Predictive Modeling of Cardiovascular Disease Progression Using Multimodal Machine Learning

Emmanouela N. Boura, Georgios D. Giannopoulos, Associate Member, IEEE, Evangelos G. Stathopoulos, Senior Member, IEEE, Andreas N. Vlachopoulos, Member, IEEE, Maria C. Zervaki, Giovanni M. T. Canevari, Marco A. Biasotti, Giovanni B. Pappone, Konstantinos A. Manisalis, Dimitrios A. Papanikolaou

Emmanouela N. Boura and Georgios D. Giannopoulos, Department of Electrical Engineering, University of Patras, Greece; Evangelos G. Stathopoulos,

***Abstract*— Nowadays, cardiovascular diseases are very common and are considered as the main causes of morbidity and mortality worldwide. Coronary Artery Disease (CAD), the most typical cardiovascular disease is diagnosed by a variety of medical imaging modalities, which have costs and complications. Therefore, several attempts have been undertaken to early diagnose and predict CAD status and progression through machine learning approaches. The purpose of this study is to present a machine learning technique for the prediction of CAD, using image-based data and clinical data. We investigate the effect of vascular anatomical features of the three coronary arteries on the graduation of CAD. A classification model is built to predict the future status of CAD, including cases of "no CAD" patients, "non-obstructive CAD" patients and "obstructive CAD" patients. The best accuracy was achieved by the implementation of a tree-based classifier, J48 classifier, after a ranking feature selection methodology. The majority of the selected features are the vessel geometry derived features, among the traditional risk factors. The combination of geometrical risk factors with the conventional ones constitutes a novel scheme for the CAD prediction.**

I. Introduction

Atherosclerosis is a disease of coronary arteries and its clinical manifestations account for a significant portion of deaths worldwide. Atherosclerotic disease is considered as a chronic inflammatory disorder of cardiovascular system, which is initiated by the deposition of low-density lipoprotein (LDL) molecules into the arterial wall [1]. Coronary artery disease (CAD) is diagnosed by invasive modalities [coronary angiography (CA), intravascular ultrasound (IVUS), optical coherence tomography (OCT)] and non-invasive modalities, such as the computed tomography coronary angiography (CTCA) and the magnetic resonance imaging (MRI) [2]. Risk and progression prediction of Coronary Artery Disease (CAD) is of high importance in cardiovascular research and aims to identify the patients for statin therapy and choose anticoagulation strategies for atrial fibrillation [3]. Different types of risk factors have been identified to contribute to the progression of CAD, such as the patient's lipid profile (total cholesterol, low-density and high-density lipoprotein cholesterol, triglykerides, lipoprotein), smoking, hypertension, diabetes, family history, obesity [4].

From the point of statistical modelling, the prediction of CAD is a widely studied problem. Several existing studies, such as the Framingham Heart Study [5] and the Systematic COronary Risk Evaluation (SCORE) [6] investigated the ability of hazard regression and logistic regression models to predict CAD progression based on the conventional risk factors. In spite of their good discriminative ability, this type of models contribute to the input association analysis and are not focused on the outcome of the predictors, contrary to machine learning approaches, whose target is to interpret how risk factors affect the outcome [7]. Thus, in the presented study, we aim to predict the occurrence of an outcome and more specifically the progression of CAD based on machine learning approaches.

In the literature, several machine learning based studies have been presented for the prediction of atherosclerosis. More specifically, Exarchos *et al*. [8] implemented typical classification schemes to predict the number of vessels' stenosis, the atherosclerosis progression, as well as a hybric score corresponding to the severity of the disease. The utilized input features were demographics, clinical data, several biochemical variables, monocytes and adhesion molecules. In another recently published study [9], demographics, clinical data, echocardiography data and 54 features of laboratory variables were used to predict the status of CAD, applying a support vector machine (SVM) algorithm with kernel fusion. Another approach, proposed by Weng et al. [10] aimed to predict a fatal or non-fatal cardiovascular event over the ten years, using typical classification algorithms.

Through extensive epidemiological studies, it has been investigated that the coronary arteriosclerosis is non-uniformly distributed in the coronary vasculature, considering not only the different human coronary arteries, but also the different sites of the coronary vessel [11]. The conventional risk factors stated and utilized in the above studies cannot explain this phenomenon, since their influence corresponds to the entire coronary vessel. This investigation has led to the concept of geometric risk factors for the evolution of atherosclerosis, having a significant influence on the mechanical environment of the coronary arterial wall. Existing studies supported that certain vasculature geometrical features could be considered as CAD risk factors [12], elucidating in this manner the mechanism in which biomechanics contributes to atherogenesis. Therefore, the different spatially distributed atherosclerotic lesion is justified in concert with the established systemic risk factors.

