A Machine Learning Model for the Atherosclerosis Prediction Based on Clinical Data

Kateryna Kolesnikova, Dariya Mochalova and Vladyslav Lavrynovych

Taras Shevchenko National University of Kyiv, Volodymyrska str., 60, Kyiv, 01033, Ukraine

Abstract

The paper outlines the relevance of diagnostics of atherosclerotic disease from the standpoint of the integrity of the research. It determines that the ahterosclerotic disease is the fist step to more sirious cardiovascular diseases, so it is verry importatnt to diagnose it in early stages. It proposes the solution - the development of technology to diagnose atherosclerotic disease. The definition of such technology has been given. It has been established that in terms of technology it is important to develop an effective and optimized model for the prediction of atherosclerosis from the standpoint of all stages of the research. It has been discovered that there are more than one efficient algorithm that can be used for such purpose. The outcome technology of atherosclerotic disease has been compiled and researched based on a dataset of 1000 patient. The solution is implemented using machine learning methods, using Python programming language as a base for the software product. The research resulted in a technology based on models with an accuracy between 98.75% and 100%. The prospects of further research - the implementation of the diagnistic system itself, which can be integrated with overviewed techniques along with computer vision and other technologies that may improve diagnosis and treatment of atherosclerosis. The paper identifies the challenges and perspectives of the research.

Keywords [1]

data science, machine learning, deep learning, binary classification, atherosclerosis, heart disease

1. Introduction

Nowadays, cardiovascular diseases (CVDs) are one of the leading causes of death all over the world. According to the World Health Organization's data [1, 2], 32% of all global deaths are caused by CVDs, which is around 17.9 million lives every year. Atherosclerosis, which is the subject of this study, tends to be one of the main underlying causes of CVDs, also playing a key role in heart stroke and peripheral artery disease (PAD). Atherosclerosis is very common, and usually followed by a set of risk factors like high cholesterol, obesity, inactivity, diabetes, etc. Atherosclerosis is a complex process, usually slow and progressing in the long-term perspective. Even today it's not completely clear what exactly causes this process and why. Atherosclerosis is often characterized by narrowing and hardening arteries, and the symptoms depend on what artery is narrowed or blocked. Atherosclerosis starts with damage to the endothelium of blood vessels, frequently caused by high cholesterol, blood pressure, inflammation and smoking. Entering the damaged area of the artery, cholesterol and other cell parts become plaque in the artery wall. As long as atherosclerosis progresses, plaque gets bigger and may create a blockage when it's big enough, causing severe consequences [3].

There were several studies for different computer-aided approaches to this issue in recent years, but for all that, the problem remains highly challenging today. Data mining and machine learning is a state-of-art technology, which allows us to discover connections between attributes of large scaled data and train models to make predictions more accurately. Machine learning has already found its application

[1] *Information Technology and Implementation (IT&I-2021), December 01–03, 2021, Kyiv, Ukraine*

amberk4@gmail.com (Kateryna Kolesnikova); daria.mochalova.02@gmail.com (Dariya Mochalova); vlad.lavrynovych@icloud.com (Vladyslav Lavrynovych)

in different medicine realms for disease prediction [4 – 6] and proved to be quite efficient, depending on tasks, algorithms and data. It is worth mentioning that a lot of studies have been performed doing classification for prediction of diagnosis of heart disease using various models and methods. Nonetheless, most of them were limited with data and were relying on a 13-feature Cleveland dataset [7], that includes only 303 records. Such an approach may not only narrow the research field, but also result in a row of inaccurate conclusions [8], caused by the lack of data needed for model training. For our model we will use a real dataset, provided by Amosov National Institute of Cardiovascular Surgery which will help us to talk not only about theoretical results but find different approaches for real cases. Unlike the other studies, we will not limit to prediction of cardiovascular disease, but will try to predict atherosclerosis, which is a precondition of CAD and is not as well researched due to the reasons described above. Therefore, the study carries both scientific and practical interest for the audience and may bring some light to crucial medicine problems.

2. Related studies and algorithms overview

Currently, a lot of available atherosclerosis prediction studies using machine learning approach rely on Cleveland heart disease dataset [7] and thus are not completely representative, as this dataset contains data only about coronary heart disease which is a bit different from the subject of our study. However, such studies claim to be describing prediction of atherosclerosis [9], which is actually a wrong statement. Indeed, a lot of studies of coronary heart disease prediction using machine and deep learning techniques were performed, but prediction of atherosclerosis is relatively poorly researched due to the lack of data. However, there are several interesting studies [10, 11] in this field and most of the techniques applied for CAD prediction may also be applied here.

