Classification of cardiac arrhythmia using machine learning techniques

M A Firyulina and I L Kashirina

Department of Applied Mathematics, Informatics and Mechanics, Voronezh State University, 1 Universitetskaya pl., Voronezh, 394006, Russia

E-mail: mashafiryulina@mail.ru

Abstract. Cardiovascular disease is one of the leading causes of death worldwide. Currently, there is an increase in the percentage of people with various heart rhythm disorders. There are permanent (chronic), persistent and paroxysmal form of atrial fibrillation, and the most severe violation is the permanent form. Since the reasons for the development of a certain form of atrial fibrillation are not completely clear, this article presents an analysis of various characteristics that affect the formation of arrhythmias of these species. The most significant signs that can potentially be predictors of different forms of the disease have been identified. Four machine learning methods were used for the analysis: classification trees, logistic regression, random forest, and gradient boosting. The highest cross-validation accuracy was obtained using logistic regression.

1. Introduction

Atrial fibrillation refers to conditions where the heart rate becomes irregular, too fast, or too slow. If the heart rate is too high, that is, above 100 beats per minute, then this situation is called tachycardia, and if the heart rate is too low, that is, below 60 beats per minute, then it is called bradycardia. \[1, 2\].

In modern medicine, it is accepted to classify atrial fibrillation (AF) into three types, depending on the duration of manifestation: paroxysmal (transient) – the attack in most cases lasts no more than a day; persistent – signs of atrial fibrillation persist for more than 7 days; permanent (or chronic) – long-lasting rhythm disturbances (for example, more than 1 year), in which electric cardioversion (the most studied and effective method of restoring heart rhythm) was ineffective. Permanent arrhythmia is characterized by the most severe course and can lead to sudden cardiac arrest. Some patients develop one of the episodic forms of arrhythmia (paroxysmal or persistent), which can lead to the fact that others initially form a permanent form of arrhythmia. The mechanisms of development of this or that form of arrhythmia are still not clear.

Progress in the field of hardware and software technologies has improved the methods of treatment and treatment of various diseases [3]. In particular, machine learning technologies have been actively applied in medicine [4]. The accumulated masses of medical data make it possible to detect previously unknown patterns [5]. The effectiveness of using various machine learning algorithms can improve the accuracy and quality of diagnosing atrial fibrillation, predicting the development of forms of this disease and making medical decisions [6].

The purpose of this study is to identify significant factors in the development of a certain type of atrial fibrillation and predict what course the disease will take in the future. Popular machine learning
methods were used for the analysis: decision trees, random variables, gradient regression. All graphs and calculations are performed in the Python programming language using the Google Colab service.

2. Materials and Methods
Depersonalized data on 112 patients diagnosed with atrial fibrillation were used for the analysis. The data were presented by Voronezh regional clinical hospital № 1. The original file contained information on 39 attributes characterizing the condition of patients. The predicted variable is a form of atrial fibrillation, has three different meanings: constant, paroxysmal, persistent. Paroxysmal and persistent arrhythmias belong to the form of episodic arrhythmias, therefore, in the framework of this study, due to the small sample size, they were combined into one class. To build the model, the predicted variable was encoded as follows: 0 - the constant form of atrial fibrillation, 1 - episodic.

Initial analysis of the initial data showed that the modal age of development of atrial fibrillation is 70 years. The average age of women with atrial fibrillation in the sample is 67 years old, men 62 years old. The percentage of women with arrhythmia in the sample is higher and makes up about 65% of the total number of patients.

Several machine learning models were used to predict the form of atrial fibrillation. For comparison, models of classification trees, linear regression, random forest and gradient boosting were constructed. Since the sample under study has a small amount of data, validation of each model was carried out using cross-validation. This method helps to estimate the accuracy of the model on a small set of data. In conditions of small volume, additional splitting of the sample into training and test subsets can negatively affect the overall representativeness of the model and, consequently, the quality of training.

Classification trees (decision trees) - a clear and easily interpreted model for solving regression and classification problems. The advantages of this method are: natural accounting of feature dependencies - in the case of complex interactions of predictors, other models can give worse results, flexibility - categorical and numerical features are considered equally. Classification trees are easy to interpret - the classification result can be represented as a chain of rules, which has an important role in solving medical problems and makes this method one of the most popular in this area.

