ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

UDC 612.741:519.237.8 **doi: 10.32620/reks.2023.3.06**

Gennady CHUIKO, Yevhen DARNAPUK, Olga DVORNIK, Yaroslav KRAINYK

Petro Mohyla Black Sea National University, Mykolaiv, Ukraine

ABDOMINAL ELECTROMYOGRAMS MINING: BREATHING PATTERNS OF ASLEEP ADULTS

The article's subject matter *is the processing of abdominal EMG recordings and finding breathing patterns. The* ***goal*** *is to automatically classify respiratory patterns into two classes, or clusters, by two breathing patterns, regular and irregular, using machine learning (ML) methods.* ***The object of the study was to obtain a dataset*** *of 40 randomly picked abdominal EMG recordings (sampling rate equal to 200 Hz) borrowed from the complete dataset published by the Computational Clinical Neurophysiology Laboratory and the Clinical Data Animation Laboratory of Massachusetts General Hospital.* ***The tasks to be solved are*** *as follows: finding ETS (errors-trend-seasonality) model for the EMG series using the exponential smoothing method; obtaining denoised and detrended signals; obtaining the Hurst exponents for EMGs using the power-law decaying of correlograms for the denoised and detrended signals; describing the variabilities, SNR, the outlier fractions, and Hurst exponents by robust statistics, performing correlation analysis, and Principal Components Analysis (PCA); analyzing the structure of the distant matrix by a graph-based technique; obtaining the periodograms in the frequency domain using the known Wiener-Khinchin theorem; and finding the best models and methods of classification and clusterization and evaluating them within modern Machine Learning methods.* ***The methods used are*** *exponential smoothing, the Wiener-Khinchin theorem, the graph theory method, principal component analysis, programing within MAPLE 2020, and data processing by Weka. The authors obtained the following* ***results****: 1) wide data variability has been rated with the median absolute deviations, which is the most robust statistic in this case; 2) most of the signals (38 of 40) showed frequent outliers: from a few percent up to 24.6 % of emissions; 3) these four variables: outliers' percentage, variability, SNR, and persistency factors – form the attributes of input vectors of the subjects for further Machine Learning with Weka software; 4) Manhattan distances matrix among subjects' vectors in 4D attributes space allows imaging the data set as a weighted graph, the vertices of which are subjects; 5) the weights of the graph's edges reflect distances between any pair of them. "Closeness centralities" of vertices allowed us to cluster the data set on two clusters with 11 and 29 subjects, and Weka clustering algorithms confirmed this result. 6) The learning curve shows that a sufficiently small data set (from 25 subjects) might be suitable for classification purposes.* ***Conclusions****. The scientific novelty of the results obtained is as follows: 1) the Error-Trend-Seasonality model was the same for all data sets. Abdominal EMG of sleeping patients had additive errors and undamped trends without any seasonality; 2) the correlograms' decaying according to power law had been set, and Hurst exponents were in the range (of 0.776–0.887). This testifies to "long memory" (high persistence) of abdominal EMGs; 3) the modified Z-scores and robust statistics with the highest breakdown values were used for the EMG parameters because of many outliers; 4) breathing patterns were set using the periodograms in the frequency domain using the Wiener-Khinchin theorem; 5) the new graph-based method was successfully exploited to cluster the dataset. Parallel clustering with Weka algorithms confirmed the graph-based clustering results.*

Keywords: *Abdominal Electromyogram; Breathing Patterns; Machine Learning; Variability; Outliers; Persistency.*

Introduction

Healthy sleep is a necessary part of the life cycle and circadian rhythm. Sleep upsets have become a conjoint field of study in physiology and medicine. For example, many shared efforts have been devoted to this topic: several databases of known biomedical portals [1], in particular [2], with detailed descriptions in [3]. The reasons for sleep disorders may be manifold. In particular, the authors [4], and even more coherently [5], have recently pointed out severe sleep maladies among various population layers caused by the COVID-19 pandemic [6].

The breathing pattern of sleeping adults shall be, as an ideal, regular and maintain some range of frequencies. The range (12–20) of breathing cycles per minute that meets frequencies (0.2–0.33) Hz is commonly accepted [7]. Less regular breathing characterizes the waking-up state for evident reasons.

Abnormal breathing rates (BR), outgoing from the bounds of the above range, are perhaps the most frequently admitted sign of sleep disorder. Although the BR can rise slightly at the so-called Rapid Eyes Moving (REM, fast sleep stage), this little climb drops again at the slow sleep. This slow stage of sleep is the prevailing one: 80 % versus 20 % [8].

Thus, asleep breathing is a low-frequency process that must be directly reflected by abdominal electromyograms (EMGs) if one considers the role of the diaphragm and abdominal muscles in breathing. Therefore, abdominal EMGs picked from vaster data [1, 2] randomly served as the dataset for mining BR in this paper.

Breathing patterns are the main but not the only subject of this study. The persistence estimation of EMGs as time series, their exponential smoothing model, variability and outlier studies, clustering, and heterogeneity of the dataset are also on the "shortlist of interests" of this report. For example, the variability of medical signals and EMGs displays the homeostasis phenomenon inherent in living organisms [9]. Outliers are a common problem with EMGs that cannot be neglected during their proper processing. It was first noted in [10] concerning EMGs and then developed in [11].

Insight into persistence grade and the Errors-Trend-Seasonality (ETS) model is vital for forecasting and interpreting the EMG series. The exponential smoothing method (or ETS modeling) prognoses the future results of a time series based on past results. Its feature is the exponentially decaying weights of past results [12]. The ETS model of the signal allows its decomposition, denoising, and detrending if needed.

The authors intend to knot these different aspects at first glance into one complete. Thus, non-trivial relations among the described subjects can define the novelty of this research if they can be convincingly shown. Data mining is often described as patterns and non-trivial knowledge discovery in databases. This approach relies on Machine Learning (ML) methods, robust statistics, and modern time series analysis tools.

Let us recall the machine learning tasks [13, 14]. One can stress clustering of data, anomalies (outliers) search, probability density function estimation from data, data visualization, and Principal Component Analysis (PCA). The EMG data mining in this paper will include all of the above listed.

The text above is the authors' attempt to base the study's relevancy, explain its subject and trend, and convince the reader that this topic is worthy of study. The reader can find the aims and tasks of this research in section 1.

1. Related works, aims, and tasks of the study

First, we should mention the papers [1 - 3] because our dataset is a small part of this richer data. Sources [2, 3] hold only common data descriptions and do not profoundly analyze their branches. These data ensured that studies of various sleep upsets. Apnoea of all types and hypopnea are dominant within the data [2, 3] (roughly in equal fractions [2]). The data include 13 physiological signals, and the abdominal EMGs of sleeping adults are only one of them.

A row (23 publications) of recent works is devoted to these data and work with them. Source [15] contains their list and relevant references. Most consider methods for automatically detecting certain relatively rare sleep disorders within complete data using all or most signals. Machine Learning methods were used in their pure form.

In contrast, we will consider a small data set picked randomly from the cited above data [2, 3] and source [15]. Moreover, this dataset will include only abdominal EMGs because they are correlated with sleeping respiratory patterns. Let us refer to the periodic movement of the diaphragm and its joined muscles during breathing to explain this bond.

Respiratory rate and breathing patterns (BPs) may serve as objective signs to identify lung diseases [16]. The commonly recognized classification of BPs, including abnormal ones, has existed for a long time [7, 8, 17]. However, these patterns are considered within the time domain most often. Meanwhile, the regularity or irregularity of breathing should manifest better in the frequency domain. Sad to say, but such frequency domain studies are rare.