2.1. Support vector machine

SVM is a set of supervised machine learning algorithms used for classification and regression analysis. This method relies on a hyperplane or a set of hyperplanes in multidimensional space which separate the data into classes. Algorithm finds points closest to the hyperplane from both classes as illustrated on Figure 1.

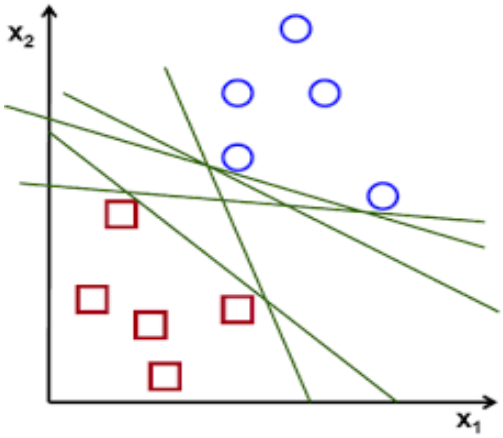

Figure 1: SVM algorithm visualization

These points are called support vectors and computing maximized distance between support vectors and hyperplane we find the optimal hyperplane (the more distance between classes, the better result). This algorithm was already used for the prediction of CAD [12] and reached up to 96.67% accuracy, thus in this paper we will also try to apply it to predict atherosclerosis.

2.2. Naive Bayes

Naive Bayes classifier is a supervised machine learning algorithm based on Bayes theorem, which "naively" assumes that all the features independently contribute to the probability of belonging to some class. Formulas (1, 2) express the calculation of posterior probability of class.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c) \quad (2)$$

Where P(c|x) - posterior probability of class, c - target, x - attributes, P(c) - prior probability of class, P(x|c) - probability of predictor given class, P(x) - prior probability of predictor [13]. Naive Bayes algorithm is known for well performance on large scaled data and sometimes outperforms even more sophisticated methods. Along with its simplicity, the algorithm found its way into prediction of CAD and atherosclerosis [14] reaching 98.60% of accuracy, and we will also test this method against our dataset for comparison purposes.

2.3. Decision tree

The next algorithm we will consider in this study is a decision tree. This algorithm is widely used for classification and regression purposes and represents a tree-structured classifier. Leaves represent target classes, each node - a test case for a particular attribute of data and edges are the result of a test case. Decision trees form nested if-else statements and the deeper the tree - the fitter the model. A brief illustration of the algorithm is shown on Figure 2.

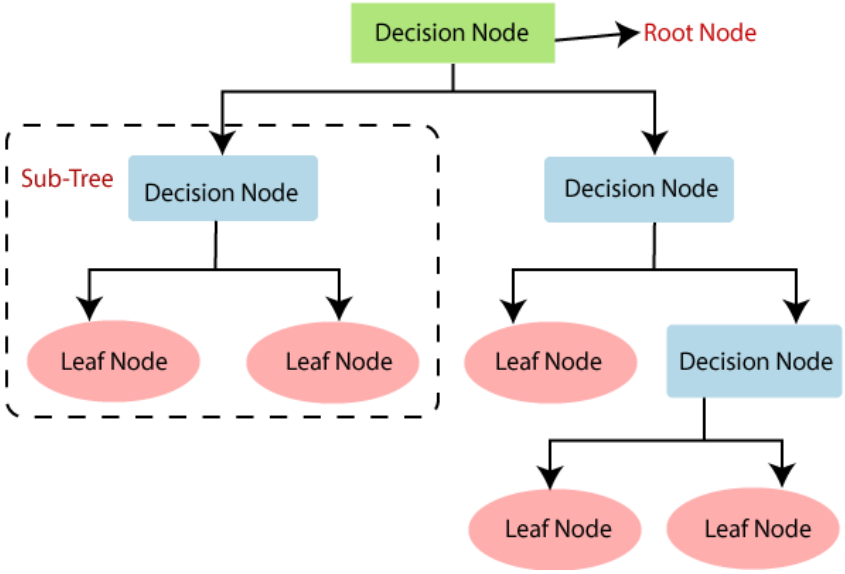

Figure 2: Schematic visualization of decision tree algorithm

Decision tree was successfully applied for atherosclerosis prediction with obtained accuracy 82.6% [15]. In this study we will analyze the accuracy of the algorithm applied to our dataset and evaluate the reasonableness of its use in this area.

