The development of the atherosclerosis diagnostic models under conditions of unbalanced classes

M V Demchenko and I L Kashirina

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

E-mail: kash.irina@mail.ru

Abstract. The main purpose of this study is to identify, using various methods of machine learning, the most effective markers of atherosclerosis and their main predictors, which can accurately determine the risk of this disease in the human body. In this study, various models and balancing techniques of the initial data set were used, which allowed us to develop the efficient classifiers according to the criteria of sensitivity, specificity and ROC AUC.

1. Introduction
Cardiovascular diseases are the most widespread cause of death worldwide and account for more than 30% of the total number of fatal outcomes. One of the leading causes of vascular damage and ischemic lesions of organs is an atherosclerosis disease (that is, damage of the arteries), which requires the research and development of high-quality and optimal methods of diagnostics and treatment. Modern machine learning methods provide a number of effective tools for solving medical diagnostic problems. The efficient algorithms are often used to classify patients based on such a criterion as the presence of a disease, based on a set of input features (predictors).

The development of machine learning models in medical diagnostics problems is often accompanied by the issue of uneven classes distribution in the training sample, due to the fact that the number of patients suffering from a particular disease in the source data is significantly lower (or, conversely, higher) than the number of healthy patients. The presence of minority and majority classes of the initial data set requires the use of special techniques to eliminate imbalances and increase the recognition of the class represented by a minority of objects.

In order to solve the problem of atherosclerosis diagnostics, we applied such methods as random forest, method of extra-trees classifier, gradient boosting, and balancing classifier. The evaluation and comparison of the applied classification algorithms were carried out using the ROC AUC criterion.

2. Materials and methods

2.1. Study population
Patient data were provided by Voronezh Regional Cardiology Dispensary. The database contained the clinical examination results of the patients in the Bogucharsky district, carried out as part of the general medical examination program. During the examination 522 patients went through a number of required procedures, including multichannel volumetric sphygmmography. It allowed to perform the simultaneous measurement of systolic (SBP) and diastolic (DBP) blood pressure on the upper and lower extremities, calculate their difference, and automatically retrieve the values of the ankle-brachial indices (ABI) and
other indicators of limbs blood pressure asymmetry. Besides that, anthropometric, clinical, hemodynamic and other indicators of patients were analyzed. After the medical experts review, the most significant features were chosen as possible atherosclerosis predictors, i.e. independent analysis variables, used for the prediction of marker values. The main 28 predictors are listed in the table 1. For more information about this research see [4,5].

Table 1. Input features.

| Feature category     | Variables                                                                 |
|----------------------|---------------------------------------------------------------------------|
| Hemodynamic          | Systolic/diastolic/ pulse arterial pressure on the right/left arm/leg (SBPra, DBPra, Ppra, SBPla, DBPla, Ppla, SBPrl, SBPll), heart rate (HR), pulse wave velocity (cfPWVV, baPWV) |
| Socio-Demographic    | Gender, age, smoker                                                       |
| Anthropometric       | Height, weight, body mass index (BMI)                                     |
| Laboratory           | Glucose, cholesterol                                                     |
| Clinical             | Arterial hypertension (AH), stenocardia, infarction, acute disorder of cerebral circulation (ADCC), coronary artery bypass grafting/percutaneous intervention (CABG/PCI), diabetes, chronic heart failure (CHF), atrial flutter and atrial fibrillation (AF), obesity |

The study included 522 patients, 115 men (mean age 52.7 ± 13.6 years) and 407 women (mean age 51.9 ± 12.5 years). Smoking was significantly more likely to occur among the men, and their level of SBP was also significantly higher. At the same time, obesity was more common among women and hence they had a higher body mass index (BMI). Moreover, signs of atrial atherosclerotic lesions were found in 37 cases for any upper limb, (|∆SBPa| > 15 mm Hg), in 33 cases for any lower limb, in 70 cases for any limb. The pathological value of the ABI on at least one lower limb was recorded in 12 cases. The visualization of the distribution between pairs of integer input characteristics (Age, SBP, BMI) depending on the presence or absence of atherosclerosis is shown in figure 1. It shows, that the presence of atherosclerosis is accompanied by increased values of SBP and BMI.

