ECG analysis based on Word2Vec model

Yurii Oliinyk, Andrii Tereschenko, Igor Baklan, Elisa Beraudo

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Prosp. Peremohy, Kyiv, 03056, Ukraine

Abstract

Heart disease accounts for a significant percentage of deaths in both Ukraine and most countries. For example, every year in Ukraine more than 68% of people die from cardiovascular disease. An important factor in the fight against the disease is the prevention and detection of the disease in its early stages. The principal technique of observing the heart is electrocardiography, so it is very important to quickly and accurately analyze the electrocardiogram (ECG). In this article propose to expand the capabilities of automatic analysis of electrocardiograms by creating a Word2Vec model based on selected waves in the ECG.

Keywords 1

ECG, Word2Vec, NLP, Word Embedding, Random Forest

1. Introduction

Heart disease accounts for a significant percentage of deaths in both Ukraine and most countries. For example, every year in Ukraine more than 68% of people die from cardiovascular disease. An important factor in the fight against the disease is the prevention and detection of the disease in its early stages. The principal technique of observing the heart is electrical heart recording, so it is very important to quickly and accurately analyze the electrocardiogram (ECG). Therefore, searching new analysis capabilities for ECG analysis is a very important task.

1.1 Related work

For an instrumental research of high activity of the heart muscle is used the electrocardiography. The research can be carried out at dormant state, during exercises and while using some special medical drugs - during the ECG determines the condition of the heart muscle, heart rhythm, blood flow in the myocardium

Electrocardiography is a method of graphically recording electrical phenomena that occur in the heart muscle during the activity, from the surface of the body. The curve that gives back the electrical magnetic energy of the blood-pumping organ is named as electrocardiogram (ECG). Accordingly, the ECG [1] is a recording of fluctuations in the possible variance that occurs in the heart throughout its fervor.

Various methods are used to analyze the ECG signal. A perspective direction is the use of the method of linguistic chains, because it is fast and is designed to compare short intervals. The main concept of this method [2] is the ratio of the numerical interval to a certain letter. The length of the numeric interval can be selected differently according to the distribution, the selected alphabet

EMAIL: oliyura@gmail.com (Y. Oliinyk); avtogol1998@gmail.com (A. Tereschenko); iaa@ukr.net (I. Baklan); elisa.beraudolive@gmail.com (E. Beraudo);
ORCID: 0000-0002-7408-4927 (Y. Oliinyk); 0000-0001-6846-8089 (A. Tereschenko); 0000-0002-5274-5261 (I. Baklan); 0000-0001-7550-3620 (E. Beraudo);

characters and the range of values on the numerical series. Article D [13] extended the application of the method using Linguistic Chain Fuzzy Sets.

Different algorithms are used for signal segmentation. The Pana-Tompkins algorithm [3] is commonly accustomed to detect QRS systems in electrocardiographic signals.

The QRS complex is a depolarization of the ventricles of both atria, so it is considered the main object for the analysis of the ECG signal. This feature makes it particularly suitable for measuring heart rate and the first tool for assessing heart health. In the first variation of the physiological view of the heart proposed by Eintoven, the QRS complex consists of a downward deviation (Q-wave), a high upward deviation (R-wave) and a final downward deviation (S-wave). According to the results, the algorithm of Pan and Tompkins showed that 99.3 percent of QRS complexes were correctly detected.

The algorithm of Engel and Zelenberg[4] was proposed in 1979. It is used to detect R-peaks in the ECG signal. At first, a differentiator is applied to the input signal and then the low-pass filters are overlapped. After this there is an evaluation in the current window of the threshold value for the R-peak. The condition is checked whether the peak is maximum in a given interval. If so, we add a value to our result. The threshold values are determined each time using the maximum signal amplitude function. After determining the QRS complex in the current window we move on to the next and perform a preliminary description of the action. As a result, we obtain the values of all R-peaks for a given ECG signal. In 2002, Hamilton proposed a comprehensive algorithm [5] for detecting a QRS complex that works by scanning an ECG signal and making an appropriate assessment.

