$	$P^{(CAD)}(\pi_1)$	*p-value*
a	0,273275	0,274386	0,907793
b	0,203286	0,196676	0,504594
c	0,236435	0,217466	0,082603
d	0,286128	0,311056	0,024375

codewords of healthy patients than in the group of CAD patients, moreover, these differences are statistically significant with high reliability (*p-value* = 0,0001).

Taking this fact into account, we formulate the decision rule, based on joint analysis of the Levenshtein distances and pattern frequency $\pi_2 = ba$.

Decision rule 4:

CAD, if $L(S_t, S_0^{(CAD)}) \le L(S_t, S_0^{(Healthy)})$ **AND**

$$\left|\frac{G_t(ba)}{W_t - 1} - P^{(CAD)}(ba)\right| \le \left|\frac{G_t(ba)}{W_t - 1} - P^{(Healthy)}(ba)\right|, \quad (16)$$

Healthy, if $L(S_t, S_0^{(CAD)}) > L(S_t, S_0^{(Healthy)})$ **AND**

$$\left|\frac{G_t(ba)}{W_t - 1} - P^{(CAD)}(ba)\right| > \left|\frac{G_t(ba)}{W_t - 1} - P^{(Healthy)}(ba)\right|, \quad (17)$$

Else uncertain,

where $G_t(ab)$ — the number of occurrences of a two-character pattern $\pi_2 = ba$ into the analyzed codeword of length W_t.

The rule (16), (17) provided decision with sensitivity $S_E = 77{,}3\%$ and specificity $S_P = 89{,}1\%$ for most of the training set codewords.

Let's consider the option of making a decision with a simultaneous estimation of the Levenstein distances and the number of occurrences of single-character and two-character patterns in a code word. Such decisions form decision rule 5.

Decision rule 5:

CAD, if $L(S_t, S_0^{(CAD)}) \le L(S_t, S_0^{(Healthy)})$ **AND**

$$\left|\frac{G_t(\pi_1)}{W_t} - P^{(CAD)}(d)\right| \le \left|\frac{G_t(\pi_1)}{W_t} - P^{(Healthy)}(d)\right|$$

OR

$$\left|\frac{G_t(ba)}{W_t - 1} - P^{(CAD)}(ba)\right| \le \left|\frac{G_t(ba)}{W_t - 1} - P^{(Healthy)}(ba)\right| \quad (18)$$

Healthy, if $L(S_t, S_0^{(CAD)}) > L(S_t, S_0^{(Healthy)})$

AND $\left|\frac{G_t(d)}{W_t - 1} - P^{(CAD)}(d)\right| \ge \left|\frac{G_t(d)}{W_t - 1} - P^{(Healthy)}(d)\right|$

Table 5. **Estimating the frequencies of occurrence of two-symbol patterns**

Pattern	$P^{(Healthy)}\|(\pi_2)$	$P^{(CAD)}\|(\pi_2)$	*p*-value
aa	0,059126	0,038845	0,000189
ab	0,091659	0,093691	0,400061
ac	0,025715	0,019925	0,154795
ad	0,094804	0,118936	0,012481
ba	**0,048478**	**0,027195**	**0,00001**
bb	0,02254	0,015407	0,024597
bc	0,059796	0,063595	0,385212
bd	0,071602	0,084538	0,166476
ca	0,082968	0,086643	0,376135
cb	0,065612	0,064056	0,208652
cc	0,026535	0,020845	0,087007
cd	0,0544	0,044895	0,010992
da	0,079259	0,120013	0,000026
db	0,021287	0,022272	0,453138
dc	0,123771	0,111342	0,044878
dd	0,057954	0,053647	0,050011

OR

$$\left|\frac{G_t(ba)}{W_t - 1} - P^{(CAD)}(ba)\right| \geq \left|\frac{G_t(ba)}{W_t - 1} - P^{(Healthy)}(ba)\right|. \quad (19)$$

Else uncertain.

Using rule (18), (19) allows making the decisions with sensitivity S_E = 76,5% and specificity S_P = 84,7% for more than 80 % ECG of training set.

