BioMedInformatics

MDPI

Systematic Review

From Explainability to Clinical Actionability in Multi-Modal AI for Cardiovascular Prediction: A Systematic Review

Hamza Nouri *, Rafae Abderrahim and Mohamed Erritali

Data4earth Laboratory Morocco, Sultan Moulay Slimane University, Beni Mellal 23000, Morocco; abderrahim.rafae@usms.ac.ma (R.A.); m.erritali@usms.ma (M.E.)

* Correspondence: nouri.hamza@usms.ac.ma

Abstract

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, driving the need for reliable tools for early risk prediction. Artificial intelligence (AI) applied to electrocardiograms (ECGs) has shown strong predictive performance for future cardiac events, yet its clinical adoption remains limited by the lack of transparency and trust associated with black-box models. This systematic review examines recent advances in AI-based cardiovascular prediction, focusing on the combined challenges of multi-modal data fusion and clinically actionable explainability. Following PRISMA 2020 guidelines, we analyzed 65 peer-reviewed studies published between 2018 and 2025, identified through a systematic search of PubMed, IEEE Xplore, Web of Science, Scopus, ACM Digital Library, and Google Scholar. The reviewed literature reveals that while most AI-ECG models achieve high predictive accuracy, typically AUC 0.85–0.95, the majority rely on post hoc explainability techniques that offer limited clinical insight, and 61.5% of included studies implement no explainability method at all. External validation remains critically underutilized, performed by only 12.3% of studies, and multi-modal approaches integrating ECG data with electronic health records, biomarkers, or genomics represent only 27.7% of the reviewed literature. While these multi-modal models demonstrate improved contextualization and predictive performance, they remain insufficiently validated and inconsistently interpretable. Among studies employing XAI techniques, attention mechanisms were the most prevalent approach (28% of XAI studies), followed by saliency maps (20%), SHAP (16%), and LIME (8%). Only 9.2% of studies were prospective or clinical trials, underscoring the gap between algorithmic development and real-world clinical deployment. Applying a pre-specified four-level clinical actionability scoring framework (Level 0–3), we found that the majority of studies (61.5%) scored at Level 0 (no actionability), with only 9.2% reaching Level 3 (demonstrated clinical impact), confirming that the clinical translation gap extends beyond trial design to encompass the broader absence of clinically contextualised evaluation of AI-ECG systems. This review highlights a persistent and critical gap between predictive performance and clinical usability, and outlines four key directions for developing AI-ECG systems that can better support trustworthy clinical decision-making: (1) developing inherently interpretable architectures, (2) advancing unified multi-modal fusion and explanation frameworks, (3) establishing standardized benchmarks for explainability evaluation, and (4) conducting robust prospective validation measuring real-world patient outcomes.

Keywords: electrocardiogram (ECG); artificial intelligence; explainable AI; multi-modal data fusion; cardiovascular risk prediction; clinical decision support; systematic review

Academic Editors: Kwang-Sig Lee and Hyuntae Park

Received: 24 May 2026
Revised: 10 July 2026
Accepted: 20 July 2026
Published: 3 August 2026

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality globally, accounting for nearly 18 million deaths annually and creating an urgent need for tools that enable early and accurate prediction to facilitate preventive care [1]. The standard 12-lead electrocardiogram (ECG), a cornerstone of cardiac diagnostics for over a century, provides a wealth of information about cardiac function. However, its interpretation relies heavily on human expertise, which is subject to inter-observer variability and often misses subtle, sub-clinical signals that precede major cardiac events. This limitation underscores the need for innovative approaches to maximize the diagnostic and predictive potential of ECGs.

Artificial Intelligence (AI), particularly deep learning (DL), has emerged as a transformative solution to these challenges. AI models excel at analyzing vast ECG datasets to identify complex patterns invisible to the human eye, demonstrating high accuracy in diagnosing existing conditions and predicting future events. For instance, AI-ECG systems have been shown to predict atrial fibrillation (AF) in patients with normal sinus rhythm [2] and identify low ejection fraction in primary care settings, enabling earlier intervention [3]. These advancements position AI-ECG as a powerful, low-cost screening tool with the potential to revolutionize cardiovascular care.

These developments sit within a broader trajectory of AI-enhanced electrocardiography in cardiovascular disease management [4], and reflect a wider vision of precision cardiovascular medicine enabled by large-scale data and artificial intelligence [5]. This trajectory is not unique to ECG: multimodal artificial intelligence has increasingly been applied across medical diagnostics more broadly, with fusion strategies, architectures, and clinical applications reviewed comprehensively by [6].

Despite these advances, a fundamental limitation persists: most AI-ECG systems remain weakly integrated into clinical decision-making processes. While predictive accuracy continues to improve, current models often fail to provide explanations that clinicians can meaningfully interpret, trust, and act upon in real-world cardiovascular care settings. The widespread clinical adoption of AI-ECG models is thus hindered by their critical "black box" nature. Clinicians are hesitant to base high-stakes treatment decisions on predictions they cannot understand or trust. This gap between predictive excellence and interpretability directly limits regulatory approval and physician adoption [7,8]. As recent studies continue to focus on direct-to-physician reporting and ambulatory monitoring [9], the need for transparency remains paramount.

Explainable AI (XAI) techniques, such as saliency maps, have been introduced to address this issue [10], but they often fall short of providing clinically meaningful explanations, showing where the model is looking but failing to explain why in terms that align with pathophysiological reasoning. In this review, clinical actionability refers to the capacity of model explanations to support clinical reasoning, inform diagnostic or therapeutic decisions, and align with established cardiovascular workflows, rather than merely providing visual or statistical interpretability.

To truly advance predictive power and clinical trust, the next frontier lies in integrating ECG data with other patient information, such as electronic health records (EHRs), biomarkers, and genomics—a concept known as multi-modal data fusion [11]. While multi-modal approaches may improve contextualization and robustness, they also increase model complexity and can further exacerbate opacity unless explainability is addressed jointly with fusion.

Although several recent reviews have examined AI-based ECG analysis from perspectives such as predictive performance, data augmentation, or general explainability, none has systematically analyzed how multi-modal data integration and explainability jointly contribute to clinically actionable cardiovascular prediction. Consequently, the relationship

between technical interpretability, clinical trust, and real-world decision support remains insufficiently structured in the existing literature.

This systematic review focuses on studies published between 2018 and 2025, a period chosen to encapsulate the "deep learning revolution" in electrocardiography, beginning with landmark studies that demonstrated AI's ability to learn from raw ECG signals. This review uniquely bridges the dual frontiers of multi-modal integration and clinically actionable explainability, aiming to transform AI-ECG models from research tools into trusted clinical partners.

To guide this investigation, we examine four key areas: (1) the deep learning architectures and explainability techniques predominantly employed for cardiovascular prediction and their adaptation for multi-modal data; (2) performance trade-offs between predictive accuracy, computational cost, and clinical interpretability; (3) the extent to which existing models address the need for clinical explainability and the limitations in providing actionable insights; and (4) the primary research gaps and emerging trends critical for enhancing real-world clinical applicability.

The remainder of this paper is organized as follows: Section 2 describes the methodology adopted for this review. Section 3 analyzes prior research in AI-ECG prediction and explainability. Section 4 presents our evaluation and key findings. Section 5 discusses the implications of our analysis, and finally, Section 6 concludes the paper and outlines directions for future research.

2. Methodology

This systematic review was conducted in accordance with the PRISMA 2020 guidelines (PRISMA Checklist can be found in the Supplementary Material) [12]. The review protocol was not registered in a public database prior to commencement due to resource constraints and the exploratory nature of the initial scoping phase; however, the search strategy, eligibility criteria, and data extraction template were all predefined before screening began, and the data extraction process adhered to the predefined template throughout, constituting the primary safeguard against post hoc analytical flexibility. We acknowledge this as a limitation and note that the predefined extraction template, dual screening procedure (detailed in Section 2.1), and transparent methodology reporting represent the compensating measures taken in the absence of formal registration (see Section 5.2).

The methodology for this systematic review was meticulously designed to ensure a comprehensive, unbiased, and reproducible analysis of the current research landscape at the intersection of AI-ECG prediction, multi-modal data fusion, and clinical explainability. This section details the protocol registration, the search and study selection strategy, the data extraction process, and the synthesis methods, ensuring the process is transparent and follows established academic standards.

2.1. Search and Study Selection Strategy

To identify high-quality, peer-reviewed articles, a systematic search was conducted across several major academic databases, including PubMed, IEEE Xplore, Web of Science, Scopus, ACM Digital Library, and Google Scholar. These databases were chosen for their extensive coverage of the literature in medicine, computer science, and engineering, a practice consistent with other systematic reviews in the AI and health domain [13]. The search was limited to articles published between January 2018 and August 2025.

A precise, multi-component search string was developed to capture the relevant literature. The query was structured around four core concepts: the patient signal (ECG), the technology (AI), the task (prediction), and the specific research focus (explainability and multi-modality). The primary search string used was as follows:

("Electrocardiogram" OR "ECG") AND ("Artificial Intelligence" OR "AI" OR "Deep Learning" OR "Machine Learning") AND ("Prediction" OR "Risk Stratification" OR "Prognosis" OR "Heart Failure" OR "Atrial Fibrillation" OR "Mortality") AND ("Explainability" OR "XAI" OR "Interpretability" OR "Multi-modal" OR "Data Fusion" OR "EHR")

The study selection process strictly adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [12], ensuring transparency and methodological rigor. To maintain scientific reliability, a stringent set of inclusion and exclusion criteria was applied.