Different methods are used for data clustering [6]. A distinctive feature of this method of K-Means [7] is the existence of the centroids of each cluster. A centroid is a point in the middle of a cluster. Each object under consideration will belong to the cluster whose centroid is the closest. On the first stage, the centroids of the clusters are chosen randomly or according to a certain rule (for example, choose centroids that maximize the initial distance between the clusters). The next step is to assign each object to a specific cluster. For each object, the distance to all centroids is calculated and then the nearest one is selected. After that there is a recalculation of the coordinates of the centroids. This is repeated at each step until the coordinates of the centroids stop changing. After that, the work of the algorithm can be considered to be complete.

Spectral clustering [8] is one of the most efficient clustering algorithms due to its ability to separate nonlinear data. The efficiency of the technique is explained by that the other data from the basic space is displayed in a new space in which they can be linearly separated. The main disadvantage of this algorithm is the cubic computational complexity.

The principal component method (PCA) is a dimensional reduction method [9] that uses the orthogonal conversion of a set of a large set of variables to a smaller one, which still contains most of the information from the previous dataset.

T-distributed Stochastic Neighbor Embedding is a neural network method [10] designed by *Laurens van der Maaten and Geoffrey Hinton*. It is a technique of uncertain dimensional devaluation, well suitable for burring high-dimensional data for result in low-dimensional space (two- or three-dimensional).

In the article [11] it is offered to carry out the analysis by means of ECG language processing (ECG), which action an ECG flag similarly to processing by a natural language of the text document. The proposed structure is suitable to different biomedical operations and can also develop the efficiency of depthless neural network method. The articulation consists of fixed/infixed sets of determination that make up a discussion.

The ECG diagram meters the height (amplitude) of a given sign or diversion. The standard normalization is 10 mm (10 small boxes), equal to 1 mV. On instance, especially when the waveforms are small, double standard is used (20 mm equals 1 mv). When the sign forms are very big, half standard may be used (5 mm equals 1 mv). Paper speed and potency are usually printed on the bottom of the ECG for testimonial.

Even though normal neural network methods with the bespoke features have reached acceptable execution for ECG analysis, AI functions with the power of computerized extraction of features and depiction learning have proven to get human-level achievement in analyzing biomedical signals [16; 17; 18]. All in all, deep learning approaches need a large amount of data and are composed of many fields to be learned. As well, most of the advised methods and workflows for analyzing ECG waves are bespoken to the specific task and are not unrealizable to other biomedical problems.

Word2Vec is a pure language processing technique. The word2vec algorithm [12] uses a AI model to study word associations from a large body of text. After studying, such a model can find synonymous words or suggest additional words for a partial sentence. As the name implies, word2vec maps each single word to a particular list of numbers called a vector. Vectors are faithfully selected so that a simple arithmetic method (cosine similarity between vectors) indicates the amount of connotation similarity between the words represented by these vectors. This method implements two main constructors- Continuous Bag of Words (CBOW) and Skip-gram.

Far from the connectionist approach, such as AI, the mathematical symbolic approach is a method that deals with association with emerging patterns for signs that are forming the sequences, and it is much more suitable to recognize [19, 20]. Classic examples include genomic series analysis, music search by flake, or pure language processing using computers. If the real number time series data, such as EEG and ECG, are defined as a sequence of symbols like sentences and studied and used for analysis, they would be helpful when explaining and understanding the causes or output of feature selection, training, pattern recognition, and classification. Moreover, if unnecessary information is blackballed and only the necessary information is left in the typification stage, the complexity of computation will lose, and in some cases, the rise of analysis performance can be expected. Furthermore, an encoding-based Wave2vec time series classifier model is proposed, which applies deep learning whereby the data are vectorized and the black boxes of the deep learning stage are minimized by converting the classification problem of real number time series data, such as bio-signals, into a sequence classification problem of symbolic approach.

In this article, an attempt is made to associate functions of extracting ECG signals, creating a glossary of words, clustering methods, creating a Word2Vec model to perform ECG signal analysis in future includes anomalies detection, disease classification, signal segmentation, etc.

