Using of machine learning methods in cardiology

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. In the modern world, artificial intelligence and machine learning affect many areas of life. Medicine is no exception. It is one of the most important areas of application of intelligent technologies. This article describes the experience of building machine learning models in cardiology and discusses the features of their development, analyses various machine learning algorithms used in problems of cardiological practice, examines ways to improve their accuracy and efficiency.

1. Introduction

As is known, among the causes of mortality in the population, the largest percentage are cardiovascular diseases (CVD) (over 30%) [1]. That is why the introduction of improved methods of examining patients using modern information technologies plays an important role in the development of cardiology. This issue is devoted to a lot of research, both in Russia and abroad [2]. The use of artificial intelligence (AI) and machine learning methods in cardiology should provide a set of tools to increase and expand the effectiveness of treatment of patients with CVD.

There are several reasons why this is necessary. Many CVDs, such as ischemic heart disease, arrhythmias, myocardial infarction, require early recognition and appropriate preventive measures. It is not always possible to establish an accurate diagnosis based on standard clinical studies. The use of AI systems will help to establish a diagnosis faster, while reducing the time for additional research. The transfer of biometric data through mobile devices will help patients to independently monitor their health and notify in case of risk factors for deterioration [3].

The introduction of artificial intelligence into cardiovascular medicine will affect all aspects of cardiology, from research and development to clinical practice and public health. Figure 1 shows a diagram of how AI is being introduced into the medical technology [4].

Many works all over the world are devoted to identifying predictors of the risk of diseases of the cardiovascular system [5]. Enough data is required for a more accurate and complete analysis. The informatization of medical institutions in Russia begun in recent years has led to the fact that right now, at the current moment, large amounts of data have accumulated. It is also now possible to use international open databases that allow tuning machine learning algorithms to predict cardiovascular risk. For example, MIMIC-III (Medical Intensive Care Information Center) brings together clinical data from patients admitted to Beth Israel Deaconess Medical Center in Boston and makes it widely available to researchers internationally [6]. According to these data, the three most common diagnoses of patients in the intensive care center are acute myocardial infarction, arterial hypertension and atherosclerosis, which are expected to be given special attention in this study.
2. Description of the solved problems

During the study «Development and research of machine learning methods in the problems of diagnostics and support of patients with diseases of the cardiovascular system», several tasks were considered and implemented.

1. Models and algorithms have been developed for predicting the development of myocardial infarction (MI) at a certain date, depending on the initial clinical characteristics of the patient and considering the monitoring of meteorological data. We analyzed the data of the Voronezh Regional Register of myocardial infarctions, containing depersonalized information about all patients admitted with a diagnosis of myocardial infarction to hospitals in the Voronezh Region in 2014–2017. The original sample was supplemented with meteorological data. The dataset contains a description of 14633 cases of myocardial infarction, of which 2451 were fatal (16.8%).

2. A complex of models and algorithms has been developed for predicting the risk of one-year mortality after MI, which differ by taking into account the regularities of the influence of socio-demographic and clinical factors, meteorological data, as well as the quality of medical care. To obtain the most reliable results, the initial sample from point 1 was supplemented with information on registered deaths after patients were discharged, based on data provided by Voronezh Region Hospital №1. Deaths within a few days after discharge are important for analysis [7].

3. A model has been developed for predicting the occurrence of certain types of atrial fibrillation (arrhythmia). The study was conducted on patient data provided by the Voronezh Regional Clinical
Hospital №1. The original file contained information on 39 attributes characterizing the patient's condition. The predicted variable, the form of arrhythmia, has three different meanings: constant, paroxysmal, persistent. Moreover, the constant is the most difficult to treat and significantly worsens the quality of life of patients, therefore it is important to distinguish the class of patients who are at risk for the development of this form of arrhythmia. The study analyzed the accuracy of various machine learning models. Based on machine learning methods, personalized risk factors for patients with arrhythmia were identified and a model was developed that considers risk factors when choosing a therapy for chronic diseases.

