diagnostics

MDPI

Article

Machine Learning Coronary Artery Disease Prediction Based on Imaging and Non-Imaging Data

Vassiliki I. Kigka [1,2], Eleni Georga [1,2], Vassilis Tsakanikas [1,2], Savvas Kyriakidis [1,2], Panagiota Tsompou [1], Panagiotis Siogkas [1,2], Lampros K. Michalis [3], Katerina K. Naka [3], Danilo Neglia [4], Silvia Rocchiccioli [5], Gualtiero Pelosi [5], Dimitrios I. Fotiadis [1,2] and Antonis Sakellarios [1,2,*]

[1] Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, GR 45110 Ioannina, Greece; kigkavaso@gmail.com (V.I.K.); egewrga@gmail.com (E.G.); vasilistsakanikas@gmail.com (V.T.); savvasik21@gmail.com (S.K.); panagiotatsompou@gmail.com (P.T.); psiogkas4454@gmail.com (P.S.); dimitris.fotiadis30@gmail.com (D.I.F.)
[2] Institute of Molecular Biology and Biotechnology, Department of Biomedical Research—FORTH, University Campus of Ioannina, GR 45110 Ioannina, Greece
[3] Department of Cardiology, Medical School, University of Ioannina, GR 45110 Ioannina, Greece; lamprosmihalis@gmail.com (L.K.M.); drkknaka@gmail.com (K.K.N.)
[4] Fondazione Toscana Gabriele Monasterio, IT 56126 Pisa, Italy; dneglia@ftgm.it
[5] Institute of Clinical Physiology, National Research Council, IT 56124 Pisa, Italy; silvia.rocchiccioli@ifc.cnr.it (S.R.); pelosi@ifc.cnr.it (G.P.)
* Correspondence: ansakel13@gmail.com; Tel.: +30-26510-07848

Abstract: The prediction of obstructive atherosclerotic disease has significant clinical meaning for the decision making. In this study, a machine learning predictive model based on gradient boosting classifier is presented, aiming to identify the patients of high CAD risk and those of low CAD risk. The machine learning methodology includes five steps: the preprocessing of the input data, the class imbalance handling applying the Easy Ensemble algorithm, the recursive feature elimination technique implementation, the implementation of gradient boosting classifier, and finally the model evaluation, while the fine tuning of the presented model was implemented through a randomized search optimization of the model's hyper-parameters over an internal 3-fold cross-validation. In total, 187 participants with suspicion of CAD previously underwent CTCA during EVINCI and ARTreat clinical studies and were prospectively included to undergo follow-up CTCA. The predictive model was trained using imaging data (geometrical and blood flow based) and non-imaging data. The overall predictive accuracy of the model was 0.81, using both imaging and non-imaging data. The innovative aspect of the proposed study is the combination of imaging-based data with the typical CAD risk factors to provide an integrated CAD risk-predictive model.

Keywords: coronary artery disease; noninvasive cardiovascular imaging; coronary artery disease risk stratification; machine learning models

Citation: Kigka, V.I.; Georga, E.; Tsakanikas, V.; Kyriakidis, S.; Tsompou, P.; Siogkas, P.; Michalis, L.K.; Naka, K.K.; Neglia, D.; Rocchiccioli, S.; et al. Machine Learning Coronary Artery Disease Prediction Based on Imaging and Non-Imaging Data. *Diagnostics* **2022**, *12*, 1466. https://doi.org/10.3390/diagnostics12061466

Academic Editors: Keun Ho Ryu and Nipon Theera-Umpon

Received: 15 April 2022
Accepted: 11 June 2022
Published: 14 June 2022

1. Introduction

Atherosclerosis is considered as a chronic inflammatory disease of arteries, and its clinical manifestation accounts for a significant number of deaths worldwide. Atherosclerotic disease is characterized by the pathologic process of lipid accumulation and inflammation in the vessel wall, leading to the vessel wall thickening, lumen stenosis, calcification, and in some cases thrombosis [1]. The most important form of atherosclerosis is coronary artery disease (CAD), which accounts for the largest portion of cardiovascular disease deaths and leads to narrowing of the arteries that carry blood to the heart muscle [2]. The recent advances in coronary imaging techniques, either invasive or noninvasive, have enabled the identification of coronary vessels features, which are considered as CAD risk factors.

However, despite the recent technological cardiovascular imaging advancements to recognize the subclinical disease and the improvement of patient's management, the

identification of high-risk patients remains a challenge due to the inherently unpredictable disease's nature [3] since the rate of major adverse cardiac events (MACE) remains high both for patients with known CAD or for asymptomatic individuals [4].

Both CAD risk prediction and its progression prediction are two issues of high importance in biomedical research that aims to identify those individuals who are associated with an increased risk of CAD and the main factors that contribute to the disease progression. Existing studies have reported the different types of CAD risk factors, such as the patient's lipid profile (total cholesterol and low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides), smoking, hypertension, diabetes mellitus, obesity, and family history [5], and have established the importance of the conventional CAD risk factors in the prediction of CAD.

In the literature, different studies have been proposed for CAD risk prediction and the classification of patients into risk categories, either taking advantage of statistical modelling or artificial intelligence-based models [6–8]. The traditional statistical-based CAD prediction models have implemented regression models, such as the Cox model utilized in Framingham Risk Score study [6] and the Weibull model applied for the Systematic Coronary Risk Evaluation (SCORE) model. In spite of their predictability, statistical models are often dedicated to interpreting the input parameters and contributing to features association input analysis [9]. On the other hand, machine-learning-based models perform an automated search in the input features, either stochastic or deterministic, for the optimal prediction outcome and, in some cases, may be found advantageous over traditional regression methods [10,11].

As far as the existing machine-learning-based studies, different studies have been presented both for the prediction of CAD and the prediction of its progression, whereas other studies are dedicated to detect the most significant biomarkers. More specifically, Exarchos et al. [10] implemented typical classification schemes to predict the number of vessels' stenosis, the atherosclerosis progression, as well as a hybrid score corresponding to the severity of the disease. The utilized input features were demographics, clinical data, several biochemical variables, monocytes, and adhesion molecules. In another recently published study [11], demographics, clinical data, echocardiography data, and 54 features of laboratory variables were used to predict the status of CAD by applying a support vector machine (SVM) algorithm with kernel fusion. Ambale et al. [12] implemented machine learning techniques to characterize cardiovascular risk, predict outcomes, and identify biomarkers in population studies. More specifically, they tested the ability of random forests (RF) to predict six different cardiovascular events and concluded that the RF technique performed better than established risk scores with increased prediction accuracy. Motwani et al. [13] found that machine learning techniques combining clinical and CTCA data predict 5-year all-cause mortality (ACM) in patients with suspected coronary artery disease better than existing clinical or CTCA metrics alone. In a study proposed by van Rosendael et al. [14], they investigated whether a machine-learning-based score incorporating only the 16-segment coronary tree information derived from CTCA provides enhanced risk stratification compared with current CTCA-based risk scores and concluded that the proposed model can improve the integration of CTCA-derived plaque information to improve risk stratification. In a more recent study, Sakellarios et al. [15] presented a multi-parametric predictive model, including traditional risk factors, plasma lipids, 3D imaging parameters, and computational data, for the prediction of site-specific plaque progression and concluded that imaging-based characteristics, such as low endothelial shear stress (ESS) and low-density lipid (LDL) accumulation, are significant predictors. On the other hand, Heo et al. [16] developed and validated machine learning models to predict patients with hidden CAD and assess long-term outcomes in patients with acute ischemic stroke.

