Detecting Heart Disease Symptoms Using Machine Learning Methods

Diera Pirova
Samara State Technical University
Samara, Russia
di.pirova@yandex.ru

Borislav Zaberzhinsky
Samara State Technical University
Samara, Russia
di.pirova@yandex.ru

Andrey Mashkov
Samara State Technical University
Samara, Russia
mavstu@list.ru

***Abstract*—This paper explores the possibility of application of machine learning methods for detecting the signs of cardiovascular diseases. The ECG Heartbeat Categorization Dataset that contains electrocardiogram data of various heart rhythms is taken for the study. The paper describes classification of five various arrhythmia types by the use of machine learning methods. The classification has used such methods as the random forest algorithm, the decision tree and the convolutional neural network. The results of the study show the highest accuracy of the neural network that equals to 0.9347.**

Keywords—machine learning, heart diseases, electrocardiography, random forest, neural network

I. INTRODUCTION

Cardiovascular diseases (CVD) are one of the most significant mortality causes in the modern world. The forecasting of cardiovascular diseases is a major concern in the area of analyzing of clinical data [1]. In modern society, heart disease prevention problems have medical and social significance and remain paramount due to the prevalence rate, large percentage of disablement and extremely heavily mortality that is specific to the people of working age. Millions of people experience irregular heartbeats that in some cases may cause death. Therefore, an accurate and inexpensive diagnosis of arrhythmic heartbeat is critical.

Neural network technologies are aimed at resolving various complicated tasks that include a number of issues in medicine.First of all, the use of neural networks relates to the fact that a researcher receives a large amount of different factual materials that do not already have any mathematical models. Application of the machine learning can improve the accuracy of medical visualization and diagnostics of pathologies. With the use of various algorithms, machine learning methods are suited to address the tasks with unknown mechanisms of situation-based developments and dependencies between inputs and outputs [2].

The objective of the paper is to analyze the application of machine learning methods for detecting the electrocardiograms with indicators of some abnormal heart rhythms. The scientific novelty of the research lies with the use of the convolutional neural network for the analysis of constantly incoming data and the testing of machine learning methods to compare the accuracy of algorithms. The experiment shows that deep neural networks are better suited for solving the problem of classification of heart rhythms.

In order to deal with the issues arising from the manual analysis of electrocardiogram signals, many scientists considered the application of machine learning methods for accurate detection of abnormalities in the signal. For example, the study [3] examined the heart disease prediction model using hybrid random forest linear model (HRFLM). The model represented various combination of signs and several recognized classification methods and included the analysis of multiple signs such as gender, weight and so on, excluding electrocardiogram analysis. The work [4] described the approaches to the design and use of deep neural networks for improving accuracy of diagnosis of heart diseases. The deep neural network architecture was developed and optimized to diagnose heart diseases. Like much of the study described in [3], that work did not consider electrocardiogram records as experimental variables. It is important to emphasize that many studies related to computer-assisted diagnosis of various diseases exist, for example [5], [6] and [7]. Dr. Ana C. Calderon and Dr. Simon Thorne, specialists from the the Department of Computing at Cardif Metropolitan University, examined the benefits of machine learning to medical diagnostics and data analysis. Scientists studied various methods of using neural networks in the diagnosis and surgical planning of diseases and came to the conclusion that with application of such methods doctors are ensured with more reliable decision making, greater productivity and fewer medical errors. The useful capacity of neural networks as the basis for supporting clinical decisions is also evident. They are capable of representing complex relationships found in data that are not immediately obvious for a human examination [8]. The research [9] considers the use of artificial neural networks for identifying and characterizing pathologies in blood vessels. According to the results of the study, the models of neural networks were designed suitable to solve the set tasks with an error of not more than 9%.

Our study consists of two stages. Data pre-processing at the first stage solves the problem of unbalanced classes. We have processed each electrocardiogram signal by Gaussian noise overlaying.

The second stage represents the resolution of the multiclass classification task with the use of various machine learning methods such as random forest, the decision tree and the convolutional neural network.

II. PRE-PROCESSING OF THE SIGNAL

For the analysis of the electrocardiogram, we have selected the MIT-BIH Arrhythmia database with 48 half-hour excerpts of two-channel ambulatory electrocardiogram recordings obtained from 47 subjects as the data source. It is important to note that this study uses pre-processed data from the MIT-BIH Arrhythmia database presented in the Kaggle platform (ECG Heartbeat Categorization Dataset) [10].