The paper [18] is worth more detailed mention as an example of frequency domain studies. A row of respiratory indicators is presented and extracted by remote photoplethysmography using a video camera. The signal power spectrum for 12 waking volunteers is shown in the frequency domain. Dominant peaks occupied the interval (0.12–0.39) Hz, matching BR from 7 to 23 breaths per minute. In addition, two patterns, regular and irregular breathing, were presented for a volunteer. Note that similar BPs concerning asleep people will be the main subject of our paper. As one can see from [18], the pattern of regular breathing differs by a mighty and narrow peak (maybe with a pair of minor satellites) within the (0.2–0.33) Hz band. It corresponds to BR in the (12–20) breaths per minute. Irregular breathing shows lower and broader peaks, extending this band.

In contrast to the work cited above, this study intends to extract these respiratory patterns from abdominal EMGs. These patterns will be bonded with these series' variability, outlier percentages, persistence factors, and signal-to-noise ratios (SNR).

Thus, the aim of this study is defined. This study aims to automatically classify respiratory patterns into two classes, or clusters, by two breathing patterns, regular and irregular, using machine learning (ML) methods. Therefore, the input vectors include outlier percentage in series, variability, persistence factor (Hurst exponent), and SNR. We need to resolve few associated problems to attain the abovementioned aim. One can arrange them in the following order:

1 Finding an error-trend-seasonality (ETS) model for EMG series using the exponential smoothing method [11] and obtaining denoised and detrended signals.

2 Obtaining the Hurst exponents for EMGs using the power-law decaying of correlograms for the denoised and detrended signals [19, 20].

3 Describing the variabilities, SNR, outlier fractions, and Hurst exponents using robust statistics, performing correlation analysis, and Principal Components Analysis (PCA).

4 Analyzing the structure of the distant matrix using a graph-based technique.

5 Obtain the periodograms in the frequency domain using the known Wiener-Khinchin theorem [21].

6 Finding the best models and methods of classification and clusterization and evaluating them within modern Machine Learning methods.

2. Methods and data

2.1. Provenance and main features of the data

The Computational Clinical Neurophysiology Laboratory and the Clinical Data Animation Laboratory of Massachusetts General Hospital have provided their data for the 2018 Computing in Cardiology Challenge conference [2]. The data include 1,985 patients who were watched at the hospital to diagnose sleep conditions. The balanced training set (n = 994) and testing set (n = 989) comprised two parts of data [2] and [3].

One can find in [2] the means and standard deviations of such parameters as age, gender, body mass index, drug use, and reasons for a clinic visit of the subjects. Sleep and arousal features of patients have also been presented. We have not exploited these "demographic" descriptions, so these are presented here by reference instead of a large table or direct citation.

Medics recorded a set of signals as the subjects slept. The offered time series included electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), and arterial blood oxygen saturation (SaO2). EMGs were recorded at the chin, chest, and abdomen. The sampling rate was equal to 200 Hz. Almost all signals were measured in microvolts, except for SaO2 given in percentage. These data can be imported into Python, Matlab (V4 or higher), and C programs [2].

We randomly selected 40 abdominal EMG recordings from the test set. The selection of subjects for our dataset was performed by the "Random Tools" software package from Maple 2020 [22]. Thus, randomness of selection was the main criterion for creating the dataset (the studied population). The size of the studied population (40 people) may have caused some complaints about its "insufficiency." In the framework of the learning curves method [23], we will return to this point later, evaluating the classifier by performance depending on the size of the population.

Each series had an initial duration of 3 min (N = 36000 samples at the sampling rate of 200 Hz). However, we later slightly reduced it to N = 32768, an integer degree of two, for convenient calculations.

2.2. Methods as a Shortlist

Let us begin with exponential smoothing, as described in electronic resources [12, 24] and in the paper [25], since it is used to handle the data first. This method serves chiefly as a forecasting tool for time series. The main idea is that older data are much less valid for prognosis than fresh data. Therefore, older data have weights with exponential decay.

On the other hand, it is also a series modeling method called the Error-Trend-Seasonality pattern. There are over 30 types of such models. It depends on additive or multiplicative errors. Is the trend existing or not? Furthermore, is it additive or multiplicative and damped if there is a trend? Whether seasonal changes are present and whether they are additive or multiplicative.

The ETS model choice is also an optimization problem. Information criteria evaluating the goodness of fit for a model for a time series are used. It may be a Bayesian criterion, as an example.

In addition, the insight of the ETS model allows us to denoise and detrend the primary signals. Detrend Fluctuation Analysis (DFA) and similar methods of persistence studies of series demand such a prepared series. Below, we will study the power-law decay of the autocorrelation functions of detrended EMGs similar to [19, 20] but using the Wiener-Khinchin theorem [21].

Because the outliers’ problem is inherent in EMG signals, notable robust statistics are required for their description [10, 11]. This touches on the statistics that evaluate the central tendency and estimate the variability. Their breakdown values [26] must be over 25% because the fraction of outliers sometimes exceeds that. Simplified, this means that statistics have to maintain reasonable meaning under conditions up to half of the outliers in a data. Robust statistical descriptors (estimators) with high breakdown values (up to 50%) must resist outliers. Fortunately, these statistics exist.

The authors built histograms according to the methods of [27, 28]. Estimates of probability density from data were performed within the density estimation theory using kernel functions (KDE) [28]. The reader can consider these KDE curves to be highly smoothed histograms or as estimations of probability density functions.

First, we selected four numeric attributes in the input vector (image) of a subject in the dataset. These were modified z-scores of variability, outlier percentage, persistency factor, and signal-to-noise (SNR) ratio. However, the correlation analysis shows the decorrelation necessity for these attributes. Principal component analysis (PCA) was performed by diagonalizing the covariation matrix. PCA allows us to rank variables (attributes) on their deposits into total variance. In addition, PCA often allows the reduction of the dimensionality of the problem.

We used the Graph theory method to divide the dataset into two unequal subsets: the core and peripheral vertices (nodes). Both subsets define the biconnected subgraphs with weighted edges (arcs) by distance matrix. Subgraphs have well-different diameters and non-overlapping ranges of the so-called closeness centrality values. We have accepted the belonging of a Graph node (point in vector space of attributes) to one of two such ranges as the additional attribute. This attribute may be nominal ("low centrality" and "high centrality") or even binary (0 and 1). One can find various graph-based clusterization methods in the book [29]. However, we could not find either there or elsewhere an approach similar enough to that used in our study

The power spectrum is the Fourier transform of a signal's ACF. This claim is the Wiener-Khinchin theorem content, also called the Wiener–Khinchin–Einstein theorem or the Khinchin Kolmogorov theorem [20]. One of the effects of this theorem is the linear correlation among DFA scaling exponent (α), power spectra decay exponent (β), and ACF decay exponent (γ). On the basis of this theorem, we have built EMG power spectra not from the signal but from its ACFs. These spectra were used to define breathing pattern classes (regular or irregular) because they were easily separated, even visually.

3. Experiments: data preprocessing

3.1. Exponential Smoothing model and clearing of EMGs series

The optimal exponential smoothing model was the same for all EMGs. Additive errors, additive undamped trends, and no seasonality are the points of this model. The uniform ETS model shows us some series affinity within the data set.

Insight into the ETS model lets one clear off the initial series' noises and trends. According to the ETS model, one can separate noise and trends from signals. One can find autocorrelation functions (ACFs, correlograms) with the plain decay for handled raw series without noises and trends. It looks like the well-known Detrended Fluctuation Analysis (DFA) method [19]. There is the same target as well: to estimate the persistency of the series.