The basic concept of the proposed study is to develop a machine learning predictive model that incorporates both noninvasive imaging data derived by CTCA and typical patient baseline characteristics to predict the CAD risk and especially the obstructive disease. The clinical focus of the proposed machine-learning-based model is to indicate the prognostic value of the combination of non-imaging and imaging features derived by CTCA imaging for the prediction of CAD high-risk patients and to compare the predictability of the combination of non-imaging and imaging features with non-imaging features alone. As for the innovative aspect of the proposed study, we assemble a variety of patient characteristics that have never been previously utilized, aiming to predict the risk for CAD. The final selected predictive model adopted both bagging and boosting ensemble modelling principles such that the model's variance and bias are treated concurrently.

2. Materials and Methods

2.1. Dataset Description

The proposed study is based on the EVINCI population [17], in which patient-specific information, both imaging and non-imaging, were collected for clinical purposes and utilized as the baseline information for the development of a CAD risk stratification methodology, whereas the follow-up data were collected after 6.22 ± 1.42 during the SMARTool project (September 2016–November 2017) [18]. More specifically, during the H2020 SMARTool project, a prospective, multicenter study in patients was conducted by 7 medical centers (Pisa, Turku, Zurich, Barcelona, Warsaw, Naples, Viareggio) from 5 European countries. All the participants signed informed consent to participate in the study and all the following procedures. Patients who previously underwent coronary CTCA during the EVINCI (Evaluation of Integrated Cardiac Imaging for the Detection and Characterization of Ischemic Heart Disease; FP7-222915; n = 152—February 2009–June 2012) [12] and ARTreat (FP7-224297; n = 18) [13] clinical studies were prospectively included to undergo follow-up CTCA. In addition to this, individuals (n = 32) who underwent CTCA in the period from 2009 to 2012 were also prospectively included. A detailed list of inclusion and exclusion criteria is provided in Supplementary Materials.

Anonymized data were acquired from 187 patients, derived by different medical centers, and the cohort data were obtained under a data protection agreement fulling all the ethical and legal requirements for data sharing posed by the General Data Protection Regulation in a third-level care setting. Table 1 below demonstrates the collected data types. The median age of the patients of our dataset is 61 years old (45–76), and at their first visit to the physician, all the participants underwent CTCA imaging regardless of the presence of symptoms. More specifically, 45% and 25% of the participants had atypical and typical angina, respectively, whereas 12% of them had other symptoms, and 16% were asymptomatic. In addition to this, as for the pharmaceutical treatment of the participants, 18%, 28%, 13%, 40%, 13%, 10%, 3%, and 48% of them received angiotensin receptor blockers (ARBs), angiotensin converting enzyme inhibitors (ACE inhibitors), diuretics, beta blockers, calcium antagonists, oral antidiabetics, insulin, and statins, respectively, at the baseline time step.

2.2. Methodology

2.2.1. CTCA Image Analysis and Three-Dimensional Reconstruction

The first step of the development of the CAD risk-prediction model was the analysis of the CTCA images. This analysis was conducted by implementing an active contour based model for the segmentation of CTCA images and aimed to provide a detailed geometry of the three major coronary arteries, the left anterior descending artery (LAD), the left circumflex artery (LCX), and the right coronary artery (RCA). This methodology is integrated in a dedicated software tool, which can semi-automatically provide the detailed 3D coronary artery anatomy [19,20]. More details for the overall three-dimensional reconstruction methodology can be found in the Supplementary Materials in the Section 2.2.1.

Table 1. Imaging and non-imaging data utilized. * Imaging data from CTCA.

Type		Features
Imaging data *	Geometrical vasculature	Degree of Stenosis, Minimal Lumen Area, Minimal Lumen Diameter, Plaque Burden, Calcified Plaque Volume, Noncalcified Plaque Volume, SmartFFR Index, Number of Calcified Plaques, Number of Non-calcified Plaques
Non-imaging data	Demographics	Age, Gender
	Risk factors	Family History of CAD, Hypertension, Diabetes, Dyslipidemia, Smoking, Obesity, Metabolic Syndrome, Past Smokers
	Biohumoral Markers	Creatinine, Uric Acid, Glucose, Total Cholesterol, HDL, LDL, Triglycerides, Insulin, Aspartate Aminotransferase, Alanine Aminotransferase, Alkaline Phosphatase, Gamma-glutamyl Transferase, Hs-C Reactive Protein, Interleukin-6, TSH, fT3, fT4, Leptin, MMP2 Protein Plasma, MMP9 Protein Plasma, hs-cardiac Troponin T, N terminal Fragment of Pro-brain Natriuretic Peptide, Lipidomics, Metabolomics

2.2.2. Calculation of the SmartFFR index

In this study, except the geometrical derived metrics and a blood-flow-based index, the SmartFFR index [21] was utilized. More details about the SmartFFR index can be found in the Supplementary Materials, in the Section 2.2.2.

2.2.3. Problem Definition

The CAD risk stratification problem has been formulated as a two-class classification problem based on the maximal coronary artery stenosis. This hypothesis is based on the findings of the Coronary Artery Disease Reporting and Data System (CAD-RADS) [22], which provides a standardized method to associate findings of the CTCA imaging modality to facilitate decision making regarding further patient management. Figure 1 shows the distribution of the population across the two CAD-severity groups. More specifically, among the total 263 patients who underwent CTCA imaging for clinical purposes, 55 patients underwent percutaneous coronary intervention stenting procedure and 10 patients coronary artery bypass grafting procedure, whereas CTCA images of 11 patients were considered as interpretable either at the baseline time step or at the follow-up time step. The annotation was based on the assessment of the obstructive disease: at least one major artery with stenosis > 50%.

The definition of these two classes is based on the quantitative degree of stenosis derived by the CTCA imaging modality according to the society of cardiovascular computed tomography guidelines committee [23]. More specifically, the first class, the no CAD—minimal CAD class (Class $1^{-}C_1$)), includes the grading scale 0, 1, and 2 (normal, minimal, and mild), whereas the obstructive CAD class (Class $2^{-}C_2$)) includes the grading scale 3, 4, and 5 (moderate, severe, and occluded), as it is shown in Table 2. This classification was selected because we want to predict the obstructive CAD disease.