4. A compliance assessment model has been developed, aimed at identifying the adherence of patients with cardiovascular diseases to treatment with certain groups of medicines. Based on the collected observations, including the age, gender of the patient, reasons for hospitalization, the side effects of medicines observed by the patient, the presence of visual recommendations issued by the doctor, a feeling of improvement in well-being against the background of taking medicines, the experience of arterial hypertension, the presence of chronic concomitant diseases, etc. The tendencies and patterns of the influence of various factors on the compliance of patients with cardiovascular diseases were revealed. The analysis was carried out for seven groups of medicines, which were recommended for further self-administration to patients after discharge from the cardiological hospital.

5. A model of prescribing treatment for hypertension in a cardiological hospital has been developed. The original file contained data for 262 patients for 83 clinical and socio-demographic indicators. Based on machine learning methods, we have selected features that have a significant impact on the prescription of a certain therapy and built models for prescribing therapy to patients in a cardiological hospital in the treatment of hypertension. The models are built for 6 main groups of medicines that are prescribed to patients during inpatient treatment.

Figure 2. Implemented tasks and methods of their implementation.
Figure 2 shows schematically the tasks implemented and some of the methods used to solve them. For the preparation of the initial data, aggregation and primary statistical analysis, the Oracle 12c DBMS was used, for the construction of ML models and neural networks, the built-in libraries of the Python 3.6 programming language were used. Graphs and charts for graphical analysis were also built using Python 3.6 powered by Google Colab. Additional data for research, such as meteorological indicators, were downloaded from the data archives of the rp5.ru website.

3. Review of the methods used to solve problems

3.1. Methods for statistical data analysis.

Statistical analysis of data is used in the initial stages of research. Different methods of statistical data analysis were applied for each problem considered. As an example, consider the results of applying statistical methods for primary data analysis in the first and second tasks.

For the first two tasks, we searched for factors that influence the dependent variable. Three external influencing factors were selected for the study: month, day of the week, season. According to the results of the study, it can be concluded that the month and season affect the average number of deaths in this month (or season), but it was not possible to identify a general relationship for four years. The average numbers of myocardial infarctions on different days of the week have significant differences (p<0.05), the maximum number of MI in the Voronezh region is recorded on Monday, and the minimum on weekends, the difference in the number of MI between Monday and weekends is on average about 25%. Correlations were found between the number of MI per day and the risk of mortality after MI from the jumps in air temperature and atmospheric pressure. Figure 3 shows the distribution of the average number of myocardial infarctions by monthly factors: season and month.

![Figure 3. Distribution of the average number of MIs by months and seasons.](image)

Chi-square test showed that the random variable characterizing the average number of myocardial infarctions per day obeys the Poisson distribution law (p>0.05), which is consistent with the results of similar studies in other regions.

One of the statistical methods for assessing the influence of factors on patient survival is the Kaplan-Meier method. The Kaplan-Meier method estimates the survival function based on the survival time for complete and censored data. The observation period was chosen as 20 days, since most patients had hospitalization periods less than or equal to 20 days. The Gehan-Wilcoxon test was used to assess the difference in the groups under consideration. The analysis showed that for the subsequent construction of the prognostic system, the following signs can be used: gender, age group, whether myocardial infarction is recurrent, localization, whether the patient has a history of diabetes mellitus, atrial fibrillation, CHF, and whether he underwent percutaneous coronary interventions (PCI). The significance of these signs is shown in table 1. The only predictor turned out to be insignificant - the presence of arterial hypertension in the patient's history.
### Table 1. Significance level of predictors.

| Predictors              | p-value | Significance |
|-------------------------|---------|--------------|
| Gender                  | 0       | Yes          |
| Age group               | 0       | Yes          |
| Arterial hypertension   | 0.64007 | No           |
| IM (repeated)           | 0       | Yes          |
| Diabetes                | 0.0001  | Yes          |
| Atrial fibrillation     | 0       | Yes          |
| ONMK                    | 0       | Yes          |
| COPD                    | 0       | Yes          |
| CHF                     | 0       | Yes          |
| Localization            | 0.01032 | Yes          |
| KILLIP                  | 0       | Yes          |
| TLT                     | 0.001   | Yes          |
| PCI                     | 0       | Yes          |

In general, patient survival is significantly lower in the first five days from the onset of myocardial infarction, as shown in figure 4. This period is the most critical.