Note that the noises are relatively low in data [1 - 3]. So, the signal-to-noise ratio (SNR) [30] belongs to the range (40–61) dB. That can slightly surprise one because such a high SNR is not often met among medical signals. For a collation: the SNR does not exceed 20 dB for tibial EMG data [9, 11]. We have chosen the SNR of our data set as one of the numeric attributes of the input vectors for Machine Learning (it will be denoted as z4 further).

3.2. Power-law decaying of correlograms and estimations of persistency

Persistent and antipersistent time series can show hyperbolic, also called power-law, decay. Fractional Gaussian noise (FGN) is a known model with hyperbolic decay [19, 20]. Detrended fluctuation analysis (DFA) is a popular method for determining the persistency factors of time series [19]. Time series with persistent "long memory" have diverged correlation times. They differ by power-law decaying autocorrelation functions (ACF, correlograms) with lags [20]:

$$\mathrm{ACF} \sim \mathrm{L}^{-\gamma}, \qquad (1)$$

where L is a lag, and γ is the decay factor. The relation (1) must be a straight line with a slope equal to γ within the double logarithmic coordinates.

There exists a simple linear relation between DFA scaling exponent (α) and the correlogram decay factor (γ) [19, 20]:

$$\alpha = 1 - \frac{\gamma}{2}. \qquad (2)$$

This relation is a result of mentioned already Wiener-Khinchin theorem [21].

Here we want to show the power law decaying according to formula (1) for one of the correlograms, but the same straight lines are also typical for the others. Let it is a correlogram for subject ID = 6 (see Fig. 1).

Note the excellent linearity of the graph: the adjusted coefficient of determination (R-squared) is 0.9968. Other correlograms also showed acceptable linearity over several lags: L = (512–32768). Thus, the decay power law (see equation (1)) is valid in this range.

The DFA scaling indices (α) were obtained by formula (2) and had a range of (0.776–0.887). Because this range was between 0 and 1, our results were Fractional Gaussian Noise (FGN), as mentioned above.

Then, and we mean the FGN case, the DFA exponent coincides with the well-known Hurst exponent [30]. The given range of these indicators indicates relatively high stability of abdominal EMG. In other words, we dealt with a time series with sufficiently long memory [20].

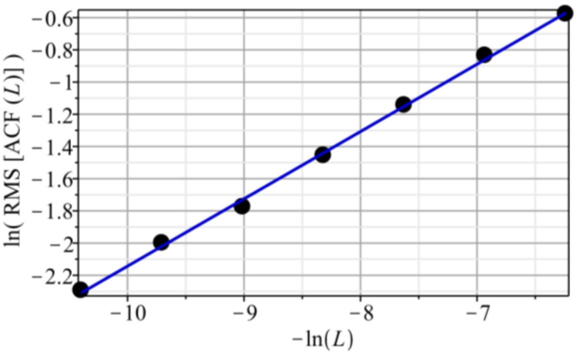

Fig. 1. Power-law decaying of a correlogram (ID=6) in double-logarithmic coordinates; here, the "RMS" abbreviation means the known operation Root of Mean Square (the arithmetic mean of the squares of a set of numbers)

The DFA exponent factor (nicknamed Hurst exponent) has also become a numeric attribute. This will be mentioned further as z3 after z-scoring.

3.3. Variability and Outliers

Variability is one of the most common traits of medical signals. This reflects the homeostasis of living things in fickle surroundings—an example of this general phenomenon is EMG [9]. This fact was first noted in [10] and then detailed in [11].

Let us consider the robust statistical indicators of variability. The interquartile ranges (IQR) are the most known among them. This indicator is stable enough for the outliers' impacts. The so-called breakdown point for IQR is equal to 1/4. In other words, IQR still has a reasonable value, even if the outliers reach a quarter of the sample [26]. Statistical box plots are a handy tool for conjoint visual analysis of IQRs and outliers [32]. Figure 2 shows the box plots for all records in our dataset.

The top whiskers show the so-called upper inner fence: (Q3+1.5IQR, where Q3 denotes the bound of the third quartile). The bottom whiskers show a lower inner fence: Q1-1.5IQR, where Q1 is the bound of the first quartile [32]. Points outside of fences are outliers (also called emissions). They can spoil the statistics but also hold valuable information [26, 33]. Therefore, one should not remove this part of the data at once without thorough analysis.

IQRs are robust enough statistics. However, fractions larger than 1/4 of outliers force us to search for even more robust descriptors for variability (see Fig. 2). Fortunately, more robust variability indexes exist with a breakdown value of up to 1/2. One is the median absolute deviation (MAD) [26, 33], which we will use further in this paper. In addition, we will prefer medians to means when analyzing the central tendencies for the same reason [10, 11]. MADs and IRQs are excellently correlated: the linear correlation coefficient is 0.9948 for our dataset.

We wrote a short program (procedure on the Maple programing language) to compute the outlier fractions in the dataset records. There are a few methods of outlier detection and box plots, and Tukey's method is the only one of them [34, 35]. We used the "median rule" [35], counting outliers.

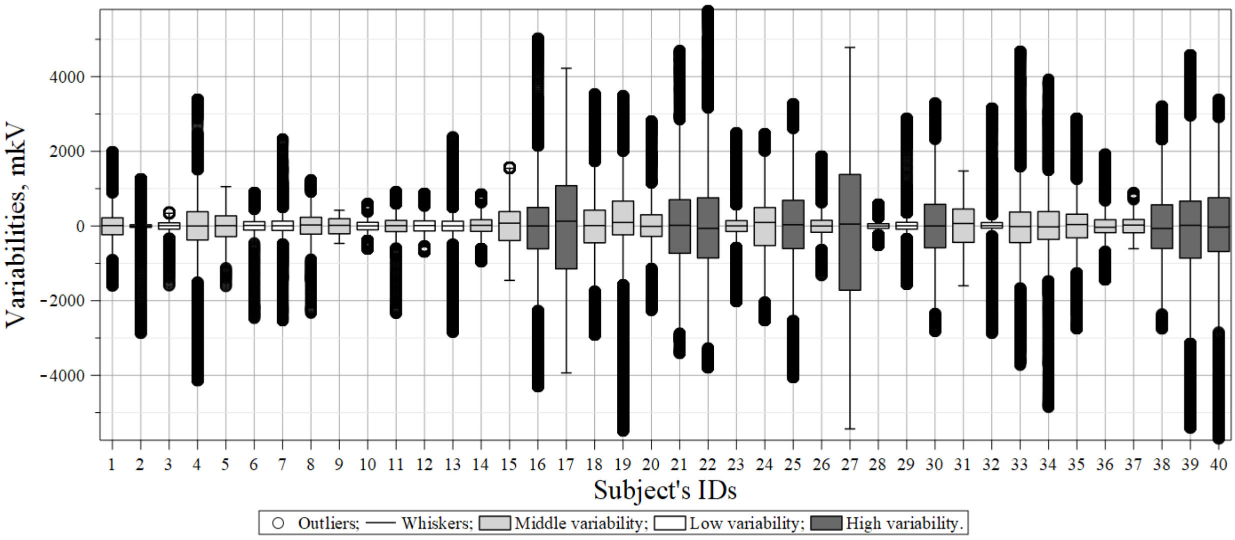

Fig. 2. The box-and-whisker plot with outliers for the abdominal EMGs for the studied data set; the horizontal axis shows the IDs of subjects (1 to 40); the vertical one is graduated in microvolts, which matches the EMG signal measure units; the height of boxes compares IQR (in mkV), the horizontal lines inside boxes show the medians; the boxes are colored according to the quartiles of variability: the first quartile (the lowest boxes highs) are white, the fourth one (highest boxes) is dark, and the second and third are middle colored

The results of outlier counting confirmed our preliminary concern: the percentage of outliers can reach up to 24.6% of the samples (ID 2, for example). The significant parts of the outliers catch the eye in Fig. 2. The individual points that mark each outlier are merged into solid bold lines. Variability (as MADs) and outlier percentages were two additional numeric attributes (z2 and z1 after z-scoring, respectively).

