Artificial Intelligence and Data Mining Methods for Cardiovascular Risk Prediction

Eleni I. Georga, Nikolaos S. Tachos, Antonis I. Sakellarios, Vassiliki I. Kigka, Themis P. Exarchos, Gualtiero Pelosi, Oberdan Parodi, Lampros K. Michalis and Dimitrios I. Fotiadis

Abstract This chapter describes the state-of-the-art in artificial intelligence and machine learning methods for cardiovascular disease diagnosis and prognosis, focusing on Coronary Artery Disease (CAD). We aim at providing a cohesive overview of the existing methodologies in the topic and the most exploitable predictors for CAD staging and evolution. Thus, the relevant literature is analysed and contrasted with respect to the acquired dataset, the examined feature space, the employed predictive modelling schemes and their discriminative or predictive capacity. Moreover, important challenges stemming from the increasing ubiquity of electronic health records, personal health records and big data are discussed and, given the limitations of current approaches, future directions are delineated.

Keywords Machine learning · Artificial intelligence · Cardiovascular disease
Coronary artery disease · Atherosclerosis · Diagnosis · Prediction

E. I. Georga · V. I. Kigka · T. P. Exarchos · D. I. Fotiadis (✉)
Unit of Medical Technology and Intelligent Information Systems, Materials Science and Engineering Department, University of Ioannina, Ioannina 45110, Greece
e-mail: fotiadis@cc.uoi.gr

N. S. Tachos · A. I. Sakellarios · V. I. Kigka · D. I. Fotiadis
Biomedical Research Department, FORTH, Institute of Molecular Biology and Biotechnology, Ioannina 45110, Greece

G. Pelosi · O. Parodi
Institute of Clinical Physiology, National Research Council, Pisa 56124, Italy

L. K. Michalis
Department of Cardiology, Medical School, University of Ioannina, Ioannina 45110, Greece

S. Golemati and K. S. Nikita (eds.), *Cardiovascular Computing—Methodologies and Clinical Applications*, Series in BioEngineering,
https://doi.org/10.1007/978-981-10-5092-3_14

List of Abbreviations

ATS	Atherosclerosis
AUC	Area Under the ROC Curve
BMI	Body Mass Index
CA	Coronary Angiography
CAD	Coronary Artery Disease
CART	Classification and Regression Trees
CFS	Correlation-based Feature Selection
CTA	Computed Tomography Angiography
CVD	Cardiovascular Disease
DBN	Dynamic Bayesian Network
EHR	Electronic Health Record
FFNN	Feed-forward Neural Network
FRS	Framingham Risk Score
FURIA	Fuzzy Unordered Rule Induction Algorithm
GAM	Generalized Additive Model
GBT	Gradient Boosted Trees
HDL	High-density Lipoprotein
IVUS	Intravascular Ultrasound
LAD	Left Anterior Descending
LCX	Left Circumflex
LDL	Low-density Lipoprotein
LR	Logistic Regression
MRI	Magnetic Resonance Imaging
NPV	Negative Predictive Value
OCT	Optical Coherence Tomography
PHR	Personal Health Record
PPV	Positive Predictive Value
RBF	Radial Basis Function Network
RCA	Right Coronary Artery
RF	Random Forest
ROC	Receiver Operating Curve
RTF	Rotation Forest
SMOTE	Synthetic Minority Oversampling Technique
SOFM	Self-organizing Feature Map
SVM	Support Vector Machine
TA	Temporal Abstraction

1 Introduction

According to the World Health Organisation approximately 45% of total deaths in Europe are caused by cardiovascular disease (CVD), while 20% of total deaths occur in coronary artery disease (CAD) patients. 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. 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 (gender, smoking, hyperlipidaemia, hypertension, diabetes), but also molecular, biohumoral and biomechanical factors, such as the low endothelial shear stress. According to the guidelines of the European Society of Cardiology and the American Heart Association, the early prevention, diagnosis and prediction of disease stage may have a potential influence to the patient health status, but also may reduce the healthcare costs for the management and treatment of CAD patients [3, 4].

Predicting the risk of CVD constitutes a widely-studied problem from the perspective of statistical modelling. The majority of existing risk models, such as the Framingham risk score (FRS) [5], the Systematic COronary Risk Evaluation (SCORE) [6] and the QRISK [7], postulate a Cox proportional hazard regression or logistic regression (LR) model of relatively few traditional predictors of the disease, focusing on CAD or CVD. Most frequently applied predictor variables describe information on family history, lifestyle, comorbidities, blood pressure, physical examinations and blood lipids; whereas, other blood variables, treatment and genetics are less frequently exploited. In spite of the reported good discrimination ability of such parametric 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 [8]. More importantly: (i) precision medicine suggests more dynamic individualized nonlinear predictive modelling approaches not being hypotheses-driven, and (ii) 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 this context, artificial intelligence and machine learning naturally arise as favourable solutions to CVD risk prediction.

A case study addressing the prediction of in hospital mortality after diagnosis of acute myocardial infarction illustrated the main shortcomings of statistical methods, including non-linearity and homogeneity of interactions, as well as the challenges introduced to machine learning by CVD risk prediction models [9]. Classical machine learning and data mining techniques can be certainly employed to solve a variety of classification, regression, clustering and rule mining problems related to personalized medicine in cardiovascular research and clinical practice [10–23]. Moreover, the potential for utilizing big data analytics to improve cardiovascular health care and the emerging literature on CVD risk predictive modelling has been discussed in [24, 25].

In this chapter, we provide an overview of the state of the art on data-driven solutions to CAD diagnosis and prognosis focusing on studies employing non-imaging data. Methodological and technical issues pertaining to the development and evaluation of such models are described in detail, whereas special emphasis is placed on the predictive value of the examined feature sets and on how complex input-output interactions can be captured by the different algorithms. Our aim is to provide a clear picture of the existing methodologies to CAD diagnosis or prediction contributing to synthesizing innovative predictive schemes.

2 Non-imaging CAD Diagnosis Based on Machine Learning Methods

The diagnosis of clinically significant (obstructive) CAD is typically formulated as a binary classification problem on the basis of a confined set of features (e.g. imaging, clinical, laboratory and demographic data), with a $\geq$50% diameter stenosis in at least one main coronary artery vessel, as assessed by CA or other imaging modality, characterizing patients with CAD. Herein, we provide an overview of the literature studies approaching the CAD diagnosis problem through artificial intelligence and non-imaging procedures of data acquisition (Table 1).