Based on the hypothesis of the combination of geometric risk factors and the traditional ones, the current study aims to

present a patient specific machine learning based model able to predict the CAD progression. More specifically, we have assembled a variety of characteristics which have never been previously used for the prediction of the atherosclerosis, including both the vasculature geometrical features based on Computed Tomography Coronary Angiography (CTCA) imaging and the conventional CAD risk factors. Herein, we outline the formulation of the problem, the basic components of the proposed model and finally its predictive capability.

II. Materials and Methods

A. Clinical Scenario

Patient-specific information have been collated from retrospective data recorded during the EVINCI study [13] and are utilized as the baseline information. In the presented study, a dataset of 48 patients was used, who underwent CTCA imaging to diagnose their risk of CAD and evaluate the percentage of coronary vessels' stenosis. This assessment has been reperformed after 5±2 years during SMARTool followup re-evaluation. Except for CTCA imaging, in both time- slices, baseline and follow-up, a variety of data was collected and analyzed, including clinical history and lifestyle of each patient, as well as molecular systemic variables and inflammatory and monocyte markers.

Based on the assembled dataset, we aim to identify the factors that affect the progression of atherosclerosis. More specifically, in this study due to the relatively small number of patients, we focused only on vessel's geometrical features based on the CTCA imaging modality in the baseline step, as well as on the patient's medical history and lifestyle. To this end, we formulate CAD risk prediction problem as a multiclass problem, representing the progression of the CAD as a nonlinear parametric function of set of features $f(x) = C_i, x = [x_1, \dots, x_d]$, $i = 1, \dots k$. In Table I, we present the utilized feature set. The principal utilized classes C_i are namely "No CAD", "Non obstructive CAD" and "Obstructive CAD", which is characterized by degree of stenosis $\geq$ 50% in at least one coronary vessel.

TABLE I. VARIABLES USED IN THE CURRENT ANALYSIS

Category	Features
Geometrical vasculature	Degree of stenosis, Minimal Lumen Area, Minimal Lumen Diameter, Plaque Burden, Calcified plaque Volume, Noncalcified plaque Volume
Risk factors	Family History of CAD, Hypertension, Diabetes, Dyslipidaemia, Smoking, Obesity, Metabolic Syndrome, Past smokers

B. Techniques and Algorithms

The steps followed for the construction of the machine learning models are illustrated in Fig.1

1) CTCA image acquisition and analysis

The geometrical vasculature features utilized in the proposed machine learning based model are estimated based on an already published methodology for the threedimensional (3D) reconstruction of coronary arteries [14]. Briefly, the implemented methodology is summarized in different steps: the preprocessing procedure using Frangi Vesselness filter, the vessel centerline extraction using minimal cost path approach, segmentation of the inner wall, outer wall and calcified plaque (CP), implementing active contour models using prior shapes on two-dimensional (2D) CTCA axial images, segmentation of noncalcified plaques (NCP) applying a dynamic threshold segmentation technique and finally the 3D models construction based on the Marching cubes approach. This methodology is integrated in a dedicated software tool, which generates the 3D models of the inner wall, outer wall, CP and NCP for each coronary artery segment and provides automatically the corresponding geometrical features. *2) Feature Selection*

In the proposed study, we employed feature selection techniques, aiming to reduce the dimension of input features and identify the redundant features. More specifically, we implemented the gain ratioG algorithm, the principal components analysis and the attribute evaluation technique. The progressive improvement of the sensitivity and specificity ratio is achieved through a proper customization of the input features, accompanied by a forward selection procedure. *3) Classification*

After feature selection algorithms implementation, we examined four different machine learning algorithms; a parametric model [artificial neural network (ANN)], a nonparametric kernel-based model [support vector machine (SVM)], an ensemble model [random forest (RF)] and J48 [15-17].

4) Evaluation

In terms of evaluation, we apply the ten-fold cross validation, which splits the initial dataset into ten subsets, whereby the nine subsets are used for training and the remaining one subset is used for testing. Moreover, it should be noted that the feature selection techniques are repeated in every ten-fold repetition, ensuring that the feature selection procedure is based exclusively on the training dataset. In addition to this, it should be highlighted that our initial dataset is considered as a balance dataset, including almost same portions of patients of the three classes.

The performance of each classification scheme is quantified using as evaluation metrics the sensitivity and the specificity of each class, which denote the model's ability to quantify the positive and the negative results, respectively.