2.4. Random Forest

Random forest represents ensemble learning, which means combining many classifiers to obtain a solution (classification trees). This is achieved by averaging the prediction of each classifier. Random Forest technique allows the model to learn complex relations and increase accuracy for predictions. Due to flexibility of the algorithm, it produces good results even without hyper-parameter tuning. Random forest was not yet widely applied to atherosclerosis prediction, but has shown relatively good results in CAD prediction (87.64%) [16].

2.5. XGBoost

XGBoost is a relatively new machine learning algorithm, which stands for eXtreme Gradient Boosting. This algorithm, as well as the random forest, is based on decision-tree ensemble learning, but

the difference is that it uses a gradient boosting framework and is highly optimized and uses less resources than plain gradient boosting. The method was introduced in 2016 and showed good performance on different tasks and also recently was applied to atherosclerosis prediction based on electronic health records [17] showing accuracy 74%, and CAD prediction with accuracy 91.8% [18].

2.6. Deep neural network

Deep learning is a machine learning technique that uses multiple layers to extract high-level relationships and features from data. Nowadays deep learning is recognized as a state-of-art technology, which is flexible and provides good results allowing to optimize the accuracy during the train process. However, deep learning requires large and well prepared datasets as well as computational resources for the training process. Neural network takes inputs and modifies the neuron weights in accordance with the error rate calculated between actual and predicted value. Hidden layers allow the network to learn nonlinear relations in the dataset. The output layer represents only one output for binary classification (which is our case), where 0 is absence of atherosclerosis and 1 is its presence. The scheme for such a neural network is represented on Figure 3.

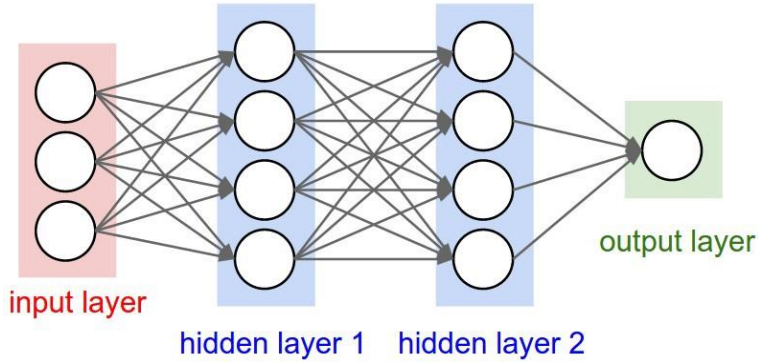

Figure 3: Visualization of binary classification ANN architecture

Different neural network architectures were applied to cardiovascular disease prediction, in particular, study of cardiovascular disease prediction using deep learning techniques [19] introduced ANN with prediction accuracy 85%. In this study we will create and apply our own ANN architecture and try to improve the precision score.

3. Methodology

3.1. Data description

The dataset used for this study was provided by Amosov National Institute of Cardiovascular Surgery. The dataset contains 14 columns and 1000 records of patients. Most of the latest research of heart diseases refer to the UCI dataset which dates back to 1988. Having such a new and accurate dataset provides a unique opportunity for atherosclerosis prediction based on already existing methods and applying new, which opens new doors to the application of machine learning algorithms in medicine. The sample of this dataset is shown on Figure 4.

The dataset does not contain empty cells, all attributes are filled and have normal distribution. The dataset includes next attributes:

- Progress - display presence (1) or absence (0) of atherosclerosis.
- OP - surgical intrusion, 1 - present, 0 - absent;
- Shunt - cardiac shunt, 1 - present, 0 - absent
- age - age, years;
- height - height in cm;
- weight - weight in kilos;
- IMT - BMI, body index mass;
- sex - 0 - male, 1 - female;
- ChSS - heart rate, beats in one minute;

- AD sist. - systolic blood pressure;
- AD diast - diastolic blood pressure
- AG therapia - antihypertensive therapy, 1- present, 0 - absent
- cholesterin - total cholesterol levels;
- diabetus melitus - diabetes, 1 - present, 0 - absent;

	Progress	OP	Shunt	age	height	weight	IMT	sex	ChSS	AD sist.	AD diast	AG therapia	cholesterin	diabetus melitus
0	0	0	0	9	110	35	28.93	0	65	122	82	0	0	0
1	0	0	0	10	120	40	27.78	0	63	115	73	0	0	0
2	0	0	0	11	121	33	22.54	1	61	118	78	0	0	0
3	0	0	0	11	125	30	19.20	0	59	119	80	0	0	0
4	0	0	0	12	142	45	22.32	1	63	120	80	0	0	0
..	...	..	...	...	...	...	...	...	...	...	...	...	...	...
995	1	0	0	91	173	75	25.06	0	63	134	118	1	3	0
996	1	1	0	91	173	75	25.06	0	63	134	118	1	3	0
997	1	1	0	91	173	75	25.06	0	63	134	118	1	3	0
998	1	1	1	92	165	54	19.83	0	67	155	71	1	3	0
999	1	1	0	92	177	88	28.09	0	54	152	74	0	3	0

[1000 rows x 14 columns]

Figure 4: The sample of dataset records

The Progress attribute is a target value of this study and the algorithms' predictions will be compared to it. The dataset contains the data of 591 patients with present atherosclerosis, and 409 rows of data from healthy people. Also it is important to check sex distribution in the dataset, which is presented on Figure 5.