The training dataset contains information on 87 thousand heartbeats. Each heartbeat falls within one of five following categories:

1. Normal beat;
2. Supraventricular premature beat;
3. Premature ventricular contraction;
4. Fusion of ventricular and normal beat;

5. Unclassifiable beat.

In all our experiments, we have used electrocardiogram lead II re-sampled with the sampling frequency of 125 Hz as the input signal.

For further interpretation of the data, we show an example of input data for each class.

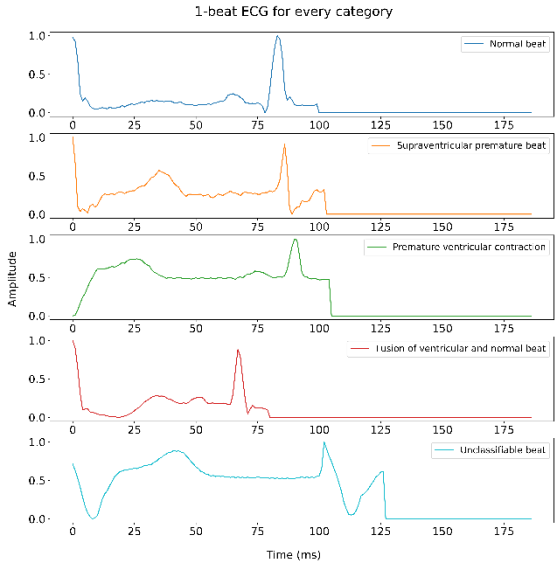

Fig. 1. Example of input data.

The major part of the dataset (72471) falls under the regular rhythm. Figure 2 shows the bar chart of number of examples depending from classes (0 - Normal beat, 1 - Supraventricular premature beat, 2 - Premature ventricular contraction, 3 - Fusion of ventricular and normal beat, 4 - Unclassifiable beat).

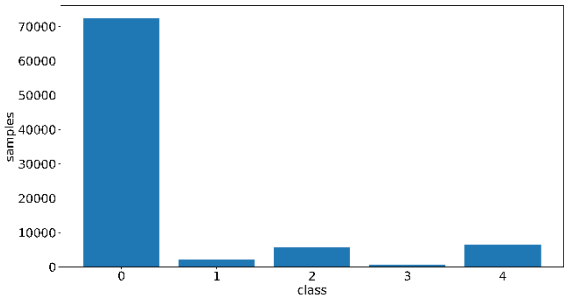

Fig. 2. Bar chart of input training sampling.

Figure 2 demonstrates the big difference between the numbers of input data for each of the classes. We have used the method described in [11] for solving the problem of unbalanced classes. We have balanced the sampling using the resampling method from the Scikit-learn library.

Figure 3 shows the bar chart of balanced sampling.

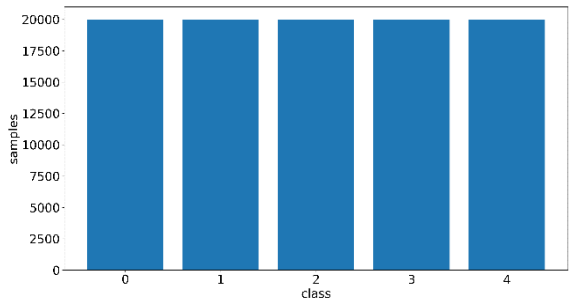

Fig. 3. Bar chart of balanced input training sampling.

At the next step, we have added Gaussian noise with the mathematical expectation of 0 and the dispersion of 0.05 [11] for transforming data of each object of sampling.

Figure 4 and Figure 5 show data before and after transformation of data respectively. Such transformation is necessary to summarize the data and remove duplicates.

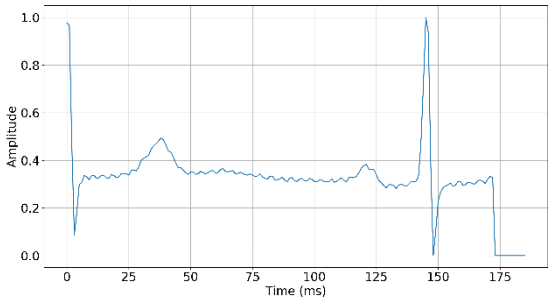

Fig. 4. Before transformation of data.