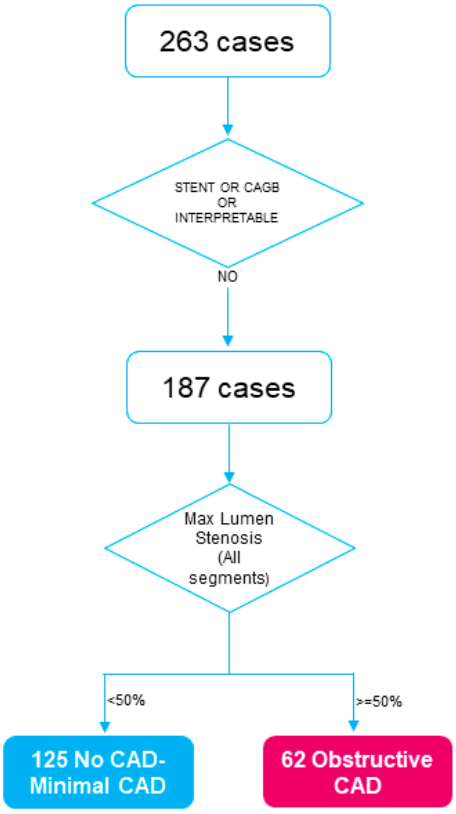

Figure 1. Flow chart depicting the distribution of the cohort in CAD-severity groups based on the CTCA imaging at the follow-up step. in total, 287 patient imaging data (125 in Class 1 and 62 in Class 2) were analyzed. (CAGB, coronary artery bypass graft surgery; CAD, coronary artery disease).

Table 2. Definition of the utilized CAD risk classes. (CAD, coronary artery disease).

Proposed Classes	Recommended Stenosis Grading Scale of CAD	Quantitative Stenosis
Class 1-C_1	0: Normal 1: Minimal 2: Mild	No luminal stenosis Plaque with <25% stenosis 25–49% stenosis
Class 2-C_2	3: Moderate 4: Severe 5: Occluded	50–69% stenosis 70–99% stenosis 100% stenosis

Baseline imaging and non-imaging characteristics were trained into a gradient boosting classification scheme, aiming to discriminate the patients at low risk (Class C_1) and those at high risk (Class C_2), concerning their follow-up time step. This predictive supervised learning approach aims to learn mapping from input features x to output Y given a labeled set of input output pairs $D = \{(x_i, Y_i)\}_{i=1}^{N}$, where D is the training set, and N is the number of training examples [24]. Each sample (x_i, Y_i) associates the input features with the risk prediction of CAD severity, Y, where $Y \in \{C_1, C_2\}$, is estimated by a non-linear parameterized function (f) of input features $x \in R^d$, $x = [x_1, x_2, \ldots, x_N]$. The goal of this supervised classification problem is to obtain an approximation $F(x)$ of the function $F^*(x)$ mapping the input x to output Y. The function $F^*(x)$ minimizes the expected value of some specified loss function $L(y, F(x))$, whereas the procedure followed in this proposed study is to restrict the function $F(x)$ to be a member of parameterized class of functions $F(x; Y)$. In addition to this, in this paper, we constructed our model based on additive

expansions of the form $F\left(x;\{\beta_m, a_m\}_1^M\right)$. $F^*(x)$ and $F\left(x;\{\beta_m, a_m\}_1^M\right)$, which are described in the Supplementary Materials in Equations (S1) and (S2), respectively [24].

The selected predictive model was nested into an easy ensemble classification scheme to overcome the class imbalance problem. To estimate the classification performance of the proposed method, an externally stratified 10-fold cross-validation was applied, with data pre-processing, a multivariate feature ranking, and a gradient boosting classification scheme being efficiently combined at each iteration of the procedure. The overall proposed model performs feature selection in the learning time since it achieves model fitting and feature selection simultaneously. Data-preprocessing and feature ranking follow the resampling procedure itself, which reduces the selection bias in the estimates of the model's performance, whereas stratification assures that each validation fold retains the class distribution in the dataset. In addition to these, randomized search optimization of the model's hyper-parameters over an internal 3-fold cross-validation contributes to the fine-tuning of the presented model.

2.2.4. Easy Ensemble Algorithm Implementation-Class Imbalance Handling

The easy ensemble algorithm [25] is a class imbalance handling approach in which P are the training instances of the minority class, whereas Q denotes the instances of the majority class. The idea of the easy ensemble algorithm is to employ random resampling to generate K subsets of $\{Q_1, Q_2, \ldots., Q_K\}$ from Q ($|Q_i| < |Q|$, $i = 1, 2, \ldots, K$). Subsequently, each $Q_i \cup P$ is trained by the classifier, and the final decision is selected by majority voting. In the proposed predictive model approach, the easy ensemble approach is combined with the gradient boosting classifier, and each individual model is trained by the Equations (S10)–(S13) in the Supplementary Materials.

2.2.5. Data Pre-Processing

In this step, one hot encoding procedure was implemented to represent all the categorical input features as binary vectors. In addition to this, a curation procedure was implemented to curate our dataset both for outliers and missing values. All the input features whose missing values were higher than 10% were removed from the dataset, whereas features with missing values lower than 10% were imputed by either the most frequent value (categorical type features) or the median value (numerical type features).

2.2.6. Recursive Feature Elimination

In this step, our aim is to reduce the dimensionality d of input features $x \in R^d$ to overcome the risk of overfitting, which basically arises when the number of d is comparatively large, and the number of the training patterns is small. In this study, a feature ranking technique with a support vector machine (SVM) with recursive feature elimination (RFE) was implemented to rank the input features. The whole SMV RFE procedure is shown in Table S1 in the Supplementary Materials [26].

2.2.7. Gradient Boosting Classification

In the first step, the gradient boosting classification algorithm [27] implements a numerical implementation minimizing the equation of $F^*(x)$ (Equation (S1)—Supplementary Materials). The whole function of the utilized classification scheme is described in the Supplementary Materials (Equations (S3)–(S13)).

The overall pipeline of the proposed machine learning methodology is shown in Figure 2 below.