1.2 Researches tasks

Main aim: rising the expanding the capabilities of the automatic analysis of electrocardiograms by creating a Word2Vec model based on selected waves in the ECG. The following tasks should be solved within the framework of this research:

- selection of a data set for processing;
- splitting the ECG signal into a heartbeat sequence;
- allocation of the waves for each heartbeat;
- clustering of selected waves and creating a dictionary;
- transfer of the ECG input signal to a set of symbols, where each symbol corresponds to a certain wave (part of the ECG cycle);
- creation of words and sentences based on the created dictionary;
- Word2Vec model creation based on created words;
- the analysis based on the WORD2VEC model.

After training the WORD2VEC model, both NLP and ML methods can be used.

2. THE USAGE OF WORD2VEC MODEL FOR ECG ANALYSIS

The following describes the steps for processing and converting data using the Word2Vec model to analyze the ECG signal.

2.1. Conversion of ECG signal into a set of symbols (translation – ECG signal transformation to symbols)

Like pure languages [11], an ECG waves consists of series of tierce or quadruple different signals, including the P-wave, the QRS network, the T-wave, and the U-wave (Figure 1). Every ECG includes different types of each wave.

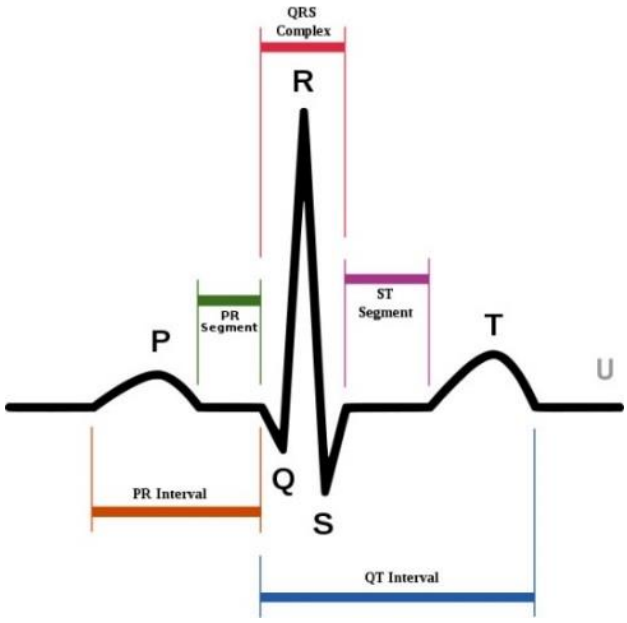

Figure 1: Schematic elements of ECG

Splitting the ECG signal into a series of the pulse. The step includes splitting the constant ECG waves into a specific heartbeat series and splitting each heartbeat into individual parts called waves. Later identifying R-waves, the existence of further components of the signals (as P, QRS and T signals) in the ECG wave can be removed using suitable search windows. To perform heartbeat signal segmentation, we identify one segment as a fixed number of instances of signal withdrawal before the location of peak R and up to a certain number of instances later on the location of peak R or from the beginning of wave P to offset of serial wave T. Figure 2 shows a segregate ECG signal stained with R- waves, P, QRS and T-waves.

