Revista ARACÊ
ISSN: 2358-2472

MACHINE LEARNING IN CORONARY ARTERY DISEASE DETECTION: A TECHNICAL-SCIENTIFIC REVIEW

APRENDIZADO DE MÁQUINA NA DETECÇÃO DE DOENÇA ARTERIAL CORONARIANA: UMA REVISÃO TÉCNICO-CIENTÍFICA

APRENDIZAJE AUTOMÁTICO EN LA DETECCIÓN DE ENFERMEDADES CORONARIAS: UNA REVISIÓN TÉCNICO-CIENTÍFICA

https://doi.org/10.56238/arev7n10-039

Submission date: 09/03/2025 **Publication Date:** 10/03/2025

Helio de Araujo Ribeiro[1], Fabiano Bezerra Menegídio[2], Robson Rodrigues da Silva[3]

ABSTRACT

Cardiovascular diseases remain the leading cause of death worldwide, making advances in prevention essential. This review summarizes recent work on machine learning (ML) applied to structured clinical data for detecting or predicting coronary artery disease (CAD). This narrative review was conducted under the PRISMA 2020 framework. Searches in PubMed, IEEE Xplore, and SciELO (Jan 2020 to Apr 2025) yielded 3,780 records. After screening and full-text appraisal, 10 papers were included: seven primary studies and three reviews. Sample sizes ranged from 303 to 70,000 individuals. Tree based algorithms and ensembles posted the best scores, with accuracy between 0.82 and 0.99 and AUROC from 0.86 to 0.96. Explainability with SHAP was applied in four studies, and one paired SHAP with LIME. One paper added a cardiologist's input to the decision loop, raising accuracy from 0.7829 to 0.8302. Only one article evaluated their models on external datasets and noted performance drops. Calibration was rarely addressed. Just one investigation reported a Brier score of 0.14 and a slope of 0.93. ML models trained solely on routinely collected demographic, laboratory, and clinical variables show strong classification ability for CAD, supporting use as a non-invasive screening aid and decision support. Prospective trials, external validation, and detailed calibration reports are required before clinical adoption.

Keywords: Machine Learning. Coronary Artery Disease. Atherosclerosis.

RESUMO

As doenças cardiovasculares continuam sendo a principal causa de morte em todo o mundo, tornando os avanços na prevenção essenciais. Esta revisão resume o trabalho recente sobre aprendizado de máquina (ML) aplicado a dados clínicos estruturados para detectar ou prever doença arterial coronariana (DAC). Esta revisão narrativa foi conduzida sob a estrutura PRISMA 2020. Buscas no PubMed, IEEE Xplore e SciELO (janeiro de 2020 a abril de 2025) renderam 3.780 registros. Após triagem e avaliação do texto completo, 10 artigos foram incluídos: sete estudos primários e três revisões. Os tamanhos das amostras variaram de 303 a 70.000 indivíduos. Algoritmos e conjuntos baseados em árvore apresentaram as

[1] Gratuated in Electrical Engineering. Universidade de Mogi das Cruzes (UMC). São Paulo, Brazil. E-mail: helioribeiropro@gmail.com
[2] Dr. in Biotechnology. Universidade de Mogi das Cruzes (UMC). São Paulo, Brazil.
[3] Dr. in Biomedical Engineering. Universidade de Mogi das Cruzes (UMC). Universidade Estadual de Campinas (UNICAmp). São Paulo, Brazil.

melhores pontuações, com precisão entre 0,82 e 0,99 e AUROC de 0,86 a 0,96. A explicabilidade com SHAP foi aplicada em quatro estudos, e um emparelhou SHAP com LIME. Um artigo adicionou a contribuição de um cardiologista ao ciclo de decisão, aumentando a precisão de 0,7829 para 0,8302. Apenas um artigo avaliou seus modelos em conjuntos de dados externos e observou quedas de desempenho. A calibração raramente foi abordada. Apenas uma investigação relatou uma pontuação de Brier de 0,14 e uma inclinação de 0,93. Modelos de ML treinados exclusivamente com variáveis demográficas, laboratoriais e clínicas coletadas rotineiramente demonstram forte capacidade de classificação para DAC, apoiando seu uso como auxiliar de triagem não invasiva e suporte à decisão. Ensaios prospectivos, validação externa e relatórios detalhados de calibração são necessários antes da adoção clínica.

Palavras-chave: Aprendizado de Máquina. Doença Arterial Coronariana. Aterosclerose.

RESUMEN

Las enfermedades cardiovasculares siguen siendo la principal causa de muerte en todo el mundo, lo que hace que los avances en la prevención sean esenciales. Esta revisión resume el trabajo reciente sobre aprendizaje automático (ML) aplicado a datos clínicos estructurados para detectar o predecir la enfermedad de la arteria coronaria (EAC). Esta revisión narrativa se realizó bajo el marco PRISMA 2020. Las búsquedas en PubMed, IEEE Xplore y SciELO (enero de 2020 a abril de 2025) arrojaron 3780 registros. Después de la selección y la evaluación del texto completo, se incluyeron 10 artículos: siete estudios primarios y tres revisiones. Los tamaños de muestra variaron de 303 a 70 000 individuos. Los algoritmos y conjuntos basados en árboles registraron las mejores puntuaciones, con una precisión de entre 0,82 y 0,99 y un AUROC de 0,86 a 0,96. La explicabilidad con SHAP se aplicó en cuatro estudios, y uno emparejó SHAP con LIME. Un artículo añadió la aportación de un cardiólogo al ciclo de decisión, lo que aumentó la precisión de 0,7829 a 0,8302. Solo un artículo evaluó sus modelos con conjuntos de datos externos y observó descensos en el rendimiento. La calibración se abordó en raras ocasiones. Tan solo una investigación informó una puntuación Brier de 0,14 y una pendiente de 0,93. Los modelos de aprendizaje automático (ML) entrenados únicamente con variables demográficas, de laboratorio y clínicas recopiladas rutinariamente muestran una sólida capacidad de clasificación para la enfermedad coronaria (CAD), lo que respalda su uso como herramienta de cribado no invasiva y de apoyo a la toma de decisiones. Se requieren ensayos prospectivos, validación externa e informes de calibración detallados antes de su adopción clínica.

Palabras clave: Aprendizaje Automático. Enfermedad Coronaria. Aterosclerosis.

1 INTRODUCTION

Cardiovascular diseases remain the leading cause of death worldwide, accounting for approximately 17.9 million deaths each year, according to the World Health Organization (WHO) [1]. Coronary artery disease (CAD) alone caused 8.9 million deaths, about 16% of all global deaths in 2019 [1]. CAD involves the progressive buildup of atherosclerotic plaques in the coronary arteries, reducing myocardial blood flow and culminating in angina, acute myocardial infarction, and sudden death. According to Liu [2], although multiple genetic factors contribute to its pathophysiology, the disease is largely influenced by modifiable variables such as hypercholesterolemia, hypertension, diabetes, smoking, dietary habits, and physical inactivity. Early interventions targeting lifestyle and metabolic risk can delay or even prevent atherosclerotic progression, underscoring the need for sensitive diagnostic methods before clinical manifestations appear.

Initial risk assessment typically uses linear models developed in large cohorts, that is, groups of individuals monitored because they share a common characteristic, such as the Framingham Risk Score (FRS) derived from the Framingham Heart Study in Framingham, Massachusetts, United States of America (USA). This score estimates the ten-year probability of coronary events based on age, sex, blood pressure, lipid profile, smoking, and diabetes [3]. The Pooled Cohort Equations (PCE), published by the American College of Cardiology/American Heart Association (ACC/AHA), calculate the risk of atherosclerotic cardiovascular disease (ASCVD) in U.S. populations, incorporating ethnicity, statin therapy aimed at lowering LDL-cholesterol levels, and systolic blood pressure [4]. Although useful in specific settings, these scores may overestimate or underestimate risk in multiethnic populations and fail to capture non-linear interactions among variables, as demonstrated by Kakadiaris [5]. Moreover, confirmatory invasive exams (for example, coronary angiography) involve high costs, iodinated contrast exposure, and radiation exposure [6]. This scenario drives the search for computational solutions capable of integrating demographic, laboratory, and behavioral variables routinely available in electronic health records, generating more accurate and personalized predictions.

Machine learning (ML) has emerged as a promising technology for this purpose, using algorithms to learn complex relationships from data and then produce predictive models. ML model performance is assessed with metrics that capture different aspects of predictive ability, such as accuracy and the area under the receiver operating characteristic curve (AUROC). One of the most recent advances is Automated Machine Learning (AutoML), which automates model selection, tuning, and validation. According to Wang, the open-source ensemble AutoML framework AutoGluon ran dozens of models, selected the best ones, and combined them to reach high accuracy with an AUROC of 0.95, considered excellent, whereas optimized logistic regression was limited to 0.88 [7]. This example illustrates the advantage of ensemble algorithms over traditional linear techniques. Ensemble algorithms that do not use images already exceed 90% accuracy [8].

