

Effectiveness of AI Models for Cardiovascular Disease Diagnosis: A Systematic Review and Meta-analysis

Efectividad de los modelos de inteligencia artificial para el diagnóstico de enfermedades cardiovasculares: una revisión sistemática y metaanálisis

Carlos Zepeda-Lugo¹  , Marcos Sanchez-Lizarraga²  , Andrea Insfran-Rivarola³  ,
 Ana Pamela Arévalos³  

¹ Faculty of Engineering, Architecture and Design, Universidad Autónoma de Baja California, Ensenada, Baja California, Mexico

² Academic Unit of Business, Universidad Autónoma de Sinaloa, Los Mochis, Sinaloa, Mexico

³ Department of Industrial Engineering, Universidad Nacional de Asunción, San Lorenzo, Paraguay



Copyright

© 2025 María Cano University Foundation. The *Revista de Investigación e Innovación en Ciencias de la Salud* provides open access to all its content under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International \(CC BY-NC-ND 4.0\)](#) license.

Correspondence

Carlos Zepeda-Lugo.
 Email: czepeda@uabc.edu.mx

Editors

Fraidy-Alonso Alzate-Pamplona, MSc. 
 Efrén Murillo-Zamora, Ph.D. 

Declaration of interests

The authors declare no conflicts of interest.

Funding

This research received no external funding.

Ethics statement

Ethical approval was not required for this study as it did not involve human or animal subjects.

Abstract

Introduction. Cardiovascular disease represents a substantial global health burden, and delays in diagnosis can further compromise clinical outcomes.

Objective. This review aimed to evaluate the diagnostic effectiveness of artificial intelligence models across major cardiovascular conditions, with a specific focus on coronary artery disease, acute coronary syndromes and myocardial infarction, and heart failure.

Method. Following PRISMA guidelines, we conducted a systematic review and applied meta-analytical techniques to pool measures of diagnostic performance. Random effects models with restricted maximum likelihood and Hartung–Knapp adjustments were used to calculate summary estimates, and analyses were stratified by cardiovascular condition. Only effects with variance (reported 95% CI or derivable denominators) were pooled. Areas under the receiver operating characteristic curve without variance were synthesized narratively.

Results. Out of 39,383 records, thirty-five studies were eligible. The pooled area under the receiver operating characteristic curve was 0.823 (95% confidence interval 0.754–0.892). Heterogeneity was considerable ($I^2 = 98.4\%$), and the 95% prediction interval was 0.59 to 1. However, 9 of 35 included studies reported external/independent validation. The pooled estimate therefore represents an overall descriptive average, reflecting the expected range of diagnostic performance rather than a single universal value. Stratified results showed pooled values of 0.88 for acute coronary

Data availability

All data supporting the findings of this study are available within the article. For additional details, please contact the corresponding author.

Author Contributions

Carlos Zepeda-Lugo:

conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft.

Marcos Sanchez-Lizarraga: data curation, formal analysis, investigation, methodology, validation, writing – review & editing.

Andrea Insfran-Rivarola: formal analysis, investigation, methodology, validation, visualization, writing – review & editing.

Ana Pamela Arévalos: investigation, methodology, resources, validation, writing – review & editing.

Generative AI declaration

The authors declare that generative AI tools (ChatGPT, OpenAI) were used solely for language editing to improve grammar, clarity, and consistency in English. The tool did not generate or modify any scientific content, analysis, or interpretation. All AI-assisted outputs were independently reviewed, verified, and fully approved by the authors, who take full responsibility for the integrity and accuracy of the final manuscript.

Cite this article

Zepeda-Lugo C, Sanchez-Lizarraga M, Insfran-Rivarola A, Arévalos AP. Effectiveness of AI models for cardiovascular disease diagnosis: A systematic review and meta-analysis. *Revista de Investigación e Innovación en Ciencias de la Salud*. 2026;8(1):1-30. e-v8n1a504. <https://doi.org/10.46634/riics.504>

Acknowledgments

The authors thank the Universidad Autónoma de Baja California and the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) for their institutional support in fostering scientific research in Mexico.

Received: 08/30/2025

Revised: 10/21/2025

Accepted: 11/28/2025

Published: 12/09/2025

Disclaimer

The content of this article is the sole responsibility of the authors and does not necessarily represent the official views of their affiliated institutions, the funding agency, or the *Revista de Investigación e Innovación en Ciencias de la Salud*.

syndromes and myocardial infarction, 0.87 for heart failure, and 0.86 for coronary artery disease. Sensitivity often reached 0.90 or higher in electrocardiography deep learning models, while specificity remained variable.

Conclusion. Artificial intelligence models demonstrated promising discrimination across cardiovascular conditions, with variable specificity and predictive value, indicating that clinical utility requires external validation and local calibration.

Keywords

Cardiovascular disease; artificial intelligence; diagnostic models; systematic review.

Resumen

Introducción. Las enfermedades cardiovasculares representan una carga sustancial a nivel mundial, y los retrasos en su diagnóstico pueden comprometer de forma directa los resultados clínicos.

Objetivo. Esta revisión tuvo como objetivo evaluar la efectividad diagnóstica de los modelos de inteligencia artificial aplicados a las principales enfermedades cardiovasculares, con énfasis en la enfermedad arterial coronaria, los síndromes coronarios agudos e infarto de miocardio, y la insuficiencia cardíaca.

Método. Siguiendo las directrices PRISMA, se realizó una revisión sistemática y se aplicaron técnicas de metaanálisis para integrar las medidas de desempeño diagnóstico. Se emplearon modelos de efectos aleatorios con estimación de máxima verosimilitud restringida y ajustes de Hartung–Knapp para obtener estimaciones globales, estratificando los análisis según la condición cardiovascular. Solo se combinaron los efectos que contaban con varianza, es decir, aquellos que reportaban o permitían calcular intervalos de confianza del 95 %. Los valores del área bajo la curva sin varianza se analizaron de forma narrativa.

Resultados. De un total de 39,383 registros, treinta y cinco estudios cumplieron con los criterios de inclusión. El área bajo la curva resumen fue de 0.823 (IC 95%: 0.754–0.892). La heterogeneidad fue considerable ($I^2 = 98.4\%$) y el intervalo de predicción del 95% osciló entre 0.59 y 1. Sin embargo, solo 9 de los 35 estudios incluidos reportaron validación externa o independiente. Por lo tanto, la estimación combinada debe interpretarse como un promedio descriptivo general, que refleja el rango esperado de desempeño diagnóstico, más que un valor universal. Los resultados estratificados mostraron valores agrupados de 0.88 para síndromes coronarios agudos e infarto de miocardio, 0.87 para insuficiencia cardíaca y 0.86 para enfermedad arterial coronaria. En los modelos de aprendizaje profundo basados en electrocardiografía, la sensibilidad alcanzó frecuentemente valores de 0.90 o superiores, mientras que la especificidad mostró mayor variabilidad.

Conclusión. Los modelos de inteligencia artificial mostraron una capacidad discriminativa prometedora en las distintas enfermedades cardiovasculares, aunque con especificidad y valor predictivo variables. Estos hallazgos sugieren que su aplicación clínica requiere validación externa y calibración local antes de su integración en la práctica médica.

Palabras clave

Enfermedad cardiovascular; inteligencia artificial; modelos diagnósticos; revisión sistemática.

Introduction

Cardiovascular diseases (CVD) remain the principal driver of global morbidity and mortality and impose a disproportionate public health burden across regions [1-3]. In light of current estimates from the World Health Organization and the Global Burden of Disease collaboration, millions of deaths each year are attributed to these conditions, which underscores the magnitude of the problem and accounts for 32% of global deaths [4,5]. Several modifiable cardiovascular risk factors directly addressable in clinical practice include smoking, physical inactivity, dyslipidemia, obesity, diabetes mellitus, and hypertension [6-9]. Epidemiologic assessments indicate that approximately 85% of the population falls within a low cardiovascular risk category and 14% within a moderate risk category, with overweight representing the most frequent factor (51%), followed by grade 1 obesity (21%) and hypertension (5%) [10,11]. With regard to the Americas, CVD persist as a dominant source of premature mortality and disability adjusted life years, reflecting sustained population impact [12,13]. Diagnostic delays or missed recognition of acute coronary syndromes (ACS), including myocardial infarction (MI), and heart failure are linked to cardiogenic shock [14], subsequent progression to advanced (stage D) heart failure, and avoidable mortality; consequently, timely and accurate diagnostic pathways remain imperative [15-17]. These disorders share a vascular and hemodynamic basis: atherosclerotic narrowing in coronary artery disease, plaque rupture with thrombosis in ACS and MI, and impaired ventricular filling or ejection in heart failure. Their overlapping pathways explain why early, accurate detection determines clinical outcomes and guides timely reperfusion and therapy.

Standard workups for coronary artery disease, ACS, and MI typically encompass electrocardiography [18,19], cardiac biomarkers such as troponins [20,21], risk stratification instruments, such as TIMI (Thrombolysis in Myocardial Infarction) and GRACE (Global Registry of Acute Coronary Events) scores [22,23], computed tomography angiography [24,25], and invasive coronary angiography [26,27]. For heart failure and the estimation of left ventricular filling pressures, routine approaches include echocardiography [28], B-type natriuretic peptide and N-terminal pro-B-type natriuretic peptide (NT-proBNP) [29], chest radiography [30,31], and cardiac magnetic resonance imaging (MRI) [32]. Notwithstanding their clinical utility, these modalities present well documented constraints operator dependent variability (notably in echocardiography), barriers to access, difficulties with atypical presentations, and delays to result availability that collectively jeopardize time-sensitive decisions [4,33,34].

Artificial intelligence (AI), including Machine Learning (ML) and Deep Learning (DL), denotes a family of computational methods with substantial potential to strengthen diagnostic performance in cardiology [35-37]. It is increasingly explored to address the limitations of conventional diagnostic pathways by integrating imaging, biomarkers, and clinical data to uncover complex associations that may not be evident through traditional analysis, thereby supporting earlier recognition and more consistent decision making in cardiovascular care. AI refers to systems capable of executing tasks typically requiring human cognition [38]; ML specifies algorithms that infer patterns from data without explicit task specific programming [39]; DL designates multilayer neural architectures that capture hierarchical representations and complex data structure [40]. Within cardiovascular practice, these techniques are deployed for supervised classification (e.g., detection of hypertrophic cardiomyopathy, atrial fibrillation) [41,42], anatomical segmentation in medical imaging [43,44], analysis of time-series signals

such as the electrocardiogram (ECG) [45], and multimodal data fusion [46,47]. From a clinical perspective, model performance is judged not only by discrimination but also by calibration and external validation, ensuring that probability estimates correspond to real world risk and maintain reliability across distinct clinical environments.

Reported results include high discriminative performance, often summarized by the area under the receiver operating characteristic curve (AUROC) for MI and left ventricular dysfunction [12]. Nevertheless, persistent methodologic limitations—including overfitting, data leakage, class imbalance, miscalibration, and the scarcity of rigorous external validation—continue to limit transportability and trustworthiness in clinical settings [4,48]. Previous systematic reviews provide important summaries of the field; notwithstanding, indication-stratified quantitative synthesis remains uncommon [49,50]. Several primary studies also omit essential statistics such as confidence intervals (CI) for reported AUROC estimates, which in turn, constrains meta-analytic aggregation [12].

This work is justified by the need for a robust evidence synthesis that not only quantifies diagnostic accuracy but also accounts for methodological gaps in primary studies. A rigorous meta-analysis can provide more reliable pooled estimates to inform both clinical practice and future model development. Therefore, given the limited number of quantitative syntheses, we conducted a systematic review and meta-analysis to evaluate the diagnostic performance of AI models for CVD. The primary objective was to estimate pooled diagnostic accuracy, reported as AUROC with 95% confidence and prediction intervals (PI), stratified by clinical indication. The remainder of the manuscript is structured as follows: Section 2 details the methods used in this systematic review; Section 3 presents the results; Section 4 provides an extended discussion; and Sections 5 and 6 outline study limitations and the conclusions, respectively.

Method

The protocol for our systematic review was prepared and registered on the International Prospective Register of Systematic Reviews. The systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [51,52]. The completed PRISMA 2020 checklist is provided in the [Supplementary Table ST1](#). The following subsections summarize the methodology.

Information sources and search strategy

For the search, we used the following databases: MEDLINE (Ovid), PubMed (National Library of Medicine), Embase (Elsevier), Scopus (Elsevier), Web of Science Core Collection/ Science Citation Index (Clarivate), and IEEE Xplore (IEEE). Additional sources included backward and forward citation tracking, reference-list screening of all included studies, and preprints from arXiv and medRxiv to identify forthcoming eligible studies. Moreover, only peer-reviewed versions were included in analyses. We also hand-searched conference proceedings from IEEE and the American Heart Association (AHA). Authors were contacted to clarify study eligibility or obtain missing information.

As part of the electronic search strategy, we combined controlled vocabulary and free-text terms, adapting syntax and subject headings to each source. The keywords used included “cardiovascular diseases”, “myocardial infarction”, “heart failure”, “coronary artery disease”, “artificial intelligence”, “machine learning”, “deep learning”, “neural network”, and “algorithm”. The final database search was conducted in August 2025. The detailed search strategies for each database are provided in [Supplementary Table ST2](#).