Finally, let's consider the diagnostic capabilities of the decision rule, which provides for the analysis of three-symbol patterns $\pi_3 = xyz$, $x, y, z \in \{a, b, c, d\}$. Table 6 shows the results confirming statistically significant ($p < 0{,}05$) differences in the probability of the appearance of such patterns in words corresponding to the first and second groups of the training sample.

Table 6 shows that the pattern $\pi_3 = dad$ with high statistical significance is more typical for the group of patients with CAD, and the pattern $\pi_3 = caa$ is more typical for the group of healthy. For illustration, fig. 7 shows parts of real ECGs that generate these patterns in code words. And although the presented fragments are visually almost indistinguishable, the proposed ECG processing algorithm provides an unambiguous assignment of such fragments to the pattern. $\pi_3 = dad$ or $\pi_3 = caa$.

This allowed to propose another decision rule based on the analysis of the Levenshtein distances and the number of occurrences of three-character patterns in the codeword.

Decision rule 6:

CAD, if $L(S_t, S_0^{(CAD)}) \leq L(S_t, S_0^{(Healthy)})$ **AND**

$$G_t(dad) \geq G_t(caa), \quad (20)$$

Healthy, if $L(S_t, S_0^{(CAD)}) > L(S_t, S_0^{(Healthy)})$ **AND**

$$G_t(dad) \leq G_t(caa), \quad (21)$$

Else uncertain,

where $G_t(dad)$ — the number of occurrences of a three-character pattern $\pi_3 = dad$ into the analyzed codeword of length W_t, a $G_t(caa)$ — the number of

Table 6. **Estimating the frequencies of occurrence of three-symbol patterns**

Pattern	$P^{(Healthy)}(\pi_3)$	$P^{(Healthy)}(\pi_3)$	*p-value*
ada	0,009104	0,029465	0,000006
add	0,031545	0,025732	0,017603
baa	0,007682	0,003707	0,045142
bad	0,023236	0,012264	0,00159
bca	0,013526	0,02479	0,003743
bda	0,021176	0,042309	0,000412
bdc	0,034448	0,025401	0,003428
caa	0,01891	0,007792	0,002103
cab	0,034292	0,026427	0,015428
cad	0,025595	0,051444	0,00015
cba	0,023889	0,011414	0,000466
cbb	0,012348	0,006269	0,021577
cbd	0,018828	0,028515	0,027272
cdc	0,019029	0,010455	0,009801
dab	0,028581	0,052221	0,00127
dad	0,01142	0,031005	0,000021
dba	0,009673	0,007151	0,11723

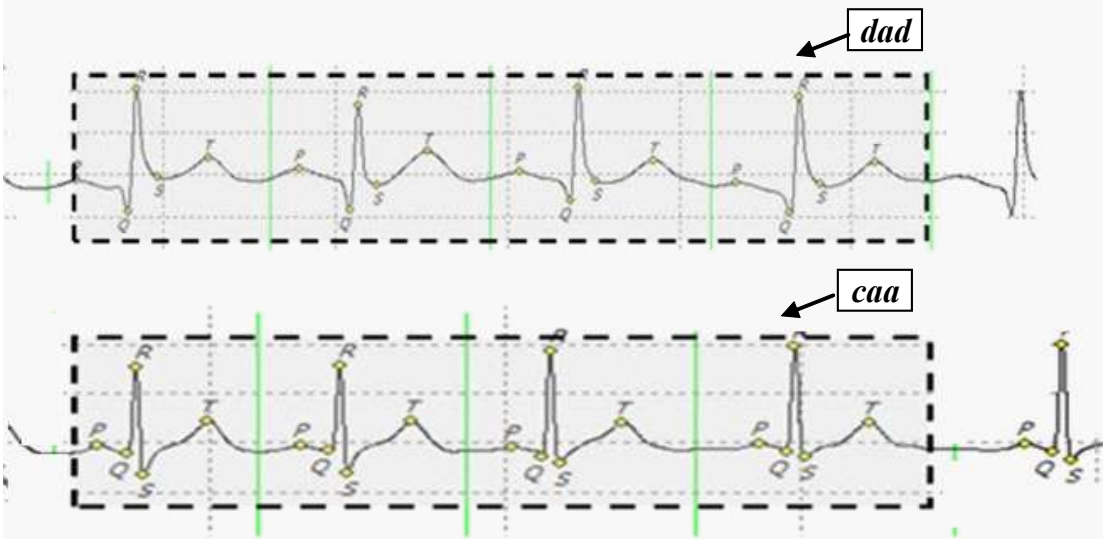

Fig. 7. Patterns for group CAD (*dad*) and healthy group (*caa*)