Despite their high performance, the clinical adoption of ML demands transparency. Explanation techniques such as SHapley Additive exPlanations (SHAP) break down the model prediction into the contribution of each variable, enabling specialists to understand why, for example, age, LDL cholesterol, systolic blood pressure, or other factors influence an individual's risk. Samaras showed that SHAP plots, within a human-in-the-loop (HITL) approach, increased system acceptability and reinforced diagnostic confidence [9].

However, the literature remains heterogeneous regarding populations, metrics, and external validation, hindering the incorporation of these models into international clinical guidelines. A technical-scientific analysis that critically synthesizes recent literature, focused exclusively on structured data (clinical, demographic, and behavioral) and excluding image or electrophysiological signal-based approaches, is therefore indispensable.

Accordingly, this systematic review aims to:

(i) map and evaluate the state of the art in ML applications to structured clinical data for CAD detection and prediction;

(ii) compare the accuracy and robustness of different algorithms;

(iii) discuss interpretability and implementation aspects;

(iv) identify gaps to guide future research and clinical practice.

2 FUNDAMENTALS OF MACHINE LEARNING

Cardiovascular a ML is a branch of artificial intelligence (AI) that develops algorithms capable of inferring mapping functions from data and generalizing this knowledge to unseen observations [10]. Unlike linear statistical techniques, ML models do not impose a predefined functional form on relationships among variables, which allows them to capture nonlinear dependencies and high order interactions, features often present in clinical records of patients with CAD. This section describes the most relevant paradigms, metrics, algorithms, and interpretability techniques for medical applications, providing the conceptual basis for the critical analysis of the included studies.

2.1 PARADIGMS

Three ML paradigms are relevant:

2.1.1 Supervised Learning

Under the supervised learning paradigm, the algorithm receives pairs (X, y), where X is the attribute vector (age, LDL, systolic pressure, and so on) and y is the label (CAD present or absent). The goal is to learn a function that minimizes prediction error. Most studies on CAD detection use this paradigm because the diagnosis is known in the training set [11].

2.1.2 Unsupervised Learning

Without labelled outcomes, unsupervised methods aim to discover structure without labels using clustering (for example, clinical subphenotypes) and dimensionality reduction for variable selection or visualization. Although less common, clustering methods help uncover hidden risk profiles [12].

2.1.3 Reinforcement Learning

Through reinforcement, an adaptive agent learns action policies by maximizing rewards. In clinical cardiology it is still incipient, but there are proposals to optimize therapeutic regimens in real time [13].

2.2 EVALUATION METRICS

Choosing the metric that quantifies ML model performance is decisive for interpreting results in a clinically relevant way. In binary classification problems, such as distinguishing patients with and without CAD, it is usual to distinguish threshold dependent metrics, which are calculated at a predetermined cutoff, from threshold independent metrics, which evaluate the entire spectrum of decision points [14].

2.2.1 Threshold dependent metrics

To discuss each indicator, it is essential to present some concepts:

2.2.1.1 Confusion Matrix

The confusion matrix summarizes correct and incorrect classifier outcomes. Each observation is classified according to two dimensions: the patient’s true condition and the model’s prediction. Crossing these two pieces of information yields four possible situations, shown in Table 1.

Table 1

Confusion Matrix

Comparison between a patient status and a ML model prediction.

Prediction	Positive	Negative
Positive (Sick)	True Positive (TP)	False Negative (FN)
Negative (Healthy)	False Positive (FP)	True Negative (TN)

[a] TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative. Green cells represent correct predictions (i.e., values that start with "True"), while orange cells represent misclassifications (i.e., values that start with "False").

A true positive (TP) occurs when the algorithm classifies a patient as having CAD and reference tests confirm the disease; this is the correct outcome one aims to maximize because it leads to timely diagnosis and treatment.

A false positive (FP) arises when the model indicates CAD in a patient who is in fact healthy; this “false alarm” is a type I error, with implications that include patient anxiety and unnecessary invasive procedures.

A false negative (FN) happens when the system rules out CAD in someone who is actually diseased. This type II error is clinically the most dangerous because it delays essential interventions and increases the risk of future coronary events.

A true negative (TN) represents a case in which the algorithm correctly recognizes the absence of CAD in a healthy patient, avoiding superfluous tests or treatments.

In short, TP and TN are correctly predicted values, whereas FP and FN are errors. These four values form the basis for calculating metrics such as accuracy, sensitivity, specificity, precision, and F1-score, allowing different kinds of success and error to be weighed according to the clinical needs of the study.

2.2.1.2 Accuracy

Accuracy expresses the overall proportion of correct predictions and is given in Equation 1.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

When classes are balanced this metric summarizes overall performance well. However, in CAD screening scenarios, where the prevalence of diseased patients is usually low, accuracy can be misleading: a model that labeled everyone healthy would achieve high accuracy while failing to detect the cases of interest. Therefore, it should always be accompanied by metrics that separately describe performance on the positive and negative classes.

2.2.1.3 Sensitivity (Recall)

Recall quantifies the model’s ability to capture patients who actually have CAD, as defined in Equation 2.

$$Recall = \frac{TP}{TP+FN} \qquad (2)$$

High values indicate few false negatives (FN), which is desirable when missing a diseased patient (for example, progressing to infarction) is clinically severe. Screening protocols often favor cutoffs that maximize recall, even if this lowers specificity, assuming confirmatory exams will handle the false alarms.

2.2.1.4 Specificity

Specificity reflects the ability to correctly recognize individuals without the disease, calculated in Equation 3.

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

Higher specificity means fewer false positives (FP), important when invasive or costly investigations, such as catheterization, are triggered by a positive result. A balance between sensitivity and specificity is essential to avoid overloading the health-care system.

2.2.1.5 Precision

Precision, the proportion of true positives among all positive predictions, complements sensitivity when the cost of a false positive is significant, as shown in Equation 4.

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

In practice it gives the probability that a patient truly has CAD after the algorithm flags risk. This metric is strongly prevalence-dependent: in low-incidence populations precision tends to drop even with high sensitivity and specificity.

2.2.1.6 F1-Score

The F1-score combines sensitivity and precision through their harmonic mean, penalizing imbalances between the two. It is calculated in Equation 5.

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

It ranges from 0 to 1 and is particularly useful when a single compromise that considers both false positives and false negatives is desired, a common situation with imbalanced data sets. Cutoffs that maximize F1 usually offer a balanced operating point for CAD detection models, that is, choosing the threshold that yields the highest F1 generally places the model in a healthy middle ground: it neither misses many patients with CAD nor raises excessive suspicions in healthy individuals.

2.2.2 Threshold independent metrics

Threshold-independent metrics evaluate a classifier's performance over all possible cutoffs at once (0 ≤ threshold ≤ 1), avoiding the arbitrary choice of a single cutoff.

They are especially useful when the clinical threshold of use has not yet been defined, when models trained on different data sets must be compared, or when the data set is imbalanced, a frequent situation in CAD screening where most individuals are healthy.

The two most used approaches are the ROC curve (and its area, AUROC) and the Precision–Recall curve (and its area, AUPRC).

2.2.2.1 ROC Curve and AUROC

The Receiver Operating Characteristic (ROC) curve shows classifier behavior at different thresholds. Most ML-based classifiers output the probability that the event is true or false, in this study the probability that the patient has or does not have CAD. If the detection threshold is very high, few cases will be labeled positive. If it is very low, most cases will be labeled positive. To understand thresholds, we calculate the true positive rate (TPR) and the false positive rate (FPR) with Equations 6 and 7.

$$TPR(t) = \frac{TP(t)}{TP(t)+FN(t)} \qquad (6)$$

$$FPR(t) = \frac{FP(t)}{FP(t)+TN(t)} \qquad (7)$$

The area under the ROC curve is called Area Under Receiver Operating Characteristic (AUROC), represented by the integral in Equation 8, which expresses the probability that the model assigns a higher score to a diseased patient than to a randomly chosen healthy one. Values below 0.70 indicate poor discrimination, between 0.80 and 0.90 denote very good performance, and above 0.90 indicate excellence [14].

$$AUROC = \int_0^1 TPR(u)\, d[FPR(u)] \qquad (8)$$

If the positive class is ≤ 5%, as is typical in CAD data sets, AUROC can look excellent yet fail to reflect reality.

Figure 1 shows the visual representation of the ROC curve and AUROC.