Study selection

We prespecified eligibility through a diagnostic patient, intervention, comparator, outcome (PICO) framework adapted to CVD [53]. Participants were adult patients (≥ 18 years) undergoing diagnostic evaluation for coronary artery disease (CAD), acute coronary syndromes/myocardial infarction (ACS/MI), or heart failure (HF). The setting included hospitals, emergency departments, outpatient clinics, or specialized cardiology centers. Index tests were AI diagnostic models such as machine-learning or deep-learning approaches, applied to human clinical data, including imaging, electrocardiography (ECG), electronic health records (EHR), and biomarkers. Comparators were accepted clinical reference standards: clinical assessment, ECG, echocardiography, coronary angiography, cardiac MRI, biomarker testing, or conventional diagnostic risk models. We included diagnostic accuracy designs. We excluded animal or pre-clinical studies, pediatric populations, traditional statistical models without AI, non-diagnostic AI applications (prognosis only), studies lacking diagnostic accuracy outcomes, reviews, editorials/opinions, and conference abstracts with insufficient data. We collected studies published between 2014 and 2025 to reflect the modern AI, restricting the search to English and Spanish. To preserve interpretability, we extracted effects against each study's stated reference standard and documented any deviations (e.g., expert consensus) as applicability concerns within PROBAST+AI. The review focused on diagnostic accuracy at the index evaluation; therefore, prognostic risk scores such as the Thrombolysis in Myocardial Infarction (TIMI) score, the Global Registry of Acute Coronary Events (GRACE) score, and the Framingham risk score were not considered diagnostic reference standards. When studies directly compared AI with a clinical risk score for the same diagnostic outcome and provided variance data, these results were described narratively. Heterogeneity in outcome definitions, thresholds, and intended use prevented a comparative meta-analysis.

Data collection process

Two reviewers independently extracted data using a piloted, standardized form. The rate of disagreements was 12% and was resolved through discussion. Then, two reviewers retrieved and assessed the full texts of the relevant studies based on the inclusion and exclusion criteria. When consensus was difficult to reach (close to 4%), a third review author assessed the study. The raw data were extracted by one reviewer and examined by a second reviewer; these data included study identifiers, methodological characteristics, and model development and performance features. Moreover, we contacted corresponding authors to obtain missing accuracy statistics (e.g., 95% CI), thresholds, validation details, and dataset sources. No individual participant data were sought; analyses used aggregate study-level data.

External or independent validation was prespecified as the preferred effect for synthesis when available; when both internal and external results were reported, external estimates were extracted preferentially. Because few external validations provided complete 95% CI, external-only performance was summarized narratively rather than pooled. The full study extraction matrix is provided in the [Supplementary Table ST3](#).

Accordingly, primary outcomes were area under the receiver operating characteristic curve (AUROC) and paired sensitivity and specificity for diagnosing CAD, MI, or HF at the index encounter. Secondary outcomes included positive predictive value (PPV), negative predictive value (NPV), accuracy, F1 score, and balanced accuracy. We sought all compatible results per outcome domain and analysis. For syntheses, one effect per study per domain was retained according to predefined rules (external validation preferred; otherwise, the prespecified primary

analysis); alternative effects informed sensitivity analyses. No changes to outcome domains or their priority occurred. Subgroups were prespecified as three non-overlapping diagnostic indications: CAD (stable/obstructive disease), ACS/MI (acute ischemic presentations, including STEMI/NSTEMI), and HF (reduced or preserved EF). When a study reported more than one indication, each effect was mapped to its corresponding subgroup and entered once per domain; no effect was counted in more than one pool. Pooling and 95% PI were computed within indication against the indication-specific reference standard.

Risk of bias assessment and effect measures

To assess the quality, risk of bias (RoB), and applicability of CVD prediction models, the Prediction model Risk of Bias Assessment Tool (PROBAST)+AI was employed [54,55]. PROBAST+AI includes four domains, 34 signaling questions (16 for model development, 18 for model evaluation), and six applicability items (three each for model development and for model evaluation) [55]. We judged certainty narratively for each outcome and subgroup by integrating RoB, consistency, precision, and applicability. Quality concerns related to model development and evaluation were judged as low, high, or unclear [55]. Two reviewers independently applied PROBAST+AI following BMJ 2025 guidance. Overall RoB was rated low only when all domains were low, high when any domain was high, and unclear when at least one domain was unclear and none were high. The same criteria were applied to overall applicability. The detailed evaluations of RoB at the item level, together with the applicability ratings (A1–A3) are presented in the [Supplementary Table ST4](#). As an internal consistency check, pooled estimates were re-examined after excluding studies rated as high overall RoB in PROBAST+AI.

For meta-analysis, we included only AUROC estimates with reported variance (95% CI or sufficient data to calculate it) [56]. Studies without variance information were summarized narratively and excluded from pooled weighting. No imputation of missing CI or standard errors was undertaken, and effects lacking variance were handled qualitatively only. For sensitivity and specificity, we used estimates at the study-defined clinical threshold [57]. These measures directly addressed comparative diagnostic performance.

We pooled AUROC values using random effects inverse variance models [58], estimating between study variance primarily with restricted maximum likelihood (REML). Where appropriate, Hartung–Knapp–Sidik–Jonkman adjustments were applied to CI [59]. Heterogeneity was quantified by means of τ^2 , I^2 , Cochran's Q , and the 95% prediction interval (PI) [60,61]. We calculated 95% PI by combining the REML between-study variance with the average within-study variance and applying Hartung–Knapp t-quantiles with degrees of freedom determined by the number of included studies. AUROC bounds were constrained to 0–1. Pooling was conducted within each clinical indication (CAD, ACS/MI, and HF) relative to its reference standard, with cross-standard differences treated as a source of heterogeneity reflected in τ^2 , I^2 , and the 95% PI. Only 12 effects fulfilled the pooling criteria. The substantial variability in study design and reference standards precluded a meaningful meta-regression. Heterogeneity was assessed through stratification by clinical condition and by the 95% PI. No additional pooled subgroup analyses based on sample size, geographic region, or algorithmic approach were performed, given the limited number of effects with variance ($n = 12$) and the presence of overlapping moderators across studies. Study setting was evaluated as a potential effect modifier; however, the small number of effects with variance and the associated risk of sparse data bias rendered stratification by setting impractical. Emphasis was therefore placed on the 95% PI. When a study included multiple clinical conditions, each effect was assigned to a single condition and entered once per analytic domain, ensuring that no effect contributed to more than one pooled estimate.

Results

As shown in Figure 1, we identified a total of 39,383 records. Before the screening stage, 17,384 records were excluded. Among these, 13,784 were duplicates, 2,500 were automatically identified as ineligible through algorithmic filters, and 1,100 were removed according to prespecified criteria related to document type, language, or publication year. After these exclusions, 21,999 records remained for title and abstract screening. From this stage, 412 full-text articles were selected for detailed evaluation. Twelve of these could not be retrieved despite repeated attempts. The remaining 400 full-text articles were assessed in detail, and 365 were subsequently excluded. The most frequent reasons for exclusion were studies addressing non-diagnostic or prognosis-only artificial intelligence applications (152 studies), ineligible populations or conditions (117 studies), and absence of diagnostic accuracy outcomes or use of inappropriate reference standards (96 studies). In total, 35 studies met all inclusion criteria and were included in the final analysis [62-96].

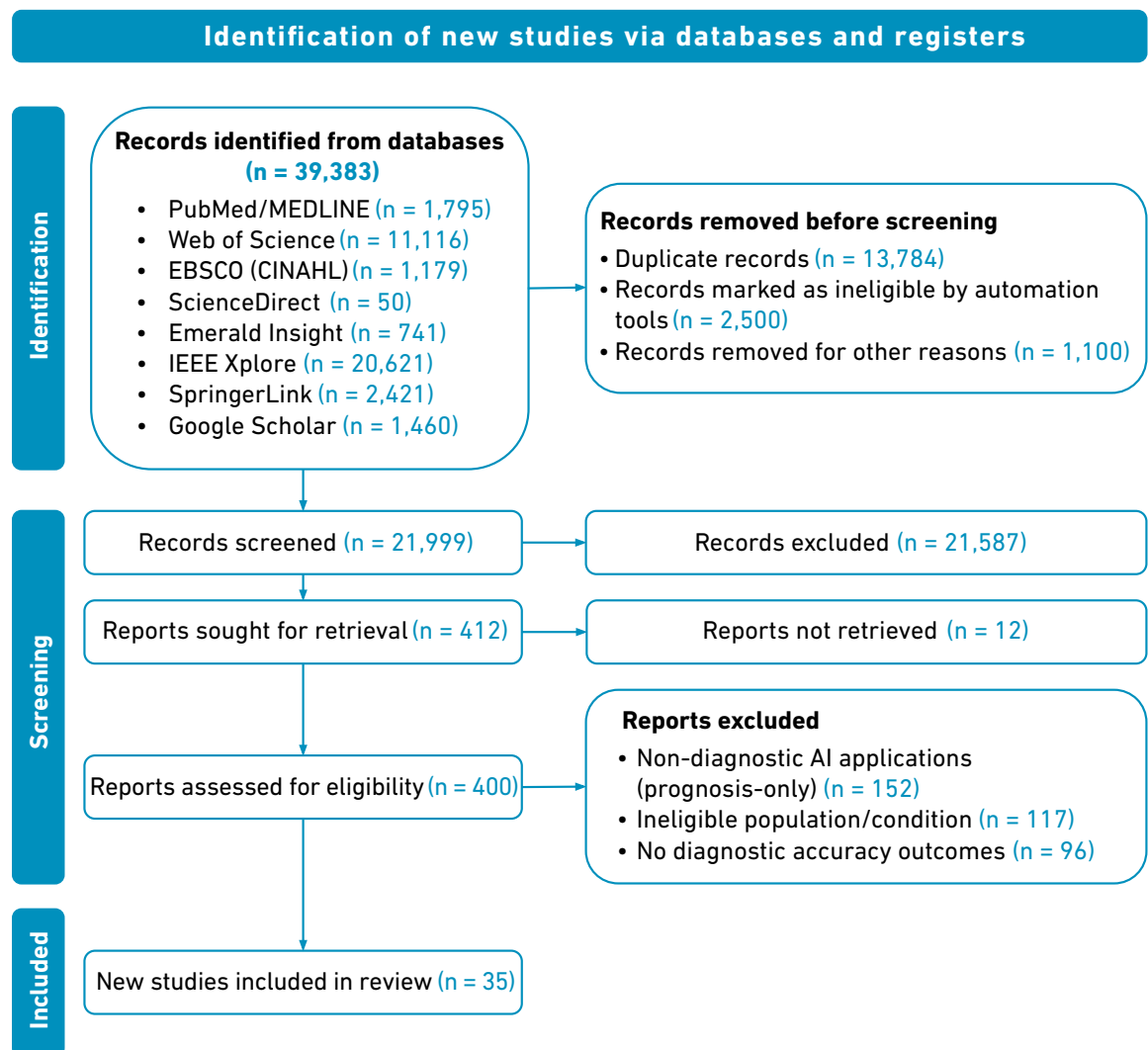


Figure 1. PRISMA Flow Diagram.

Note. Study selection process following PRISMA 2020.

Across the included studies, sample size descriptors were heterogeneous (patients, ECGs, beats, or image instances). Among studies reporting patient-level denominators, the cumulative number of participants was 94,356, with per-study cohorts ranging from 90 to 27,652. When records rather than participants were reported, the largest datasets comprised 202,945 to 259,789 ECG records, while the smallest comprised 130 to 573 individuals or examinations.

Characteristics of included studies

Table 1 presents the methodological features of the included studies. Retrospective cohorts were most frequent (n = 23, 65.7%), followed by prospective cohorts (n = 4, 11.4%), cross-sectional designs (n = 3, 8.6%), and a diagnostic validation study (n = 1, 2.9%). Moreover, one case-control study was identified (n = 1, 2.9%), and three reports summarized mixed or not explicitly stated designs (n = 3, 8.6%). Regarding study setting, these comprised single-center (n = 21, 60.0%) and multi-center designs (n = 9, 25.7%), with five studies using national or registry-like sources (n = 5, 14.3%). Enrolment periods ranged from single-year snapshots (e.g., 2013) to multi-year spans exceeding a decade (e.g., 2001–2012). Populations included emergency, inpatient, and outpatient cohorts. In terms of diagnostic indications, CAD was the most populated stratum (n = 18), HF contributed 9 studies, and ACS/MI contributed 4; the remaining studies addressed valvular heart disease (VHD) (n = 1), carotid artery stenosis (CAS) and major adverse cardiovascular events (MACE) (n = 1), cardiac rhythm classification (n = 1), and congenital heart disease (n = 1).

Table 1. General characteristics of included studies.

