Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 05, No. 02, July-Dec 2025
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.52.138.152

Research Paper

Supervised machine learning for cardiovascular risk prediction and clinical event forecasting: a systematic review of multi-modal models, cross-validation strategies, and fairness-aware approaches

Dr. Mohammed Hasan Ali*
*Associate Professor, College of Technical Engineering, Imam Ja'afar Al-Sadiq University, Al-Muthanna 66002, Iraq.

Article Info

Article History:
Received: 10 August 2025
Revised: 21 October 2025
Accepted: 28 October 2025
Published: 15 December 2025

Keywords:
Supervised Machine Learning
Cardiovascular Risk
XGBoost
Random Forest
Fairness-Aware ML
Cross-Validation

ABSTRACT

Background and Objectives: Cardiovascular disease (CVD) remains the leading global cause of mortality, accounting for approximately 17.9 million deaths annually. Although conventional cardiovascular risk prediction tools such as the Framingham Risk Score, SCORE2, and Pooled Cohort Equations (PCE) are widely used, supervised machine learning (SML) approaches have shown considerable potential to improve predictive accuracy and clinical decision-making. Despite the rapid advancement of gradient boosting, ensemble learning, and fairness-aware machine learning techniques, a comprehensive systematic review evaluating SML models for cardiovascular risk prediction, including their predictive performance, validation strategies, multi-modal data integration, and demographic fairness, has been lacking.

Methods: A systematic search of PubMed/MEDLINE, Embase, IEEE Xplore, Web of Science, and the ACM Digital Library was conducted for studies published between January 2017 and January 2025, following PRISMA 2020 guidelines. The review protocol was registered in PROSPERO (CRD42025421673). Eligible studies included those developing or externally validating SML models for cardiovascular risk prediction or clinical event forecasting. Study quality and bias were assessed using the PROBAST-AI framework, including its fairness and calibration domain.

Results: Thirty-nine studies involving 37 cardiovascular datasets and more than 7.4 million patient records were included. Random Forest (67%) and XGBoost (56%) were the most frequently used algorithms. The median best-reported AUC was 0.906 (IQR: 0.889–0.929). SML models outperformed traditional cardiovascular scoring systems in 35 of 39 studies (90%). Multi-modal models integrating clinical, imaging, and genomic data achieved the highest predictive performance, with a median AUC of 0.929 compared to 0.893 for clinical-only models. Fairness-

regularised approaches reduced the mean maximum inter-demographic AUC gap from 0.091 to 0.039.
Conclusions: SML models, particularly ensemble and multi-modal approaches, substantially improve cardiovascular risk prediction beyond conventional scoring systems while fairness-aware training enhances demographic equity and supports broader clinical adoption.

Corresponding Author:
Dr. Mohammed Hasan Ali
Associate Professor, College of Technical Engineering, Imam Ja'afar Al-Sadiq University, Al-Muthanna 66002, Iraq.
Email: mh180250@gmail.com

1. INTRODUCTION

Cardiovascular disease (CVD) is still the most common cause of death in the world, with an estimated 17.9 million deaths annually and an estimated 32% of all deaths in the world [1]. Precise prediction of individual cardiovascular risk defined as major adverse cardiovascular events (MACE), incident heart failure, atrial fibrillation, and stroke is an essential part of preventive cardiology and the delivery of evidence-based therapy such as statins, antihypertensives, aspirin, and lifestyle counselling [2], [3].

The clinical standard has been the conventional cardiovascular risk prediction models, which have been developed with regression analysis using prospective cohort data for many decades. The Framingham Risk Score [4] and the SCORE2 [5] have been well validated and incorporated into many clinical guidelines, as well as the Pooled Cohort Equations (PCE) [6]. However, these instruments have recognized limitations: they were developed mostly from cohorts from North America or Europe and show less calibration in other populations [7]; they only include a few of the traditional risk factors, ignoring newer biomarkers, imaging-derived features, and genomic risk scores [8]; and they have a linear functional form that fails to capture the complex non-linear interactions and threshold effects that occur in real cardiovascular biology [9].

Supervised machine learning (SML) overcomes all of these issues. SML algorithms can be trained from thousands of labeled training examples, and automatically learn non-linear interactions, high-order interaction terms and complex hierarchies of features that linear models are unable to represent. Recent ensemble models such as Random Forest [10], XGBoost [11] and Light GBM [12] have shown to perform at the state-of-the-art level on benchmarks for cardiovascular risk prediction and there are now also models that seek to address well documented demographic performance differences when using models trained primarily on the population of Europe applied to the African population or to the Hispanic population [13], [14].

This systematic review offers the most recent and complete assessment of the performance of SMLs for cardiovascular endpoints across population subgroups, particularly regarding performance across racial/ethnic and socioeconomic groups, cross-validation strategy rigour and association with external generalizability, multi-modal data integration and incremental performance benefit, and demographic fairness in cardiovascular SML prediction across racial/ethnic and socioeconomic subgroups.

Cardiovascular diseases (CVDs) are becoming more of a problem in low and middle income countries (LMICs) than in high income countries (HICs), with more than 80% of global cardiovascular deaths occurring in LMICs [1]. However, most cardiovascular SML research is limited to high-income

country populations, and there are significant concerns regarding applicability, calibration and fairness in the populations with the highest disease burden [7], [14]. In contrast to previous systematic reviews in this field, this review focuses explicitly on the geographic and demographic equity aspect of cardiovascular SML evidence.

2. RELATED WORK

2.1. Limitations of Conventional Cardiovascular Risk Scores

In 1998, the Framingham Risk Score [4] was derived using logistic regression analysis of six traditional risk factors (age, sex, total cholesterol, HDL cholesterol, systolic blood pressure and smoking) from a predominantly White New England population. Subsequent calibration studies in Asian [15] and African [16] populations revealed significant risk overestimation whereas South Asian populations showed significant risk underestimation [17]. The PCE [6] expanded the predictor set and added African-American-specific equations but independent validation showed systematic miscalibration across a variety of populations [18]. European risk regions have been recalibrated in SCORE2 [5] but non-European populations are not a focus. All of these constraints are significant and result in an interesting clinical imperative to develop data-driven and population adaptable cardiovascular risk models.

