Prediction of Coronary Plaque Progression Using Data Driven Approach

Bojana Andjelkovic Cirkovic[1,2(✉)], Velibor Isailovic[1,2], Dalibor Nikolic[2], Igor Saveljic[1,2], Oberdan Parodi[3], and Nenad Filipovic[1,2]

[1] Faculty of Engineering, University of Kragujevac, Kragujevac, Serbia
{abojana, velibor, isaveljic, fica}@kg.ac.rs
[2] Research and Development Center for Bioengineering "BioIRC", Kragujevac, Serbia
markovac85@kg.ac.rs
[3] CNR Clinical Physiology Institute, Pisa, Italy
oberpar@tin.it

Abstract. Coronary artery disease or coronary atherosclerosis (CATS) is the most common type of cardiovascular disease and the number one cause of death worldwide. Early identification of patients who will develop progression of disease is beneficial for treatment planning and adopting the strategy for reduction of risk factors that could cause future cardiac events. In this paper, we propose the data mining model for prediction of CATS progression. We exploit patient's health record by using various machine learning methods. Predictor variables, including heterogenious data from cellular to the whole organism level, are initially preprocessed by feature selection approaches to select only the most informative features as inputs to machine learning algorithms. Results obtained and features selected within this study indicate the high potential of machine learning to be used in clinical practice as well as that specific monocytes are important markers impacting the plaque progression.

Keywords: Coronary artery disease · Atherosclerosis progression
Machine learning · Feature selection

1 Introduction

Coronary artery disease, specifically coronary atherosclerosis, is one of the leading causes of death worldwide. This enduring inflammatory disease of the arterial wall is result of chronic formation of fatty streaks and atheromas, followed by acute thrombotic events. The number of patients with coronary atherosclerosis was rapidly increased over the past decade and the disease has tendency to be an epidemic in the near future [1]. Pharmacological treatment and revascularization are standard treatment modalities used to control and prevent future cardiac events. Several studies [2–4] showed that, although the revascularization of significant hemodynamic stenosis in combination with pharmacological therapy is superior in comparison to medicaments only, not all patients will benefit from revascularization. Understanding of the

O. Fratu et al. (Eds.): FABULOUS 2017, LNICST 241, pp. 227–233, 2018.
https://doi.org/10.1007/978-3-319-92213-3_33

hemodynamic significance of atherosclerotic lesions via additional assessment is crucial in order to better manage therapy and optimize patient outcomes.

Many artery specific and patient-specific factors are known to contribute to the process of atherosclerosis formation and progression by a complex interaction between biological and mechanical elements. These factors were mostly explored by computing correlations between single features and the angiography outcome. However, several moderate risk factors may, in combination, result in a much higher risk than an impressively raised single factor [5]. For this reason, it is necessary to examine the disease from cellular to the whole organism level and collect the expanded list of factors in order to find the proper combination among them that are the most correlated with the progression of CATS. This will significantly improve strategies for identification of patients with high-risk CATS and early intervention.

The number of computational techniques has been developed in order to understand the mechanisms of CATS progression. On the other side, machine learning methods (ML) provide the framework to learn complex patterns from data. Due to ability to handle the high-dimensionality nature of patients' data [6, 7] and make accurate patient specific predictions in a real time, this approach has become a popular research tool for medical researchers who seek to gain insights and exploit the information in a more effective manner.

In this study, we employ a machine learning approach based on heterogeneous data collected within the SMARTool project [8], which include both imaging data (Coronary Computed Tomography Angiography - CCTA) and clinical, molecular and cellular data, to predict progression of CATS. More specifically, we analyzed information coming from the patients' medical records, biochemical analyzes, adhesion molecules, and monocyte markers with the aim to model the progression of atherosclerosis in coronary vessels. The problem was defined as binary classification problem since we wanted to predict whether the progression of disease will occur or not for a specific patient.

2 Methods

2.1 Patients and Dataset

In this study, we considered medical records of 55 patients obtained within the H2020 SMARTool project. These patients underwent angiographic assessment by coronary computed tomography angiography in order to evaluate the percentage of stenosis. In this study, results were based on the assessment of disease for each patient performed in two time moments, with the average time between two assessment of 10.2 ± 3 months (median time = 11 months). Moreover, wide series of data obtained during the second time-slice were collected, namely, the patient's medical history, relevant risk factors, biochemical markers, therapy and monocyte markers and each patient is described with 72 features (Table 1).