2.2. Inclusion Criteria

- Peer-reviewed. journal articles or full conference papers published in English.
- Studies published between January 2018 and August 2025.
- Studies that proposed or evaluated an AI model whose primary predicted outcome was a cardiovascular event, condition, or cardiac risk state—including, but not limited to, atrial fibrillation, heart failure, coronary artery disease, sudden cardiac death, low ejection fraction, arrhythmia, hypertension, long QT syndrome, and AV block—using ECG data as a primary or contributing input signal. Studies in which ECG served as a signal modality for a primarily non-cardiovascular prediction task (e.g., seizure detection, stress monitoring, cognitive load estimation, user authentication, sleep staging, or diabetes prediction) were not eligible under this criterion. Such studies are referenced in Section 3 (Background) only where their XAI or multi-modal methodology is technically informative for the cardiovascular prediction domain. Approximately 18 studies with non-cardiovascular primary outcomes were retained in the full dataset solely for their methodological contributions to multi-modal fusion and XAI techniques; they are excluded from all cardiovascular-specific prevalence claims and are clearly flagged in the sensitivity analysis (Section 4).
- Studies that explicitly addressed or implemented techniques for either model explainability or the fusion of ECG data with at least one other data modality (e.g., EHR, imaging, or genomics).

2.3. Exclusion Criteria

- Review articles, editorials, abstracts, pre-prints, or book chapters.
- Studies focusing solely on the detection of concurrent arrhythmias without a predictive component.
- Studies that used AI on signals other than the ECG (e.g., EEG or EMG).
- Studies were explicitly excluded if they lacked an explainability component, were not predictive in nature, or had insufficient performance data reported to allow for meaningful analysis. For example, a paper describing a novel architecture without reporting metrics like AUC or accuracy was excluded.

These exclusion criteria were applied to ensure that the review remained focused on AI-based models addressing prospective cardiovascular prediction rather than retrospective diagnosis, and on studies providing sufficient methodological detail to enable meaningful comparison of predictive performance, explainability, and clinical applicability.

The screening process was conducted in two stages. At the title and abstract screening stage, two authors (H.N. and R.A.) independently assessed all 1250 retrieved records against the eligibility criteria. At the full-text assessment stage, the same two reviewers independently assessed all 225 articles selected for full-text review. The overall study selection process, from initial record identification through to the final 65 included studies, is summarized in the PRISMA flow diagram (Figure 1). Disagreements at both stages

were resolved through discussion between the two reviewers, with the third author (M.E.) adjudicating any cases that could not be resolved by consensus. Inter-rater agreement at the full-text screening stage was Cohen's $\kappa = 0.84$, indicating strong agreement. Conference papers were assessed against the same eligibility criteria as journal articles and were required to be full peer-reviewed proceedings papers; short or extended abstracts were excluded regardless of conference prestige.

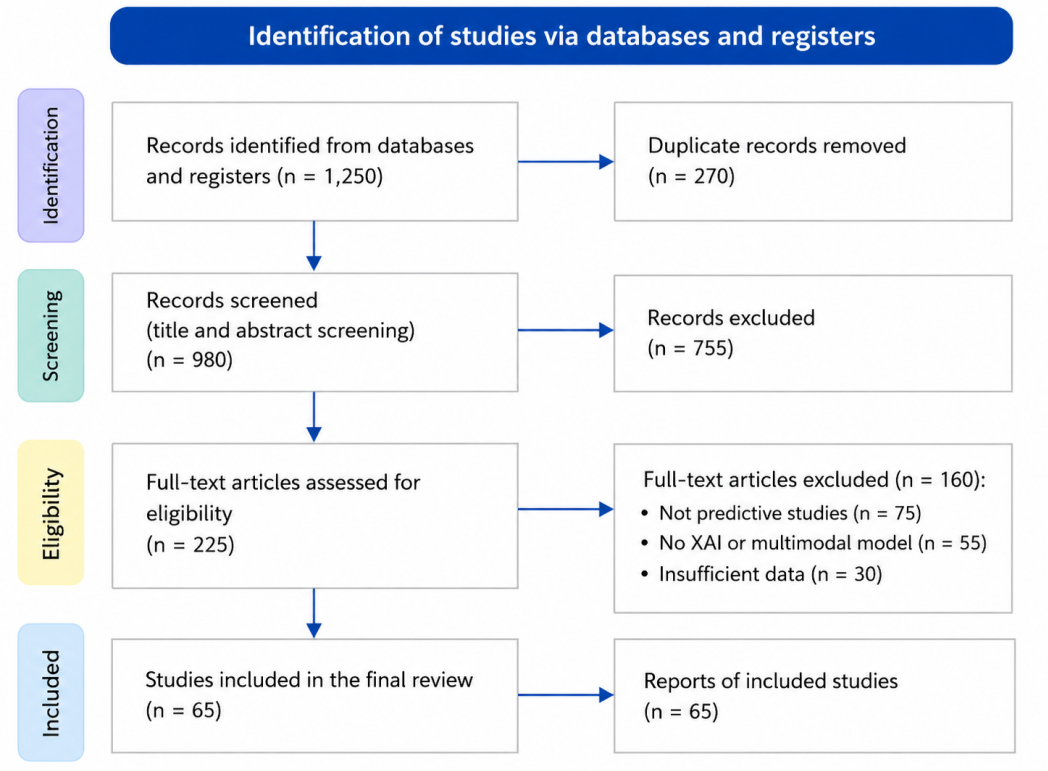

Figure 1. The PRISMA flow diagram illustrates the systematic filtering process applied in this review. The process began with 1250 articles identified from multiple databases. After removing duplicates and screening titles and abstracts, 225 articles were selected for full-text review. Of these, 160 were excluded for not meeting the specific inclusion criteria, culminating in a final set of 65 studies for qualitative synthesis. This rigorous process ensures the relevance and quality of the included studies.

2.4. Data Extraction and Synthesis

Data from the 65 included studies were systematically extracted using a predefined template (see Appendix A Table A1) to ensure consistency and accuracy. This structured approach facilitates a coherent comparison of methodologies and outcomes across diverse studies [14]. The synthesis of findings was structured around five critical dimensions, chosen to holistically evaluate not just the technical performance but also the clinical readiness and trustworthiness of the models. These dimensions were as follows:

- **Accuracy:** The model's reported ability to correctly predict the target cardiovascular outcome. This is typically measured by metrics such as Area Under the Receiver Operating Characteristic Curve (AUC), sensitivity, specificity, and F1-score. For example, a study might report an AUC of 0.95 for predicting incident AF [2] or an overall classification accuracy of 98.8% for distinguishing normal from anomalous signals [15]. For the purpose of descriptive synthesis, reported AUC values were classified into three performance bands: high (AUC $\geq$ 0.90), moderate (AUC 0.85–0.90), and lower (AUC < 0.85). These bands were defined to facilitate consistent characterization of performance trends across the included studies and are used throughout Section 4.1.

They were applied as descriptive categories only and did not constitute a quality threshold for study inclusion or exclusion: studies reporting AUC values in any band remained eligible for synthesis. In addition to AUC, extraction recorded whether each study reported precision, recall, and/or F1-score alongside AUC, given the well-documented limitations of AUC as a sole performance metric in the context of the imbalanced class distributions typical of cardiovascular event prediction tasks.

- **Explainability:** The degree to which the model's decision-making process is transparent and interpretable to a human user, particularly a clinician. This ranges from basic methods like saliency maps that highlight influential ECG regions, to more advanced techniques like SHAP or LIME that attribute importance to specific features. For instance, a study might compare the utility of GradCAM, SHAP, and LIME for uncovering which ECG features are most critical for a given classification [16].
- **Robustness and Generalizability:** The model's ability to maintain performance when applied to new, unseen data from different populations, healthcare systems, or ECG acquisition devices. This is often assessed through external validation. A key aspect is whether a model trained on one population can generalize to others, as the need to "collect diverse ECG data across different populations" is critical for improving AI efficacy [17].
- **Clinical Actionability:** The practical utility of the model's output in a real-world clinical setting. This dimension assesses whether a prediction and its explanation can directly inform a clinical decision, alter a clinical workflow, or improve patient outcomes. To move beyond a purely narrative assessment of this dimension, a four-level ordinal clinical actionability scoring framework was developed and applied descriptively to each of the 65 included studies (Section 2.4 and Appendix A Table A1). This framework distinguishes between studies that offer only a prediction with no explanation or validation context (Level 0), studies that implement a technical explainability method without clinical evaluation (Level 1), studies that explicitly align their explanations with clinical or pathophysiological reasoning (Level 2), and studies that provide prospective or interventional evidence of measurable impact on clinical decisions or patient outcomes (Level 3). The full framework definition and application criteria are specified in Section 2.4.
- **Multi-modal Integration:** The sophistication of the method used to fuse ECG data with other information sources and the demonstrated benefit of this fusion. This includes combining ECG data with structured EHR data, demographics, lab results, or imaging reports. For example, a model described as "Co-CNN (with incorporated demographic data and clinical variables)" was shown to achieve higher sensitivity than models using ECG alone, demonstrating the value of multi-modal fusion [17].

2.5. Quality Assessment

To assess the methodological robustness of the included studies, a qualitative quality assessment was conducted based on four criteria: (1) dataset size and diversity, (2) presence of external validation, (3) clarity of model description and explainability method, and (4) clinical evaluation or validation setting. Each study was assessed descriptively across these dimensions to identify common methodological strengths and limitations without excluding studies based on quality scores. In addition, signal acquisition and preprocessing parameters—including ECG sampling rate, band-pass filter range, notch filter specification, denoising method, baseline drift correction approach, and lead configuration—were extracted for each study where reported, or recorded as "Not Reported" where absent. The aggregate non-reporting rates for these parameters are quantified in Section 4.4.

The extracted data and quality indicators were subsequently synthesized to support a comparative analysis across five dimensions: predictive accuracy, explainability, robustness and generalizability, clinical actionability, and multi-modal integration, as detailed in Section 4. For completeness, we note that the five synthesis dimensions defined in Section 2.2 did not include systematic extraction of: (1) model calibration metrics (e.g., Brier score, calibration curve, or Platt scaling), which assess the correspondence between predicted probabilities and observed event rates; (2) fairness or equity metrics (e.g., performance stratified by sex, age group, or race/ethnicity), which assess whether model performance is equitable across clinically and demographically defined subgroups; or (3) uncertainty quantification approaches (e.g., Bayesian deep learning, Monte Carlo dropout, and conformal prediction), which characterize model confidence and are increasingly recognized as important for safe clinical deployment. These dimensions were outside the predefined scope of the current review but represent an important complementary lens on clinical AI trustworthiness that future systematic reviews should address. Where these dimensions were incidentally noted in the primary studies during extraction, they are reported narratively in the relevant sections; they were not systematically extracted across all 65 studies. Regarding foundation models and self-supervised pre-training approaches for ECG analysis: no primary studies using foundation model or self-supervised pre-training approaches were identified among the 65 included studies, though our search identified relevant survey literature on this topic [18]. The rapidly evolving nature of ECG foundation models in 2024–2025 means that this review may undercapture the most recent developments in this subfield, as discussed further in Section 5.2.