Machine learning algorithms, ranging from parametric (e.g. neural networks, dynamic Bayesian networks [DBN], decision trees) to non-parametric (e.g. kernel methods) ones, have been examined towards discriminating subjects with respect to CAD existence. Feature evaluation techniques, such as filter to wrapper approaches are used, in conjunction with classification or regression algorithms, to identify the most informative features with respect to the CAD diagnosis or prediction. Kurt et al. demonstrated that a feature set comprised of traditional heart disease risk factors (i.e. age, sex, body mass index [BMI], smoking status, diabetes, hypertension, hypercholesterolemia, family history of CAD) yields predictions of low specificity, though a high sensitivity is obtained, irrespective of the employed classification algorithm [15]. More specifically, the overall accuracy of LR, classification and regression trees (CART), and feed-forward neural networks (FFNN) was comparable (~80%), whereas radial basis function network (RBF) exhibited a slightly lower performance; on the other hand, self-organizing feature maps (SOFM) behaved inaccurately regarding the identification of negative samples resulting in 7.4% specificity.

More comprehensive datasets, exploited by purely nonlinear classifiers, can improve substantially the accuracy of predictions. In that case, feature subset selection becomes a prerequisite for avoiding overfitting stemming from the increased input size. Correlation-based feature selection (CFS) using particle swarm optimization identified Duke Treadmill Score and post exercise recovery period with

Table 1 Characteristic non-imaging CAD diagnosis methods based on machine learning methods

Study	Dataset	Outcome	Methods	Feature set	Performance		
Kurt et al. [15]	n = 1245 subjects with angina associated with evidence for myocardial ischemia **Exclusion criteria**: Non-atherosclerotic CAD	**Class I—CAD (n = 865)**: ≥50% stenosis in at least one coronary artery vessel in CA **Class II—Normal (n = 380)**: Otherwise	**Classification**: LR, CART, FFNN, RBF, SOFM **Evaluation**: Training, test, validation sets (60%–20%–20%)	Age, sex, BMI, smoking status, diabetes mellitus, systemic hypertension, hypercholesterolemia, family history of CAD	**LR** Acc. (%) 79.5 Se. (%) 92.3 Sp. (%) 45.6	**CART** Acc. (%) 79.9 Se. (%) 92.3 Sp. (%) 47.1	**FFNN** Acc. (%) 79.1 Se. (%) 91.7 Sp. (%) 45.6
					RBF Acc. (%) 76.7 Se. (%) 89.5 Sp. (%) 42.6	**SOFM** Acc. (%) 73.9 Se. (%) 98.9 Sp. (%) 7.4	
Tsipouras et al. [26]	n = 199 subjects who were suspected for CAD **Exclusion Criteria**: Acute coronary syndrome, known CAD, or more than mild valvular heart disease	**Class I—Significant CAD (n = 110)**:≥ 50% diameter stenosis in at least one coronary artery vessel. **Class II—Absence of CAD (n = 89)**: Completely smooth epicardial coronary arteries without any narrowing visible in CA	**Optimized fuzzy model** 1. Decision tree (C4.5) induction 2. Extraction of the rule base from the tree 3. Development of a fuzzy model 4. Optimization of the fuzzy model's parameters **Evaluation**: ten-fold stratified cross-validation	Age, sex, family history, smoking, diabetes mellitus, hypertension, hyperlipidaemia, creatinine, glucose, total cholesterol, HDL, Triglycerides, BMI, waist, heart rate, systolic blood pressure, diastolic blood pressure, carotid femoral pulse wave velocity, augmentation index	Acc. (%) 73.4 Se. (%) 80.0 Sp. (%) 65.2		

(continued)

Table 1 (continued)

Study	Dataset	Outcome	Methods	Feature set	Performance
Anooj 2012 [27]	The UCI heart disease dataset ($n = 303$)—Cleveland data Hungarian data Switzerland data	**Class I—Existence of heart disease**:≥ 50% diameter stenosis in at least one coronary artery vessel Cleveland data: 46% positive cases Hungarian data: 37.5% positive cases Switzerland data: 93.5% positive cases **Class II—Absence of heart disease**: otherwise Cleveland data: 54% positive cases Hungarian data: 62.5% positive cases Switzerland data: 6.5% positive cases	**Automated generation of weighted fuzzy rules**: Mamdani fuzzy inference system **Evaluation**: Training-Test sets	**Cleeveland data** Age, resting blood pressure, serum cholesterol, maximum heart rate achieved (Thalach), ST depression induced by exercise relative to rest, Thal (Categorical variable, normal: 3; fixed defect: 6; reversible defect:7)	**Cleeveland data** Acc. (%) 62.4 Se. (%) 44.7 Sp. (%) 76.6
				Hungarian data Age, resting blood pressure, serum cholesterol, resting electrocardiographic results, maximum heart rate achieved (Thalach), exercise induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment	**Hungarian data** Acc. (%) 46.9 Se. (%) 74.3 Sp. (%) 31.7

(continued)

Table 1 (continued)

Study	Dataset	Outcome	Methods	Feature set	Performance
				Switzerland data Age, sex, chest pain type, resting blood pressure, fasting blood glucose, resting electrocardiographic results, maximum heart rate achieved (Thalach), ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels, Thal (Categorical variable, normal: 3; fixed defect: 6; reversible defect:7)	**Switzerland data** Acc. (%) 51.2 Se. (%) 52.6 Sp. (%) 33.3
Karabulut and Ibrikci [28]	The UCI heart disease dataset ($n = 303$)	**Class I—Existence of heart disease (*n* = 165)** :≥ 50% diameter stenosis **Class II—Absence of heart disease (*n* = 138)**: otherwise	**Classification**: RTF with FFNNs as the base classifier **Evaluation**: ten-fold cross-validation	Age, sex, chest pain type, resting systolic blood pressure, serum cholesterol, fasting blood glucose, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels coloured by fluoroscopy, exercise thallium scintigraphic defects	Acc. (%) 91.2 Se. (%) 95.6 Sp. (%) 86.7 AUC 0.915
Nahar et al. [29]	The UCI heart disease dataset ($n = 303$)	**Class I—Existence of heart disease**:≥ 50% diameter stenosis **Class II—Absence of heart disease**: otherwise	**Feature Selection**: CFS, knowledge-based feature selection (MFS) **Classification**: SMO **Evaluation**: Approach I ten-fold cross-validation Approach II training—test sets (66–33%) combined with internal ten-fold cross-validation for hyper-parameter optimization	**CFS** Old peak, number of vessels coloured, Thal (categorical variable, normal: 3; fixed defect: 6; reversible defect: 7)	Approach II - MFS: Acc. (%) 77.95 Se. (%) 81.1