The CCTA metrics adopted to define clinically relevant disease progression at each artery site are the following:

Table 1. List of features in dataset.

Clinical			
Gender (Binary)	Age (Numeric)	BMI (Numeric)	Alcohol consumption (Numeric)
Current Smoking (Binary)	Past Smoking (Binary)	Physical Activity (Binary)	Diet/Vegetable (Binary)
Diabetes Mellitus (Binary)	Dyslipidemia (Binary)	Hypertension (Binary)	Metabolic Sindrome (Binary)
Family History CHD (Binary)	Current Symptoms (Binary)	Infarct (Binary)	CABG (Binary)
PICStenting (Binary)			
Therapy			
ARB (Binary)	Diuretics (Binary)	Statins (Binary)	oral Antidiabetics (Binary)
Aspirin (Binary)	Insulin (Binary)	BETA Blockers (Binary)	ACE Inhibitors (Binary)
DAPT (Binary)	Nitrates (Binary)	Calcium Antagonists (Binary)	Other Drugs (Binary)
Blood Test			
Creatinine (Numeric)	HDL (Numeric)	Leukocytes (Numeric)	Total Cholesterol (Numeric)
Erythrocytes (Numeric)	Hemoglobin (Numeric)	MCH (Numeric)	Triglycerides (Numeric)
Fasting Glucose (Numeric)	INR (Numeric)	MCV (Numeric)	Uric acid (Numeric)
Fibrinogen (Numeric)	LDL (Numeric)	Platelets (Numeric)	aPTT (Numeric)
HCT (Numeric)			
Inflammatory			
ICAM1 (Numeric)	VCAM1 (Numeric)		
Monocytes			
CCR2_% (Numeric)	CD14(++/+)_% (Numeric)	CD163_ RFI (Numeric)	CX3CR1_ RFI (Numeric)
CCR2_RFI (Numeric)	CD14(++/+)_RFI (Numeric)	CD16_% (Numeric)	CXCR4_% (Numeric)
CCR5_% (Numeric)	CD14++/CD16+/CCR2+_% (Numeric)	CD16_ RFI (Numeric)	CXCR4_ RFI (Numeric)
CCR5_ RFI (Numeric)	CD14++/CD16-/CCR2+_% (Numeric)	CD18_% (Numeric)	HLA-DR_% (Numeric)
CD11b_% (Numeric)	CD14+/CD16++/CCR2-_% (Numeric)	CD18_ RFI (Numeric)	HLA-DR_ RFI (Numeric)
CD11b_ RFI (Numeric)	CD163_% (Numeric)	CX3CR1_% (Numeric)	MONOCYTE COUNT_ RFI (Numeric)
Progress (Binary)			

- Plaque progression: further luminal diameter narrowing > 20% at the site of pre-existing stenosis;
- De novo plaque formation: development of a new plaque causing a luminal narrowing >20% in a previously normal segment.

2.2 Feature Selection

In machine learning and data mining, feature selection is the process of selecting a subset of features for use in model construction. There are plenty of studies pointing out that learning algorithms may be adversely affected by the presence of irrelevant and/or redundant features and that feature selection in most of cases improves the classification accuracy [9, 10]. Additionally, feature selection significantly reduces computation complexity, prevents over-fitting and facilitates data understanding as well as data acquiring.

Specifically, in this study we used three state-of-the-art filter algorithms for feature selection: ReliefF [11], mRMR [12] and Gain ratio [13] widely used for extraction of useful features from data.

1. The ReliefF algorithm is based on the idea that useful features should differentiate between instances from different classes and have similar values from instances from the same class. It randomly choses an instance from the dataset, finds its nearest neighbor from the same and opposite class, and updates relevance score for each feature by comparing the values of nearest neighbors to the sampled instance.
2. The mRMR algorithm maintains the features that have high correlation with the class attribute and low inter-correlation among themselves.
3. Gain ratio (GR) normalizes the information gain score of splitting on an attribute by the entropy of this attribute.

2.3 Classification

A set of seven representative supervised machine learning algorithms were used for knowledge extraction and performing of classification [13]: naive Bayes classifier (NB), Bayesian Network (BN), decision tree (DT) based on C4.5 algorithm, Random Forest (RF), Support Vector Machine (SVM), K-Nearest neighbor (K-NN) and artificial neural network (ANN). ANN was constructed as a multi-layer perceptron with one hidden layer, unipolar sigmoid activation function was set in all neurons, and learning algorithm was back propagation with momentum. SVM was run with a polynomial as well as the RBF kernel.