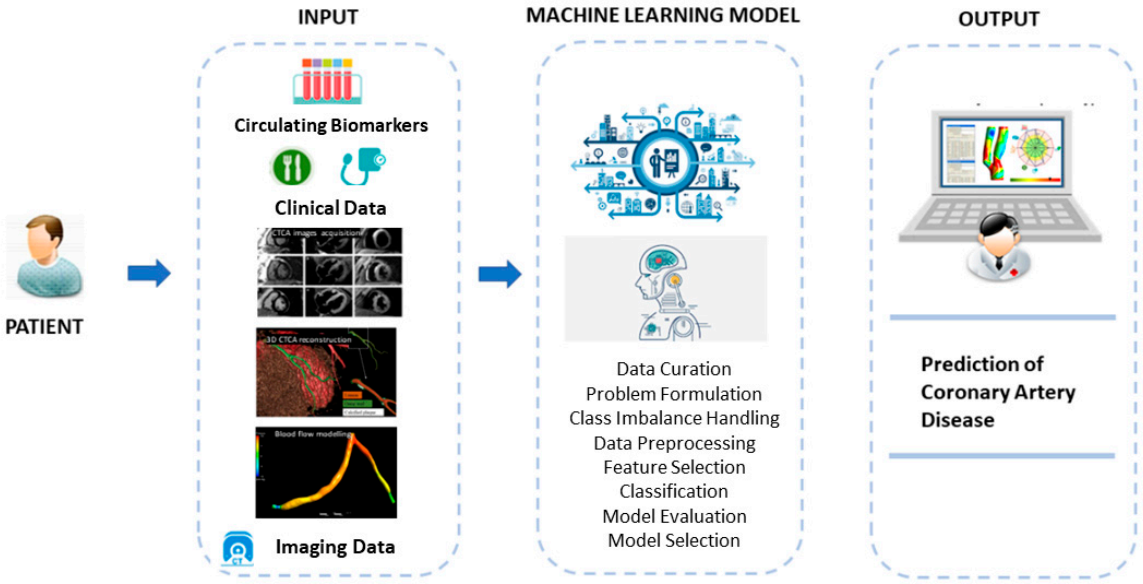

Figure 2. Overall pipeline of the proposed methodology. The input is based on clinical data, laboratory test, and imaging data provided by the three-dimensional reconstruction of the artery and the blood-flow modelling. Different machine learning models were implemented for the prediction of coronary artery disease presence.

3. Results

CAD Risk-Prediction Model Performance Evaluation

The utilized CAD risk-prediction model performance metrics are the balanced accuracy, the negative predictive value, the positive predictive value, the area under the receiver operating curve (ROC AUC), and the sensitivity and specificity. The values of the adopted performance metrics and their mean value and the 10-fold standard deviation are given in Table 3. The average balanced accuracy of the selected predictive model is 0.81, while its sensitivity and specificity is 0.88 and 0.73, respectively. In Figure 3, we demonstrate the normalized confusion matrix regarding the selected gradient boosting classification algorithm combined with an SVM RFE feature selection technique. In addition to this, in Table 4, the respective performance metrics over the different folds and their mean and standard deviation values using only non-imaging data are shown. The average balanced accuracy of the predictive model trained only by non-imaging features is 0.69, while its sensitivity and specificity are both 0.69.

Table 3. Evaluation of the CAD risk-prediction problem over 10-fold using imaging and non-imaging data (AUC, area under curve).

Folds	Balanced Accuracy	Negative Predictive Value	Positive Predictive Value	ROC AUC	Sensitivity	Specificity
Fold #0	0.73	0.78	0.67	0.60	0.67	0.78
Fold #1	0.75	0.86	0.63	0.82	0.84	0.67
Fold #2	0.89	1	0.72	0.92	1	0.78
Fold #3	0.69	0.78	0.6	0.72	0.6	0.78
Fold #4	0.84	1	0.63	0.83	1	0.67
Fold #5	0.84	1	0.63	0.89	1	0.67
Fold #6	0.95	1	0.84	1	1	0.89
Fold #7	0.78	1	0.56	0.78	1	0.56
Fold #8	0.8	0.86	0.72	0.88	0.84	0.75
Fold #9	0.8	0.86	0.72	0.75	0.84	0.75
Mean ± std	**0.81 ± 0.08**	**0.92 ± 0.1**	**0.68 ± 0.08**	**0.82 ± 0.11**	**0.88 ± 0.15**	**0.73 ± 0.09**

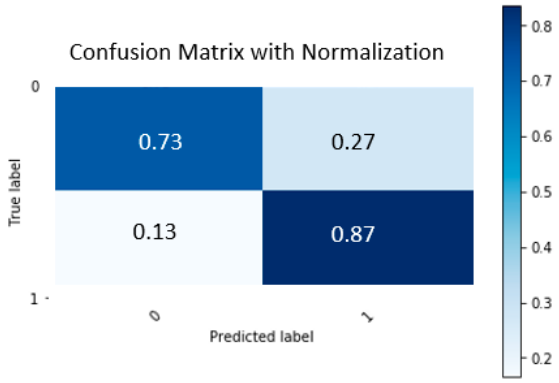

Figure 3. Normalized Confusion Matrix regarding the Gradient Boosting Classification algorithm for the CAD risk prediction using imaging and non-imaging data. The percentage of the true negative predicted cases is 73%, whereas the percentage of the true positives cases is 87%.

Table 4. Evaluation of the CAD risk-prediction problem over 10-fold using only non-imaging data (AUC, area under curve).

Folds	Balanced Accuracy	Negative Predictive Value	Positive Predictive Value	ROC AUC	Sensitivity	Specificity
Fold #0	0.5	0.6	0.4	0.33	0.33	0.67
Fold #1	0.56	0.64	0.5	0.65	0.33	0.78
Fold #2	0.72	1	0.5	0.89	1	0.44
Fold #3	0.69	0.78	0.6	0.69	0.6	0.78
Fold #4	0.79	0.88	0.67	0.87	0.8	0.78
Fold #5	0.78	1	0.56	0.78	1	0.56
Fold #6	0.79	0.88	0.67	0.8	0.8	0.78
Fold #7	0.72	1	0.5	0.76	1	0.44
Fold #8	0.6	0.64	0.67	0.79	0.33	0.88
Fold #9	0.71	0.75	0.67	0.83	0.67	0.75
Mean ± std	**0.69 ± 0.1**	**0.82 ± 0.16**	**0.57 ± 0.1**	**0.74 ± 0.16**	**0.69 ± 0.28**	**0.69 ± 0.15**

Additionally, a SHAPley Additive exPlanations (SHAP) analysis was implemented for explaining the prediction of the proposed model by computing the contribution of each feature to the prediction [28]. The most important predictors of the proposed model are presented in Figure 4 below. Mean absolute SHAP values for the 10 most significant features are estimated to illustrate the global feature importance. As it is shown in Figure 4, the most significant feature is the number of the existing calcified plaques and the highest coronary degree of stenosis at the baseline step. In addition to this, input features such as pro-brain natriuretic peptide (NT-proBNP), matrix metalloproteinase-2 and 9 (MMP-2, MMP-9), leptin, low-density lipoprotein (LDL), and patient characteristics such as weight, age, and height are highly ranked as significant features for the prognosis of coronary artery disease (CAD).