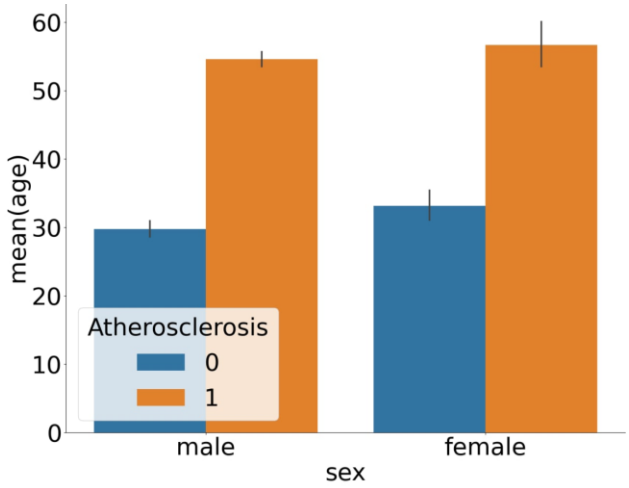

Figure 5: Distribution of data by sex, age mean and atherosclerosis presence.

As we can see on Figure 5, the dataset contains a little bit more records of females than males. Also the age mean of healthy people is 30.96, while the mean of age of patients with atherosclerosis is 54.78, and this tendency is true for both males and females. Given that information we can conclude that atherosclerosis is more common for older people, and to prove that, we will check the age distribution of patients in our dataset. As we can see on Figure 6, atherosclerosis mostly is not present in young people, and on the contrary, was diagnosed in the majority of middle-aged and older people. This conclusion is also proved by the result of a recent study [20] which states that atherosclerosis rapidly develops between ages 40 to 50. The Figure shows a large gap in disease occurrences between 37 and 38 years , and after that, the tendency of disease increasing occurs, which represents the general heart disease statistics and risk factors impact.

Also the number of healthy patients constantly decreases after the age around 30, which means that people are often diagnosed with disease when it is too late. As was mentioned in the introduction, atherosclerosis is usually a long-term progressing disease, however sometimes it might progress more aggressively. The disease usually shows its symptoms in the later stages, when arteries are narrowed or blocked which is followed by pain in chest or other manifestations. Early atherosclerosis diagnosis may help to avoid a set of heart diseases and save a lot of lives as a consequence.