Author(s), Year	Country Setting	Study Design, n, Time Frame	Main Intervention	Outcomes
Stankovic et al., 2014 [62]	Serbia, Single-center	Retrospective cohort; n = 269 (181 with analyzable strain); NR	Echocardiography, GLS strain	CAD diagnosis
Cella et al., 2015 [63]	Italy, Multi-center	Retrospective cohort; n = 90; 2001–2012	Clinical data, classical ML/ANN	VHD diagnosis
Kim et al., 2015 [64]	South Korea, EHR	Cross-sectional; n = 748; 2013	Clinical data, classical ML/ANN	CAD diagnosis
Mülayim & Alaybeyoğlu, 2016 [65]	Turkey, Single-center	Retrospective cohort; n = 100; NR	Clinical data, classical ML/ANN	CAD diagnosis
Orphanou et al., 2016 [66]	Czechia, Multi-center	Retrospective cohort; n = 709; 2001–2002	Clinical data, classical ML/ANN	CAD diagnosis
Ribeiro et al., 2016 [67]	USA, Multi-center	Cross-sectional; n = 11,178; NR	Clinical data, classical ML/ANN	HF diagnosis
Cheng & Chiu, 2017 [68]	Taiwan, EHR	Retrospective cohort; n = 317; NR	Clinical data, classical ML/ANN	CAS\MACE
Dekamin & Sheibatolhamdi, 2017 [69]	Iran, Single-center	Retrospective cohort; n = 303; 2011–2013	Clinical data, classical ML/ANN	CAD diagnosis
Kumar et al., 2017 [70]	India, EHR	Retrospective cohort; n = 182,013 beats; NR	ECG/PPG, classical ML	CAD diagnosis
Mansoor et al., 2017 [71]	USA, EHR	Retrospective cohort; n = 12,047 (Dev/Int) + 24,959 (Ext); 2011–2013	Clinical data, classical ML/ANN	MACE/MI diagnosis
Prakash et al., 2017 [72]	USA, MIMIC-III	Retrospective cohort; n = 2,852; NR	Clinical text, NLP	CAD diagnosis
Wiharto et al., 2017 [73]	Indonesia, EHR	Retrospective cohort; n = 303; NR	Clinical data, classical ML/ANN	CAD diagnosis
Banerjee et al., 2018 [74]	India, MIMIC II + Single-center hospital	Retrospective cohort; n = 130 (MIMIC-II) + 99 (hospital); NR	ECG/PPG, classical ML	CAD diagnosis

Acharya et al., 2019 [75]	Singapore, Multi-center	Retrospective cohort; n = 140,000 ECG segments; NR	ECG, DL	HF diagnosis
Panicacci et al., 2019 [76]	Italy, EHR	Retrospective cohort; n = 14,616; 2010–2017	Clinical data, classical ML/ANN	HF diagnosis
Zhang et al., 2019 [77]	China, Single-center	Diagnostic validation; n = 8,486 ECGs; NR	ECG, DL	ACS/MI diagnosis
Banerjee et al., 2020 [78]	USA, MIMIC II	Retrospective cohort; n = 200 (development) + 200 (in-house); NR	ECG, DL	CAD diagnosis
Rjoob et al., 2020 [79]	UK & Ireland, Single-center	Retrospective cohort; n = 188; NR	Clinical data, classical ML/ANN	MI diagnosis
Zhang & Xia, 2020a [80]	China, Multi-center	Retrospective cohort; n = 91 subjects (33 CHF, 58 controls); NR	ECG, DL	HF diagnosis
Zhang & Xia, 2020b [81]	China, Multi-center	Retrospective cohort; n = 91 subjects (33 CHF, 58 healthy); 1988–2000	ECG, DL	HF diagnosis
Zhang et al., 2020 [82]	China, Single-center	Retrospective cohort; n = 259,789 train; 18,018 test; 2018–2019	ECG, DL	Cardiac rhythm classification
Cherradi et al., 2021 [83]	Morocco, EHR	Retrospective cohort; n = 573; NR	Clinical data, classical ML/ANN	CAD diagnosis
Li et al., 2021 [84]	China, Single-center	Retrospective cohort; n = 5,819; 2015–2018	Clinical data, classical ML/ANN	CAD diagnosis
Muhammad & Algehyne, 2021 [85]	Nigeria, Multi-center	Cross-sectional; n = 506; NR	Clinical data, classical ML/ANN	CAD diagnosis
Zahiri et al., 2021 [86]	Iran, Single-center	Retrospective cohort; n = 3,318; 2018–2020	Imaging, DL	CAD diagnosis
Chen, 2024 [87]	Japan, Single-center	Retrospective cohort; n = 28,000 ECGs (training); NR	ECG, DL	HF diagnosis
Karabayir et al., 2024 [88]	USA, Multi-center	Retrospective cohort; n = 202,945 ECGs; to 2024	ECG, DL	HF diagnosis
Li et al., 2024 [89]	China, Multi-center	Retrospective cohort; n = 303 (UCI Cleveland) + 1,024 (Kaggle); NR	Clinical data, classical ML/ANN	CAD diagnosis
Shao et al., 2024 [90]	China, Single-center	Case-control; n = 3,039; 2009–2022	Imaging, DL	Congenital heart disease diagnosis
Than et al., 2024 [91]	New Zealand, Multi-center	Retrospective cohort; n = NR; NR	Clinical data, classical ML/ANN	ACS/MI diagnosis
Ye et al., 2025 [92]	China, Single-center and multi-center	Prospective cohort; n = 3,271; 2017–2021	Angiography, classical ML	CAD diagnosis
Andersson et al., 2025 [93]	Sweden, Single-center	Prospective cohort; n = 157 train; 53 test; NR	Clinical data, classical ML/ANN	HF diagnosis
Dou et al., 2025 [94]	China, Multi-center	Prospective cohort; n = 27,652; 2017–2019	Clinical data, classical ML/ANN	CAD diagnosis
Shin et al., 2025 [95]	South Korea, EHR	Retrospective cohort; n = 3,591; NR	Multimodal, DL	HF diagnosis
Zhang et al., 2025 [96]	China, Single-center	Prospective cohort; n = 1,418; 2022	Imaging, classical ML	CAD diagnosis

Note. n: Sample size; **NR**: Not reported; **ML**: Machine learning; **ANN**: Artificial neural network; **MACE**: Major adverse cardiovascular events; **DL**: Deep Learning; **CAD**: Coronary artery disease; **GLS**: Global longitudinal strain; **ECG**: Electrocardiogram; **NLP**: Natural language processing; **PPG**: Photoplethysmography; **MIMIC-III**: Medical Information Mart for Intensive Care III; **MIMIC-II**: Medical Information Mart for Intensive Care II; **ACS**: Acute coronary syndrome; **MI**: Myocardial infarction; **PCI**: Percutaneous coronary intervention; **EHR**: Electronic health record; **HF**: Heart failure. **CAS**: Carotid artery stenosis; **VHD**: Valvular heart disease.

In terms of publication timeline, the earliest included study was published in 2014 and the most recent in 2025. Accordingly, panel A of Figure 2 displays the annual distribution of the 35 studies. Output was limited before 2016 (2014, $n = 1$; 2015, $n = 2$; 2016, $n = 3$), then increased from 2017 onward, with 2017–2019 contributing 10 studies in total. A distinct peak was observed in 2020 ($n = 5$), followed by sustained activity in 2021 ($n = 4$). Production increased again in 2024–2025, which together accounted for 10 studies (five in 2024 and five in 2025). The lowest annual counts occurred at the beginning of the period (2014–2016), and the highest combined two-year count was observed in 2024–2025.

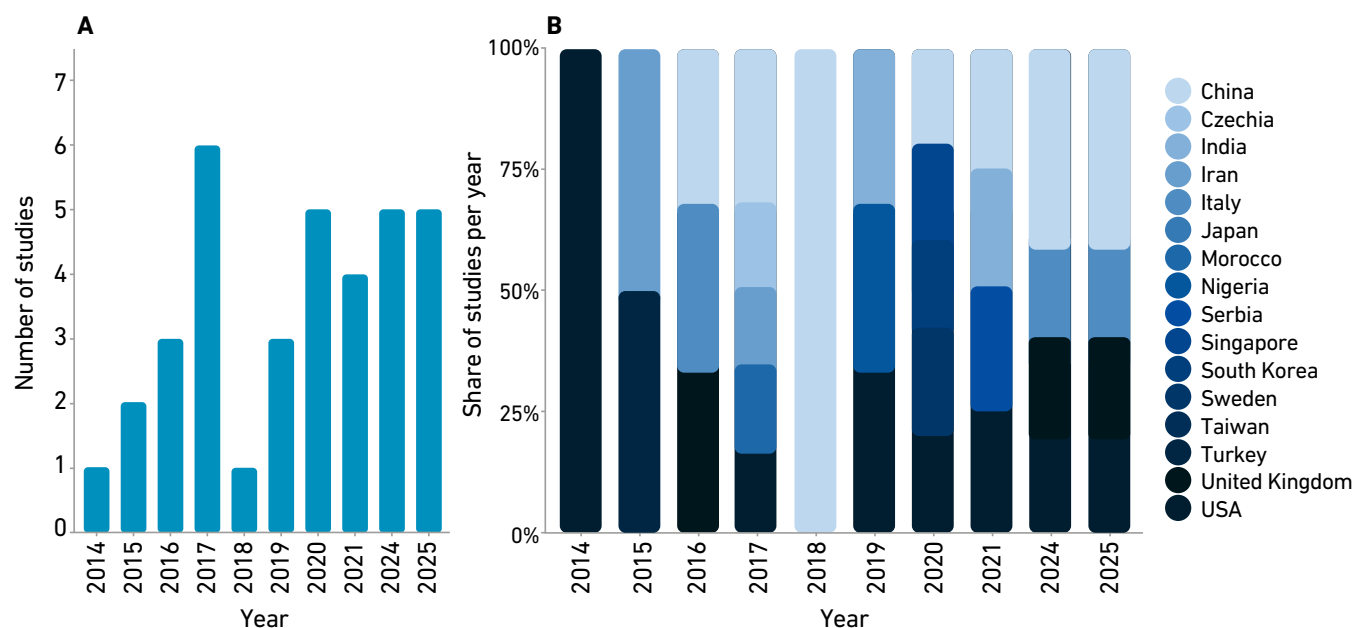


Figure 2. Temporal and geographic distribution of included studies.

Note. Panel A shows the annual number of included studies (darker blue indicates higher counts). Panel B displays the country distribution by year as stacked bars.

In parallel, panel B of Figure 2 summarizes the country contribution per year. China provided the largest number of studies ($n = 10$), followed by the United States ($n = 6$) and the United Kingdom ($n = 3$). Countries contributing two studies each were India ($n = 2$), Iran ($n = 2$), and Italy ($n = 2$). Single-study contributors included Czechia, Japan, Morocco, Nigeria, Serbia, Singapore, South Korea, Sweden, Taiwan, and Turkey ($n = 1$ each). Multi-center networks were noted across China, the United Kingdom, the United States, and Italy, while single-center cohorts predominated in Serbia, Turkey, Morocco, Nigeria, Japan, Sweden, and Taiwan.

AI Models and Validation in Cardiovascular Disease

An overview of AI model characteristics and validation methods in cardiovascular disease is provided in Table 2. Input modalities were clinical/EHR variables ($n = 14$, 40.0%), ECG or photoplethysmography (PPG) signals ($n = 12$, 34.3%), medical imaging (echocardiography, chest radiographs, angiography; $n = 6$, 17.1%), and multimodal pipelines ($n = 3$, 8.6%). Classical ML such as logistic regression, support vector machine (SVM), k-nearest neighbors

(k-NN), Naïve Bayes, and gradient boosting, appeared in 17 studies (48.6%). Additionally, DL methods such as convolutional neural networks (CNN), recurrent neural networks (RNN), and transformers were used in 14 studies (40.0%), while rule-based or risk score approaches were applied in 4 studies (11.4%). Candidate predictors ranged from raw signals (no handcrafted features) to high-dimensional EHR matrices (e.g., 1,514 candidates reduced to 572) and very large vocabularies in text/natural language processing (NLP) models (~ 20,000 tokens). Final model parsimony varied from ≤ 10 variables after recursive feature elimination (RFE)/least absolute shrinkage and selection operator (LASSO) to end-to-end feature learning.

Table 2. Characteristics of AI models and validation approaches.

Author(s), Year	Modeling method	No. of predictors		Selection		Missing data	Validation
		Candidate	Final	Candidate	Final		
Stankovic et al., 2014 [62]	Echocardiography, GLS metric	18 segments	12 segments	Protocol-driven segmentation	Protocol-based exclusion	Quality-based exclusion (33%)	ROC analysis (no resampling)
Cella et al., 2015 [63]	Classical ML, logistic regression (LASSO)	26	12	Filter: correlation/collinearity	Embedded: LASSO	NR	k-fold CV (k = 10, repeated)
Kim et al., 2015 [64]	Classical ML, CART (fuzzy rules)	9	9	Clinical/expert driven	Embedded: tree-based	Exclusion of uncertain responses	Hold-out split (70/30)
Mülayim & Alaybeyoğlu, 2016 [65]	Classical ML, SVM (meta-optimized)	13	13	Clinical/expert driven	Meta-optimization	NR	Hold-out split
Orphanou et al., 2016 [66]	Classical ML, Naive Bayes (+ temporal rules)	NR	9	Clinical/expert driven	Filter: discretization/entropy	Imputation, random forest (missForest)	k-fold CV (k = 10)
Ribeiro et al., 2016 [67]	Classical ML, decision path (AUROC)	21	Per-patient (avg 3.4)	Data-driven (available variables)	Filter: AUROC	Imputation, iterative (nonparametric)	k-fold CV (k = 20)
Cheng & Chiu, 2017 [68]	ANN, MLP	13	13	Clinical/expert driven	None	NR	Hold-out split (80/20)
Dekamin & Sheibatolhamdi, 2017 [69]	Classical ML, k-NN	53	53	Clinical/expert driven	None	NR	Hold-out split (80/20)
Kumar et al., 2017 [70]	Classical ML, SVM (LS-SVM)	5	5	Signal-derived features	None	NR	k-fold CV (k = 10)
Mansoor et al., 2017† [71]	Classical ML, logistic regression	32	11	Clinical/expert driven (literature)	Wrapper: backward elimination	NR	Hold-out (20%) + external validation
Prakash et al., 2017 [72]	DL, memory network (NLP)	20,000 vocabulary	20,000	Frequency-based vocabulary	None	NR	Hold-out split (80/10/10)
Wiharto et al., 2017 [73]	ANN, MLP	13	7	Clinical/expert driven	Wrapper: logistic screening	NR	Hold-out split (70/30)
Banerjee et al., 2018† [74]	Classical ML, SVM (RBF)	28	21	Clinical/expert driven	Filter: MIC + CV tuning	Clean-signal selection	External validation (in-house)
Acharya et al., 2019 [75]	DL, CNN (ECG)	Raw	Raw	N/A (raw signals)	N/A (end-to-end)	Signal preprocessing	k-fold CV (k = 10)