Random forest is a machine learning algorithm that uses an ensemble of decision trees [7]. This algorithm uses a bootstrap mechanism that allows you to create several samples of the same size based on the original training data set using random selection with repetitions. In this case, some observations may fall into one of the samples several times, and some do not fall once. Each of the decision trees is trained on one of these samples using some random subset of the input features. Each tree of the ensemble refers the object to be classified to one of the classes, and the class for which the largest number of trees voted is determined. The advantage of this algorithm is the high accuracy of training (compared to the accuracy of individual trees).

Logistic regression is one of the most popular methods of linear binary classification. The prediction result of this method is the probability that the input object belongs to a certain class. This property is important in medical applications where, along with object classification, the associated risk of misclassification must be assessed.

Gradient boosting is a machine learning technique whose main idea is the iterative process of sequentially constructing particular models. Each new model is trained using information about the errors made in the previous step, and the resulting function is a linear combination of the entire ensemble of models, taking into account the minimization of some penalty function [8]. This algorithm is distinguished by high accuracy, in most cases superior to the accuracy of other methods.

3. Results and Discussion
The first models built using the libraries of the Python programming language to solve this problem were the classification tree and the random forest. The depth of trees and the number of trees for a random forest were selected using the GridSearchCV procedure, a method for finding the best model parameters by completely sorting through the grid of possible values. Table 1 presents the quality
metrics of the evaluation of models that were obtained on the basis of true and false answers of the model in the training sample: sensitivity, specificity, the proportion of correct answers, as well as the proportion of correct answers in the final cross-check. As you can see, the accuracy of the algorithms during validation falls significantly. The reason for this is that the training sample is too small (as is often the case in medical tasks), and as a result, the models are too fit to the initial data. At the same time, it can be noted that a random forest demonstrates higher accuracy compared to the decision tree method. However, despite the low accuracy, the decision tree method is often used in medical practice, as it allows you to get intuitive and visual classification rules. An example of one of the rules found using the decision tree method is: “If (GH1 <38.5) and (Age <72.5) then the arrhythmia form is constant.” This rule is true in 100% of cases (for the available data sample) and it is satisfied by about 11% of the total number of patients.

### Table 1. Quality metrics. Classification trees and random forest.

| Metrics                          | Classification trees | Random forest |
|----------------------------------|----------------------|---------------|
| Sensitivity, Recall              | 0.9059               | 0.9035        |
| Specificity, True Negative Rate  | 0.8971               | 0.9139        |
| The proportion of correct answers, Accuracy | 0.9019               | 0.9117        |
| Average percentage of correct answers in cross-validation | 0.589               | 0.688         |

Table 2 shows the significant features identified by the constructed models. The significance of each factor in tree models is calculated as a normalized result of a decrease in the branching criterion caused by this factor. The branching criterion calculates a measure of uncertainty in tree nodes. The Gini index was used as such a criterion.

GH1 was the most significant trait in the decision tree model and the second most important trait in the random forest model. This is an indicator of quality of life, which is revealed by questioning patients. Many somatic diseases affect not only the physical condition of a person, but also the psychology of behavior, emotional reactions. Part of the work [9] is devoted to assessing the impact of psychological problems on the development of cardiovascular diseases. Subjective assessment of quality of life is very important for patients with chronic and life-threatening diseases, including cardiac arrhythmias. The specificity of these diseases creates a real threat for a sudden attack with a possible lethal outcome.

### Table 2. Feature importance. Classification trees and random forest.

| Classification trees | Random forest |
|----------------------|---------------|
| Weight               | Feature       | Weight | Feature |
| 0.2303               | GH1           | 0.1049 | Age     |
| 0.1415               | LA            | 0.0936 | GH1     |
| 0.1394               | Age           | 0.0806 | Cholesterol |
| 0.1175               | Index Kerdo   | 0.0786 | LA      |
| 0.0964               | LV ESV        | 0.0659 | Glucose |
| 0.0827               | LV EDV        | 0.0627 | Index Kerdo |
| 0.0535               | HR            | 0.0566 | EF (ejection fraction) |

The expected significant factor in both models is age. Another important feature is left atrium (LA) size. In the medical literature, an increase in the size of LA in atrial fibrillation was recorded. However, there is no detailed analysis of the effect of LA volume on various forms of arrhythmia.

Index Kerdo is an indicator used to assess the activity of the autonomic nervous system. Decoding of indicators of a sign: 1 - eitonia, 2 - sympathetic influence, 3 - parasympathetic. Such indicators as end-diastolic volume (EDV), end-systolic volume (ESV) of the left ventricle (LV) and heart rate (HR) are recommended for consideration when examining patients with heart disease.
At the next stage of the study, models of classification of atrial fibrillation forms were constructed using gradient boosting methods over decision trees and logistic regression. Table 3 shows the quality metrics obtained for these models.