2.6. Clinical Actionability Scoring Framework

To operationalize the central construct of clinical actionability beyond a qualitative narrative, a four-level ordinal scoring framework was developed and applied descriptively to each of the 65 included studies prior to the synthesis in Section 4.5. The framework was designed to be transparent, parsimonious, and grounded solely in evidence available within the primary studies—it does not constitute an independent clinical assessment of explanation quality, which would require prospective clinician evaluation beyond the scope of a literature-based systematic review. The four levels are defined as follows:

Level 0—No actionability: The study presents a predictive model producing a prediction score or classification label, with no explainability method implemented and no evidence of integration into a clinical decision pathway. Model evaluation is conducted exclusively on internal retrospective metrics.

Level 1—Technical interpretability only: The study implements a post hoc explainability method (e.g., saliency map, Grad-CAM, SHAP, LIME, or attention visualization) or an intrinsically transparent architecture (e.g., a decision tree or linear model), but provides no evidence that the explanation was evaluated by practicing clinicians, linked to pathophysiological reasoning, or embedded in a clinical workflow. The explainability output is assessed only by technical criteria.

Level 2—Clinically contextualized explainability: The study explicitly aligns model explanations with clinical or pathophysiological reasoning—for example, by mapping highlighted ECG features to established diagnostic criteria, involving clinician evaluation of explanation plausibility, or using a concept-based or knowledge-aligned architecture designed to generate clinically interpretable outputs. Prospective impact on clinical decisions is not required for Level 2 classification, but the study must demonstrate that the explanation crosses from technical attribution into clinical intelligibility.

Level 3—Demonstrated clinical impact: The study reports prospective, interventional, or trial-based evidence that the AI model measurably altered clinical decisions, workflow

efficiency, or patient health outcomes in a real-world or controlled deployment setting. This level requires a study design beyond retrospective analysis—specifically, a prospective trial, a randomized controlled trial, a pre-registered non-randomized interventional study, or a deployment study with controlled outcome measurement.

Scoring was applied by two authors (H.N. and R.A.) independently, with disagreements resolved by consensus with the third author (M.E.). The resulting scores are reported for each study in Appendix A Table A1 and are summarised in Section 4.5.

3. Background and Related Work

Over the past decade, artificial intelligence has profoundly transformed the analysis of electrocardiograms, evolving from handcrafted signal-processing pipelines to complex deep learning architectures. This section traces this evolution chronologically, highlighting how successive methodological advances addressed the limitations of earlier approaches while introducing new challenges culminating in today's dual frontier of multi-modal integration and clinically actionable explainability.

3.1. Early Machine Learning Approaches (Pre-2018): Automation Without Context

The earliest applications of AI to ECG analysis relied on classical machine learning models such as support vector machines (SVMs) and random forests. These models operated on predefined, human-engineered features like QRS duration, heart rate variability, and wave amplitudes. While these systems successfully automated the detection of certain arrhythmias, their performance was heavily dependent on the quality of feature engineering and they struggled to generalize across diverse patient populations or capture the subtle, non-linear patterns that often precede major cardiac events. This era laid the foundation for automation but lacked the capacity for deep contextual understanding.

3.2. The Deep Learning Revolution (2018–2020): Accuracy Without Transparency

A paradigm shift occurred with the introduction of deep learning, particularly Convolutional Neural Networks (CNNs), which learn features directly from raw ECG signals. The landmark study by [2] demonstrated that a CNN could predict future atrial fibrillation from a normal sinus rhythm ECG, revealing that neural networks could uncover latent electrophysiological signatures invisible to human interpretation. This breakthrough was quickly followed by models that could predict low ejection fraction [19] and other conditions with clinical-grade accuracy. However, these powerful models operated as opaque "black boxes." Their decisions could not be traced back to physiological reasoning, and as highlighted by [7], this absence of interpretability became the principal barrier to clinical trust and adoption.

3.3. The Rise of Explainable AI (2020–2023): Transparency Without Actionability

In response to skepticism about the black-box problem, researchers began integrating Explainable AI (XAI) techniques. Early efforts by [10] employed saliency maps (Grad-CAM) to generate heatmaps highlighting the ECG regions influencing a model's output. While these visualizations offered a first step toward transparency, they failed to explain why a decision was made in clinically relevant terms. Later studies applied methods like LIME and SHAP [20], but as noted by [8], these post hoc explanations remained technically descriptive but clinically uninformative—highlighting signal segments rather than pathophysiology. This generation of models achieved partial transparency but fell short of providing actionable insights for clinicians.

3.4. The Current Frontier (2023–Present): The Quest for Multi-Modal, Actionable Intelligence

The current research frontier seeks to resolve these tensions by simultaneously pursuing two goals: enhancing predictive power through multi-modal data fusion and achieving genuine clinical trust through actionable explainability. Recent studies show that integrating ECG data with EHR, biomarkers, or genomics significantly improves predictive accuracy [21]. Concurrently, there is a push toward developing inherently interpretable models that can provide justifications for their predictions in natural language, moving beyond simple heatmaps. This dual focus defines the cutting edge of the field.

Taken together, prior work suggests that advances in AI-ECG prediction can be broadly characterized along three interacting dimensions: predictive performance, model explainability, and clinical actionability. While predictive accuracy has progressed rapidly, explainability has largely remained technical in nature, and its translation into clinically actionable decision support—particularly in multi-modal settings—remains limited. This conceptual framing informs the analytical synthesis presented in the following sections.

3.5. Positioning This Review Within the Existing Literature

Several systematic reviews have mapped parts of this landscape. However, as shown in Table 1, they have focused on specific aspects in isolation. Ref. [22] provided a broad overview of AI applications, while [23] focused narrowly on data augmentation techniques. While valuable, no existing review has systematically analyzed the critical intersection of multi-modal data fusion and clinically actionable explainability. Unlike prior surveys, this review explicitly examines how explainability and multi-modal data fusion jointly shape the clinical actionability of AI-ECG systems, thereby providing a structured analytical perspective on the transition from high-performing predictive models to trustworthy clinical decision support tools. Beyond the four reviews summarized in Table 1, several additional systematic and narrative reviews have appeared recently and further illustrate the fragmented state of the field. Ref. [24] conducted a systematic literature review of explainable deep learning techniques for ECG-based heart disease classification, cataloguing XAI methods in a manner consistent with our own findings but without addressing multi-modal integration or an operationalized actionability framework. Ref. [25] similarly reviewed explainable AI methods, applications, and evaluation frameworks across clinical decision support systems more broadly, echoing our call for standardized explainability benchmarks (Section 5.3) but not focusing specifically on cardiovascular prediction. At a broader methodological level, refs. [26,27] each reviewed machine learning approaches to heart disease prediction, covering trends, datasets, and clinical applicability; neither, however, isolates explainability or multi-modal fusion as a primary analytical axis, reinforcing the gap this review is positioned to address.

Table 1. Comparison of prior systematic reviews.

Review	Year	Primary Focus	Gap Remaining for This Review
Grün et al. [28]	2021	Meta-analysis of AI for heart failure detection.	Limited to a single condition; does not analyze explainability or multi-modality.
Rahman et al. [23]	2023	Data augmentation techniques for ECG signals.	Technical focus on data preparation; does not evaluate predictive models or clinical trust.
Oke & Cavus [22]	2024	Broad impact of AI on ECGs in cardiology.	General overview; lacks a deep, structured analysis of explainability and multi-modal models.
Ose et al. [29]	2024	State-of-the-art review of AI for ECG interpretation.	Focuses on interpretation accuracy; does not systematically compare models on clinical actionability.
This Review	2026	Systematic analysis of explainable, multi-modal AI models for cardiovascular prediction.	Provides the first structured synthesis at the intersection of multi-modal fusion and clinical explainability.

Table 2 further contrasts standard single-modal and multi-modal/explainable AI-ECG paradigms across the five evaluation dimensions introduced in Section 2.2, summarizing the key insights that motivate the synthesis presented in Section 4.

Table 2. Comparative evaluation of AI-ECG model paradigms.

Metric	Standard Single-Modal AI-ECG	Multi-Modal/Explainable AI-ECG	Key Insights
Accuracy	High (AUC $>$ 0.85). Performance is well-established for specific tasks within a single population.	High–Very High (AUC $>$ 0.90). Multi-modal fusion can provide a marginal but significant boost in accuracy.	High accuracy is a necessary but insufficient condition for clinical utility; the focus is shifting to other dimensions.
Multi-modal Integration	Low. Relies exclusively on the ECG signal, ignoring valuable clinical context.	Moderate. Begins to fuse ECG with clinical risk scores, EHR data, or genetics to create a more holistic patient view.	This is a key growth area, promising more personalized and robust predictions by contextualizing the ECG signal.
Explainability	Low (“Black Box”). Provides a prediction with no rationale, which is a major barrier to clinical trust.	Moderate. Primarily uses saliency maps to show “where” the model looks, but lacks deep clinical reasoning.	There is a critical gap between basic interpretability and providing explanations that are truly actionable for a clinician.
Robustness	Low to Moderate. Often validated only on internal datasets, with performance drop-off on external cohorts.	Moderate. Increasing emphasis on external validation across different populations and healthcare systems.	Demonstrating generalizability is essential for widespread adoption and ensuring equitable performance.
Clinical Actionability	Low. Clinicians are hesitant to act on predictions they cannot understand or trust.	Low to Moderate. The strongest actionability evidence comes from single-modal ECG-only systems [2,9]. No multi-modal AI-ECG system has been evaluated in a prospective interventional trial; actionability evidence for multi-modal systems is currently absent.	This dimension bridges the gap from algorithm to patient care, though evidence for multi-modal actionability remains unestablished.