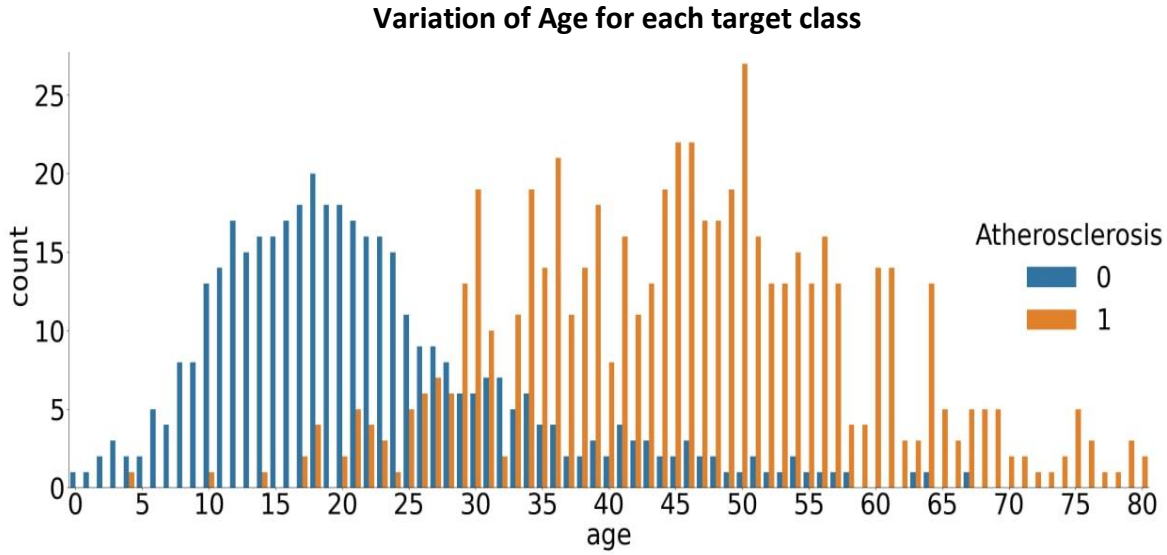

Figure 6: Disease distribution by age

It is important to define relations between the attributes and their impact on the target value before we start application of machine learning algorithms. For that purpose we will calculate correlation between all attributes and build a heat map, which is displayed at Figure 7.

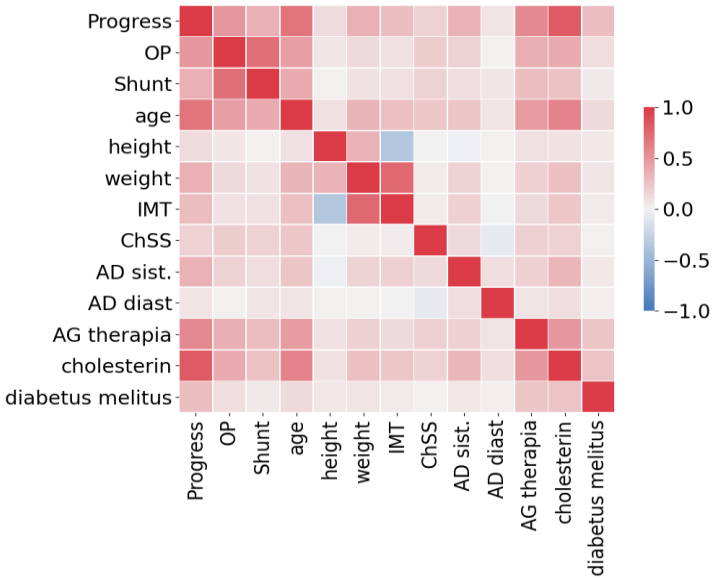

Figure 7: Attribute correlation heatmap

Based on the heatmap, we can define correlation between attributes of our data set. The first thing we should look at is the first row of the diagram, which represents correlation of each separate attribute with our target data. The next thing is to define which attributes impact the target via other attributes indirectly. Thus, cholesterin, age and AG therapia have strong positive correlation with target. OP, Shunt, weight and AD sist. have moderate positive relationships, no strong indirect correlated attributes were found. Now, let's take the most correlated parameters and display data distribution in this 3D plane. Because AG thrapia has binary values we will replace it with weight which has moderate

correlation, but more diverse value distribution. The visualisation of the data for such a plane is displayed on Figure 8. In Figure 8, we can see that most patients with atherosclerosis have high cholesterol level, are middle-aged and older, and overweight, while healthy patients mostly have weight under 80 kilos, low cholesterol level and are under 40. This conclusion reflects the general idea of atherosclerosis and corresponds to the risk factors: atherosclerosis is more common for older people, people with overweight and high cholesterol levels. Based on the plot above we can see that atherosclerosis and non-atherosclerosis records can be distinguished by these three parameters, however some atherosclerosis cases occur even in people with normal weight and low cholesterol levels, but those are minor.

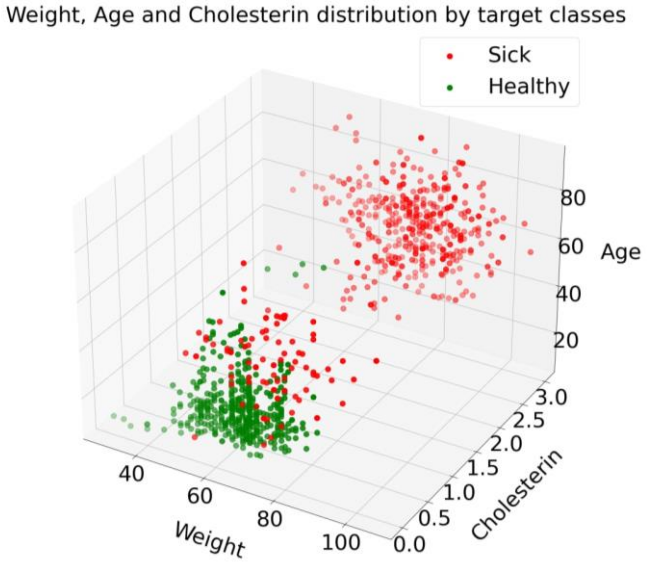

Figure 8: Data distribution in 3d plane by most correlated attributes and target class

1.1. Applied software technology

For this study we used Python programming language of 3.8 version and its ecosystem. Nowadays Python is the most popular programming language for data analysis and machine learning, and offers a lot of libraries and solutions for solving such issues. Python provides a lot of utilities which reduces development time and provides highly efficient results.