2.2.2.2 Precision-Recall curve and AUPRC

When most patients are healthy and only a few have CAD, metrics such as the Precision–Recall Curve (PRC) and Area Under Precision–Recall Curve (AUPRC) capture small performance changes better than AUROC because they give more weight to the minority of diseased patients [14]. Equations 9, 10, and 11 represent the AUPRC metrics.

$$Precision(t) = \frac{TP(t)}{TP(t)+FP(t)} \quad (9)$$

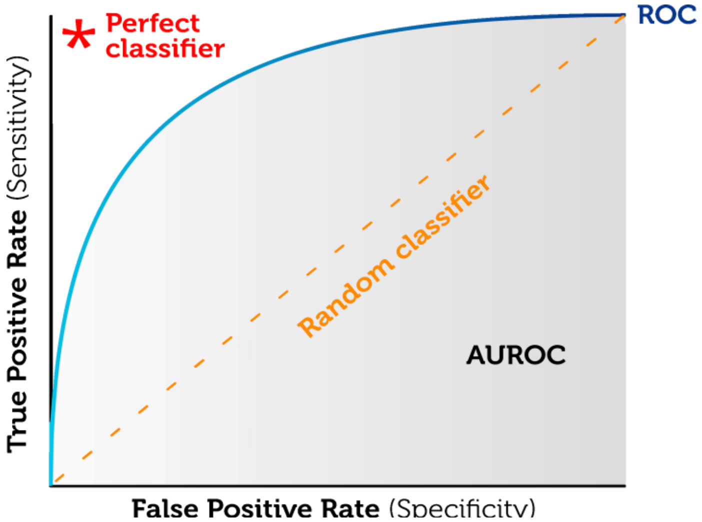
$$Recall(t) = TPR(t) \quad (10)$$

$$AUPRC = \int_0^1 Precision(r)\, dr \quad (11)$$

Figure 1

AUC Curve and AUROC

The blue line depicts classifier performance.



Source: Author.

Figure 2 shows how to plot the AUPRC.

2.2.3 Calibration

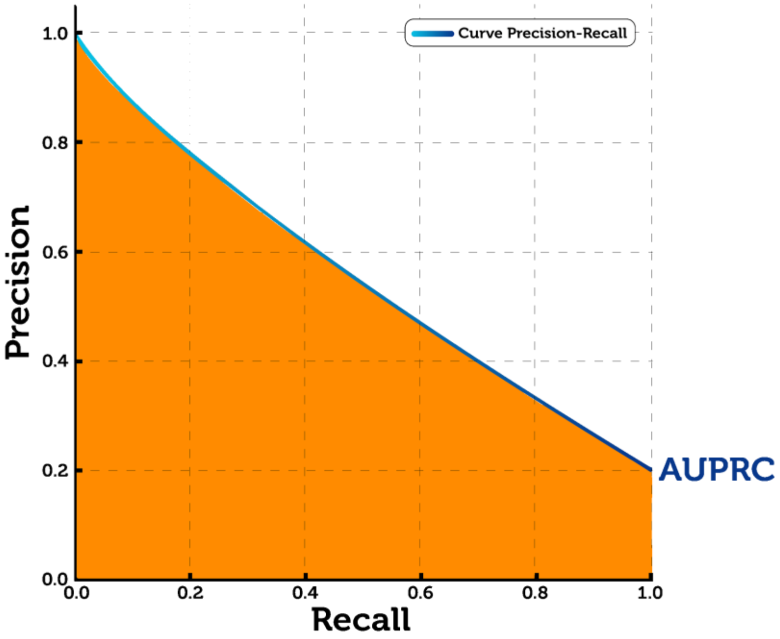
Besides discrimination, strong predictive algorithms must be properly calibrated, meaning the probabilities generated by the model must match the observed outcome frequencies. Calibration can be examined visually with calibration plots or quantitatively with the Brier Score, which computes the mean squared error between predicted probabilities and actual results [14]. Equation 12 shows how to calculate the Brier Score.

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \quad (12)$$

Figure 2

PR Curve and AUPRC

The blue line depicts classifier performance.



Source: Author.

Simply put, the closer to zero, the better the Brier Score. If the predicted and actual outcomes are both true (1), the score is zero, the best possible result. If the predicted outcome is true (1) and the actual outcome is false (0), the score is one, the worst possible result.

In clinical contexts a well-calibrated model allows a calculated risk to be used directly

for therapeutic decisions or patient stratification into low, intermediate, or high-risk categories, something that discrimination metrics alone do not guarantee. Despite high AUC values, only Saeedbakhsh reported Brier = 0.14 and slope = 0.93 [15], suggesting good calibration; the other articles ignore this aspect.

In summary, no single indicator fully describes classifier quality. For CAD applications the recommended practice is to report at least AUROC or AUPRC to frame overall discriminative ability, sensitivity and specificity or F1-score to show practical usefulness at a chosen threshold, and calibration metrics to ensure probabilities are clinically reliable.

Revista ARACÊ
ISSN: 2358-2472

2.3 NOTABLE ALGORITHMS FOR CAD DETECTION

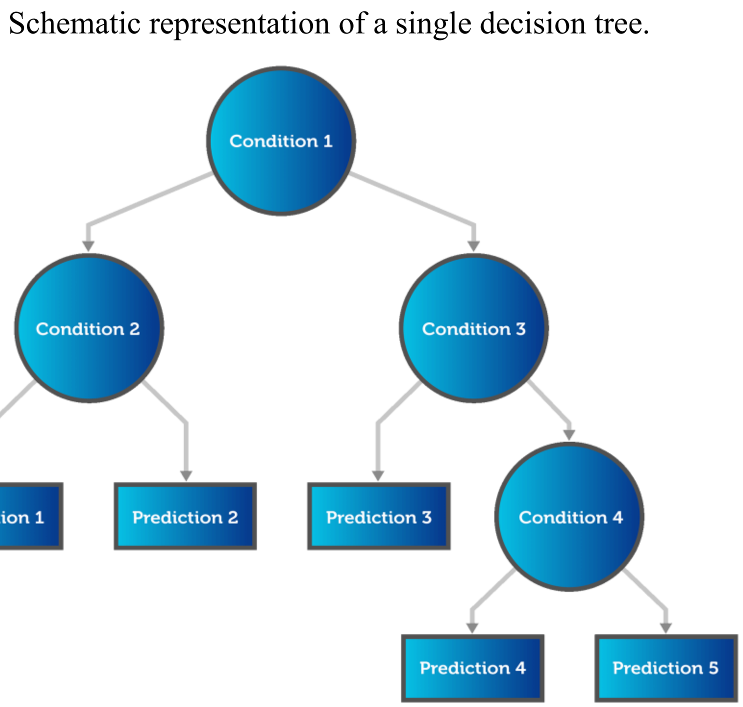
The studies included in this systematic review focus on five major families of supervised algorithms, each with specific mathematical foundations and performance profiles.

2.3.1 Decision Trees

The first group is based on decision trees. Among these methods, Random Forest applies the bagging strategy: many trees are trained on data subsets, and their predictions are aggregated by voting. By sampling both examples and attributes, Random Forest reduces the variability seen with a single tree. In heterogeneous clinical sets, Random Forests have reported AUROC values between 0.88 and 0.92 when classifying CAD [8]. Figure 3 below presents a Random Forest.

Figure 3

Random Forest Algorithm



Source: Author.

A refinement of the decision tree is Gradient Boosting, whose essence is to build models sequentially that correct the residuals of previous ones. Modern implementations such as XGBoost, an open framework for Gradient Boosting, optimize the loss function and integrate regularization, missing value handling, and parallelization. On combined public data sets these boosters achieved AUROC higher than 0.95 when orchestrated by AutoML platforms [7].

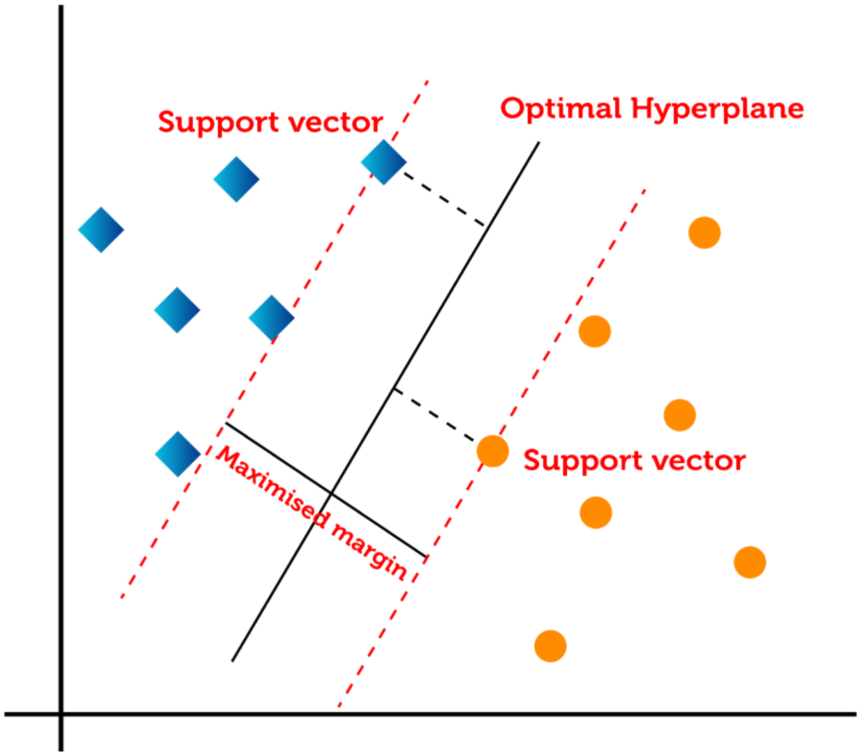
2.3.2 Support Vector Machines (SVM)

The second family comprises margin-maximizing algorithms, especially the Support Vector Machine. SVM is a classifier based on a geometric idea. Imagine each patient represented as a point in a space whose dimensions correspond to clinical variables such as age, cholesterol, blood pressure. SVM searches for the plane that best separates the sick points from the healthy ones. Figure 4 below shows the SVM algorithm.