Panicacci et al., 2019 [76]	Classical ML, random forest (+ SMOTE)	1,514	572	Administrative/EHR-derived	Wrapper: Boruta	Temporal exclusion (pre-2010)	k-fold CV (k = 10) + independent test (20%)
Zhang et al., 2019† [77]	DL, CNN (ECG, 12-lead)	Raw	Raw	Raw signals	N/A (automated feature extraction)	NR	External validation (independent cohort)
Banerjee et al., 2020† [78]	DL, CNN, LSTM (ECG + HRV)	NR	NR	Hybrid (raw + expert features)	Filter: MIC	NR	k-fold CV (tuning) + external validation
Rjoob et al., 2020 [79]	Classical ML, logistic regression	327	20	All candidate variables	Filter + wrapper (JMI + backward)	NR	k-fold CV (k = 5)
Zhang & Xia, 2020a [80]	DL, CNN (DenseNet, ECG)	Raw	Automated features	Raw signals	N/A (end-to-end)	Signal preprocessing (noise removal/resampling)	Hold-out split (70/30; inter-patient)
Zhang & Xia, 2020b [81]	DL, RNN (GRU, attention; ECG)	180/time-step	Attention-weighted	Window-based segmentation	N/A (attention)	Signal preprocessing	Hold-out split (70/30)
Zhang et al., 2020 [82]	DL, CNN (ECG)	Raw	Automated features	Raw signals	N/A (end-to-end)	Preprocessing/augmentation	Independent test set
Cherradi et al., 2021† [83]	Classical ML, k-NN	30	17	Clinical/expert driven	Expert ranking	Complete-case analysis (Cleveland)	k-fold CV (k = 10) + independent test
Li et al., 2021† [84]	Ensemble, stacked classical ML + deep learning	11	11	Clinical/expert driven	Stacking (logistic combiner)	Imputation, mode/mean	k-fold CV (k = 10) + independent validation (n = 582)
Muhammad & Algehyne, 2021 [85]	Rule-based, fuzzy expert system	12	12	Clinical/expert driven	Embedded: rule induction	Data cleansing	Hold-out split (fixed test set)
Zahiri et al., 2021 [86]	DL, CNN (medical imaging)	Image features	Image features	Raw images	N/A (automated deep features)	Complete-case analysis	k-fold CV (k = 5)
Chen, 2024† [87]	DL, CNN (ECG, Lead I)	Raw	Deep features	Automated extraction	N/A (CNN layers)	NR	k-fold CV + independent test
Karabayir et al., 2024 [88]	DL, CNN (ECG, wearable Lead I)	Raw	Lead-I features	Simulated wearable data	Automated feature extraction	NR	Hold-out split (10%)
Li et al., 2024 [89]	ANN, BP network (meta-optimized)	13 to 32 (one-hot)	UCI: 9; Kaggle: 13	Clinical/expert driven	Dimensionality reduction (PCA)	Encoding/normalization	k-fold CV (k = 10) + independent test (25%)
Shao et al., 2024 [90]	DL, CNN (chest radiographs)	NR	NR	Raw images	N/A (deep features)	NR	k-fold CV (k = 4) + independent test (20%)
Than et al., 2024† [91]	Risk score, CoDE-ACS	10	10	Prespecified variables	Prespecified	NR	External validation

Ye et al., 2025† [92]	Classical ML, gradient boosting (XGBoost)	NR	NR	Clinical/expert driven	None (all features)	Exclusion >30% missing; FCS imputation	Hold-out (70/30) + k-fold CV (k=10) + external validation
Andersson et al., 2025 [93]	Classical ML, gradient boosting (XGBoost)	Initial: NR	10 (after RFE)	Clinical + echocardiographic variables	Wrapper: RFE	Imputation, Bayesian ridge/ k-NN	k-fold CV (k = 10) + independent test (25%)
Dou et al., 2025 [94]	Classical ML, gradient boosting (XGBoost)	11	11	Clinical/expert driven	Post-hoc importance (SHAP)	Multiple imputation (3.5% missing)	Hold-out split (70/30)
Shin et al., 2025 [95]	DL, multimodal (Transformer + MLP)	15,133	313	QC filters (SNPs); all EHR	Filter: chi- square (SNPs)	Imputation, k-NN (EHR)	Repeated hold-out (80/20)
Zhang et al., 2025 [96]	Classical ML, logistic regression (fundus + clinical)	15	15	Clinical/expert driven	All included	Complete-case analysis	k-fold CV (k = 5)

Note: †: External or independent validation performed; **ML:** Machine learning; **AUROC:** Area under the receiver operating characteristic curve; **CV:** Cross-validation; **CART:** Classification and regression tree; **GLS:** Global longitudinal strain; **ROC:** Receiver operating characteristic; **LASSO:** Least absolute shrinkage and selection operator; **SVM:** Support vector machine; **NR:** Not reported; **LS-SVM:** Least-squares support vector machine; **k-NN:** k-nearest neighbors; **ANN:** Artificial neural network; **MLP:** Multilayer perceptron; **DL:** Deep learning; **NLP:** Natural language processing; **RBF:** Radial basis function; **MIC:** Maximal information coefficient; **CNN:** Convolutional neural network; **ECG:** Electrocardiogram; **SMOTE:** Synthetic minority oversampling technique; **EHR:** Electronic health record; **DenseNet:** Densely connected convolutional network; **JMI:** Joint mutual information; **LSTM:** Long short-term memory; **HRV:** Heart rate variability; **RNN:** Recurrent neural network; **GRU:** Gated recurrent unit; **CoDE-ACS:** Collaboration for the Diagnosis and Evaluation of Acute Coronary Syndrome; **XGBoost:** Extreme Gradient Boosting; **SHAP:** Shapley additive explanations; **FCS:** Fully conditional specification; **BP:** Backpropagation; **PCA:** Principal component analysis; **UCI:** University of California, Irvine; **RFE:** Recursive feature elimination; **QC:** Quality control; **SNP:** Single-nucleotide polymorphism.

Internal validation predominantly used k-fold cross-validation (≥ 10 -fold in several cases), recorded in at least 18 studies (51.4%). Fixed hold-out splits appeared in 10 studies (28.6%). External/independent validation was reported in nine studies (25.7%), including independent cohorts or temporally separated test sets. Furthermore, missing-data handling comprised complete-case analysis ($n = 3$), deterministic exclusion of uncertain entries ($n = 2$), and explicit imputation (iterative non-parametric, missForest, mode/mean, Bayesian ridge, k-NN, or fully conditional specification) in seven studies. Feature selection strategies included filter (correlation, entropy, maximal information coefficient (MIC), embedded (LASSO, tree-based), and wrapper (backward elimination, Boruta, RFE), with several deep models bypassing handcrafted features.

Risk of Bias and Applicability Assessment

As presented in Figure 3, the methodological quality of the included studies was evaluated using the PROBAST+AI framework [55]. The distribution of judgments indicated variability across domains: participant selection and predictors were most frequently rated as low or unclear, whereas outcome assessment and analysis domains showed higher proportions of high risk. Overall, 11 studies (31.4%) were judged to have one or more domains rated as high risk, and only 6 studies (17.1%) were consistently classified as low risk across all domains. Applicability concerns were generally minimal in reference standard and analysis domains, while 9 studies (25.7%) were judged as unclear due to incomplete reporting of data sources, predictor handling, or validation strategy.

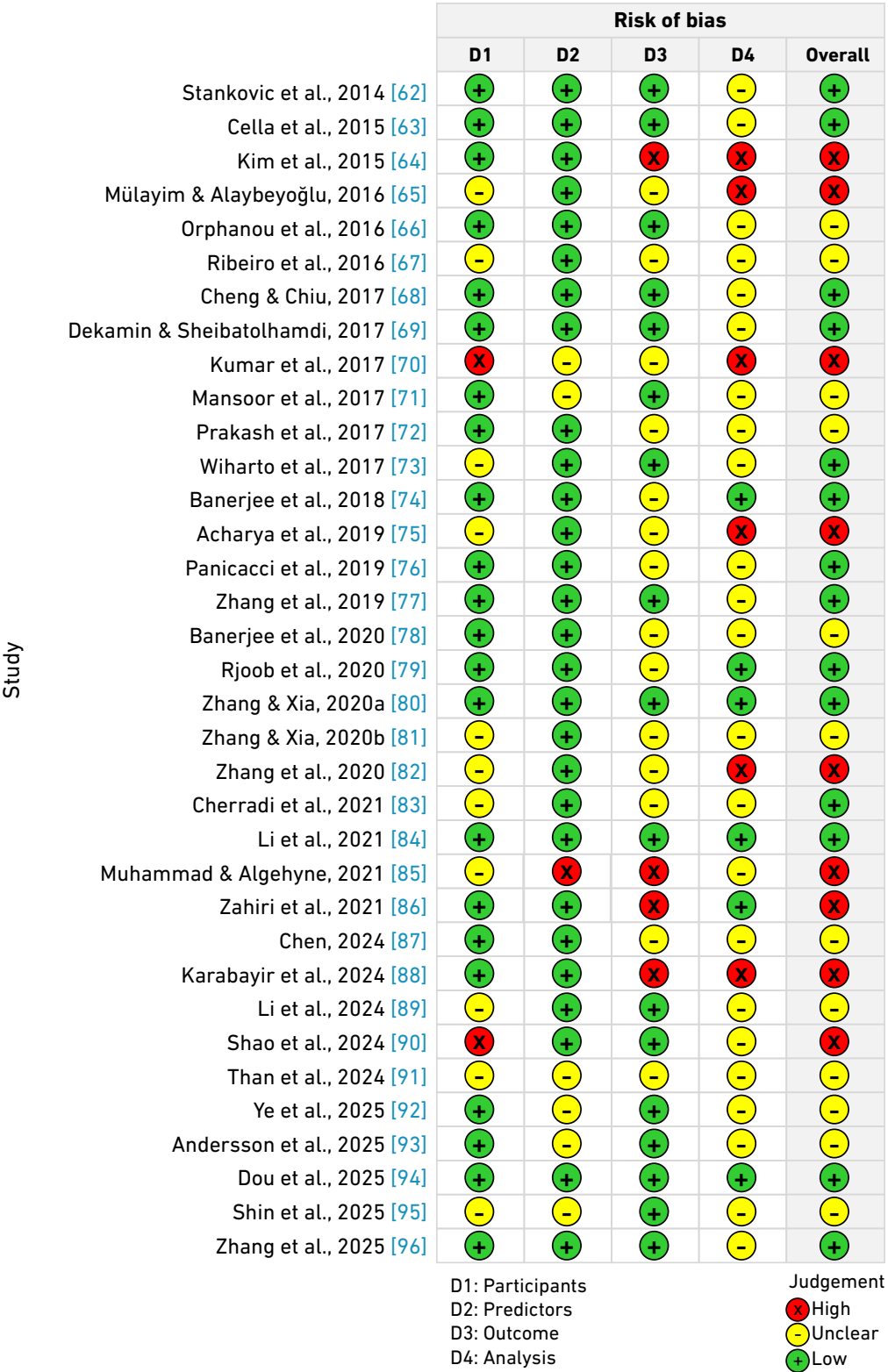


Figure 3. Risk of Bias Assessment of Included Studies.

Note. Domain level and overall assessment of risk of bias in the included studies.

Diagnostic Accuracy of AI Models

Table 3 provides an overview of diagnostic accuracy metrics for AI models in cardiovascular disease. Specifically, AUROC was reported in 22 out of 35 studies (62.9%), with eight providing 95% CI. Across all studies reporting AUROC, values ranged from 0.594 (lower bound: 0.513–0.675) to 0.9991, with multiple reports ≥ 0.95 and several between 0.75–0.90. Furthermore, accuracy was available for 23 studies (65.7%), with values ranging from 0.694 to 0.996, while sensitivity and specificity were reported in 21 and 19 studies, respectively. Sensitivities ranged from 0.686 to 0.996, with 12 out of 21 studies (57.1%) reporting sensitivity ≥ 0.90 . Specificities ranged from 0.256 to 1.000, with 8 out of 19 (42.1%) at or above 0.95. The narrowest AUROC CI among studies reporting CIs was 0.012 (0.9458–0.9671), and the widest was 0.206 (0.791–0.997). Point estimates ≥ 0.95 frequently coincided with CNN ECG or imaging inputs, while lower estimates (≤ 0.77) were observed in classical-ML models using limited clinical inputs.

Table 3. Diagnostic performance of AI models.