We used the next set of python libraries for this study:

- pandas - the library which provides functionality for creating and operating with datasets;
- numpy - allows to perform sophisticated calculations on high-performance multidimensional arrays, and operate with them;
- matplotlib - offers a software interface different visualizations of data;
- seaborn - data visualization library based on matplotlib, which provides high-level interface and a lot of presets for drawing more user-friendly plots as well as a variety of diagrams, heatmaps, color themes, etc;
- scikit-learn - offers various unsupervised and supervised ready-to-use machine learning algorithms, built upon numpy, pandas and plotlib;
- xgboost - optimized distributed library that provides gradient boosting algorithm implementations;
- keras - high-level neural network API that provides functionality for developing and evaluating deep learning models;
- tensorflow - open-source platform that provides a backend engine for keras
- ann_visualizer - visualization library that is used to work with keras, uses graphviz library to create a graph of the neural network;
- graphviz - an open source graph visualization software that provides functionality to represent structural information;

1.2. Application of algorithms

The goal of this research is to predict whether patients have atherosclerosis or not. The research was done using supervised machine learning techniques: naive bayes, decision tree, random forest, XGBoost and neural network as a deep learning technique. We will elaborate on the neural network, as it is more complicated in configuration and tuning. For the neural network we used a set of dense layers with dropout to avoid overfitting and ReLU as an activation function. For the output layer we used sigmoid function, binary cross entropy loss function, because the task of the model is binary classification, and adam optimizer. The dataset was divided into test and training sets, 20% and 80% accordingly. The training process consisted of 500 epochs to reach better accuracy of the result. The visualizations of model loss and model accuracy improvement are shown on Figure 9. As we can see, the neural network training process was balanced without overfitting and the model reached good accuracy. One more thing that should be mentioned before comparison of results is the decision tree structure. Decision tree classifier creates rules based on parameters that allow it to classify data. Thus, this structure may help figure out what parameters affect the classification result most. The structure of the received decision tree is displayed on Figure 10. Based on the Figure 10 we can conclude that most important parameters for classification result are cholesterin, AD sist. and weight. The general performance results comparison are reflected on Table 1. To evaluate the precision score of all algorithms, a confusion matrix was used. Among all applied algorithms, Random Forest and Neural Network showed best performance for both training and test process. Also we should notice that all applied models reached a very high accuracy score, which makes them applicable for atherosclerosis prediction in medical institutions.

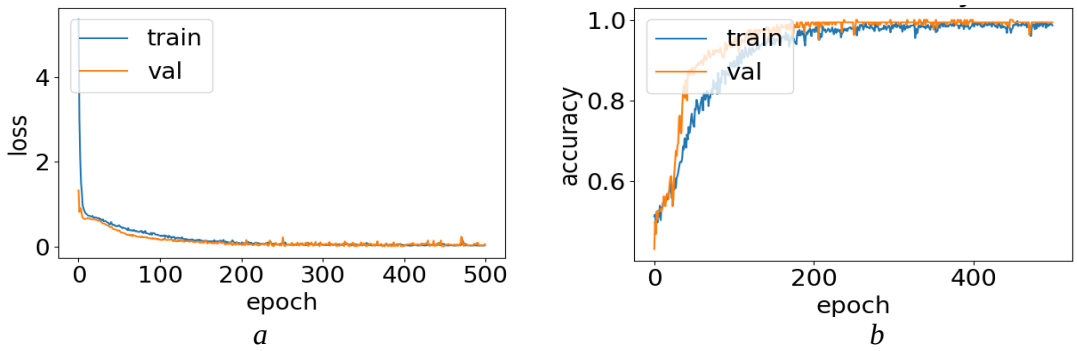

Figure 9: Neural network training plot for loss (a) and accuracy (b)

Table 1
Algorithms' accuracy scores

Algorithm	Training accuracy	Test accuracy
SVM	0.9975	1.0
Naive Bayes	0.9875	1.0
Decision Tree	1.0	0.995
Random Forest	1.0	1.0
XGBoost	1.0	0.995
Neural network	0.99875	1.0

2. Conclusion

In this paper 5 machine learning methods had been analyzed for atherosclerosis prediction. Our team trained and tested all the algorithms against the clinical data. It achieved promising results after what the accuracy of models have been compared. All the models showed extremely high performance scores, and performed better in this study in comparison with overviewed application cases with CAD disease dataset.