### Table 3. Quality metrics. Gradient boosting and logistic regression.

| Metrics                                | Gradient boosting | Logistic regression |
|----------------------------------------|------------------|---------------------|
| Sensitivity, Recall                     | 0.8916           | 0.7273              |
| Specificity, True Negative Rate         | 0.9058           | 0.7603              |
| The proportion of correct answers,     | 0.9019           | 0.7549              |
| Accuracy                               |                  |                     |
| AUC ROC                                | 0.952            | 0.8192              |
| Average percentage of correct answers in cross-validation | 0.747 | 0.754 |

It can be noted that the logistic regression model has the worst values of the metrics on the training sample relative to the other developed models. However, according to the results of validation, it showed the best proportion of correct answers, which is practically the same as the value of this metric in the training sample. This is because logistic regression, like other linear models, does not have a significant propensity to retrain.

To create an optimal diagnostic system, it is necessary to find a compromise between the obtained sensitivity and specificity of the models. A common way of visualizing the relationship between these metrics is to use the ROC curve — graphic characteristics of the quality of a binary classifier, sensitivity of rate (1-specificity) while varying the threshold of decision rule model. The optimal position for the ROC curve is the maximum approximation to the upper left corner, where specificity and sensitivity are at their maximum. Figure 1 shows a graph of ROC curves for gradient boosting models (left) and logistic regression models (right). The value of the AUC ROC—the area under the ROC curve is a compromise metric widely used in medical research. As shown in table 3, the areas under these curves are 0.952 and 0.8192, respectively.

![ROC-curve](image.png)

**Figure 1.** Graphs of ROC-curves models gradient boosting (left) and logistic regression (right).

Table 4 contains indicators of the significance of the features of these models. For gradient boosting, significance is calculated using the approach described above. The logistic regression method uses the inverse exclusion approach to assess the significance of features. In this case, a statistical p-value of significance is calculated for each input indicator. Then the attributes with the highest value of p are removed one by one, followed by repeated recalculation of the p-values of the remaining indicators, until all the attributes have a value of p less than 0.05 (for other models, on the contrary, signs with a significance coefficient above 0.95 are significant).
Table 4. Feature importance. Gradient boosting and logistic regression.

| Gradient boosting | Logistic regression |
|-------------------|--------------------|
| Weight            | Feature | P-value (p<0.05) | Odds ratio | Feature |
| 0.1871            | CH1     | 0.017            | 1.062774   | Age     |
| 0.1037            | LA      | 0                | 1.09523    | GH1     |
| 0.0889            | Thyroid lesions | 0.022            | 0.648386   | Urea    |
| 0.0882            | Index Kerdo | 0                | 0.24914    | RA      |

Based on the results of logistic regression model, it is possible to calculate the odds ratios for significant features. Odds ratio is one way to describe how the absence or presence of a particular form of arrhythmia is related to the presence or absence of a given factor. The odds for age ratio suggest that if all other indicators remain constant, the risk of developing an episodic form of AF increases by 6% with an increase in age by one year. In contrast, a high right atrial volume (RA) significantly increases the chances of having a permanent form of AF.

To analyze the influence of significant signs on a particular form of atrial fibrillation, the signs that stood out in the construction of different models were selected: GH1, LA, age, Index Kerdo, electric cardioversion, LV ESV, LV EDV, heart rate, urea, RA volume, duration of arrhythmia, cholesterol. Figure 2 shows a correlation matrix that reflects the linear relationship between features. The form of arrhythmia is most strongly correlated with GH1, a subjective assessment of the quality of life of patients.

![Figure 2. Correlation matrix.](image)

Figures 3-7 show the distribution curves of the dependent predictor values for each form of atrial fibrillation: 0 – constant, 1 – episodic (includes paroxysmal and persistent). The indicators of Index Kerdo, right atrium volume (RA) and whether electrical cardioversion was performed are presented in the form of bar charts in figures 7-8, as they are less selective.

Graph analysis (Fig. 3, left) shows that the peak age value for patients of both forms is the same-70 years, however, we see that the permanent form of AF in the age category of 50-65 years is present more often than episodic. The risk of developing episodic arrhythmia increases significantly after 65 years.