Author(s), Year	AUROC (95% CI)	Accuracy	Sensitivity	Specificity
Stankovic et al., 2014 [62]	0.850 (0.79–0.91)	NR	NR	NR
Cella et al., 2015 [63]	0.840 (0.77–0.93)	NR	NR	NR
Kim et al., 2015 [64]	0.594 (0.513–0.675)	0.695	0.931	0.256
Mülayim & Alaybeyoğlu, 2016 [65]	NR	0.9	0.927	0.86
Orphanou et al., 2016 [66]	0.850 (0.780–0.910)	0.844	0.972	NR
Ribeiro et al., 2016 [67]	0.709 (0.684–0.735)	NR	NR	NR
Cheng & Chiu, 2017 [68]	0.77	0.825	0.894	0.574
Dekamin & Sheibatolhamdi, 2017 [69]	NR	0.909	0.933	0.857
Kumar et al., 2017 [70]	NR	0.996	0.996	0.996
Mansoor et al., 2017 [71]	0.84	0.890	NR	NR
Prakash et al., 2017 [72]	0.767	NR	NR	NR
Wiharto et al., 2017 [73]	0.921	0.863	0.848	0.882
Banerjee et al., 2018 [74]	NR	NR	0.830 (avg)	0.860 (avg)
Acharya et al., 2019 [75]	NR	0.99	0.989	0.99
Panicacci et al., 2019 [76]	NR	NR	0.896	0.987
Zhang et al., 2019 [77]	0.9610 (0.933–0.989)	0.91	0.929	0.893
Banerjee et al., 2020 [78]	NR	NR	0.940	0.920
Rjoob et al., 2020 [79]	0.890 (0.788–0.992)	0.825	0.852	0.8
Zhang & Xia, 2020a [80]	NR	0.95	0.894	0.995
Zhang & Xia, 2020b [81]	NR	0.934	0.931	0.937
Zhang et al., 2020 [82]	NR	0.95	NR	NR
Cherradi et al., 2021 [83]	NR	0.968	0.96	1
Li et al., 2021 [84]	0.75	0.694	0.686	0.805
Muhammad & Algehyne, 2021 [85]	NR	0.946	0.954	0.95
Zahiri et al., 2021 [86]	0.8450 (0.832–0.857)	0.7562	0.7856	0.7434
Chen, 2024 [87]	0.9576 (0.9458–0.9671)	NR	NR	NR
Karabayir et al., 2024 [88]	0.890 (0.89–0.90)	NR	0.81	0.81

Li et al., 2024 [89]	0.9991	0.9550	0.9703	0.9732
Shao et al., 2024 [90]	0.9922	0.963	0.965	0.963
Than et al., 2024 [91]	NR	NR	0.9830	NR
Ye et al., 2025 [92]	0.984	NR	0.965	0.975
Andersson et al., 2025 [93]	0.894 (0.791–0.997)	NR	NR	NR
Dou et al., 2025 [94]	0.7690 (0.753–0.784)	NR	NR	NR
Shin et al., 2025 [95]	0.834	NR	0.7810 (weighted recall)	NR
Zhang et al., 2025 [96]	0.710 (0.64–0.78)	NR	NR	NR

Meta-analysis of AI performance in cardiovascular disease

As shown in Figure 4, the pooled AUROC was 0.823 (95% CI 0.754–0.892). Twelve effects entered the random effects model, and AUROCs lacking variance were excluded from pooling. Heterogeneity was very high ($I^2 = 98.4\%$, $\tau^2 = 0.0108$, Q -test $p < 0.0001$). The 95% PI was 0.59–1, with the upper bound truncated to the AUROC scale, which supports a descriptive interpretation of the summary estimate. Individual AUROCs ranged from 0.594 to 0.961. When grouped into predefined AUC intervals, one estimate was < 0.70 , three were 0.70–0.79, six were 0.80–0.89, and two were ≥ 0.95 . CI lower limits spanned 0.513 to 0.947 and upper limits 0.675 to 0.997. CI widths varied from 0.021 to 0.206. Study weights were narrowly distributed from 7.2% to 9.1%, and the forest plot displays square sizes proportional to weight with horizontal bars representing the 95% CI. External validation with complete 95% CI was reported in 9 of 35 studies. Accordingly, we did not undertake a synthesis restricted to externally validated estimates. The overall estimate prioritizes external effects when available and otherwise adheres to the prespecified primary analyses. Excluding high risk studies did not materially alter the pooled AUROC, and the 95% PI remained similar.

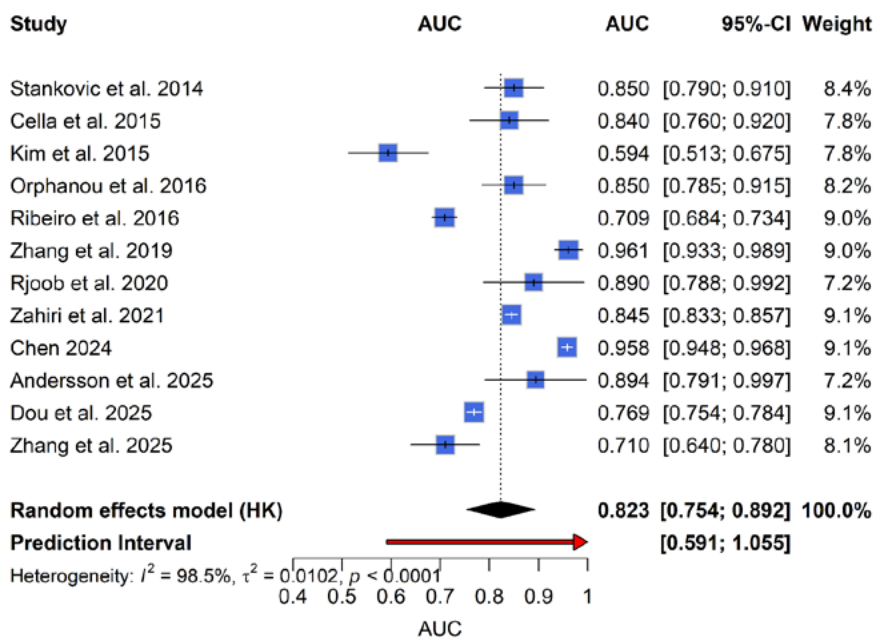


Figure 4. Pooled AUROC meta-analysis of AI in cardiovascular disease
Note. Forest plot of pooled AUROC for AI in cardiovascular disease.

Diagnostic Performance by cardiac condition

Figure 5 presents pooled discriminative performance by cardiac disease category. For CAD, the pooled AUROC was 0.86 (95% CI 0.83–0.89), synthesizing studies using clinical/EHR features, ECG-derived features, and angiography-linked models. For ACS/MI, the pooled AUROC was 0.88 (95% CI 0.84–0.91). For heart failure, the pooled AUROC was 0.87 (95% CI 0.83–0.90). The numeric ordering was ACS/MI (highest), HF (intermediate), and CAD (lowest), with overlapping CI and close absolute differences (≤ 0.02). Category-specific heterogeneity mirrored the overall forest, with broader intervals in CAD (multiple modalities and variable reference standards) and narrower intervals in ECG-centric ACS/MI studies. As depicted in Figure 4, the spread within categories was substantial: CAD AUROCs extended from 0.594 to 0.9991; ACS/MI from 0.710 to 0.9610; and HF from 0.834 to 0.9576. Sensitivities in CAD studies ranged from 0.686 to 0.989 (≥ 0.95 in several ECG/DL pipelines), while specificities ranged from 0.256 to 0.987. In ACS/MI, sensitivity estimates reached 0.983 (single-analyte pathway), with specificities up to 0.937. HF models reported AUROCs near 0.834–0.9576, with sensitivity frequently ≥ 0.90 in 1-lead ECG CNNs and specificities spanning approximately 0.81–0.9732 when clinical variables were integrated.

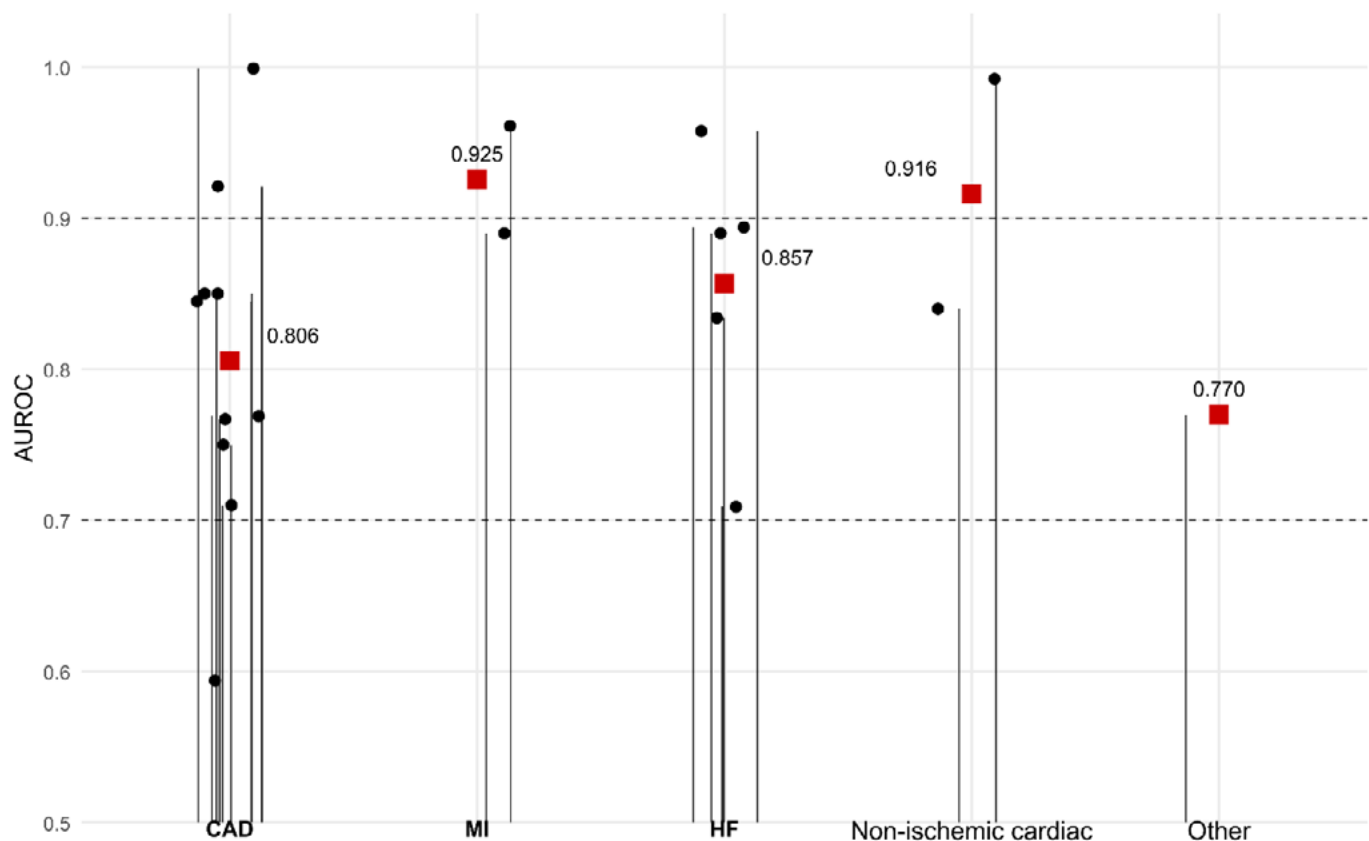


Figure 5. Discriminative performance of AI across cardiac disease categories.

Note. AUROC by cardiac disease category; red squares show means (n=35).

Discussion

The broad search process reflects the selective nature of diagnostic AI evidence, where only a small subset of available studies provides analyzable accuracy data, consistent with prior cardiovascular AI reviews emphasizing stringent inclusion criteria to ensure diagnostic validity [97-99]. By encompassing ECG, imaging, and EHR pipelines, our sample captured diverse diagnostic contexts, allowing condition-specific synthesis while retaining a clinically oriented translational perspective [97,99,100].

These findings support a cautious, context-dependent interpretation of AI, where use is not inherently beneficial but depends on external validation, local calibration, and explicit decision-threshold specification. We do not propose these models as standalone diagnostics. Among studies reporting specificity, 8 of 19 (42.1%) achieved values of 0.95 or higher. Where specificity is modest, the intended role is high-sensitivity screening to minimize missed events, with false positives resolved through standard confirmatory testing and locally calibrated thresholds.

We observed predominance of retrospective, single-center cohorts with heterogeneity in denominators (e.g. patients, ECGs, beats), which plausibly reflects differences in data capture across ECG-centric pipelines, imaging labs, and administrative/EHR sources [75,76,81,84]. Multicenter and registry designs were present but less frequent, mirroring the limited number of externally validated pipelines and highlighting transportability concerns when moving from local development to broader deployment [28,87,92,94]. This limitation also has equity implications, as most development cohorts originated from high-income regions, and diagnostic performance may differ in low- and middle-income settings due to variations in prevalence, device type, and data quality. Populations included emergency triage for ST-elevation myocardial infarction (STEMI), pre-test probability evaluations for CAD, and chronic care (HF detection and monitoring), reflecting diverse bedside applications ranging from rapid rule-in/rule-out to longitudinal surveillance [28,87,88,94]. This distribution supports clinical integration of AI as augmentation, with ECG models for front-door triage, EHR models for routing to imaging, and imaging models serving as second readers, rather than replacing clinician judgment [28,84,86,93].