2.2. Atherosclerosis markers
The variables reflecting the systolic blood pressure (SBP) asymmetry on the upper and lower extremities were used as diagnostic signs indicating the presence of atherosclerotic lesions in the arteries of the limbs. The medical experts often determine the pressure difference between the arms and legs (ArmsIndex and LegsIndex, respectively), as well as the ankle-brachial index (ABI) as the main atherosclerosis markers [6]. These markers represented by equations (1) - (3) are the output parameters of the initial classification problem.

\[
\text{ArmsIndex} = \begin{cases} 
1, & \text{if } |\Delta \text{SBP}| \geq \Delta_1 \\
0, & \text{otherwise}
\end{cases},
\]

\[
\text{LegsIndex} = \begin{cases} 
1, & \text{if } |\Delta \text{SBP}| \geq \Delta_2 \\
0, & \text{otherwise}
\end{cases},
\]

\[
\text{ABI} = \begin{cases} 
1, & \text{if } \text{ABI} \leq \Delta_3 \\
0, & \text{otherwise}
\end{cases},
\]

where \(\Delta_1\) and \(\Delta_2\) are the asymmetry coefficients of the SBP on the arms and legs, respectively and \(\Delta_3\) is the coefficient of ABI.
Figure 1. Pairwise features distribution.

2.3. Model building

The dataset under study represents a small sample in terms of machine learning approaches, since it contains 522 examples, which corresponds to the number of examined patients. In order to solve this problem, the stratified cross-validation procedure with shuffling [7] was applied during the building of classification models, which allowed, on the one hand, to operate with the whole dataset in the training process, and, on the other hand, to accurately evaluate the model’s generalization ability. Moreover, the stratified cross-validation procedure produces the uniform classes distribution at each iteration, which allows, in conjunction with the shuffling procedure, to achieve the highest possible representativeness of the initial dataset.

The feature of the problem under consideration is the imbalance of the initial sample classes, due to the fact that the number of healthy patients in the sample significantly exceeds the number of the patients who suffer from atherosclerosis, which is shown in table 2.

Table 2. The ratio of diseased patients to healthy patients.

| Marker   | The portion of patients with True markers, % | The portion of patients with False markers, % |
|----------|---------------------------------------------|---------------------------------------------|
| ArmsIndex| 8                                           | 92                                          |
| LegsIndex| 7                                           | 93                                          |
| ABI      | 2                                           | 98                                          |

The classical strategy of minimizing the total number of classification errors in this case can have a negative impact on the quality of recognition of a class represented by a minority of objects. To solve
this problem, various methods of machine learning (in particular, the method of decision trees, random forest), employ a classification matrix reflecting the costs associated with all possible outcomes \( \{C_{11}, C_{10}, C_{01}, C_{00}\} \), where \( C_{ij} \) is the cost of an error defining an element of a class \( i \) to a class \( j \).

In this study, such methods as random forest, extra-trees classifier, gradient boosting and balancing classifier were applied. These models are ensembles of machine learning models, which can significantly improve the classification accuracy when standard methods do not achieve the required efficiency.

RandomForest is the bagging over the decision trees. During the random forest training the features are selected from some random subset of variables obtained by bootstrap of the original sample \( X \) for each of the splits. I.e. random forest is an ensemble-type classifier, choosing the class for which most of the ensemble's algorithms voted as a classification response. This procedure is called majority voting. Bagging prevents the model’s overfitting due to the fact that errors of basic algorithms trained on different sub-samples are mutually compensated during voting, and also by the fact that the outliers may not be included into some training sub-samples.