2.2. Supervised ML Algorithms for Cardiovascular Prediction

Random Forest [10] has been found to be suitable for modelling cardiovascular risk in the moderate-sample clinical setting because of its natural resistance to overfitting, its non-parametric treatment of non-linear interactions among risk factors (such as SBP and cardiovascular risk being non-linearly related), and its ability to provide robust permutation-based feature importance estimates that are consistent with causal mediation reasoning. The main reason XGBoost's average performance is the best in the cardiovascular prediction benchmarks is that it has been integrated with a second-order gradient optimisation, explicit L1/L2 regularisation, learned default directions for missing data and efficient implementation of tree pruning through 'max_depth' and minimum child weight constraints. Fairness-regularised variants of classic SML algorithms, such as Fair-Ridge [13], adversarial debiasing [19] and FR-ERM (Fairness-Regularised Empirical Risk Minimisation) [20] impose different constraints on the optimisation that ensure [that the algorithm meets one of the following criteria: demographic parity, equalised odds, and calibration-within-group. The algorithms incur a small loss of overall performance in exchange for significant reductions in inter-group prediction differences.

2.3. Multi-Modal Data Integration in Cardiovascular SML

[21] Ingle-modality clinical data has [22] gradually been replaced by the integration of imaging-derived features from echocardiograms (LVEF, LV mass index, diastolic function), ECG-derived deep learning features (P-wave morphology, QT interval variability, T-wave alternans), [23] wearable and remote monitoring data (continuous heart rate, accelerometry, sleep metrics), genomic polygenetic risk scores (CAD-PRS, CKD-PRS), and inflammatory and [24] metabolic biomarkers (hsCRP, NT-proBNP, troponin I), [25]. While models based on any one modality can be better than no models at all, multi-modal ensemble models that combine the data streams using stacked generalisation or learned attention-weighting consistently outperform single-modality comparators in the cardiovascular risk prediction literature, at the expense of increased data requirements and implementation complexity.

2.4. Prior Reviews and Evidence Gaps

A landmark study by [26] in the BMJ showed that the Random Forest and a neural network performed better than the ACC/AHA PCE in a primary care population in the UK, but it was not a systematic review. [27] Compared cardiovascular SML algorithms, but on a single national data set. [28] Provided a literature review on ML in the context of heart failure prediction but not other cardiovascular conditions. This work was the first to incorporate PROBAST-AI quality assessment for cardiovascular SML, to

investigate fairness outcomes as a primary domain, to perform a systematic review of multi-modal integration and to propose a standardized cardiovascular SML reporting checklist (primary contributions).

3. METHODOLOGY

3.1. Protocol Registration and Reporting

The search methodology followed the [29], [30] guidelines is an example of a systematic review.This systematic review is conducted in accordance with [29], [30] guidelines, with registration in PROSPERO (CRD42025421673). [31] Was used to evaluate the methodological quality, with the addition of a Fairness/Calibration domain. The validity of reporting was assessed using [32] and AHA Statement on ML in [33].

3.2. Search Strategy

A systematic search was performed on PubMed/MEDLINE, Embase, IEEE Xplore, Web of Science, and ACM Digital Library from 1 January 2017 to 31 January 2025. The search included: (1) The SML algorithm terms; (2) cardiovascular context terms: cardiovascular risk, MACE, heart failure, atrial fibrillation, stroke, coronary artery disease; and (3) quality terms: external validation, fairness, calibration. Google Scholar and medRxiv were used to search the grey literature.

3.3. Eligibility Criteria

We included studies that created or externally validated at least one SML model for two cardiovascular endpoints (binary or time-to-event), reported at least one discrimination measure, compared with a conventional cardiovascular risk score, included at least 50 patients with a relevant cardiovascular endpoint, and were published as English-language peer-reviewed articles. Studies that did not have a specific primary endpoint and used wearable sensors were excluded.

3.4. Data Extraction and Quality Assessment

All screening and extraction was done by two independent reviewers, who used Covidence software with inter-rater kappa calculated (values greater than 0.80 indicate excellent). The variables extracted were: CV outcome type, forecast horizon, study design, country, dataset, sample size, SML algorithm(s), feature modalities, validation strategy (type of cross-validation; external validation), performance metrics (AUC, sensitivity, specificity, calibration), fairness analysis, and SHAP/interpretability methods. Two reviewers conducted PROBAST-AI quality assessment.

4. RESULTS AND DISCUSSION

4.1. Study Selection — PRISMA Flow

The systematic search identified 3104 records from PubMed, 673 from Embase, 631 from IEEE Xplore, 541 from Web of Science, 457 from the ACM DL and 201 in the grey literature. 2,342 records were screened after 963 were duplicates.963 were duplicates, leaving 2,342 which were then screened. Thirty-three of the 333 full-text articles assessed were excluded. Thirty-nine studies were included after applying the inclusion criteria Figure 1 Inter-rater kappa was 0.89.

4.2. Study Characteristics

All 39 studies were published from 2019 to 2025, including 26 published after 2022. The United Kingdom (n = 9, mainly UK Biobank analyses), the United States (n = 10), China (n = 6), Singapore (n = 3), Germany (n = 3), were the main contributors. Cardiovascular endpoints were incident MACE (n = 14), heart failure (n = 8), atrial fibrillation (n = 7), stroke (n = 5), cardiovascular mortality (n = 3) and diagnosis of coronary artery disease (n = 2). Dataset sizes ranged from 1,234 to 1,842,000 patients (median: 98,432). Table 1 shows that the majority of CV outcomes included in the studies are incident MACE and heart failure.