(continued)

Table 1 (continued)

Study	Dataset	Outcome	Methods	Feature set	Performance	
				Knowledge-based feature selection (MFS) Age, chest pain type, resting blood pressure, cholesterol, fasting blood glucose, resting heart rate, maximum heart rate, and exercise induced angina	Approach II - MFS combined with CFS: Acc. (%) 83.83 Se. (%) 91.9	
Alizadehsani et al. [30]	$n = 303$ subjects	**Class I—CAD (*n* = 865)**:$\geq 50\%$ stenosis in at least one coronary artery vessel in CA **Class II—Normal (*n* = 380)**: Otherwise	**Feature selection**: Embedded in SVM weights **Classification**: SVM, Naïve Bayes, bagging of SVMs, FFNN **Association rule mining**: Apriori **Evaluation**: ten-fold cross-validation	Typical chest pain, region with regional wall motion abnormality*, age, ejection fraction*, hypertension, diabetes, T inversion, erythrocyte sedimentation rate, Q wave, ST elevation, pulse rate, BMI, lymph, blood pressure*, dyspnoea, HDL, creatinine*, white blood cell*, weight, valvular heart disease, function class, airway disease, haemoglobin, triglycerides*, bundle branch block, Na*, sex, Left ventricular hypertrophy, haemoglobin*, family history	**SVM** Acc. (%) 93.39 ± 5.14 Se. (%) 95.37 Sp. (%) 88.51	**Bagging SVM** Acc. (%) 92.74 ± 6.43 Se. (%) 95.37 Sp. (%) 86.21
					FFNN Acc. (%) 87.13 ± 5.84 Se. (%) 90.28 Sp. (%) 79.31	**Naïve Bayes** Acc. (%) 55.37 ± 9.62 Se. (%) 38.89 Sp. (%) 96.55

(continued)

Table 1 (continued)

Study	Dataset	Outcome	Methods	Feature set	Performance	
Alizadehsani et al. [30]	$n = 303$ subjects	**Class I—CAD (*n* = 865)**:$\geq 50\%$ stenosis in at least one coronary artery vessel in CA **Class II—Normal (*n* = 380)**: Otherwise	**Feature selection**: Embedded in SVM weights **Classification**: SVM, Naïve Bayes, bagging of SVMs, FFNN **Association rule mining**: Apriori **Evaluation**: ten-fold cross-validation	Typical chest pain, region with regional wall motion abnormality*, age, ejection fraction*, hypertension, diabetes, T inversion, erythrocyte sedimentation rate, Q wave, ST elevation, pulse rate, BMI, lymph, blood pressure*, dyspnoea, HDL, creatinine*, white blood cell*, weight, valvular heart disease, function class, airway disease, haemoglobin, triglycerides*, bundle branch block, Na*, sex, Left ventricular hypertrophy, haemoglobin*, family history		
Alizadehsani et al. [31]	$n = 303$	**Problem I** Class I—LAD stenotic: $\geq 50\%$ stenosis in LAD artery Class II—LAD normal: Otherwise **Problem II** Class I—LCX stenotic: $\geq 50\%$ stenosis in LCX artery Class II—LCX normal: Otherwise **Problem III** Class I—RCA stenotic: $\geq 50\%$ diameter stenosis in RCA artery Class II—RCA normal: Otherwise	**Feature selection** Approach I Different feature set for each artery: SVM weights Approach II Common feature set for all arteries: Information Gain **Classification**: SVM with kernel fusion **Association rule mining**: Apriori **Evaluation**: ten-fold cross-validation	**Feature selection approach II** Typical chest pain, atypical chest pain, ejection fraction, region with regional wall motion abnormality, age, valvular heart disease, diabetes, hypertension, T inversion, lymphocyte, fasting blood glucose*, neutrophil, blood pressure, Nonanginal chest pain, fasting blood glucose, erythrocyte sedimentation rate, Na, K, creatinine, creatinine*, blood urea nitrogen, ST elevation, white blood cell count, neutrophil*, Q wave, white blood cell*, weight	**Feature selection approach I**	**Feature selection approach II**

(continued)

Table 1 (continued)

Study	Dataset	Outcome	Methods	Feature set	Performance	
					LAD Acc. (%) 85.81 ± 1.7 Se. (%) 92.66 ± 1.9 Sp. (%) 76.19 ± 1.2	LAD Acc. (%) 86.14 ± 1.1 Se. (%) 90.96 ± 0.9 Sp. (%) 79.37 ± 1.4
					LCX Acc. (%) 77.23 ± 1.6 Se. (%) 69.75 ± 1.9 Sp. (%) 82.07 ± 1.9	LCX Acc. (%) 83.17 ± 0.4 Se. (%) 90.96 ± 0.3 Sp. (%) 72.22 ± 0.6
					RCA Acc. (%) 81.85 ± 0.4 Se. (%) 68.42 ± 0.6 Sp. (%) 89.95 ± 0.1	RCA Acc. (%) 83.50 ± 0.8 Se. (%) 87.01 ± 1.2 Sp. (%) 78.57 ± 0.7
Verma et al. [32]	$n = 335$ subjects who were suspected for CAD	**Class I—CAD (48.9%)** **Class II—No CAD (51.1%)**	**Feature Selection**: CFS with particle swarm optimization **Clustering**: k-means **Classification**: FFNN, LR, Fuzzy unordered rule induction algorithm (FURIA), Decision tree (C4.5) **Evaluation**: ten-fold cross-validation	Smoking, diabetes, HDL, duke treadmill score, duration of recovery with persistent ST changes	**FFNN** Acc. (%) 88.40	**FURIA** Acc. (%) 82.80
					LR Acc. (%) 84.11	**C4.5** Acc. (%) 80.68