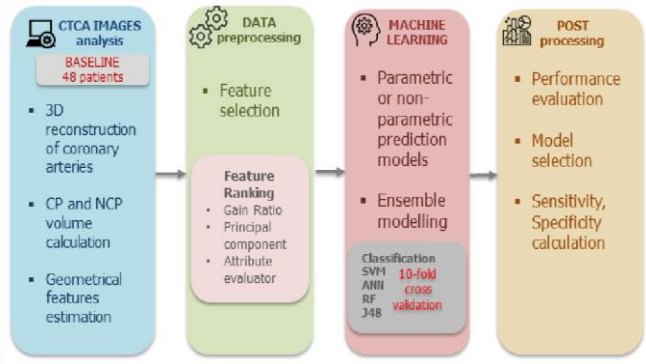

Figure 1. Flowchart of the prediction model

III. Results

In the Table II, we present the obtained sensitivity and specificity for each of the three different classes, after implementing the classification schemes. As far as the class of non-obstructive CAD is concerned, its sensitivity is observed lower than these of the other classes.

TABLE II. CLASSIFICATION PERFORMANCE

		Class No CAD		Class Nonobstructive CAD		Class Obstructive CAD	
		Se.	*Sp.*	*Se.*	*Sp.*	*Se.*	*Sp.*
Case 1	SVM	0.67	0.73	0.42	0.81	0.44	0.85
	ANN	0.6	0.79	0.49	0.65	0.42	0.78
	RF	0.69	0.85	0.48	0.71	0.47	0.75
	J48	0.67	0.73	0.42	0.81	0.63	0.75
	SVM	0.54	0.82	0.53	0.59	0.44	0.85
	ANN	0.8	0.79	0.48	0.88	0.57	0.75
	RF	0.6	0.82	0.53	0.78	0.63	0.79
	J48	0.6	0.82	0.58	0.81	0.59	0.82
	SVM	0.74	0.85	0.48	0.71	0.46	0.75
	ANN	1	0.5	0.42	0.91	0.32	0.97
	RF	0.74	0.76	0.36	0.81	0.63	0.79
	J48	0.87	0.79	0.36	0.91	0.82	0.82

Se.: Sensitivity, Sp.: Specificity

It is noted that in case we implement the J48 classification algorithm, after the class attribute evaluation feature selection technique, we achieve a sensitivity of 0.87 and 0.82 for the healthy class, the No CAD class and the unhealthy class, the Obstructive class, respectively. The confusion matrix of the aforementioned approach is showed in Table III. In addition to this the ranking of the selected features, maintained by this classification scheme is illustrated in Table IV and the best performance of the proposed model is achieved by maintaining the eight first ranking features.

TABLE III. CONFUSION MATRIX OF CASE 3-J48

		Prediction		
		Class No CAD	**Class Nonobstructive CAD**	**Class Obstructive CAD**
Annotation	**No CAD**	13	1	1
	Nonobstructive CAD	6	6	5
	Obstructive CAD	1	2	13

TABLE IV. FEATURES RANKING SELECTION BY CASE 3-J48

Class Attribute evaluator	**Features**
	Current symptoms {0,1}
	volume of the most significant calfified plaque (mm^3)
	volume of the most significant non-calcified plaque (mm^3)
	existence of non-calcified plaques {0,1}
	Minimal lumen area (mm^2)
	Past smokers {0,1}
	existence of calcified plaques {0,1}
	Minimal lumen area (mm)

IV. Discussion

In this study, multiple models were performed to predict the progression of CAD, taking into account non-invasive image-based features and traditional atherosclerosis risk factors, such as the medical history and lifestyle. The combination of the input features as well as the problem formulation as a multiclass problem, integrating the graduation of atherosclerosis, constitute a crucial novelty of the presented study. This study is considered as a preliminary approach and it is important to further validate our results in larger datasets.

After the implementation of different classification schemes, we conclude that the tree-based algorithms and more specifically C4.5 (J48) algorithm combined by a ranking feature selection method, outperforms the other classification models [17]. In general, tree-based algorithms constitutes a typical type of machine-learning approaches and are able to handle the non-linearities, the heterogeneous data, and many predictors, by searching through the input variables to find this one, which seperates the outcome into two groups [7].

In this point, it should be highlighted that the ability to foresee a future health condition of the most patients is a

significant aspect of the proposed model. The sensitivity for the first class is 0.71±0.13, based on the different implementation, whereas for the most accurate model is 0.87. In this manner, the patient may avoid undergo a multiple CTCA imaging in the follow-up time slice step, considering the risks related to radiation from medical imaging. In addition to this, another aspect of the proposed study worth mentioning is the model's ability to accurately identify the non-healthy patients, those of the third class. In the last case of J48 algorithm, the sensitivity is 0.82. An accurate prediction of such a status contributes significantly to the clinical patient management, allowing the timely diagnosis of CAD and the safe selection of treatment.

In addition to this, it is notable that the input geometric features derived from a non-invasive imaging modality, the CTCA. In this manner, the double inference of the presented approach is the ability of CTCA to visualize the inner, outer wall and atherosclerotic plaques accurately, highly contributing to the cardiovascular research and clinical area as well as the high performance of the utilized algorithm, concerning the construction of the coronary arteries 3D models.