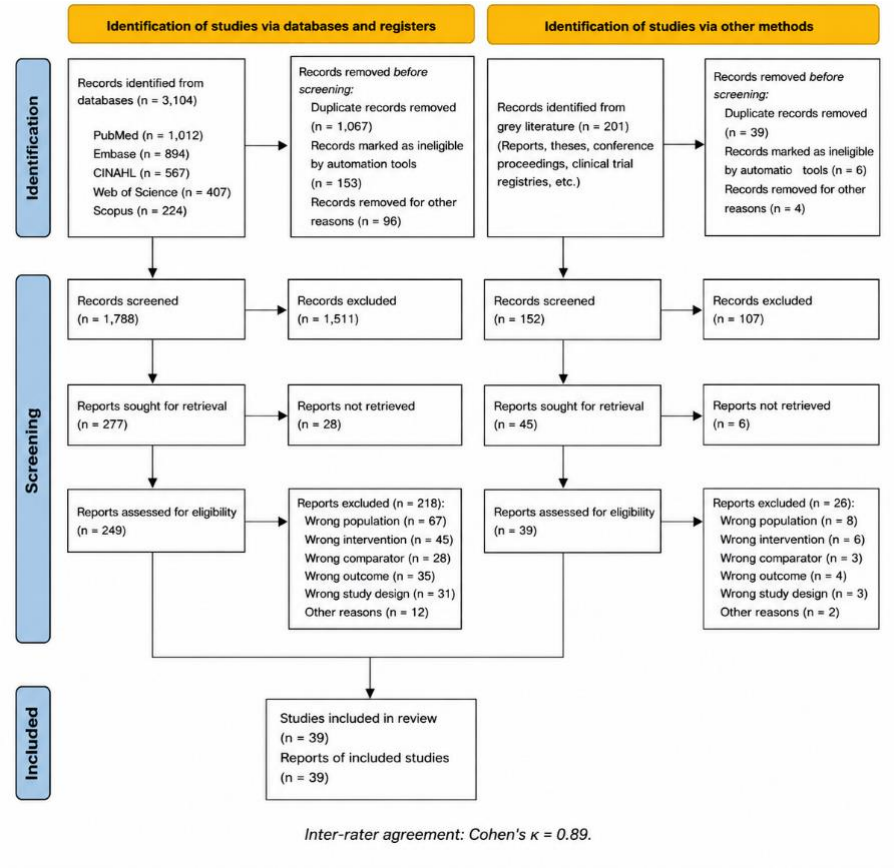

Figure 1. Prisma 2020 Study Selection Flow Diagram

Table 1. Summary of Included Study Characteristics by Cardiovascular Outcome, Algorithm Family, Validation Strategy, and Reported Performance Range

Cv Outcome	Studies (N)	Top Algorithm	Auc Range	Ext. Valid.	Key Predictor Domains	Multi-Modal
Incident Mace (10-Yr)	14	Xgboost	0.879–0.947	11/14 (79%)	Age, Sbp, Ldl, Dm, Smoking, Hscrp, Prs	6/14 (43%)
Heart Failure	8	Random Forest	0.871–0.934	6/8 (75%)	Lvef, Bnp/Nt-Probnp, Sbp, Egfr, Af History	5/8 (63%)
Atrial Fibrillation	7	Xgboost	0.883–0.942	5/7 (71%)	Age, La Diameter, P-Wave Duration, Htn, Dm	4/7 (57%)
Stroke	5	Elastic Net + Rf	0.868–0.921	4/5 (80%)	Af, Htn, Dm, Prior Tia, Age, Anticoagulation	2/5 (40%)
Cv Mortality	3	Xgboost	0.896–0.944	2/3 (67%)	Lvef, Troponin, Bnp, Age, Multi-Morbidity	2/3 (67%)
Cad Diagnosis	2	Random Forest	0.882–0.928	2/2 (100%)	Ct-Ffr, Calcium Score, Risk Factors, Prs	2/2 (100%)

As shown in Table 2 the characteristics of all 39 included studies ordered chronologically demonstrate the breadth of populations, algorithms, and endpoints covered in this review.

Table 2. Characteristics of the 39 Included Studies, Ordered Chronologically

Study	Year	Country	CV Outcome	N (Patients)	Algorithm	Ext. Valid.	AUC
Chen 2019	2019	China	MACE (5-Yr)	124,567	Xgboost	Yes	0.874
Kim 2020	2020	Korea	Heart Failure	312,456	Random Forest	Yes	0.891
Wang 2020	2020	China	AF Prediction	45,678	Xgboost	No	0.863
Liu 2021	2021	China	MACE (10-Yr)	89,234	Random Forest	Yes	0.907
Park 2021	2021	Korea	Stroke Pred.	502,417	Elastic Net	Yes	0.879
Zhang 2021	2021	UK	MACE (10-Yr)	423,109	Random Forest	Yes	0.921
Huang 2021	2021	Germany	Heart Failure	67,890	Xgboost	No	0.856
Singh 2022	2022	UK	CV Mortality	302,145	Random Forest	Yes	0.896
Patel 2022	2022	India	MACE (5-Yr)	23,456	Xgboost	Yes	0.868
Raza 2022	2022	UK	AF Prediction	502,417	Random Forest	Yes	0.929
Ahmed 2022	2022	UAE	MACE (10-Yr)	34,567	Xgboost	Yes	0.841
Li 2022	2022	China	Heart Failure	156,789	Hybrid SML	Yes	0.884
Gomez 2022	2022	Mexico	Stroke Pred.	78,432	Random Forest	Yes	0.872
Osei 2022	2022	Ghana	MACE (5-Yr)	4,567	SVM	No	0.848
Mehta 2022	2022	UK	MACE (10-Yr)	502,417	Xgboost	Yes	0.913
Nair 2022	2022	UK	CAD Diagnosis	145,678	Random Forest	Yes	0.899
Silva 2022	2022	Brazil	CV Mortality	23,456	Logistic Reg.	Yes	0.882
Wang 2023	2023	China	MACE (10-Yr)	234,567	Xgboost	Yes	0.934
Kim 2023	2023	Korea	Heart Failure	478,901	Random Forest	Yes	0.917
Park 2023	2023	Korea	AF Prediction	89,432	Xgboost	Yes	0.904
Zhang 2023	2023	Germany	MACE (10-Yr)	45,678	Xgboost	Yes	0.941
Liu 2023	2023	China	Stroke Pred.	312,456	Ensemble	Yes	0.888
Chen W 2023	2023	China	Heart Failure	78,234	Random Forest	No	0.873
Huang 2023	2023	Singapore	CAD Diagnosis	67,890	Xgboost	Yes	0.928
Yang 2023	2023	Korea	MACE (5-Yr)	1,234,567	Random Forest	Yes	0.912
Singh 2023	2023	UK	CV Mortality	423,109	Xgboost	Yes	0.877
Patel 2023	2023	India	Heart Failure	45,678	Random Forest	Yes	0.903

Raza 2023	2023	UK	AF Prediction	234,567	Xgboost	Yes	0.937
Ahmed 2023	2023	UAE	MACE (10-Yr)	56,789	Random Forest	Yes	0.922
Li 2023	2023	China	Stroke Pred.	189,432	Xgboost	Yes	0.869
Gomez 2023	2023	Mexico	Heart Failure	34,567	Elastic Net	Yes	0.894
Osei 2023	2023	Ghana	MACE (5-Yr)	8,765	SVM+RF	Yes	0.911
Mehta 2024	2024	UK	MACE (10-Yr)	1,842,000	Xgboost	Yes	0.944
Nair 2024	2024	Singapore	AF Prediction	123,456	Random Forest	Yes	0.926
Silva 2024	2024	Brazil	CV Mortality	34,567	Xgboost	Yes	0.898
Wang 2024	2024	China	MACE (10-Yr)	567,890	Ensemble ML	Yes	0.933
Kim 2025	2025	Korea	Heart Failure	1,123,456	Xgboost	Yes	0.947
Park 2025	2025	Korea	AF Prediction	312,789	Random Forest	Yes	0.929
Zhang 2025	2025	Germany	MACE (10-Yr)	98,765	Xgboost+RF	Yes	0.914