Classical ML (logistic regression, tree-based, k-NN, Naïve Bayes) and DL (CNN/RNN/transformer) coexisted across modalities, with hand-engineered features dominating smaller tabular datasets and end-to-end learning prevailing for ECG/imaging where signal richness supports representation learning [67,73,75,81]. Internal k-fold validation was common, while external/temporal validation remained comparatively scarce, and when performed it frequently revealed performance attenuation consistent with domain shift [28,87,92,94]. Real world applicability is constrained by limited external testing. External or independent validation with complete 95% CI was reported in 9 of 35 studies, which corresponds to 25.7%. We therefore interpret the pooled AUROC as a descriptive average rather than a universally transportable benchmark. Consistent with the protocol, external estimates were prioritized when available, a wide PI was reported, and clinical deployment should follow explicit external validation, calibration (including thresholds), and ideally impact evaluation.

Imputation and feature selection were variably reported, with embedded methods (e.g., LASSO, tree-importance) aiding parsimony in EHR models and signal pipelines bypassing handcrafted steps, which may explain higher discrimination but uncertain calibration without explicit post-hoc adjustment [67,76,84,86]. At the bedside, this argues for local calibration and threshold setting, particularly where disease prevalence and acquisition protocols differ from those of development cohorts [28,84,87,94].

Our PROBAST+AI appraisal found frequent high-risk judgments in outcome assessment and analysis, which is clinically meaningful: label quality (expert consensus vs. reference standard) and analytic flexibility (multiple testing without correction) can inflate AUROC and lead to miscalibrated probabilities used at the point of care [28,76,84,86,90]. Small single-center datasets and public benchmarks were common, introducing spectrum effects and selection bias that may not reflect real world case-mix, particularly for borderline CAD and mild HF [65,73,83,87]. External testing repeatedly showed drops in discrimination, underscoring the need for site-specific recalibration and workflow-anchored evaluation before adoption [28,87,92,94]. Such validation, calibration, and ongoing monitoring are consistent with the principles typically required for regulatory evaluation of clinical AI systems. These concerns parallel review-level observations of heterogeneity, limited calibration reporting, and overoptimistic internal validation in cardiovascular AI, reinforcing the priority of transparent validation and reporting [98,99,101].

Across studies, discrimination was generally high, particularly in ECG/DL pipelines, while specificity showed wider variability. This pattern reflects methodological and clinical priorities: deep models trained on continuous signals are optimized to avoid missed events, naturally enhancing sensitivity, whereas specificity depends more on data balance, labeling consistency, and threshold calibration. In clinical terms, these models function mainly as rule-out or triage tools rather than confirmatory diagnostics, emphasizing the importance of local validation and threshold adjustment before deployment [28,75,78,86]. High-end performance clustered in CNN-based ECG/imaging models, whereas lower estimates were typical of classical ML on limited clinical inputs, a pattern consistent with representational capacity and the need for richer inputs when signal quality is limited [67,73,76,84]. Our distribution of AUROCs aligns with review summaries showing strong discrimination for CAD and HF with boosting/RF or DL, while also echoing warnings that pooled accuracy can be overstated without standardized definitions and calibration [97,98,100]. For clinicians, this means thresholds and PPV/NPV should be set against local prevalence and workflow impact, not solely against headline AUROC, particularly where specificity is modest [28,84,86,94].

Our pooled AUROC was 0.823 (95% CI 0.754–0.892). Heterogeneity was very high ($I^2 = 98.4\%$, $\tau^2 = 0.0108$), and the 95% PI was 0.59 to 1. Under a random effects framework with restricted maximum likelihood and Hartung–Knapp adjustments, the pooled value is interpreted as a descriptive average across heterogeneous contexts rather than a universal benchmark for any single setting. The wide dispersion largely reflects structural differences across datasets and model types. ECG and imaging pipelines trained on raw signals tended to yield higher AUROC and narrower internal CIs than EHR models reliant on tabular variables. Likewise, studies using invasive angiography or objective imaging as the reference standard showed lower uncertainty than those using expert consensus or incomplete ground truth. Deep-learning (DL) architectures captured richer temporal or spatial features and generally outperformed shallow or rule-based models, but at the cost of higher variability when externally tested. These factors collectively explain the magnitude of between-study variance observed, underscoring that heterogeneity arises from both data modality and methodological depth rather than random noise [97,99,102]. Clinically, a pooled AUROC near 0.82 supports augmentation in triage and test selection, yet the wide interval and high heterogeneity call for local validation, calibration, and predefined thresholds before deployment [103,104]. Accordingly, the pooled estimate is not intended to define actionable cutoffs but to summarize discriminative performance, acknowledging that clinical thresholds must be locally derived to avoid false reassurance or patient harm. Including studies from highly diverse clinical and technical settings likely increased heterogeneity. The summary AUROC is descriptive, with a wide 95% PI. Stratification by setting was not feasible given the small number of effects.

Pooling by condition showed ACS/MI (0.88), HF (0.87), and CAD (0.86) with overlapping intervals, a hierarchy that is consistent with reviews noting strong diagnostic performance in CAD imaging pipelines and robust discrimination in EHR risk stratification, while cautioning about heterogeneity and calibration gaps [94,98,103]. Within category spread can be interpreted by modality: ECG/DL excels for STEMI triage with externally validated AUROC 0.961, favoring rapid rule-in/rule-out and reduced delay to reperfusion; conversely, clinical CAD models show modest gains and benefit from integration with imaging or richer predictors to lift specificity [78,86,99]. For HF, single lead models reached high discrimination but fell on external data, indicating utility for screening/monitoring with mandatory site calibration and consideration of device variability in ambulatory workflows [75,81,87,88]. Finally, imaging AI as a second reader (e.g., single photon emission computed tomography (SPECT) polar maps, echo strain, chest radiography for congenital heart disease) supports targeted pathways and may reduce downstream testing when paired with clear thresholds and human oversight [86,90,93]. Added value over existing tools was not quantified since direct externally validated comparisons and decision curve analyses were rarely available. The pooled estimates describe diagnostic discrimination rather than superiority. Overall, these results highlight how AI can improve diagnostic precision and shorten time to clinical decision, while its broader contribution will depend on equitable validation and adaptation to diverse healthcare contexts.

Conclusions

AI diagnostic models show promising discrimination, yet generalizability remains uncertain given that only 9 of 35 studies (25.7%) reported external validation. The pooled AUROC is an overall descriptive average and should not be interpreted as evidence of reliability for deployment. The integration of these models into existing diagnostic pathways entails local validation and threshold calibration to account for site-specific prevalence and data acquisition practices. Furthermore, the translation of model outputs into clinical action requires continuous monitoring for performance drift and human oversight to ensure alignment with established reference standards. Ultimately, these findings indicate potential to enhance diagnostic efficiency and accuracy in cardiovascular care. This potential is contingent on rigorous external validation, local calibration, and continuous monitoring after implementation. Key limitations include potential sample bias from small single-center cohorts, reporting bias due to incomplete variance and calibration reporting, and inconsistent reference standards in several primary studies. These constraints may inflate some accuracy estimates and contribute to the wide prediction interval (PI); therefore, results should be interpreted with caution. Future research should prioritize multi-center external validation, the use of standardized diagnostic endpoints, and open benchmark datasets to enable reproducibility and comparative evaluation.

References

1. Ahmad A, Lim LL, Morieri ML, Tam CH ting, Cheng F, Chikowore T, et al. Precision prognostics for cardiovascular disease in Type 2 diabetes: a systematic review and meta-analysis. *Commun Med* [Internet]. 2024;4(1):11. doi: <https://doi.org/10.1038/s43856-023-00429-z>
2. Yu J, Liu X, Zhu Z, Yang Z, He J, Zhang L, et al. Prediction models for cardiovascular disease risk among people living with HIV: A systematic review and meta-analysis. *Front Cardiovasc Med* [Internet]. 2023;10:1138234. doi: <https://doi.org/10.3389/fcvm.2023.1138234>

3. Rodríguez Perón JM, Rodríguez Izquierdo MM. Síndrome metabólico y su correlación con ecuaciones de predicción del riesgo global de enfermedad cardiovascular. *Rev Cubana Med Gen Integr* [Internet]. 2022;38(3):e1909. Available from: <https://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=119564>
4. Chowdhury MA, Rizk R, Chiu C, Zhang JJ, Scholl JL, Bosch TJ, et al. The Heart of Transformation: Exploring Artificial Intelligence in Cardiovascular Disease. *Biomedicines* [Internet]. 2025;13(2):427. doi: <https://doi.org/10.3390/biomedicines13020427>
5. Lopez Gonzalez AA, Riutord Sbert P, Arroyo Bote S, Gonzalez San Miguel H, Rigo Vives M del M, Ramirez Manent JI. Determination of different cardiovascular risk scales in Spanish teaching staff: associated variables. *Med Segur Trab* [Internet]. 2021;67(263):99-111. doi: <https://doi.org/10.4321/s0465-546x2021000200003>
6. Lecarnaque-Rojas CG, Guerrero-Cueva JI, Guillen-López OB. Knowledge about cardiovascular diseases in a first-level healthcare center in Lima, Peru. *Rev Peru Med Exp Salud Publica* [Internet]. 2024;41(3):281-6. doi: <https://doi.org/10.17843/rpmesp.2024.413.13575>
7. Buitrago-Gomez N, Figueroa L, Casanova ME. La prediabetes y su impacto sobre la salud cardiovascular: artículo de revisión. *Univ Salud* [Internet]. 2022;24(2):170-83. doi: <https://doi.org/10.22267/rus.222402.270>
8. León-Samaniego GF, Vásquez Espinoza G de J, Encalada Campos GE, Bustamante Silva JS. Prevalencia de obesidad y dislipidemias, y su relación con la hipertensión arterial en trabajadores universitarios en Ecuador. *Salud Bienestar Colect* [Internet]. 2020;4(1):33-43. Available from: <https://pesquisa.bvsalud.org/portal/resource/pt/biblio-1179914>
9. Arredondo Buce AE, de Quezada Lopez F, Guerrero Jimenez G, Arredondo Buce AE, de Quezada López F, Guerrero Jiménez G. Avances actuales en la reducción de los factores de riesgo cardiovasculares. *Revista Med Electrón* [Internet]. 2020;42(1):1657-68. Available from: <https://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=95504>
10. Morales MJQ. Riesgo cardiovascular en personal médico y paramédico del Hospital Nacional Nicolasa Cruz de Jalapa. *Rev Cienc Multidiscip CUNORI* [Internet]. 2024;8(2):1-14. doi: <https://doi.org/10.36314/cunori.v8i2.255>
11. Dattoli-Garcia CA, Jackson-Pedroza CN, Gallardo-Grajeda AL, Gopar-Nieto R, Araiza-Garygordobil D, Arias-Mendoza A. Infarto agudo de miocardio: revisión sobre factores de riesgo, etiología, hallazgos angiográficos y desenlaces en pacientes jóvenes. *Arch Cardiol Mex* [Internet]. 2021;91(4):485-92. doi: <https://doi.org/10.24875/acm.20000386>
12. Uzun Ozsahin D, Ozgocmen C, Balcioglu O, Ozsahin I, Uzun B. Diagnostic AI and Cardiac Diseases. *Diagnostics* [Internet]. 2022;12(12):2901. doi: <https://doi.org/10.3390/diagnostics12122901>
13. Rincon Rueda ZR, Carrillo Vasquez AC, Cepeda Florez KT, Quintero Santos LS, Sanchez-Delgado JC. Impacto de la rehabilitación cardiaca de 8 vs 12 semanas sobre la calidad de vida de personas con angioplastia. *Ensayo clínico aleatorizado. Rev Investig Innov Cienc Salud* [Internet]. 2023;5(1):91-102. doi: <https://doi.org/10.46634/riics.151>

14. Ralapanawa U, Sivakanesan R. Epidemiology and the Magnitude of Coronary Artery Disease and Acute Coronary Syndrome: A Narrative Review. *J Epidemiol Glob Health* [Internet]. 2021;11(2):169-77. doi: <https://doi.org/10.2991/jegh.k.201217.001>
15. Taylor C, Hobbs R. Diagnosing Heart Failure: Experience and 'Best Pathways.' *European Cardiology* [Internet]. 2010;6(3):10-2. doi: <https://doi.org/10.15420/ecr.2010.6.3.10>
16. Tongpeth J, Du H, Clark R. The Effectiveness of an Avatar-Based Education Application on Improving Knowledge and Response to Heart Attack Symptoms: A Pragmatic Randomised Trial in Patients with Acute Coronary Syndrome. *Heart Lung Circ* [Internet]. 2018;27(Suppl 2):S388. doi: <https://doi.org/10.1016/j.hlc.2018.06.776>
17. Xiao C, Guo Y, Zhao K, Liu S, He N, He Y, et al. Prognostic Value of Machine Learning in Patients with Acute Myocardial Infarction. *J Cardiovasc Dev Dis* [Internet]. 2022;9(2):56. doi: <https://doi.org/10.3390/jcdd9020056>
18. Tang P, Wang Q, Ouyang H, Yang S, Hua P. The feasibility of early detecting coronary artery disease using deep learning-based algorithm based on electrocardiography. *Aging (Albany NY)* [Internet]. 2023;15(9):3524-37. doi: <https://doi.org/10.18632/aging.204688>
19. Alhamaydeh M, Gregg R, Ahmad A, Faramand Z, Saba S, Al-Zaiti S. Identifying the most important ECG predictors of reduced ejection fraction in patients with suspected acute coronary syndrome. *J Electrocardiol* [Internet]. 2020;61:81-5. doi: <https://doi.org/10.1016/j.jelectrocard.2020.06.003>
20. Mohsen M, Shawky A. The diagnostic utility of High-Sensitivity Cardiac Troponin T in acute coronary syndrome. *Egypt Heart J* [Internet]. 2016;68(1):1-9. doi: <https://doi.org/10.1016/j.chj.2014.12.003>
21. Ouyang M, Tu D, Tong L, Sarwar M, Bhimaraj A, Li C, et al. A review of biosensor technologies for blood biomarkers toward monitoring cardiovascular diseases at the point-of-care. *Biosens Bioelectron* [Internet]. 2021;171:112621. doi: <https://doi.org/10.1016/j.bios.2020.112621>
22. Kirigaya J, Iwahashi N, Tahakashi H, Minamimoto Y, Gohbara M, Abe T, et al. Impact of Cardio-Ankle Vascular Index on Long-Term Outcome in Patients with Acute Coronary Syndrome. *J Atheroscler Thromb* [Internet]. 2020;27(7):657-68. doi: <https://doi.org/10.5551/jat.51409>
23. Kumar D, Saghir T, Zahid M, Ashok A, Kumar M, Shah AA, et al. Validity of TIMI score for predicting 14-day mortality of non-ST elevation myocardial infarction patients. *Cureus* [Internet]. 2021;13(1):e12518. doi: <https://doi.org/10.7759/cureus.12518>
24. Alhailiy AB, Brennan PC, McEntee MF, Kench PL, Ryan EA. Diagnostic reference levels in cardiac computed tomography angiography: A Systematic Review. *Radiat Prot Dosimetry* [Internet]. 2018;178(1):63-72. doi: <https://doi.org/10.1093/rpd/ncx075>
25. Kauw F, Velthuis BK, Takx RAP, Guglielmo M, Cramer MJ, van Ommen F, et al. Detection of Cardioembolic Sources With Nongated Cardiac Computed Tomography Angiography in Acute Stroke: Results From the ENCLOSE Study. *Stroke* [Internet]. 2023;54(3):821-30. doi: <https://doi.org/10.1161/STROKEAHA.122.041018>