However, in some cases overfitting is hard to avoid, especially when the original sample is unbalanced. Therefore, some classification tasks require the application of the efficient machine learning techniques, such as extra-trees classifier and gradient boosting. Extra-trees classifier models have a lot in common with random forests, except for the procedure of features thresholds selection during the nodes splitting. In the extra-trees classifier algorithm, the thresholds are arbitrarily selected for each possible feature, and the best of these randomly generated thresholds is chosen as the best rule for node splitting.

The gradient boosting is one of the most efficient algorithms due to the optimal approach to the trees composition into the final classifier. This method assumes that each subsequent ensemble tree is constructed in such a way that it minimizes the error of all previous trees. This approach is represented by the equation (4):

\[
\sum_{i=0}^{l} L(y_i, \hat{y}_i) = \sum_{i=0}^{l} L(y_i, a_{N-1}(x_i) + T_N(x_i)) \rightarrow \min \ ,
\]

\( T_N(x_i) \) is a new decision tree for the dataset with elements \( x_i \), \( L(y_i, \hat{y}_i) \) is a loss function, \( a_{N-1}(x_i) \) is a decision tree, built at a previous iteration \( N-1 \).

The sampling procedure was applied during the process of building the extra-trees, random forest and gradient boosting models, which allowed to reduce the degree of the sample imbalance. There are two main approaches to deal with imbalanced data: removing the instances of majority class (undersampling) and increasing the instances number of minority (oversampling) class. The deletion and duplication of class instances can be carried out not only in a random way, but also by certain rules (Tomek Links, Condensed Nearest Neighbor Rule, One-side sampling for undersampling, SMOTE, ASMO algorithms for oversampling). In this study the random undersampling strategy was applied. This algorithm generates the random \( K \) number of the majority class instances that are to be removed in order to achieve the required ratio of classes. \( K \) majority instances are then randomly selected and removed.

The application of the sampling procedure in conjunction with the methods of random forest, gradient boosting and extra-trees classifier made it possible to significantly improve the performance of the models. However, in addition to these methods, this study also addressed the Blagging classifier with the built-in class balancing procedure. The implementation of this algorithm includes the bootstrapping of the original sample, as well as the subsequent reduction of the majority class size for each sample generated at the bootstrap step. Thus, this model is a bagging algorithm over the decision trees with the built-in sampling procedure of removing the instances of a larger class, which results in the high classification efficiency without fitting the additional model settings. This algorithm is illustrated in figure 2 [8].
Classification models were built using the Python scikit-learn, imbalanced-learn, learning-from-imbalanced-classes libraries [9,10]. The base machine learning techniques are described in detail in [11-13], and the Python language basics are provided in [14].

2.4. Model performance
The values of the main metrics of the binary classification were used as performance indicators of the built models.

Table 3. Classification matrix.

| Actual class | Predicted class «+» (diseased) | Predicted class «-» (healthy) |
|--------------|---------------------------------|------------------------------|
| Class «+» (diseased) | TP | FN |
| Class «-» (healthy) | FP | TN |

Figure 2. Balancing classifier [8].

Based on the classification matrix (table 3), it is possible to calculate the main classification quality metrics, in particular Accuracy, Precision, Recall, and True Negative Rate, represented by equations (5) - (8). In terms of medical diagnostics, Recall and True Negative Rate have the meaning of sensitivity and specificity, i.e. reflect the ability of the model to classify diseased and healthy patients respectively.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}; \quad (5)
\]

\[
\text{Precision} = \frac{TP}{TP + FP}; \quad (6)
\]

\[
\text{Recall} = \frac{TP}{TP + FN}; \quad (7)
\]
In medical diagnostics, the method of study is optimal, which would be a priori both highly specific and highly sensitive, but in reality this is difficult to achieve.

In order to solve the problem of identifying the single most important classification metric we applied the integral ROC-curve instrument which takes into account both the sensitivity and specificity values. The area under the ROC curve is often selected by medical experts as an optimal and reliable criterion for diagnostic quality.