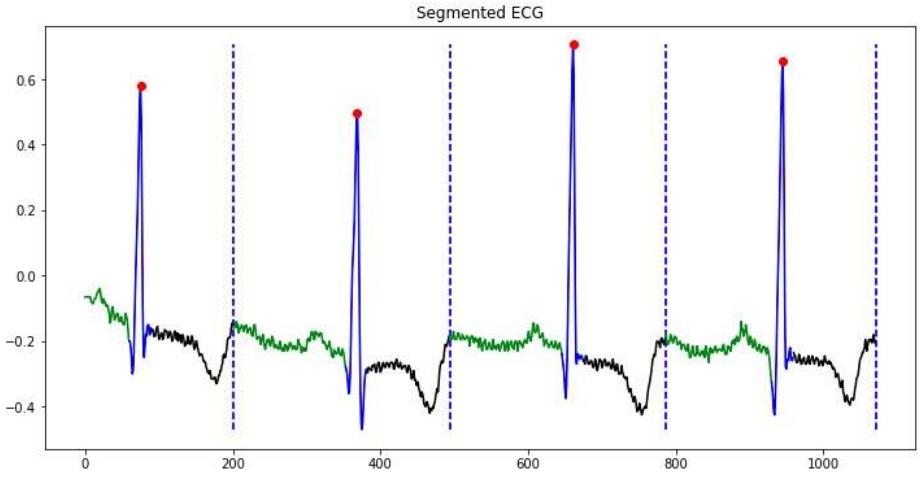

Figure 2: Segmented ECG signal

Creating a dictionary of waves. The step involves constructing a dictionary of signs found on the extricated signals from the ECG waves. By grouping the waves, we form the average value of every group as an input to the dictionary. This can be done by sustaining all signals to the input of clustering method, such as K-means, spectral clustering or agglomerative clustering algorithms. After clustering waves, the average value of every cluster may show a separate signals of dictionary. Illustration 3 shows the clustering of an ECG dataset using the t-distributed probabilistic neighborhood (t-SNE) method.

Wave segmentation. The process of segmentation of broken waves creates a series of signs for each ECG waves. Then, the clutch of each sign in the order is picked out by the return of the foregoing step (as the previous stage of the conduit). Namely, it allocates a specific symbol (which corresponds to a specific cluster) to each sign in the sequence. Thus, every ECG wave is ciphered by a sequence of characters, so that each character [a-z] constitutes a specific sign (or cluster) in the dictionary.