3.4. Norming by Z-scores and principal components analysis (PCA)

Every subject from the dataset has a specific vector representation (input vector of attributes) with the four numeric attributes mentioned above. These numeric attributes (genes in accord with another terminology) are the following: outlier fraction (in percent), variability (shown as MAD in mkV), persistence exponent, and SNR (both are dimensionless).

Good practice in pattern recognition and Machine Learning areas [13, 29] requires the norming of such vectors to obtain dimensionless and comparable numeric attributes. There are various ways to achieve this norm [13]. We have done these using modified Z-scores, which use medians and MADs [26, 33]:

$$\mathrm{z} = \frac{0.6745 \cdot (\mathrm{x} - \mathrm{Median}(\mathbf{x}))}{\mathrm{MAD}(\mathbf{x})}, \tag{3}$$

where x is an attribute mentioned above. Statistics trials, which are histograms [27], probability density functions [28], and normal plots, claim that all z-attributes have probability distributions that resemble Gauss ones skewed approximately. The highest modes for distributions were observed near zero z-scores, which was expected.

The four normalized attributes (z1, z2, z3, and z4) were pairwise correlated; therefore, they cannot be considered entirely independent. The complete correlation coefficient matrix for them is as follows:

$$\mathrm{C} = \begin{pmatrix} 1 & -0.3096 & -0.3538 & -0.0245 \\ & 1 & 0.2078 & 0.6167 \\ & & 1 & -0.1419 \\ & & & 1 \end{pmatrix}$$

These correlation coefficients might be statistically significant or not. This depends on the absolute value of the so-called critical correlation coefficient. This value separates significant and insignificant correlations. The critical values table, by Student statistics, has two inputs: the number of subjects in the population (40) and the confidence level (let it be standard 0.95). In such a case, the critical correlation coefficient is 0.2635. Thus, at least half of the matrix (4) correlations are statistically significant. Hence, the decorrelation by PCA (Principal Components Analysis) seems to be well-grounded.

PCA is not only decorrelating but also ranking principal components. The chart in Fig. 3 shows that the contribution of the z_4 principal component to the total variance is the weakest. We can neglect z_4 further because it affects only about 4% of the total variance. Such a dimensionality reduction does not mean that we ignore SNR or some of the other data because the PCA procedure has already included them all in other principal components (z_1, z_2, and z_3). Therefore, we leave only the three above principal components in a further study as numeric attributes.

There is also a cast of outliers among principal component values. Table 1 shows the IDs of such subjects, which were recognized as outliers, using three different methods [33, 34].

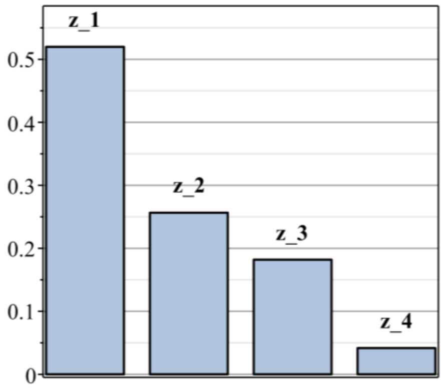

Fig. 3. Ranking of principal components: the vertical axis shows the relative contribution of the principal component to the total variance

Table 1

Subjects with abnormal values (outliers) of the principal components

Method of detection	IDs
Modified z-scoring	{2, 5, 9, 15, 19, 27, 31, 32}
Tukey's (box plot)	{2, 5, 9, 15, 19, 27, 31, 32}
Median rule	{2, 5, 9, 15, 27, 31} (4)

4. Results

4.1. Graph-based clustering of the dataset and one more categorial attribute

Various clustering methods in machine learning use the distance matrix in one way [13] and [14] or another [29]. Here, we refer to the distances among the subjects within the metric attribute vector space. Thus, the dilemma "Manhattan or Euclidean distances?" appears at the beginning of the process.

We have chosen Manhattan (also called Leming's or L1-metric). The point is that our variables are sufficiently different by their nature. In addition to being statistically independent, they have different ranges of z-scores: approximately 14.0, 8.9, and 6.0, respectively. In addition, statistics recommend considering any modified z-score with modules greater than 3.5 as potential excess (outlier). The dataset has a cast of outliers (see Table 1).

Therefore, we obtained a matrix holding 780 distances between 40 subjects, defined by points in the three-dimensional (3D) attribute vector space. One can show the points and distances among them in the mentioned space as a graph in the graph theory sense. This is a completely biconnected graph with 40 vertices and 780 weighted edges. The center of this graph is the vertex with ID 16, which has minimal eccentricity (12.854). The diameter of the graph is 20.103.

Closeness centrality is, within graph theory, the inverse of the average shortest distance between the vertex and all other vertices in the graph. The inversion is used because a higher closeness centrality indicates a greater centrality, resulting in a shorter average distance to other vertices [36]. Thus, smaller values are typical for peripheral vertices (nodes) with mostly high edge (arc) weights. The compacter core nodes have shorter edges and larger values of closeness centralities (see Fig.4).

Fig. 4 lets us consider two ranges of closeness centralities at least. The first of two is the range (0.158–0.202), which matches the right-shifted higher part of the histogram and determines the central (core) vertex subgraph. The second range (0.068–0.143) corresponds to a subgraph of the peripheral vertices. Table 2 shows some parameters of the complete graph and the two above subgraphs.

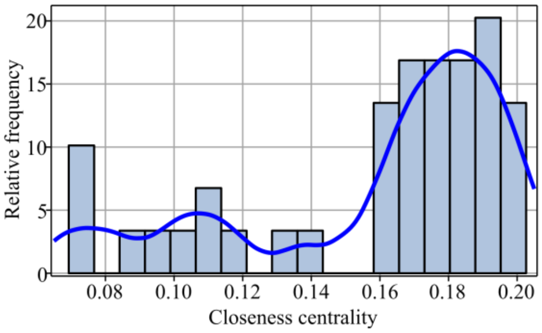

Fig. 4. Histogram and estimation of the probabilities density function [28] for closeness centrality of the vertices of the complete graph

Note that all outliers from Table 1 fall into the peripheral subgraph as a subset of its vertices. The diameter of the complete graph is equal to that of its peripheral subgraph but roughly two and a half times larger than the diameter of the central core subgraph. Thus, we can now introduce the fourth categorical (or nominal) attribute for a vertice, depending on whether it belongs to the closeness centrality ranges: low or high. This attribute may also be binary: 0 or 1, with an average weighted value of 0.75.

4.2. Power spectra and breathing patterns

Fig. 5 shows the visual odds between the breathing pattern spectra of the two representatives. The high and narrow peaks, possibly with a few minor satellites, are specific for regular breathing (left-hand side plot of Fig. 5). The irregular one finds many far lower tops over a far broader frequency range. Thus, visual breathing pattern recognition within the frequency domain is quite accessible to clinicians.

