

Predicting coronary artery disease severity through genomic profiling and machine learning modelling: The GENetic SYNTAX Score (GESS) trial

Fani Chatzopoulou^{1,2}, Nikolaos Mittas³, Dimitrios Chatzidimitriou^{1,2}, Alexandros Giannopoulos-Dimitriou⁴, Aikaterini Saiti⁴, Maria Ganopoulou⁵, Efstratios Karagiannidis⁶, Andreas S. Papazoglou⁷, Nikolaos Stalikas⁷, Anna Papa¹, George Giannakoulas⁷, Georgios Sianos⁷, Lefteris Angelis⁵, Ioannis S. Vizirianakis^{4,8*}

¹Laboratory of Microbiology, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

²Labnet Laboratories, Thessaloniki, Greece

³Hephaestus Laboratory, School of Chemistry, Faculty of Sciences, Democritus University of Thrace, Kavala, Greece

⁴Laboratory of Pharmacology, School of Pharmacy, Aristotle University of Thessaloniki, Thessaloniki, Greece

⁵School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

⁶Second Department of Cardiology, Hippokration General Hospital, Aristotle University of Thessaloniki, Thessaloniki, Greece

⁷AHEPA University General Hospital of Thessaloniki, 1st Cardiology Department, Thessaloniki, Greece

⁸Department of Health Sciences, School of Life and Health Sciences, University of Nicosia, Nicosia, Cyprus

Running title: ML prediction of genetic SYNTAX score

Keywords: Coronary artery disease, CAD, SYNTAX score, SNPs, biomarkers, cardiovascular disorders, CVDs, risk prediction, machine learning, personalized (precision) medicine, pharmacogenomics

Type of manuscript: Original Research Article

Main text: 9,225 words, (title page, abstract, manuscript text, references, and figure legends)

Tables: 1

Figures: 5

* To whom correspondence should be addressed: Ioannis S. Vizirianakis, Laboratory of Pharmacology, School of Pharmacy, Aristotle University of Thessaloniki, Thessaloniki, Greece; Email: ivizir@pharm.auth.gr; vizirianakis.i@unic.ac.cy; Tel.: 0030 23 10 997658

Abstract

Cardiovascular diseases (CVDs) present multifactorial pathophysiology and produce immense health and economic burdens globally. The most common type, coronary artery disease (CAD), shows a complex etiology with multiple genetic variants to interplay with various clinical features and demographic traits affecting CAD risk and severity. The development and clinical validation of machine learning (ML) algorithms that integrate genetic biomarkers and clinical features can improve diagnostic accuracy for CAD avoiding, thereby, unnecessary invasive procedures. To this end, we present, here, the development of a data-driven ML approach able to predict the existence and severity of CAD based on the analysis of 228 single nucleotide polymorphisms (SNPs) and clinical and demographic data of 953 patients enrolled in the Genetic Syntax Score (GESS) trial (NCT03150680). Two competing ensemble models (one with clinical predictors and another with clinical plus genetic predictors) were built and evaluated to infer their prediction capabilities. The ensemble model with both clinical and genetic predictors exhibited superior diagnostic performance compared to the competing model with only clinical predictors. The proposed ML framework identified a total of eight contributing SNPs as predictors for the existence of obstructive CAD and seven significant SNPs for the severity of CAD. Such algorithms positively contributes to global efforts aiming to predict the risk and severity of CAD in early stages, thus lowering the cost as well as achieving prognostic, diagnostic, and therapeutic benefits in healthcare and improving patient outcomes in a non-invasive way. Overall, the design and execution of this trial reinforces clinical decision-making and facilitate the harmonization in digitized healthcare within the concept of precision medicine.

Clinical Trial Registration: NCT03150680;

<https://clinicaltrials.gov/study/NCT03150680?cond=NCT03150680&rank=1>

Introduction

Several machine learning (ML) approaches have already been applied in the clinical setting, including cardiovascular disorders (CVDs). In particular, the utility of ML methodologies has been shown in imaging practices used for the diagnosis of acute coronary syndrome (ACS) and in the prediction of coronary artery disease (CAD) severity. Indeed, emerging evidence suggests that ML methodologies may accurately predict CAD severity as well as short- and long-term outcomes in patients presenting with ACS¹⁻⁵. Specifically, ML algorithms have been applied in the CAD setting to: (i) predict the occurrence of obstructive CAD by evaluating multiple clinical variables and the coronary artery calcium score⁶; (ii) improve functional coronary assessment, as well as to detect lesion-specific ischemia by using computational fluid dynamics algorithms^{7,8}; (iii) estimate the pre-test probability of CAD⁹; and (iv) evaluate the automatic prediction of obstructive CAD from myocardial perfusion imaging^{10,11}.

A previous effort by our research group led to the development of a ML clinical risk-stratification framework aiming to the prediction of the severity of CAD based on the assessment of SYNTAX score¹²⁻¹⁵. By combining anatomic and clinical prognostic variables, the SYNTAX score is used to predict the prognostic course in patients with CAD and guide practitioners in choosing percutaneous coronary intervention (PCI) or coronary artery bypass graft (CABG) surgery. Higher scores are associated with more complex CAD¹³. At this end, we proposed an ensemble ML approach that integrates via a two-stage model both classification and regression techniques into a unified risk-score assessment process in order to model: (i) the existence of obstructive CAD through a binary classifier (patients with zero vs. non-zero SYNTAX score); and (ii) the severity of CAD (the expected SYNTAX score) through a regression model given that a patient is diagnosed with the existence of obstructive CAD (non-zero SYNTAX score).

The rationale behind this work was to further extend our previous research efforts through the leverage of genomic information hidden into single nucleotide polymorphisms (SNPs) analyzed using next generation sequencing (NGS) technology. The aim of the current study was to investigate whether introducing the genetic information of specific SNPs into the development of ML prediction models [i.e., including clinical biomarkers and demographic

parameters (Supplementary Table 1)] could bolster SYNTAX score prediction in relatively large patient populations. This effort was achieved by evaluating 228 SNPs (Supplementary Table 2, previously recognized as contributing factors to CAD from genome-wide association studies (GWAS), in 953 patients enrolled in the GESS trial (NCT03150680)¹³. The findings obtained contribute to non-invasive evaluation of CAD severity through ML approaches that allows practitioners very early to carry out personalized intervention in the clinical setting, stratification, and therapeutic guidance of their patients. It also provides a way in which the integration of genetic data of SNPs into ML models, may reinforce clinical decision-making and facilitate the harmonization in digitized healthcare¹⁶ in compliance with the American Heart Association recent statement on AI/ML use in healthcare¹⁷.

Methods and Materials

Clinical study design and data collection

The detailed design of the GESS trial (ClinicalTrials.gov Identifier: NCT03150680) has been previously published¹³. Adult patients undergoing invasive coronary angiography were enrolled in this trial after providing written informed consent. Two experienced interventionalists blinded to the study protocol assessed the obtained angiographic images, and, thereby, the SYNTAX score was calculated for every study participant. Patients with history of prior revascularization procedure and patients with cardiopulmonary arrest on admission were not deemed eligible for enrolment. Ethical approval was obtained from the Scientific Committee of the AHEPA University Hospital of Thessaloniki, Greece (reference number 309/11-05-2017); all trial procedures complied with the Declaration of Helsinki¹⁸.

Pre-specified clinical data, including demographic characteristics, medical history, clinical presentation (Supplementary Table 1) and medication, were recorded for the entire study population under standardized methods. Additionally, peripheral blood samples were drawn on the enrolment day -prior to coronary angiography- for genomic profiling. The vials of drawn blood were aliquoted and stored as whole blood, plasma, serum, and buffy coat, accordingly.

NGS analysis of SNP variations in patient DNA samples

Genomic DNA was extracted from blood using the QIAamp DNA Blood Midi kit (Qiagen) following the manufacturer's instructions. The extracted samples were quantified using Nanodrop 1000 (Thermo Scientific).

Library preparation

A custom-made panel consisting of 228 SNPs was used (Supplementary Table 2). SNP selection was based on their potential correlation to the pathophysiology of CVDs, as previously presented^{13,14}. DNA libraries were built using the QIAseq Targeted DNA Panel kit (Qiagen) following the manufacturer's protocol. In brief, DNA (40 ng) was enzymatically fragmented, end repaired and A-tailed followed by adapter ligation. These adapters (12 different indices, IL-N701–N715) are unique molecular indices (UMIs) that integrate on each samples sequence. Then, double cleanup was performed using QIAseq beads and freshly prepared ethanol. Following the cleanup, target enrichment was performed using one region-specific primer and one universal primer complementary to the adapter. After a cleanup step, a universal PCR was carried out to amplify the library and add index primer (IL-S502–S511). The reaction was performed directly in the QIAseq 96-index I set A plate, that contains sample index primer and universal PCR primers. After completion of universal PCR, a final cleanup step using QIAseq beads was performed.

Sequencing and data analysis

Amplified libraries were quantified using the Qubit dsDNA HS (High Sensitivity) assay kit and the Qubit 2.0 Fluorometer (Thermo Fisher Scientific). In addition to fluorometry, libraries were further quantified by qPCR, using the Universal qPCR Master Mix (KAPA Library Quantification Kit) for Illumina Platforms in two working dilutions. The libraries were normalized, pooled, and diluted to 1 nM. After denaturation with 0.1N NaOH the pool was further diluted to 1.3 pM. The library was loaded onto a Mid Output cartridge on the MiniSeq system (Illumina) according to manufacturer's instructions and sequenced using paired-ends (2×151 bp) and the QIAseq custom primer Read 1 provided with the QIAseq library kit. Upon completion of the sequencing run, FASTQ files were downloaded for further bioinformatic analysis. NGS data (FASTQ files) were analyzed using the CLC Genomics Workbench v.20.0.3 (Qiagen) with the Biomedical

Genomics Analysis Plugin v. 20.0.1. A custom workflow has been created for trimming and annotation of UMIs, mapping to hg19 (reference genome), removal of ligation artefacts, detection, and annotation of indels and structural variants, and local realignment. The identified variants were called, and the three different genotypes (homozygous for the reference allele, heterozygous, homozygous for the alteration) were recognized.