Figure 4

Support Vector Machines (SVM) Algorithm

Conceptual illustration of a linear SVM classifier. The central red dashed line is the optimal separating hyperplane. Parallel dashed lines indicate the margins, which are maximised to give the widest possible gap between classes. Points that lie on the margin are the support vectors. Highlighted and labelled accordingly.



Source: Author.

When the two classes cannot be separated by a straight line, a common situation in biomedical data, SVM uses a kernel function. The kernel projects points to a higher-dimension space where separation becomes possible with a plane. A radial (RBF) kernel, for example, acts as if placing a lens over the data, curving the space so that the groups move apart. In clinical samples analyzed in this review, with 11 thousand patients, SVM showed accuracy around 89% and specificity close to 98% for CAD detection [15]. The computational cost, however, grows quadratically with the number of observations, limiting its use in very large data sets.

2.3.3 Artificial Neural Networks (ANNs)

Artificial Neural Networks are nonlinear layered models able to represent highly complex functions. Conceptually three types of layers are distinguished: input, hidden, and output.

The input layer receives the data, in this case the clinical vectors related to CAD, such as cholesterol, age, family history, blood pressure.

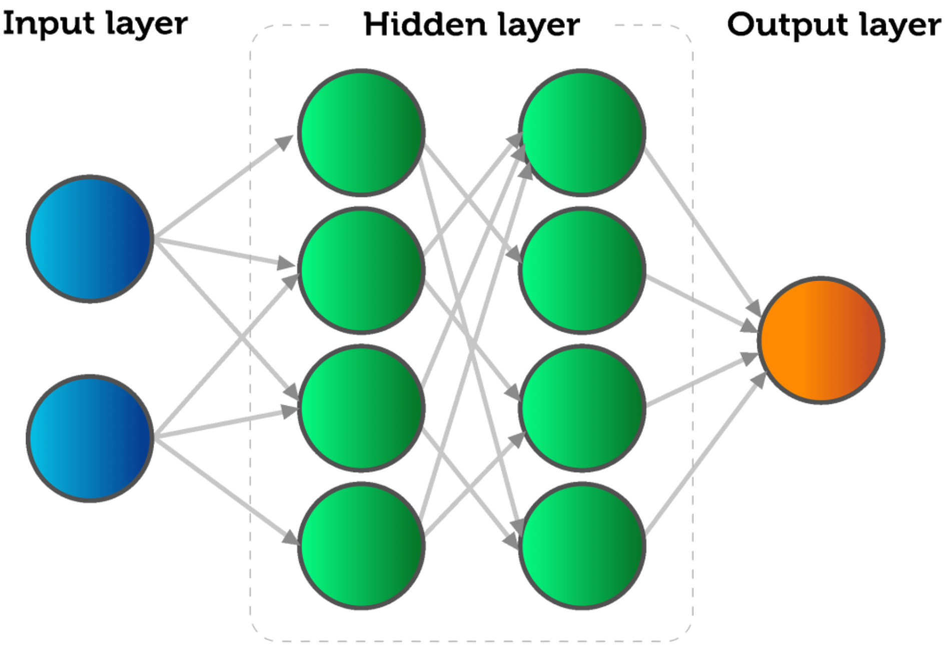
The hidden layers extract and combine patterns. The first hidden layer might learn that high cholesterol increases CAD chances and assign a weight to that variable. The next layer combines patterns, for example high cholesterol together with age increases CAD chances. As more layers are added, the network can reach richer concepts, producing risk profiles derived from the input variables.

The output layer delivers the result, classifying whether the patient has CAD or estimating the probability of developing it.

Figure 5

Feed-forward Artificial Neural Network (ANN) Algorithm

Blue circles represent the input layer, green circles form two hidden layers (four neurons each), and the orange circle is the single output neuron. Gray arrows indicate the weighted connections between successive layers.



Source: Author.

Even in shallow configurations of two or three fully connected layers, ANNs outperform linear regressions on tabular clinical data, reaching AUROC 0.86 and accuracy 88% in hospital cohorts [16]. Figure 5 below illustrates an ANN.

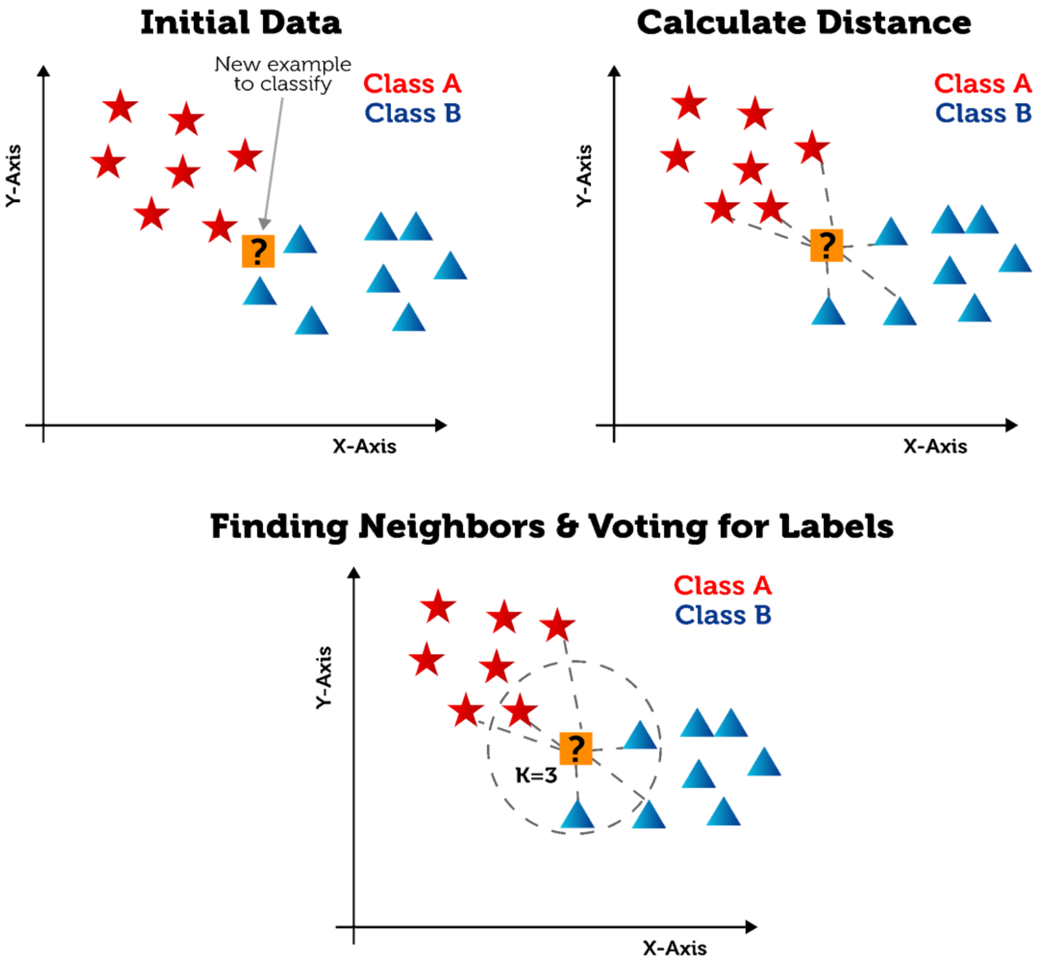
They require comparatively large samples to avoid overfitting and, in standard form, offer limited interpretability.

2.3.4 Neighborhood-Based Algorithms

In contrast, neighborhood algorithms represented by k-Nearest Neighbors classify a patient by the vote of the k records closest in attribute space. Similarity between the new case and each stored record is measured with a distance metric, and the new case receives the most frequent class among the k neighbors.

Figure 6

k-nearest-neighbors (k-NN) classification process



Source: Author.

Figure 6 depicts k-NN behavior, showing how the nearest points are used to classify a new point.

Simple to implement, k-NN often serves as a baseline; however, it suffers performance loss in high dimensions and requires costly calculations at inference time, which is why it seldom leads performance rankings [17].