4. Results and Analysis

The systematic evaluation of the 65 included studies reveals a field of rapid advancement but also of significant, persistent challenges. While AI models demonstrate high technical performance, their journey toward robust clinical integration is shaped by critical limitations in generalizability, explainability, and proven patient impact. This section synthesizes findings across five key dimensions—predictive performance, explainability strategy, degree of multi-modal integration, robustness and validation setting, and clinical actionability—highlighting how emerging models are beginning to address the shortcomings of first-generation “black box” approaches. A quantitative summary of key methodological trends is presented in Table 3 and visualized in Figures 2–4. Following reviewer feedback, we conducted a sensitivity analysis restricting the dataset to studies whose primary predicted outcome was a cardiovascular event, condition, or cardiac risk state as specified in the revised Section 2.1 inclusion criteria. This cardiac-specific subset comprises 47 of the 65 included studies (72.3%). Of these 47 studies, 20 satisfied both the cardiovascular prediction criterion and at least one of the explainability or multi-modal criteria that motivated the review’s dual-focus design. Key prevalence figures for the cardiac-specific subset are reported throughout Section 4 in parentheses alongside the full-sample figures. The non-cardiac studies retained in the full 65-study dataset contributed methodologically informative examples of multi-modal fusion and XAI architectures; their outcomes and findings are used only for technical illustration in Sections 3 and 4 where explicitly noted, and they are excluded from all claims about cardiovascular prediction specifically.

Table 3. Quantitative summary of methodological trends in included studies ($n = 65$).

Methodological Aspect	Count	Prevalence %	Key Details
External Validation	8/65	12.3%	Validated on independent cohorts including multi-center trials [30], federated learning across institutions [31], and prospective interventional trials [32].
Multi-modal Approach	18/65	27.7%	Most common fusions: ECG + biosignals (PPG/IMU) 8 studies; ECG + EEG 3 studies; ECG + clinical/demographic data 4 studies; ECG + genomics 1 study; ECG + imaging 1 study; other combinations 1 study.
Dominant XAI Technique	7/25 *	28% of XAI studies	Attention mechanisms most frequent, followed by saliency maps/GradCAM (5 studies, 20%), SHAP (4 studies, 16%), Gini index (2 studies, 8%), LIME (2 studies, 8%), multi-layered XAI (2 studies, 8%), other (3 studies, 12%).
Studies with No XAI	40/65	61.5%	Majority of studies do not implement any explainability method, highlighting a critical transparency gap in the field.
Code/Data Availability	3–4/65	∼5%	Very few studies provide public access to source code or datasets, indicating a significant reproducibility crisis.
Prospective/Interventional Studies	6/65	9.2%	Disaggregated by evidence tier: Randomized controlled trials—2 studies [2] EAGLE, process outcome; [9] Lin et al., mortality outcome—the only study measuring a hard patient health outcome); Prospective non-randomized interventional trial—1 study ([32], AF diagnosis-rate outcome); Deployment/process study—1 study ([33], workload reduction); Multi-centre validation without intervention—1 study [30]; Prospective multimodal risk stratification study—1 study ([34], arrhythmic death prediction in hypertrophic cardiomyopathy, demonstrating high AUC in a prospective cohort). Only [9] measured a hard patient outcome in a randomized controlled design. No multi-modal AI-ECG system has been evaluated in a prospective RCT measuring patient health outcomes.
Single-center Studies	∼51/65	∼78%	Vast majority rely on data from a single institution, limiting generalizability.
ECG-only (Single-modal)	47/65	72.3%	Most studies rely exclusively on ECG signals without integrating additional clinical context.
Target Condition: AF	∼18/65	∼27.7%	Atrial fibrillation is the most studied target condition.
Target Condition: Arrhythmia	∼14/65	∼21.5%	General arrhythmia detection/classification.
Target Condition: Other CVD	∼15/65	∼23%	Including CAD, LVEF, LQTS, AV block, hypertension.
Target Condition: Non-cardiac	∼18/65	∼27.7%	Including stroke, stress, seizure, diabetes, cognitive load.
Reported AUC ≥ 0.90	∼22/65	∼34%	High-performing models concentrated in AF and arrhythmia prediction tasks.
Reported AUC 0.85–0.90	∼18/65	∼28%	Moderate-high performance, mostly single-modal ECG models.
Reported AUC < 0.85 *	∼25/65	∼38%	Lower-performing models; includes lightweight, classical ML, and resource-constrained architectures.
Deep Learning Architecture	∼48/65	∼74%	CNN, LSTM, CNN-LSTM, and transformer-based models dominate.
Classical ML Architecture	∼10/65	∼15%	SVM, random forest, SVR used mainly in interpretable or lightweight models.
Federated/Privacy-preserving	3/65	4.6%	Emerging approach for multi-institution training without raw data sharing; includes [31,35,36].
Studies reporting Precision/Recall/F1	10/65	15.4%	Multi-metric reporting is rare across the reviewed literature despite the imbalanced class distributions inherent in cardiovascular event prediction.
Imbalance handling: Oversampling/SMOTE	8/65	12.3%	Applied to address class imbalance in rare-event prediction tasks; includes SMOTE and related augmentation techniques.
Imbalance handling: Undersampling	3/65	4.6%	Used to balance training data by reducing the majority class; less common than oversampling approaches.
Imbalance handling: Cost-sensitive/Class-weighted	4/65	6.2%	Applied class-weighted loss functions to penalise misclassification of the minority class.
Imbalance handling: Not reported	50/65	76.9%	Majority of studies do not disclose any strategy for managing class imbalance, limiting interpretability of reported performance metrics.
Actionability: Level 0 (No actionability)	40/65	61.5%	No explainability method or deployment context reported.
Actionability: Level 1 (Technical XAI only)	15/65	23.1%	XAI method implemented but no clinician evaluation or pathophysiological reasoning.
Actionability: Level 2 (Clinically contextualized)	4/65	6.2%	Explanations aligned with clinical reasoning; includes [37,38].

Table 3. *Cont.*

Methodological Aspect	Count	Prevalence %	Key Details
Actionability: Level 3 (Demonstrated clinical impact)	6/65	9.2%	Prospective or interventional evidence; includes [2,9,30,32–34].
Target Condition: Non-cardiac primary outcome	∼18/65	∼27.7%	Including stress monitoring, seizure detection, cognitive load estimation, user recognition, hydration monitoring, insomnia staging, diabetes prediction, and fall-risk prediction. These studies are excluded from the cardiac-specific sensitivity analysis reported throughout Section 4.

* AUC performance bands (≥0.90 = high; 0.85–0.90 = moderate; <0.85 = lower) were pre-specified as descriptive categories in Section 2.2. They were not used as eligibility thresholds.

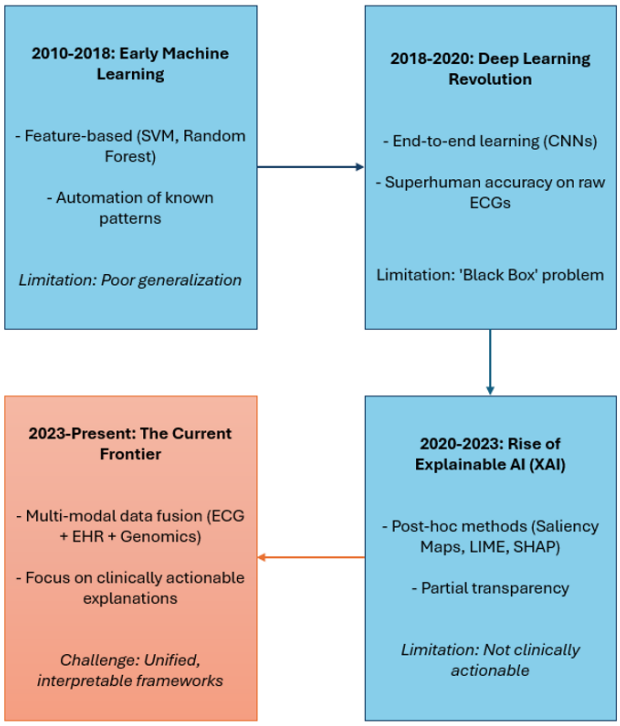

Figure 2. A timeline of methodological evolution in AI-ECG prediction.

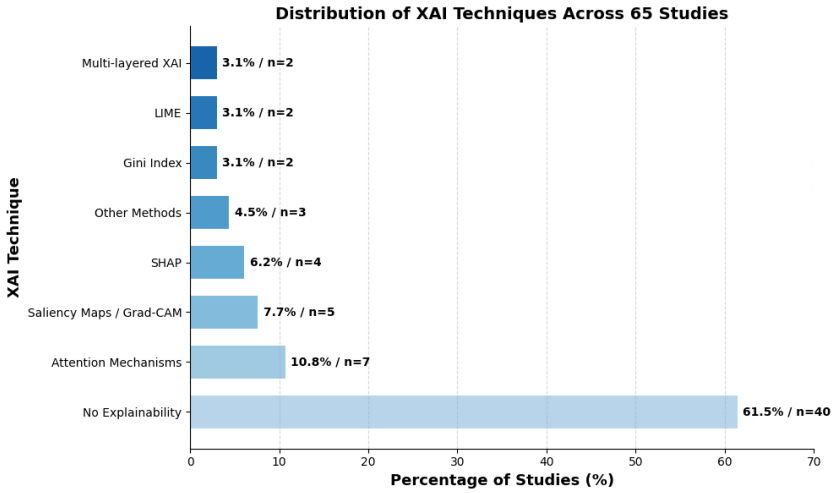

Figure 3. Distribution of explainability (XAI) techniques across all 65 included studies, expressed as a percentage of all studies ($n = 65$). The dominant category is No Explainability (61.5%, $n = 40$). Among the 25 studies implementing at least one XAI method (38.5%), methods are distributed as follows: Attention Mechanisms (10.8%, $n = 7$), Saliency Maps/Grad-CAM (7.7%, $n = 5$), SHAP (6.2%, $n = 4$), Gini Index (3.1%, $n = 2$), LIME (3.1%, $n = 2$), Multi-layered XAI (3.1%, $n = 2$), and Other Methods (4.5%, $n = 3$).

