A Multimodal Machine Learning Approach to Omics-based Risk Stratification in Coronary Artery Disease

Eleni I. Georga\(^1\), Nikolaos S. Tachos\(^2\), Antonis I. Sakellarios\(^3\), Gualtiero Pelosi\(^3\), Silvia Rocchiccioli\(^3\), Oberdan Parodi\(^3\), Lampros K. Michalis\(^4\), and Dimitrios I. Fotiadis\(^1,2\)

\(^1\) Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ioannina, Greece
\(^2\) Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology - Hellas (FORTH), Ioannina, Greece
\(^3\) Institute of Clinical Physiology, National Research Council, Pisa, Italy
\(^4\) Department of Cardiology, Medical School, University of Ioannina, Ioannina, GR 45110, Greece

Abstract. This study aims at developing a personalized model for coronary artery disease (CAD) risk stratification based on machine learning modelling of non-imaging data, i.e. clinical, molecular, cellular, inflammatory and omics data. A multimodal architectural approach is proposed whose generalization capability, with respect to CAD stratification, is currently evaluated. Different data fusion techniques are investigated, ranging from early to late integration methods, aiming at designing a predictive model capable of representing genotype-phenotype interactions pertaining to CAD development. An initial evaluation of the discriminative capacity of the feature space with respect to a binary classification problem (No CAD, CAD), although not complete, shows that: (i) kernel-based classification provides more accurate results as compared with neural network-based and decision tree-based modelling, and (ii) appropriate input refinement by feature ranking has the potential to increase the sensitivity of the model.

Keywords: Coronary Artery Disease, Patient Risk Stratification, Multi-modal machine learning.

1 Introduction

Coronary artery disease (CAD) is a multi-factorial disease characterized by the accumulation of lipids into the arterial wall and the subsequent inflammatory response \([1, 2]\). The phenotype of disease progression is affected by several factors, including clinical risk factors (e.g. gender, smoking, hyperlipidaemia, hypertension, diabetes) as well as molecular, biohumoral and biomechanical factors (e.g. low endothelial shear stress). CAD diagnosis is validated through invasive coronary angiography (CA); however, different invasive [e.g. intravascular ultrasound (IVUS), optical coherence tomography (OCT)] and non-invasive imaging modalities [e.g. computed tomography angiography
[CTA], magnetic resonance imaging (MRI)] are nowadays available to visualize the vessel wall, quantify the plaque burden and characterize the type of the atherosclerotic plaque.

Predicting the risk of CAD constitutes a widely-studied problem from the perspective of statistical modelling. The majority of existing risk models, such as the Framing-ham risk score (FRS) [3], the Systematic CORonary Risk Evaluation (SCORE) [4] and the QRISK [5], postulate a Cox proportional hazard regression or logistic regression model of relatively few traditional predictors of the disease, focusing on CAD or cardiovascular disease (CVD). In spite of the reported good discrimination ability of parametric linear regression models, a recent systematic review demonstrated the paucity of external validation and head-to-head comparisons, the poor reporting of their technical characteristics as well as the variability in outcome variables, predictors and prediction horizons, which limits their applicability in evidence-based decision making in healthcare [6]. Precision medicine suggests individualized dynamic predictive modelling approaches not being hypotheses-driven [7-9]. Moreover, the increasing availability of electronic health records (EHRs), personal health records (PHRs) and omics big data give rise to multiscale multi-parametric predictive big data analytics in personalized medicine in cardiovascular research and clinical practice [10-12].

The purpose of this study is to design and develop a machine learning-based model effectively integrating multiple categories of biological data towards precise risk stratification in coronary artery disease. Herein, we outline the formulation of the problem, present the main components of the model architecture, and investigate the predictive power of the currently available feature set.

## 2 CAD Risk Stratification Methodology

### 2.1 Problem Formulation

CAD risk stratification is formulated as a multiclass classification problem, representing the severity of the disease as a nonlinear parametric function of a confined set of features \( f(x) = C_i, x = [x_1, \ldots, x_k], i = 1, \ldots, k \). The utilized feature set is provided in Table 1. Three dominant classes \( C_i, i = 1, \ldots, k \) have been defined, namely “No CAD”, “Non Obstructive CAD”, and “Obstructive CAD”, with a ≥50% diameter stenosis in at least one main coronary artery vessel, as assessed by computed tomography coronary angiography (CTCA), characterizing patients with obstructive CAD.

| Category       | Features                                         |
|----------------|--------------------------------------------------|
| Demographics   | Age, Gender                                      |
| Risk Factors   | Family History of CAD, Hypertension, Diabetes, Dyslipidaemia, Smoking, Obesity, Metabolic Syndrome |
### Molecular Systemic Variables