Machine learning framework

The data-driven ML framework along with its core elements aiming at the prediction of SYNTAX score is illustrated in Fig. 1. The methodology follows previous efforts of our group^{12,15}, involving nine phases with important actions and decisions that should be cautiously considered to design and implement a ML solution with the aim of efficiently achieving the goal of CAD risk and severity prediction at real time and non-invasively in the point-of-care. Following the structured process for the design, development and deployment of any ML solution, the workflow of the proposed approach can be summarized into the following nine phases: (i) Problem Definition, (ii) Data Collection, (iii) Data Preparation, (iv) Data Exploration, (v) Model Selection, (vi) Feature Selection (vii) Model Building, (viii) Model Evaluation and (ix) Practical Implications of model deployment (Fig. 1). Note that the detailed description of the ML workflow along with its phases and decisions made for each phase are presented in the Supplementary data section (Supplementary note). In brief, the proposed ML solution adopts the idea of ensemble learning, where multiple models (base learners) co-operate with the aim of achieving superior performance compared to any individual model. In our framework, the ensemble mechanism exploits (a) a classification base learner (or zero-part model) that is responsible for discriminating patients into two mutually exclusive groups (patients with SYNTAX score higher than zero versus patients with SYNTAX score equals to zero) and (b) a regression base learner (count-part model) that is dedicated to the prediction of the expected SYNTAX score given that a patient is classified into the non-zero group (SYNTAX score higher than zero).

Results

ML model functionality and practical utility

In this section, we mainly focus on the presentation of the results concerning the identification of the set of the important clinical- and genome- related factors (Feature Selection phase) that were used, in turn, as inputs into the fitting of the competing ensemble models (Model Building). Supplementary Fig. S1 presents the distribution of SYNTAX score for the total number of patients participated in the current study (red violin plot) compared to the SYNTAX score distribution for the subgroup of patients that presenting SYNTAX score higher than zero (blue violin plot). Also, Supplementary Table 3 summarizes the descriptive statistics for the selected clinical and demographic variables from various types of Electronic Health Records (EHRs).

Supplementary Tables 4 and 5 show the initial set of candidate risk-factors that exhibited a statistically significant effect on the SYNTAX score response variables for the zero- and count-part models, respectively. In the case of the zero-part model (Supplementary Table 4), the Likelihood Ratio (LR) test after the fitting of separate univariate logistic models for each candidate predictor indicated that a set of forty-one predictors (nineteen clinical risk-factors and twenty-two SNPs) seem to meet the first criterion for insertion into the second round of the feature selection protocol. Concerning the count-part model, thirty-three predictors (Supplementary Table 5) (twelve clinical risk-factors and twenty-one SNPs) were identified as significant risk-factors for further investigation into the second round of the feature selection protocol. At the second step, the Boruta algorithm was executed on the candidate sets of risk-factors that were extracted from the previous selection round to decide upon the final predictors that were inserted into the zero- and count-part models of the ensemble mechanism. The algorithm resulted into the variable importance measure (VIM) value for each predictor assessing the contribution of features to the modeling process of the training phase¹⁹. More specifically, the wrapper feature selection technique was applied, separately, for the subsets of clinical risk-factors and SNPs after omitting clinical predictors that either (a) presented a high proportion of missing observations that would lead, in turn, to well-known problems related to the fitting of multivariate models or (b) participated into the computation formulae of RATIO1 (monocyte to HDL cholesterol ratio), RATIO2 (lymphocyte to monocyte ratio), RATIO3 [atherogenic index of plasma levels, $\log(TG / HDL)$] and RATIO4 [liver cell injury

indicator, serum aspartate aminotransaminase (SGOT) to serum alanine aminotransaminase (SGPT) ratio], so as to mitigate the arisen multicollinearity issues, due to the inclusion of intercorrelated features.

The results related to the distributions of VIM after the execution of the Boruta algorithm on the qualified sets of the clinical and SNPs predictors for the zero- and count-part models are graphically displayed in Fig. 2a. From the total set of the nineteen clinical predictors, fourteen were confirmed as statistically significant for insertion into the zero-part of the model (upper left, green-colored boxplots in Fig. 2a). Additionally, the Boruta algorithm signified eight important SNPs (upper right, green-colored boxplots in Fig. 2a) that will be used for the model building phase of the zero-part model. Regarding the regression base learner (count-part model), nine out of twelve clinical risk factors (lower left, green-colored boxplots in Fig. 2a) and seven out of twenty-one SNPs (lower right, green-colored boxplots in Fig. 2a) were confirmed as informative.

The above sets of important clinical risk-factors and SNPs constituted the basis for the training of the two competing ensemble models, called Model A (ensemble model with clinical predictors) and Model B (ensemble model with clinical plus SNPs predictors). The examination of the performance metrics for the classification base learner (zero-part model) (Supplementary Table 6) indicates superior predictive capabilities for Model B that exploited both clinical and SNPs information for all performance indicators compared to Model A. The graphical inspection via receiver operating characteristics (ROC) analysis (Fig. 2b, left figure) that visualizes the trade-off between true positive rate (TPR) (or sensitivity) and false positive rate (FPR) (or 1-specificity) suggests that both classifiers present sufficient discrimination capacity compared to the baseline random classifier (points lying along the diagonal reference line, where FPR = TPR), whereas Model B seems to dominate Model A, since it is closer to the optimal top-left corner of the plot. Indeed, the computation of the area under the curve (AUC) measure for the two competing classifiers confirms that Model B presents a higher AUC value ($AUC_{ROC_{Model\ B}} = 0.798$) compared to the corresponding value derived from the model incorporating only the set of clinical risk-factors ($AUC_{ROC_{Model\ A}} = 0.757$), a difference that was statistically signified via the execution of the Delong's test ($Z = 2.451, p = 0.014$). Moreover, the investigation of the

performance capabilities of the two classifiers through precision-recall (PR) curves (Fig. 2b, right figure) showcases similar findings, since Model B is closer to the top-right corner presenting, again, higher AUC value ($AUCPR_{\text{Model B}} = 0.912$) in comparison with Model A ($AUCPR_{\text{Model A}} = 0.893$). We note that in the case of PR curves, the baseline model (horizontal reference line) represents the performance of a model classifying all instances to the positive class.

The performance evaluation concerning the competing regression base learners that are responsible for the prediction of the expected SYNTAX score for patients with obstructive CAD revealed, again, that Model B yielded superior prediction performance compared to Model A in terms of both MdAE and MdMRE values (Supplementary Table 6). The conduction of appropriate inferential mechanisms via the Wilcoxon Signed Rank test demonstrated statistically significant differences for the distributions of both the absolute error ($V = 9696, p = 0.008$) and the magnitude of relative error ($V = 10095, p < 0.001$). Finally, the comparison related to the prediction capabilities of the ensemble learning synergy between the classification and regression base learners showed, again, that the model with both clinical and SNPs predictors (Model B) outperformed Model A that leveraged only clinical risk-factors (Model A) ($V = 16731, p = 0.002$).

Network analysis of the important SNPs predictors of the zero- and count-part models

The genetic etiology of CAD includes multiple genetic variants associated with CAD risk and severity²⁰⁻²³. The eight and seven important SNPs (Table 1) that were inserted into the zero- and count-part models, respectively, were subjected to bioinformatics analysis to determine their corresponding genes accompanied by their genetic/protein interactors, as well as to unveil the association between a variant with human diseases and drug phenotype.

Regarding the zero-part model (classifier), three out of eight SNPs are located either within the intronic (rs11984041) or the 3' prime untranslated (3'-UTR) (rs2023938) or the intergenic region (rs2107595) of the *HDAC9* gene. The genetic variant rs2107595, localized between the *HDAC9* and *TWIST1* genes, is a known risk factor associated with CAD, large artery intracranial occlusive disease²⁴, and aortic calcification^{25,26}. Different studies, however, contradict whether

this SNP regulates the *HDAC9* or the *TWIST1* mRNA expression²⁶. In addition, rs3732379, rs1800562, and rs41291556 are nonsynonymous variants of the genes *CX3CR1*, *HFE*, and *CYP2C19*, respectively, causing alterations in the amino acid sequence of the corresponding proteins. Besides, the variant rs964184 is localized in the 3'-UTR of the *ZPR1* gene and the rs216172 variant inside the intronic region of the *SMG6* gene.

In the count-part model (regressor), six out of seven important SNPs are intronic. Specifically, (i) the variants rs4845625 and rs6689306 are localized in the *IL6R* gene; (ii) the variants rs2046934 and rs6801273 are localized in the overlapping genes (OLGs) *MED12L* and *P2RY12*; (iii) the variant rs870142 is located in the interval between *STX18* and *MSX1* and, according to a previous report, potentially affects the expression of the latter²⁷; and (iv) the variant rs1332844 which is located in the *PHACTR1* gene. Finally, rs663129 is an intergenic SNP located downstream of *PMAIP1* and *MC4R* genes.

According to the network analysis, based on DisGeNET database²⁸, the important SNP predictors of the zero-part model are mainly associated with CVDs, including CAD and atrial fibrillation (Fig. 3a). However, compared to the variant-disease network of the zero-part model, the SNP predictors of the count-part model participate in a more compact variant-disease network which highlights the strong association of these variants particularly with CVDs (Fig. 4a). The expert-curated interactions of the genes corresponding to the important SNPs predictors of the zero- and count-part models were retrieved from BioGRID database of protein, genetic and chemical interactions²⁹ and integrated into graphical network representations (Fig. 3b and Fig. 4b, respectively).

The bioinformatic analysis of the genes associated to the zero-part model revealed the following associations: (i) the *HDAC9* gene interacts with the non-selective histone deacetylase inhibitor panobinostat, and with anticonvulsive drug valproic acid (Fig. 3b); (ii) the Gene Ontology (GO) enrichment analysis based on Biological Process (BP) of the genes that interact with *HDAC9* revealed that the interacting genes are significantly enriched in processes involved in protein deacetylation, histone modification, regulation of gene silencing by RNA and nuclear transport (Supplementary Fig. S2); (iii) a large number of genes that interact with *HFE* are mainly involved in glycoprotein biosynthetic and metabolic process, (Supplementary Fig. S3);

(iv) the genes interacting with *ZPR1* are significantly enriched in biological processes involved in regulation of protein catabolic process and regulation of protein ubiquitination (Supplementary Fig. S4), whilst the variant rs964184 of *ZPR1* gene is associated with phenotypic modifications of the fenofibrate drug and antiretroviral agents via affecting *APOA1* and *APOA5* genes (Fig. 3b); (v) the genes that interact with *SMG6* are mainly enriched in biological processes related with RNA catabolic process, RNA localization/transport and regulation of translation (Supplementary Fig. S5); however, no drug interactions are recorded for the *SMG6* rs216172 (Fig. 3b); (vi) the gene *CYP2C19* which encodes a cytochrome P450 enzyme involved in the metabolism of xenobiotics³⁰ interacts with other genes of cytochrome P450 superfamily of enzymes (Fig. 3b) and; (vii) the SNP rs41291556 located in the *CYP2C19* gene affects the efficacy, toxicity and metabolism/pharmacokinetics of several drugs belonging to different pharmacologic classes including aspirin, clopidogrel, celecoxib, fluconazol etc. (Fig. 3b).