In addition to this, in Figure 5 below, we demonstrate the global interpretability of the proposed model by representing how much each input feature, either positively or negatively, contributes to the target variable. In Figure 5, we show with yellow columns the input features that contribute positively to the output target (detection of Class 2, CAD class). On the other hand, with blue columns, we indicate the input predictors that contribute negatively to the output target (detection of Class 1, no-CAD class). As it is shown in Figure 5, most of the input features contribute negatively to the output target and contribute to the prognosis of Class 1. Indicatively, the most significant features that contribute positively to the output target are thyroid stimulating hormone, medication therapy of beta blockers, aspartate aminotransferase, diabetes, and minimum lumen area. In the presented model, we observe that the most significant predictor for the prognosis of CAD is thyroid stimulating hormone, which confirms the effect of the thyroid hormones

on the cardiovascular system [29,30]. Thyroid hormone is considered is a significant regulator of cardiovascular system function and hemodynamics through different mechanisms. More specifically, inadequate thyroid hormone levels impair the relaxation of vascular smooth muscle cells and decrease cardiac contractility by regulating calcium uptake and the expression of several contractile proteins in cardiomyocytes. Additionally, low thyroid hormone levels also increase systemic vascular resistance and induce endothelial dysfunction by reducing nitric oxide availability [31,32]. As for the imaging-based input predictors, minimum lumen area has the most significant positive effect on the proposed model. As it is shown in Figure 5, the most significant feature with negative effect on the output is the number of the calcified plaques at the baseline analysis of patient imaging. Different studies in the literature have confirmed the prognostic capability of the presence of calcified atherosclerotic plaques [33]. Calcification of the coronary arteries plays a key role in the pathophysiology of atherosclerosis, and these lesions are considered advanced lesions [34]. In addition to this, patient height contributes negatively to the prognosis of the output target, confirming the genetic relationship between height and coronary artery disease [35,36]. As for the biochemical predictors for the no-CAD class (Class 1), we observed that pro-brain natriuretic peptide, low density lipoprotein, and matrix metalloproteinase 2 have a high negative effect on the output target.

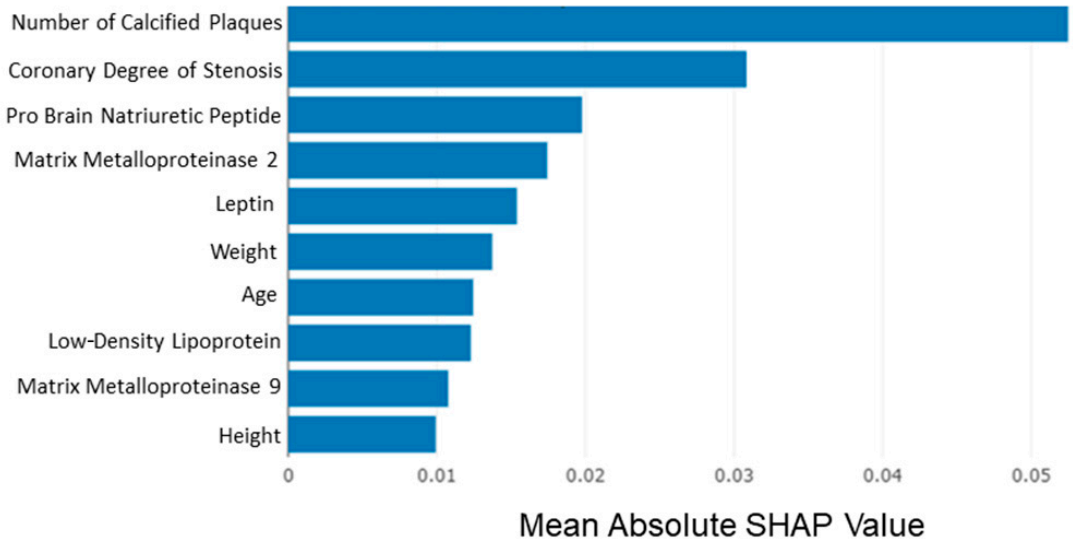

Figure 4. Feature importance based on mean SHAP values. The number of the existing calcified plaques and the highest coronary degree of stenosis are indicated as the most significant features.