We found 10 (25%) subjects with regular breathing and 30 (75%) persons with irregular breathing among 40 sleeping adults. Curious that this percentage

Table 2

Some parameters of the complete graph and its subgraphs with low and high closeness centralities of the vertices

Object	Complete graph, biconnected	Core subgraph, biconnected	Peripheral subgraph, biconnected
Parameters	Vertices: from 1 to 40; total 40	Vertices: {1, 4, 6, 7, 8, 10, 11, 12, 13, 14, 16, 18, 20, 21, 23, 24, 25, 26, 28, 29, 30, 33, 34, 35, 36, 37, 38, 39, 40}; total 29	Vertices: {2, 3, 5, 9, 15, 17, 19, 22, 27, 31, 32}; total 11
	780 weighted edges	406 weighted edges;	55 weighted edges
	Diameter: 20.1035	Diameter: 8.0450	Diameter: 20.1035
	Central vertex: 16	Central vertex: 34	Central vertex: 22
	Closeness centrality range: (0.068–0.202)	Closeness centrality range: (0.158–0202)	Closeness centrality range: (0.068–0.143)

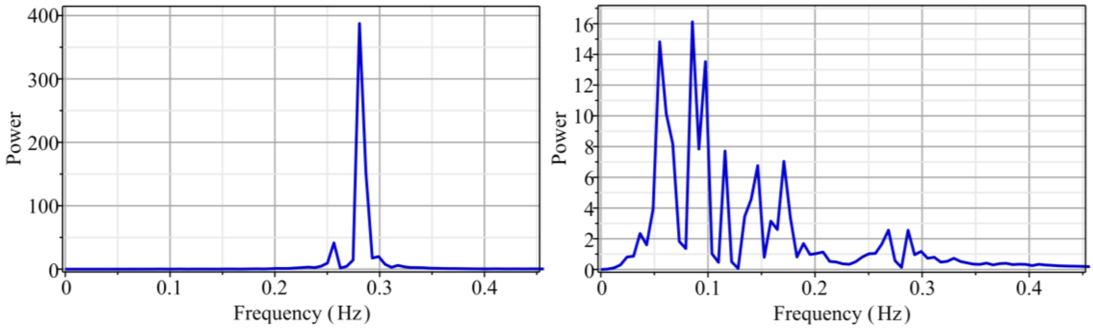

Fig. 5. Power spectra or periodograms in arbitrary units for two representatives of the regular breathing pattern (left-hand side, ID 5) and the irregular one (right-hand, ID 34); pay attention to the different scales for the power axes

relation roughly matches the number ratio of vertices in subgraphs (clusters) of the previous section (11 to 29). Therefore, the distribution into two breathing patterns is as follows:

1. Ten subjects with IDs {3, 5, 6, 9, 11, 15, 20, 27, 31, 37} belong to the second class (regular breathing pattern);

2. Thirty subjects with IDs {1, 2, 4, 7, 8, 10, 12, 13, 14, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 28, 29, 30, 32, 33, 34, 35, 36, 38, 39, 40} belong to the first one (irregular breathing).

This permits us to form a categorial attribute of the class belonging: irregular and regular. Let us summarize: we have three numeric and two nominal attributes. The last of them is the class attribute. This permits us to form a categorial attribute of the class belonging: irregular and regular.

Now we can create an ARFF (Attribute-Relation file format) file with five attributes and forty instances, input for Weka software specified to Machine Learning [37]. This file, with the extension ".txt" instead of ".arff," is attached to the paper for the readers working with Weka. They can perform their experiments with that.

4.3. Machine Learning with Weka software: classification, learning curves, clustering, and variables rank

Weka (we have worked with version 3-9-6) is an open-source data mining software that many researchers use. Weka has GUI (graphic user interface) that is handy and user-friendly. Weka Explorer allows preprocessing, classifying, clustering, filtering, and visualizing data.

The data set was preprocessed by a Weka- filter, which converts nominal attributes to binary (Weka.filters.unsupervised.attribute.NominalToBinary). This filter was used for the nominal attribute "close_centrality."

The classification test option was cross-validation with eight folds (the data set was segmented into eight parts). All trials had this option varying according to the classifiers listed in the first column of Table 3.

Table 3

Comparison of classifier performances for the dataset

Classifier	Confusion matrix		Precision	Sensitivity (Recall)	F-measure	ROC area	Kappa statistics
Bayes Network	26	4	0.800	0.800	0.800	0.757	0.4667
	4	6					
Filtered Classifier	26	4	0.800	0.800	0.800	0.810	0.4667
	4	6					
J48	27	3	0.825	0.825	0.822	0.815	0.5172
	4	6					
Random tree	29	1	0.875	0.875	0.867	0.809	0.6296
	4	6					
IMT (logistic model tree)	30	0	0.950	0.950	0.948	0.994	0.8571
	2	8					
Voted Perceptron	30	0	0.975	0.975	0.975	0.997	0.931
	1	9					
Multilayer Perceptron	30	0	0.975	0.975	0.975	1.000	0.931

Confusion matrices have an upper row matching the irregular pattern, whereas the lower one – is the regular pattern. The F-measure is the harmonic mean of the precision and sensitivity (recall). Receiver operating characteristic (ROC) is a plot illustrating a binary classifier's diagnostic ability at various decision-making thresholds. Many researchers use the area under ROC in the framework of Machine Learning for classifier comparison. The closer the ROC area is to 1.00, the better the classifier's performance. Usually, ROC area greater than 0.85 may be considered acceptable. The kappa statistics (Cohen's statistics, Interrater reliability measure [38]) is also determined by confusion matrix elements. It estimates the agreement between the two classes' evaluations. Table 3 shows various estimations from moderate agreement (0.41 - 0.60) to near perfect (0.81-0.99).

Learning curves in ML show the predictive performance as a function of the size of the population (or the number of instances), sometimes it may be the number of instances of a training set [23]. Fig. 6 shows the dependence of the ROC area as a measure of performance on the number of instances in the study population.

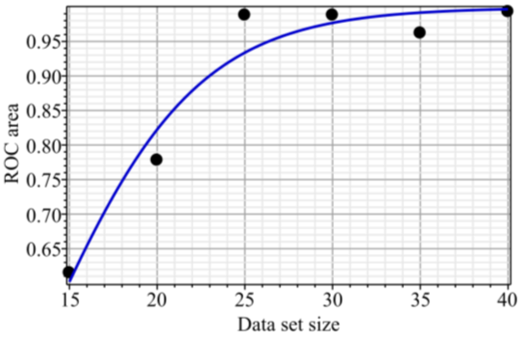

Fig. 6. The Learning curve for LMT-classifier (see Table 3): points show the results of the experiments within Weka; the logistic curve is their interpolation (fitting)

Hence, it is a learning curve. The logistic curve fitted the experimental results (points) obtained in the LMT classifier framework (logistic model tree, see Table 3). The LMT classifier performs well enough (ROC area > 0.85), already at 25 or more instances in the dataset. Note that the LMT classifier is not the champion of performance among those listed in Table 3.

The clustering algorithm finds groups of similar instances in the entire dataset. WEKA supports several clustering algorithms. Some of them are displayed in Table 4. A new, graph-based method of clustering is also presented, described above in subsection 4.1 for comparison.

One can see that the clustering result obtained by the suggested graph-based method is closest to the outcomes of the EM algorithm of Weka. Although the outputs of Simple K-mean and Density Based clusters also are near enough to those predictions. Agreement among the four methods from Table 4 appears entirely satisfactory, approving the technique suggested in section 4.1.

Weka also allows the ranking of variables (attributes). We have exploited Evaluator named "weka. Attribute selection. InfoGainAttributeEval" that worked on all data set with "Information Gain Ranking Filter." Attributes were ranked in the following order: 1. z_1(0.467); 2. z_2(0.456); 3. z_4 (close centrality, 0.118); 4. z_3 (0).

After this rating, the third numerical attribute (z_3) seems unnecessary for classification or clustering purposes (see also Fig.3). The reader can check this suspicion by changing the quantity of the attributes in the ARFF file for the dataset, which is an appendix to this article.