Concerning the count-part model, the significant associations extracted via the bioinformatics analysis are (i) the genes that interact with *MED12L* are significantly enriched in GO biological processes involved in the RNA biosynthetic processes and the initiation of DNA-dependent transcription initiation; the vast majority of the *MED12L* gene-interactors are subunit proteins of the Mediator complex which fundamental role is to communicate regulatory signals from DNA-bound transcription factors (TFs) to the RNA polymerase II (Pol II) enzyme³¹ (Supplementary Fig. S6). According to the PharmGKB database, the SNPs rs2046934 and rs6801273 mapped in the *MED12L* gene region are associated with phenotypic modifications of clopidogrel, aspirin, prasugrel (rs2046934), and clopidogrel (rs6801273), respectively (Fig. 4b); (ii) the genes that interact with *P2RY12* are involved mainly in oxidative phosphorylation biological process, and enrich pathways related to the respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins, as well as to the citric acid cycle and the mitochondrial biogenesis (Supplementary Fig. S7). Since the genes *MED12L* and *P2RY12* are OLGs in the human genome, the drug phenotype associated with the variants rs2046934 and rs6801273 are maintained the same as described above; (iii) the *IL6R* gene interacts with 25 genes, whilst rs4845625 affects the tocilizumab response of patients with

rheumatoid arthritis³² (Fig. 4b) and; (iv) the *PHACTR1* gene interacts with eight genes, amongst them *HIST1*, *H2BH*, *PPP1CA*, *EIF4B*, and *TRIM25* (Fig. 4b).

The clinical features and the genetic characteristics found to be associated in the developed data-driven ML framework for the prediction and severity of CAD are depicted in Fig. 5.

Discussion

The ML-based analysis presented herein is one of the first to demonstrate the enhanced predictive value of over 200 SNPs in addition to a clinically relevant predictive model for predicting angiographically confirmed obstructive CAD. Our results highlight that the combination of genetic information and clinical parameters, using a robust ML ensemble algorithm could effectively distinguish patients with and without obstructive CAD. As far as the genetic polymorphisms are concerned, the data obtained from the current study demonstrated for the first time eight out of 228 SNPs analyzed by NGS significantly associated with obstructive CAD occurrence and additional seven associated to the severity as expressed by higher SYNTAX score and, thus, increased CAD severity. To this end, the Boruta algorithm, implemented herein, yielded that adding these eight SNPs to a model consisting of fourteen clinical parameters could significantly bolster its predictive capability of obstructive CAD (zero-part model). Moreover, the addition of the second set of seven SNPs to a model consisting of nine clinically relevant risk-factors could also add to the predictive performance of the model for the prediction of the SYNTAX score (count-part model).

Regarding the zero-part model, the clinical variables associated with obstructive CAD were: patients' age, gender, smoking history, atrial fibrillation history, the presence of chest pain or atypical angina symptoms, fasting glucose levels, neutrophil, eosinophil and basophil percentages, the ratio of monocyte to HDL-cholesterol (RATIO1), the ratio of lymphocyte to monocyte (RATIO2), the atherogenic index of plasma levels (RATIO3), and the ratio of SGOT to SGPT; (RATIO4). Of those variables, the presence of CAD-related symptoms (typical or atypical), higher patient's age, gender, smoking history and higher lipidemic levels have been consistently used in most pre-test probability algorithms to serve as an effective gatekeeper for non-invasive testing and eliminate the likelihood of obstructive CAD without the need for invasive

procedures^{33,34}. The predictive significance of atrial fibrillation history has not been yet established; however, the overall incidence of obstructive CAD in patients presenting with atrial fibrillation is relatively high³⁵. Additionally, higher fasting glucose levels could predict the occurrence of obstructive CAD in a chronic setting even if the observed hyperglycemia might be stress-induced^{36,37} in patients with suspected ACS. Moreover, the observed association between elevated levels of almost all subtypes of white blood cell counts (neutrophil, eosinophil and basophil percentages, and the ratio of lymphocyte to monocyte) and increased risk of obstructive CAD has been previously demonstrated as well^{38,39}. This association might denote the underlying inflammation and/or hypersensitivity associated with CAD occurrence since leukocyte count is a widely available marker of inflammation.

As far as the count-part model of the Boruta algorithm is concerned, higher patients' age, history of diabetes mellitus, elevated fasting glucose and galectin-3 levels, higher CRUSADE score, lower hemoglobin levels and chronic kidney disease (lower GFR, higher urea and creatinine values) were all associated with higher SYNTAX score. Patients' age, diabetes mellitus history and dysglycemic status have been associated with increased CAD severity⁴⁰⁻⁴². On the other hand, lower hemoglobin levels may have been already linked with poor prognostic course; however, no study has linked yet anemia (a potential inflammation marker) with CAD severity⁴³. Furthermore, serum galectin-3 levels have been positively linked with CAD severity⁴⁴, and renal dysfunction has been also associated with more severe CAD⁴⁵ as demonstrated in our study. Finally, the positive association between CRUSADE and SYNTAX scores might be attributed to the already shown surrogate prognostic utility of both scores⁴⁶.

Several ML-based predictive models have been already created aiming to predict CAD occurrence based on traditional or non-traditional CVD risk factors⁴⁷. Promising approaches towards precision medicine currently investigate the predictive capacity of combined clinical and genomic data forced into ML-based algorithms. These investigations assessed genomic data (analysis of different SNPs) from real-world cohort studies⁴⁸⁻⁵³ or other registries such as U.K. BioBank^{54,55} and GWAS⁵⁶. Although, most genetic risk score models are based on the genotyping of dozens of genetic loci, however, many of them do not show high specificity for the diagnosis

of CAD⁴⁸ and some of them do not add significant predictive utility on top of clinical CVD risk factors⁵³.

As it was revealed from the bioinformatic and network analysis presented in this work, for the zero-part model, the significant eight SNPs, rs3732379, rs1800562, rs11984041, rs2023938, rs2107595, rs41291556, rs964184, and rs216172, have been previously associated with CAD, cerebrovascular events, hypertension, and diabetes mellitus⁵⁷. Notably, three (rs11984041, rs2107595 and rs2023938) of the most important SNPs identified are associated with *HDAC9* gene that participates in deacetylation of histones, and, thus, regulates gene transcription⁵⁸. Especially rs11984041 has been associated with large vessel occlusion in acute stroke²⁴ and is in linkage disequilibrium with rs2107595. Both SNPs are found to elevate the risk for ischemic stroke via promoting carotid atherosclerosis⁵⁹. However, a latter study in 733 large artery atherosclerotic stroke patients from China failed to confirm the significance of rs11984041⁶⁰. In addition, the CARDIoGRAMplusC4D Consortium study identified 15 new susceptibility loci for CAD, amongst them rs2023938⁶¹. In a previous study of the same consortium rs964184 and rs216172 were denoted as significant susceptibility loci for CAD with risk allele frequencies of 0.13 and 0.37, respectively⁶². Of importance is also the rs1800562 in *HFE* gene, where a Mendelian randomization study highlighted its association between higher iron status and reduced CAD risk⁶³.

Several GWASs performed between 2010 and 2015, as well as meta-analysis of their data, revealed several risk loci associated with CAD^{22,27,61,64}. From our analysis, in the count-part model, the significant seven SNPs, rs6689306, rs4845625, rs2046934, rs6801273, rs870142, rs1332844, rs1332844 and rs663129, have been previously shown to be associated with CVDs, including CAD and atrial fibrillation. For instance, variants of the *IL6R* gene are associated with common diseases including inflammatory diseases such as CVDs⁶⁵ and type 2 diabetes mellitus⁶⁶. Specifically, rs6689306 is associated with atrial fibrillation, while rs4845625 is associated with the risk of CVDs including atrial fibrillation⁶⁷, CAD⁶¹, and aortic aneurysm⁶⁵. Two of the significant SNPs (rs2046934, rs6801273) are located at the overlapping genes *MED12L* and *P2RY12*. The latter is one of the most important pharmacological antiplatelet aggregation drug targets of crucial clinical value for CVD patients, where multiple SNPs, as well as different

haplotypes, have been studied in terms of their significance with the risk of restenosis⁶⁸, high on-treatment platelet reactivity⁶⁹ and CAD⁷⁰. The intergenic rs663129 is also associated with higher risk for CAD and diabetes mellitus, and with obesity-associated risk factors including lower HDL and higher triglyceride concentrations^{71,72}. The rs870142 is involved in congenital heart diseases including ostium secundum atrial septal defect^{27,73}. Finally, rs1332844, an intronic variant of the gene *PHACTR1* is also associated with CAD risk⁷⁴ and major adverse cardiovascular events⁷⁵.

In our study, the addition of SNPs to a clinical predictive model (Model A) offered statistically significant yet maybe not clinically relevant improvement in the diagnostic indices. This could be attributed to the fact that clinical CVD risk factors (i.e., dyslipidemia, diabetes mellitus, hypertension, obesity, and family history of CAD) *per se* represent the disease state and are often accompanied by multiple genetic correlations⁵⁶. Dysregulation in any of those clinical parameters is probably a comprehensive manifestation of many genetic variations rather than the outcome of individual SNPs. Genetic variations eventually cause disease through CAD complexity (i.e., disturbed cellular balance, accumulation of harmful substances and damaged genetic cellular information). Hence, individuals with higher genetic risk scores are at higher risk of suffering from genotoxicity, which is not inevitable though therapeutic lifestyle change⁷⁶, indicating that it must be the combined effect of various parameters determining the final phenotype, and CAD severity.

For the developed ML-risk stratification model to gain broad clinical applicability and practical utility, the external validation of our findings along with their clinical applicability in different patient populations must be demonstrated. To this end and in parallel to external validation, a definite feature analysis must be performed to allow practical scoring and interpretation of each predictor in a manner useful for clinicians in determining the risk of obstructive CAD. Also, future work must provide an explicit explanation for each predictor, thus, allowing the healthcare practitioner to read the result and decide according to the optimal cut point.

Despite the high-performance that our ML-risk stratification model shows in predicting the risk and severity of CAD by evaluating genetic, clinical, and demographic data, some study

limitations warrant mention. Firstly, our data derived from a single center of similar ethnicity and the sample size was relatively small limiting the generalizability and universality of our findings. Secondly, this was a retrospective analysis of a prospectively designed study. Although the models could accurately classify individuals into obstructive and non-obstructive CAD, the predictive power of the models requires further validation in larger prospective studies.