Acc accuracy, *Se* sensitivity, *Sp* specificity

* Discretized according to “Braunwald’s heart disease: a textbook of cardiovascular medicine”

persistent electrocardiographic ST-segment changes, following a treadmill stress testing, amongst the most informative features with respect to CAD diagnosis [32]. In particular, a FFNN fed additionally with information on smoking, diabetes, and high-density lipoprotein (HDL) attains 88.4% accuracy. Besides filter-based feature selection methods, feature selection embedded in learning algorithms have been applied to reduce the dimensionality of the feature space. The two-stage methodology by Alizadehsani et al. encompassed: (i) an evaluation of the discriminative capability of 54 features concerning demographic, clinical, electrocardiographic, echocardiographic, and laboratory data based on the support vector machine (SVM) weight vector, and (ii) a comparative study of the performance of four algorithms including naïve Bayes, SVM, bagging SVM, and FFNN [30]. The kernel-based methods (i.e. SVM and bagging SVM) outperformed both FFNN and naïve Bayes, exhibiting 93.4 and 92.7% accuracy as well as high sensitivity and specificity rates. In a subsequent study, Alizadehsani et al. examined the diagnostic accuracy of the same feature set with respect to the level of stenosis of each coronary artery [i.e. left anterior descending (LAD) artery, left circumflex (LCX) artery and right coronary artery (RCA)] separately, formulating a 2-class problem where a $\geq$50% diameter stenosis characterizes a stenotic artery [31]. In particular, (i) a common feature set was used for the diagnosis of the stenosis of each coronary artery, encompassing the 24 top ranked features according to a combined info-gain index, and (ii) a new multiple kernel learning algorithm was proposed to define the most appropriate hyperplane which may classify the dataset. The stenosis of LAD, LCX and RCA is diagnosed with 86.14%, 83.17% and 83.5% accuracy, respectively. On the other hand, Nahar et al. [29], using the UCI Cleveland heart disease dataset, showed that knowledge-based feature selection is an asset for the diagnosis of heart disease [33]. Nahar et al. decomposed the 5-class problem into 5 binary classification problems, which were solved employing well-known classification algorithms, i.e. naïve Bayes, SVM, k-nearest neighbour algorithm, Adaboost.M1, J48 decision tree, and PART rule-based classifier. The results indicated that: (i) the best performing algorithm in the case where the whole feature set is considered was SVM, and (ii) feature selection enhances the accuracy for the majority of algorithms and for all binary problems.

Ensemble learning of the UCI Cleveland heart disease dataset, when focusing on the heart disease diagnosis problem (Class 0 vs. Class1–4), has been shown to improve the accuracy of FFNN [29, 34]; an ensemble of three FFNNs yielded 89.01% accuracy, 80.95% sensitivity and 95.91% specificity, whereas rotation forest (RTF) using FFNN as the base classifier improved its accuracy by 7% reaching 91.2%.

Unlike most machine learning techniques, fuzzy rule-based classifiers provide interpretable decision making. To that end, Tsipouras et al. proposed an optimized fuzzy model for the diagnosis of CAD considering traditional cardiovascular risk factors as well as two non-invasive indices of pulse wave velocity, namely carotid–femoral and augmentation index. A four-stage methodology was developed including: (i) induction of a decision tree, (ii) extraction of the rule base from the decision tree, in disjunctive normal form and formulation of a crisp model, (iii) transformation of the crisp set of rules into a fuzzy model, and (iv) optimization of the parameters of the fuzzy model [26]. The optimized fuzzy model resulted in 73.4% accuracy, 80.0% sensitivity and 65.2% specificity, exhibiting comparable performance with a FFNN (73.9% accuracy) and significantly better results than an adaptive neuro-fuzzy inference system (56.8% accuracy), both applied to the same task.

3 Non-imaging CAD Prediction Based on Machine Learning Methods

Prediction of CAD development or CAD progression can be also viewed as a classification problem which involves a temporal dimension. The existing machine learning predictive modelling approaches of CAD, which are based on non-imaging data, utilize information obtained either at a specific time instance *t* (at baseline) or up to a specific time instance *t* in order to predict one patient's status at time $t + h$ (at follow-up), where *h* is the prediction horizon, typically, expressed in years. Well-designed prospective clinical studies constitute the standard data source of CAD prediction machine learning methods. Nevertheless, the consolidation of EHRs have inspired researchers to explore longitudinal patient health information from EHRs towards constructing data-driven CAD risk prediction models. The studies presented in this section are representative of the spectrum of methodologies which are employed in the related literature (Table 2).

Exarchos et al. assembled and analyzed a multivariate dataset aiming at: (i) identifying the most significant features towards the progression of atherosclerosis (ATS), and (ii) developing a decision support system inferencing the prognosis of the disease [35]. Patients underwent angiographic assessment by CTA or CA both at the baseline visit as well as during the follow-up, whereas demographic data, clinical data, standard biohumoral analytes, adhesion molecules, markers of monocyte activation, and therapy, were measured at the same time-slices. To this end, Exarchos et al. defined two binary outcome variables capturing the severity and progression of ATS: (i) number of stenoses: Binary variable indicating whether any coronary vessels exhibit stenosis >50%, (ii) ATS progression: Binary variable indicating whether the number or percentage of stenosis in any vessel increased from the baseline to the follow-up visit. A hybrid score is also utilized according to which each patient is assigned a severity level in the range [0, . . . , 17], with 17 denoting the most severe condition. In addition, two analysis axes were defined. The first one concerns the solution of the binary classification problem employing baseline data and encompasses: (i) class imbalance handling through the synthetic minority oversampling technique (SMOTE), (ii) feature selection by the CFS, gain ratio and wrapper algorithms, and (iii) evaluation and comparison of a multitude of classification algorithms (i.e. Bayesian network, naïve Bayes, FFNN, SVM, decision tree, and random forest [RF]). The second axis of analysis considers temporal modelling of the information obtained both at baseline and follow-up visits by DBN. The results pertaining to the first analysis axis indicated that naïve Bayes yields the highest performance, 91.7%

Table 2 Characteristic non-imaging CAD prediction methods based on machine learning methods

Study	Dataset	Outcome	Methods	Feature Set	Performance
Exarchos et al. [35]	$n = 39$ subjects **Average follow-up time**: 31.4 ± 17.2 months	**Problem I—Number of stenoses**: Binary variable indicating whether one or more coronary vessels exhibit stenosis > 50% **Problem II—ATS progression**: Binary variable indicating whether the number or percentage of stenosis in any vessel increased from the baseline to the follow-up visit **Problem III—Hybrid score**: Each patient is assigned a severity level in the range [0,…17], with 17 denoting the most severe condition	**Axis I—Baseline analysis**: Baseline information is used to predict the progression of ATS **Resampling**: SMOTE algorithm **Feature selection**: CFS, gain ratio algorithm, wrapper algorithm **Classification**: Bayesian network, Naïve Bayes, FFNN, SVM, Decision tree, RF **Evaluation**: ten-fold cross-validation, leave 1 patient out	**Axis I**: **Problem I** Age, sex, weight, diabetes, family history, left ventricular ejection fraction, cholesterol, HDL, creatinine clearance, glucose. E-selectin, vascular cell adhesion molecule 1 (VCAM-1), beta-blockers, statins **Problem II** Weight, diabetes, hypertension, smoke, FRS, infarct site, cholesterol, statins **Problem III** Hypercholesterolemia, hypertension, monocytes, ca antagonists	**Axis I**: **Problem I** Wrapper and Naïve Bayes Acc. (%) 91.7 Se. (%) 93.3 Sp. (%) 90% PPV(%) 90.3 AUC 0.944 **Problem II** CFS and Naïve Bayes Acc. (%) 93.3 Se. (%) 96.7 Sp. (%) 90% PPV (%) 90.6 AUC 0.937