3. Results and discussions
In this study such models as random forest, extra-trees classifier, balancing classifier and gradient boosting were applied to each of the atherosclerosis markers and various sets of input features (clinical, hemodynamic, anthropometric, socio-demographic). The models have been compared using the ROC AUC metric, hence the best classifiers had the highest AUC values.

The most important goal of the medical diagnostics is to identify the groups of patients with the high risk of disease. Hence, the key requirement for the models developed was the acceptable value of the sensitivity metric, however the specificity and accuracy of the built classifiers were also taken into account.

3.1. Calibration of asymmetry coefficients
The calibration of the blood pressure asymmetry coefficients Δ₁, Δ₂, Δ₃ (equations (1)-(3)) was an important step of the models building and selection. At the first stage the values Δ₁=Δ₂=15 mm Hg, Δ₃=0.9 were suggested by medical experts and accepted as possible asymmetry coefficients. However the Δ₃ coefficient is well established in medical practice in contrast to Δ₁ and Δ₂, that are not generally accepted indicators and there are often discrepancies in their determination. In particular, the authors [6] consider the significant blood pressure difference as 10 < Δ₁ < 20 mm Hg for arms and Δ₂ > 20 mg Hg for legs.

The values of indicators Δ₁ =14 and Δ₂=15, which showed the best performance during the calculation experiment according to the AUC criterion, were chosen as the most informative asymmetry coefficients of SBP, indicating the presence of atherosclerosis.

3.2. Model comparison
In order to eliminate the problem of uneven distribution of the classes of the original dataset, such models as random forest, extra-trees classifier and gradient boosting were constructed using the random undersampling strategy. The usage of the misclassification costs for the random forest and extra-trees classifier models also resulted in the higher models performance. The balancing classifier model didn't require any additional adjustments, however resulted in the most efficient classification results.

The main metrics values are provided in the tables 4 – 6 for each of the model and atherosclerosis marker. The corresponding ROC-curves are illustrated in figures 3-5.

The models built using the methods of balancing classifier and extra-trees classifier have the best performance, since they result in the highest AUC metric values, and the random forest models most often have the lowest AUC (figures 3-5). The increased sensitivity of random forests, extra-trees classifier and gradient boosting is often accompanied by a significant drop in the specificity index, in contrast to the balancing classifier models, which have the highest specificity and accuracy, while slightly inferior in sensitivity criteria.

3.3. Model selection
The ROC curves show that the best models for the LegsIndex and ABI markers have been built using the balancing classifier. The model built for ABI and the set of hemodynamic features had the highest
AUC value (0.89). However the extra-trees classifier model outperformed the other classifiers for the ArmsIndex marker.

Table 7 contains the features of the most efficient models, built for each of the markers under consideration.

### Table 4. Models base metrics values for ArmsIndex.

| Method         | Sensitivity | Specificity | Accuracy | AUC  |
|----------------|-------------|-------------|----------|------|
| RandomForest   | 0.58        | 0.66        | 0.65     | 0.72 |
| ExtraTrees     | 0.6         | 0.64        | 0.64     | 0.73 |
| GradientBoosting| 0.63        | 0.61        | 0.62     | 0.7  |
| Blagging       | 0.46        | 0.78        | 0.75     | 0.72 |

### Table 5. Models base metrics values for LegsIndex.

| Method         | Sensitivity | Specificity | Accuracy | AUC  |
|----------------|-------------|-------------|----------|------|
| RandomForest   | 0.43        | 0.62        | 0.6      | 0.59 |
| ExtraTrees     | 0.59        | 0.69        | 0.68     | 0.67 |
| GradientBoosting| 0.59        | 0.64        | 0.64     | 0.69 |
| Blagging       | 0.45        | 0.79        | 0.76     | 0.71 |

### Table 6. Models base metrics values for ABI.

| Method         | Sensitivity | Specificity | Accuracy | AUC  |
|----------------|-------------|-------------|----------|------|
| RandomForest   | 0.58        | 0.9         | 0.89     | 0.79 |
| ExtraTrees     | 0.58        | 0.88        | 0.88     | 0.85 |
| GradientBoosting| 0.58        | 0.76        | 0.75     | 0.83 |
| Blagging       | 0.77        | 0.91        | 0.91     | 0.89 |

![Figure 3. ROC – curves for ArmsIndex and the full features set.](image)
Figure 4. ROC – curves for LegsIndex and the full features set.