All these algorithms are implemented in WEKA software packet [14] which is used for the resolving of the defined classification problem.

For evaluation of the classifiers, the 10-fold cross validation procedure was used, where, in each repetition, instances belonging to the left-out fold were used for testing purposes (test set), while the remaining observations were used for feature selection and classifier training (training set). In each iteration of the cross validation procedure we applied the classifier‘s parameter selection algorithm (CVParameterSelection algorithm from the Weka machine learning software) [15] which explores all possible combinations for a set of classifier's parameters by applying nested 10-fold cross validation procedure on the training set. For given classifiers the following parameters were optimized:

- For SVM: C ∈ $\{10^{-3}, 10^{-2}, \ldots, 10^{3}\}$ and degree of kernel k ∈ {1, 2} for polynomial, and gamma ∈ $\{10^{-2}, 10^{-1}, 0, 10^{0}, 10^{1}\}$ for RBF kernel.
- For KNN: K ∈ {1, 3, 5, 7}.
- For ANN: number of nodes in hidden layer.

It should be noted that test data were never used for feature selection and training of the classifiers, including the tuning of classifiers.

For measuring the performances of classifiers, we adopted the following metrics: accuracy, sensitivity, specificity, precision and the area under the ROC curve (AUC). Accuracy (AC) is the proportion of the total number of predictions that were correct. Sensitivity (SENS) is defined as the proportion of true positives that are correctly identified by the classifier (e.g., patients with ATS progression), while specificity (SPEC) is a measure of how accurate a test is against false positives (e.g., patients without ATS progression). Precision (PREC) reflects the percentage of patients who actually have the disease among all tested positive. The AUC is an effective and combined measure of sensitivity and specificity that describes ability of model to discriminate between samples that belong to positive and negative classes.

Before the features selection and evaluation of classifiers' results, all data values were normalized to the range [0, 1] by using min-max transformation.

3 Results and Discussion

In this section we present the best results of classifiers' performances obtained by applying the previously described feature selection procedure. The filter algorithms for features selection rank the features from the most important to the less important one according to the statistics they use. We analyzed the results for top ranked k = 5, 10, 20, 30, 40 features. Table 2 clearly shows that MRMR algorithm selection outperformed other algorithms for feature selection. More precisely, the best result was achieved for 20 selected attributes and ANN classifier: **accuracy = 0.891, sensitivity = 0.897, specificity = 0.885, precision = 0.897, AUC = 0.971**.

The most informative selected features are: Diabetes Mellitus, Hypertension, Metabolic Syndrome, Current Symptoms, Aspirin, Oral Antidiabetics, Creatinine, Fibrinogen, INR, MCH, MCV, Total Cholesterol, Triglycerides, CD14(++/+)_RFI, CD14++/CD16-/CCR2+_%, CD163_%, CD16_RFI, CX3CR1_%, CXCR4_%, HLA-DR_%. The study also indicated that among these parameters, CD16_RFI and Hypertension demonstrated the highest correlation with CATS progression. Additionally, a great number of monocytes are selected which in combination with other selected features favors the progression of existing plaque. It is well known that monocytes are directly involved in the process of atherosclerosis formation and a number of other chronic inflammatory conditions. Also, a plenty of studies in literature investigated the monocyte subsets and cardiovascular risk factors. Other selected

Table 2. The best results obtained using classification algorithms in combination with different features selection algorithms.