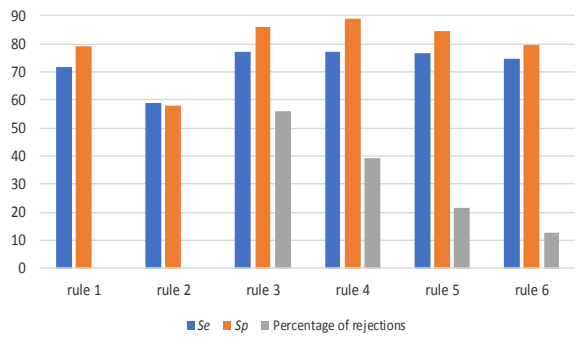

Fig. 8. Comparative characteristics of the developed decision rules

occurrences of a three-character pattern $\pi_3 = caa$ the number of occurrences of a three-character pattern W_t.

Rule (20), (21) provided decision with sensitivity $S_E = 74{,}7\%$ and specificity $S_P = 79{,}5\%$, and the number of decisions increased to 87,5%.

In the table 7 and fig. 8 summarized the results of assessing the diagnostic capabilities of the developed decision rules.

Table 7. **Operational characteristics of decision rules**

Decision rules	S_E, %	S_P,%	Percentage of rejections, %
1	72	79	0
2	59	58	0
3	77,2	86,2	56
4	77,3	89,1	39,5
5	76,5	84,7	21,5
6	74,7	79,5	12,5

CONCLUSIONS

The developed approach is based on the analysis of the dynamics of two ECG indicators calculated on the sequence of cardiac cycles. The first indicator (traditional) represents the duration of an individual cycle, and the second (original) indicator characterizes the symmetry of the *T* wave.

The proposed method is based on the transition from the calculated sequence of the specified indicators to the codeword encoding the analyzed ECG. This coding method made it possible to use the techniques of mathematical linguistics to solve the problem of analyzing and interpreting real ECGs.

Six decision rules based on the analysis of the Levenshtein distances and the number of occurrences of characteristic patterns of codewords are considered and investigated. Based on the carried out studies, it is researched that the proposed approach allows obtaining additional diagnostic information on real ECGs, which are missing the electrocardiographic signs of CAD, adopted in traditional electrocardiology.