![Survival function graph](image)

**Figure 4.** Survival function graph.

### 3.2. Machine learning methods.

The main perspective of machine learning in medicine is to incorporate data from various sources (clinical measurements and observations, biological data, experimental results, environmental information, wearable devices) in models to describe and predict human diseases. The typical machine learning workflow described in figure 5, starts with collecting data, proceeding the initial data, then choosing an algorithm and developing a model, and finally leads to model evaluation and application.

To solve the problem of predicting the type of arrhythmia, 39 attributes were used, including the following categories of signs: hemodynamic (heart rate), socio-demographic (gender, age, smoker, GH – a subjective indicator of quality of life), laboratory (cholesterol, glucose, creatinine, urea), functional (ejection fraction, end-diastolic size of the left and right ventricles, thickness of the posterior wall of the left ventricle), clinical (duration of arrhythmia, arterial hypertension, the presence of concomitant diseases).
Several machine learning models were used to solve problems 1–5. Most often, models of classification trees, logistic regression, random forest and gradient boosting were built.

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

A random forest is a machine learning algorithm that uses an ensemble of decision trees [8]. This algorithm uses a bootstrap method, which makes it possible to generate several samples of the same size based on the initial training dataset using random selection with repeats. Some observations may be included in one of the samples several times, and some may not be included even once. Each of the decision trees is trained on one of these samples, using some random subset of the input features. Each tree in the ensemble assigns 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 training accuracy (compared to the accuracy of individual trees).

Logistic regression is one of the most popular linear classification methods. The result of the prediction of this method is the probability that the input object belongs to a class. This property is important in medical applications, where, along with object classification, it is required to assess the associated risk of misclassification. In addition, this method is less to overfitting than other methods.

Gradient boosting is a machine learning technique whose main idea is the iterative process of sequentially building partial models. Each new model is trained using information about the errors made at the previous stage, and the resulting function is a linear combination of the entire ensemble of models,
considering the minimization of some penalty function [9]. This algorithm is distinguished by its high accuracy, which in most cases exceeds the accuracy of other methods.

Table 2 shows, as an example, the results of the comparative effectiveness of the described methods in the problem of classifying the form of arrhythmia [10] (calculated for the class «constant arrhythmia»). The accuracy of the method was assessed by the method of five-fold cross-validation, as the size of the provided training sample is not large enough (only 112 patients). Sensitivity (proportion of true positives that are correctly identified), specificity (the proportion of true negatives) and AUC_ROC, a metric that aggregates sensitivity and specificity, are usually used as basic metrics in medical research in classification problems [11].

The best cross-validation metric AUC_ROC in this problem was shown by the logistic regression method (0.754). Perhaps this result is just related to the small sample size, and logistic regression is a method less to overfitting than the others. The decision tree model predicted the shape of arrhythmia worse than other methods, however, this algorithm revealed several clear and accurate rules for predicting the development of persistent arrhythmia.

At the stage of developing each model, an assessment of the significance of features was carried out to reduce the number of input predictors. This step is necessary to simplify the model and improve its quality.

