2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Machine Learning Model for the Identification of High risk Carotid Atherosclerotic Plaques

Vassiliki I. Kigka, Antonis I. Sakellarios, Michalis D. Mantzaris, Vassilis D. Tsakanikas, Vassiliki T. Potsika, Domenico Palombo, Fabrizio Montecucco and Dimitrios I. Fotiadis, Fellow, IEEE

Abstract— **Carotid artery disease is an inflammatory condition involving the deposition and accumulation of lipid species and leucocytes from blood into the arterial wall, which causes the narrowing of the carotid arteries on either side of the neck. Different imaging modalities can by implemented to determine the presence and the location of carotid artery stenosis, such as carotid ultrasound, computed tomography angiography (CTA), magnetic resonance angiography (MRA), or cerebral angiography. However, except of the presence and the degree of stenosis of the carotid arteries, the vulnerability of the carotid atherosclerotic plaques constitutes a significant factor for the progression of the disease and the presence of disease symptoms. In this study, our aim is to develop and present a machine learning model for the identification of high risk plaques using non imaging based features and non-invasive imaging based features. Firstly, we implemented statistical analysis to identify the most statistical significant features according to the defined output, and subsequently, we implemented different feature selection techniques and classification schemes for the development of our machine learning model. The overall methodology has been trained and tested using 208 cases of 107 cases of low risk plaques and 101 cases of high risk plaques. The highest accuracy of 0.76 was achieved using the relief feature selection technique and the support vector machine classification scheme. The innovative aspect of the proposed machine learning model is both the different categories of the utilized input features and the definition of the problem to be solved.**

I. Introduction

Carotid artery disease (CAD), the build-up of atherosclerotic plaques in carotid bifurcations, is a highly prevalent and devastating disease of our times, with enormous socioeconomic burden. It constitutes the primary cause of cerebrovascular events and ischaemic stroke, and accounts for up to 30% of all strokes. In the European Union, this translates to more than 150.000 deaths annually and over €12 billion per year in direct and indirect costs. Despite its high prevalence and enormous socioeconomic burden, carotid artery disease is still treated with criteria established in the 90s that do not take into account any of the advances of the molecular evolution we have witnessed since. Thus, at an era where targeted biological therapies, assisted by genomics and pharmacogenomics, already constitute the standard method of care for most other prevalent diseases including autoimmune and allergic diseases, and cancer, carotid artery disease treatment is still managed according to the level of carotid stenosis [1].

Carotid artery disease is characterized by the development of an atherosclerotic plaque inside the artery wall that reduces blood flow to the brain and increases the risk for transient ischemic attack (TIA) or stroke. The carotid atherosclerotic plaque can be asymptomatic, or it can be a source of embolization-symptomatic as emboli can break off from the plaque and block arteries in the brain, causing a transient ischemic attack or more permanent damage manifested as stroke. In compliance with the current guidelines, the treatment of carotid artery disease depends on the degree of the carotid artery stenosis. More specifically, if the stenosis is mild to moderate, lifestyle changes and medication is proposed to slow the progression of atherosclerosis, whereas the most common treatment of severe carotid artery stenosis is either the carotid endarterectomy or the carotid angioplasty and stenting [2].

Different attempts have been undertaken to identify the most significant biomarkers, which are correlated with CAD and different data driven models have been proposed in the literature for the risk stratification of CAD. These models are based either on statistical analysis or on machine learning models. More specifically, Jamthikar *et al.* [3] proposed a machine learning (ML) based algorithm for the development of stroke risk stratification tool, and concluded that ML-based integrated model with the event-equivalent gold standard as artery degree of stenosis is powerful and offers low cost and high performance for stroke risk assessment. In a same attempt, Verde *et al.* [4] investigated and compared the performance of conventional machine learning classifiers, capable of identifying the presence of CAD In another study, Greco *et al.* [5] taking advantage of statistical based models, developed a model for the prediction of individuals of high risk for carotid degree of stenosis. In another study proposed by de Weerd *et al*. [6], a CAD prediction model was developed for the identification of individuals with a carotid artery stenosis

* This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 755320, as part of the TAXINOMISIS project.