REFERENCES

1. Connolly D. C., Elveback L. R., Oxman H. A., 1984. "Coronary heart disease in residents of Rochester, Minnesota. IV. Prognostic value of the resting electrocardiogram at the time of initial diagnosis of angina pectoris", Mayo. Clin. Proc., 59, pp. 247–250. DOI: 10.1016/s0025-6196(12)61257-9.
2. Gritsenko V. I., Fainzilberg L. S., 2019. Intellektualnyye informatsionnyye tekhnologii v tsifrovoy meditsine na primere fazagrafii [Intelligent information technologies in digital medicine on the phase-graphy example], Naukova Dumka, Kyiv, 423 p. (In Russian).
3. Dyachuk D. D., Kravchenko A. N., Faynzilberg L. S., Stanislavskaya S. S., Korchinskaya Z. A., Orikhovskaya K. B., Pasko V. S., Mikhalev K. A., 2016. "Skrining ishemii miokarda metodom otsenki fazy repolyarizatsii" ["Screening of myocardial ischemia by the method of assessing the phase of repolarization], Ukrainian Journal of Cardiology, 6, pp. 82–89. (In Russian).
4. Dyachuk D. D., Gritsenko V. I., Fainzilberg L. S., Kravchenko A. M., et. al., 2017. "Zastosuvannya metodu fazahrafiyi pry provedenni skryninhu ishemichnoyi khvoroby sertsya" ["The use of the method of fasegraphy in the screening of coronary artery disease"], Methodological Recommendations of the Ministry of Health of Ukraine № 163.16/13.17, Ukrainian Center for Scientific Medical Information and Patent and License Work, Kyiv, 32 p. (In Ukrainian).
5. Fainzilberg L. S., 2020. "New Approaches to the Analysis and Interpretation of the Shape of Cyclic Signals", Cybernetics and Systems Analysis, 56 (4), pp. 665–674. DOI: 10.1007/s10559-020-00283-0.
6. Fainzilberg L. S., Dykach Ju. R., 2019. "Linguistic approach for estimation of electrocardiograms's subtle changes based on the Levenstein distance", Cybernetics and Computer Engineering, 2 (196), pp. 3–26. DOI: 10.15407/kvt196.02.003.
7. Uspenskiy V. M., 2012. "Diagnostic System Based on the Information Analysis of Electrocardiogram", Proceedings of Mediterranean Conference on Embedded Computing, MECO 2012, June 19–21, Montenegro, pp. 74–76.
8. Kolesnikova O. V., Krivenko S. S., 2018. "Informatsiynyy analiz elektrokardiosyhnaliv: ob-hruntuvannya i mozhlyvosti" ["Information analysis of electrocardiosignals: rationale and possibilities"], Proceedings 1st International Scientific and Practical conference "Information systems and technologies in medicine", ISM-2018, KHNURE, Kharkiv, pp. 161–163. (In Ukrainian).
9. Levenshteyn V. I., 1965. "Dvoichnyye kody s ispravleniyem vypadeniy, vstavok i zameshcheniy simvolov" ["Binary codes with correction of occurrences, inserts and symbol substitutions"], dokl. Academy of Sciences of the USSR, 163 (4), pp. 845–848. (In Russian).
10. Wagner R. A., Fischer M. J., 1971. "The String-to-String Correction Problem", Journal of the ACM, 21 (1), pp. 168–173. DOI: 10.1145/321796.321811.
11. Faynzilberg L. S., 2010. Matematicheskiye metody otsenki poleznosti diagnosticheskikh priznakov [Mathematical methods for evaluating the usefulness of diagnostic features], Osvita Ukrainy, Kyiv, 152 p. (In Russian).
12. Senkevich Yu. I., 2008. "Lingvisticheskiy analiz fiziologicheskikh signalov" ["Linguistic analysis of physiological signals"], Digital Signal Processing, 2, pp. 54–57. (In Russian).

Received 06.04.2021

ЛІТЕРАТУРА

1. Connolly D. C., Elveback L. R., Oxman H. A. Coronary heart disease in residents of Rochester, Minnesota. IV. Prognostic value of the resting electrocardiogram at the time of initial diagnosis of angina pectoris. Mayo. Clin. Proc. 1984. 59. P. 247–250. DOI: https://doi.org/10.1016/s0025-6196(12)61257-9.
2. Гриценко В. И., Файнзильберг Л. С. Интеллектуальные информационные технологии в цифровой медицине на примере фазаграфии. Киев : Наукова Думка, 2019. 423 с.
3. Дячук Д. Д., Кравченко А. Н., Файнзильберг Л. С., Станиславская С. С., Корчинская З. А., Ориховская К. Б., Пасько В. С., Михалев К. А. Скрининг ишемии миокарда методом оценки фазы реполяризации. Український кардіологічний журнал. 2016. 6. С. 82–89.
4. Дячук Д. Д., Гриценко В. І., Файнзільберг Л. С., Кравченко А. М. и др. Застосування методу фазаграфії при проведенні скринінгу ішемічної хвороби серця. Методичні рекомендації МОЗ України № 163.16/13.17. Київ : Український центр наукової медичної інформації і патентно-ліцензійної роботи, 2017. 32 с.
5. Fainzilberg L. S. New Approaches to the Analysis and Interpretation of the Shape of Cyclic Signals. Cybernetics and Systems Analysis. 2020. 56 (4). P. 665–674. DOI: https://doi.org/10.1007/s10559-020-00283-0.
6. Fainzilberg L. S., Dykach Ju. R. Linguistic approach for estimation of electrocardiograms's subtle changes based on the Levenstein distance. Cybernetics and Computer Engineering. 2019. 2 (196). P. 3–26. DOI: https://doi.org/10.15407/kvt196.02.003.
7. Uspenskiy V. M. Diagnostic System Based on the Information Analysis of Electrocardiogram. MECO 2012 : Proceedings of Mediterranean Conference on Embedded Computing (Montenegro, June 19–21). 2012. P. 74–76.
8. Колеснікова О. В., Крівенко С. С. Інформаційний аналіз електрокардіосигналів: обгрунтування і можливості. ISM–2018 : збірник наукових праць Першої Міжнародної науково-практичної конференції «Інформаційні системи та технології в медицині». Харків : ХНУРЕ, 2018. С. 161–163.
9. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов : докл. АН СССР. 1965. 163 (4). С. 845–848.
10. Wagner R. A., Fischer M. J. The String-to-String Correction Problem. Journal of the ACM. 1971. 21 (1). P. 168–173. DOI: https://doi.org/10.1145/321796.321811.
11. Файнзильберг Л. С. Математические методы оценки полезности диагностических признаков. Киев : Освита Украины, 2010. 152 с.
12. Сенкевич Ю. И. Лингвистический анализ физиологических сигналов. Цифровая обработка сигналов. 2008. 2. С. 54–57.