(continued)

Table 2 (continued)

Study	Dataset	Outcome	Methods	Feature Set		Performance	
			Axis II—Temporal analysis: Snapshots of the patient's status over the follow-up period are analyzed to model ATS evolvement **Resampling**: SMOTE algorithm Statistical testing: Chi-square test, Fischer's test **Temporal analysis**: DBN	**Axis II**: **Problem I** Diabetes, hypercholesterolemia, total cholesterol to HDL ratio, triglycerides, Glucose **Problem II** Diabetes, low-density lipoprotein (LDL), infarct site, creatinine, creatinine clearance, monocytes, Total cholesterol to HDL ratio, white blood cell, Smoke		**Axis II**: **Problem I** Acc. (%) 79 **Problem II** Acc. (%) 83	
Weng et al. [19]	$n = 378256$ subjects free from cardiovascular disease at outset	The first recorded diagnosis of a fatal or non-fatal cardio-vascular event over 10 years ($n = 24970$)	**Feature selection**: Ranking mechanism embedded into machine learning classification algorithms **Classification**: RF, LR, gradient boosting machines, FFNN **Evaluation**: Training-Test sets	**RF** Age, gender, ethnicity, smoking, HDL, glycated haemoglobin (HbA1c), triglycerides, townsend deprivation index, BMI, total cholesterol	**LR** Ethnicity, age, townsend deprivation index, gender, smoking, atrial fibrillation, chronic kidney disease, rheumatoid arthritis, family history of premature CAD, chronic obstructive pulmonary disease	**RF** Se. (%) 65.3% Sp. (%) 70.5% PPV(%) 17.8% NPV(%) 95.4% AUC 0.745	**Gradient boosting machines** Se. (%) 67.5% Sp. (%) 70.7% PPV(%) 18.4% NPV(%) 95.7% AUC 0.761

(continued)

Table 2 (continued)

Study	Dataset	Outcome	Methods	Feature Set		Performance		
				Gradient boosting machines Age, gender, ethnicity, smoking, HDL, triglycerides, total cholesterol, glycated haemoglobin (HbA1c), systolic blood pressure, townsend deprivation index	**FFNN** Atrial fibrillation, ethnicity, oral corticosteroid prescribed, age, severe mental illness, townsend deprivation index, chronic kidney disease, bmi, smoking, gender	**LR** Se. (%) 67.1% Sp. (%) 70.7% PPV(%) 18.3% NPV(%) 95.6% AUC 0.760	**FFNN** Se. (%) 67.5% Sp. (%) 70.7% PPV(%) 18.4% NPV(%) 95.7% AUC 0.764	
Kennedy et al. [36]	$n = 113973$ subjects **Exclusion criteria**: Cerebrovascular disease or CVD diagnosis or event during the baseline year	Occurrence of a fatal cerebrovascular or CVD event over a 5-year period ($n = 4995$)	**Feature selection**: Three feature sets were explored: 1. Feature set 1: Traditional risk predictors 2. Feature set 2: Traditional risk predictors and Medication 3. Feature set 3: Traditional risk predictors, medication, and labs/vital signs/diagnoses/other **Classification**: FRS, LR, GAM, GBT **Evaluation**: ten-fold cross-validation	**Traditional risk predictors**: Age, male, Systolic BP, Total cholesterol to HDL ratio, Diabetes **Medication**: Hypertension, Lipids, Diabetes, Narcotics or opiates, Benzodiazepines, Levothyroxines, Anticoagulants **Labs/vital signs/diagnoses/other**: Albumin, blood urea nitrogen, LDL, serum creatinine, pulse, pulse pressure, baseline diagnoses: chronic obstructive pulmonary disease, periodontitis, inflammatory arthritis, sleep apnea, body mass index, number of visits		**Feature set 1**	**Feature set 2**	**Feature set 3**

(continued)

Table 2 (continued)

Study	Dataset	Outcome	Methods	Feature Set	Performance		
					FRS AUC(%) 71.3 ± 1.0 ***LR*** AUC(%) 72.6 ± 1.0 ***GAM*** AUC(%) 73.1 ± 0.9 ***GBT*** AUC(%) 73.1 ± 0.9	***FRS*** AUC(%) - ***LR*** AUC(%) 74.3 ± 0.8 ***GAM*** AUC(%) 74.8 ± 0.7 ***GBT*** AUC(%) 74.9 ± 0.7	***FRS*** AUC(%) - ***LR*** AUC(%) 76.3 ± 1.0 ***GAM*** AUC(%) 77.5 ± 0.9 ***GBT*** AUC(%) 77.8 ± 0.9
Orphanou et al. [37]	STULONG dataset—849 men monitored from 2 to 21 years	Occurence of CAD event in the last 3 years of the total observation period (21 years)	Prediction of the risk of a patient suffering a CAD event during a particular time period t, based on the patient's medical history up to time t-1 **Resampling**: SMOTE-N oversampling with clustering undersampling **Feature selection**: Knowledge-based **Temporal abstractions derivation** (State, trend and persistence TAs) **Classification**: Extended DBN **Evaluation**: k-fold cross-validation	Age, blood pressure, dyslipidemia levels, obesity history, diabetes history, cholesterol and hypertension medication, smoking, diet, exercise	Precision: 0.7207 Recall: 0.75 F1 score: 0.7351 AUC: 0.778		

Acc accuracy, *Se* sensitivity, *Sp* specificity and 93.3%, for the prediction of both the number of stenoses and ATS progression, respectively. With regard to the temporal analysis, DBN provided a satisfactory accuracy of 87 and 84% for the two aforementioned outcome variables. Nevertheless, Exarchos et al. note that the application of the SMOTE algorithm might have introduced an overestimation of the performance metrics.