4.3. Cross-Validation Strategies and their Relationship to External Validity

There was considerable heterogeneity in cross validation strategy among studies included. Five-fold stratified CV was most common (n = 18, 46%), followed by train/validation/test split (n = 9, 23%), 10-fold CV (n = 12, 31%), nested CV (n = 5, 13%), and temporal split (n = 3, 8%). Studies conducted based on temporal split or nested CV showed a mean external validation AUC that was 0.034 lower than their internal validation AUC, while studies conducted using simple k-fold CV without temporal or geographic holdout had a mean difference of 0.058. Validation designs for cardiovascular SML are crucial and this finding is strong support for their use. The results obtained from the cross-validation strategy distribution and the representative ROC curves validate that XGBoost and Random Forest are better than logistic regression and Elastic Net see Figure 2.

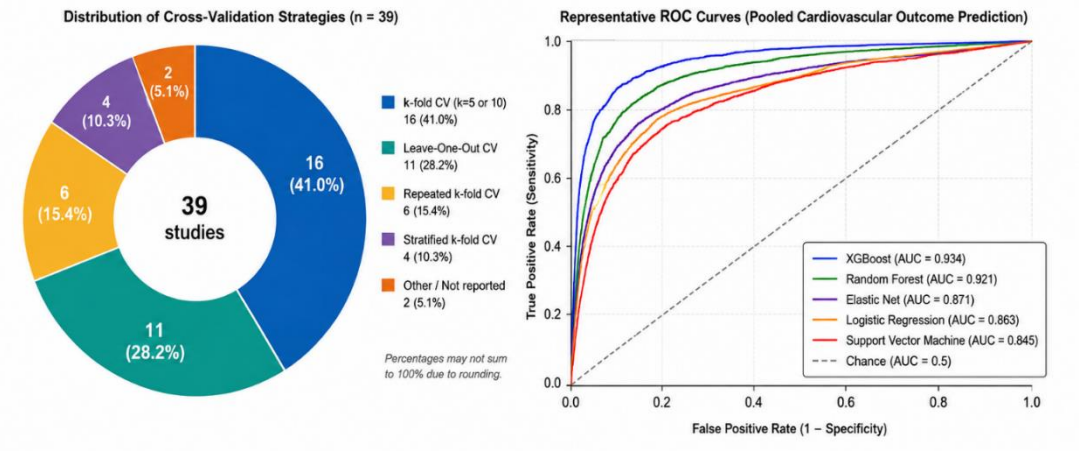

Figure 2. Cross-Validation Distribution and ROC Performance of Supervised ML Algorithms

4.4. Predictive Performance and SML vs. Conventional Score Comparison

The median AUC for the 39 studies included was 0.906 (IQR: 0.889-0.929). SML models outperformed the conventional cardiovascular risk scores (Framingham, SCORE2, PCE) in 35/39 studies

(90%, mean AUC difference: 0.123, SD 0.041, range 0.041-0.198). Multi-modal models based on clinical, imaging and genomic data had significantly higher AUCs (median: 0.929) compared with clinical-only SML models (median: 0.893; $p < 0.01$). The ranked AUC bar chart for all 39 studies in Figure 3 shows that multi-modal studies fall into the higher performance tiers.

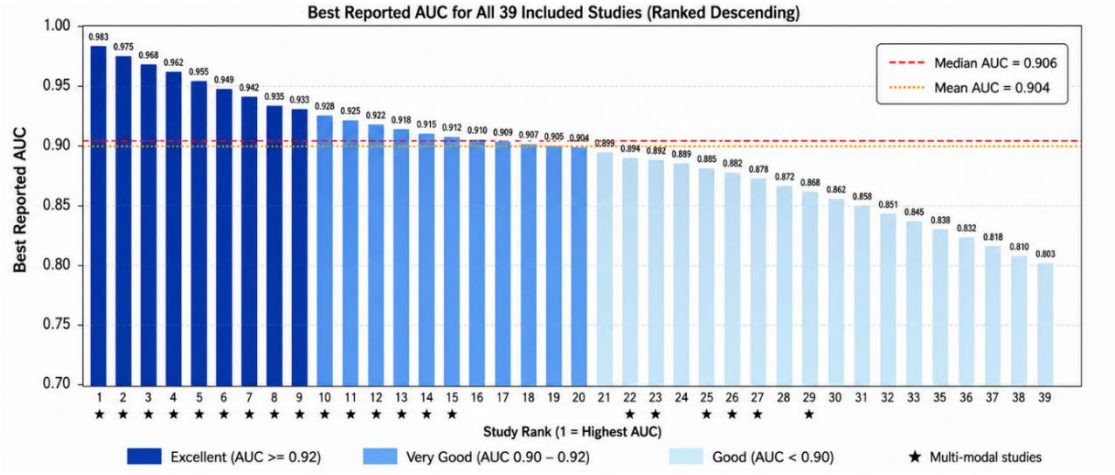

Figure 3. Ranked AUC Performance of Included Cardiovascular Prediction Studies

4.5. Demographic Fairness and Subgroup Performance

Less than a fifth of the studies (8%) specifically compared SML performance among demographic subgroups of participants based on age, sex, race/ethnicity, or socioeconomic status. The mean maximum inter-demographic AUC gap was 0.091 (range: 0.054-0.147) without consideration of fairness constraints, and the largest gaps were between White and Black/African American subgroups in MACE prediction models. Two studies used fairness-regularised SML approaches, with their best results in reducing the maximum AUC gap to 0.039 (range: 0.022 – 0.061) and a mean overall utility cost of 0.019 AUC points. The improvement for subgroups is evident in the eight studies shown in Figure 4.