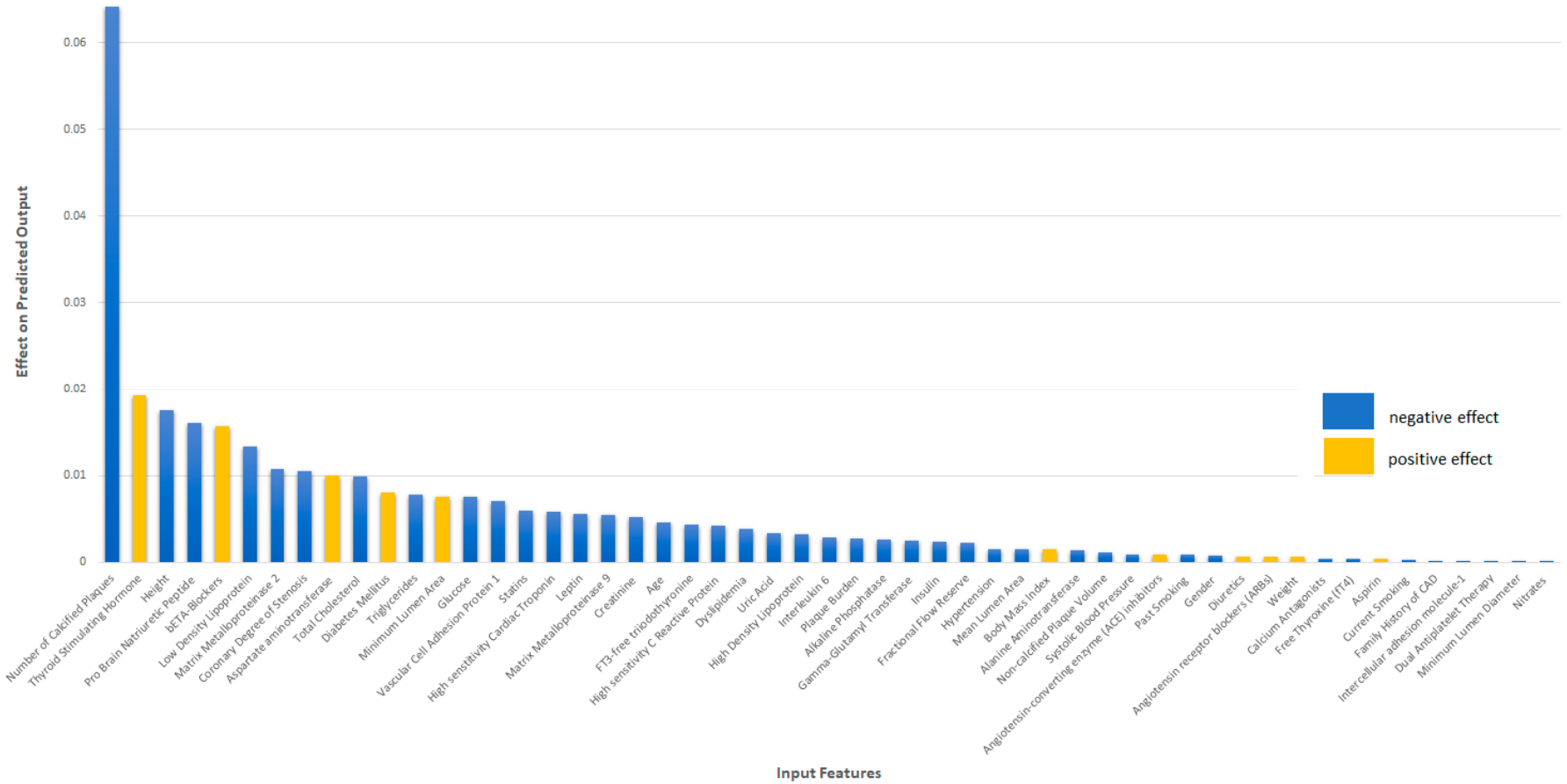

Figure 5. Input Features Contribution Table (blue, features with negative effect; yellow, features with positive effect). The most significant features that contribute positively to the output target are thyroid stimulating hormone, medication therapy of beta blockers, aspartate aminotransferase, diabetes, and minimum lumen area, whereas the most significant feature with negative effect on the output is the number of the calcified plaques at the baseline analysis of patient imaging.

4. Discussion

In this study, a novel approach for the prediction of obstructive CAD is presented. The aim of this study is to develop a machine learning model for the CAD risk prediction, which takes into account different types of data, including both imaging and non-imaging data. To our knowledge, our approach to combine the imaging and blood-flow-based characteristics with typical CAD risk factors constitutes the novelty of the presented study.

Different methodologies have been presented for the prediction of CAD and the identification of the major CAD risk factors. Most of these studies are concentrated on the different CAD-related risk-prediction outputs and are based either on statistical analysis [6,37,38] or machine learning classification schemes [39]. Our proposed study in comparison with these ones is more concentrated on the CAD risk prediction and its future presence and achieves a higher AUC.

Additionally, recent studies have indicated that non-invasive cardiovascular imaging and especially the CTCA imaging modality utilized in this study provides useful prognostic information of atherosclerosis progression since it permits the accurate quantification of luminal area and the detection of plaque burden region and the characterization of its composition. Moreover, the overall plaque burden, which can be provided by CTCA imaging, is highly relevant to the degree and characteristics of atherosclerosis [40]. In addition to this, the clinical relevance of the overall coronary plaque burden has been also emphasized by studies showing that increased non-calcified plaque volumes is directly linked with acute coronary syndrome (ACS) patients [41]. Furthermore, the latest technological advancements in patient-specific blood-flow modeling have introduced alternative CAD progression risk factors, such as fractional flow reserve (FFR) index and wall shear stress (WSS).

The prognostic capability of CTCA imaging modality and its derived imaging features has also been confirmed by the proposed study, in which the overall accuracy of the proposed predictive model using both imaging and non-imaging data is 0.81. Moreover, the prognostic significance of imaging-derived features is also indicated by the collected results, shown in Table 4. More specifically, the predictive model trained by the non-imaging-based features achieved a comparatively lower accuracy of 0.69.

Furthermore, another notable point of the proposed CAD risk-predictive model is that the input geometrical features are derived by an automated CTCA image analysis tool [19,20], able to detect accurately the inner and outer wall and atherosclerotic plaques and provide an accurate 3D model of coronary arteries and the atherosclerotic plaques distribution over the 3D space. As far as the SmartFFR index is concerned, it is also calculated automatically by the developed software tool in the 3D-reconstructed coronary artery.

In addition to this, another innovative aspect of the presented predictive model is the implementation of the easy ensemble algorithm, which constitutes a random resampling scheme, which mainly handles the class imbalance problem. Except for the class imbalance handling, the applied easy ensemble scheme allows the progressive correction of the model's decision hyperplane and subsequently the reduction of the classification error. In addition to this, the predictive capability of the proposed model is evaluated based on nested stratified cross-validation, which provides and unbiased estimation of the predictive model's capability. Moreover, except for the innate hyperparameters of the classification algorithm, the input features are also treated as a hyper-parameter, and an SVM RFE feature selection technique is implemented to eliminate the input features' dimension. The particular machine learning algorithm was selected after the implementation of different classification schemes in combination with different feature selection techniques, and the highest accuracy was provided by the combination of the extreme gradient boosting algorithm and the support vector machine (SVM) feature selection technique.

However, except for the prediction of CAD presence, the prediction of the CAD-related events is also a very important task both for the clinical research area and for patients' management. However, the proposed methodology was trained using an existing dataset of 187 participants, in which there were only few CAD-related events. This is a low–medium-risk population, and we have few major CAD events to use for the development

of such an event-prediction model. On the other hand, thanks to the advantage of our intermediate CAD risk population, we were able to build a model that can be used as a prognostic decision-support tool by clinicians to properly monitor and manage patients of intermediate CAD risk for the next years after a first imaging is available.

A future step of the proposed methodology will include the integration of additional imaging-based features, which will be either based on CTCA image analysis or on blood-flow modeling. Additionally, another future step will be the development of predictive models aiming to predict CAD-related events either using imaging and non-imaging data and the investigation of the predictability of these features when the desired outcome is the CAD-related events.

5. Conclusions

This proposed CAD risk-predictive model highlights the clinical utility of machine learning models to identify individuals of high CAD risk and those at risk of a potential CAD clinical event. In this investigation, we conclude that both imaging-derived features, combined with typical CAD risk factors and typical biochemical markers, have a significant predictive capability considering risk-prediction problem for CAD presence. In clinical practice, the utilization of such machine-learning-based approaches could improve CAD risk stratification and contribute to better strategies for patients' management.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/diagnostics12061466/s1. Reference [42] is cited in the supplementary materials.

Author Contributions: Conceptualization and investigation, V.I.K., V.T., E.G., S.K., P.T., P.S., L.K.M., K.K.N., D.N., S.R., G.P., and D.I.F.; supervision, A.S. All authors have read and agreed to the published version of the manuscript.