Table 2. Quality metrics of various machine learning models.

|               | Gradient boosting | Logistic regression | Decision trees | Random forest |
|---------------|-------------------|---------------------|----------------|---------------|
| Sensitivity   | 0.759             | 0.727               | 0.590          | 0.595         |
| Specificity   | 0.840             | 0.760               | 0.589          | 0.691         |
| AUC_ROC       | 0.747             | 0.754               | 0.589          | 0.688         |

4. Problems in building machine learning models

4.1. Unbalanced data.

During working with a real data sample, a situation often arises when the values of one class in it are much more (or less) than the values of other classes. Such data is called unbalanced. Building a model in such a situation may turn out to be ineffective and the error on the test data will be large. For medical tasks, the data balancing stage is important. The advantage of the number of cases of one class leads to a shift in the model towards the majority class. For example, task 2 (predicting mortality after myocardial infarction) belongs to the type of «early warning» tasks of an event. It is more important to predict the onset of death than to suppose that the patient will survive and get the opposite result.

Figure 6. Data processing methods for undersampling and oversampling.
In the initial sample, the number of patients who died was much less than the number of survivors, and it was necessary to balance the data to correctly build a predictive model and reduce false negative results. Rebalancing can be done in two ways: undersampling and oversampling. Oversampling is a duplication of minority class examples. Undersampling is the removal of majority class examples. Figure 6 shows the data processing strategies of various sampling methods.

Undersampling is the simplest and at the same time the most correct in the tasks of medical research. Duplication of examples can potentially lead to loss of information and inaccurate results. Therefore, it was this method that was chosen to solve the problem. The undersampling technique can be implemented by several methods [1-2]. To achieve the highest forecasting accuracy, five algorithms were considered, and the accuracy of their work was compared. Table 3 shows the results of the study.

| Methods                          | Sensitivity | Specificity | AUC_ROC |
|---------------------------------|-------------|-------------|---------|
| Random undersampling            | 0.5         | 0.91        | 0.81    |
| Cluster centroids undersampling | 0.68        | 0.75        | 0.79    |
| TomekLinks undersampling        | 0.4         | 0.98        | 0.82    |
| NearMiss undersampling          | 0.15        | 0.99        | 0.75    |
| Scale pos-weight                | 0.48        | 0.88        | 0.76    |

Based on the results of the model quality metrics, we can conclude that the best indicators of accuracy AUC_ROC are achieved when using Random undersampling for the majority class and TomekLinks method. At the same time, however, the Cluster Centroids method provides the greatest sensitivity (that is, it found 68% of patients in the test sample who subsequently had a fatal outcome). But he also showed poorer specificity (only 75% of the patients he predicted death from myocardial infarction died).

4.2. Interpretability requirement.

Another problem is that sometimes accuracy is not enough to effectively use machine learning models. Medicine is one of the areas where the interpretability of the results is important. It is not enough for doctors to use the system as a «black box»; it is important to understand why the model predicted a certain result.

![Figure 7. Compliance of patients with calcium antagonists.](image)
During building a model for studying patient compliance (task 4), doctors set a prerequisite for the interpretability of the results obtained. Therefore, to solve this problem, a method for constructing decision trees was chosen. Compliance refers to the patient's adherence to treatment and compliance with medical prescriptions. In the process of interpreting the data of compliance of patients with CVD, the most significant signs were revealed for each of the seven studied groups of medicines. Figure 7 shows an example of the interpretation of the obtained model for calcium antagonist medicines (class 0 – patients who will comply with the recommendations for treatment with these drugs, class 1 – no).

Figure 7 shows that patients at the age of no more than 76 years who received these medicines on an outpatient basis tend to take them independently for a long time. The visual representation of results in the form of decision trees helps clinicians understand the cause of the problem and improve the effectiveness of treatment, using expert opinion [13].

5. Using the results in practical medicine

The final result of the study described in this work is the development of an automated application for a cardiologist that helps in the examination and treatment of patients with profile pathology (diseases of the cardiovascular system), reducing the time for performing additional operations related to input, storage, search and analysis of medical data, with elements of support for making medical decisions based on artificial intelligence and the function of remote monitoring of the condition of patients.