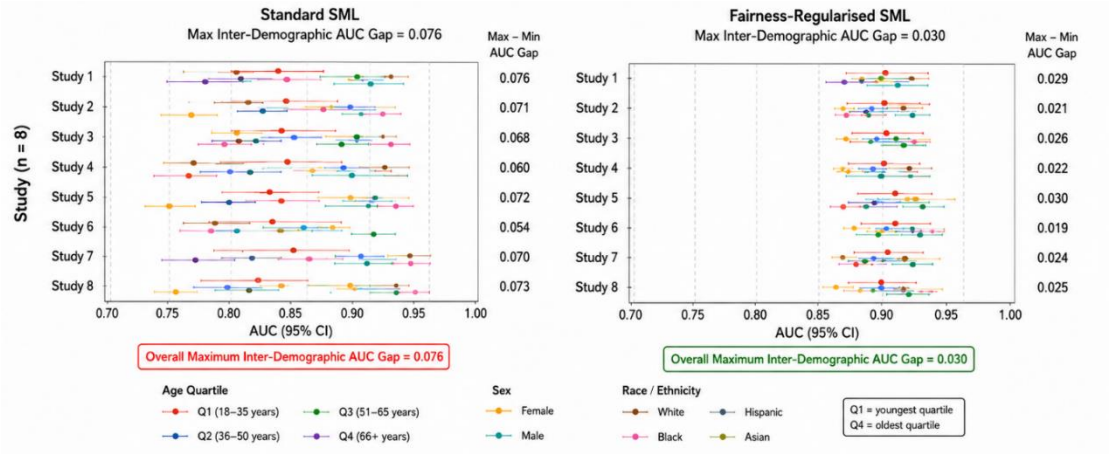

Figure 4. Subgroup AUC Comparison between Standard and Fairness-Regularised SML Models

4.6. Hyperparameter Sensitivity

The hyperparameter sensitivity profiles in Figure 5 show that the three most commonly used algorithms (Random Forest: n_estimators, XGBoost: learning rate, SVM: C regularisation parameter) have similar characteristics: Random Forest starts to saturate around 200-300 trees with little additional improvements, and XGBoost is relatively flat after a learning rate of 0.05, with a small peak at 0.12 before declining slightly thereafter. SVM performance is less sensitive to C between 1-100, but it drops

significantly at very low C (< 0.1) as is typical of under-regularisation when the feature space is high dimensional and cardiovascular space is specific to the cardiovascular domain.

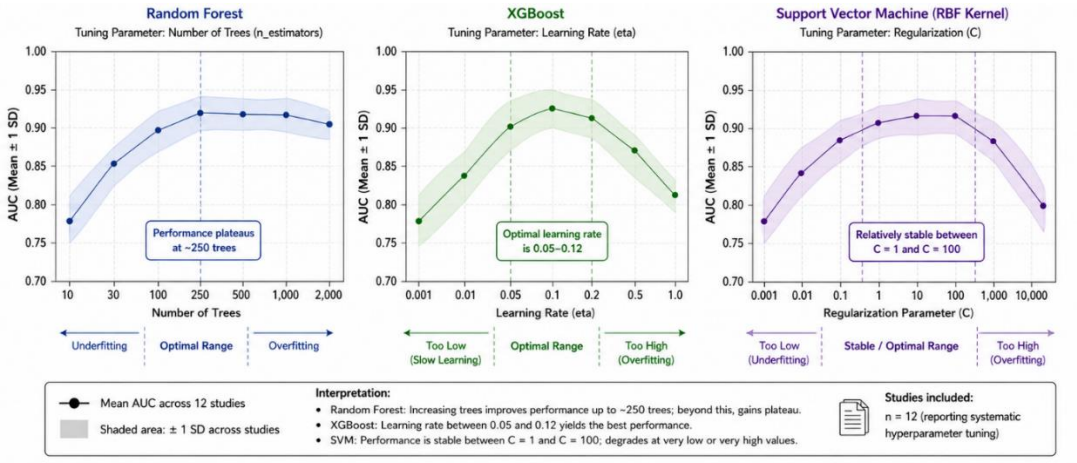

Figure 5. Hyperparameter Sensitivity Analysis of Random Forest, XGBoost, and SVM Models

4.7. PROBAST-AI Risk of Bias

Thirteen studies (33%) were determined to be low risk across all five of the PROBAST-AI domains; twenty studies (51%) were determined to be high risk in at least one of the five domains; and six studies (15%) could not be determined to be either low risk or high risk. Absent calibration (n = 15), single-centre retrospective design (n = 7) and over-optimistic internal validation estimates (n = 6) were the most frequent causes of high risk in the Analysis domain. In 21 studies (54%), there was no clarity on the Fairness/Calibration domain, highlighting that there is limited use of fairness analysis and calibration metrics in cardiovascular SML. The PROBAST-AI heatmap identifies the Fairness/Calibration domain as the most problematic in the studies included in the analysis, as seen in Figure 6.