Conclusion

This study emphasizes that integrating clinical information with specific SNPs through ML can potentially assist in risk stratification for patients with suspected CAD. The selection of features required to build our ML-risk stratification model is likely to be the key determinant of its clinical performance. The addition of the genetic information to the clinical parameters further improves the performance of the model to predict the risk of CAD. Subsequently, the combination of such features can be used to build high-performance models for the prediction of CAD severity. Such an approach may enable healthcare professionals to identify patients not only at an increased risk for CAD but also for more complex forms of the disease, thereby facilitating targeted therapeutic interventions and preventive measures. In any case, however, this work also addresses the existing challenges in developing digitized precision medicine interventions related to the design of the trials, the data acquisition manner, as well as the clinical translation and implementation of molecular knowledge in healthcare. The development of such predictive models is envisaged to improve the accuracy and effectiveness of CVD prediction, ultimately leading to better patient outcomes broadly, positively affecting societal discrepancies and ultimately to a healthier population globally.

Acknowledgements

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Programme Competitiveness, Entrepreneurship, and Innovation, under the call RESEARCH—CREATE—INNOVATE (project code: T1EDK-02354).

Conflicts of Interest statement

Fani Chatzopoulou is employed by Labnet Laboratories. Dimitrios Chatzidimitriou is CEO of Labnet Laboratories. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

SNPs included in the ML model are covered by the international patent application No PCT/GR2022/000067: Development of “GESScore Calculator” as predictive risk tool of cardiovascular events by the implementation of an algorithm using genetic factors and the complexity of coronary disease (International filing date: 30 November 2022) and the Greek patent Hellenic Industrial Property Organisation (OBI) - efile number: 2410-0004615550 OBI-Ref. Num. 24652/2022 (efiling date: 30 November 2022).

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Authors' contribution

Conceptualization of the study and ML model development: NM, LA, FC, ISV; Clinical study and data evaluation: GS, EK, ASP, NS; NGS analysis and data curation: FC, DC, ISV. Network and bioinformatics analysis: AD-G, AS, ISV; ML model validation and data interpretation: All authors contributed; Writing of the manuscript: All authors contributed; Supervision of the study: ISV; All authors have read and approved the final version of the manuscript.

Note

Part of the data included in this article has been presented in an abstract in the “American Heart Association Scientific Sessions 2023” held in Philadelphia, PA, USA, 11-13 November 2023 (Chatzopoulou F, Mittas N, Karagiannidis E, Papazoglou AS, Stalikas N, Moysidis D, Giannopoulos-Dimitriou A, Saiti A, Ganopoulou M, Papa A, Chatzidimitriou D, Giannakoulas G, Sianos G, Angelis L, Vizirianakis IS. Abstract 17570: Combining genomic profiling and clinical data through machine learning modeling for the prediction of coronary artery disease severity: Insights from the GEnetic SYNTAX Score (GESS) trial. Circulation, 2023;148:A17570, published 6

502 Nov 2023); as well as in the 44th Panhellenic Congress of Cardiology held in Thessaloniki,
503 Greece, 12-14 October 2023.
504

Supplementary Material

List of Supplementary Tables

Supplementary Table 1. Selected clinical and demographic variables (features) from various types of Electronic Health Records (EHRs).

Supplementary Table 2. SNPs studied and their characteristics.

Supplementary Table 3. Descriptive statistics of elected clinical and demographic variables.

Supplementary Table 4. Results of statistical hypothesis testing procedures (first round of feature selection) for the subset of predictors with statistically significant effect on the distribution of SYNTAX score for the zero-part model.

Supplementary Table 5. Results of statistical hypothesis testing procedures (first round of feature selection) for the subset of predictors with statistically significant effect on the distribution of SYNTAX score for the count-part model.

Supplementary Table 6. Evaluation of prediction performances for competing models.

List of Supplementary Figures

Supplementary Fig. S1. Distribution of SYNTAX score for the patients of the study (red) and for the patients with SYNTAX score higher than zero (blue).

Supplementary Fig. S2. Gene ontology (GO) and pathway enrichment analysis of the genes interacting with HDAC9 gene, according to BioGRID database.

Supplementary Fig. S3. Gene ontology (GO) and pathway enrichment analysis of the genes interacting with HFE gene, according to BioGRID database.

Supplementary Fig. S4. Gene ontology (GO) and pathway enrichment analysis of the genes interacting with ZPR1 gene, according to BioGRID database.

Supplementary Fig. S5. Gene ontology (GO) and pathway enrichment analysis of the genes interacting with SMG6 gene, according to BioGRID database.

Supplementary Fig. S6. Gene ontology (GO) and pathway enrichment analysis of the genes interacting with MED12L gene, according to BioGRID database.

Supplementary Fig. S7. Gene ontology (GO) and pathway enrichment analysis of the genes interacting with P2RY12 gene, according to BioGRID database.

Supplementary Note

Machine Learning Framework

Bioinformatic and Network Analysis

References

1. Gibson WJ, Nafee T, Travis R, Yee M, Kerneis M, Ohman M, et al. Machine learning versus traditional risk stratification methods in acute coronary syndrome: a pooled randomized clinical trial analysis. *J Thromb Thrombolysis* 2020;49:1–9. [https://doi: 10.1007/s11239-019-01940-8](https://doi.org/10.1007/s11239-019-01940-8).
2. Lee W, Lee J, Woo SI, Choi SH, Bae JW, Jung S. et al. Machine learning enhances the performance of short and long-term mortality prediction model in non-ST-segment elevation myocardial infarction. *Sci Rep* 2021;11:12886. [https://doi: 10.1038/s41598-021-92362-1](https://doi.org/10.1038/s41598-021-92362-1).
3. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from Imbalanced Data Sets. Springer Nature, 2018. [https://doi:10.1007/978-3-319-98074-4](https://doi.org/10.1007/978-3-319-98074-4).
4. D’Ascenzo F, De Filippo O, Gallone G, Mittone G, Deriu MA, Iannaccone M. et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *The Lancet* 2021;397:199–207. [https://doi: 10.1016/S0140-6736\(20\)32519-8](https://doi.org/10.1016/S0140-6736(20)32519-8).
5. Doudesis D, Lee KK, Boeddinghaus J, Bularga A, Ferry AV, Tuck C. et al. Machine learning for diagnosis of myocardial infarction using cardiac troponin concentrations. *Nat Med* 2023;29:1201–1210. [https://doi: 10.1038/s41591-023-02325-4](https://doi.org/10.1038/s41591-023-02325-4).
6. Al’Aref SJ, Maliakal G, Singh G, van Rosendael AR, Ma X, Xu Z, Alawamlh OAH. et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. *Eur Heart J* 2020;41:359–367. [https://doi: 10.1093/eurheartj/ehz565](https://doi.org/10.1093/eurheartj/ehz565).
7. Coenen A, Kim YH, Kruk M, Tesche C, De Geer J, Kurata A. et al. Diagnostic Accuracy of a Machine-Learning Approach to Coronary Computed Tomographic Angiography–Based Fractional Flow Reserve. *Circ Cardiovasc Imaging* 2018;11(6):e007217. [https://doi: 10.1161/CIRCIMAGING.117.007217](https://doi.org/10.1161/CIRCIMAGING.117.007217).
8. Tang CX, Liu CY, Lu MJ, Schoepf UJ, Tesche C, Bayer RR 2nd. et al. CT FFR for Ischemia-Specific CAD With a New Computational Fluid Dynamics Algorithm. *JACC Cardiovasc Imaging* 2020;13:980–990. [https://doi: 10.1016/j.jcmg.2022.07.012](https://doi.org/10.1016/j.jcmg.2022.07.012).
9. Hou Z, Lu B, Li ZN, An YQ, Gao Y, Yin WH. et al. Machine Learning for Pretest Probability of Obstructive Coronary Stenosis in Symptomatic Patients. *JACC Cardiovasc Imaging* 2019;12:2584–2586. [https://doi: 10.1016/j.jcmg.2019.07.030](https://doi.org/10.1016/j.jcmg.2019.07.030).
10. Betancur J, Commandeur F, Motlagh M, Sharir T, Einstein AJ, Bokhari S. et al. Deep Learning for Prediction of Obstructive Disease From Fast Myocardial Perfusion SPECT. *JACC Cardiovasc Imaging* 2018;11:1654–1663. [https://doi: 10.1016/j.jcmg.2018.01.020](https://doi.org/10.1016/j.jcmg.2018.01.020).
11. Betancur J, Hu LH, Commandeur F, Sharir T, Einstein AJ, Fish MB. et al. Deep Learning Analysis of Upright-Supine High-Efficiency SPECT Myocardial Perfusion Imaging for Prediction of Obstructive Coronary Artery Disease: A Multicenter Study. *J Nucl Med* 2019; 60:664–670. [https://doi: 10.2967/jnumed.118.213538](https://doi.org/10.2967/jnumed.118.213538).

12. Mittas N, Chatzopoulou F, Karagiannidis E, Chatzidimitriou D, Sianos G, Angelis L. et al. CRISPAC: A web-based platform for predicting the SYNTAX Score and severity of coronary artery disease. *SoftwareX* 2023;21:101310. <https://doi.org/10.1016/j.softx.2023.101310>.
13. Vizirianakis IS, Chatzopoulou F, Papazoglou AS, Karagiannidis E, Sofidis G, Stalikas N. et al. The GENetic Syntax Score: a genetic risk assessment implementation tool grading the complexity of coronary artery disease—rationale and design of the GESS study. *BMC Cardiovasc Disord* 2021;21(1):284. <https://doi.org/10.1186/s12872-021-02092-5>.
14. Chatzopoulou F, Kyritsis KA, Papagiannopoulos CI, Galatou E, Mittas N, Theodoroula NF. et al. Dissecting miRNA–Gene Networks to Map Clinical Utility Roads of Pharmacogenomics-Guided Therapeutic Decisions in Cardiovascular Precision Medicine. *Cells* 2022;11:607. <https://doi.org/10.3390/cells11040607>.
15. Mittas N, Chatzopoulou F, Kyritsis KA, Papagiannopoulos CI, Theodoroula NF, Papazoglou AS. et al. A Risk-Stratification Machine Learning Framework for the Prediction of Coronary Artery Disease Severity: Insights From the GESS Trial. *Front Cardiovasc Med* 2022;8:812182. <https://doi.org/10.3389/fcvm.2021.812182>.
16. Collin CB, Gebhardt T, Golebiewski M, Karaderi T, Hillemanns M, Khan FM. et al. Computational Models for Clinical Applications in Personalized Medicine—Guidelines and Recommendations for Data Integration and Model Validation. *J Pers Med* 2022;12(2):166. <https://doi.org/10.3390/jpm12020166>.
17. Aroundas AA, Narayan SM, Arnett DK, Spector-Bagdady K, Bennett DA, Celi LA. et al. Use of Artificial Intelligence in Improving Outcomes in Heart Disease: A Scientific Statement From the American Heart Association. *Circulation* 2024;149(14):e1028-e1050. <https://doi.org/10.1161/CIR.0000000000001201>.
18. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 2013;310(20):2191-2194. <https://doi.org/10.1001/jama.2013.281053>.
19. Kursu MB, Rudnicki WR. Feature Selection with the Boruta Package. *J Stat Softw* 2010; 36(11):1–13. <https://doi.org/10.18637/jss.v036.i11>.
20. Kessler T, Schunkert H. Coronary Artery Disease Genetics Enlightened by Genome-Wide Association Studies. *JACC Basic Transl Sci* 2021;6:610–623. <https://doi.org/10.1016/j.jacbts.2021.04.001>.
21. Erdmann J, Kessler T, Munoz Venegas L, Schunkert H. A decade of genome-wide association studies for coronary artery disease: The challenges ahead. *Cardiovasc Res* 2018;114(9):1241-1257. <https://doi.org/10.1093/cvr/cvy084>.
22. The Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat Genet* 2011; 43: 339–344. <https://doi.org/10.1038/ng.782>.
23. Soranzo N, Spector TD, Mangino M, Kühnel B, Rendon A, Teumer A. et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet* 2009;41:1182–1190. <https://doi.org/10.1038/ng.467>.