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

Institutional Review Board Statement: Ethical approval was provided by each participating center (National Research Council, University of Turku, University of Zurich, Fondazione Toscana Gabriele Monasterio, Warsaw National Institute of Cardiology) through the approval of the clinical study by the Ethics Committee Vast Area Northwest of Tuscany (CEAVNO), Pisa, Italy, and all subjects gave written informed consent. Our clinical study follows the Declaration of Helsinki. Approval Code: 1003/2016, Approval Date: 14 April 2016.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Subcommittee on Arteriosclerosis; Andrus, E.C.; Allen, E.V.; Merritt, H.H.; Duff, G.L.; Moore, R.A.; Kendall, F.E.; Shumacker, J.H.B.; Levy, R.L.; Wright, I.S. The pathogenesis of arteriosclerosis 1. *Int. J. Epidemiol.* **2015**, *44*, 1791–1793. [CrossRef] [PubMed]
2. Wong, N.D. Epidemiological studies of CHD and the evolution of preventive cardiology. *Nat. Rev. Cardiol.* **2014**, *11*, 276. [CrossRef] [PubMed]
3. Wexler, L.; Brundage, B.; Crouse, J.; Detrano, R.; Fuster, V.; Maddahi, J.; Rumberger, J.; Stanford, W.; White, R.; Taubert, K. Coronary artery calcification: Pathophysiology, epidemiology, imaging methods, and clinical implications: A statement for health professionals from the American Heart Association. *Circulation* **1996**, *94*, 1175–1192. [CrossRef] [PubMed]
4. Papafaklis, M.; Mavrogiannis, M.; Stone, P. Identifying the progression of coronary artery disease: Prediction of cardiac events. *Contin. Cardiol. Educ.* **2016**, *2*, 105–114. [CrossRef]
5. Roeters van Lennep, J.E.; Westerveld, H.T.; Erkelens, D.W.; van der Wall, E.E. Risk factors for coronary heart disease: Implications of gender. *Cardiovasc. Res.* **2002**, *53*, 538–549. [CrossRef]
6. D'agostino, R.B.; Vasan, R.S.; Pencina, M.J.; Wolf, P.A.; Cobain, M.; Massaro, J.M.; Kannel, W.B. General cardiovascular risk profile for use in primary care: The Framingham Heart Study. *Circulation* **2008**, *117*, 743–753. [CrossRef]

7. Conroy, R.M.; Pyörälä, K.; Fitzgerald, A.E.; Sans, S.; Menotti, A.; De Backer, G.; De Bacquer, D.; Ducimetiere, P.; Jousilahti, P.; Keil, U.; et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: The SCORE project. *Eur. Heart J.* **2003**, *24*, 987–1003. [CrossRef]
8. Miller, R.J.; Huang, C.; Liang, J.X.; Slomka, P.J. Artificial intelligence for disease diagnosis and risk prediction in nuclear cardiology. *J. Nucl. Cardiol.* **2022**, 1–9. [CrossRef]
9. Goldstein, B.A.; Navar, A.M.; Carter, R.E. Moving beyond regression techniques in cardiovascular risk prediction: Applying machine learning to address analytic challenges. *Eur. Heart J.* **2017**, *38*, 1805–1814. [CrossRef]
10. Exarchos, K.P.; Carpegianni, C.; Rigas, G.; Exarchos, T.P.; Vozzi, F.; Sakellarios, A.; Marraccini, P.; Naka, K.; Michalis, L.; Parodi, O. A multiscale approach for modeling atherosclerosis progression. *IEEE J. Biomed. Health Inform.* **2015**, *19*, 709–719. [CrossRef]
11. Alizadehsani, R.; Zangooei, M.H.; Hosseini, M.J.; Habibi, J.; Khosravi, A.; Roshanzamir, M.; Khozeimeh, F.; Sarrafzadegan, N.; Nahavandi, S. Coronary artery disease detection using computational intelligence methods. *Knowl. Based Syst.* **2016**, *109*, 187–197. [CrossRef]
12. Ambale-Venkatesh, B.; Yang, X.; Wu, C.O.; Liu, K.; Hundley, W.G.; McClelland, R.; Gomes, A.S.; Folsom, A.R.; Shea, S.; Guallar, E. Cardiovascular event prediction by machine learning: The multi-ethnic study of atherosclerosis. *Circ. Res.* **2017**, *121*, 1092–1101. [CrossRef] [PubMed]
13. Motwani, M.; Dey, D.; Berman, D.S.; Germano, G.; Achenbach, S.; Al-Mallah, M.H.; Andreini, D.; Budoff, M.J.; Cademartiri, F.; Callister, T.Q. 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.* **2017**, *38*, 500–507. [CrossRef] [PubMed]
14. van Rosendael, A.R.; Maliakal, G.; Kolli, K.K.; Beecy, A.; Al'Aref, S.J.; Dwivedi, A.; Singh, G.; Panday, M.; Kumar, A.; Ma, X. Maximization of the usage of coronary CTA derived plaque information using a machine learning based algorithm to improve risk stratification; insights from the CONFIRM registry. *J. Cardiovasc. Comput. Tomogr.* **2018**, *12*, 204–209. [CrossRef]
15. Sakellarios, A.I.; Tsompou, P.; Kigka, V.; Siogkas, P.; Kyriakidis, S.; Tachos, N.; Karanasiou, G.; Scholte, A.; Clemente, A.; Neglia, D. Non-invasive prediction of site-specific coronary atherosclerotic plaque progression using lipidomics, blood flow, and LDL transport modeling. *Appl. Sci.* **2021**, *11*, 1976. [CrossRef]
16. Heo, J.; Yoo, J.; Lee, H.; Lee, I.H.; Kim, J.-S.; Park, E.; Kim, Y.D.; Nam, H.S. Prediction of Hidden Coronary Artery Disease Using Machine Learning in Patients with Acute Ischemic Stroke. *Neurology* **2022**. [CrossRef]
17. Liga, R.; Vontobel, J.; Rovai, D.; Marinelli, M.; Caselli, C.; Pietila, M.; Teresinska, A.; Aguadé-Bruix, S.; Pizzi, M.N.; Todiere, G. Multicentre multi-device hybrid imaging study of coronary artery disease: Results from the EValuation of INtegrated Cardiac Imaging for the Detection and Characterization of Ischaemic Heart Disease (EVINCI) hybrid imaging population. *Eur. Heart J. Cardiovasc. Imaging* **2016**, *17*, 951–960. [CrossRef]
18. Sakellarios, A.I.; Rigas, G.; Kigka, V.; Siogkas, P.; Tsompou, P.; Karanasiou, G.; Exarchos, T.; Andrikos, I.; Tachos, N.; Pelosi, G. SMARTool: A tool for clinical decision support for the management of patients with coronary artery disease based on modeling of atherosclerotic plaque process. In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea, 11–15 July 2017; pp. 96–99.
19. Kigka, V.I.; Rigas, G.; Sakellarios, A.; Siogkas, P.; Andrikos, I.O.; Exarchos, T.P.; Loggitsi, D.; Anagnostopoulos, C.D.; Michalis, L.K.; Neglia, D.; et al. 3D reconstruction of coronary arteries and atherosclerotic plaques based on computed tomography angiography images. *Biomed. Signal Process. Control* **2018**, *40*, 286–294. [CrossRef]
20. Kigka, V.I.; Sakellarios, A.; Kyriakidis, S.; Rigas, G.; Athanasiou, L.; Siogkas, P.; Tsompou, P.; Loggitsi, D.; Benz, D.C.; Buechel, R. A three-dimensional quantification of calcified and non-calcified plaques in coronary arteries based on computed tomography coronary angiography images: Comparison with expert's annotations and virtual histology intravascular ultrasound. *Comput. Biol. Med.* **2019**, *113*, 103409. [CrossRef]
21. Siogkas, P.K.; Anagnostopoulos, C.D.; Liga, R.; Exarchos, T.P.; Sakellarios, A.I.; Rigas, G.; Scholte, A.J.H.A.; Papafaklis, M.I.; Loggitsi, D.; Pelosi, G.; et al. Noninvasive CT-based hemodynamic assessment of coronary lesions derived from fast computational analysis: A comparison against fractional flow reserve. *Eur. Radiol.* **2019**, *29*, 2117–2126. [CrossRef]
22. Cury, R.C.; Abbara, S.; Achenbach, S.; Agatston, A.; Berman, D.S.; Budoff, M.J.; Dill, K.E.; Jacobs, J.E.; Maroules, C.D.; Rubin, G.D. CAD-RADSTM coronary artery disease–reporting and data system. An expert consensus document of the Society of Cardiovascular Computed Tomography (SCCT), the American College of Radiology (ACR) and the North American Society for Cardiovascular Imaging (NASCI). Endorsed by the American College of Cardiology. *J. Cardiovasc. Comput. Tomogr.* **2016**, *10*, 269–281. [PubMed]
23. Raff, G.L.; Abidov, A.; Achenbach, S.; Berman, D.S.; Boxt, L.M.; Budoff, M.J.; Cheng, V.; DeFrance, T.; Hellinger, J.C.; Karlsberg, R.P. SCCT guidelines for the interpretation and reporting of coronary computed tomographic angiography. *J. Cardiovasc. Comput. Tomogr.* **2009**, *3*, 122–136. [CrossRef] [PubMed]
24. Robert, C. *Machine Learning, A Probabilistic Perspective*; Taylor & Francis: Abingdon-on-Thames, UK, 2014.
25. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. Part B* **2009**, *39*, 539–550. [CrossRef]
26. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
27. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
28. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777.