Identifying patients at high risk of a CVD event in the follow-up period constitutes a different endpoint than estimating asymptomatic CAD progression. Recently, Weng et al. evaluated four machine-learning algorithms (i.e. RF, LR, gradient boosting machines and FFNNs) with respect to the prediction of first CVD event over a 10-year follow-up period on EHR data of a cohort of patients ($n = 378256$), who were free from cardiovascular disease at outset [19]. In total, 30 variables, concerning patient's characteristics, clinical and laboratory data, CVD risk factors, history, lifestyle and medications, with potential to be associated with CVD were examined. Their importance was determined by the embedded in each algorithm mechanisms of feature ranking, and the overall ranking was consistent with the standard risk factors included in the American College of Cardiology/American Heart Association (ACC/AHA) model. Compared with the established recommendations on the assessment of cardiovascular risk by the ACC/AHA [38], a considerable improvement in the area under the receiver operating curve (AUC) measure was obtained: RF +1.7% (AUC 0.745), LR +3.2% (AUC 0.760), gradient boosting +3.3% (AUC 0.761), FFNN +3.6% (AUC 0.764). More specifically, the highest achieving algorithm, i.e. FFNN, featured 67.5% sensitivity, 70.7% specificity, 18.4% positive predictive value (PPV) and 95.7% negative predictive value (NPV), resulting in a net increase of 355 true positive CVD cases (4,998 out of 7,404 total CVD cases) as compared with ACC/AHA model (sensitivity 62.7%, specificity 70.3%, PPV 17.1%, NPV 95.1%).

Similarly, a systematic comparative study of modelling approaches for predicting the risk of a fatal cardiovascular event over a 5-year period based on comprehensive EHR data demonstrated the predominance of gradient boosted trees over the FRS; the AUC increased from 71 to 78% [36]. In particular, the predictive capacity of traditional risk factors (i.e. age, gender, systolic blood pressure, total cholesterol to HDL ratio, diabetes) along with medication information, laboratory and clinical data, was examined, with non-parametric algorithms (i.e. generalized additive model [GAM], gradient boosted trees [GBT]) capturing better the relationships in the feature set as its size increases. Nevertheless, we should note that in the two aforementioned studies the values of all features were recorded during the baseline year, without exploring the longitudinal nature of EHR data.

From a different perspective, Orphanou et al. proposed a dynamic approach to CAD prognosis integrating DBN and temporal abstractions (TAs) and which has been applied to a longitudinal benchmark dataset [37]. In particular, the STULONG dataset comprises from 1 to 20 examinations for each patient, which corresponds to 1–24 years of clinical monitoring. Essentially, the proposed approach consisted of the following steps: (i) data pre-processing and knowledge-based feature selection, (ii) derivation of basic TAs (state, trend, and persistence TAs), and (iii) deployment and evaluation of the extended DBN. The selected feature set, which was incorporated into the extended DBN, contained information on well-established CAD risk factors; namely, hypertension, smoking status, dyslipidaemia level, obesity, diabetes, patient's and family history, age, hypertension and high-cholesterol medication, diet, and exercise. The maximum observation period was set equal to 21 years, whereas the outcome variable describes the occurrence of CAD event in the last 3 years of the total observation period (19–21 years). Therefore, the examined problem is postulated as follows: prediction of the risk of a patient suffering a CAD event during a particular time period $\Delta t = [t, t + 2]$, based on the patient's medical history up to time t. Orphanou et al. applied two oversampling methods (SMOTE, random oversampling of the minority class) as well as a combination of oversampling with undersampling (SMOTE combined with $k = 2$-means clustering undersampling, random oversampling combined with $k = 2$-means clustering undersampling), aiming at addressing the class imbalance problem. A 72% precision accompanied with a 75% recall and 74% F1 score were obtained for the combination of random oversampling with $k = 2$-means clustering undersampling. Moreover, the extended DBN model outperformed a DBN model without TAs applied to the same task.

4 Discussion and Future Trends

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 baseline clinical, laboratory and anthropometric features, assuming linearity as well as time-invariance of the underlying input-output relationships. Nonlinearity is addressed by black-box parameterizations (neural networks and kernel-based models) or more transparent architectures (decision trees, DBN) or ensembles of classification models (RF, RTF), 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, no consensus has been reached on feature learning and model identification and validation.

New research approaches to CVD risk prediction can be enhanced as follows:

i. First, the input space can be partitioned into coherent and well-separated clusters which portray the innate data similarities or structures. Unsupervised learning (k-means, expectation maximization clustering, hierarchical agglomerative clustering) can be investigated towards identifying groups of patients with similar characteristics, especially for omics data, or organize patients into a hierarchy of clusters. Profile analysis can also rely on pattern mining aiming at identifying dynamic dependencies into genomics, clinical, biohumoral, molecular/cellular, and environmental/lifestyle information, which may have a prognostic relevance in CAD. Especially longitudinal data trajectories (PHRs, EHRs) has to be explored for co-occurrence relationships (static data analytics) as well as sequences of events (dynamic data analytics) aiming at inferring high-level context describing a patient or a group of individuals [39]. For this purpose, innovative temporal pattern mining algorithms have been proposed that consider the temporal dimension of the data [40–45] as well as deep-learning approaches to EHRs representation [46].

ii. Second, special emphasis should be placed to the identification of a minimum subset of the most informative features, aiming at, eventually, refining the existing stratification scores and, in parallel, increasing their accuracy. Modality and feature learning should be addressed such that conditional dependencies between input and output variables are effectively detected in the quantized space even in the presence of groups of highly-correlated features. To this end, sequential (backward or forward) feature selection, evaluating the incremental predictive value of the input space, would allow the adoption of only those parameters that contribute to the improvement in accuracy of CAD stratification.

iii. Third, the core of predictive modelling ought to be built upon adaptive non-linear regression or classification solutions on the basis of the results of patient's profiling analysis, feature learning and dynamic pattern analysis. In this direction, contemporary powerful learning methods (e.g. deep-learning, DBN and continuous time Bayesian networks) and big data solutions can be employed to identify novel correlations and causal relationships, strongly related with the onset of CAD. In addition, the discriminative/predictive capacity of the extracted clusters or temporal patterns (grouping of patients), can be studied, resulting to a hybrid prediction scheme. On top of these, a comprehensive pre-processing procedure has to be applied in order to resolve issues related with data heterogeneities, missing data unbalanced classes and sampling times, and assure a high-quality adequately-synchronized dataset.

iv. Finally, the expected generalization performance of the computational model should be evaluated on large-scale multivariate datasets using well-established statistical measures and approaches aiming at balancing the trade-off among accuracy, interpretability and time and space complexity. The efficient integration of personalized behavioural and psychosocial data with health data can provide a better understanding of the effect of patient's daily context on clinical health outcomes.