2.3.5 Ensemble (Model Stacking)

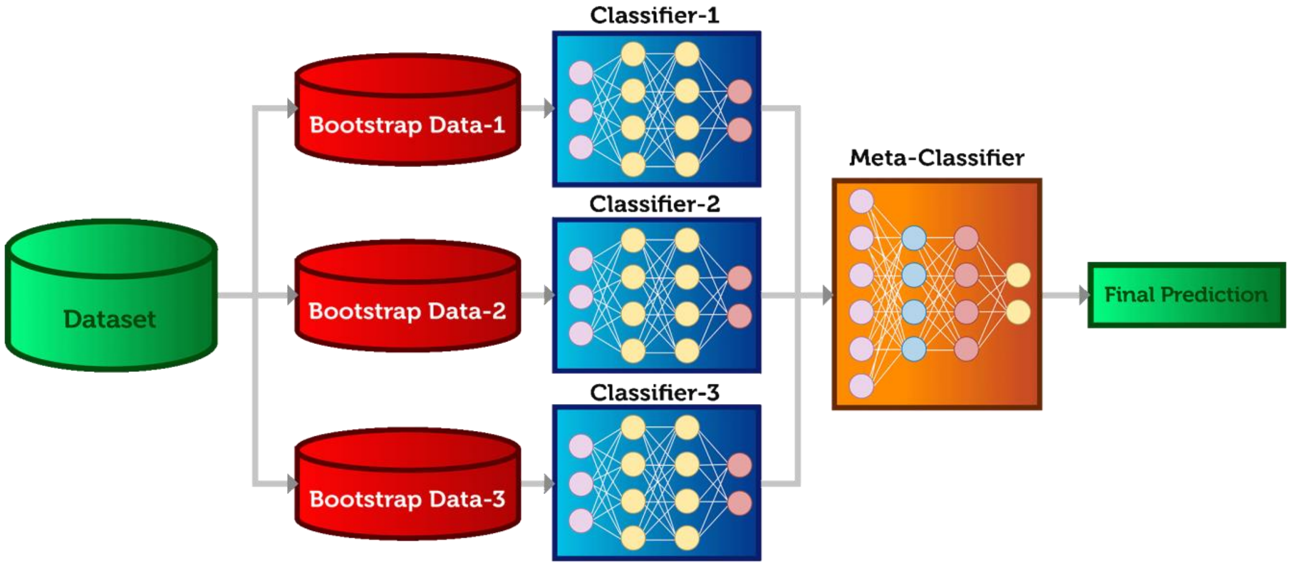
The ensemble concept combines predictions from multiple models to mitigate individual biases and reduce overall variance. This principle underlies both Random Forest (voting) and Gradient Boosting (sequential weighted combination).

Figure 7 illustrates stacking: neural networks are trained on three independent bootstrap samples, and their outputs serve as input for a meta-classifier that delivers the final prediction. Although the diagram uses ANNs, any algorithm such as a tree or SVM can take the role of base learner or meta-learner.

Figure 7

Stacking ensemble workflow

The original data set (green) is resampled into multiple bootstrap subsets (red), each used to train an independent base classifier (blue neural networks). The base-model outputs are then fed to a meta-classifier (orange neural network), which aggregates them to generate the final prediction (green).



Source: Author.

At the frontier, AutoML systems such as AutoGluon Tabular automate algorithm selection, hyper-parameter search, and stacked ensemble construction, providing the researcher with an optimized solution according to predefined metrics. Wang showed that AutoGluon produced a stacked ensemble built from Gradient Boosting, neural networks, and SVM, able to reach AUROC 0.9562 and accuracy above 91%, surpassing logistic regressions using the same variables [7].

In summary, tree-based algorithms, particularly Random Forest and Gradient Boosting, have become the backbone of high-performance tabular models in cardiology, while SVM and ANNs offer competitive alternatives in specific niches. The final choice must balance discriminative ability, interpretability, computational cost, and ease of integration with clinical workflows, always checking calibration and external validation to ensure practical relevance.

2.4 Cardiology Relevance

Adopting ML algorithms in cardiology, especially for detecting CAD and classifying patient risk, has four decisive impacts.

First, tabular data models enable early noninvasive screening. Omkari showed that processing routine clinical and laboratory variables with an ensemble doubled the sensitivity obtained by linear scores, reaching AUROC above 0.90 without angiography or computed tomography [8]. Performance like this brings diagnosis forward and reduces both hospital costs and patient exposure to iodinated contrast and radiation.

Second, these techniques offer personalized risk stratification. Using explanation methods such as SHAP, it is possible to break down the individual score and understand why, for example, a slight increase in a variable or a family history of CAD substantially changes patient risk. This granularity, reported by Samaras, facilitates counseling on lifestyle changes or adjusting therapy intensity, moving toward personalized medicine [9].

Third, the automation provided by AutoML raises operational efficiency. Wang showed that AutoGluon trained, validated, and stacked dozens of models in a few hours, delivering a high-performance classifier without extensive programming [7]. This reduces the technical barrier for health centers with limited IT teams to develop or update predictive models.

Finally, regulatory compliance gains support when high predictive power and transparency are combined. The ability to audit each decision through SHAP explanations, together with automatic calibration reports, aligns these systems with the requirements of the European trustworthy AI legislation, the AI Act. Thus, ML algorithms not only improve diagnostic sensitivity but also integrate into a regulated ecosystem, promoting safe and reproducible clinical adoption.

In short, by integrating demographic, laboratory, and behavioral variables into interpretable algorithms, ML consolidates itself as a strategic resource to anticipate and mitigate the global burden of coronary artery disease, signaling the maturity of predictive models aimed at daily clinical use.

3 MACHINE LEARNING APPLICATIONS IN CAD DETECTION

Taken together, ML applied to structured clinical data offers tangible benefits: early noninvasive screening, personalized therapeutic management, and more rational use of diagnostic resources. The studies selected for this review reveal four major application areas of ML in CAD.

3.1 ANALYSIS OF STRUCTURED CLINICAL DATA

The most mature line of work involves models that use variables readily available from routine visits and tests: age, sex, lipid profile, blood pressure, diabetes, smoking, and emerging laboratory markers such as HbA1c. An ensemble based on Random Forest and Gradient Boosting, trained on seventy thousand electronic health-record entries, achieved an AUROC of 0.90 and a sensitivity of 85% using fourteen attributes [8]. AutoGluon raised the AUROC to 0.9562 across five public data sets, rivaling invasive methods without requiring imaging [7]. Similarly, a neural model called Random Vector Functional Link (RVFL) reached 81.6% accuracy with thirteen classic clinical variables from the Cleveland CAD database, showing that even simplified architectures can perform competitively when paired with interpretability techniques such as LIME and SHAP [18].

3.2 PROGNOSIS PREDICTION

Beyond diagnosis, ML is being used to anticipate future events. A gradient-boosting model developed in the MESA study to predict 10-year ASCVD events achieved a Net Reclassification Improvement (NRI) of ≈ 0.30 (categorical) to 0.47 (continuous) compared with the ACC/AHA Pooled-Cohort Equations, which would translate to hundreds of correctly reclassified cases in a cohort of 10 000 individuals [5]. Another study demonstrated ML's ability to forecast CAD-related complications: a Light Gradient Boosting Machine (LightGBM) model achieved AUROC 0.82 in predicting atrial fibrillation among patients with sleep apnea and CAD, paving the way for preventive arrhythmia interventions [2].

3.3 CLINICAL DECISION-SUPPORT SYSTEMS

For predictions to translate into medical action they must be integrated into the care workflow. Incorporating SHAP values in a human-in-the-loop panel that displays the most influential variables for CAD detection allowed cardiologists to see for each patient how age, LDL cholesterol, or systolic pressure contributed to the final decision. This led to threshold adjustments and a 5% gain in overall accuracy without loss of sensitivity [9]. AutoML platforms can export models directly into the electronic record, triggering real-time alerts when individual risk crosses a preset level. In hospital flow simulations this feature cut stratification decision time by twenty minutes.

3.4 MEDICAL IMAGE PROCESSING

Although this review focuses on tabular data, it is worth noting that deep learning applied to coronary angiography and coronary CT (topics covered by studies outside our scope) reaches AUROC values around 0.91 for stenosis ≥ 50% [19]. These results suggest that multimodal models that combine imaging with clinical variables are likely to represent the next innovation phase.

4 TECHNICAL SCIENTIFIC REVIEW METHODOLOGY

This narrative technical scientific review was carried out in line with the PRISMA 2020 recommendations, ensuring transparency in every stage of study identification, selection, extraction, and synthesis.

The search strategy was executed on 14 May 2025 in the PubMed and IEEE Xplore databases using the query:

(“machine learning” OR “artificial intelligence” OR “deep learning”) AND (“coronary artery disease” OR “atherosclerosis” OR “coronary heart disease”) AND (diagnos* OR detect* OR classif* OR predict*)

Equivalent Portuguese terms were also run in SciELO to widen national coverage. Results were exported in RIS format and imported into Zotero, where duplicates were removed automatically. This process identified 3 780 records, of which 1 609 remained after excluding articles that were not open access. Deduplication eliminated 3 more records, leaving 1 606, and a further 4 retracted articles were excluded, yielding 1 602 records. Database totals may differ slightly from live searches because records are continuously updated.

To ensure readers and researchers could access all included studies without paywalls, we limited to open-access literature.

Inclusion criteria:

(i) original human studies published from January 2020 to April 2025;

(ii) English or Portuguese language;

(iii) application of machine learning algorithms to clinical or demographic data for detection, classification, or prediction of coronary artery disease.