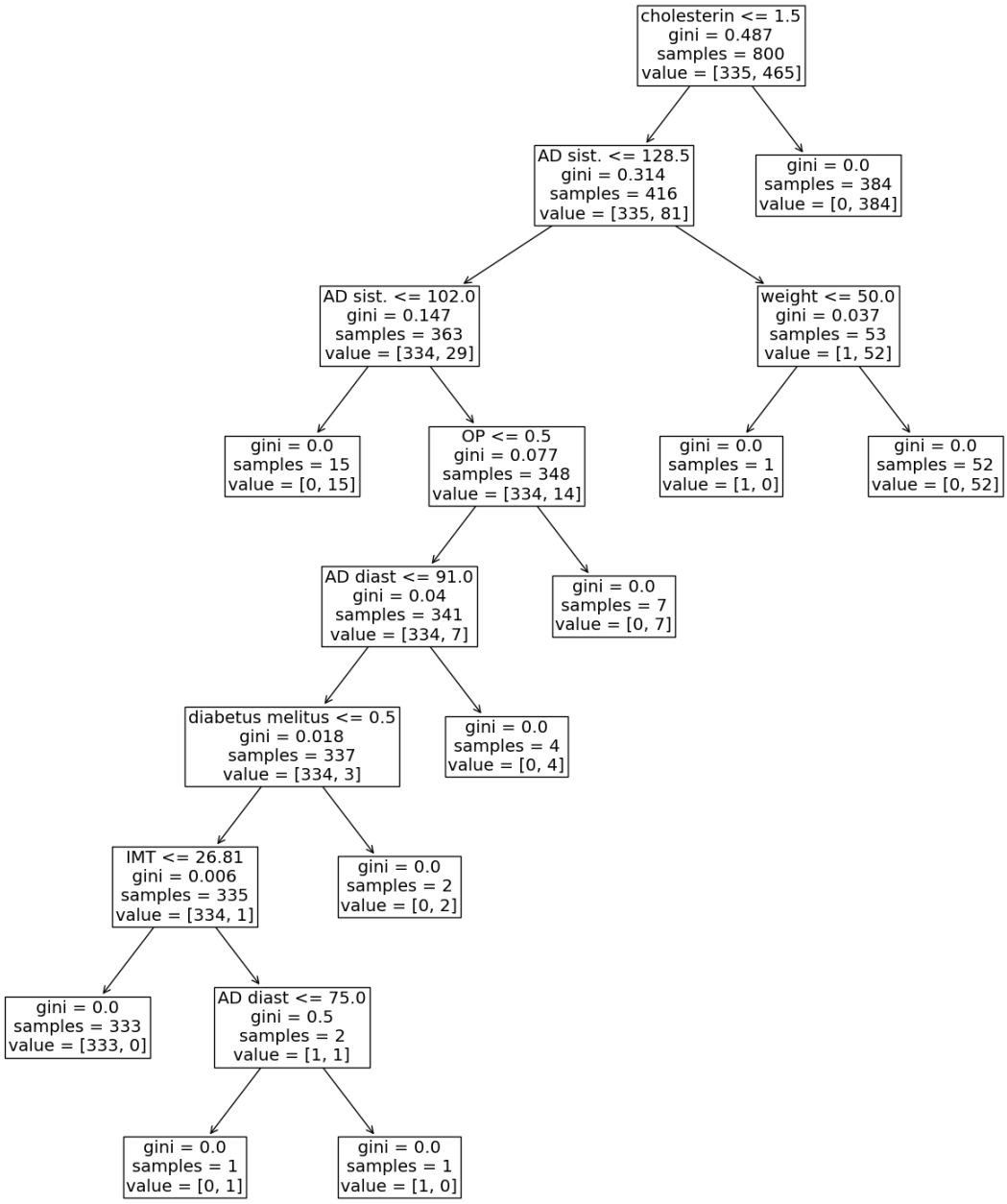

Figure 10: Decision tree structure

We used a confusion matrix for comparison of ML algorithms' performance for training and testing sets. Many researchers note that ML algorithms show better performance for not large datasets, whereas deep learning neural networks are better for large scaled data. However with right hyperparameter tuning and architecture can be reached good results even for small-sized datasets which was proved in this research.