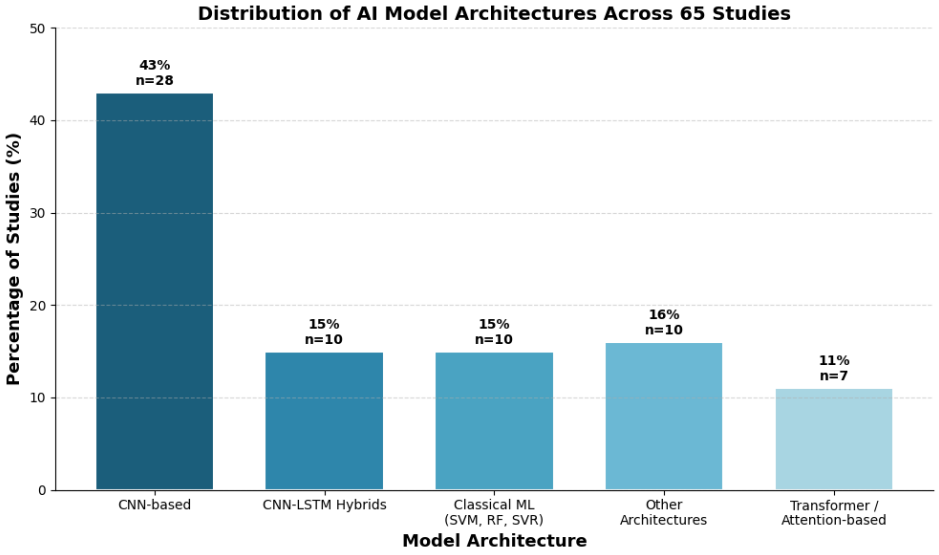

Figure 4. Distribution of AI model architectures across all 65 included studies, expressed as a percentage of studies ($n = 65$). Categories are ordered by decreasing frequency: CNN-Based Architectures (43%, $n = 28$), CNN-LSTM Hybrids (15%, $n = 10$), Classical ML Approaches—SVM, Random Forest, SVR (15%, $n = 10$)—Transformer/Attention-Based models (11%, $n = 7$), and Other Architectures (16%, $n = 10$), which includes reinforcement learning, generative models, and other miscellaneous approaches. Deep learning architectures collectively account for approximately 74% of included studies.

4.1. Predictive Accuracy

Across the reviewed literature, AI-ECG models consistently demonstrate high predictive accuracy in controlled settings. For tasks such as predicting incident atrial fibrillation (AF) or identifying low left ventricular ejection fraction (LVEF), most models report an Area Under the Curve (AUC) between 0.85 and 0.95 on internal validation sets [2,39]. Approximately 34% of reviewed studies reported AUC values exceeding 0.90, with the highest performance concentrated in AF and arrhythmia prediction tasks. Classical deep learning architectures, particularly CNN-based models, dominate the landscape, accounting for approximately 43% of all included studies, followed by CNN-LSTM hybrids (15%), transformer and attention-based models (11%), and classical ML approaches such as SVM and Random Forest (15%) (Figure 4).

However, a major limitation of many studies is their reliance on large but homogeneous single-center datasets, which account for approximately 78% of all included studies (Table 3). While internal validation metrics are impressive, external validation—performed by only 12.3% of studies—often reveals performance degradation, underscoring the critical need for more diverse and representative training data.

A growing number of studies are now directly addressing this generalizability gap. For instance, ref. [11] successfully validated their AF prediction model on the independent, multi-center MIMIC-IV dataset, demonstrating retained—though slightly lower—performance. Similarly, ref. [40] validated their multi-modal model for predicting AF on an international population combining ECG, clinical variables, and genomics, confirming its utility beyond its initial development cohort. Ref. [30] further demonstrated multi-center validation for detecting Long QT Syndrome and AV block using a single-lead AI-ECG system. These studies are pivotal, as they provide essential evidence of real-world viability. High accuracy, however, remains a necessary but insufficient condition for clinical utility: the field is increasingly shifting its focus toward explainability, robustness, and actionability. Consistent with the transformer-based trend noted in Section 3.4, several additional transformer architectures for cardiovascular prediction

have emerged during the review period. Ref. [41] proposed a transformer-based approach combining ECG signals with clinical variables for cardiovascular disease prediction, while [42] demonstrated multi-label transformer-based classification for early detection of cardiac abnormalities. Ref. [43] introduced a hybrid transformer architecture (DeepECG-Net) for real-time ECG anomaly detection and classification. These studies reinforce the broader architectural shift toward transformer and attention-based models identified in Section 4.1 and Figure 4, though, consistent with the pattern observed across the reviewed literature, none report external validation or prospective clinical evaluation. A notable limitation in the reporting practices of the reviewed literature concerns metric completeness and class-imbalance transparency. Of the 65 included studies, only 10 (15.4%) reported precision, recall, or F1-score alongside AUC, despite the fact that cardiovascular event prediction tasks—such as incident atrial fibrillation, sudden cardiac death, and acute coronary syndrome—are inherently characterized by imbalanced outcome distributions in which positive events are rare relative to the negative class. In such settings, AUC alone can overestimate practical model utility by not reflecting performance on the minority positive class. With regard to imbalance-handling strategies, eight studies applied oversampling or synthetic minority oversampling (SMOTE) techniques, three used undersampling, four applied cost-sensitive or class-weighted loss functions, and the remaining 50 studies did not report any explicit imbalance-handling strategy (Table 3). This represents a significant transparency gap: without knowledge of how class imbalance was managed, the high AUC values reported across the literature cannot be straightforwardly interpreted as evidence of strong performance on the clinically critical positive class.

4.2. Multi-Modal Integration

A clear trend in recent research is the shift from single-modal (ECG only) to multi-modal models that enrich the ECG signal with clinical context. However, as shown in Table 3, ECG-only studies still constitute 72.3% of the reviewed literature, and multi-modal approaches—while growing—represent only 27.7% of included studies (18 out of 65). Integrating heterogeneous data sources introduces significant challenges in standardizing formats, handling missing or noisy data, and ensuring computational efficiency.

Despite these hurdles, successful multi-modal models demonstrate clear and measurable benefits. Ref. [20] showed that combining AI-ECG with clinical risk factors significantly improved AF risk stratification over single-modality models. Ref. [40] further demonstrated that fusing ECG data with a polygenic risk score (genomics) and a clinical risk score improved AF prediction beyond any single modality, pointing toward a future where models can create a holistic patient profile by combining the ECG's subtle electrical signals with biomarkers and genomics. The most common fusion combinations identified in this review were ECG with biosignals such as PPG and IMU (eight studies), ECG with clinical or demographic data (four studies), ECG with EEG (three studies), and ECG with genomics (one study).

More recently, the frontier has expanded toward integrating imaging data. The newly included study by [34] demonstrated that multimodal AI combining ECG, cardiac imaging, and clinical variables could forecast arrhythmic death in hypertrophic cardiomyopathy with high accuracy, representing one of the most clinically impactful multi-modal applications identified in this review. Similarly, ref. [44] proposed a precision medicine framework integrating genomics, imaging, and ECG data, further illustrating the potential of comprehensive data fusion for personalized cardiovascular risk prediction.

This shift toward richer data fusion is consistent with the broader narrative reviewed by [45], who surveyed multimodal data fusion strategies combining ECG with electronic health records for cardiovascular risk prediction; their narrative synthesis identifies similar

integration challenges around missing data and format standardization to those discussed above, though without the systematic, PRISMA-based quantification provided by the present review.

These findings collectively suggest that multi-modal integration is a key growth area for improving predictive performance and contextualising the ECG signal within a broader clinical picture. However, a critical distinction must be drawn between evidence of improved predictive accuracy—which is present for multi-modal approaches, as demonstrated by [20,34,40]—and evidence of improved clinical actionability, which requires prospective or interventional evaluation rather than retrospective AUC comparison. No multi-modal AI-ECG system identified in this review has been evaluated in a prospective randomized or interventional trial measuring patient health outcomes or clinical decision changes. The most robust actionability evidence in this review—the EAGLE trial [2] and the mortality-reducing trial of Lin et al. [9]—both derive from single-modal, ECG-only systems. The relationship between multi-modal integration and clinical actionability therefore remains an empirically open and important hypothesis for future prospective investigation, rather than a demonstrated finding in the current literature.

4.3. Explainability and Interpretability

Explainability remains the most significant barrier to clinical trust in AI-ECG systems. As detailed in Table 3 and visualized in Figure 3, the field is overwhelmingly characterized by a lack of transparency: 61.5% of included studies (40 out of 65) implemented no explainability method whatsoever. Among the 25 studies that did incorporate XAI techniques, attention mechanisms were the most prevalent approach (28% of XAI studies, 10.8% of all studies), followed by saliency maps and GradCAM (20% of XAI studies), SHAP (16%), Gini index (8%), LIME (8%), and multi-layered XAI (8%).

The classification of attention mechanisms as an explainability technique in this review follows the framing used by the original study authors and is not an endorsement of their explanatory validity. A substantive and unresolved methodological debate exists in the machine learning literature regarding whether attention weights constitute genuine explanations of model reasoning—faithfully reflecting the causal drivers of model output—or merely represent learned feature salience without a verified link to the model's decision process [46,47]. None of the attention-based studies identified in this review formally tested the faithfulness of attention weights to model decisions through perturbation-based or counterfactual methods. Accordingly, the reported prevalence of attention mechanisms as the leading XAI approach (28% of XAI-implementing studies) should be interpreted as an upper bound on the field's transparency, acknowledging that the explanatory validity of attention-based outputs has not been independently verified in the primary literature reviewed here.