D. I. Fotiadis is with the Unit of Medical Technology and IntelligentInformation Systems, Department of Materials Science and Engineering, University of Ioannina, and with the Department of Biomedical Research, Institute of Molecular Biology and Biotechnology–FORTH, University Campus of Ioannina, 45110 Ioannina, Greece (phone: +30 26510 09006;email: fotiadis@uoi.gr).

Vassiliki I. Kigka, Antonis I. Sakellarios, Michalis D. Mantzaris, Vassilis D. Tsakanikas, Vassiliki T. Potsika are with the Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering University of Ioannina, Ioannina, Greece

Domenico Palombo is with Division of Vascular and Endovascular Surgery, IRCCS Ospedale Policlinico San Martino Genoa – Italian Cardiovascular Network, Genoa, Italy

Fabrizio Montecucco is with Clinic of Internal Medicine I, IRCCS Ospedale Policlinico San Martino Genoa – Italian Cardiovascular Network, Genoa Italy.

>50% and >70% and to identify the most efficient predictors of the carotid artery stenosis >50 % and 70%. Araki *et al.* [7] proposed an innovative risk stratification model for the classification of high and risk carotid plaques or the classification between symptomatic and asymptomatic plaques. The overall model was based on the grayscale morphology of the ultrasound carotid wall. Zhu *et al.* [8] in a study published in 2019 attempted to investigate whether quantitative imaging based features derived by carotid computed tomography angiography (CTA) can predict the 10 year cardiovascular events. In a same attempt, Tecellioglu *et al.* [9] aims to compare the histomorphological characteristics of carotid plaques and glycosylated hemoglobin (HbA1c), which are risk factors for ischemic stroke, in patients who underwent carotid endarterectomy for carotid artery stenosis. On the other hand, Hao-wen Li *et al*. [10] conducted an observational study to distinguish the high risk plaques (vulnerable plaques) and the low risk (stable) carotid artery plaques, which were classified using high-resolution magnetic resonance imaging (MRI) modality.

The aim of this manuscript is to present a machine learning based model, which is able to identify the vulnerable-high risk carotid atherosclerotic plaques. The vulnerable plaques have been defined based on the histology based features, as clinical doctors proposed. The steps of the overall methodology are the following: i) the data curation, ii) the problem definition, iii) the feature selection and classification scheme implementation and finally iv) the model evaluation and performance metrics computation.

II. MATERIALS & METHODS

A. Dataset Description

The utilized dataset includes the following 5 distinct data views, utilized as input for the carotid artery disease risk stratification model:

i) View 1 – Clinical Data

ii) View 2 – Risk Factors

iii) View 3 – Laboratory Data

iv) View 4 – Serum Markers

v) View 5 –Imaging.

The utilized dataset includes totally 208 cases and the input features are described in detail, in Table 1, below.

B. Statistical Analysis Implementation

In this section, we performed typical statistical analysis of the utilized dataset, which aimed to check its quality and identify the most significant statistically features. Appropriate parametric and nonparametric univariate two-sample and k-sample statistical tests, based on the type (continuous numeric or categorical) and the per class distribution (Gaussian or non-Gaussian) of the dependent variables were implemented as the first analysis of our dataset.

TABLE I. DESCRIPTION OF THE FEATURE DATASET UTILIZED AS INPUT.