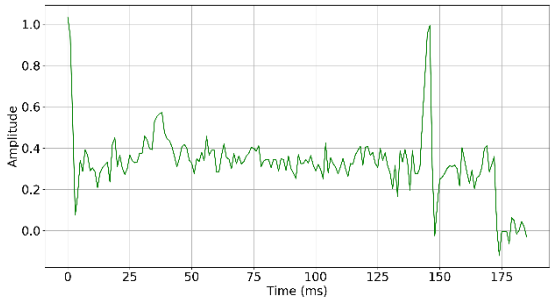

Fig. 5. After transformation of data.

As the result, after data pre-processing, we have received the sampling consisting of 100 thousand heartbeats uniformly ranged into classes.

III. Development and Training of Machine Learning Methods

We have built the models based on such methods as the random forest algorithm, the decision tree and the convolutional neural network for detecting the signs of cardiovascular diseases. The paper describes in detail the operation principle of each of the methods:

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [12].

The random forest consists of many decision trees. The random forest is used in statistics, data analysis and machine learning. Figure 6 demonstrates the model of this algorithm. Each individual tree is a rather simple model with branches, nodes and leaves. Attributes are recorded in the nodes and objective function depends on the attribute values. Further, values of the objective function get into the leaves through the branches. As part of the classification process for a new case we have to go down the tree through the branches all the way until a leaf, moving through all attribute values based on the logical principle “if-then” In accordance with these conditions, a target variable receives one or another value or class (the target variable gets into the specific list). The objective of the decision tree construction is to develop a model that predicts the value of the target variable depending on several variables at the input [2].

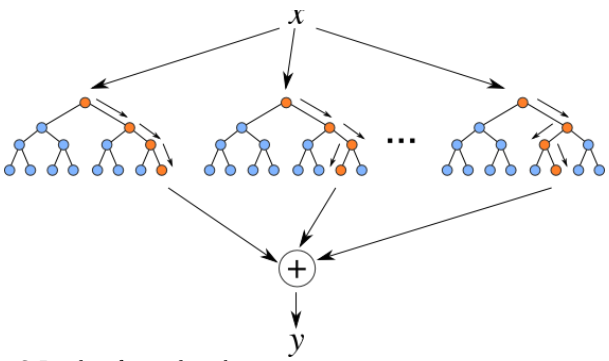

Fig. 6. Random forest algorithm.

We apply decision tree and random forest algorithms with the use of the Scikit-learn library for the Python programming language. The paper considers the n_estimators parameter (the number of trees in the forest) for random forest. We have selected this parameter in experimental studies and it equals to 50.

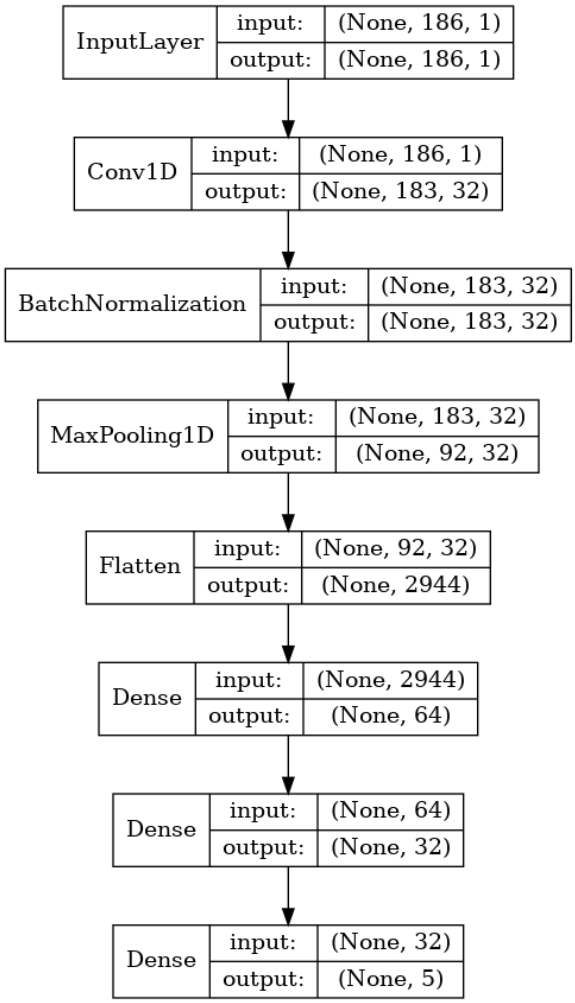

Fig. 7. Model of artificial neural network.