The most prevalent method—the saliency or class activation map—has been heavily critiqued. While saliency maps highlight influential ECG regions, they fundamentally fail to provide pathophysiological reasoning, limiting their clinical utility. They answer where the model is looking but not why in terms a clinician can act upon [7,19]. To address these limitations, recent studies have explored more advanced explainability approaches. Post hoc techniques such as SHAP and LIME, applied by [14,48], provide more granular feature attribution compared to saliency-based methods; however, their ability to deliver intuitive and clinically actionable insights remains contested. The newly included study [38] specifically addressed this gap by conducting a rigorous clinical validation of SHAP and LIME explanations for cardiovascular prediction, demonstrating that while these methods enhance technical interpretability, they rarely translate directly into actionable clinical reasoning without additional contextual scaffolding.

A further notable gap in the current literature is the near-complete absence of concept-based and causal explanation frameworks applied to AI-ECG prediction. Concept Bottleneck Models, Testing with Concept Activation Vectors (TCAVs), and causal graph-based explanation methods were not identified in any of the 65 included studies as implemented approaches. This absence is not incidental: concept-based methods require pre-specified, clinically meaningful concept ontologies—for example, systematically defined categories of T-wave morphology, QRS duration thresholds, or P-wave abnormality patterns—that have not been standardised for ECG-based AI research. Similarly, causal explanation frameworks require interventional datasets or validated causal graph structures that are rarely available in retrospective ECG repositories. These structural prerequisites represent substantive barriers to adoption rather than simply an oversight in the research community, and their development is identified as a priority direction in Section 5.3. With respect to clinician evaluation of explanations, of the 25 XAI-implementing studies identified in this review, only two reported any form of evaluation by practicing clinicians of the plausibility or clinical utility of the model's explanations. This figure—which includes the clinical validation study [38] and the knowledge-aligned transformer study [37]—suggests that the high prevalence of XAI method implementation substantially overstates the prevalence of clinically contextualised or clinician-validated explainability across the reviewed literature.

In parallel, a smaller subset of studies has investigated intrinsically interpretable model architectures. The newly included study [37] proposed an interpretable transformer architecture explicitly aligned with clinical ECG knowledge, representing a significant step toward models that are transparent by design rather than explained after the fact. Similarly, refs. [8,19] demonstrated that classical ML models using Gini index feature importance can achieve competitive accuracy (94.4%) while maintaining full transparency. These approaches aim to align model reasoning with clinically meaningful concepts rather than retrospective feature attribution, yet their adoption remains limited and their real-world clinical impact is still insufficiently validated.

4.4. Robustness and Generalizability

The robustness of AI-ECG models is a frequently cited concern and a persistent methodological weakness in the reviewed literature. As quantified in Table 3, only 12.3% of studies (8 out of 65) performed external validation on an independent cohort—a striking finding that underscores the fragility of many reported performance metrics. The vast majority of studies (78%) relied on single-center datasets, raising serious questions about the generalizability of their findings to diverse clinical populations and healthcare systems.

To proactively improve robustness, several strategies are being explored. Federated learning—identified in three studies (4.6% of included studies; [31,35,36])—represents a particularly promising approach, enabling models to train on decentralized data from multiple institutions without sharing raw patient information, thereby preserving privacy while improving generalizability.

Among the cardiovascular-focused federated studies, ref. [31] reported robust cross-institution generalizability across multiple institutions without raw data sharing, achieving AUROC values of 0.90–0.96 across validation cohorts, including AUROC = 0.93 on the external Brigham and Women's Hospital validation set. Alasmari et al. [35] proposed a privacy-preserving federated multimodal framework integrating ECG, cardiac imaging, EHR, and nutritional data for cardiovascular disease detection, achieving high classification accuracy (97.76–99.12%) across multiple datasets; no direct head-to-head comparison with a centralized baseline was reported. Ref. [36] reviewed explainable AI and federated learning architectures for privacy-preserving healthcare, highlighting their potential to improve model transparency, security, and clinical trust, though without reporting a specific

performance metric. It should be noted that direct comparisons between federated and centralized training performance are rare in the reviewed literature; the available evidence is therefore preliminary, and future work should report such comparisons explicitly to substantiate the generalizability advantage of federated approaches.

Two further federated-learning architectures, published after our primary search window closed but relevant to this discussion, illustrate the continued trajectory of this approach. Ref. [49] proposed Mamba-fusion, a privacy-preserving multimodal disease-prediction framework integrating federated learning with state-space models, extending the architectures identified in [31,35,50] toward more expressive sequence models. Ref. [51] similarly developed an explainable federated learning framework for multi-class heart disease classification using ECG signals, combining privacy preservation with interpretability in a single architecture—directly addressing the joint fusion-explainability gap highlighted throughout Section 4.3. As with the federated studies already included in this review, neither reports a direct centralized-training comparator.

Ref. [32] provided additional evidence of robustness through a prospective non-randomized interventional trial, demonstrating that AI-guided AF screening maintained performance in a real-world clinical deployment.

Collectively, these findings highlight that demonstrating generalizability is not merely a methodological nicety but an essential prerequisite for widespread clinical adoption and equitable performance across diverse patient populations. Reproducibility is further constrained by widespread non-reporting of signal acquisition and preprocessing parameters across the included literature. Among the 49 primary empirical studies (excluding 16 reviews, editorials, framework papers, and text/LLM-only studies), 42 (85.7%) did not report ECG sampling rate, 38 (77.6%) did not specify band-pass filter parameters, 41 (83.7%) did not describe their denoising method, 44 (89.8%) did not report baseline drift correction procedures, and 35 (71.4%) did not specify lead configuration. These omissions are not incidental: sampling rate and filtering parameters directly determine the frequency content of the ECG signal available to the model, and their non-disclosure prevents independent verification of whether reported performance metrics are reproducible on other hardware or signal acquisition systems. Taken together with the finding that only approximately 5% of studies provide public access to source code or datasets (Table 3), the preprocessing reporting gap identified here points to a systemic reproducibility deficit across the AI-ECG literature that substantially limits the translational utility of the reported models.

4.5. Clinical Actionability

Applying the four-level clinical actionability scoring framework specified in Section 2.4, the distribution of scores across the 65 included studies was as follows: Level 0 (no actionability)—40 studies (61.5%); Level 1 (technical interpretability only)—15 studies (23.1%); Level 2 (clinically contextualized explainability)—four studies (6.2%); Level 3 (demonstrated clinical impact)—six studies (9.2%). This distribution confirms quantitatively that clinical actionability, as distinct from technical prediction performance or explainability method implementation, is the least developed dimension across the reviewed literature. The majority of included studies fall within Levels 0 and 1, indicating that most AI-ECG research produces predictions and, where present, technical explanations that have not been contextualized within clinical reasoning or subjected to real-world evaluation. Full study-level scores are provided in Appendix A Table A1.

The ultimate measure of an AI model's value is its ability to positively impact clinical practice. Several pragmatic trials included in this review have demonstrated that AI-ECG models can successfully alter clinical workflows. The landmark EAGLE trial [2] showed that providing clinicians with an AI-generated alert for low LVEF in a primary

care setting significantly increased the rate of formal diagnosis. Ref. [32] similarly demonstrated that AI-guided AF screening during sinus rhythm increased new AF diagnoses in a prospective interventional trial. Ref. [33] further showed that AI-based direct-to-physician reporting of ambulatory ECG significantly reduced physician workload while maintaining diagnostic accuracy.

Crucially, the field is now moving beyond measuring process changes to measuring direct impacts on patient health outcomes. Ref. [9] provided powerful evidence in this direction, demonstrating that an AI-ECG alert intervention was associated with a significant reduction in all-cause mortality over a median follow-up of 2.5 years—representing the highest level of clinical evidence identified in this review. The newly included study [34] on multimodal AI for arrhythmic death prediction in hypertrophic cardiomyopathy further illustrates the potential of combining multiple data modalities to achieve clinically actionable risk stratification for high-stakes decisions.

Clinical actionability remains the least developed dimension across the reviewed literature. The aggregate figure of 9.2% (6/65) of studies categorised as prospective or clinical trials requires disaggregation by evidence tier to be informative. Of the six studies in this category: two were randomized controlled trials—the EAGLE trial [2], which demonstrated increased new LVEF diagnosis rates (a process outcome) in a primary care setting, and the pragmatic trial of Lin et al. [9], which demonstrated a significant reduction in all-cause mortality over a median follow-up of 2.5 years (the only study in this review to measure a hard patient health outcome in a randomized design). One additional study [32] was a prospective, non-randomized interventional trial demonstrating increased new AF diagnosis rates in real-world deployment. One study [33] evaluated AI-based ambulatory ECG reporting as a deployment study, demonstrating reduced physician workload (a process outcome) without a controlled patient-outcome endpoint. One study [30] reported multi-centre performance validation across multiple hospitals without an interventional design.

This disaggregation reveals that genuine randomized outcome-level evidence for AI-ECG clinical impact is limited to two RCTs within the reviewed literature, of which only one [9] measured a hard patient health outcome rather than a diagnostic process metric. Moreover, both RCTs involved single-modal, ECG-only systems—no multi-modal AI-ECG system has yet been evaluated in a prospective RCT measuring patient health outcomes, representing the most important evidence gap in the field.

A Delphi study identified that "poor model performance" was the most important barrier to adoption, while "faster recognition of subtle ECG abnormalities" was the top facilitator [52], highlighting the direct and critical link between technical performance, explainability, and clinical action. The need for prospective trials measuring downstream effects on hard patient outcomes—including reduced mortality and hospitalizations—remains paramount and represents the most important frontier for future research.

5. Discussion

This systematic review provides comprehensive and structured insights into the evolving landscape of AI models for cardiovascular prediction, focusing on the critical intersection of multi-modal data integration and clinical explainability. The findings presented in Section 4 offer a nuanced understanding of the field's performance, strengths, limitations, and future trajectory. This section addresses the four formulated research questions in greater depth, contextualizes the review's contributions and limitations, and outlines profound implications for both researchers and clinicians.

5.1. Addressing Research Questions

RQ1: What specific architectures and explainability techniques are employed, and how are they adapted for multi-modal data? Our review confirms the dominance of CNN-based architectures for ECG analysis, accounting for approximately 43% of all included studies. Transformer and attention-based models are emerging as a significant secondary trend (11%), reflecting the broader shift in deep learning toward sequence modeling paradigms better suited to temporal ECG signals. For explainability, the field relies heavily on post hoc visualization techniques, with attention mechanisms (10.8% of all studies) and saliency maps (7.7%) being the most prevalent. However, a striking finding is that 61.5% of studies implement no explainability method at all, revealing a profound transparency deficit across the field.