View #	Category	Features
View 1	Clinical Data	Age, Gender
View 2	Risk Factors	hypertension, Diabetes II, Dyslipidemia, Chronic CAD, Smoking, waist circumference
View 3	Laboratory Data	Total WBC, Neutrophils, Monocytes, Lymphocytes, Platelets, RBC, D-dimer, Fibrinogenemia, Fasting Insulinemia, Fasting C peptidemia, Lpa, total cholesterol, high density lipoprotein, low density lipoprotein, Triglyceridemia, Fasting Glycemia
View 4	Serum Markers	C-C motif chemokine ligand 2, Osteoprotegerin, High-sensitivity C-reactive protein, P-selectin, Intercellular Adhesion Molecule 1, vascular cell adhesion molecule 1, Adiponectin, Serum, E-selectin, Resistin, Leptin, Interleukin 1 alpha, Tumour Necrosis Factor alpha, C-C Motif Chemokine Ligand 5, C-C Motif Chemokine Ligand 4, C-C Motif Chemokine Ligand 3, CD40L, Interleukin 6, Soluble Interleukin-6 receptor, Insulin-like growth factor 1, L-selectin, Matrix metallopeptidase 9, pro Matrix metallopeptidase 9, gelatinolytic activity, Matrix metallopeptidase 8, Tissue Inhibitor of Metalloproteinase 1,2,3,4, Matrix Metalloproteinase-9/ Tissue inhibitor of metalloproteinase-1, FAP, Osteopontin, Receptor activator of nuclear factor kappa-B ligand, Myeloperoxidase, Neutrophil elastase, Total Vitamin D, Proprotein convertase subtilisin/kexin type 9
View 5	Imaging	% Stenosis ipsilateral, Plaque size

C Methodology

i) Data Curation

Data cleaning procedure, also known as data curation procedure, is considered as a key aspect, prior to the development of any data analytics services. In this step, the data curation framework proposed by Pezoulas *et al*. [11] is implemented in each utilized dataset and the framework

outputs two different documents: (i) the data quality assessment report, and (ii) the curated dataset.

ii) Problem Definition

The carotid artery disease risk stratification problem has been formulated as a multivariate 2 class classification problem based on plaque histology related feature, as it is shown in Figure 1 More specifically, clinical doctors have identified significant plaque related features, which are directly associated with the vulnerability of the atherosclerotic plaques and proposed a cut off value for these plaque histology related features, in order to binarize the output of our proposed model. Clinical doctors have statistically analyzed the dataset to identify accurate cut- off values for the classification of non-vulnerable-low risk plaques (Class 0) and vulnerable-high risk plaques (Class 1).

This overall predictive supervised learning approach aims to learn a mapping from input features x to output Y, given a labeled set of input output pairs $D = \{(x_i, y_i)\}_{i=1}^{N}$, where D is the training set and N is the number of training examples [12]. Each sample (x_i, y_i) associates the input features with the carotid artery disease risk prediction Y, where $Y \in \{C_1, C_2\}$, is estimated by a non-linear parameterized function (f) of input features $x \in R^d$, $x = [x_1, x_2, \dots, x_d]$. The goal of this supervised classification problem is to obtain an approximation $F(x)$ of the function $F^*(x)$ mapping the input x to output Y.

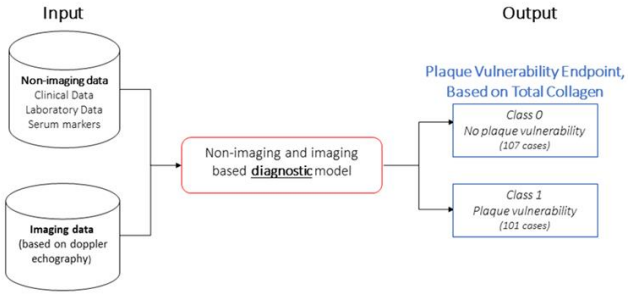

Figure 1. Problem Definition of the proposed machine learning based model

iii) Feature selection and classification scheme implementation

As far as the classification procedure is concerned, we feed our input data directly to classification algorithms. More specifically, we examine the performance of five popular classification algorithms to discriminate the high risk group patients. The implemented algorithms are J48 algorithm, Random Forests (RF), Naive Bayes (NB), Support Vector Machine (SVM) and Artificial Neural Networks (ANN). In combination with the implementation of these classification schemes, we integrate in our methodology the implementation of feature selection techniques, to discard redundant features and maintain the more informative input subset, achieving in this manner high classifier performance. For this purpose, eight different feature selection techniques are implemented, the Correlation based Feature Selection (case 1), the Class Attribute Evaluation (case 2), the Correlation Attribute Evaluation (Case 3), the Gain Ratio (case 4), the InfoGain Ratio (case 5), the OneR (case 6), the principal components analysis (case 7) and the relief technique (case 8).

iv) Model Evaluation and Performance Metrics Computation

For evaluation purposes, the ten-fold cross validation is applied, which splits the initial subset into ten subsets, whereby the nine subsets are used for training and the remaining one subset is used for testing and after a full rotation, the results over the 10 testing sets are averaged in order to procure the overall performance of the algorithm. 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. The evaluation metrics used to compare the employed classification schemes is the total accuracy of the model which denotes the quotient of the correctly classified instances among the total instances..