In recent years, convolutional neural networks are gaining popularity. In this paper, the model of convolutional neural network contain 1 convolution consisting of 32 feature maps. Then the batch-normalization layer comes that normalizes previously obtained data. The normalized data have mathematical expectation of 0 and dispersion of 1. Sub Nyquist sampling layer (MaxPooling1D) compresses the obtained data twice and the Flatten layer shrinks the compressed data into 1D dimension array. Three cascade fully connected layers containing 64, 32 and 5 neurons respectively follow these layers. At the output, we have five neurons that show the probability of belonging to each class. Figure 7 shows the described neural network architecture.

The neural network is introduced through the Keras library in Python programming language and we have trained it with the use of Adam (Adaptive Moment Estimation) optimizer.

Adam is an update to the RMSProp optimizer. This optimization algorithm uses running averages of both the gradients and the second moments of the gradients. Given parameters of the weight - $w^{(t)}$, where t indexes current training iteration are updated according to the following algorithm [12]:

$$m_w^{(t+1)} \leftarrow \beta_1 m_w^{(t)} + (1-\beta_1)\nabla_w L^{(t)};$$

$$v_w^{(t+1)} \leftarrow \beta_2 v_w^{(t)} + (1-\beta_2)(\nabla_w L^{(t)})^2;$$

$$\hat{m}_w = \frac{m_w^{(t+1)}}{1-\beta_1^{t+1}};$$

$$\hat{v}_w = \frac{v_w^{(t+1)}}{1-\beta_2^{t+1}};$$

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \frac{\hat{m}_w}{\varepsilon + \sqrt{\hat{v}_w}},$$

where ϵ is a small scalar used to prevent division by 0, and β_1 and β_2 are the forgetting factors for gradients and second moments of gradients, respectively. Squaring and square-rooting is done elementwise [13].

We select ReLU function as the activation function on each layer except the last one. For the last layer we use Softmax function that is often applied in neural networks to show unnormalized network outputs into a probability distribution over the predicted output classes. Softmax function is as follows:

$$Soft\max(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{5} e^{z_j}}, \quad (1)$$

where z_i is an unnormalized input vector.

We use logloss as an error function:

$$\log loss = -\frac{1}{q}\sum_{i=1}^{q}\sum_{j=1}^{l} y_{ij} \log a_{ij}, \quad (2)$$

where q is a number of elements in sampling, l is a number of classes, a_{ij} is a response (probability) of the algorithm on the object i to the question on its belonging to class j, $y_{ij} = 1$ if object i belongs to class j, if not $y_{ij} = 0$.

Each of the above-mentioned methods has its own advantages and disadvantages. In this paper, we present the training of all models mentioned above to identify which model solves the classification problem more accurately.

IV. Experimental studies

For experimental studies, we have launched the program based on test sampling consisting of 21,892 heartbeat records.

We have started up the training process of the neural network in 15 epochs. Figure 8 shows a graph of the accuracy of neural network training depending upon an epoch for each training and validation set.

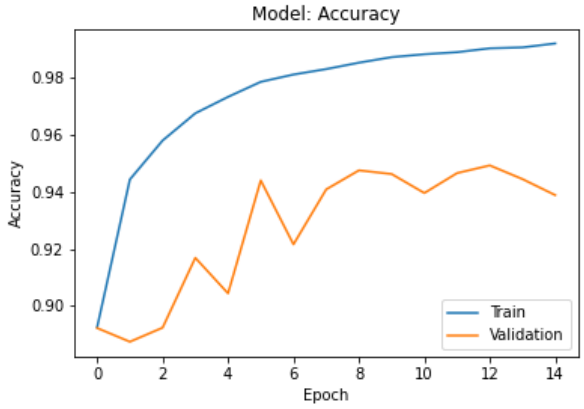

Fig. 8. Neural network training.

For each model, we consider Accuracy characteristic that is formed through the ratio of the total number of correctly predicted objects (P) to the total number of objects (N):

$$Accuracy = \frac{P}{N}.$$

Table 1 shows the training results (Accuracy for validation data).

TABLE I. Comparing the accuracy of algorithms

Method	Accuracy
Decision tree	0.7870
Random forest	0.9090
Neural network	0.9347

Table 1 demonstrates that the convolutional neural network has showed the best result.