29. Cappola, A.R.; Desai, A.S.; Medici, M.; Cooper, L.S.; Egan, D.; Sopko, G.; Fishman, G.I.; Goldman, S.; Cooper, D.S.; Mora, S. Thyroid and cardiovascular disease: Research agenda for enhancing knowledge, prevention, and treatment. *Circulation* **2019**, *139*, 2892–2909. [CrossRef]
30. Inoue, K.; Ritz, B.; Brent, G.A.; Ebrahimi, R.; Rhee, C.M.; Leung, A.M. Association of subclinical hypothyroidism and cardiovascular disease with mortality. *JAMA Netw. Open* **2020**, *3*, e1920745. [CrossRef]
31. Klein, I.; Danzi, S. Thyroid disease and the heart. *Circulation* **2007**, *116*, 1725–1735. [CrossRef]
32. Galli, E.; Pingitore, A.; Iervasi, G. The role of thyroid hormone in the pathophysiology of heart failure: Clinical evidence. *Heart Fail. Rev.* **2010**, *15*, 155–169. [CrossRef]
33. Jin, H.-Y.; Weir-McCall, J.R.; Leipsic, J.A.; Son, J.-W.; Sellers, S.L.; Shao, M.; Blanke, P.; Ahmadi, A.; Hadamitzky, M.; Kim, Y.-J. The relationship between coronary calcification and the natural history of coronary artery disease. *Cardiovasc. Imaging* **2021**, *14*, 233–242. [CrossRef] [PubMed]
34. Virmani, R.; Burke, A.P.; Farb, A.; Kolodgie, F.D. Pathology of the vulnerable plaque. *J. Am. Coll. Cardiol.* **2006**, *47*, C13–C18. [CrossRef] [PubMed]
35. Nelson, C.P.; Hamby, S.E.; Saleheen, D.; Hopewell, J.C.; Zeng, L.; Assimes, T.L.; Kanoni, S.; Willenborg, C.; Burgess, S.; Amouyel, P. Genetically determined height and coronary artery disease. *N. Engl. J. Med.* **2015**, *372*, 1608–1618. [CrossRef] [PubMed]
36. Moon, J.; Hwang, I.C. The link between height and cardiovascular disease: To be deciphered. *Cardiology* **2019**, *143*, 114–115. [CrossRef]
37. Stone, P.H.; Saito, S.; Takahashi, S.; Makita, Y.; Nakamura, S.; Kawasaki, T.; Takahashi, A.; Katsuki, T.; Nakamura, S.; Namiki, A. 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* **2013**, *127*, e489–e490. [CrossRef]
38. Liu, X.; Wu, G.; Xu, C.; He, Y.; Shu, L.; Liu, Y.; Zhang, N.; Lin, C. Prediction of coronary plaque progression using biomechanical factors and vascular characteristics based on computed tomography angiography. *Comput. Assist. Surg.* **2017**, *22*, 286–294. [CrossRef]
39. Weng, S.F.; Reps, J.; Kai, J.; Garibaldi, J.M.; Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* **2017**, *12*, e0174944. [CrossRef]
40. Bittencourt, M.S.; Hulten, E.; Ghoshhajra, B.; O'leary, D.; Christman, M.P.; Montana, P.; Truong, Q.A.; Steigner, M.; Murthy, V.L.; Rybicki, F.J. Prognostic value of nonobstructive and obstructive coronary artery disease detected by coronary computed tomography angiography to identify cardiovascular events. *Circ. Cardiovasc. Imaging* **2014**, *7*, 282–291. [CrossRef]
41. Dey, D.; Achenbach, S.; Schuhbaeck, A.; Pflederer, T.; Nakazato, R.; Slomka, P.J.; Berman, D.S.; Marwan, M. Comparison of quantitative atherosclerotic plaque burden from coronary CT angiography in patients with first acute coronary syndrome and stable coronary artery disease. *J. Cardiovasc. Comput. Tomogr.* **2014**, *8*, 368–374. [CrossRef]
42. Papafaklis, M.I.; Muramatsu, T.; Ishibashi, Y.; Lakkas, L.S.; Nakatani, S.; Bourantas, C.V.; Ligthart, J.; Onuma, Y.; Echavarria-Pinto, M.; Tsirka, G. Fast virtual functional assessment of intermediate coronary lesions using routine angiographic data and blood flow simulation in humans: Comparison with pressure wire-fractional flow reserve. *EuroIntervention* **2014**, *10*, 574–583. [CrossRef]