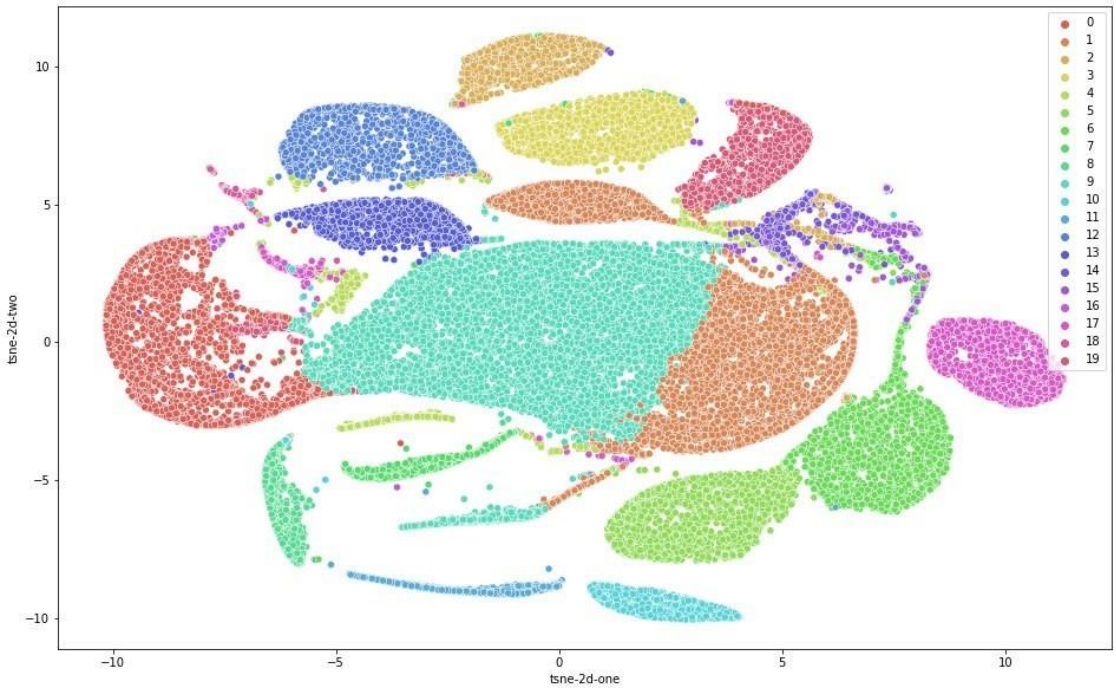

Figure 3: The result of clustering the waves by the t-SNE method

Wave Vectorization. At this stage, we take the coded dictionary of selected waves and build an embedding vector (like, a transmitter with a given distance) for each bid of the lexis. The basic case for implanting words is that permit us to try an improved model based on both AI and entity coded ECG signals for a particular task. Using NLP, we can use different approaches, such as a graph vectorizer, in which a sequence of waves is converted into a vector of fixed length with the size of the vocabulary (Figure 4). The value at each position in the vector will be the count of each wave in the encoded signal, or the Word2Vec way, which uses machine learning methods to show signs in vector area. The closing viewpoint is more effective because it acknowledges the circumstances, relationships, and similarities between waves.

```
('dys', 0.33801624178886414),
('jvf', 0.3240267038345337),
('ivg', 0.30510637164115906),
('jye', 0.29708006978034973),
('nyd', 0.2733161151409149),
('gwe', 0.2686147093772888),
('rvq', 0.26842424273490906),
('oul', 0.25311803817749023),
('nve', 0.24161389470100403),
('rum', 0.23714938759803772)]
```

Fig. 4. An example of the linguistic chains vectorization

2.2 Model training

After converting the entire ECG signal into a sentence, we can proceed to the creation and practicing of the Word2Vec model. We use the skip-gram architecture as the basis of our model, which, unlike CBOW, considers the central word from the wreath and provides contextual words.

After learning the neural network throughout the heartbeat of the ECG, we get a ready-made word2vec model with a linguistic chain representation of the connection with the heartbeat in the form of a word and the corresponding vector representation. An example of a linguistic chain is observed in figure 4. With the help of the created model we can find similarities between heartbeats, which we want to, analyze and with heartbeats, which are marked by some labels (arrhythmia, disease, no anomalies).

```
kwb : [-4.52855631e-04  2.69200653e-03 -2.85770698e-03 -3.71513655e-03
  7.24595389e-04 -6.08999457e-04  2.23849644e-03 -3.87491775e-03
 -1.76306057e-03  6.52205083e-04  7.47220067e-04 -2.37480062e-03
  2.80323718e-03  1.15808658e-03 -1.81753351e-03 -2.80798832e-03
 -5.11458609e-03  3.07700131e-03  2.54009705e-04  4.26528323e-03
 -1.63862924e-03  4.70669661e-03  3.97165120e-03  3.91787663e-03
 -3.11121764e-03  3.04562040e-03 -2.35919980e-03  1.19900316e-04
 -5.36466995e-03  2.48389272e-03 -1.44891499e-03 -7.83059513e-04
  3.53254564e-03 -2.12145061e-03 -1.40378485e-03  3.31551279e-03
  4.04210994e-03  1.01519178e-03  1.57758989e-03  4.68220329e-03
  1.86871411e-03 -6.41508261e-04  2.16125301e-03  1.22731004e-03
 -1.09590031e-03 -6.07335009e-03  1.30223471e-03 -3.83058796e-03
  2.22818181e-03 -4.63103503e-03 -1.52695086e-03  3.62753542e-03
 -2.10220553e-03  4.25343588e-03  2.28840276e-03  3.60550778e-03
 -5.99312079e-05 -3.19302292e-03 -4.41611465e-03  5.54119237e-04
 -5.55756828e-03  4.10656631e-03  2.68298457e-03  4.14208882e-03
 -8.60912609e-04  5.34155697e-04  3.25926510e-03  2.76880083e-03
  1.42918539e-03 -3.18572158e-03 -2.72700260e-03 -2.76432396e-03
  2.55474891e-03  2.83807237e-03 -8.61959648e-04  4.05401969e-03
  3.96387372e-03 -2.75863800e-03 -4.49991878e-03 -1.00883329e-03
  4.77602240e-03  2.75431201e-03 -1.71090907e-03 -1.92280975e-03
  4.18488542e-03  6.76358759e-04 -7.16106326e-04  1.82076625e-03
  3.56791681e-03 -1.83787569e-03 -8.79865547e-05  1.55338924e-03
 -2.37935386e-03 -4.98656929e-03 -4.83540772e-03 -3.00369668e-03
  1.35593917e-04  3.89454677e-03  3.50713078e-03 -5.40652266e-03]
```