As the neural network results further show, we have calculated the normalized confusion matrix that presents the proportion of correct and incorrect predictions for each class. Table 2 demonstrates the values of the confusion matrix.

TABLE II. Confusion matrix for the neural network model

N	0.94	0.03	0.02	0.01	0
S	0.16	0.80	0.03	0.01	0
V	0.02	0.01	0.95	0.02	0
F	0.06	0.01	0.08	0.86	0
Q	0.01	0	0.01	0	0.98
	N	S	V	F	Q

Table 2 indicates that the neural network classifies Supraventricular premature beat much worse. This is simply ause of the small data amount of this class. The neural network identifies Unclassifiable beat more accurately. This appears to be due to the specificity of the electrocardiogram data in this class.

V. Conclusion

In this paper, we compare various machine learning algorithms for detecting the signs of cardiovascular diseases. We have pre-processed all signals in accordance with Section 2 before applying the algorithms. The study shows that the convolutional neural network classifies heart rhythms better (0.9347) than other algorithms. The random forest algorithm also demonstrates excellent results (0.9090) that are a bit worse than the results of the neural networks. The decision tree presents the worst results (0.7870). The significance of the study lies in the feasibility of application of the achieved results in predicting the possible development of cardiovascular diseases. The findings obtained in the present study can be used to create medical or other expert systems based on artificial neural networks under conditions of low volume of statistical material.

References

[1] L.Yang, H. Wu, X. Jin, P. Zheng, S. Hu, W. Yu and J. Yan, “Study of cardiovascular disease prediction model based on random forest in eastern China,” Sci Rep, vol. 10, pp. 5240-5245, 2020. DOI: 10.1038/s41598-020-62133-5.

[2] O. Obulesu, M. Mahendra and M. ThrilokReddy, "Machine Learning Techniques and Tools: A Survey," International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, pp. 605-611, 2018. DOI: 10.1109/ICIRCA.2018.8597302.

[3] S. Mohan, C. Thirumalai and G. Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” IEEE Access, vol. 7, pp. 81542-81554, 2019.

[4] N. Tomov and S. Tomov, “On Deep Neural Networks for Detecting Heart Disease,” CoRR abs/1808.07168, pp. 1-10, 2018.

[5] N.Y. Ilyasova, A.V. Kupriyanov and R.A. Paringer, “Formation of features for improving the quality of medical diagnosis based on discriminant analysis methods,” Computer Optics, vol. 38, no. 4, pp. 851-855, 2014.

[6] A.V. Gaidel and S.S. Pervushkin, “Research of the textural features for the bony tissue diseases diagnostics using the roentgenograms,” Computer Optics, vol. 37, no. 1, pp. 113-119, 2013.

[7] A.D. Bragin and V.G. Spitsyn, “Motor imagery recognition in electroencephalograms using convolutional neural networks,” Computer Optics, vol. 44, no. 3, pp. 482-487, 2020. DOI: 10.18287/2412-6179-CO-669.

[8] A.Calderon and S. Thorne, “Neural Networks for Medical Applications,” ITNOW, vol. 60, no. 1, pp. 46-49, 2018. DOI: 10.1093/itnow/bwy022.

[9] S. Moccia, E.D. Momi, L.S. Mattos and S. Hadji, “Blood vessel segmentation algorithms — Review of methods, datasets and evaluation metrics,” Computer Methods and Programs in Biomedicine, pp. 158, 2018. DOI: 10.1016/j.cmpb.2018.02.001.

[10] S. Fazeli, “ECG Heartbeat Categorization Dataset,” 2020 [Online]. URL: https://www.kaggle.com/shayanfazeli/heartbeat.

[11] D. Gregoire, “Arrhythmia on ECG Classification using CNN”, 2020 [Online]. URL: https://www.kaggle.com/gregoiredc/arrhythmia-on-ecg-classification-using-cnn/notebook.

[12] L. Tian, H. Ning, L. Kong, K. Chen, H. Qi and Z. Han, “Sentense Paraphrase Detection Using Classification Models,” FIRE International Workshop, Kolkata, India, pp. 171-172, 2016. DOI: 10.1007/978-3-319-73606-8.

[13] L. Balles and P. Hennig, “Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients,” 6th International Conference on Learning Representations, Canada, pp. 16-18, 2018.