III. RESULTS

I) Statistical Analysis results

The high risk plaques based on the composition of Total Collagen is associated with higher mean values of Age (years), D-dimer, P-selectin, Soluble Interleukin-6 receptor (sIL-6R), L-selectin, Tissue Inhibitor of Metalloproteinase 2 (TIMP-2), Matrix Metalloproteinase-9/ Tissue inhibitor of metalloproteinase-1 (MMP-9/TIMP-1) and with lower mean values of Sistolic arterial pressure, Fasting Insulinemia, C-C Motif Chemokine Ligand 3 (CCL3) and Tissue Inhibitor of Metalloproteinase 3 (TIMP-3).

II) Classification schemes results

In Table II, we present the results obtained implementing different feature selection techniques and classification schemes. We observe that the highest accuracy (Acc.) and area under the curve (AUC) was 0.76 and 0.73, respectively and were achieved implementing the relief feature selection technique (Case 8) and the support vector machine (SVM) classification scheme.

IV. DISCUSSION

Most of the existing studies in the literature focused on the carotid artery disease stratification based on the carotid artery degree of stenosis. However, in our development of carotid artery risk stratification model, we focused on the development of a machine learning model based on the presence of high risk atherosclerotic plaques, directly associated with the carotid artery disease symptoms and cardiovascular events.

In addition to this, our proposed machine learning based model is primarily based on clinical data, biomarkers, serum markers and imaging data, derived by non-invasive imaging modality, the carotid ultrasound.

Moreover, except of the development of the machine learning model for the identification of high risk carotid atherosclerotic plaques, through this study, we have investigated new biomarkers, directly associated with the presence of high risk plaques. Some of them are matrix degrading biomarkers (TIMP-2, MMP-9/TIMP-1, TIMP-3), inflammatory biomarkers (CCL3, sIL-6R), whereas as others are endothelial and call adhesion biomarkers (P-selectin). Thus, in this direction new attempts could be undertaken to explore innovative biomarkers not only associated with the presence of the CAD, as the percentage of carotid stenosis define, but also to define new biomarkers directly associated with the mechanism of atherosclerosis pathophysiology and the non-stable vulnerable atherosclerotic plaques.

TABLE II. RESULTS OBTAINED IMPLEMENTING DIFFERENT FEATURE SELECTION TECHNIQUES AND CLASSIFICATION SCHEMES

	Case 1		Case 2		Case 3		Case 4	
	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
J48	0.69	0.74	0.63	0.62	0.62	0.6	0.62	0.62
RF	0.65	0.72	0.69	0.75	0.68	0.74	0.66	0.75
NB	0.66	0.72	0.58	0.62	0.58	0.62	0.58	0.62
SVM	0.64	0.63	0.71	0.71	0.71	0.71	0.7	0.7
ANN	0.7	0.77	0.68	0.74	0.65	0.71	0.68	0.72
	Case 5		Case 6		Case 7		Case 8	
	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
J48	0.62	0.62	0.61	0.6	0.59	0.61	0.62	0.61
RF	0.67	0.76	0.69	0.75	0.63	0.7	0.68	0.74
NB	0.58	0.62	0.58	0.62	0.63	0.63	0.58	0.62
SVM	0.7	0.7	0.7	0.7	0.66	0.66	0.76	0.73
ANN	0.68	0.72	0.68	0.74	0.66	0.69	0.69	0.72

RF: Random Forests, NB: Naive Bayes, SVM: Support Vector Machine, ANN: Artificial Neural Networks, Case 1: Correlation based Feature Selection, Case 2: Class Attribute Evaluation, Case 3: Correlation Attribute Evaluation, Case 4: Gain Ratio, Case 5: InfoGain Ratio, Case 6: OneR, Case 7: principal components analysis, Case 8: relief technique.