	Algorithms	ACC	Sens	Spec	Prec	AUC
ReliefF	NB	0.636	0.828	0.423	0.615	0.653
	BN	0.564	0.966	0.115	0.549	0.554
	DT	**0.727**	**0.828**	**0.615**	**0.706**	**0.69**
	RF	0.618	0.552	0.692	0.667	0.64
	SVM	0.618	0.828	0.385	0.6	0.606
	KNN	0.527	0.655	0.385	0.543	0.534
	ANN	0.6	0.793	0.385	0.59	0.612
GainRatio	NB	0.636	0.828	0.423	0.615	0.653
	BN	0.564	0.966	0.115	0.549	0.554
	DT	**0.655**	**0.828**	**0.462**	**0.632**	**0.624**
	RF	0.655	0.759	0.538	0.647	0.664
	SVM	0.618	0.828	0.385	0.6	0.606
	KNN	0.545	0.586	0.5	0.567	0.582
	ANN	0.6	0.655	0.538	0.613	0.553
MRMR	NB	0.745	0.897	0.577	0.703	0.726
	BN	0.564	0.966	0.115	0.549	0.554
	DT	0.727	0.828	0.615	0.706	0.718
	RF	0.673	0.759	0.577	0.667	0.722
	SVM	0.764	0.897	0.615	0.722	0.756
	KNN	0.745	0.724	0.769	0.778	0.827
	ANN	**0.891**	**0.897**	**0.885**	**0.897**	**0.971**

features are already known from the literature as important factors associated to the atherosclerosis disease.

ANN classification algorithm with learning rate set to 0.1 and 6 neurons in hidden layer demonstrated the highest predictive ability. Although our database was relatively limited, results we achieved are encouraging and indicate that ML models can be efficiently developed to predict the prognosis on plaque progression. However, further validation of our results is needed in order to be useful in clinical practice.

4 Conclusion

We presented approach based on medical data analysis aiming to assess the progression of coronary atherosclerosis. Wide series of data obtained from patients' medical records were assembled in order to provide the comprehensive data set able to capture the possible disease manifestations. The results showed that ANN classifier applied on 20 selected features is the most reliable among all tested ML algorithms for prediction whether the patient will be faced with the CATS progression. In the future, we will focus on testing the model on new unseen data and knowledge extraction in order to demonstrate generalization capability and provide explanation of predictions.

Acknowledgments. This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under grant III41007, and by EU H2020 SMARTool project, project ID: 689068. The authors also thank to SES, the world-leading operator of ASTRA satellites, for their kind support in presentation and publishing of this paper.

References

1. Moran, A.E., Forouzanfar, M.H., Roth, G.A., et al.: Temporal trends in ischemic heart disease mortality in 21 world regions, 1980 to 2010: the Global Burden of Disease 2010 study. Circulation **129**(14), 1483–1492 (2014)
2. Boden, W.E., O'Rourke, R.A., Teo, K.K., et al.: Optimal medical therapy with or without PCI for stable coronary disease. N. Engl. J. Med. **356**(15), 1503–1516 (2007)
3. Tonino, P.A., De Bruyne, B., Pijls, N.H., et al.: Fractional flow reserve versus angiography for guiding percutaneous coronary intervention. N. Engl. J. Med. **360**(3), 213–224 (2009)
4. De Bruyne, B., Pijls, N.H., Kalesan, B., et al.: Fractional flow reserve-guided PCI versus medical therapy in stable coronary disease. N. Engl. J. Med. **367**(11), 991–1001 (2012)
5. Papafaklis, M.I., Mavrogiannis, M.C., Stone, P.H.: Identifying the progression of coronary artery disease: prediction of cardiac events. Continuing Cardiol. Educ. **2**, 105–114 (2016)
6. Bolón-Canedo, V., Remeseiro, B., Alonso-Betanzos, A., Campilho, A.: Machine learning for medical applications. In: ESANN proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), pp. 27–29, April 2016. ISBN 978-287587027
7. Thottakkara, P., Ozrazgat-Baslanti, T., Hupf, B.B., Rashidi, P., Pardalos, P., Momcilovic, P., Bihorac, A.: Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. PLOS ONE **11**(5) (2016), https://doi.org/10.1371/journal.pone.0155705
8. EU H2020 project: Simulation Modeling of coronary ARTery disease: a tool for clinical decision support. http://www.smartool.eu/
9. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: Foundations of feature selection. Feature Selection for High-Dimensional Data. AIFTA, pp. 13–28. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21858-8_2
10. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87481-2_21
11. Robnik-Sikonja, M., Kononenko, I.: An adaptation of Relief for attribute estimation in regression. In: Fourteenth International Conference on Machine Learning, pp. 296–304 (1997)
12. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
13. Kononenko, I., Kukar, M.: Machine Learning and Data Mining: Introduction to Principles and Algorithms, Horwood Publ. (2007)
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1) (2009)
15. Kohavi, R.: Wrappers for Performance Enhancement and Oblivious Decision Graphs. Department of Computer Science, Stanford University (1995)