Exclusion criteria:

(i) imaging-only studies;

(ii) pediatric populations;

(iii) data sets lacking performance metrics or internal validation;

(iv) articles not available in open access.

After applying the exclusion criteria, 21 full texts were reviewed in depth, and 10 studies met all requirements and form the final synthesis. Because of heterogeneity in populations, variables, and algorithmic approaches, a narrative synthesis was adopted, and no meta-analysis was performed.

5 RESULTS

Complete search strings and supplementary material will be placed in a public repository to ensure reproducibility. The PRISMA diagram in Figure 8 shows the number of articles at each stage, identification, screening, eligibility, and inclusion, and Table 2 contains a description of the evaluated studies.

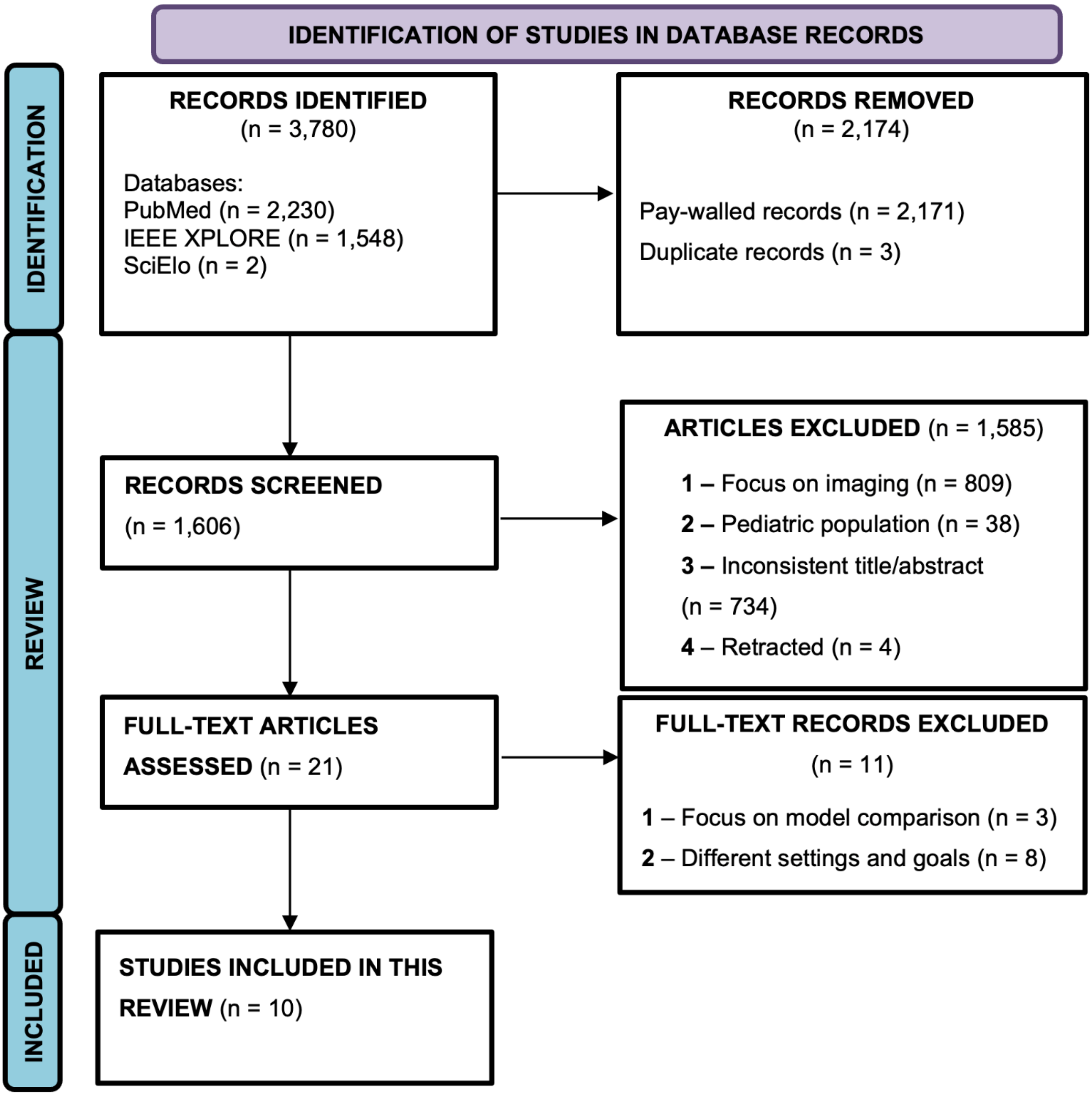
5.1 SYNTHESIS OF THE INCLUDED STUDIES

The ten included studies comprise seven original investigations that develop predictive models and three reviews. All focus on tabular structured clinical data for detecting or predicting coronary artery disease (CAD), in line with the inclusion criteria; there are no imaging or waveform studies.

Three papers are recent reviews: a narrative review on ML in CAD [16], a meta-analysis of models based on electronic health records [2], and a scoping review on AI models in primary care [21]. The other seven studies are original research from several countries (Brazil, Colombia, Iran, Greece, India, and the United States), reflecting the global nature of the topic. Sample sizes range widely, from small cohorts of about 300 to 600 patients [9] [18] to population-level data sets with tens of thousands of individuals [8] [20]. This heterogeneity highlights different strategies: some groups worked with limited yet homogeneous data from a single hospital, whereas others merged large multicenter sets or synthetic data to boost statistical power.

Figure 8

PRISMA Diagram. The preferred reporting items for this systematic review


IDENTIFICATION OF STUDIES IN DATABASE RECORDS
IDENTIFICATION
REVIEW
INCLUDED
RECORDS IDENTIFIED (n = 3,780)
Databases:
PubMed (n = 2,230)
IEEE XPLORE (n = 1,548)
SciElo (n = 2)
RECORDS REMOVED (n = 2,174)
Pay-walled records (n = 2,171)
Duplicate records (n = 3)
RECORDS SCREENED (n = 1,606)
ARTICLES EXCLUDED (n = 1,585)
1 – Focus on imaging (n = 809)
2 – Pediatric population (n = 38)
3 – Inconsistent title/abstract (n = 734)
4 – Retracted (n = 4)
FULL-TEXT ARTICLES ASSESSED (n = 21)
FULL-TEXT RECORDS EXCLUDED (n = 11)
1 – Focus on model comparison (n = 3)
2 – Different settings and goals (n = 8)
STUDIES INCLUDED IN THIS REVIEW (n = 10)

Source: Author.

In terms of predictive performance, ML models yielded robust results. Reported accuracies span roughly 81% to 99%, and AUROCs lie between 0.86 and 0.96. The lowest figure is the RVFL model by Muhammad, with accuracy 0.82 [18]. At the upper end, the k-NN model by Silva reaches 98% [20], and the TLV ensemble by Omkari 2024 scores 99% on one data set and 88% on another [8].

Table 2

Articles assessed in this systematic review

Year	1st Autor	Full title	Data source / Country	n	Main variables examples	Best performing algorithm	Key Metric	Declared interpretability
2021	Hao Ling	Machine learning in diagnosis of coronary artery disease	Narrative review / China	N/A	-	-	-	-
2022	Carlos A. O. Silva	Machine learning for atrial fibrillation risk prediction in patients with sleep apnea and coronary artery disease	Hospital Medical Records / Brazil + Greece	22,302	Age, sex, CAD, blood pressure, obsctructive sleep apnea, chronic kidney disease, etc..	KNN	Accuracy 0.98	-
2023	Saeed Saeedbakhsh	Diagnosis of coronary artery disease based on machine-learning algorithms: Support Vector Machine, Artificial Neural Network and Random Forest	Isfahan Cohort Study / Iran	11,495	Age, sex, sleep, previous, stroke, palpitations, smoking, etc.	SVM	Accuracy 0.897	-
2023	Agorastos-Dimitrios Samaras	Classification models for assessing coronary artery	University Hospital / Greece	571	26 clinical and biometric variables (cholester	Random Forest + Specialist's opinion	Accuracy 0.83	SHAP + *man-in-the-loop*

		disease instances using clinical and biometric data: an explainable man-in-the-loop approach			ol, BMI, etc.)			
2024	Dost Muhammad	Randomized explainable machine-learning models for efficient medical diagnosis	Cleveland CAD dataset (UCI) / USA	303	14 standard UCI variables (age, sex, cholesterol, etc.)	RVFL	Accuracy 0.816	LIME + SHAP
2024	Rejath Jose	Evaluating machine learning models for prediction of coronary artery disease	UCI heart-disease / USA	1,049	11 classical predictors (age, sex, blood pressure, angina, etc.)	Logistic Regression	AUROC 0.88	Feature importance