24. International Stroke Genetics Consortium (ISGC), Wellcome Trust Case Control Consortium 2 (WTCCC2), Bellenguez C, Bevan S, Gschwendtner A, Spencer CC. et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat Genet* 2012;44:328–333. [https://doi: 10.1038/ng.1081](https://doi.org/10.1038/ng.1081).
25. Malhotra R, Mauer AC, Lino Cardenas CL, Guo X, Yao J, Zhang X. et al. HDAC9 is implicated in atherosclerotic aortic calcification and affects vascular smooth muscle cell phenotype. *Nat Genet* 2019;51:1580–1587. [https://doi: 10.1038/s41588-019-0514-8](https://doi.org/10.1038/s41588-019-0514-8).
26. Ma L, Bryce NS, Turner AW, Di Narzo AF, Rahman K, Xu Y. et al. The HDAC9-associated risk locus promotes coronary artery disease by governing TWIST1. *PLoS Genet* 2022;18(6):e1010261. [https://doi: 10.1371/journal.pgen.1010261](https://doi.org/10.1371/journal.pgen.1010261).
27. Cordell HJ, Bentham J, Topf A, Zelenika D, Heath S, Mamasoula C. et al. Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nat Genet* 2013;45:822–824. [https://doi: 10.1038/ng.2637](https://doi.org/10.1038/ng.2637).
28. Piñero J, Saüch J, Sanz F, Furlong LI. The DisGeNET cytoscape app: Exploring and visualizing disease genomics data. *Comput Struct Biotechnol J* 2021;19:2960–2967. [https://doi: 10.1016/j.csbj.2021.05.015](https://doi.org/10.1016/j.csbj.2021.05.015).
29. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A. et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* 2021;30:187–200. [https://doi: 10.1002/pro.3978](https://doi.org/10.1002/pro.3978).
30. Shubbar Q, Alchakee A, Issa KW, Adi AJ, Shorbagi AI, Saber-Ayad M. From genes to drugs: CYP2C19 and pharmacogenetics in clinical practice. *Front Pharmacol* 2024;15:1326776. [https://doi: 10.3389/fphar.2024.1326776](https://doi.org/10.3389/fphar.2024.1326776).
31. Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol* **16**, 155–166 (2015).
32. Sainz L, Riera P, Moya P, Bernal S, Casademont J, Díaz-Torné C. et al. Role of IL6R Genetic Variants in Predicting Response to Tocilizumab in Patients with Rheumatoid Arthritis. *Pharmaceutics* 2022;14(9):1942. [https://doi: 10.3390/pharmaceutics14091942](https://doi.org/10.3390/pharmaceutics14091942).
33. Di Carli MF, Gupta A. Estimating Pre-Test Probability of Coronary Artery Disease. *JACC Cardiovasc Imaging* 2019;12:1401–1404. [https://doi: 10.1016/j.jcmg.2018.04.036](https://doi.org/10.1016/j.jcmg.2018.04.036).
34. Mincarone P, Bodini A, Tumolo MR, Vozzi F, Rocchiccioli S, Pelosi G. et al. Discrimination capability of pretest probability of stable coronary artery disease: a systematic review and meta-analysis suggesting how to improve validation procedures. *BMJ Open* 11(7):e047677. [https://doi: 10.1136/bmjopen-2020-047677](https://doi.org/10.1136/bmjopen-2020-047677).
35. Kravlev S, Schneider K, Lang S, Süselbeck T, Borggrefe, M. Incidence and Severity of Coronary Artery Disease in Patients with Atrial Fibrillation Undergoing First-Time Coronary Angiography. *PLoS One* 2011;6(9):e24964. [https://doi: 10.1371/journal.pone.0024964](https://doi.org/10.1371/journal.pone.0024964).

36. Stalikas N, Karagiannidis E, Papazoglou AS, Panteris E, Didagelos M, Ziakas A. et al. Added prognostic value of stress-induced hyperglycemia to the GRACE 2.0 risk score for prediction of 1-year major adverse cardiovascular events in patients with ST-elevation myocardial infarction. *Hellenic Journal of Cardiol* 2023;73:81-83. doi: 10.1016/j.hjc.2023.04.002. doi:10.1016/j.hjc.2023.04.002.
37. Stalikas N, Papazoglou AS, Karagiannidis E, Panteris E, Moysidis D, Daios S. et al. Association of stress induced hyperglycemia with angiographic findings and clinical outcomes in patients with ST-elevation myocardial infarction. *Cardiovasc Diabetol* 2022;21(1):140. [https://doi: 10.1186/s12933-022-01578-6](https://doi.org/10.1186/s12933-022-01578-6).
38. Kounis NG, Soufras GD, Tsigkas G, Hahalis G. White Blood Cell Counts, Leukocyte Ratios, and Eosinophils as Inflammatory Markers in Patients With Coronary Artery Disease. *Clin Appl Thromb Hemost* 2015;21:139–143. [https://doi: 10.1177/1076029614531449](https://doi.org/10.1177/1076029614531449).
39. Chen H, Li M, Liu L, Dang X, Zhu D, Tian G. et al. Monocyte/lymphocyte ratio is related to the severity of coronary artery disease and clinical outcome in patients with non-ST-elevation myocardial infarction. *Medicine* 2019;98:e16267. [https://doi: 10.1097/MD.00000000000016267](https://doi.org/10.1097/MD.00000000000016267).
40. Strisciuglio T, Izzo R, Barbato E, Di Gioia G, Colaioni I, Fiordelisi A. et al. Insulin Resistance Predicts Severity of Coronary Atherosclerotic Disease in Non-Diabetic Patients. *J Clin Med* 2020;9:2144. [https://doi: 10.3390/jcm9072144](https://doi.org/10.3390/jcm9072144).
41. Garg N, Moorthy N, Kapoor A, Tewari S, Kumar S, Sinha A. et al. Hemoglobin A1c in Nondiabetic Patients: An Independent Predictor of Coronary Artery Disease and Its Severity. *Mayo Clin Proc* 2014;89:908–916. [https://doi: 10.1016/j.mayocp.2014.03.017](https://doi.org/10.1016/j.mayocp.2014.03.017).
42. Karagiannidis E, Moysidis DV, Papazoglou AS, Panteris E, Deda O, Stalikas N. et al. Prognostic significance of metabolomic biomarkers in patients with diabetes mellitus and coronary artery disease. *Cardiovasc Diabetol* 2022;21(1):70. [http://doi: 10.1186/s12933-022-01494-9](http://doi.org/10.1186/s12933-022-01494-9).
43. Arant CB, Wessel TR, Olson MB, Bairey Merz CN, Sopko G. et al. Hemoglobin level is an independent predictor for adverse cardiovascular outcomes in women undergoing evaluation for chest pain. *J Am Coll Cardiol* 2004;43:2009–2014. [https://doi: 10.1016/j.jacc.2004.01.038](https://doi.org/10.1016/j.jacc.2004.01.038).
44. Li M, Guo K, Huang X, Feng L, Yuan Y, Li J. et al. Association Between Serum Galectin-3 Levels and Coronary Stenosis Severity in Patients With Coronary Artery Disease. *Front Cardiovasc Med* 2022;9:818162. [http://doi: 10.3389/fcvm.2022.818162](http://doi.org/10.3389/fcvm.2022.818162).
45. Kiyosue A, Hirata Y, Ando J, Fujita H, Morita T, Takahashi M. et al. Relationship Between Renal Dysfunction and Severity of Coronary Artery Disease in Japanese Patients. *Circulation J* 2010;74(4):786–791. [http://doi: 10.1253/circj.cj-09-0715](http://doi.org/10.1253/circj.cj-09-0715).
46. Atique S, Schultz C, Rankin J, Knuiman M, Nguyen M, Newman M. et al. The CRUSADE score is useful in stratifying risk of major bleeding and death following STEMI PCI. *Heart Lung Circ* 2015;24:S306–S307. <http://dx.doi.org/10.1016/j.hlc.2015.06.455>

47. Kim J, Lee SY, Cha BH, Lee W, Ryu J, Chung YH . et al. Machine learning models of clinically relevant biomarkers for the prediction of stable obstructive coronary artery disease. *Front Cardiovasc Med* 2022; 9:933803. <https://doi:10.3389/fcvm.2022.933803>.
48. Liu B, Fang L, Xiong Y, Du Q, Xiang Y, Chen X. et al. A Machine Learning Model Based on Genetic and Traditional Cardiovascular Risk Factors to Predict Premature Coronary Artery Disease. *Front Biosci (Landmark Ed)* 2022;27(7):211. <https://doi:10.31083/j.fbl2707211>.
49. Vaara S, Tikkanen E, Parkkonen O, Lokki ML, Ripatti S, Perola M. et al. Genetic Risk Scores Predict Recurrence of Acute Coronary Syndrome. *Circ Cardiovasc Genet* 2016;9:172–178. <https://doi:10.1161/CIRCGENETICS.115.001271>.
50. Natarajan P, Young R, Stitzel NO, Padmanabhan S, Baber U, Mehran R. et al. Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation* 2017;135:2091–2101. <https://doi:10.1161/CIRCULATIONAHA.116.024436>.
51. Pattarabanjird T, Cress C, Nguyen A, Taylor A, Bekiranov S, McNamara C. et al. A Machine Learning Model Utilizing a Novel SNP Shows Enhanced Prediction of Coronary Artery Disease Severity. *Genes (Basel)* 2020;11:1446. <https://doi:10.3390/genes11121446>.
52. Tikkanen E, Havulinna AS, Palotie A, Salomaa V, Ripatti S. Genetic Risk Prediction and a 2-Stage Risk Screening Strategy for Coronary Heart Disease. *Arterioscler Thromb Vasc Biol* 2013;33:2261–2266. <https://doi:10.1161/ATVBAHA.112.301120>.
53. Beaney KE, Cooper JA, Drenos F, Humphries S E. Assessment of the clinical utility of adding common single nucleotide polymorphism genetic scores to classical risk factor algorithms in coronary heart disease risk prediction in UK men. *Clin Chem Lab Med* 2017;55(10):1605-1613. <https://doi:10.1515/cclm-2016-0984>.
54. Howe LJ, Dudbridge F, Schmidt AF, Finan C, Denaxas S, Asselbergs FW. et al. Polygenic risk scores for coronary artery disease and subsequent event risk amongst established cases. *Hum Mol Genet* 2020;29:1388–1395. <https://doi:10.1093/hmg/ddaa052>.
55. Manduchi E, Le TT, Fu W, Moore JH. Genetic Analysis of Coronary Artery Disease Using Tree-Based Automated Machine Learning Informed By Biology-Based Feature Selection. *IEEE/ACM Trans Comput Biol Bioinform* 2022;19(3):1379-1386. <https://doi:10.1109/TCBB.2021.3099068>.
56. LeBlanc M, Zuber V, Andreassen BK, Witoelar A, Zeng L, Bettella F. et al. Identifying Novel Gene Variants in Coronary Artery Disease and Shared Genes With Several Cardiovascular Risk Factors. *Circ Res* 2016;118:83–94. <https://doi:10.1161/CIRCRESAHA.115.306629>.
57. Sun Z, Pan X, Tian A, Surakka I, Wang T, Jiao X . et al. Genetic variants in HFE are associated with non-alcoholic fatty liver disease in lean individuals. *JHEP Rep* 2023;5(7):100744. <https://doi:10.1016/j.jhepr.2023.100744>.