The adaptation of XAI techniques for multi-modal data remains a critical and largely unresolved weakness. In models fusing ECG with EHR data, feature importance scores are often used for the non-ECG components while saliency maps are applied to the ECG branch—creating a disjointed and incomplete picture for the clinician. Unified explanations that articulate how different data modalities interact to inform the final prediction remain exceptionally rare. The newly included interpretable transformer study [37] represents a promising step toward architectures that are transparent by design, while [44] illustrates the potential of attention mechanisms to weigh the importance of different modalities in a unified framework. Nevertheless, the development of truly unified, multi-modal explainability frameworks remains an open and urgent research challenge.

RQ2: What are the key performance trade-offs in current AI-ECG models? The primary trade-off identified in this review is not a simple loss in accuracy for explainability, but rather a more complex and multidimensional tension between model complexity, clinical interpretability, and generalizability. As models become more powerful by incorporating multiple data modalities—as demonstrated by [40] with ECG, clinical variables, and genomics, and by [34] with ECG, imaging, and clinical data—their internal complexity grows, making them even more opaque. This highlights that simply applying basic post hoc explanation methods to increasingly complex models is an insufficient and ultimately counterproductive strategy. Furthermore, the trade-off between performance and generalizability is equally critical. Models achieving AUC > 0.90 on internal validation sets frequently demonstrate significant performance degradation on external cohorts, as evidenced by the fact that only 12.3% of studies performed external validation.

The federated learning approaches identified in this review [31,35,36] offer a promising pathway to address this trade-off by enabling training on diverse, decentralized datasets without compromising patient privacy. However, quantitative comparisons with centralized training are not consistently available across these studies; the cross-institution performance reported in [31,35] supports the feasibility of federated architectures for cardiovascular AI, but explicit centralized-versus-federated benchmarks are needed to confirm the claimed generalizability advantage.

These findings collectively underscore the urgent need for inherently interpretable architectures that are transparent by design and validated across diverse populations, rather than attempting to explain a black box after the fact on a single homogeneous dataset.

A further underappreciated trade-off is the distinction between predictive performance gains and clinical actionability gains from multi-modal integration. The reviewed literature provides reasonable evidence that multi-modal fusion improves AUC—particularly for AF prediction [20,40] and complex cardiovascular risk stratification [34]—but provides no direct evidence that multi-modal integration improves clinical actionability specifically. Disentangling these two dimensions is essential for setting appropriate research priorities:

future work should include prospective outcome-level evaluation of multi-modal systems, rather than relying on retrospective AUC improvement as a proxy for clinical benefit.

RQ3: To what extent do existing models address the critical need for clinical explainability? Our analysis reveals that the need for true clinical explainability is largely unmet. There is a profound and persistent gap between technical interpretability and clinical trust. The prevalent use of saliency maps identified in 7.7% of all studies is a prime example of this disconnect. These heatmaps fail to translate into actionable clinical insights, such as identifying subtle markers of ventricular hypertrophy or early signs of repolarization abnormalities. They show where a model is looking but critically fail to explain why in the pathophysiological language clinicians use for diagnosis and treatment planning [7,19].

More advanced post hoc methods such as SHAP and LIME, while providing more granular feature attribution, face similar limitations in clinical translation. The clinical validation study [38] included in this review specifically demonstrated that while SHAP and LIME improve technical interpretability, their explanations rarely provide sufficient context to directly influence clinical decision-making without additional scaffolding. The most promising approaches identified are intrinsically interpretable architectures such as the clinical knowledge-aligned transformer [37] and concept-based models that generate explanations as an integral part of the prediction process rather than as a post hoc approximation. However, these approaches remain a small minority of the reviewed literature and require substantially more clinical validation before widespread adoption.

Furthermore, concept-based and causal explanation methods—which would more directly address the pathophysiological-language gap identified throughout this review—are structurally absent from the current AI-ECG literature due to the absence of standardized ECG-specific concept ontologies and the rarity of the interventional datasets needed for causal analysis. This represents an important direction for future research investment rather than a limitation of the review's scope. The prevalent classification of attention mechanisms as explainability methods should also be treated with caution in light of the ongoing debate in the machine learning literature regarding whether attention weights faithfully represent model reasoning [46,47]; the field has not yet established the explanatory validity of attention-based outputs in AI-ECG systems, and future work should evaluate attention faithfulness explicitly rather than assuming it.

RQ4: What are the primary research gaps and emerging trends? The most significant research gap identified by this review is the development of AI models that are simultaneously multi-modal, clinically explainable, and externally validated. The most promising emerging trends to address this are: (1) the shift toward concept-based learning where models learn to identify and report on clinically relevant concepts rather than abstract features; (2) the application of large language models (LLMs) and transformers to ECG analysis, as surveyed in [18], enabling natural language rationales for predictions; (3) the integration of imaging data with ECG and genomics to create comprehensive digital patient profiles, as demonstrated by [34,44]; and (4) the adoption of federated learning frameworks [31] to enable robust, privacy-preserving multi-institution validation. These trends collectively define the cutting edge of the field and represent the most productive directions for future research investment.

Additionally, the emergence of large-scale self-supervised and foundation models for ECG analysis—in which models are pre-trained on millions of unlabelled ECG recordings before fine-tuning on specific prediction tasks—represents a structural shift in the field's methodology that is partially captured in this review [18] but will require dedicated systematic evaluation as more primary outcome studies become available.

5.2. Contributions and Limitations of This Review

This review offers the first structured synthesis at the intersection of multi-modal data fusion and clinically actionable explainability in AI-ECG systems, providing a dual-lens analytical framework that prior reviews have not addressed. By systematically evaluating 65 studies across five analytical dimensions—predictive accuracy, explainability, multi-modal integration, robustness, and clinical actionability—this review provides both a comprehensive snapshot of the current state of the field and a structured roadmap for future research priorities.

However, several limitations must be acknowledged.

First, by focusing only on English-language studies, this review may have missed innovations from non-English-speaking research communities, particularly from China, Japan, and South Korea, which are highly active in AI-ECG research.

This restriction likely introduces a directional bias in several of the prevalence estimates reported in Table 3. Specifically, the dominance of CNN-based architectures (43% of included studies) and post hoc explainability methods such as attention mechanisms and saliency maps may be overestimated relative to the true global distribution, as lightweight and edge-deployable architectures are disproportionately represented in Asian engineering venues. Conversely, the relatively low rate of multi-modal fusion studies (27.7%) and prospective clinical trials (9.2%) may underestimate actual global prevalence, as some research in these areas is published in regional journals not comprehensively indexed by PubMed or Web of Science. Accordingly, the quantitative figures reported throughout Section 4 and Table 3 should be interpreted as characterizing the English-language AI-ECG literature specifically, rather than as definitive global estimates.

Second, our conclusions rely on reported performance metrics within the selected literature, which may not fully capture real-world efficacy due to differences in clinical workflows, patient populations, and implementation challenges.

Third, the review protocol was not pre-registered in a public database such as PROSPERO, which, while mitigated by our predefined extraction template and transparent methodology, represents a limitation in terms of prospective bias prevention. Fourth, the rapidly evolving nature of the field means that significant studies published after August 2025 are not captured in this review.

Fifth, the reporting practices of the included primary literature—specifically, near-universal reliance on AUC as the sole performance metric (only 10 of 65 studies, 15.4%, reported precision, recall, or F1), combined with infrequent disclosure of class-imbalance handling strategies (50 of 65 studies, 76.9%, did not report any strategy)—limit the depth of the quantitative synthesis that this review can provide on model performance for rare-event prediction tasks. We have characterized these reporting gaps explicitly in Table 3 and Section 4.1 rather than treating them as a limitation of the review's extraction, as they represent a systemic transparency deficit in the underlying primary literature rather than an omission in our data collection.

Sixth, approximately 27.7% of the 65 included studies ($n \approx 18$) had a primary predicted outcome that is non-cardiovascular in nature. These studies entered the dataset because the search strategy was designed to capture ECG-based AI systems employing multi-modal fusion or XAI techniques regardless of outcome domain, and their methodological contributions to explainability and fusion architecture are relevant to the technical scope of this review. However, to ensure that prevalence statistics are not misleadingly applied to a cardiovascular prediction claim, a cardiac-specific sensitivity analysis was conducted (Section 4), and all key figures are reported alongside the full-sample figures. We acknowledge that a stricter application of the original inclusion criteria at the screening

stage would have excluded these studies and reduced the dataset to 47 studies. The impact of this restriction on core conclusions is reported in the sensitivity analysis.

Seventh, the data extraction template for this review did not include systematic extraction of model calibration metrics, fairness or subgroup performance indicators, or uncertainty quantification approaches. These dimensions represent an increasingly important complementary framework for evaluating the trustworthiness of clinical AI systems—one that extends beyond the explainability and actionability dimensions addressed here—and their absence from this review reflects a scoping decision rather than a judgement about their importance. A dedicated systematic review examining calibration, fairness, and uncertainty quantification in AI-ECG systems would be a valuable and timely companion to the present work.

5.3. Implications and Future Work

The findings of this review underscore that the future of AI in cardiovascular medicine hinges fundamentally on building clinical trust. To fully realize the transformative potential of AI-ECG, future work must prioritize four interconnected directions:

1. *Developing Inherently Interpretable Models*

The field must move decisively beyond post hoc explanations toward architectures that are transparent by design. This includes developing concept bottleneck models that can explicitly state they are flagging a patient due to "T-wave inversion patterns" or "prolonged QT interval," case-based reasoning systems that justify predictions by referencing similar verified patient cases, and natural language generation systems that produce clinician-readable rationales [18]. The interpretable transformer proposed in [37] represents an important step in this direction, demonstrating that high predictive performance and clinical interpretability need not be mutually exclusive.