- Alanine Aminotransferase, Alkaline Phosphatase, Aspartate Aminotransferase, Creatinine, Gamma-Glutamyl Transferase, Glucose, HDL, High-Sensitivity C-Reactive Protein, Interleukin-6, LDL, Leptin, Total Cholesterol, Triglycerides, Uric Acid

### Symptoms

- Typical Angina, Atypical Angina, Non Angina Chest Pain, Other Symptoms, No Symptoms

### Exposome

- Alcohol Consumption, Vegetable Consumption, Physical Activity, Home Environment, Exposition to Pollutants

### Inflammatory Markers

- ICAM1, VCAM1

### Monocyte Markers

- CCR2_val1, CCR2_val2, CCR5_val1, CCR5_val2, CD11b_val1, CD11b_val2, CD14(+/-)_val1, CD14(+/-)_val2, CD14+/CD16+/CCR2+_val1, CD14+/CD16-/CCR2+_val1, CD14+/CD16++/CCR2-_val1, CD163_val1, CD163_val2, CD16_val1, CD16_val2, CD18_val1, CD18_val2, CX3CR1_val1, CX3CR1_val2, CXCR4_val1, CXCR4_val2, HLA-DR_val1, HLA-DR_val2, MONOCYTE COUNT_val2

### Omics Data

- Lipid Profile, MRNA Sequencing, Exome Sequencing

## 2.2 Multimodal Machine-Learning CAD Stratification Model

A multimodal architecture was specified relying on two processing layers which are defined according to late or intermediate data integration strategies [13]. First, the following feature classes (or views) were defined: (View 1) demographics, (View 2) clinical data, risk factors, symptoms, (View 3) molecular variables (i.e. biohumoral, inflammatory markers and lipids profile), (View 4) gene expression data, (View 5) exposome, and (View 6) monocytes. As it is shown in Fig.1, late data integration consists in the construction of: (i) an ensemble of decision tree-based prediction models (i.e. random forests, boosted decision trees) for each data view, whose individual decisions are effectively merged using simple mechanisms (e.g. weighted voting), or (ii) a multimodal deep neural network comprising of appropriate deep learning subnetworks for each separate data view and, unifying their output into higher network layers.

![Fig. 1. Integrative CAD risk stratification model.](image-url)
Intermediate data integration is based on multiple kernel learning (Fig. 2). Kernel matrices are computed for each data view, and then they are combined, through a parametric linear function, in order to generate the final kernel matrix. Kernel-based classification (i.e. support vector machine, relevance vector machine) is subsequently applied to predict CAD risk stratification.

The skeleton and individual modules of the integrative model (i.e. merging mechanisms, machine learning algorithms, metric learning, regularization, and feature extraction) are implemented in R.

## 3 Results

Currently, the dataset is confined to demographics, risk factors, biohumoral markers and symptoms, which led us to concatenate all features into a single vector. In particular, three machine learning algorithms have been examined; a parametric model [i.e. feed-forward neural network (FFNN)], a non-parametric kernel-based model [i.e. support vector machine (SVM)] and an ensemble model [i.e. random forest (RF)]. In addition, the discriminative capacity of the available data categories has been evaluated via:

(i) a knowledge-based approach consisting in the a priori definition of 3 input cases (Case 1: Demographics, Risk Factors; Case 2: Demographics, Risk Factors, Symptoms; Case 3 Demographics, Risk Factors, Symptoms, Molecular Systemic Variables), and
(ii) feature ranking according to the InfoGain criterion accompanied by a forward selection procedure (Case 4).

Table 2 reports classification results on 101 patients (No CAD: $n=25$, Age: 58.36±7.45; Mild to Severe CAD: $n=76$, Age: 63.61±7.43) by 10-fold cross-validation.
The gradual improvement of accuracy with the enhancement of the input space is apparent, with proper customization of the input via feature ranking (\( d = 20 \)) better balancing the sensitivity to specificity ratio. SVM outperforms FFNN and RF resulting in an overall accuracy 85.1\% and a nearly perfect sensitivity (98.7\%), whereas specificity remains low (44.0\%), presumably due to the class imbalance in the dataset. The confusion matrices corresponding to SVM output in Case 3 and Case 4 are reported in Table 3 and Table 4, respectively.

| Case  | FFNN Acc. | FFNN Se. | FFNN Sp. | SVM Acc. | SVM Se. | SVM Sp. | RF Acc. | RF Se. | RF Sp. |
|-------|-----------|----------|----------|----------|---------|---------|---------|--------|--------|
| Case 1| 66.3      | 78.9     | 28.0     | 97.4     | 16.0    | 73.3    | 85.5    | 36.0   |
| Case 2| 70.3      | 81.6     | 36.0     | 94.7     | 40.0    | 75.2    | 88.2    | 36.0   |
| Case 3| 74.3      | 84.2     | 44.0     | 97.4     | 44.0    | 77.2    | 97.4    | 16.0   |
| Case 4| 78.2      | 90.8     | 40.0     | 98.7     | 44.0    | 81.2    | 92.1    | 48.0   |

\( \text{Acc.}: \) Accuracy, \( \text{Se.}: \) Sensitivity, \( \text{Sp.}: \) Specificity.