Figure 5. ROC – curves for ABI and the hemodynamic features set.

Table 7. The most efficient models.

| Marker     | Features set                                                                 | Classifier  | AUC  |
|------------|-------------------------------------------------------------------------------|-------------|------|
| ABI        | SBPra, SBPll, HR, DBPra, cfPWV_{calc}> 10 m/s                                | Blagging    | 0.89 |
| ArmsIndex  | SBPla, SBPrl, SBPll, PPra, cfPWV, HR, Gender, AH, diabetes, CCF, Age, Height, Weight | ExtraTrees  | 0.73 |
| (Δ₁ = 14)  |                                                                             |             |      |
| LegsIndex  | SBPll, SBPla, SBPra, DBPra, HR, cfPWV, DBPla, Gender, AH, diabetes, CCF, Age, Height, Weight | Blagging    | 0.71 |
| (Δ₂ = 15)  |                                                                             |             |      |
4. Conclusion
As a result of this study, the comprehensive analysis of atherosclerosis markers and their predictors was carried out. We determined the asymmetry values of systolic blood pressure on hands and legs, which are the most informative signs of atherosclerosis. The problem was solved using the random forest, extra-trees classifier, gradient boosting and balancing classifier methods. As a result we built a number of efficient models with the high values of performance criterions. In particular, this result was achieved due to the random downsampling approach, which allowed to solve the problem of data imbalance. However, the models built with a balancing classifier had the best classification quality according to the AUC criterion.

References

[1] Demchenko M V and Kashirina I L 2017 Proc. Int. Conf. «Applied Mathematics, Computational Science and Mechanics: Current Problems» (Voronezh) p 636
[2] Kashirina I L and Demchenko M V 2019 Proc. Int. Conf. «Computer science: problems, methodology, technologies» (Voronezh) p 1161
[3] Demchenko M V and Kashirina I L 2018 Proc. Int. Conf. «Computer science: problems, methodology, technologies» (Voronezh) p 141
[4] Khokhlov R A, Ostroushko N I, Gaydashev A E, Kirsanov D V and Akhmedzhanov N M 2015 Rational Pharmacotherapy in Cardiology 11(4) 470
[5] Khokhlov R A, Gaydashev A E and Akhmedzhanov 2015 Rational Pharmacotherapy in Cardiology 11(5) 371
[6] Leng G C, Fowkes F G, Lee A J, Dubnar J, Housley E and Ruckley C V 1996 Use of ankle brachial pressure index to predict cardiovascular events and death: a cohort study BMJ 313(7070)
[7] Sholle F 2017 Deep learning with Python (New York: Manning Publications)
[8] Fawcett T 2016 Learning from imbalanced classes Silicon Valley Data Science https://www.svds.com/learning-imbalanced-classes/ 
[9] VanderPlas J 2016 Python Data Science Handbook: Essential Tools for Working with Data (Sebastopol: O'Reilly Media)
[10] Garreta R and Moncechi G 2013 Learning scikit-learn: Machine Learning in Python (Birmingham: Packt Publishing)
[11] Goodfellow I, Bengio Y and Courville A 2016 Deep learning (Cambridge: The MIT Press)
[12] Kane F 2019 Machine Learning, Data Science and Deep Learning with Python (New York: Manning Publications)
[13] Geron A 2019 Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (Sebastopol: O'Reilly Media)
[14] Ramalho L 2015 Fluent Python: Clear, Concise, and Effective Programming (Sebastopol: O'Reilly Media)