Concluding, predictive modelling of CAD diagnosis or CAD progression should aim to develop hybrid multi-level multi-scale schemes, combining unsupervised and supervised adaptive learning systems and being built upon novel multi-sensor, multi-source and multi-process information fusion schemes. Intelligent data mining and machine learning algorithms integrating previous clinical risk stratification models and refining novel ones using new knowledge coming from big data sources (e.g. molecular, cellular, inflammatory and omics data) could advance existing modelling methods in terms of accuracy, precision and interpretability. New paradigms should emphasize on both architecture and algorithms of the predictive model aiming at promoting the synergism among different information analysis levels.

Acknowledgements This work is partially funded by the European Commission: Project SMARTOOL, "Simulation Modeling of coronary ARTery disease: a tool for clinical decision support—SMARTool" GA number: 689068.

References

1. Stone PH, Saito S, Takahashi S, Makita Y, Nakamura S, Kawasaki T, Takahashi A, Katsuki T, Nakamura S, Namiki A, Hirohata A, Matsumura T, Yamazaki S, Yokoi H, Tanaka S, Otsuji S, Yoshimachi F, Honye J, Harwood D, Reitman M, Coskun AU, Papafaklis MI, Feldman CL (2012) 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 126(2):172–181. https://doi.org/10.1161/circulationaha.112.096438
2. Sakellarios A, Bourantas CV, Papadopoulou S-L, Tsirka Z, de Vries T, Kitslaar PH, Girasis C, Naka KK, Fotiadis DI, Veldhof S, Stone GW, Reiber JHC, Michalis LK, Serruys PW, de Feyter PJ, Garcia-Garcia HM (2017) Prediction of atherosclerotic disease progression using LDL transport modelling: a serial computed tomographic coronary angiographic study. Eur Heart J Cardiovasc Imaging 18(1):11–18. https://doi.org/10.1093/ehjci/jew035
3. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, Cooney MT, Corra U, Cosyns B, Deaton C, Graham I, Hall MS, Hobbs FD, Lochen ML, Lollgen H, Marques-Vidal P, Perk J, Prescott E, Redon J, Richter DJ, Sattar N, Smulders Y, Tiberi M, van der Worp HB, van Dis I, Verschuren WM (2016) European Guidelines on cardiovascular disease prevention in clinical practice: the Sixth Joint Task Force of the European Society of Cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). Eur Heart J 37(29):2315–2381. https://doi.org/10.1093/eurheartj/ehw106
4. Ferguson JF, Allayee H, Gerszten RE, Ideraabdullah F, Kris-Etherton PM, Ordovas JM, Rimm EB, Wang TJ, Bennett BJ (2016) Nutrigenomics, the microbiome, and gene-environment interactions: new directions in cardiovascular disease research, prevention, and treatment: a scientific statement from the American Heart Association. Circ Cardiovasc Genet 9(3):291–313. https://doi.org/10.1161/hcg.0000000000000030
5. D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB (2008) General cardiovascular risk profile for use in primary care: the Framingham Heart Study. Circulation 117(6):743–753. https://doi.org/10.1161/circulationaha.107.699579
6. Conroy RM, Pyorala K, Fitzgerald AP, Sans S, Menotti A, De Backer G, De Bacquer D, Ducimetiere P, Jousilahti P, Keil U, Njolstad I, Oganov RG, Thomsen T, Tunstall-Pedoe H, Tverdal A, Wedel H, Whincup P, Wilhelmsen L, Graham IM (2003) Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J 24(11):987–1003
7. Hippisley-Cox J, Coupland C, Robson J, Brindle P (2010) Derivation, validation, and evaluation of a new QRISK model to estimate lifetime risk of cardiovascular disease: cohort study using QResearch database. BMJ (Clinical research ed) 341:c6624. https://doi.org/10.1136/bmj.c6624
8. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, Schlussel MM, Gerry S, Black JA, Heus P, van der Schouw YT, Peelen LM, Moons KG (2016) Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ (Clinical research ed) 353:i2416. https://doi.org/10.1136/bmj.i2416
9. Goldstein BA, Navar AM, Carter RE (2017) Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J 38(23):1805–1814. https://doi.org/10.1093/eurheartj/ehw302
10. Volzke H, Schmidt CO, Baumeister SE, Ittermann T, Fung G, Krafczyk-Korth J, Hoffmann W, Schwab M, Meyer zu Schwabedissen HE, Dorr M, Felix SB, Lieb W, Kroemer HK (2013) Personalized cardiovascular medicine: concepts and methodological considerations. Nat Rev Cardiol 10(6):308–316. https://doi.org/10.1038/nrcardio.2013.35
11. Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart WF (2016) Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. Circ Cardiovasc Qual Outcomes 9(6):649–658. https://doi.org/10.1161/circoutcomes.116.002797

12. Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS (2010) Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Trans Inf Technol Biomed 14(3):559–566. https://doi.org/10.1109/TITB.2009.2038906
13. Nahar J, Imam T, Tickle KS, Chen Y-PP (2013) Association rule mining to detect factors which contribute to heart disease in males and females. Expert Syst Appl 40(4):1086–1093. https://doi.org/10.1016/j.eswa.2012.08.028
14. Austin PC, Tu JV, Ho JE, Levy D, Lee DS (2013) Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. J Clin Epidemiol 66(4):398–407. https://doi.org/10.1016/j.jclinepi.2012.11.008
15. Kurt I, Ture M, Kurum AT (2008) Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst Appl 34(1):366–374. https://doi.org/10.1016/j.eswa.2006.09.004
16. Choi E, Schuetz A, Stewart WF, Sun J (2017) Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc JAMIA 24(2):361–370. https://doi.org/10.1093/jamia/ocw112
17. Hassan N, Sayed OR, Khalil AM, Ghany MA (2016) Fuzzy soft expert system in prediction of coronary artery disease. Int J Fuzzy Syst. https://doi.org/10.1007/s40815-016-0255-0
18. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, Andreini D, Budoff MJ, Cademartiri F, Callister TQ, Chang H-J, Chinnaiyan K, Chow BJW, Cury RC, Delago A, Gomez M, Gransar H, Hadamitzky M, Hausleiter J, Hindoyan N, Feuchtner G, Kaufmann PA, Kim Y-J, Leipsic J, Lin FY, Maffei E, Marques H, Pontone G, Raff G, Rubinshtein R, Shaw LJ, Stehli J, Villines TC, Dunning A, Min JK, Slomka PJ (2017) Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J 38(7):500–507. https://doi.org/10.1093/eurheartj/ehw188
19. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 12(4):e0174944. https://doi.org/10.1371/journal.pone.0174944
20. Rao VSH, Kumar MN (2013) Novel approaches for predicting risk factors of atherosclerosis. IEEE J Biomed Health Inform 17(1):183–189. https://doi.org/10.1109/TITB.2012.2227271
21. Kukar M, Kononenko I, Grošelj C (2011) Modern parameterization and explanation techniques in diagnostic decision support system: a case study in diagnostics of coronary artery disease. Artif Intell Med 52(2):77–90. https://doi.org/10.1016/j.artmed.2011.04.009
22. Shouman M, Turner T, Stocker R (2012) Using data mining techniques in heart disease diagnosis and treatment. In: 2012 Japan-Egypt conference on electronics, communications and computers, 6–9 March 2012, pp 173–177. https://doi.org/10.1109/jec-ecc.2012.6186978
23. Melillo P, Izzo R, Orrico A, Scala P, Attanasio M, Mirra M, De Luca N, Pecchia L (2015) Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. PLoS One 10(3):e0118504. https://doi.org/10.1371/journal.pone.0118504
24. Rumsfeld JS, Joynt KE, Maddox TM (2016) Big data analytics to improve cardiovascular care: promise and challenges. Nat Rev Cardiol 13(6):350–359. https://doi.org/10.1038/nrcardio.2016.42
25. Groeneveld PW, Rumsfeld JS (2016) Can big data fulfill its promise? Circ Cardiovasc Qual Outcomes 9(6):679–682. https://doi.org/10.1161/circoutcomes.116.003097
26. Tsipouras MG, Exarchos TP, Fotiadis DI, Kotsia AP, Vakalis KV, Naka KK, Michalis LK (2008) Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. IEEE Trans Inf Technol Biomed 12(4):447–458. https://doi.org/10.1109/TITB.2007.907985
27. Anooj PK (2012) Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. J King Saud Univ-Comput Inf Sci 24(1):27–40. https://doi.org/10.1016/j.jksuci.2011.09.002

28. Karabulut EM, Ibrikci T (2012) Effective diagnosis of coronary artery disease using the rotation forest ensemble method. J Med Syst 36(5):3011–3018. https://doi.org/10.1007/s10916-011-9778-y
29. Nahar J, Imam T, Tickle KS, Chen Y-PP (2013) Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. Expert Syst Appl 40(1):96–104. https://doi.org/10.1016/j.eswa.2012.07.032
30. Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, Bahadorian B, Sani ZA (2013) A data mining approach for diagnosis of coronary artery disease. Comput Methods Programs Biomed 111(1):52–61. https://doi.org/10.1016/j.cmpb.2013.03.004
31. Alizadehsani R, Zangooei MH, Hosseini MJ, Habibi J, Khosravi A, Roshanzamir M, Khozeimeh F, Sarrafzadegan N, Nahavandi S (2016) Coronary artery disease detection using computational intelligence methods. Knowl-Based Syst 109:187–197. https://doi.org/10.1016/j.knosys.2016.07.004
32. Verma L, Srivastava S, Negi PC (2016) A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. J Med Syst 40(7):178. https://doi.org/10.1007/s10916-016-0536-z
33. Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, Guppy KH, Lee S, Froelicher V (1989) International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol 64(5):304–310
34. Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. Expert Syst Appl 36(4):7675–7680. https://doi.org/10.1016/j.eswa.2008.09.013
35. Exarchos KP, Carpegianni C, Rigas G, Exarchos TP, Vozzi F, Sakellarios A, Marraccini P, Naka K, Michalis L, Parodi O, Fotiadis DI (2015) A multiscale approach for modeling atherosclerosis progression. IEEE J Biomed Health Inform 19(2):709–719. https://doi.org/10.1109/jbhi.2014.2323935
36. Kennedy EH, Wiitala WL, Hayward RA, Sussman JB (2013) Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med Care 51(3):251–258. https://doi.org/10.1097/MLR.0b013e31827da594
37. Orphanou K, Stassopoulou A, Keravnou E (2016) DBN-extended: a dynamic Bayesian network model extended with temporal abstractions for coronary heart disease prognosis. IEEE J Biomed Health Inform 20(3):944–952. https://doi.org/10.1109/jbhi.2015.2420534
38. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson J, Schwartz JS, Shero ST, Smith SC, Sorlie P, Stone NJ, Wilson PWF (2013) 2013 ACC/AHA guideline on the assessment of cardiovascular risk. A report of the American College of Cardiology/American Heart Association Task Force on practice guidelines. https://doi.org/10.1161/01.cir.0000437741.48606.98
39. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP (2017) Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 24(1):198–208. https://doi.org/10.1093/jamia/ocw042
40. Batal I, Valizadegan H, Cooper GF, Hauskrecht M (2013) A temporal pattern mining approach for classifying electronic health record data. ACM Trans Intell Syst Technol 4(4). https://doi.org/10.1145/2508037.2508044
41. Batal I, Cooper GF, Fradkin D, Harrison J, Moerchen F, Hauskrecht M (2016) An efficient pattern mining approach for event detection in multivariate temporal data. Knowl Inf Syst 46(1):115–150. https://doi.org/10.1007/s10115-015-0819-6
42. Moskovitch R, Shahar Y (2015) Fast time intervals mining using the transitivity of temporal relations. Knowl Inf Syst 42(1):21–48. https://doi.org/10.1007/s10115-013-0707-x
43. Moskovitch R, Shahar Y (2009) Medical temporal-knowledge discovery via temporal abstraction. AMIA Annu Symp Proc 2009:452–456
44. Orphanou K, Stassopoulou A, Keravnou E (2014) Temporal abstraction and temporal Bayesian networks in clinical domains: a survey. Artif Intell Med 60(3):133–149. https://doi.org/10.1016/j.artmed.2013.12.007

45. Bellazzi R, Sacchi L, Concaro S (2009) Methods and tools for mining multivariate temporal data in clinical and biomedical applications. In: Conference proceedings: annual international conference of the IEEE engineering in medicine and biology society IEEE engineering in medicine and biology society annual conference 2009:5629–5632. https://doi.org/10.1109/iembs.2009.5333788
46. Miotto R, Li L, Kidd BA, Dudley JT (2016) Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 6:26094. https://doi.org/10.1038/srep26094