### Table 3. Confusion matrix of SVM results in Case 3.

| Annotation | Prediction                |
|------------|---------------------------|
| No CAD     | 11                        |
| Mild to Severe CAD | 14                    |
| Mild to Severe CAD | 2                      |
| No CAD     | 74                        |

### Table 4. Confusion Matrix of SVM Results in Case 4.

| Annotation | Prediction                |
|------------|---------------------------|
| No CAD     | 11                        |
| Mild to Severe CAD | 14                    |
| Mild to Severe CAD | 1                      |
| No CAD     | 75                        |

### 4 Discussion & Conclusions

CAD diagnosis is currently performed according to well-known screening strategies (i.e. CA, IVUS, OCT, CTA, MRI), whereas CVD risk can be assessed by linear regression models of clinical, laboratory and anthropometric features, assuming linearity as well as time-invariance of the underlying input-output relationships. Non-linearity is addressed by black-box parameterizations (neural networks and kernel-based models) or more transparent architectures (decision trees, dynamic Bayesian networks) or ensembles of classification models (random forests), which feature space, however, resembles that of linear approaches (i.e. established risk factors). The generalization capability of the existing machine learning models for the diagnosis of CAD or the estimation of eventful or asymptomatic CAD progression is promising; however, new knowledge coming from big data sources (e.g. molecular, cellular, inflammatory and omics data) requires more integrative machine learning solutions.
In this study, a new machine-learning approach to CAD risk stratification has been proposed relying on multimodal data integration. Its deployment and evaluation are ongoing by: (i) integrating new features concerning the lipid profile, the exome and mRNA sequencing, the exposome, and inflammatory and monocyte markers, and (ii) selecting the most effective multimodal predictive modelling scheme. Moreover, the multiclass classification problem is going to be refined by considering established risk scores of coronary atherosclerosis combining markers of stenosis severity, plaque location and composition, as assessed by computed tomography angiography.

Acknowledgement

This work is funded by the European Commission: Project SMARTOOL, “Simulation Modeling of coronary ARTery disease: a tool for clinical decision support — SMART-Tool” GA number: 689068.

Compliance with Ethical Standards

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethical approval: “All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.”

Informed consent: “Informed consent was obtained from all individual participants included in the study.”

References

1. Stone, P.H., et al., Prediction of progression of coronary artery disease and clinical outcomes using vascular profiling of endothelial shear stress and arterial plaque characteristics: the PREDICTION Study. Circulation, 2012. 126(2): p. 172-81.
2. Sakellarios, A., et al., Prediction of atherosclerotic disease progression using LDL transport modelling: a serial computed tomographic coronary angiographic study. European Heart Journal - Cardiovascular Imaging, 2017. 18(1): p. 11-18.
3. D'Agostino, R.B., Sr., et al., General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation, 2008. 117(6): p. 743-53.
4. Conroy, R.M., et al., Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J, 2003. 24(11): p. 987-1003.
5. Hippisley-Cox, J., et al., Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. Bmj, 2010. 341: p. c6624.
6. Damen, J.A., et al., Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ, 2016. 353: p. i2416.
7. Weng, S.F., et al., Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE, 2017. 12(4): p. e0174944.

8. Choi, E., et al., Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association: JAMIA, 2017. 24(2): p. 361-370.

9. Motwani, M., et al., Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. European Heart Journal, 2017. 38(7): p. 500-507.

10. Goldstein, B.A., 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, 2017. 38(23): p. 1805-1814.

11. Rumsfeld, J.S., K.E. Joynt, and T.M. Maddox, Big data analytics to improve cardiovascular care: promise and challenges. Nat Rev Cardiol, 2016. 13(6): p. 350-9.

12. Groeneveld, P.W. and J.S. Rumsfeld, Can Big Data Fulfill Its Promise? Circ Cardiovasc Qual Outcomes, 2016. 9(6): p. 679-682.

13. Li, Y., F.X. Wu, and A. Ngom, A review on machine learning principles for multi-view biological data integration. Brief Bioinform, 2016.