Figure 4: An example of a vector representation based on a single heartbeat

2.3 Data analysis based on Word2Vec model

For the vectorization of data and encouragement based on Word2Vec models, it is possible to use the methods of machine translation and NLP for analysis of data. The Word2Vec model allows to find the TF-IDF metrics for the value of the statistical characteristics appropriate to the surroundings (sertsebits), knows the most important suspicions, and the data for the ECG students. In this last update, the Random Forest method has been converted for the classification of arrhythmias, both for the response to the ECG signal and for the vectorized representation in the Word2Vec model.

3 Results

The analysis was performed on a dataset including 23 durable ECG audiotapes of themes mainly with AFIB atrial-fibrillation[14]. Every AFIT MIT-BIH subject contains twain 10-hour ECG audiotapes. ECG recordings are selected at a prevalence of 250 Hz with a 12-bit settlement in the area of ± 10 mill volts.

The change in the data volume of the information ECG signal and the indicator converted into a sentence was studied using the linguistic method. The transformation occurs by replacing each selected wave from the heartbeat with the symbol of the cluster to which the wave belongs. Thereby, each heartbeat is converted in a word, and the entire ECG signal into a sequence of sentences.

Table 1
Testing of volume changes in the of ECG data

Initial signal size, B	Number of clusters to represent the waves	Signal size after processing, B
1000	25	9
2000	25	21
3000	25	30
5000	25	51
10000	25	105
20000	25	207
30000	25	312
40000	25	417
50000	25	522
75000	25	786
100000	25	1047

The classification of heartbeat by the Random Forest method was performed. After creating a Word2Vec model based on the ECG signal, we classify using a vector representation of the heartbeat from the created model. We will classify the heartbeat as a sign of arrhythmia. To do this, we will test and determine the execution time and accuracy of classification for different parameters (Table 2).

To compare the classification results, a dataset with marked heartbeats before converting the ECG signal into a sentence and a dataset with heartbeats after using a vector representation of the heart rate using the linguistic chain based on the Word2Vec model are taken. Datasets were divided into training and test samples. The classification was performed by the Random Forest method with a configuration of 15 trees.

Table 2
Research Results

The number of heartbeats	Signal size, kilobytes	ECG presentation method	Classification execution time, ms	Accuracy, %
200	57,2	Numeric time series	68	91,3
200	20	Word2Vec model	37	85,1
500	143	Numeric time series	152	96,7
500	50	Word2Vec model	68	90,3
1000	286	Numeric time series	336	97,9
1000	100	Word2Vec model	182	92,9

The results of the comparison shows that the accuracy of classification based on the heartbeats presented in the usual way is slightly higher (on average 5 percent) than the accuracy of classification appropriate to a vector representation of heartbeat using the linguistic chain with the created model Word2Vec[15]. However, it can be seen that the size of the ECG signal has decreased because the smaller amount of data represents the same number of heartbeats in the signal. It is also seen that the classification time at the same parameters of the number of heartbeats is less for the converted

heartbeat. Therefore, due to the fact that the speed is inversely proportional to the time of execution, you get a gain in speed (Table 2).

4 Discussion and Conclusion

The new technique of constructing the Word2Vec model based on the ECG signal has been developed, which unlike the existing ones, allows the use of NLP methods for the analysis of electrocardiograms and significantly reduces the amount of data without significant loss of analysis accuracy. ECG signal transform into words that in 100 times fewer that initial data. Word2Vec model-based classification accuracy is slightly reduced (~ 5%) than Random Forest method classification but it is quicker and processing small data that can be used in small devices like fitness tracker or smart watches. For increasing classification accuracy need to tune number of clusters and try another classification method like SVM, GBT.