58. Haberland M, Montgomery RL, Olson EN. The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nat Rev Genet* 2009;10:32–42. [https://doi: 10.1038/nrg2485](https://doi.org/10.1038/nrg2485).
59. Markus HS, Mäkelä KM, Bevan S, Raitoharju E, Oksala N, Bis JC. et al. Evidence HDAC9 Genetic Variant Associated With Ischemic Stroke Increases Risk via Promoting Carotid Atherosclerosis. *Stroke* 2013;44:1220–1225. [https://doi: 10.1161/STROKEAHA.111.000217](https://doi.org/10.1161/STROKEAHA.111.000217).
60. Liu J, Hu Z, Chen R, Yang H, Zheng W, Liu D. et al. Gene polymorphism of rs556621 but not rs11984041 is associated with the risk of large artery atherosclerotic stroke in a Xinjiang Uyghur population. *J Stroke Cerebrovasc Dis* 2014;23(10):2641–2645. [https://doi: 10.1016/j.jstrokecerebrovasdis.2014.06.015](https://doi.org/10.1016/j.jstrokecerebrovasdis.2014.06.015).
61. ARdioGRAMplusC4D Consortium; Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL. et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2013;45(1):25–33. [https://doi: 10.1038/ng.2480](https://doi.org/10.1038/ng.2480).
62. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* 2011;43(4):333–8. [https://doi: 10.1038/ng.784](https://doi.org/10.1038/ng.784).
63. Gill D, Del Greco M F, Walker AP, Srai SKS, Laffan MA, Minelli C. et al. The Effect of Iron Status on Risk of Coronary Artery Disease. *Arterioscler Thromb Vasc Biol* 2017;37:1788–1792. [https://doi: 10.1161/ATVBAHA.117.309757](https://doi.org/10.1161/ATVBAHA.117.309757).
64. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S. et al. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* 2015;47:1121–1130. [https://doi: 10.1038/ng.3396](https://doi.org/10.1038/ng.3396).
65. Swerdlow DI, et al. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet* 2012;379(9822):1214–1224. [https://doi: 10.1016/S0140-6736\(12\)60110-X](https://doi.org/10.1016/S0140-6736(12)60110-X).
66. Hamid YH, Urhammer SA, Jensen DP, Glümer C, Borch-Johnsen K, Jørgensen T. et al. Variation in the Interleukin-6 Receptor Gene Associates With Type 2 Diabetes in Danish Whites. *Diabetes* 2004;53:3342–3345. [https://doi: 10.2337/diabetes.53.12.3342](https://doi.org/10.2337/diabetes.53.12.3342).
67. Schnabel RB, Kerr KF, Lubitz SA, Alkylbekova EL, Marcus GM, Sinner MF. et al. Large-Scale Candidate Gene Analysis in Whites and African Americans Identifies IL6R Polymorphism in Relation to Atrial Fibrillation. *Circ Cardiovasc Genet* 2011;4:557–564. [https://doi: 10.1161/CIRCGENETICS.110.959197](https://doi.org/10.1161/CIRCGENETICS.110.959197).
68. Rudež G, Pons D, Leebeek F, Monraats P, Schrevel M, Zwinderman A. et al. Platelet receptor P2RY12 haplotypes predict restenosis after percutaneous coronary interventions. *Hum Mutat*. 2008;29:375–380. [https://doi: 10.1002/humu.20641](https://doi.org/10.1002/humu.20641).
69. Nie X, Li JL, Zhang Y, Xu Y, Yang XL, Fu Y. et al. Haplotype of platelet receptor P2RY12 gene is associated with residual clopidogrel on-treatment platelet reactivity. *J Zhejiang Univ Sci B* 2017;18:37–47. [https://doi: 10.1631/jzus.B1600333](https://doi.org/10.1631/jzus.B1600333).
70. Yang H-H, Chen Y, Gao C-Y. Associations of P2Y12R gene polymorphisms with susceptibility to coronary heart disease and clinical efficacy of antiplatelet

- treatment with clopidogrel. *Cardiovasc Ther* 2016;34:460–467. [https://doi: 10.1111/1755-5922.12223](https://doi.org/10.1111/1755-5922.12223).
71. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL. et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* 2016;167:1415–1429.e19. [https://doi: 10.1016/j.cell.2016.10.042](https://doi.org/10.1016/j.cell.2016.10.042).
72. Nikpay M, Turner AW, McPherson R. Partitioning the Pleiotropy Between Coronary Artery Disease and Body Mass Index Reveals the Importance of Low Frequency Variants and Central Nervous System–Specific Functional Elements. *Circ Genom Precis Med* 2018;11(2):e002050. [https://doi: 10.1161/CIRCGEN.117.002050](https://doi.org/10.1161/CIRCGEN.117.002050).
73. Lahm H, Jia M, Dreßen M, Wirth F, Puluca N, Gilsbach R. et al. Congenital heart disease risk loci identified by genome-wide association study in European patients. *J Clin Invest* 2021;131(2):e141837. [https://doi: 10.1172/JCI141837](https://doi.org/10.1172/JCI141837).
74. Pereira A, Mendonça MI, Borges S, Freitas S, Henriques E, Rodrigues M. et al. Genetic Risk Analysis of Coronary Artery Disease in a Population-based Study in Portugal, Using a Genetic Risk Score of 31 Variants. *Arq Bras Cardiol* 2018 Jul;111(1):50–61. [https://doi: 10.5935/abc.20180107](https://doi.org/10.5935/abc.20180107).
75. Mendonça MI, Henriques E, Borges S, Sousa AC, Pereira A, Santos M. et al. Genetic information improves the prediction of major adverse cardiovascular events in the GENEMACOR population. *Genet Mol Biol* 2021;44(2):e20200448. [https://doi: 10.1590/1678-4685-GMB-2020-0448](https://doi.org/10.1590/1678-4685-GMB-2020-0448).
76. Whayne TF, Saha SP. Genetic Risk, Adherence to a Healthy Lifestyle, and Ischemic Heart Disease. *Curr Cardiol Rep* 2019;21(1):1. [https://doi: 10.1007/s11886-019-1086-z](https://doi.org/10.1007/s11886-019-1086-z).
77. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R. et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 2023;51(D1):D638–D646. doi: 10.1093/nar/gkac1000.61.

List of Tables

Table 1. Genomic characteristics of the important SNP predictors in the ML risk-stratification model

Model	SNP	Gene (within/nearby)	Gene name	Position (GRCh38)	Consequence	MAF*
Zero-part	rs3732379	CX3CR1	C-X3-C Motif Chemokine Receptor 1	chr3:39265765	missense variant	0.14 (T)
	rs1800562	HFE	Homeostatic Iron Regulator	chr6:26092913	missense variant	0.01 (A)
	rs11984041	HDAC9	Histone Deacetylase 9	chr7:18992312	intron variant	0.10 (T)
	rs2023938	HDAC9	Histone Deacetylase 9	chr7:18997152	3' prime UTR variant	0.11 (C)
	rs2107595	Mapped genes: HDAC9, TWIST1	Histone Deacetylase 9, Twist family bHLH transcription factor 1	chr7:19009765	regulatory region variant	0.25 (A)
	rs41291556	CYP2C19	Cytochrome P450 Family 2 Subfamily C Member 19	chr10:94775416	missense variant	< 0.01 (C)
	rs964184	ZPR1	ZPR1 Zinc Finger	chr11:116778201	3' prime UTR variant	0.22 (G)
	rs216172	SMG6	SMG6 Nonsense Mediated MRNA Decay Factor	chr17:2223210	intron variant	0.34 (C)
Count-part	rs6689306	IL6R	Interleukin 6 Receptor	chr1:15442347	intron variant	0.42 (A)
	rs4845625	IL6R	Interleukin 6 Receptor	chr1:154449591	intron variant	0.44 (T)
	rs2046934	MED12L - P2RY12	Mediator Complex Subunit 12L - Purinergic Receptor P2Y12	chr3:151339854	intron variant	0.13 (G)
	rs6801273	MED12L - P2RY12	Mediator Complex Subunit 12L - Purinergic Receptor P2Y13	chr3:151345042	intron variant	0.42 (C)
	rs870142	STX18-AS1	STX18 Antisense RNA 1	chr4:4646320	intron variant	0.20 (T)
	rs1332844	PHACTR1	Phosphatase and Actin Regulator 1	chr6:12888772	intron variant	0.32 (C)
	rs663129	Mapped genes: RNU4-17P, MC4R	RNA U4 small nuclear 17 pseudogene, Melanocortin 4 receptor	chr18:60171168	intergenic variant	0.25 (A)

* MAF, minor allele frequency (minor allele)