V. CONCLUSION

Through this study, an overall methodology for the identification of high risk atherosclerotic plaques has been presented. The proposed methodology takes advantage of machine learning models and is not dedicated only to identify the relation between the input features and the defined output. The clinical significance of this model is very high both for the patient management and the selection of medical therapy.

REFERENCES

[1] A. Abbas, P. Aukrust, D. Russell, K. Krohg-Sørensen, T. Almås, D. Bundgaard, *et al.*, "Matrix metalloproteinase 7 is associated with symptomatic lesions and adverse events in patients with carotid atherosclerosis," *PloS one,* vol. 9, p. e84935, 2014.

[2] A. L. Abbott, K. I. Paraskevas, S. K. Kakkos, J. Golledge, H.-H. Eckstein, L. J. Diaz-Sandoval, *et al.*, "Systematic review of guidelines for the management of asymptomatic and symptomatic carotid stenosis," *Stroke,* vol. 46, pp. 3288-3301, 2015.

[3] A. Jamthikar, D. Gupta, N. N. Khanna, L. Saba, T. Araki, K. Viskovic, *et al.*, "A low-cost machine learning-based cardiovascular/stroke risk assessment system: integration of conventional factors with image phenotypes," *Cardiovascular Diagnosis and Therapy,* vol. 9, pp. 420-430, 2019.

[4] L. Verde and G. De Pietro, "A Machine Learning Approach for Carotid Diseases using Heart Rate Variability Features," in *HEALTHINF*, 2018, pp. 658-664.

[5] G. Greco, N. N. Egorova, A. J. Moskowitz, A. C. Gelijns, K. C. Kent, A. J. Manganaro, *et al.*, "A model for predicting the risk of carotid artery disease," *Annals of surgery,* vol. 257, pp. 1168-1173, 2013.

[6] M. de Weerd, J. P. Greving, B. Hedblad, M. W. Lorenz, E. B. Mathiesen, D. H. O'Leary, *et al.*, "Prediction of asymptomatic carotid artery stenosis in the general population: identification of high-risk groups," *Stroke,* vol. 45, pp. 2366-2371, 2014.

[7] T. Araki, P. K. Jain, H. S. Suri, N. D. Londhe, N. Ikeda, A. El-Baz, *et al.*, "Stroke risk stratification and its validation using ultrasonic echolucent carotid wall plaque morphology: a machine learning paradigm," *Computers in biology and medicine,* vol. 80, pp. 77-96, 2017.

[8] G. Zhu, Y. Li, V. Ding, B. Jiang, R. L. Ball, F. Rodriguez, *et al.*, "Semiautomated Characterization of Carotid Artery Plaque Features From Computed Tomography Angiography to Predict Atherosclerotic Cardiovascular Disease Risk Score," *Journal of computer assisted tomography,* vol. 43, pp. 452-459, 2019.

[9] M. Tecellioglu, S. Alan, S. Kamisli, F. Tecellioglu, O. Kamisli, and C. Ozcan, "Hemoglobin A1c-related histologic characteristics of symptomatic carotid plaques," *Nigerian journal of clinical practice,* vol. 22, p. 393, 2019.

[10] H.-w. Li, M. Shen, P.-y. Gao, Z.-r. Li, J.-l. Cao, W.-l. Zhang, *et al.*, "Association between ADAMTS7 polymorphism and carotid artery plaque vulnerability," *Medicine,* vol. 98, p. e17438, 2019.

[11] V. C. Pezoulas, K. D. Kourou, F. Kalatzis, T. P. Exarchos, A. Venetsanopoulou, E. Zampeli, *et al.*, "Medical data quality assessment: On the development of an automated framework for medical data curation," *Computers in biology and medicine,* vol. 107, pp. 270-283, 2019.

[12] C. Robert, "Machine learning, a probabilistic perspective," ed: Taylor & Francis, 2014.