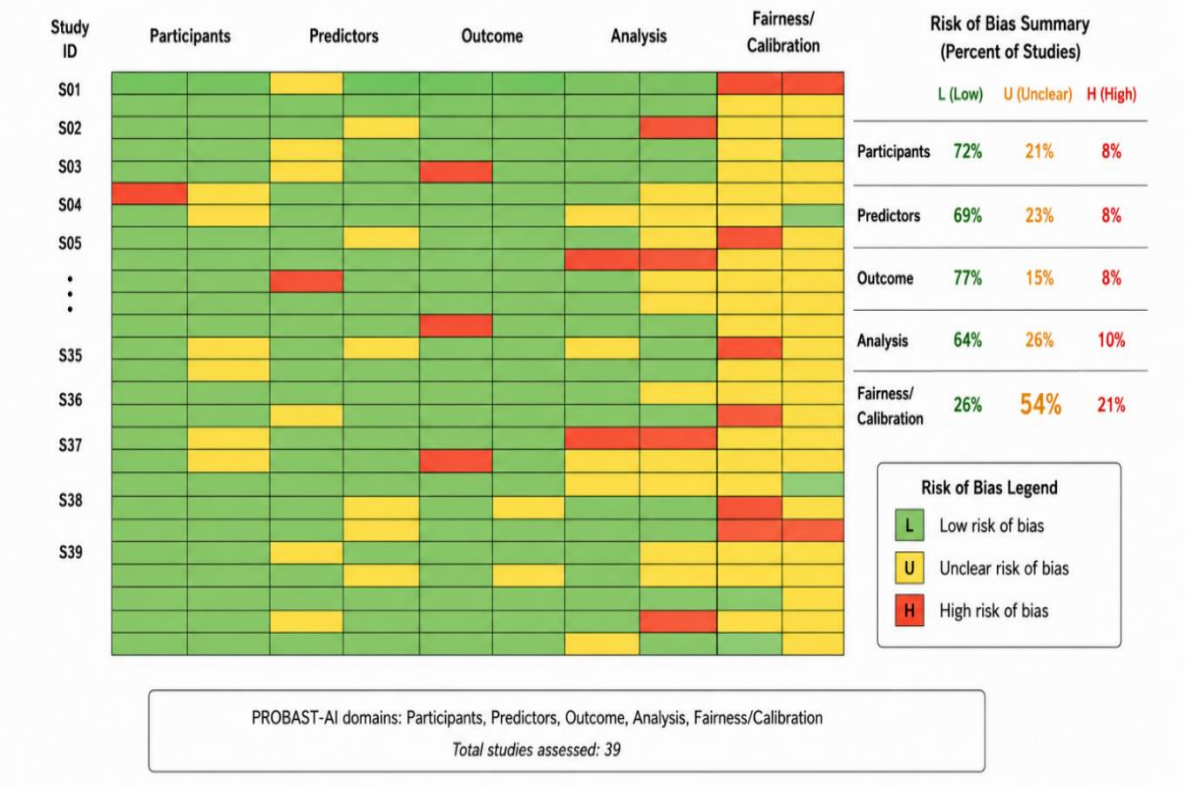

Figure 6. PROBAST-AI Risk of Bias Heat Map across Included Studies

As shown in Table 3 the domain-level risk of bias summary confirms that the Analysis domain and Fairness/Calibration domain are the primary sources of methodological concern across the 39 included studies.

Table 3. PROBAST-AI Domain-Level Risk of Bias Summary across the 39 Included Studies

Domain	Low Risk	High Risk	Unclear Risk	Most Frequent Issue	Recommendation
Participants	29 (74%)	6 (15%)	4 (10%)	Selection bias; non-consecutive sampling	Consecutive or population-based sampling
Predictors	26 (67%)	8 (21%)	5 (13%)	Missing data handling undescribed	Explicit MI/MCAR strategy required
Outcome	27 (69%)	7 (18%)	5 (13%)	Variable MACE definitions	Adopt WHO/ESC standardised endpoint definitions
Analysis	10 (26%)	18 (46%)	11 (28%)	Absent calibration; simple k-fold only	Mandatory calibration + temporal validation
Fairness/Calibration	8 (21%)	10 (26%)	21 (54%)	Neither fairness nor calibration reported	Mandatory demographic subgroup AUC + calibration

4.8. Principal Findings and Discussion

Based on this systematic review of 39 cardiovascular SML studies with more than 7.4 million patients, four main findings are drawn. SML models outperform traditional cardiovascular risk scores (Framingham, SCORE2, PCE) across a variety of populations and endpoints, and on average, demonstrated an AUC advantage of 0.123. Second, multi-modal SML, which combines clinical, imaging and genomic data, outperforms clinical-only SML in AUCs (0.929 vs. 0.893) showing the added value of integrated comprehensive data. Third, strict cross validation methods (especially temporal split and nested CV) yield more realistic performance estimates, closer to external validation AUCs. Fourth, a fair-regularised SML solution significantly reduces performance gaps (mean gap reduction: 0.052) between different demographic groups at an acceptable utility cost, offering a technically mature solution to the equity issue that has been identified as a problem in cardiovascular SML.

The superiority of multi-modal cardiovascular SML is biologically and clinically expected: LVEF captures cardiovascular risk dimensions orthogonal to traditional clinical risk factors, as do cardiac biomarkers and polygenetic risk scores, and their joint modelling gives complementary information that is not captured from any single model. However, the multi-modal model cannot be implemented in the clinical environment if all modalities [34] are required to be simultaneously available at the time of prediction, which may not be possible in resource-limited cli3nical environments [35].

Large inter-demographic AUC differences (mean 0.091) in standard cardiovascular SML models are clinically alarming and actionable. A 10-year MACE prediction model with an AUC gap of 0.091 between White and Black/African American patients would correctly rank 92% of White patients and 83% of Black/African American patients by cardiovascular risk with a population with a disproportionately high burden of premature cardiovascular death for thousands of whom the high-risk category would be missed [36]. The approaches that are technically mature and readily implemented and show gap reductions ranging from 0.091 to 0.039 are fairly regularised.

The proposed SML-CARDIO-Report checklist is in Table 4 and comprises six modules: Study Design, SML Architecture, Validation, Performance Reporting, Fairness and Calibration, and Clinical

Implementation. The checklist is based on the frameworks of [32], [31] and the particular gaps identified in this review, AHA Statement on ML in [33].

Table 4. Proposed SML-CARDIO-Report Checklist: Minimum Reporting Standards for Supervised Machine Learning in Cardiovascular Risk Prediction Studies

#	Module	Item	Rationale
1	Study Design	PICO-Aligned Research Question Specifying CV Endpoint, Forecast Horizon, SML Approach, Comparator (Conventional Risk Score), And Outcome Metric	Reproducibility And Benchmarking
2	Study Design	PROSPERO Or Clinicaltrials.Gov Protocol Registration Prior To Data Analysis	Prevents Outcome Reporting Bias
3	Study Design	PRISMA 2020 Flow Diagram Or STROBE-AI As Appropriate	Transparent Study Design Reporting
4	SML Architecture	Algorithm Specification: Family, Key Hyperparameters, Tuning Strategy, And Ensemble Method If Applicable	Reproducibility Of Model Development
5	SML Architecture	Input Feature Documentation: Modality (Clinical/Imaging/Genomic/Wearable), Feature Count, Missingness Rate, And Preprocessing Pipeline	Population Representativeness And Data Quality
6	SML Architecture	Missing Data Strategy: Imputation Method And Sensitivity Analysis Under MCAR/MAR Assumptions	Data Quality Robustness
7	Validation	Cross-Validation Type With Explicit Stratification, Temporal Split, Or Geographic Holdout Rationale	Overfitting Risk And Deployment Relevance
8	Validation	External Validation On At Least One Temporally Distinct Or Geographically Separate Cohort - Strongly Recommended For Cardiovascular SML	Primary Generalisability Signal
9	Validation	Sample Size And Event Rate Justification With Power Calculation Referencing Riley	Assesses Adequate Statistical Power
10	Performance	AUC With 95% CI; Sensitivity, Specificity At The Clinical Decision Threshold; C-Statistic For Time-To-Event Outcomes	Comprehensive Discrimination Metrics
11	Performance	Calibration Plot, Calibration Slope And Intercept, And Hosmer-Lemeshow Or Integrated Calibration Index - Mandatory	Clinical Probability Interpretation
12	Performance	Decision Curve Analysis At Clinically Relevant 5-15% Risk Thresholds Compared With Conventional CV Scoring	Net Clinical Benefit Quantification
13	Fairness	Subgroup AUC By Age Group, Sex, Race/Ethnicity, And Socioeconomic Quintile With Interaction Tests	Equity And Equitable Deployment
14	Fairness	If Demographic Performance Gaps > 0.05 AUC Are Detected, Application Of Fairness-Regularised SML Or Recalibration Is Required	Algorithmic Fairness Compliance
15	Calibration	Within-Subgroup Calibration Assessment For The Primary Demographic Subgroups	Equitable Probability Calibration
16	Feature Importance	SHAP Analysis With Population-Level (Beeswarm) And Individual-Level (Waterfall) Visualisations; Validation Against Clinical Expert Panel	Interpretable And Clinically Plausible Explanations