5. Discussions

The special processing of the abdominal EMG series described in this paper has permitted us to classify asleep adults' regular or irregular breathing patterns. In principle, this is possible by a few characteristics of their abdominal EMG using the classifiers by the Machine Learning method, with a small percentage of diagnostic errors. The initial set of such characteristics (attributes) includes variability, outlier percentage, SNR, and persistency factor (all as numeric). One more nominal (or binary) attribute belongs to one of two clusters within the dataset.

Table 4

Comparison of clustering methods

Method	Simple K-mean		EM (expectation maximization)		Density Based		Suggested Graph-Based	
Clusters	1	2	1	2	1	2	1	2
Z_1	3.9866	-0.6924	2.5046	-0.2905	3.9866	-0.6924	2.4985	-0.2916
Z_2	1.4067	0.0513	1.1049	0.1195	1.4067	0.0513	1.0645	0.0658
Z_3	-1.0150	-1.2436	-2.5759	-0.6602	-1.0150	-1.2436	-2.5834	-0.6611
Z_4	0.4	0.8333	0.0073	0.9968	0.4	0.8333	0.0	1.0
Capacity	10	30	11	29	11	29	11	29
Inter-cluster distance	6.2630		5.6962		6.2630		5.7111	

The new clustering met is suggested in this paper based on the Graph theory category as "the closeness centrality."

Another new nontrivial result is that these EMG features are linked with breathing patterns. This confirms earlier hypotheses [9, 10, 11, 29] about maintaining outliers in EMGs as carriers of specific information. Meanwhile, the haste of outliers ruling out at the start of research is typical not only for this paper [39] but also for many others devoted to biomedical signal statistics [10, 11].

Let us note another type of "outlier" mentioned in the text (see Table 1 in subsection 3.4). These subjects belong to the far periphery of the weighted biconnected graph of the dataset (Table 2 in subsection 4.1), which is quite far from the graph center. Still, it does not prevent them from being classified within the dataset later. There is a question: how often do researchers reject such data simultaneously because they spoil the orthodox statistics?

Although the outliers are responsible for some adverse effects, we obtained excellent performance indicators, at least for some classifying algorithms (see Table 3) using Weka 3-9-5 software [37]. The experimentally obtained learning curve [23], the dependence of ROC area on the population size, testifies that acceptable classification performance might be achieved even for relatively small datasets (25 and more subjects).

An essential argument suggested by the new graph-based method is that the main cluster features (number of clusters, their capacities, centers coordinates, and distances between centers) are close to the results obtained within Weka algorithms of clustering.

A novelty and the advantage of this study is the more profound insight into the role of outliers in EMGs. This paper is one of the few first steps in this direction [9, 10, 28]. As an additional essential and promising direction, we consider studies of power-law decaying of autocorrelation functions for EMGs reflecting the "long-memory" of these series. In the Introduction, we promised to stress this paper’s novelty elements. It is the following shortlist:

– finding the first uniform ETS model for abdominal EMGs that allows denoising and detrending of the signal;

– determining the persistency grade and Hurst exponent for series cleared out of noises and trends;

– a consequent and accurate account of outliers and their effects;

– applying robust statistics to abdominal EMGs with many outliers;

– the new bond finding among the breathing patterns of asleep adults on the one hand and shares of outliers, variabilities, persistency factors, and SNR of abdominal EMGs on the other.

A new method of graph-based clustering for datasets.

Returning to the aims and tasks of this study, mentioned above in section 1, we are convinced that all of them have been achieved.

Conclusions

Machine Learning with Weka software is a powerful and promising tool for medical signal processing and associated diagnostics. The detailed physiological relationships between abdominal EMG parameters and sleeping breathing patterns have not yet confirmed except for some general lore. Nonetheless, it did not prevents us from finding this bond and even using it for sleep disorder diagnostics with acceptable precision.

We classified 40 subjects' breathing patterns using cross-validations (k=8) and a few classifiers. The best performance shows classifiers on perceptions (multilayer and voted ones). At this, we did not neglect outliers, which are inherent in most EMGs. Sure, it needs special processing of the raw EMGs. The development of such a technique is one of the results of this paper. Perhaps, these details of signal handling are hardly understandable, curious, and accessible to most clinicians. However, these aspects might be "packed" into computer programs. Then clinicians become just users of them, without needing to "deep diving" inside specific details.

Let us draw a few short conclusions, recalling our tasks from section 1.

1. The Error-Trend-Seasonality model was the same for all data sets. Abdominal EMG of sleeping patients showed additive errors and undamped trends without any seasonality.

2. The correlograms’ decaying according to the power law had been set, and Hurst exponents are in the range (of 0.776–0.887). It testifies to "long memory" (high persistence) of abdominal EMGs.

3. The modified Z-scores and robust statistics with the highest breakdown values were used for the EMG parameters because of many outliers.

4. Breathing patterns were set using periodograms in the frequency domain using the Wiener-Khinchin theorem.

5. The new graph-based method was successfully exploited for clustering of the dataset. Parallel clustering with Weka algorithms confirmed the graph-based clustering results.

Acknowledgment

The authors are grateful to Maplesoft (a division of Waterloo Maple Inc.) for granting access to the licensed copy of Maple 2020.

Author Contributions: conceptualization – **Gennady Chuiko**, **Yevhen Darnapuk**, **Olga Dvornik**, **Yaroslav Krainyk**; methodology – **Gennady Chuiko**, **Yevhen Darnapuk**, **Olga Dvornik**, **Yaroslav Krainyk**; simulation – **Gennady Chuiko**, **Yevhen Darnapuk**, **Olga Dvornik**, **Yaroslav Krainyk**; validation – **Gennady Chuiko**; formal analysis – **Gennady Chuiko**, **Yevhen Darnapuk**, **Olga Dvornik**, **Yaroslav Krainyk**; investigation – **Gennady Chuiko**, **Yevhen Darnapuk**; resources – **Gennady Chuiko**, **Yevhen Darnapuk**; data curation – **Gennady Chuiko**; writing–original draft – **Gennady Chuiko**, **Yevhen Darnapuk**; writing–review and editing – **Gennady Chuiko**; supervision – **Gennady Chuiko**; project administration – **Gennady Chuiko**; funding acquisition – **Gennady Chuiko**.

All authors have read and agreed to the published version of this manuscript.

This report is a part of the research project *"Development of hardware and software complex for non-invasive monitoring of blood pressure and heart rate of dual-purpose"* with registration number *0120U101266*. Ministry of Education and Science of Ukraine financially supports this project, and the authors are grateful for that.

References

1. Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. PhysioBank, PhysioToolkit, and PhysioNet. Components of a New Research Resource for Complex Physiologic Signals. *Circulation,* 2000, vol. 101, iss. 23, pp. e215-e220. DOI: 10.1161/01.cir.101.23.e215.

2. *PhysioNet /Computing in Cardiology Challenge 2018: Training/Test Sets.* Available at: https://archive.physionet.org/physiobank/database/challenge/2018. (accessed 10.02.2023).

3. Ghassemi, M. M., Moody, B. E., Lehman, L. H., Song, C., Li, Q., Sun, H., Mark, R. G., Westover, M. B., & Clifford, G. D. You Snooze, You Win: The PhysioNet/Computing in Cardiology Challenge 2018. *Computing in Cardiology*, 2018, vol. 45. 4 p. DOI: 10.22489/cinc.2018.049.

4. Bhat, S., & Chokroverty, S. Sleep disorders and COVID-19. *Sleep Medicine,* 2022, vol. 91, pp. 253-261. DOI: 10.1016/j.sleep.2021.07.021.