Considering limitations of the research, there are a lot of broad opportunities for applying mentioned methods to the data of larger size, which however may lead to more technical challenges such as complex data preprocessing and algorithms tuning. Also a lot of other neural network architectures may be applied as well as ML methods for achieving better results. Though there is a very limited number of datasets that are available for atherosclerosis analysis nowadays (which makes the field attractive for many researchers), there are a lot of possible integrations of overviewed techniques along with computer vision and other technologies that may improve diagnosis and treatment of atherosclerosis.

3. References

[1] Fact sheets Cardiovascular diseases (CVDs) / World Health Organization, 2021. Mode of access: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] Cardiovascular diseases Overview / World Health Organization, 2021. Mode of access: https://www.who.int/health-topics/cardiovascular-diseases.

[3] Rafieian-Kopaei M, Setorki M, Doudi M, Baradaran A, Nasri H. Atherosclerosis: Process, Indicators, Risk Factors and New Hopes. Int J Prev Med (2014) 5(8):927–46.

[4] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis "Machine learning applications in cancer prognosis and prediction." Computational and Structural Biotechnology Journal, Volume 13, 8-17, 2015.

[5] D. P. Yadav and S. Rathor, "Bone Fracture Detection and Classification using Deep Learning Approach," 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), 2020, pp. 282-285.

[6] Verma AK, Pal S, Kumar S. Classification of Skin Disease using Ensemble Data Mining Techniques. Asian Pac J Cancer Prev. 2019 Jun 1;20(6):1887-1894.

[7] UCI Machine Learning Repository, "Heart disease data set," 2021, Mode of access: http://archive.ics.uci.edu/ml/datasets/heart+disease.

[8] Singh P, Singh S, Pandi-Jain GS. Effective heart disease prediction system using data mining techniques. Int J Nanomedicine. 2018 Mar 15;13(T-NANO 2014 Abstracts):121-124.

[9] O. Terrada, B. Cherradi, A. Raihani and O. Bouattane, "Atherosclerosis disease prediction using Supervised Machine Learning Techniques," 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), 2020, pp. 1-5.

[10] Munger E, Hickey JW, Dey AK, Jafri MS, Kinser JM, Mehta NN. Application of machine learning in understanding atherosclerosis: Emerging insights. APL Bioeng. 2021 Feb 16;5(1):011505.

[11] Y.Khlevna, D.Mochalova. Prediction of atherosclerosis disease with artificial neural network. Sciences of Europe. Technical sciences. VOL 1, No 50 (2020) pp. 53 –58.

[12] Zhu Y, Wu J, Fang Y. [Study on application of SVM in prediction of coronary heart disease]. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi. 2013 Dec 30(6):1180-5. Chinese.

[13] 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R / Sunil Ray, 2017. Mode of access: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained

[14] Oumaima Terrada, Bouchaib Cherradi, Abdelhadi Raihani, Omar Bouattane, A novel medical diagnosis support system for predicting patients with atherosclerosis diseases, Informatics in Medicine Unlocked, Volume 21, 2020.

[15] Qawqzeh, Y.K.; Otoom, M.M.; Al-Fayez, F.; Almarashdeh, I.; Alsmadi, M. and Jaradat, G. A Proposed Decision Tree Classifier forAtherosclerosis Prediction and Classification. IJCSNS, 2019,19(12), p.197.

[16] Akella A, Akella S. Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution. Future Sci OA. 7(6):FSO698. March 2021.

[17] Fan, J., Chen, M., Luo, J. et al. The prediction of asymptomatic carotid atherosclerosis with electronic health records: a comparative study of six machine learning models. BMC Med Inform Decis Mak 21, 115 (2021).

[18] Kartik Budholiya, Shailendra Kumar Shrivastava, Vivek Sharma, An optimized XGBoost based diagnostic system for effective prediction of heart disease, Journal of King Saud University - Computer and Information Sciences, 2020.

[19] Syed Nawaz Pasha, Dadi Ramesh, Sallauddin Mohmmad, A. Harshavardhan and Shabana / Cardiovascular disease prediction using deep learning techniques, Mode of access: https://iopscience.iop.org/article/10.1088/1757-899X/981/2/022006

[20] Journal Article, Beatriz López-Melgar, Leticia Fernández-Friera, Belén Oliva, José Manuel García-Ruiz, Fátima Sánchez-Cabo, Héctor Bueno, José María Mendiguren, Enrique Lara-Pezzi, Vicente Andrés, Borja Ibáñez, Antonio Fernández-Ortiz, Javier Sanz, Valentín Fuster Short-Term Progression of Multiterritorial Subclinical Atherosclerosis, 2020, Journal of the American College of Cardiology, 1617-1627