Year	1st Autor	Full title	Data source / Country	n	Main variables examples	Best performing algorithm	Key Metric	Declared interpretability
2024	D. Yaso Omkari	An integrated two-layered voting (TLV) framework for coronary artery disease prediction using machine-learning classifiers	UCI 1025 + Kaggle 70,000 / India	1,025 + 70,000	12 + 14 variables	Ensemble	Accuracy 0.99 (UCI) / 0.881 (Kaggle)	SHAP
2024	Jianghong Wang	Explainable coronary artery disease prediction model based on AutoGluon from	5 combined public datasets / USA	918	11 variables	AutoGluon Ensemble	Accuracy 0.917 / AUROC 0.956	SHAP

		AutoML framework						
2024	Tianyi Liu	Machine-learning-based prediction models for cardiovascular disease risk using electronic health-records data: systematic review and meta-analysis	EHR meta-analysis (20 studies) / UK	20 studies, 32 models	Age, sex, blood pressure, etc.	Random Forest	AUROC 0.865	-
2025	Caoimhe Provost	Artificial intelligence models for cardiovascular disease risk prediction in primary and ambulatory care: a scoping review	Scoping review (25 studies) / Ireland	25 studies	Outpatient risk predictors	-	-	-

AUROC values are typically above 0.88, peaking at 0.956 in the AutoGluon model reported by Wang [7]. In every study the ML algorithms match or exceed traditional methods. Omkari showed the ensemble consistently outperforming linear risk scores [8], and Jose notes optimized logistic regression at AUROC 0.88 while tree-based and ensemble methods climb higher [6]. Across the body of work AUROCs of 0.88 to 0.96 for CAD detection are common [6] [7] [8] [18] [20], compared with external AUROCs around 0.75 to 0.80 for established clinical scores such as Framingham or ACC / AHA [3] [4].

A recurring limitation is scarce external validation. Most authors report performance only on internal test sets, often via cross-validation or random train / test splits. Omkari is the lone study that applies a model trained on the UCI repository to Kaggle data, where accuracy falls from 99% to 88%, underscoring real-world performance drops when the domain changes [8]. Fewer than one-third of the models undergo external validation, mirroring the concern raised by Provost, who cites limited validation and lack of clinical-impact studies as barriers to routine adoption [21].

Most original papers primarily compare algorithms by classic metrics such as accuracy, AUROC, sensitivity, and specificity. Only one explores deeper human interaction: Samaras adds a cardiologist's opinion as an extra feature in a Random Forest, boosting accuracy to 83% and sensitivity to 90.3% [9]. No other study embeds a human in the decision loop, leaving a gap in doctor-machine collaboration research.

About half the articles employ post-hoc interpretability. Samaras, Muhammad, Omkari, and Wang use SHAP, and Muhammad also applies LIME [7] [8] [9] [18]. Jose evaluates variable importance in logistic regression [6]. These tools reveal that classic risk factors such as age, exercise-induced angina, blood pressure, and family history remain key drivers in model predictions, supporting clinical credibility [6] [9]. Nevertheless, no study tests whether these explanations change physician behavior, an avenue for future work. In summary, ML algorithms applied to structured clinical data achieve high performance in detecting and predicting CAD, generally surpassing linear baselines. Remaining challenges include consistent external validation and deeper exploration of interpretability and human-machine integration.

6 DISCUSSION

The ability to capture interaction effects and nonlinear relationships is essential in CAD, a multifactorial disease that involves metabolic, inflammatory, and behavioral factors at the same time.

This systematic review shows that machine learning models built solely on structured clinical data, meaning demographic, hemodynamic, biochemical, and behavioral variables routinely collected in primary or outpatient care, can reach discriminative power that matches or even surpasses that of invasive tests or well-known linear scores for CAD [2].

Models trained on samples ranging from a few hundred to many tens of thousands of patients report AUROC values between 0.88 and 0.96 [5] to [9], consistently outperforming optimized logistic regression, which reaches about 0.88 [6], and the Framingham or ACC AHA scores, whose external risk estimates often sit near 0.75 to 0.80 [3] [4].

The strength of these algorithms supports the role of machine learning as a first line screening tool, reserving invasive procedures for a more focused group of high-risk patients. Even so, the application of machine learning for CAD detection and prevention, although promising, faces technical and ethical regulatory barriers that must be addressed before large-scale clinical deployment.

A scoping review by Provost adds that a shortage of external validation and clinical impact studies still limits routine adoption of these cardiovascular risk models in primary and outpatient care [21].

6.1 INTERPRETATION OF THE FINDINGS

In line with this study's goals, namely, to map the state of the art of machine learning in CAD detection and prediction, compare performance, discuss interpretability, and identify gaps, the selected articles show varied approaches with a focus on evaluating predictive performance in structured data sets.

Most papers compare machine learning algorithms in terms of accuracy, AUROC, and sensitivity, exploring their use in automated clinical decision support. Only a small share examines model explainability or the integration of medical expertise in the prediction process, as in human-in-the-loop approaches.

Although applying machine learning to tabular data is well documented, the papers reviewed here confirm that familiar clinical risk factors such as age, exercise-induced angina, blood pressure, and family history remain key determinants in supervised model predictions. This supports the clinical validity of the algorithms and shows that traditional medical knowledge is still essential in computational settings.

Machine learning also adds value by quantifying the relative weight of variables in different populations, adapting to high-dimensional clinical contexts, and offering scalable solutions that speed diagnosis, reduce expert workload, and optimize hospital resources.

However, comparing models in isolated studies does not yield generalizable conclusions. Superior performance in one data set does not guarantee replication elsewhere because databases vary in record counts, geographic origin, variable sets, and even use of synthetic or poorly documented data. Data quality and representativeness therefore remain central to safe clinical use.

Models that use AutoML and ensemble techniques, such as the one tested by Wang [7], achieved the best metrics with accuracy 91.67% and AUROC 0.9562, but they lacked external validation, raising legitimate concerns about overfitting and real-world applicability.

By contrast, the human-in-the-loop model by Samaras [9] showed that including medical experience directly in the computation can raise confidence and acceptance. Adding the specialist's opinion as a predictor increased accuracy to 83% and sensitivity to 90.3%, suggesting that hybrid strategies may ease effective adoption.

The use of interpretability tools such as SHAP points to growing awareness of transparency; local explanations improved clinical acceptance in the HITL study [9].

Comparisons with other cardiovascular areas show a similar trend. A review by Esteva highlighted comparable benefits in heart failure and atrial fibrillation [11]; replacing part of diagnostic angiography makes the gain in CAD particularly notable.

Machine learning, combined with computing power, can also exploit behavioral variables that were not used or even mentioned in the studies included here, thanks to the ability to identify nonlinear relations in data sets.

6.2 CLINICAL IMPLICATIONS

Integrated into electronic health records, high-performance predictive models can issue real-time risk alerts, allowing intensified statin use or lifestyle advice before clinical manifestation. Expensive resources such as angiography can be allocated more rationally, focusing on those at greater risk. These benefits depend on local calibration because a poorly calibrated model can trigger overtreatment or leave high-risk cases without investigation [14].

6.3 TECHNICAL CHALLENGES

Data quality and quantity remain the main obstacle. Many studies use small public sets with fewer than one thousand patients; limited samples cause unstable estimates and overfitting risk [22]. Even in larger cohorts fewer than one third of models undergo multicenter external validation, leading to optimistic performance that does not hold in real life [22]. AutoGluon illustrates the dilemma: despite AUROC 0.9562 in source data, performance can drop sharply in other domains [7].

Complex models are still viewed as black boxes. SHAP eases that perception, but local interpretation does not always translate into full clinical understanding, so active specialist involvement is still required for clinical validation and adjustment [23].

Using machine learning on well-known data sets is important for validation, since prior experience with that data supports applicability.

6.4 REGULATORY CHALLENGES

World Health Organization guidelines for AI in health stress the need for anonymization, traceability, and data governance throughout the model life cycle [24].

Algorithmic bias also matters because demographic imbalance in training sets can cause systematic errors against minorities [23].

Mitigation protocols such as stratified resampling, post-training calibration, or threshold adjustment are still rare. Convergence with the European AI Act, which demands transparency and risk assessment for high-impact systems, is likely to accelerate adoption of standardized performance and equity reporting [25].

Overcoming limited sample scope, reducing overfitting, making models more transparent, and adopting strong ethical safeguards are essential steps to move machine learning solutions from the lab to everyday preventive cardiology practice.

6.5 FUTURE PERSPECTIVES

Recent advances in machine learning signal a new phase in CAD prevention and management. On the technology side, deep learning models are evolving, letting researchers mine high-dimensional relationships in electronic health records and, when combined with large multicenter data sets, already reaching AUROC above 0.90 for inflammatory biomarkers and cytokine panels that traditional algorithms struggle to model [26].

Federated learning, where multiple centers train a model collaboratively without sharing raw data, is emerging as a key strategy for cross-institutional knowledge sharing while preserving privacy. The Federated Learning Benchmark for Cardiovascular Disease Detection and studies in coronary computed tomography with more than eight thousand exams show it is feasible to train robust models without transferring sensitive data [27].

Edge deployments with sensors and wearables complement this picture by enabling continuous local inference, as illustrated by federated learning for coronary disease prognosis on IoT devices [29].