In the first stage, our study could be considered as a promising approach able to stratify the risk of CAD. Its

deployment and validation is ongoing by the integration of new features, concerning molecular systemic variables, inflammatory and monocyte markers, the lipid profile, exposome as well as mRNA sequencing. Therefore, a more detailed input space and a larger dataset of patients ensure a more effective multimodal prediction scheme and potentially a refined formulation of the classification problem.

V. CONCLUSIONS

Undoubtedly, as we enter the age of precision medicine, risk assessment and prediction models are considered more notable. In our study, the principal aim is to generate a model that accurately predicts the status of CAD, focusing more to the patients of the "No CAD" and "Obstructive CAD" class. Furthermore, it is worth mentioned that this study pose another aspect of the risks factors the clinicians have to think about, which are the image-based features. The capability of machine learning models, combined with a detailed input set of parameters and a large balanced dataset of patients may provide novel and promising CAD stratification approaches, contributing to the clinical and research cardiovascular area.

REFERENCES

[1] A. J. Lusis, "Atherosclerosis," *Nature,* vol. 407, p. 233, 09/14/online 2000.

[2] J. M. Tarkin, M. R. Dweck, N. R. Evans, R. A. Takx, A. J. Brown, A. Tawakol*, et al.*, "Imaging atherosclerosis," *Circulation research,* vol. 118, pp. 750-769, 2016.

[3] N. J. Stone, J. Robinson, A. H. Lichtenstein, C. N. B. Merz, C. B. Blum, R. H. Eckel*, et al.*, "2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults," *A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines,* 2013.

[4] J. E. Roeters van Lennep, H. T. Westerveld, D. W. Erkelens, and E. E. van der Wall, "Risk factors for coronary heart disease: implications of gender," *Cardiovascular Research,* vol. 53, pp. 538-549, 2002.

[5] R. B. D'Agostino, R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro*, et al.*, "General cardiovascular risk profile for use in primary care: the Framingham Heart Study," *Circulation,* vol. 117, pp. 743-753, 2008.

[6] R. Conroy, K. Pyörälä, A. e. Fitzgerald, S. Sans, A. Menotti, G. De Backer*, et al.*, "Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project," *European heart journal,* vol. 24, pp. 987-1003, 2003.

[7] B. A. Goldstein, A. M. Navar, and R. E. Carter, "Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges," *European Heart Journal,* vol. 38, pp. 1805-1814, 2017.

[8] K. P. Exarchos, C. Carpegianni, G. Rigas, T. P. Exarchos, F. Vozzi, A. Sakellarios*, et al.*, "A Multiscale Approach for Modeling Atherosclerosis Progression," *IEEE Journal of Biomedical and Health Informatics,* vol. 19, pp. 709-719, 2015.

[9] R. Alizadehsani, M. H. Zangooei, M. J. Hosseini, J. Habibi, A. Khosravi, M. Roshanzamir*, et al.*, "Coronary artery disease detection using computational intelligence methods," *KnowledgeBased Systems,* vol. 109, pp. 187-197, 2016/10/01/ 2016.

[10] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PloS one,* vol. 12, p. e0174944, 2017.

[11] H. Zhu, Z. Ding, R. N. Piana, T. R. Gehrig, and M. H. Friedman, "Cataloguing the geometry of the human coronary arteries: a potential tool for predicting risk of coronary artery disease," *International journal of cardiology,* vol. 135, pp. 43-52, 07/01 2009.

[12] M. H. Friedman, O. J. Deters, F. F. Mark, C. Brent Bargeron, and G. M. Hutchins, "Arterial geometry affects hemodynamics: A potential risk factor for atherosclerosis," *Atherosclerosis,* vol. 46, pp. 225-231, 1983/02/01/ 1983.

[13] R. Liga, J. Vontobel, D. Rovai, M. Marinelli, C. Caselli, M. Pietila*, et al.*, "Multicentre multi-device hybrid imaging study of coronary artery disease: results from the EValuation of INtegrated Cardiac Imaging for the Detection and Characterization of Ischaemic Heart Disease (EVINCI) hybrid imaging population," *European Heart Journal - Cardiovascular Imaging,* vol. 17, pp. 951-960, 2016.

[14] V. I. Kigka, G. Rigas, A. Sakellarios, P. Siogkas, I. O. Andrikos, T. P. Exarchos*, et al.*, "3D reconstruction of coronary arteries and atherosclerotic plaques based on computed tomography angiography images," *Biomedical Signal Processing and Control,* vol. 40, pp. 286-294, 2018/02/01/ 2018.

[15] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*: Pearson Addison Wesley, 2006.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explor. Newsl.,* vol. 11, pp. 10-18, 2009.

[17] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Machine Learning,* vol. 16, pp. 235-240, September 01 1994.