26. Takagi H, Tanaka R, Nagata K, Ninomiya R, Arakita K, Schuijf JD, et al. Diagnostic performance of coronary CT angiography with ultra-high-resolution CT: Comparison with invasive coronary angiography. *Eur J Radiol* [Internet]. 2018;101:30-7. doi: <https://doi.org/10.1016/j.ejrad.2018.01.030>
27. Dobric M, Furtula M, Tesic M, Timcic S, Borzanovic D, Lazarevic N, et al. Current status and future perspectives of fractional flow reserve derived from invasive coronary angiography. *Front Cardiovasc Med* [Internet]. 2023;10:1181803. doi: <https://doi.org/10.3389/fcvm.2023.1181803>
28. Zhang Y, Wang Y, Shi J, Hua Z, Xu J. Cardiac output measurements via echocardiography versus thermodilution: A systematic review and meta-analysis. *PLOS ONE* [Internet]. 2019;14(10):e0222105. doi: <https://doi.org/10.1371/journal.pone.0222105>
29. Rorth R, Jhund PS, Yilmaz MB, Kristensen SL, Welsh P, Desai AS, et al. Comparison of BNP and NT-proBNP in Patients with Heart Failure and Reduced Ejection Fraction. *Circ Heart Fail* [Internet]. 2020;13(2):e006541. doi: <https://doi.org/10.1161/CIRCHEARTFAILURE.119.006541>
30. Bradley SH, Abraham S, Callister ME, Grice A, Hamilton WT, Lopez RR, et al. Sensitivity of chest X-ray for detecting lung cancer in people presenting with symptoms: a systematic review. *Br J Gen Pract* [Internet]. 2019;69(689):e827-35. doi: <https://doi.org/10.3399/bjgp19X706853>
31. Alghamdi HS, Amoudi G, Elhag S, Saeedi K, Nasser J. Deep Learning Approaches for Detecting COVID-19 From Chest X-Ray Images: A Survey. *IEEE Access* [Internet]. 2021;9:20235-54. doi: <https://doi.org/10.1109/ACCESS.2021.3054484>
32. Pan JA, Kramer CM. Cardiac Magnetic Resonance Imaging in Heart Failure. *CVIA* [Internet]. 2024;9(1). doi: <https://doi.org/10.15212/CVIA.2024.0037>
33. Tam CCF, Cheung KS, Lam S, Wong A, Yung A, Sze M, et al. Impact of Coronavirus Disease 2019 (COVID-19) Outbreak on ST-Segment-Elevation Myocardial Infarction Care in Hong Kong, China. *Circ Cardiovasc Qual Outcomes* [Internet]. 2020;13(4):e006631. doi: <https://doi.org/10.1161/CIRCOUTCOMES.120.006631>
34. Rajpurkar P, Joshi A, Pareek A, Chen P, Kiani A, Irvin J, et al. CheXpedition: Investigating Generalization Challenges for Translation of Chest X-Ray Algorithms to the Clinical Setting. *arXiv* [Internet]; 2020. doi: <https://doi.org/10.48550/arXiv.2002.11379>
35. Nechita LC, Nechita A, Voipan AE, Voipan D, Debita M, Fulga A, et al. AI-Enhanced ECG Applications in Cardiology: Comprehensive Insights from the Current Literature with a Focus on COVID-19 and Multiple Cardiovascular Conditions. *Diagnostics* [Internet]. 2024;14(17):1839. doi: <https://doi.org/10.3390/diagnostics14171839>
36. Olawade DB, Aderinto N, Olatunji G, Kokori E, David-Olawade AC, Hadi M. Advancements and applications of Artificial Intelligence in cardiology: Current trends and future prospects. *J Med Surg Public Health* [Internet]. 2024;3:100109. doi: <https://doi.org/10.1016/j.glmedi.2024.100109>

37. Stamate E, Piraianu AI, Ciobotaru OR, Crassas R, Duca O, Fulga A, et al. Revolutionizing Cardiology through Artificial Intelligence—Big Data from Proactive Prevention to Precise Diagnostics and Cutting-Edge Treatment—A Comprehensive Review of the Past 5 Years. *Diagnostics* [Internet]. 2024;14(11):1103. doi: <https://doi.org/10.3390/diagnostics14111103>
38. Wang H, Zu Q, Chen J, Yang Z, Ahmed MA. Application of Artificial Intelligence in Acute Coronary Syndrome: A Brief Literature Review. *Adv Ther* [Internet]. 2021;38(10):5078-86. doi: <https://doi.org/10.1007/s12325-021-01908-2>
39. Seetharam K, Balla S, Bianco C, Cheung J, Pachulski R, Asti D, et al. Applications of Machine Learning in Cardiology. *Cardiol Ther* [Internet]. 2022;11(3):355-68. doi: <https://doi.org/10.1007/s40119-022-00273-7>
40. Chen C, Qin C, Qiu H, Tarroni G, Duan J, Bai W, et al. Deep Learning for Cardiac Image Segmentation: A Review. *Front Cardiovasc Med* [Internet]. 2020;7:25. doi: <https://doi.org/10.3389/fcvm.2020.00025>
41. Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms* [Internet]. 2023;16(2):88. doi: <https://doi.org/10.3390/a16020088>
42. Wegner FK, Plagwitz L, Doldi F, Ellermann C, Willy K, Wolfes J, et al. Machine learning in the detection and management of atrial fibrillation. *Clin Res Cardiol* [Internet]. 2022;111(9):1010-7. doi: <https://doi.org/10.1007/s00392-022-02012-3>
43. Sengan S, Mehbodniya A, Bhatia S, Saranya SS, Alharbi M, Basheer S, et al. Echocardiographic Image Segmentation for Diagnosing Fetal Cardiac Rhabdomyoma During Pregnancy Using Deep Learning. *IEEE Access* [Internet]. 2022;10:114077-91. doi: <https://doi.org/10.1109/ACCESS.2022.3215973>
44. Wu Y, Wang J, Wang R, Wang J, Zhou K, Zhang X. Research on Cardiac Medical Image Segmentation Based on an Improved U-Net Model. In: *Proceedings of 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)* [Internet]. 2024 November 22-24; Shenzhen, China: IEEE; 2024. p. 206-12. doi: <https://doi.org/10.1109/ICICML63543.2024.10957994>
45. Ma C, Lei M. Time series prediction for monitoring cardiovascular health in autistic patients. *Front Psychiatry* [Internet]. 2025;16:1623986. doi: <https://doi.org/10.3389/fpsyt.2025.1623986>
46. Amal S, Safarnejad L, Omiye JA, Ghanzouri I, Cabot JH, Ross EG. Use of Multi-Modal Data and Machine Learning to Improve Cardiovascular Disease Care. *Front Cardiovasc Med* [Internet]. 2022;9:840262. doi: <https://doi.org/10.3389/fcvm.2022.840262>
47. Moshawrab M, Adda M, Bouzouane A, Ibrahim H, Raad A. Reviewing Multimodal Machine Learning and Its Use in Cardiovascular Diseases Detection. *Electronics* [Internet]. 2023;12(7):1558. doi: <https://doi.org/10.3390/electronics12071558>
48. Moffat LM, Xu D. Accuracy of Machine Learning Models to Predict In-hospital Cardiac Arrest: A Systematic Review. *Clinical Nurse Specialist* [Internet]. 2022;36(1):29-44. doi: <https://doi.org/10.1097/NUR.0000000000000644>

49. Li B, Feridooni T, Cuen-Ojeda C, Kishibe T, de Mestral C, Mamdani M, et al. Machine learning in vascular surgery: a systematic review and critical appraisal. *npj Digit Med* [Internet]. 2022;5(1):7. doi: <https://doi.org/10.1038/s41746-021-00552-y>
50. Makimoto H, Kohro T. Adopting artificial intelligence in cardiovascular medicine: a scoping review. *Hypertens Res* [Internet]. 2024;47(3):685-99. doi: <https://doi.org/10.1038/s41440-023-01469-7>
51. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* [Internet]. 2021;372:n71. doi: <https://doi.org/10.1136/bmj.n71>
52. Sarkis-Onofre R, Catala-Lopez F, Aromataris E, Lockwood C. How to properly use the PRISMA Statement. *Syst Rev* [Internet]. 2021;10:117. doi: <https://doi.org/10.1186/s13643-021-01671-z>
53. Saputra D, Suhartini S, Sujianto U, Ismail R. A viewpoint on evaluating ECG findings to assess heart issues: A clinical nursing review. *Journal of Holistic Nursing Science* [Internet]. 2023;10(2):89-95. doi: <https://doi.org/10.31603/nursing.v0i0.8374>
54. Barker J, Li X, Khavandi S, Koeckerling D, Mavilakandy A, Pepper C, et al. Machine learning in sudden cardiac death risk prediction: a systematic review. *Europace* [Internet]. 2022;24(11):1777-87. doi: <https://doi.org/10.1093/europace/euac135>
55. Moons KGM, Damen JAA, Kaul T, Hooft L, Navarro CA, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ* [Internet]. 2025;388:e082505. doi: <https://doi.org/10.1136/bmj-2024-082505>
56. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* [Internet]. 2017;356:i6460. doi: <https://doi.org/10.1136/bmj.i6460>
57. Jones HE, Gatsonsis CA, Trikalinos TA, Welton NJ, Ades AE. Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. *Stat Med* [Internet]. 2019;38(24):4789-803. doi: <https://doi.org/10.1002/sim.8301>
58. Huang M, Li J, Zhao X, Fu R, Li X, Jiang W. Global and regional prevalence and cardiovascular risk of primary aldosteronism: a systematic review and meta-analysis. *Curr Probl Cardiol* [Internet]. 2024;49(10):102791. doi: <https://doi.org/10.1016/j.cpcardiol.2024.102791>
59. Rozanski A, Bavishi C, Kubzansky LD, Cohen R. Association of Optimism with Cardiovascular Events and All-Cause Mortality: A Systematic Review and Meta-analysis. *JAMA Netw Open* [Internet]. 2019;2(9):e1912200. doi: <https://doi.org/10.1001/jamanetworkopen.2019.12200>
60. Zhao L, Tian Z, Fang Q. Diagnostic accuracy of cardiovascular magnetic resonance for patients with suspected cardiac amyloidosis: A Systematic Review and Meta-analysis. *BMC Cardiovasc Disord* [Internet]. 2016;16:129. doi: <https://doi.org/10.1186/s12872-016-0311-6>