The volume distribution of the left atrium (Fig. 3, right) shows that at constant AF, the increase in LA volume over 5 cm occurs much more often. The norm of LA size is up to 3.3 cm. In episodic arrhythmia, the subjective indicator of quality of life GH1 (Fig. 4, left) higher than at constant, meaning patients feel better.
The heart rate is higher on average at a constant form of AF. Heart rate is one of the most indicative factors indicating the presence of diseases associated with heart rate (Fig. 4, right).

Indicators of LV ESV (norm 3.1 – 4.3 cm) are underestimated more often in the episodic form of arrhythmia and overestimated more often in the constant (Fig. 5, left).

LV EDV (norm 4.6 – 5.7 cm) is also more often overestimated with a constant form of atrial fibrillation (Fig. 5, right). The duration of the manifestation of atrial fibrillation is longer with a constant form. (Fig. 6, left).
Figure 6. The distribution of duration of AF (left) and measure of urea (right).
Distribution of indicators of urea and cholesterol have expressed peak values at the constant form of arrhythmia, for urea-6.2 for cholesterol-5.4. For the episodic form, it is impossible to allocate such values, since the indicators of patients are distributed more evenly.

Figure 7. Distribution of cholesterol index (left) and EC of patients (right).
If conservative therapy is ineffective, treatment of atrial fibrillation is carried out by applying an electric pulse discharge to the heart (electric cardioversion). The main distinguishing feature of the permanent form of arrhythmia is the inefficiency of electrical cardioversion [10]. Therefore, among those to whom it was carried out, patients with episodic form of AF more.

Figure 8. Distribution of Index Kerdo (left) and RA volume (right) of patients depending on the form of arrhythmia.

The autonomic Kerdo index is one of the simplest indicators of the functional state of the autonomic nervous system. The Kerdo index is calculated based on the values of pulse and diastolic pressure. According to the diagram in Fig. 8 it is seen that most patients with a score of 2 - sympathetic influence.
The distribution of patients with different forms of atrial fibrillation is almost the same with hypertension and sympathetic influence. There are more patients with the sporadic form of arrhythmia in parasympathetic influence. The increase in the volume of the right atrium, as previously noted, is more typical for patients with a permanent form of arrhythmia. With episodic arrhythmia, the size of the right atrium in most patients is normal.

4. Conclusion
Atrial fibrillation refers to conditions where the heart rate becomes irregular, too fast, or too slow. If the heart rate is too high, that is, above 100 beats per minute, then this situation is called tachycardia, and if the heart rate is too low, that is, below 60 beats per minute, then it is called bradycardia. [2, 3].

The classification tree model predicted the form of arrhythmia worse than other methods, but this algorithm revealed several clear and accurate classification rules.

Using the constructed models, significant factors for the classification of the form of atrial fibrillation were found. The most significant (that is, selected as significant by two or more models): GH1, LA volume, age, Index Kerdo.

It should be noted that the models presented in the article are built so far on a relatively small sample. Now, the study is ongoing, and as observations accumulate, the models will be refined.

References
[1] Thompson-Paul AM and Palella FJ Jr 2018 Heart Disease and Stroke. Chronic Disease Prevention and Health Promotion (NCCDPHP)
[2] Andersson C and Vasan RS 2018 Epidemiology of cardiovascular disease in young individuals (Nat Rev Cardiol) 15 pp 230-240
[3] Firyulina M A 2018 Analysis of mortality indicators of the Voronezh region in comparison with developed countries (VESTNIK of VIHT) 2 pp 150-153
[4] Chariatis A 2007 Very fast online learning of highly nonlinear problems (The Journal of Machine Learning Research) vol 8 pp 2017-45
[5] Canlas and Ruben D 2009 Data Mining in Healthcare : Current Applications and Issues By (Master’s Thesis, Carnegie Mellon University, Pittsburgh)
[6] Mohan K R., Paramasivam I and Narayan S S 2014 Prediction and diagnosis of cardiovascular disease–a critical survey (WCCCT) pp 246–251
[7] Leo Breiman 2001 Random Forests (Machine Learning) pp 5-32.
[8] Shitikov V K and Mastitsky SE 2017 Classification, regression, Data Mining algorithms using R.
[9] Serdechnaya E V 2007 Assessment of quality of life using the SF-36 questionnaire in patients with cardiac arrhythmias (Human Ecology) pp 36-39
[10] Kashirina I L and Firyulina M A 2019 Analysis of the significance of predictors of survival after myocardial infarction using the Kaplan-Meyer method (Modeling, optimization and information technologies) pp 7-20