Legends and Figures

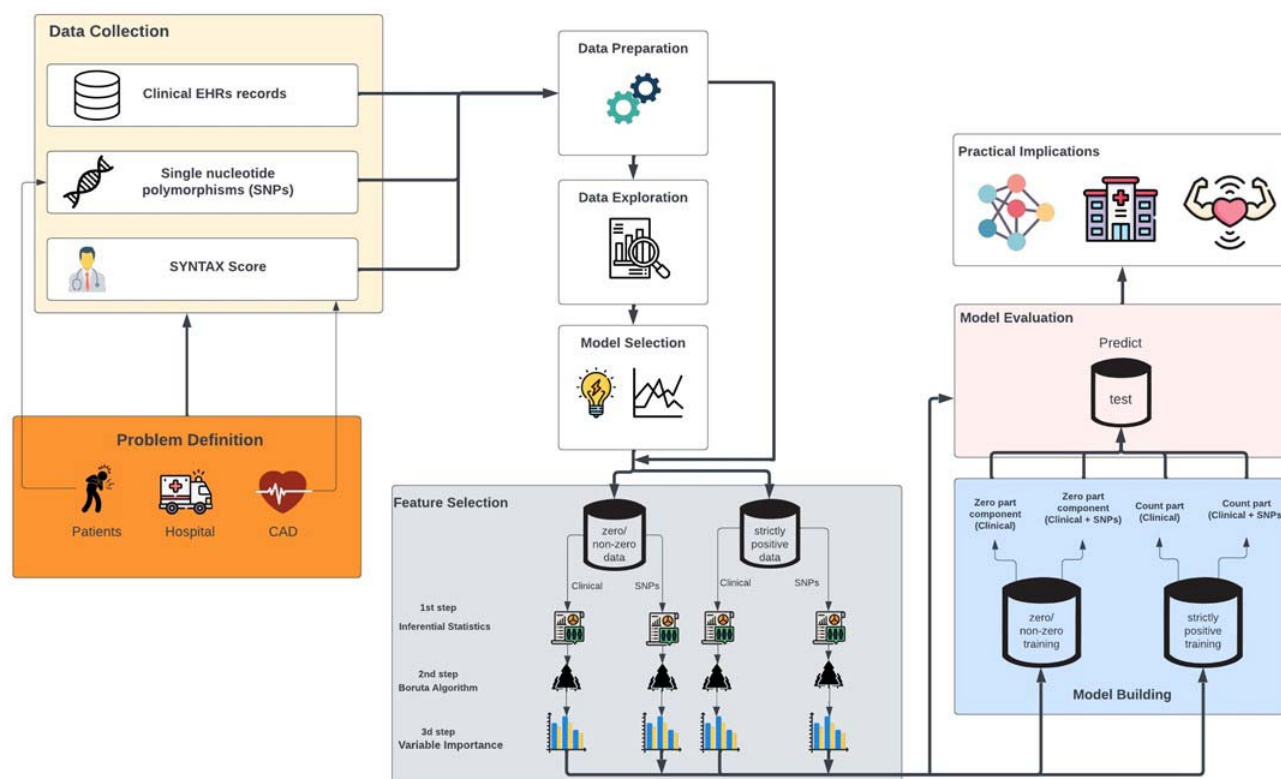


Figure 1. Data-driven ML approach of the study.

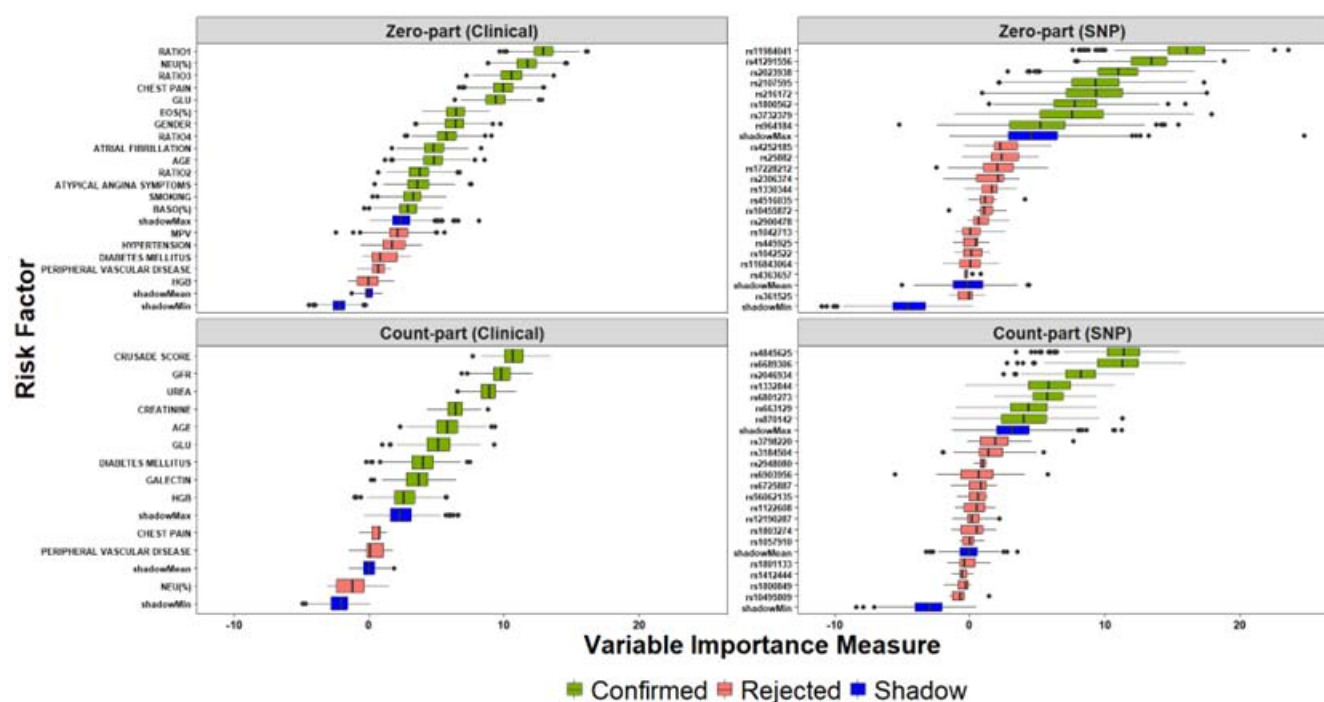


Figure 2a. Distributions of variable importance measure for the set of Clinical and SNPs predictors for the zero- and count-part models.

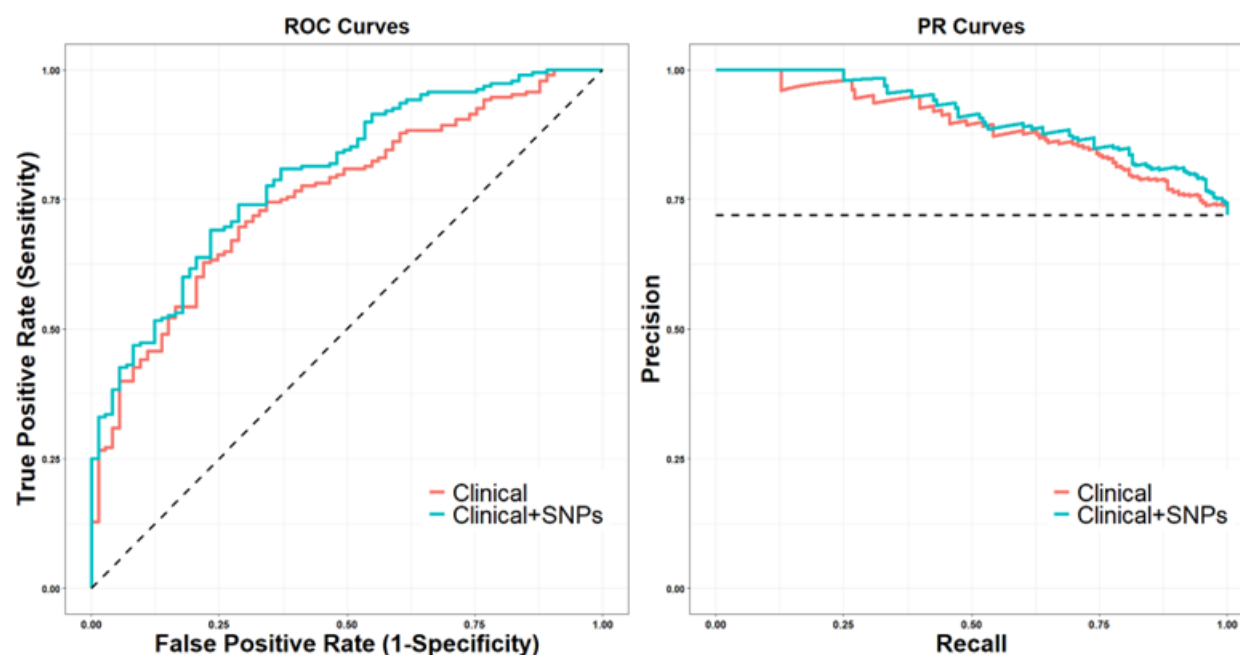


Figure 2b. ROC and PR curves for the evaluation of the classifier (zero-part model) for the competing models (Clinical (Model A)/Clinical+SNPs (Model B)).