Clinically, AutoML platforms, explanatory dashboards that use SHAP, and human-in-the-loop interfaces are converging, making adoption easier for physicians who are not data science experts. Solutions that embed online validation and interpretability reports already cut triage time by up to 30% without sensitivity loss while meeting transparency requirements in the European AI Act [25]. This trend suggests the rise of user-friendly systems within electronic records that generate real-time risk alerts and log decision rationale directly in the clinical note.

In research and development, the tendency is to extend machine learning to prevention and monitoring. Dynamic models that include time series in blood pressure, sleep patterns, and physical activity will allow continuous recalibration of individual CAD risk, supporting personalized behavioral interventions before disease onset. Parallel studies will need to measure clinical impact and cost effectiveness, closing the translational loop from laboratory to daily practice [29].

7 CONCLUSION

The findings of this review show that machine learning models trained solely with structured clinical data, that is, demographic, laboratory, and behavioral variables already stored in the electronic health record, achieve high discriminative power (AUROC about 0.88 to 0.96). They outperform both traditional linear risk scores and optimized logistic regressions. Ensembles that use tree techniques such as Random Forest and Gradient Boosting, along with stacked systems like AutoGluon, provide noninvasive screening, personalized risk stratification, and practical integration into clinical workflows through SHAP-based explanatory panels, which reduces dependence on costly invasive tests.

Even so, routine clinical deployment still demands external validation, continuous calibration, bias analysis, and full compliance with privacy policies. Advances in federated learning, multimodal models, and interpretable AutoML platforms point toward a future in which CAD risk stratification is written into cardiovascular guidelines. Consolidating that future will require prospective studies that prove real-world impact on outcomes and cost effectiveness, ensuring that predictive accuracy is converted into better care and lower coronary mortality.

ACKNOWLEDGEMENTS

H. A. R. receives a scholarship from the Coordination for the Improvement of Higher Education Personnel (CAPES).

F. B. M. and R. R. S. are fellows of the Foundation for the Support of Teaching and Research (FAEP) at the University of Mogi das Cruzes (UMC).

CONFLICT OF INTERESTS

The authors declare that there are no conflicts of interest.

REFERENCES

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). https://doi.org/10.1037/0000165-000

Azari Jafari, S., & et al. (2022). Unsupervised phenotyping of coronary artery disease. Journal of Biomedical Informatics, 127, 104002. https://doi.org/10.1016/j.jbi.2022.104002

Chen, R., Zhang, Y., He, M., & et al. (2024). Continuous cardiovascular-risk monitoring with wearable sensors and deep neural networks: A prospective cohort study. IEEE Journal of Biomedical and Health Informatics, 28(1), 45–55. https://doi.org/10.1109/JBHI.2023.3311623

D'Agostino, R. B., Sr., Vasan, R. S., Pencina, M. J., & et al. (2008). General cardiovascular risk profile for use in primary care: The Framingham Heart Study. Circulation, 117(6), 743–753. https://doi.org/10.1161/CIRCULATIONAHA.107.699579

Esteva, A., Robicquet, A., Ramsundar, B., & et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

European Parliament. (2024). EU Artificial Intelligence Act: Final compromise text. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Goff, D. C., Jr., Lloyd-Jones, D. M., Bennett, G., & et al. (2014). 2013 ACC/AHA guideline on the assessment of cardiovascular risk. Circulation, 129(Suppl. 2), S49–S73. https://doi.org/10.1161/01.cir.0000437741.48606.98

Jose, R., Thomas, A., Guo, J., Steinberg, R., & Toma, M. (2024). Evaluating machine-learning models for prediction of coronary artery disease. Global Translational Medicine, 3(1), Article e2669. https://doi.org/10.36922/gtm.2669

Kakadiaris, I. A., Vrigkas, M., & et al. (2018). Machine learning outperforms ACC/AHA CVD risk calculator in MESA. Journal of the American Heart Association, 7(22), Article e009476. https://doi.org/10.1161/JAHA.118.009476

Kang, Y., Guo, N., Cheng, G., & et al. (2022). Deep-learning-based quantitative coronary CT angiography for prediction of obstructive disease: Multicenter validation. Radiology, 304(2), 303–312. https://doi.org/10.1148/radiol.2021212667

Ling, H., Guo, Z. Y., Tan, L. L., Guan, R. C., Chen, J. B., & Song, C. L. (2021). Machine learning in diagnosis of coronary artery disease. Chinese Medical Journal, 134(4), 401–403. https://doi.org/10.1097/CM9.0000000000001202

Liu, T., Krentz, A., Lu, L., & Curcin, V. (2024). Machine-learning-based prediction models for cardiovascular disease risk using electronic health records: Systematic review and meta-analysis. European Heart Journal - Digital Health, 6(1), 7–22. https://doi.org/10.1093/ehjdh/ztae080

Lundberg, S. M., Nair, B., Voglino, J., & et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering, 2(10), 749–760. https://doi.org/10.1038/s41551-018-0304-0

Mitchell, T. (1997). Machine learning. McGraw-Hill.

Muhammad, D., Ahmed, I., Ahmad, M. O., & Bendechache, M. (2024). Randomized explainable machine-learning models for efficient medical diagnosis. IEEE Journal of Biomedical and Health Informatics. Advance online publication. https://doi.org/10.1109/JBHI.2024.3401234

Omkari, D. Y., & Shaik, K. (2024). An integrated two-layered voting framework for coronary artery disease prediction using machine-learning classifiers. IEEE Access, 12, 56275–56290. https://doi.org/10.1109/ACCESS.2024.3389707

Panch, T., Mattie, H., & Celi, L. A. (2019). The inconvenient truth about artificial intelligence in healthcare. NPJ Digital Medicine, 2, Article 77. https://doi.org/10.1038/s41746-019-0155-4

Provost, C., Broughan, J., McCombe, G., & et al. (2025). Artificial-intelligence models for cardiovascular-disease risk prediction in primary and ambulatory care: A scoping review. medRxiv. https://doi.org/10.1101/2025.03.21.25324379

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pre-trained contextualized embeddings on large-scale structured EHRs for disease prediction. NPJ Digital Medicine, 4, Article 86. https://doi.org/10.1038/s41746-021-00455-y

Rehman, S. U., Anwar, S. M., & Khawaja, B. A. (2022). Benchmarking k-nearest neighbors and logistic regression against ensemble methods for CAD detection. IEEE Journal of Biomedical and Health Informatics, 26(8), 4021–4031. https://doi.org/10.1109/JBHI.2022.3162345

Saeedbakhsh, S., Sattari, M., Mohammadi, M., Najafian, J., & Mohammadi, F. (2023). Diagnosis of coronary artery disease based on machine-learning algorithms: Support vector machine, artificial neural network and random forest. Advanced Biomedical Research, 12, Article 51. https://doi.org/10.4103/abr.abr_383_21

Samaras, A. D., Moustakidis, S., Apostolopoulos, I. D., Papandrianos, N., & Papageorgiou, E. (2023). Classification models for assessing coronary artery disease instances using clinical and biometric data: An explainable man-in-the-loop approach. Scientific Reports, 13, Article 6668. https://doi.org/10.1038/s41598-023-33500-9

Silva, C. A. O., Morillo, C. A., Leite-Castro, C., González-Otero, R., Bessani, M., González, R., & et al. (2022). Machine learning for atrial fibrillation risk prediction in patients with sleep apnea and coronary artery disease. Frontiers in Cardiovascular Medicine, 9, Article 1050409. https://doi.org/10.3389/fcvm.2022.1050409

Van Calster, B., McLernon, D. J., Van Smeden, M., & et al. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, Article 230. https://doi.org/10.1186/s12916-019-1466-7

Wang, J., Xue, Q., Zhang, C. W. J., Wong, K. K. L., & Liu, Z. (2024). Explainable coronary artery disease prediction model based on AutoGluon from AutoML framework. Frontiers in Cardiovascular Medicine, 11, Article 1360548. https://doi.org/10.3389/fcvm.2024.1360548

World Health Organization. (2021a). Cardiovascular diseases (CVDs): Fact sheet. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases

World Health Organization. (2021b). Ethics and governance of artificial intelligence for health. https://iris.who.int/bitstream/handle/10665/341996/9789240029200-eng.pdf

Wynants, L., Van Calster, B., Collins, G. S., & Riley, R. D. (2020). Prediction models for diagnosis and prognosis of COVID-19 infection: Systematic review and critical appraisal. BMJ, 369, Article m1328. https://doi.org/10.1136/bmj.m1328

Yu, M., & et al. (2022). Reinforcement learning for dynamic treatment regimes in cardiovascular care. Frontiers in Cardiovascular Medicine, 9, Article 1012456. https://doi.org/10.3389/fcvm.2022.1012456

Zhang, Y., Chen, G., Xu, Z., & et al. (2024). FedCVD: A federated learning benchmark for cardiovascular disease detection. arXiv. https://doi.org/10.48550/arXiv.2411.07050