At the moment, an information system has been partially implemented that allows monitoring data on the patient's clinical condition (blood pressure, heart rate, body weight, medication intake, meteorological factors, etc.) and, based on artificial intelligence algorithms, to develop individual data for each person connected to the patient's system, recommendations for correcting therapy and warning an exacerbation of cardiovascular disease, including by notifying the doctor at the right time. The diagram of interaction between users and the main components of the considered automated system is shown in Figure 8.

The web application is developed using the .NET Framework 4.6.1 and Angular 6.0.2. To store and process medical data, the MS SQL Server 2016 DBMS was used. Currently, the functionality of the application is to create a medical card with the minimum necessary information about the patient; the ability to track patient visits to doctors for the purpose of consultation, observation, diagnostic or other medical procedures; the ability to maintain a list of medical appointments for the patient's treatment. This system assumes the revision of the existing functionality, as well as further expansion, considering the initial requirements for the full version of the software product [14].

![Diagram of interaction between users and the main components of the system.](image-url)
6. Conclusions
With each new stage of technological progress, cardiology and medicine in general are becoming more independent of humans and are approaching an automated sphere controlled by artificial intelligence. Despite all the difficulties of using complexes based on machine learning, this will greatly facilitate the work of doctors, increase the effectiveness of diagnostics and contribute to the choice of effective treatment measures.

This article provides an overview of the tasks implemented in the study «Development and research of machine learning methods in the problems of diagnostics and support of patients with diseases of the cardiovascular system». Now, the research continues, and as the observations accumulate, the machine learning models will be refined, new tasks for prescribing therapy will be solved and the functionality of the developed automated system will be added.

Acknowledgments
The study was carried out with the financial support of the Russian Foundation for Basic Research within the framework of the scientific project № 20-37-90029 Postgraduates.

References
[1] Firyulina M A 2018 Analysis of mortality indicators of the Voronezh region in comparison with developed countries Bulletin of VIHT 2 pp 150-153 (in Russian)
[2] Glazkova T G 1994 Assessment of the quality of diagnostic and prognosis methods in medicine Bulletin of ONTs AMS of Russia 2 pp 3-11 (in Russian)
[3] Gensini G F, Alderighi C, Rasoini R and Mazzanti M 2017 Value of Telemonitoring and Telemedicine in Heart Failure Management Card Fail Rev 3(2) pp 116-121
[4] Kipp W Johnson and Jessica Torres Soto 2018 Artificial Intelligence in Cardiology Journal of the american college of cardiology 23(17) pp 2668-2679
[5] Enquselassie F 1993 Seasons, temperature and coronary disease Int J Epidemiol 22 pp 632-636
[6] Johnson, A E W.et al. 2016 MIMIC-III, a freely accessible critical care database Sci. Data 3 160035
[7] Kashirina I L, Kazakova A O and Hohlov R A 2016 Forecasting the development of myocardial infarction based on the analysis of meteorological factors and data from the regional register Bulletin of Voronezh State University. Series: Systems analysis and information technologies 3 pp 116-123 (in Russian)
[8] Leo Breiman 2001 Random Forests (Machine Learning) pp 5-32.
[9] Shitikov V K and Mastitsky SE 2017 Classification, regression, Data Mining algorithms using R.
[10] Firyulina M A and Kashirina I L 2020 Classification of cardiac arrhythmia using machine learning techniques J. Phys.: Conf. Ser 1614 012086
[11] Gusev A V, Gavrilov D V and Korsakov I N 2019 Prospects for the use of machine learning methods for predicting cardiovascular disease Artificial intelligence in healthcare 3 pp 41-47 (in Russian)
[12] Garcia S 2009 Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy Evolutionary Computation 17 pp 275-306
[13] Firyulina M A, Kashirina I L and Konobeeva E V 2019 Compliance assessment of patients with cardiovascular diseases based on data mining methods Systems analysis and control in biomedical systems 3 pp 177-183 (in Russian)
[14] Firyulina M A, Demchenko M V and Hohlov R A 2020 Development of an automated workstation for a cardiologist Informatics: problems, methods, technologies pp 1311-1318 (in Russian)