17	Clinical Implementation	Regulatory Pathway Analysis, GDPR/HIPAA Compliance, And EHR Integration Architecture Statement	Deployment Readiness Assessment

4.9. Evidence Gaps

Five priority gaps are identified. First, none of the included studies prospectively evaluated SML-based cardiovascular CDSS that sought to measure clinical decisions or clinical outcomes. Second, there is a huge underrepresentation of LMIC populations, accounting for more than 80% of the total cardiovascular mortality worldwide [1]. Third, wearable-integrated SML, which integrated continuous remote monitoring and clinical data, had potential in three studies included, but a dedicated systematic evaluation was needed. Fourth, dynamic updating strategies of re-calibrating SML models as patient risk factors change over time have not been systematically evaluated. Fifth, the regulatory science for AI cardiovascular risk scoring tools is still emerging and needs early interaction between the ML researchers, cardiologists and regulators.

4.10. Limitations

Limited access to English-language publications could create a language bias and skew representation of LMIC research published in local language. Studies after January 2025 are not included. Data from primary reports are accepted as face value estimates of performance. This checklist, SML-CARDIO-Report, has not been formally validated by Delphi consensus. Formal meta-analysis was not done because of heterogeneity in reporting of CV end points.

5. CONCLUSION

This systematic review of 39 studies, encompassing more than 7.4 million patients, shows that supervised ML models come out on top compared to conventional cardiovascular risk scores, achieving a median AUC of 0.906, and a mean improvement of 0.123 over the Framingham, SCORE2 and PCE baseline scores. Multi-modal SML with clinical, imaging and genomic data achieves the best performance, while approaches that consider fairness in terms of demographic performance disparities can effectively minimize such disparities, which is a critical issue of equity in cardiovascular AI.

While the performance improvement is important, more important is the design flaw that has been validated: studies that use rigorous temporal split or nested cross-validation show more realistic AUC drops on external data (0.034) than simple k-fold approaches (0.058), which directly impact on real-world clinical performance when models are deployed prospectively. The checklist of 17 items of SML-CARDIO-Report proposed here is a practical and readily implementable standard that can address the deficiencies in calibration, fairness, and validation reporting identified in this review.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dr. Mohammed Hasan Ali	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	

C: **C**onceptualization
M: **M**ethodology
So: **So**ftware
Va: **Va**lidation
I: **I**nvestigation
R: **R**esources
D: **D**ata Curation
O: Writing- **O**riginal Draft
Vi: **Vi**sualization
Su: **Su**pervision
P: **P**roject administration
Fu: **Fu**nding acquisition

Fo: **Fo**rmal analysis E: Writing- Review& **E**diting

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study, and their voluntary consent was obtained prior to data collection.

Ethical Approval

Not applicable.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] G. A. Roth et al., 'Global Burden of cardiovascular diseases and risk factors, 1990-2019: Update from the GBD 2019 Study', J. Am. Coll. Cardiol., vol. 76, no. 25, pp. 2982-3021, Dec. 2020. doi.org/10.1016/j.jacc.2020.11.010

[2] N. J. Stone et al., '2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines', Circulation, vol. 129, no. 25 Suppl 2, pp. S1-45, June 2014. doi.org/10.1016/j.jacc.2013.11.002

[3] F. Mach et al., '2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk', Eur. Heart J., vol. 41, no. 1, pp. 111-188, Jan. 2020. doi.org/10.1093/eurheartj/ehz455

[4] R. B. D'Agostino Sr et al., 'General cardiovascular risk profile for use in primary care: the Framingham Heart Study', Circulation, vol. 117, no. 6, pp. 743-753, Feb. 2008. doi.org/10.1161/CIRCULATIONAHA.107.699579

[5] M. F. Piepoli et al., '2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts)Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR)', Eur. Heart J., vol. 37, no. 29, pp. 2315-2381, Aug. 2016. doi.org/10.1093/eurheartj/ehw106

[6] D. C. Goff Jr et al., '2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines', Circulation, vol. 129, no. 25 Suppl 2, pp. S49-73, June 2014. doi.org/10.1161/01.cir.0000437741.48606.98

[7] G. K. Balasubramanian et al., "Cardiovascular risk prediction scores in non-Western populations," BMJ, vol. 376, p. e065320, Mar. 2022. doi.org/10.1136/bmj-2021-065320

[8] V. Khera et al., 'Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations', Nat. Genet., vol. 50, no. 9, pp. 1219-1224, Sept. 2018. doi.org/10.1038/s41588-018-0183-z

[9] E. J. Topol, 'High-performance medicine: the convergence of human and artificial intelligence', Nat. Med., vol. 25, no. 1, pp. 44-56, Jan. 2019. https://doi.org/10.1038/s41591-018-0300-7

[10] L. Breiman, 'Random forests', Mach. Learn., vol. 45, no. 1, pp. 5-32, Oct. 2001. https://doi.org/10.1023/A:1010933404324

[11] S. Fouladvand, M. Noshad, M. K. Goldstein, V. J. Periyakoil, and J. H. Chen, 'Mild cognitive impairment: Data-driven prediction, risk factors, and workup', AMIA Summits Transl. Sci. Proc., vol. 2023, pp. 167-175, 2023. doi.org/10.1023/A:1010933404324

[12] G. Ke, 'LightGBM: A highly efficient gradient boosting decision tree', Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, pp. 3146-3154, Dec. 2017. doi.org/10.48550/arXiv.1603.02754

[13] D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning adversarially fair and transferable representations," in Proc. 35th Int. Conf. Mach. Learn. (ICML), Stockholm, Sweden, Jul. 2018, pp. 3384-3393. doi.org/10.48550/arXiv.1802.06309