5. Jahrami, H., BaHammam, A. S., Bragazzi, N. L., Saif, Z., Faris, M. A., & Vitiello, M. V. Sleep problems during COVID-19 pandemic by population: a systematic review and meta-analysis. *Journal of Clinical Sleep Medicine*, 2020, vol.17, iss. 2, pp. 299-313. DOI: 10.5664/jcsm.8930.

6. Chuiko, G., & Darnapuk, Ye. Fractal nature of arterial blood oxygen saturation data. *Radioelektronni i komp'uterni sistemi – Radioelectronic and computer systems*, 2022, no. 1, pp. 206–215. DOI: 10.32620/reks.2022.1.16.

7. Capodilupo, E. *What is the Respiratory Rate?* Available at: https://www.whoop.com/thelocker/what-is-respiratory-rate-normal/. (accessed 5.02.2023).

8. Lovell, K., & Liszewski, C. *Normal Sleep Patterns*. Available at: https://learn.chm.msu.edu/neuroed/neurobiology_disease/content/otheresources/sleepdisorders.pdf. (accessed 5.02.2023).

9. Chuiko, G., Dvornik, O., Darnapuk, Y., & Baganov, Y. Devising a new filtration method and proof of self-similarity of electromyograms. *Eastern-European Journal of Enterprise Technologies*, 2021, vol. 4, no. 9(112), pp. 15–22. DOI: 10.15587/1729-4061.2021.239165.

10. Rose, W. *KAAP686 Mathematics and Signal Processing for Biomechanics. Electromyogram analysis.* Available at: http://www1.udel.edu/biology/rosewc/kaap686/notes/EMG%20analysis.pdf. (accessed 5.02.2023).

11. Chuiko, G., Darnapuk, Ye., Dvornik, O., & Krainyk, Ya. Improved robust handling of electromyograms with mining of new diagnostic signs. *Proceedings of the 1st International Workshop on Information Technologies: Theoretical and Applied Problems 2021*, Ternopil, 16 November 2021, pp. 55–62. Available at: https://ceur-ws.org/Vol-3039/short6.pdf. (accessed 5.02.2023).

12. De Livera, A. M., & Hyndman, R. J. *Forecasting time series with complex seasonal patterns using exponential smoothing.* Monash University, Working Paper 15/09, 2009. 28 p. Available at: https://www.monash.edu/business/econometrics-and-business-statistics/research/publications/ebs/wp15-09.pdf. (accessed 5.02.2023).

13. Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer Publ., 2016. 758 p.

14. *Machine learning tasks in ML.NET.* Available at: https://learn.microsoft.com/en-us/dotnet/machine-learning/resources/tasks. (accessed 5.02.2023).

15. *Publications from You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018*. Available at: https://archive.physionet.org/challenge/2018/papers/. (accessed 5.02.2023).

16. Schertel, A., Funke-Chambour, M., Geiser, T., & Brill, A.-K. P231 Respiratory breathing patterns and cough in idiopathic pulmonary fibrosis: awake, asleep and over time. *Chest*, 2017, vol. 151, iss. 5, article no. A131. DOI: /10.1016/j.chest.2017.04.138.

17. Frank, J. I., & Hanley, D. F. *Abnormal Breathing Patterns*. In: Hacke, W., Hanley, D. F., Einhäupl, K. M., Bleck, T. P., Diringer, M. N., Ropper, A.H. (eds)

Neurocritical Care. Springer, Berlin, Heidelberg, 1994, pp. 366-373. DOI: 10.1007/978-3-642-87602-8

18. Massaroni, C., Lopes, D. S., Lo Presti, D., Schena, E., & Silvestri, S. Contactless Monitoring of Breathing Patterns and Respiratory Rate at the Pit of the Neck: A Single Camera Approach. *Journal of Sensors*, 2018, vol. 2018, article id 4567213, pp. 1–13. DOI: 10.1155/2018/4567213.

19. Lan, T.-H., Gao, Z.-Y., Abdalla, A. N., Cheng, B., & Wang, S. Detrended fluctuation analysis as a statistical method to study ion single channel signal. *Cell Biology International*, 2008, vol. 32, iss. 2, pp. 247–252. DOI: 10.1016/j.cellbi.2007.09.001.

20. Veenstra, J. Q. *Persistence and Anti-persistence: Theory and Software*. Electronic Thesis and Dissertation Repository. Western University, 2013. 131 p. Available at: https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=2414&context=etd. (accessed 5.02.2023).

21. *Wiener-Khinchin Theorem* - from Wolfram MathWorld. Available at: https://mathworld.wolfram.com/Wiener-KhinchinTheorem.html. – Title from screen. (accessed 27.11.2022).

22. Bernardin, L., Chin, P., DeMarco, P., Geddes, K. O., Hare, D. E. G., Heal, K. M., Labahn, G., May, J. P., McCarron, J., Monagan, M. B., Ohashi, D., & Vorkoetter, S. M. *Maple Programming Guide*. Springer-Verlag Berlin and Heidelberg GmbH & Co. K, 2011. 678 p. Available at: https://www.maplesoft.com/view.aspx?SF=103828/337201/ProgrammingGuide.pdf. (accessed 27.11.2022).

23. Perlich, C. *Learning Curves in Machine Learning*. In: Sammut, C., Webb, G. I. (eds) *Encyclopedia of Machine Learning*. Springer, Boston, MA., 2011, pp. 577-580. DOI: 10.1007/978-0-387-30164-8_452.

24. *Exponential Smoothing for Time Series Forecasting.* Statistics By Jim. Available at: https://statisticsbyjim.com/time-series/exponential-smoothing-time-series-forecasting/. (accessed 5.02.2023).

25. Shastri, S., Sharma, A., Mansotra, V., Sharma, A., Bhadwal, A. S., & Kumari, M. A Study on Exponential Smoothing Method for Forecasting. *International Journal of Computer Sciences and Engineering*, 2018, vol. 6, iss. 4, pp. 482–485. DOI: 10.26438/ijcse/v6i4.482485.

26. Wilcox, R. Chapter 1 – Introduction. *In Statistical Modeling and Decision Science, Introduction to Robust Estimation and Hypothesis Testing (Third Edition).* Academic Press, 2012, pp. 1–22. DOI: 10.1016/b978-0-12-386983-8.00001-9

27. Scott, D. W. Averaged shifted histogram. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2010, vol. 2, iss. 2, pp. 160–164. DOI: 10.1002/wics.54.

28. Węglarczyk, S. Kernel density estimation and its application. *ITM Web of Conferences*, 2018, vol. 23, article no. 00037. DOI: 10.1051/itmconf/20182300037.

29. Vathy-Fogarassy, Á., & Abonyi, J. *Graph-Based Clustering Algorithms.* In: Graph-Based Clustering and Data Visualization Algorithms. Springer Briefs in Computer Science. Springer, London, 2013, pp. 17–41. DOI: 10.1007/978-1-4471-5158-6_2.

30. Murphy, A., & Feger, J. *Signal-to-noise ratio (radiography).* Radiopaedia.org, 2020. DOI: 10.53347/rID-75580.

31. Movahed, M. S., Jafari, G. R., Ghasemi, F., Rahvar, S., & Tabar, M. R. R. Multifractal detrended fluctuation analysis of sunspot time series. *Journal of Statistical Mechanics: Theory and Experiment*, 2006, vol. 2006, no. 02, article no. P02003. DOI: 10.1088/1742-5468/2006/02/P02003.

32. *Chapter 152. Box Plots.* NCSS Statistical Software. 10 p. Available at: https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Box_Plots.pdf. (accessed 5.02.2023).