Надійшла 06.04.2021

Л.С. Файнзільберг, доктор технічних наук, професор, головний науковий співробітник,
Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України,
03187, м. Київ, просп. Академіка Глушкова, 40, Україна,
fainzilberg@gmail.com

Ю.Р. Дикач, студентка факультету біомедичної інженерії, Національний техн. ун-т України
«Київський політехнічний інститут імені Ігоря Сікорського» НТУУ «КПІ ім. І. Сікорського»,
03056, м. Київ, просп. Перемоги, 37, Україна,
jul.dykach@gmail.com

РОЗВИТОК ЛІНГВІСТИЧНОГО ПІДХОДУ ДО ЗАДАЧІ КОМП'ЮТЕРНОЇ КЛАСИФІКАЦІЇ ЕЛЕКТРОКАРДІОГРАМ

Вступ. Лінгвістичний підхід, що заснований на переході від спостережуваного циклічного сигналу до послідовності символів (кодового слова), які характеризують динаміку показників від циклу до циклу, дає змогу використовувати процедури математичної лінгвістики для підвищення достовірності прийнятих рішень.

Мета статті — розширення діагностичних можливостей лінгвістичного підходу до аналізу та інтерпретації електрокардіограм (ЕКГ).

Методи. Кожен цикл ЕКГ кодують одним з чотирьох символів, що характеризують зміни двох показників: традиційного (тривалість циклу) і оригінального (симетрія ділянки реполяризації).

Результати. На основі обробки реальних клінічних даних верифікованих пацієнтів і здорових волонтерів побудовано еталони хворих на хронічну форму ішемічної хвороби серця (ІХС) і здорових пацієнтів. Еталони розроблено з використанням обчислювальних процедур, прийнятих в математичній лінгвістиці — відстані Левенштейна, що являє собою мінімальну кількість операцій редагування (вставки, видалення та заміни символу), що забезпечує перехід від одного слова до іншого і частоти входження підрядка в аналізоване слово. На основі цих процедур розроблено вирішальні правила, що дають змогу ухвалювати діагностичні рішення, виходячи з відстані Левенштейна до еталонів і частоти входження одно-, дво- і трисимвольних патернів в кодові слова . Встановлено, що поєднання цих двох методів розширює діагностичні можливості лінгвістичного підходу до аналізу та інтерпретації ЕКГ.

Висновки. Показано, що застосування розроблених вирішальних правил дає змогу підвищити чутливість і специфічність діагностики навіть тоді, коли на ЕКГ відсутні традиційні електрокардіологічні ознаки ішемії міокарда.

***Ключові слова**: ЕКГ, відстань Левенштейна, частота входження підрядка в кодове слово, вирішальне правило.*