This technique makes it achievable to find the difference between the ECG signals by calculating the cosine of similarity and the use based on the analysis methods such as TextRank to find keywords. At the same time, it is necessary to continue the research on the use of fuzzy sets techniques to increase the possibilities of analysis.

References

[1] O. Y. Zharinov, V. O. Kuts (2017) Fundamentals of electrocardiography: [textbook. way. for doctors-listeners (Ph.D.) postgraduate. education / Zharinov OY and others] - Bibliogr .: p. 235—236

[2] Baklan I., Mukha I., Oliinyk Y., Lishchuk K., Nedashkivsky E., Gavrilenko O. (2020) Anomalies Detection Approach in Electrocardiogram Analysis Using Linguistic Modeling. In: Hu Z., Petoukhov S., Dychka I., He M. (eds) Advances in Computer Science for Engineering and Education II. ICCSEEA 2019. Advances in Intelligent Systems and Computing, vol 938. Springer, Cham; pp 513-522, DOI - https://dx.doi.org/10.1007/978-3-030-16621-2_48

[3] J. Pan and W. J. Tompkins, “A real-time qrs detection algorithm,” IEEE Trans. Biomed. Eng, vol. 32, no. 3, pp. 230–236, 1985.

[4] Engelse, W. A. H. and Zeelenberg, C. (1979). A single scan algorithm for QRS-detection and feature extraction. Computers in Cardiology, 6:37–42.

[5] Hamilton, P. (2002). Open source ecg analysis. Computersin Cardiology

[6] Tryon, Robert C. (1939). Cluster Analysis: Correlation Profile and Ortho-metric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers

[7] J. A. Lozano J. M. Pena and P. Larranaga, An empirical comparison of four initialization methods for the k-means algorithm," Pattern Reognition Letters, vol. 20, pp. 1027-1040,1999.

[8] M. Brand and K. Huang. A unifying theorem for spectral embedding and clustering. 9th International Conference on Artificial Intelligence and Statistics, 2002.

[9] Abdi H., Williams L.J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2: 433–459

[10] Van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). Visualizing Data Using t-SNE. Journal of Machine Learning Research 9: 2579–2605

[11] Sajad Mousavi, Fatemeh Afghah, Fatemeh Khadem, and U. Rajendra Acharya ECG Language Processing (ELP): New Technique to Analyze ECG Signals, 2020, pp. 2-4

[12] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // Proc. of Workshop at ICLR. 2013. P. 1301-3781.

[13] Igor Baklan, Alina Oliinyk, Iryna Mukha, Kateryna Lishchuk, Olena Gavrilenko, Svitlana Reutska, Anna Tsytsyliuk, Yurii Oliinyk: ECG Signal Processing Based on Linguistic Chain Fuzzy Sets. COLINS 2021: 1731-1741

[14] Munger TM, Wu LQ, Shen WK (2014). "Atrial fibrillation". Journal of Biomedical Research. pp. 1–17

[15] Andrii Tereschenko: Software for creating Word2vec model based on ECG signal, master degree thesis, Kyiv, 2020. – 86 p. URI https://ela.kpi.ua/handle/123456789/39928
[16] P. Rajpurkar, A.Y. Hannun, M. Haghpanahi, C. Bourn, A.Y. Ng Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836 (2017)
[17] Ö. Yıldırım, P. Pławiak, R.-S. Tan, U.R. Acharya Arrhythmia detection using deep convolutional neural network with long duration ecg signals. Comput. Biol. Med., 102 (2018), pp. 411-420
[18] F. Murat, O. Yildirim, M. Talo, U.B. Baloglu, Y. Demir, U.R. Acharya Application of deep learning techniques for heartbeats detection using ecg signals-analysis and review. Comput. Biol. Med. (2020), p. 103726
[19] Sun R., Alexandre F. Connectionist-Symbolic Integration: From Unified to Hybrid Approaches. Psychology Press; London, UK: 2013
[20] Hall L.O., Romaniuk S.G. A Hybrid Connectionist, Symbolic Learning System; Proceedings of the AAAI; Boston, MA, USA. 29 July–3 August 1990; pp. 783–788.