61. Rao G, Lopez-Jimenez F, Boyd J, D'Amico F, Durant NH, Hlatky MA, et al. Methodological Standards for Meta-Analyses and Qualitative Systematic Reviews of Cardiac Prevention and Treatment Studies: A Scientific Statement from the American Heart Association. *Circulation* [Internet]. 2017;136(10):e172-94. doi: <https://doi.org/10.1161/CIR.0000000000000523>
62. Stankovic I, Putnikovic B, Cvjetan R, Milicevic P, Panic M, Kalezic-Radmili T, et al. Visual assessment vs. strain imaging for the detection of critical stenosis of the left anterior descending coronary artery in patients without a history of myocardial infarction. *Eur Heart J Cardiovasc Imaging* [Internet]. 2014;16(4):402-9. doi: <https://doi.org/10.1093/ehjci/jeu206>
63. Cella L, Oh JH, Deasy JO, Palma G, Liuzzi R, D'avino V, et al. Predicting radiation-induced valvular heart damage. *Acta Oncol* [Internet]. 2015;54(10):1796-804. doi: <https://doi.org/10.3109/0284186X.2015.1016624>
64. Kim J, Lee J, Lee Y. Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree. *Healthc Inform Res* [Internet]. 2015;21(3):167-74. doi: <https://doi.org/10.4258/hir.2015.21.3.167>
65. Mulayim N, Alaybeyoglu A. Designing of an expert system based on firefly algorithm for diagnosis of Heart Disease. In: 2016 20th National Biomedical Engineering Meeting (BIYOMUT). Conference Record # 39623 [Internet]. 2016 November 03-05; Izmir, Turkey; 2016. p. 1-4. doi: <https://doi.org/10.1109/BIYOMUT.2016.7849405>
66. Orphanou K, Dagliati A, Sacchi L, Stassopoulou A, Keravnou E, Bellazzi R. Combining Naive Bayes Classifiers with Temporal Association Rules for Coronary Heart Disease Diagnosis. In: Proceedings. 2016 IEEE International Conference on Healthcare Informatics (ICHI) [Internet]. 2016 October 04-07; Chicago, IL, USA; 2016. p. 81-92. doi: <https://doi.org/10.1109/ICHI.2016.15>
67. Ribeiro G, Oliveira A, Ferreira A, Visweswaran S, Cooper G. Patient-Specific Modeling Algorithm for Medical Data Based on AUC. *International Journal of Computer and Information Engineering* [Internet]. 2016;10(1):105-10. doi: <https://doi.org/10.5281/zenodo.1338644>
68. Cheng CA, Chiu HW. An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [Internet]. 2017 July 11-15; Jeju, Korea (South); 2017. p. 2566-9. doi: <https://doi.org/10.1109/EMBC.2017.8037381>
69. Dekamin A, Sheibatolhamdi A. A Data Mining Approach for Coronary Artery Disease Prediction in Iran. *Journal of Advanced Medical Sciences and Applied Technologies* [Internet]. 2017;3(1):29-38. doi: <https://doi.org/10.18869/nrip.jamsat.3.1.29>
70. Kumar M, Pachori RB, Acharya UR. Characterization of coronary artery disease using flexible analytic wavelet transform applied on ECG signals. *Biomedical Signal Processing and Control* [Internet]. 2017;31:301-8. doi: <https://doi.org/10.1016/j.bspc.2016.08.018>

71. Mansoor H, Elgendy IY, Segal R, Bavry AA, Bian J. Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: A machine learning approach. *Heart & Lung* [Internet]. 2017;46(6):405-11. doi: <https://doi.org/10.1016/j.hrtlng.2017.09.003>
72. Prakash A, Zhao S, Hasan S, Datla V, Lee K, Qadir A, et al. Condensed Memory Networks for Clinical Diagnostic Inferencing. *Proceedings of the AAAI Conference on Artificial Intelligence* [Internet]. 2017;31(1). doi: <https://doi.org/10.1609/aaai.v31i1.10964>
73. Wiharto W, Kusnanto H, Herianto H. Hybrid System of Tiered Multivariate Analysis and Artificial Neural Network for Coronary Heart Disease Diagnosis. *International Journal of Electrical and Computer Engineering (IJECE)* [Internet]. 2017;7(2):1023-31. doi: <https://doi.org/10.11591/ijece.v7i2.pp1023-1031>
74. Banerjee R, Bhattacharya S, Alam S. Time Series and Morphological Feature Extraction for Classifying Coronary Artery Disease from Photoplethysmogram. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. *Proceedings* [Internet]. 2018 April 15-20; Calgary, AB, Canada; 2018. p. 950-4. doi: <https://doi.org/10.1109/ICASSP.2018.8462604>
75. Acharya UR, Fujita H, Oh SL, Hagiwara Y, Tan JH, Adam M, et al. Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. *Appl Intell* [Internet]. 2019;49(1):16-27. doi: <https://doi.org/10.1007/s10489-018-1179-1>
76. Panicacci S, Donati M, Fanucci L, Bellini I, Profili F, Francesconi P. Exploring Machine Learning Algorithms to Identify Heart Failure Patients: the Tuscany Region Case Study. In: *Proceedings. 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* [Internet]. 2019. p. 417-22. doi: <https://doi.org/10.1109/CBMS.2019.00088>
77. Zhang Y, Zhao Y, Xu Y. Diagnosis of ST-elevated myocardial infarction by Artificial Intelligence with 12-lead Electrocardiograph. The Alert-Pilot Study. *Journal of Hypertension* [Internet]. 2019;37:e234. doi: <https://doi.org/10.1097/01.hjh.0000573000.85159.94>
78. Banerjee R, Ghose A, Muthana Mandana K. A Hybrid CNN-LSTM Architecture for Detection of Coronary Artery Disease from ECG. In: 2020 International Joint Conference on Neural Networks (IJCNN) [Internet]. 2020 July 19-24; Glasgow, United Kingdom: IEEE; 2020. p. 1-8. doi: <https://doi.org/10.1109/IJCNN48605.2020.9207044>
79. Rjoob K, McGilligan V, Bond R, Watterson S, Chemaly M, McAlister R, et al. Improving the Detection of Acute Coronary Syndrome Using Machine Learning of Blood Biomarkers. *Computing in Cardiology* [Internet]. 2020;47:1-4. doi: <https://doi.org/10.22489/CinC.2020.337>
80. Zhang Y, Xia M. Application of Deep Neural Network for Congestive Heart Failure Detection Using ECG Signals. *J Phys Conf Ser* [Internet]. 2020;1642:012021. doi: <https://doi.org/10.1088/1742-6596/1642/1/012021>

81. Zhang Y, Xia M. An Effective Method for CHF Diagnosis via Attention-based RNN Using ECG Signals. In: The 5th International Conference on Communication, Image and Signal Processing (CCISP) 2020 [Internet]. 2020 November 13-15; Chengdu, China; 2020. p. 209-13. doi: <https://doi.org/10.1109/CCISP51026.2020.9273489>
82. Zhang X, Gu K, Miao S, Zhang X, Yin Y, Wan C, et al. Automated detection of cardiovascular disease by electrocardiogram signal analysis: a deep learning system. *Cardiovasc Diagn Ther* [Internet]. 2020;10(2):227-35. doi: <https://doi.org/10.21037/cdt.2019.12.10>
83. Cherradi B, Terrada O, Ouhmida A, Hamida S, Raihani A, Bouattane O. Computer-Aided Diagnosis System for Early Prediction of Atherosclerosis using Machine Learning and K-fold cross-validation. In: 2021 International Congress of Advanced Technology and Engineering (ICOTEN). Virtual Conference [Internet]. 2021 July 4-5; Taiz, Yemen; 2021. p. 1-9. doi: <https://doi.org/10.1109/ICOTEN52080.2021.9493524>
84. Li D, Xiong G, Zeng H, Zhou Q, Jiang J, Guo X. Machine learning-aided risk stratification system for the prediction of coronary artery disease. *Int J Cardiol* [Internet]. 2021;326:30-4. doi: <https://doi.org/10.1016/j.ijcard.2020.09.070>
85. Muhammad LJ, Algehyne EA. Fuzzy based expert system for diagnosis of coronary artery disease in nigeria. *Health Technol* [Internet]. 2021;11(2):319-29. doi: <https://doi.org/10.1007/s12553-021-00531-z>
86. Zahiri N, Asgari R, Razavi-Ratki SK, parach AA. Deep Learning Analysis of Polar Maps from SPECT Myocardial Perfusion Imaging for Prediction of Coronary Artery Disease. *Research Square*; 2021. doi: <https://doi.org/10.21203/rs.3.rs-1153347/v1>
87. Chen Y. NIHA-HF: An Artificial Intelligence Solution for Heart Failure Diagnosis and Monitoring Based on Lead-I Electrocardiogram. *JACC Basic Trans Sci* [Internet]. 2024;9(4):454-5. doi: <https://doi.org/10.1016/j.jacbts.2024.02.011>
88. Karabayir I, Davis R, Chinthala L, Butler L, Celik T, Kilic U, et al. Single lead wearable ECG simulation augmented with AI for heart failure identification. *JACC* [Internet]. 2024;83(Suppl 13):2393. doi: [https://doi.org/10.1016/S0735-1097\(24\)04383-3](https://doi.org/10.1016/S0735-1097(24)04383-3)
89. Li C, Wang Y, Meng L, Zhong W, Zhang C, Liu T. Integrating EPSOSA-BP neural network algorithm for enhanced accuracy and robustness in optimizing coronary artery disease prediction. *Sci Rep* [Internet]. 2024;14:30993. doi: <https://doi.org/10.1038/s41598-024-82184-2>
90. Shao H, Ling Y, Pang J, Zhao Q, Qian Y. Identification of Atrial Septal Defects in Congenital Heart Disease Patients Using Deep Learning-based Chest Radiographs. In: 2024 IEEE 5th International Conference on Pattern Recognition and Machine Learning (PRML) [Internet]. 2024 July 19-21; Chongqing, China; 2024. p. 386-91. doi: <https://doi.org/10.1109/PRML62565.2024.10779637>
91. Than MP, Pickering JW, Mair J, Mills NL, Study Group on Biomarkers of the Association for Acute CardioVascular Care of the ESC. Clinical decision support using machine learning and cardiac troponin for the diagnosis of myocardial infarction. *Eur Heart J Acute Cardiovasc Care* [Internet]. 2024;13(8):634-6. doi: <https://doi.org/10.1093/ehjacc/zuac085>

92. Ye Z, Lin Z, Xie E, Song C, Zhang R, Wang HY, et al. Prediction of Percutaneous Coronary Intervention Success in Patients with Moderate to Severe Coronary Artery Calcification Using Machine Learning Based on Coronary Angiography: Prospective Cohort Study. *J Med Internet Res* [Internet]. 2025;27:e70943. doi: <https://doi.org/10.2196/70943>
93. Andersson P, Lindow T, Lindqvist P, Venkateshvaran A. Utilizing echocardiographic findings and machine learning to predict elevated left ventricular filling pressures in patients with preserved ejection fraction. *European Heart Journal Cardiovascular Imaging* ([Internet]. 2025;26(Suppl 1):jeac333.455. doi: <https://doi.org/10.1093/ehjci/jeac333.455>
94. Dou G, Zhou J, Guo Z, Shan D, Wang X, Li T, et al. Developing an ML-Based Pretest Probability Model of Obstructive CAD in Patients With Stable Chest Pain. *JACC: Asia* [Internet]. 2025 Jul;5(7):865-80. doi: <https://doi.org/10.1016/j.jacasi.2025.03.015>
95. Shin S, Lee S, Kim K, Lee H, Song G. SNP Position Filtering and Multimodal Approach in Heart Failure Risk Prediction. In: *Proceedings. 2025 IEEE International Conference on Big Data and Smart Computing. BigComp 2025* [Internet]. 2025 February 09-12; Kota Kinabalu, Malaysia; 2025. p. 337-40. doi: <https://doi.org/10.1109/BigComp64353.2025.00067>
96. Zhang G, Ding Y, Ma T, Zhou S, Wang B, Chang X, et al. Predicting the Degree of Coronary Artery Stenosis Through Retinal Vascular Characteristics and Minimal Clinical Information. *Int J Gen Med* [Internet]. 2025;18:585-91. doi: <https://doi.org/10.2147/IJGM.S507016>
97. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep* [Internet]. 2020;10(1):16057. doi: <https://doi.org/10.1038/s41598-020-72685-1>
98. Liu T, Krentz A, Lu L, Curcin V. Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis. *Eur Heart J Digit Health* [Internet]. 2025;6(1):7-22. doi: <https://doi.org/10.1093/ehjdh/ztac080>
99. McGranaghan P, Schoeppenthau D, Popp A, Saxena A, Kothakapu S, Rubens M, et al. Machine Learning Application in Different Imaging Modalities for Detection of Obstructive Coronary Artery Disease and Outcome Prediction: A Systematic Review and Meta-Analysis. *Hearts* [Internet]. 2025;6(3):21. doi: <https://doi.org/10.3390/hearts6030021>
100. Baashar Y, Alkaws G, Alhussian H, Capretz LF, Alwadain A, Alkahtani AA, et al. Effectiveness of Artificial Intelligence Models for Cardiovascular Disease Prediction: Network Meta-Analysis. *Computational Intelligence and Neuroscience* [Internet]. 2022;2022:5849995. doi: <https://doi.org/10.1155/2022/5849995>
101. Triantafyllidis A, Kondylakis H, Katehakis D, Kouroubali A, Koumakis L, Marias K, et al. Deep Learning in mHealth for Cardiovascular Disease, Diabetes, and Cancer: Systematic Review. *JMIR mHealth and uHealth* [Internet]. 2022 Apr 4;10(4):e32344. doi: <https://doi.org/10.2196/32344>

102. Mallikarjunamallu K, Syed K. A Review on Heart Diseases Using Machine Learning and Deep Learning Techniques. In: Pant M, Deep K, Nagar A, editors. Proceedings of the 12th International Conference on Soft Computing for Problem Solving [Internet]. Singapore: Springer Nature; 2024. p. 651-79. doi: https://doi.org/10.1007/978-981-97-3292-0_45
103. Nadarajah R, Younsi T, Romer E, Raveendra K, Nakao YM, Nakao K, et al. Prediction models for heart failure in the community: A systematic review and meta-analysis. Eur J Heart Fail [Internet]. 2023;25(10):1724-38. doi: <https://doi.org/10.1002/ehf.2970>
104. Al Hinai G, Jammoul S, Vajihi Z, Afilalo J. Deep learning analysis of resting electrocardiograms for the detection of myocardial dysfunction, hypertrophy, and ischaemia: a systematic review. Eur Heart J Digit Health [Internet]. 2021;2(3):416-23. doi: <https://doi.org/10.1093/ehjdh/ztab048>