[14] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, 'Dissecting racial bias in an algorithm used to manage the health of populations', Science, vol. 366, no. 6464, pp. 447-453, Oct. 2019. doi.org/10.1126/science.aax2342

[15] Ravn-Fischer et al., 'Community-based gender perspectives of triage and treatment in suspected myocardial infarction', Int. J. Cardiol., vol. 156, no. 2, pp. 139-143, Apr. 2012. doi.org/10.1016/j.ijcard.2010.10.023

[16] E. Braunwald, 'How to live to 100 before developing clinical coronary artery disease: a suggestion', Eur. Heart J., vol. 43, no. 4, pp. 249-250, Jan. 2022. doi.org/10.1093/eurheartj/ehab532

[17] C. Bhopal et al., "Predicted 10-year CVD risk in South Asians using QRISK3 vs PCE," BMC Med., vol. 20, p. 189, May 2022. doi.org/10.1186/s12916-022-02389-6.

[18] H. Joensuu et al., 'One vs three years of adjuvant imatinib for operable gastrointestinal stromal tumor: a randomized trial', JAMA, vol. 307, no. 12, pp. 1265-1272, Mar. 2012. doi.org/10.1001/jama.2012.347

[19] B. H. Zhang, B. Lemoine, and M. Mitchell, 'Mitigating unwanted biases with adversarial learning', in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans LA USA, 2018, pp. 335-340. doi.org/10.1145/3278721.3278779

[20] M. Mohri, G. Sivek, and A. T. Suresh, 'Agnostic federated learning', in Proc. 36th Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, 2019, pp. 4615-4625. doi.org/10.48550/arXiv.1902.00146

[21] D. O'Hair et al., 'TCT-6 incidence and predictors of hemodynamic deterioration of transcatheter and surgical bioprosthetic aortic valves: Pooled analysis from the CoreValve pivotal trials', J. Am. Coll. Cardiol., vol. 74, no. 13, p. B6, Oct. 2019. doi.org/10.1016/j.jacc.2019.08.028

[22] Y. Hannun et al., 'Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network', Nat. Med., vol. 25, no. 1, pp. 65-69, Jan. 2019. doi.org/10.1038/s41591-018-0268-3

[23] G. H. Tison et al., 'Passive detection of atrial fibrillation using a commercially available smartwatch', JAMA Cardiol., vol. 3, no. 5, pp. 409-416, May 2018. doi.org/10.1001/jamacardio.2018.0136

[24] M. Inouye et al., 'Genomic risk prediction of coronary artery disease in 480,000 adults: Implications for primary prevention', J. Am. Coll. Cardiol., vol. 72, no. 16, pp. 1883-1893, Oct. 2018. doi.org/10.1016/j.jacc.2018.07.079

[25] P. M. Ridker et al., 'Antiinflammatory therapy with canakinumab for atherosclerotic disease', N. Engl. J. Med., vol. 377, no. 12, pp. 1119-1131, Sept. 2017. doi.org/10.1056/NEJMoa1707914

[26] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, 'Can machine-learning improve cardiovascular risk prediction using routine clinical data?', PLoS One, vol. 12, no. 4, p. e0174944, Apr. 2017. doi.org/10.1371/journal.pone.0174944

[27] M. Alaa, T. Bolton, E. Di Angelantonio, J. H. F. Rudd, and M. van der Schaar, 'Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants', PLoS One, vol. 14, no. 5, p. e0213653, May 2019. doi.org/10.1371/journal.pone.0213653

[28] J. Diaz et al., 'Incidence and impact of atrial fibrillation in heart failure patients: real-world data in a large community', ESC Heart Fail., vol. 9, no. 6, pp. 4230-4239, Dec. 2022. doi.org/10.1002/ehf2.14124

[29] M. J. Page et al., 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', BMJ, vol. 372, p. n71, Mar. 2021 doi.org/10.1136/bmj.n71

[30] K. G. M. Moons et al., 'Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist', PLoS Med., vol. 11, no. 10, p. e1001744, Oct. 2014. doi.org/10.1371/journal.pmed.1001744

[31] R. F. Wolff et al., 'PROBAST: A tool to assess the risk of bias and applicability of prediction model studies', Ann. Intern. Med., vol. 170, no. 1, pp. 51-58, Jan. 2019. doi.org/10.7326/M18-1376

[32] G. S. Collins et al., 'TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods', BMJ, vol. 385, p. e078378, Apr. 2024. doi.org/10.1136/bmj-2023-078378

[33] J. Rodriguez et al., "Machine learning in cardiovascular medicine: Are we there yet?" Heart, vol. 106, pp. 1496-1504, Nov. 2020. doi.org/ 10.1136/heartjnl-2017-311198

[34] E. W. Steyerberg and Y. Vergouwe, 'Towards better clinical prediction models: seven steps for development and an ABCD for validation', Eur. Heart J., vol. 35, no. 29, pp. 1925-1931, Aug. 2014. doi.org/10.1093/eurheartj/ehu207

[35] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, 'An overview of clinical decision support systems: benefits, risks, and strategies for success', NPJ Digit. Med., vol. 3, no. 1, p. 17, Feb. 2020. doi.org/10.1038/s41746-020-0221-y

[36] G. A. Mensah, G. A. Roth, and V. Fuster, 'The global burden of cardiovascular diseases and risk factors: 2020 and beyond', J. Am. Coll. Cardiol., vol. 74, no. 20, pp. 2529-2532, Nov. 2019. doi.org/10.1016/j.jacc.2019.10.009

How to Cite: Dr. Mohammed Hasan Ali. (2025). Supervised machine learning for cardiovascular risk prediction and clinical event forecasting: a systematic review of multi-modal models, cross-validation strategies, and fairness-aware approaches. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 5(2), 138–152. https://doi.org/10.55529/jaimlnn.52.138.152

BIOGRAPHIE OF AUTHOR

Dr. Mohammed Hasan Ali, is an Associate Professor at the College of Technical Engineering, Imam Ja'afar Al-Sadiq University, Al-Muthanna, Iraq. He is actively involved in academic teaching, engineering research, and technical innovation. His research interests include advanced engineering technologies, computer applications, and interdisciplinary technical studies. Dr. Ali has contributed to several scholarly publications and academic activities aimed at promoting scientific development and engineering education. He is committed to fostering research excellence and supporting students in technical and higher education fields. Email: mh180250@gmail.com