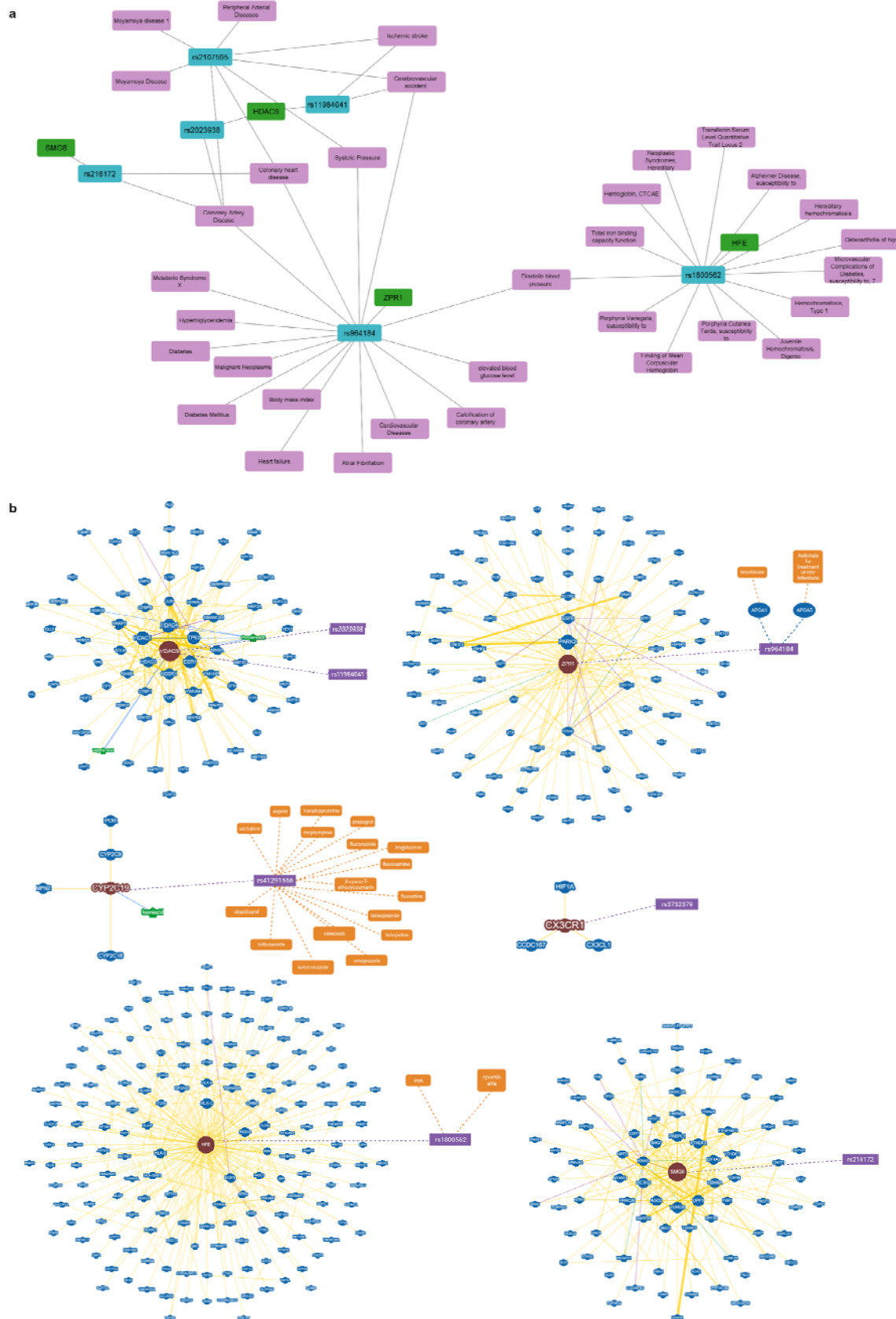


Figure 3. (a) Network visualization of the zero-part SNPs and their association with diseases according to DisGeNET database. The disease annotations, the SNPs, and the genes corresponding to the SNPs are depicted in purple, blue, and green round rectangles, respectively. The network was constructed using Cytoscape tool. **(b)** Gene-gene interaction network of the genes associated with the zero-part SNPs and their interaction genes, as retrieved by BioGRID database. The genes associated with the zero-part SNPs and their interaction genes are depicted in round red and blue shapes, respectively, while the SNPs in purple rectangles. Larger node sizes indicate higher connectivity, while thicker edge sizes represent stronger evidence supporting the association. Drugs associated with the SNPs, according to PharmGKB variant annotations, are depicted in orange round rectangle shapes, and the chemical interactions of the genes, according to BioGRID database, are shown in green round shapes.

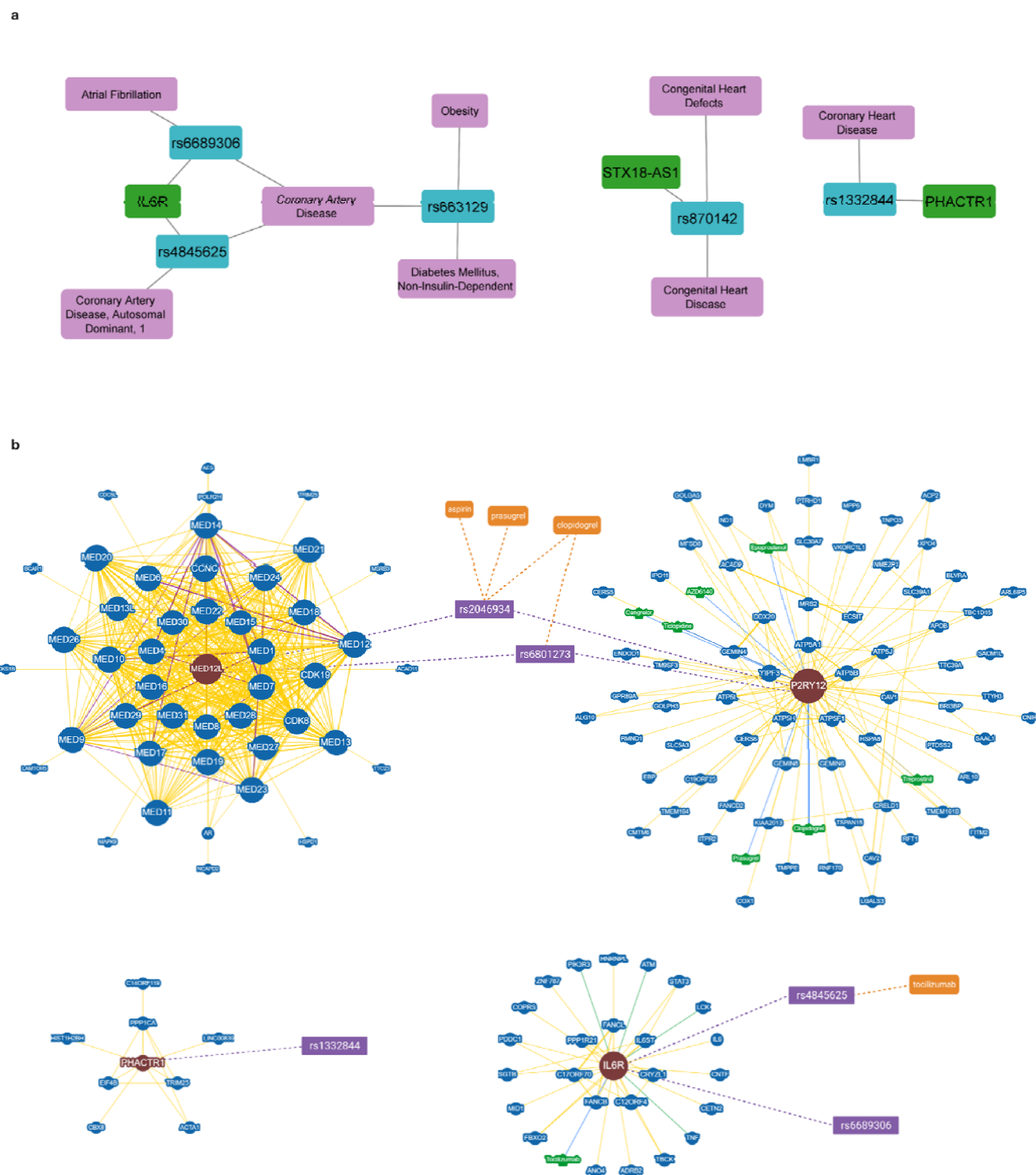


Figure 4. (a) Network visualization of the count-part SNPs and their association with diseases according to DisGeNET database. The disease annotations, the SNPs, and the genes corresponding to the SNPs are depicted in purple, blue, and green round rectangles, respectively. The network was

constructed using Cytoscape tool. **(b)** Gene-gene interaction network of the genes associated with the count-part SNPs and their interaction genes, as retrieved by BioGRID database. The genes associated with the count-part SNPs and their interaction genes are depicted in round red and blue shapes, respectively, while the SNPs in purple rectangles. Larger node sizes indicate higher connectivity, while thicker edge sizes represent stronger evidence supporting the association. Drugs associated with the SNPs, according to PharmGKB variant annotations, are depicted in orange round rectangle shapes and the chemical interactions of the genes, according to BioGRID database, are shown in green round shapes.

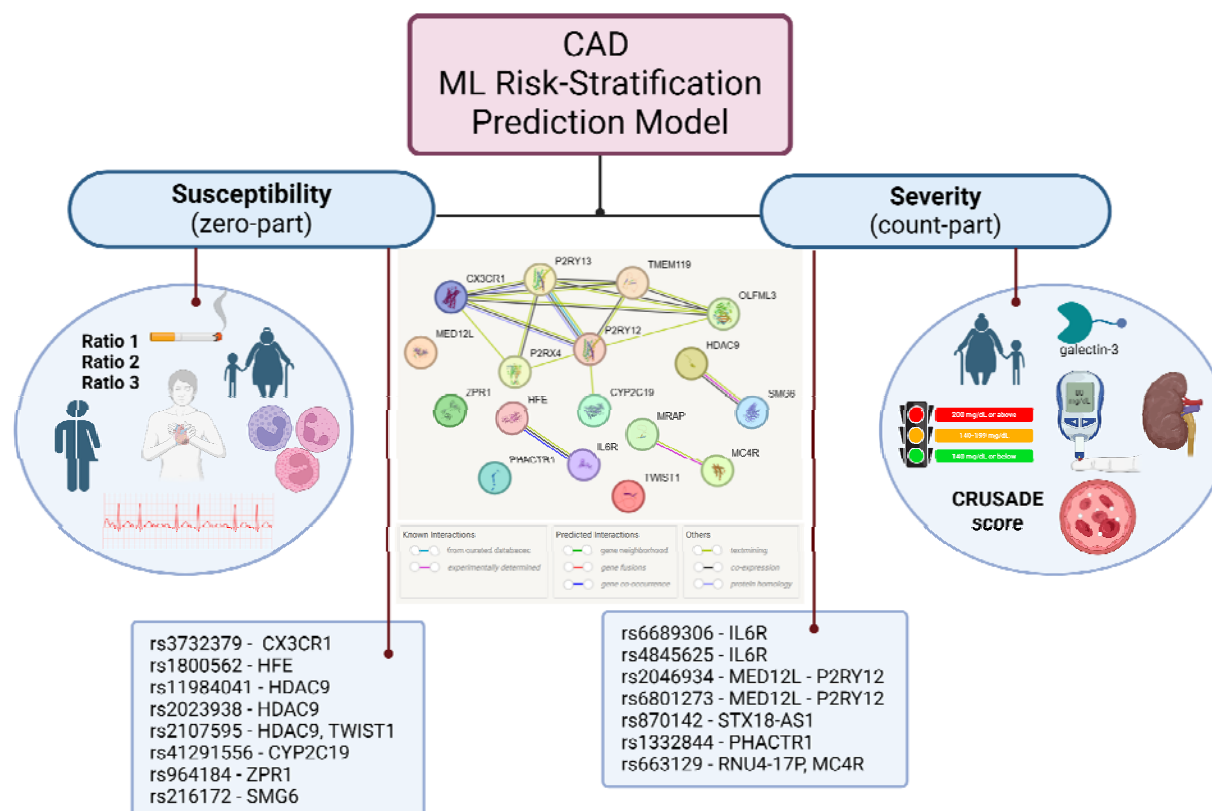


Figure 5. Schematic overview of the critical clinical, demographic and genetic features of the competing two-part ML risk-stratification model for the prediction of CAD. Note that the interactions between the genes of the identified important SNPs (shown in Table 1) were created through the protein-protein interaction networks functional enrichment analysis using the STRING database⁷⁷.