33. Rousseeuw, P. J., & Hubert, M. Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, 2011, vol. 1, iss. 1, pp. 73–79. DOI: 10.1002/widm.2.

34. Kamble, B., & Doke, K. Outlier Detection Approaches in Data Mining. *International Research Journal of Engineering and Technology*, 2017, vol. 04, iss. 3, pp. 634-638. Available at: https://www.irjet.net/archives/V4/i3/IRJET-V4I3171.pdf. (accessed 5.02.2023).

35. Seo, S. *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets*. Univariate Data Sets. Master's Thesis, University of Pittsburgh, 2006. 53 p. Available at: http://d-scholarship.pitt.edu/7948/1/Seo.pdf. (accessed 5.02.2023).

36. Hansen, D. L., Shneiderman, B., Smith, M. A., & Himelboim, I. Chapter 6 - Calculating and visualizing network metrics. *Calculating and visualizing network metrics. Analyzing Social Media Networks with NodeXL (Second Edition)*, 2020, pp. 79–94. DOI: 10.1016/b978-0-12-817756-3.00006-6.

37. *Weka - Machine Learning Software in Java. WEKA Manual for Version 3-9-5*, 2020. Available at: https://osdn.net/projects/sfnet_weka/downloads/documentation/3.9.x/WekaManual-3-9-5.pdf. (accessed 5.02.2023).

38. Grant, M. J., Button, C. M., & Snook, B. An Evaluation of Interrater Reliability Measures on Binary Tasks Using d-Prime. *Applied Psychological Measurement*, 2016, vol. 41, iss. 4, pp. 264–276. DOI: 10.1177/0146621616684584.

39. Yang, X., Fan, D., Ren, A., Zhao, N., & Alam, M. 5G-Based User-Centric Sensing at C-Band. *IEEE Transactions on Industrial Informatics*, 2019, vol. 15, iss. 5, pp. 3040–3047. DOI: 10.1109/tii.2019.2891738.

Received 26.04.2023, Accepted 20.09.2023

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

МАЙНИНГ АБДОМІНАЛЬНИХ ЕЛЕКТРОМІОГРАМ: ПАТЕРНИ ДИХАННЯ СПЛЯЧИХ ДОРОСЛИХ

Геннадій Чуйко, Євген Дарнапук, Ольга Дворник, Ярослав Крайник

Предметом статті є обробка записів електроміограм (ЕМГ) черевної порожнини та визначення патернів дихання. **Мета** полягає в тому, щоб автоматично класифікувати патерни дихання у двох класах або кластерах за двома патернами дихання, регулярним і нерегулярним, використовуючи методи машинного навчання (ML). **Об'єктом** дослідження був набір даних із 40 випадково вибраних записів ЕМГ черевної порожнини (частота дискретизації дорівнює 200 Гц), запозичених із повного набору даних, опублікованого Лабораторією обчислювальної клінічної нейрофізіології та Лабораторією анімації клінічних даних Массачусетської загальної лікарні. Завдання, які вирішуються: знайти модель ETS (errors-trend-seasonality) для серії ЕМГ методом експоненціального згладжування; отримання знешумлених і детрендованих сигналів; отримання показників Херста для ЕМГ з використанням степеневого закону спаду корелограм для сигналів зі зниженим шумом і з виключеним трендом; опис варіабельності, SNR, фракцій викидів і показників Херста за допомогою робастної статистики, виконання кореляційного аналізу та аналізу головних компонентів (PCA); аналіз структури віддаленої матриці графовим методом; отримання періодограм у частотній області за відомою теоремою Вінера-Хінчина; пошук найкращих моделей і методів класифікації та кластеризації та їх оцінка в рамках сучасних методів машинного навчання. Використовувані методи: експоненціальне згладжування, теорема Вінера-Хінчіна, метод теорії графів, аналіз головних компонент, програмування в MAPLE 2020 та обробка даних Weka. Автори отримали наступні результати: 1) широка мінливість даних була оцінена за допомогою медіанних абсолютних відхилень, які є найбільш надійною статистикою в цьому випадку; 2) більшість сигналів (38 із 40) показали часті викиди: від кількох відсотків до 24,6 % викидів; 3) ці чотири змінні: відсоток викидів, мінливість, SNR і фактори стійкості – формують атрибути вхідних векторів суб'єктів для подальшого машинного навчання за допомогою програмного забезпечення Weka; 4) матриця манхеттенських відстаней серед векторів суб'єктів у просторі 4D атрибутів дозволяє представити набір даних у вигляді зваженого графа, вершинами якого є суб'єкти; 5) ваги ребер графа відображають відстані між будь-якою їх парою. «Центральність близькості» вершин дозволила нам кластеризувати набір даних на два кластери з 11 і 29 суб'єктами, і алгоритми кластеризації Weka підтвердили цей результат; 6) крива навчання показує, що достатньо малий набір даних (з 25 суб'єктів) може підійти для класифікації. **Висновки.** Наукова новизна отриманих результатів полягає в наступному: 1) модель Error-Trend-Seasonality була однаковою для всіх наборів даних. Абдомінальна ЕМГ пацієнтів уві сні мала додаткові помилки та незатухаючі тренди без будь-якої сезонності; 2) розпад корелограм за степеневим законом встановлено, показники Херста знаходяться в діапазоні (0,776–0,887). Це свідчить про «довгу пам'ять» (високу стійкість) абдомінальних ЕМГ; 3) модифіковані Z-показники та надійна статистика з найвищими значеннями розбивки використовувалися для параметрів ЕМГ через багато викидів; 4) патерни дихання встановлювалися за періодограмами в частотній області з використанням теореми Вінера-Хінчіна; 5) новий метод на основі графів успішно використано для кластеризації набору даних. Паралельна кластеризація за допомогою алгоритмів Weka підтвердила результати кластеризації на основі графів.

Ключові слова: абдомінальна електроміограма; паттерни дихання; машинне навчання; варіативність; викиди; постійність.

Чуйко Геннадій Петрович – д-р фіз.-мат. наук, проф., проф. каф. комп'ютерної інженерії, Чорноморський національний університет ім. Петра Могили, Миколаїв, Україна.

Дарнапук Євген Сергійович – ст. викл. каф. комп'ютерної інженерії, Чорноморський національний університет ім. Петра Могили, Миколаїв, Україна.

Дворник Ольга Василівна – канд. фіз.-мат. наук, доц. каф. комп'ютерної інженерії, Чорноморський національний університет ім. Петра Могили, Миколаїв, Україна.

Крайник Ярослав Михайлович – канд. техн. наук, доц. каф. комп'ютерної інженерії, Чорноморський національний університет ім. Петра Могили, Миколаїв, Україна.

Gennady Chuiko – D.Sc. in Physics and Mathematics, Professor of Computer Engineering Department, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine,
e-mail: genn47@meta.ua, ORCID: 0000-0001-5590-9404.

Yevhen Darnapuk – Senior Lecturer, Department of Computer Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine,
e-mail: yevhen.darnapuk@chmnu.edu.ua, ORCID: 0000-0002-7099-5344.

Olga Dvornik – PhD in Physics and Mathematics, Associate Professor of Department of Computer Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine,
e-mail: olga.dvornik@chmnu.edu.ua, ORCID: 0000-0002-4545-1599.

Yaroslav Krainyk – PhD in Computer Science, Associate Professor of Department of Computer Engineering, Petro Mohyla Black Sea National University, Mykolaiv, Ukraine,
e-mail: yaroslav.krainyk@chmnu.edu.ua, ORCID: 0000-0002-7924-3878.