This need for cognitively aligned model design has recently been articulated explicitly by [53], who argue for predictive cardiovascular AI models structured specifically around what clinicians can understand, trust, and act upon—a framing that closely mirrors the clinical actionability construct operationalized in Section 2.4 of this review. Figure 5 illustrates this target architecture conceptually, depicting how heterogeneous inputs—raw ECG, electronic health records, and genomic data—would be integrated through a unified multi-modal fusion and explainability engine to produce clinically actionable outputs.

2. *Advancing Multi-modal Fusion and Explanation*

The next generation of models must not only fuse heterogeneous data—including ECG, EHR, biomarkers, genomics, and imaging—but must also provide unified explanations that articulate how each modality contributes to the final prediction.

Critically, the value of these multi-modal explanations must be demonstrated through prospective clinical evaluation measuring decision-making impact or patient outcomes—not solely through retrospective AUC comparisons—given that no multi-modal AI-ECG system identified in this review has yet been subjected to such evaluation.

The multimodal AI system for arrhythmic death prediction [34] and the precision medicine framework integrating genomics and imaging [44] represent the current frontier of this direction. Attention mechanisms offer a particularly promising architectural basis for unified multi-modal explainability, as they can learn to weigh the importance of different data sources in a manner that is both computationally efficient and potentially interpretable.

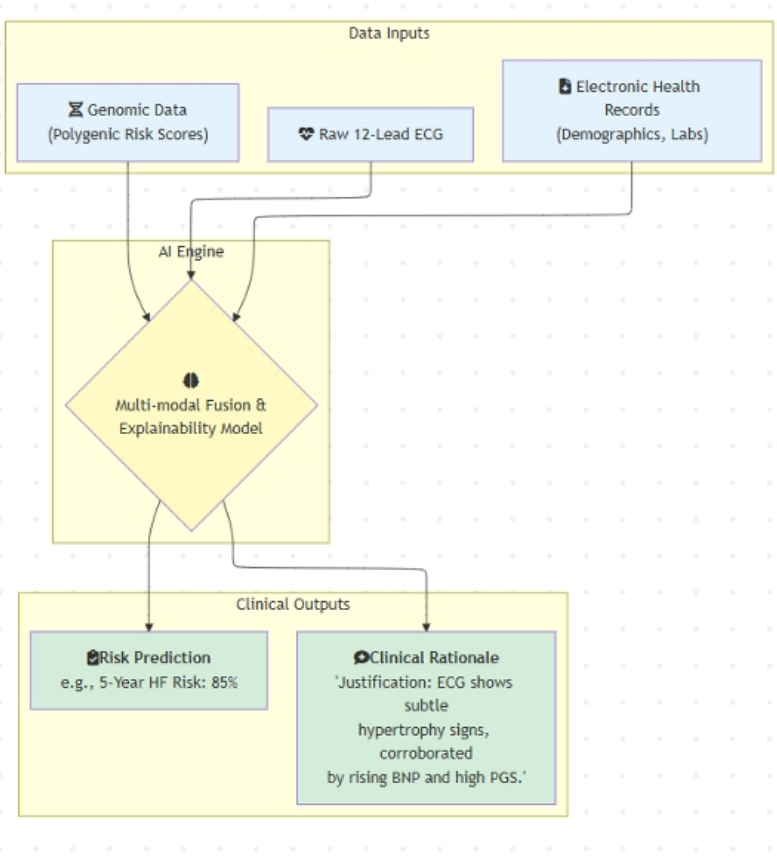

Figure 5. Conceptual framework for multi-modal AI-ECG-based cardiovascular risk prediction. The figure illustrates the integration of heterogeneous data inputs (raw 12-lead ECG, electronic health records, and genomic data) through a multi-modal fusion and explainability AI engine, producing clinically actionable outputs including risk prediction and a clinical rationale with pathophysiological justification.

3. *Establishing Standardized Benchmarks for Explainability*

The evaluation of explanations is currently subjective and inconsistent across the reviewed literature. The field urgently needs standardized metrics to assess explanation quality across at least three axes: fidelity (how accurately does the explanation reflect the model's internal logic?), comprehensibility (can a clinician easily understand the explanation?), and clinical actionability (does the explanation provide sufficient information to influence a clinical decision?). The clinical validation framework proposed in [38] offers a methodological template for this type of standardized evaluation.

4. *Conducting Robust Prospective Validation*

The value of explainable, multi-modal models must ultimately be proven in practice through rigorous prospective trials. Future trial designs should measure not only improvements in clinical processes—such as faster diagnosis, as demonstrated in the EAGLE trial [2]—but also downstream effects on hard patient outcomes, including reduced mortality and hospitalizations, as demonstrated by the landmark pragmatic trial of [9]. The federated learning framework [31] offers a pathway to conduct such trials across multiple institutions while preserving patient privacy, potentially enabling the large-scale, diverse validation that the field urgently needs.

6. Conclusions

Cardiovascular diseases remain the foremost cause of mortality worldwide, and the electrocardiogram—a century-old clinical tool—continues to serve as one of the most accessible and information-rich windows into cardiac health. The integration of artificial intelligence with ECG analysis holds transformative potential to shift cardiovascular medicine from a reactive to a proactive paradigm, enabling earlier detection, more accurate risk stratification, and ultimately a meaningful reduction in the global burden of CVD mortality. This systematic review, conducted in accordance with PRISMA 2020 guidelines and synthesizing 65 peer-reviewed studies published between 2018 and 2025, has provided the first structured analysis at the critical intersection of multi-modal data fusion and clinically actionable explainability in AI-ECG systems.

Our findings confirm that the field has achieved remarkable technical progress. AI-ECG models consistently demonstrate high predictive accuracy, with approximately 34% of reviewed studies reporting AUC values exceeding 0.90, and deep learning architectures—particularly CNN-based models (43% of studies) and emerging transformer-based approaches (11%)—have established themselves as the dominant paradigm for ECG analysis. Multi-modal integration, while still representing a minority of the reviewed studies (27.7% of studies), has demonstrated clear and measurable benefits in predictive performance, with models fusing ECG data with clinical variables, genomics, and imaging achieving higher AUC values than single-modality counterparts in retrospective studies. However, direct outcome-level evidence that multi-modal integration specifically improves clinical actionability—as distinct from predictive accuracy—is currently absent from the reviewed literature: the most robust clinical impact evidence identified in this review derives from single-modal ECG-only systems [2,9].

The landmark studies identified in this review—including the EAGLE trial [2], the mortality-reducing pragmatic trial of [9], and the multimodal arrhythmic death prediction system [34]—represent the highest level of clinical evidence and illustrate the genuine transformative potential of AI-ECG when rigorously validated.

Yet this review also reveals a field whose clinical promise is fundamentally constrained by three persistent and interrelated challenges. First, the explainability deficit is severe: 61.5% of included studies implement no explainability method at all, and among those that do, the dominant approaches—attention mechanisms (10.8% of all studies) and saliency maps (7.7%)—provide technical transparency without clinical actionability. They reveal where a model looks but fail to explain why in the pathophysiological language that clinicians require for high-stakes decision-making. Second, the generalizability crisis is equally concerning: only 12.3% of studies performed external validation on an independent cohort, and 78% relied on single-center datasets, raising fundamental questions about the real-world applicability of reported performance metrics. Third, the clinical translation gap remains wide: only 9.2% of included studies were prospective or clinical trials, meaning that the vast majority of AI-ECG systems have never been evaluated in the environments where they are intended to operate.

Bridging these gaps requires a fundamental reorientation of research priorities—away from incremental accuracy improvements on homogeneous internal datasets, and toward the development of AI-ECG systems that are simultaneously powerful, transparent, robust, and clinically validated. This review has identified four concrete and interconnected directions that must define the next generation of AI-ECG research:

1. *Prioritize Inherently Interpretable Architectures*

The field must prioritize inherently interpretable architectures over post hoc explanation methods. Models such as the clinical knowledge-aligned transformer [37], concept bottleneck models, and prototype-based networks represent a paradigm shift toward

systems that generate clinically meaningful explanations as an integral part of their prediction process—not as an afterthought. The ultimate goal should be AI-ECG systems capable of producing natural language rationales that align with established cardiovascular pathophysiology, enabling clinicians to understand, trust, and act upon model outputs.

2. *Develop Unified Multi-modal Fusion and Explanation Frameworks*

The development of unified multi-modal fusion and explanation frameworks is essential. As models increasingly integrate ECG data with EHR, biomarkers, genomics, and imaging—as demonstrated by [34,44]—the explanations they provide must reflect the contributions of all modalities in a coherent and clinically interpretable manner. Fragmented, modality-specific explanations create an incomplete and potentially misleading picture for clinicians. Attention-based architectures offer a promising foundation for unified multi-modal explainability, but substantial methodological innovation is still required.

3. *Establish Standardized Benchmarks for Explainability*

The field urgently needs standardized benchmarks for explainability evaluation. The current absence of agreed-upon metrics for assessing explanation quality across dimensions of fidelity, comprehensibility, and clinical actionability makes it impossible to systematically compare explainability methods or track progress over time. Establishing such benchmarks, informed by clinical validation frameworks such as that proposed in [38], is a prerequisite for the field to move beyond subjective assessments of interpretability.

4. *Conduct Robust Prospective Validation*

Most critically, the value of explainable multi-modal AI-ECG systems must be proven through robust prospective validation measuring real-world patient outcomes. The federated learning frameworks identified in this review [31] offer a pathway to conduct such validation across multiple institutions at scale, while preserving patient privacy and enabling the diverse, representative datasets that generalizability requires. Future trials must measure not only process improvements such as increased diagnosis rates but hard clinical endpoints, including reduced mortality, hospitalizations, and cardiovascular events.

In conclusion, the path forward for AI-ECG is not a singular focus on optimizing predictive accuracy, but on meeting the pivotal and multidimensional challenge of building clinical trust. The most important work to be done lies at the intersection of interpretability, multi-modal integration, standardized evaluation, and prospective clinical validation. AI-ECG systems that successfully navigate this intersection will not merely be high-performing algorithms—they will be trusted clinical partners, capable of transforming the immense promise of artificial intelligence into a cardiovascular medicine that is not only more accurate, but more transparent, more equitable, and more fundamentally human.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/biomedinformatics6040055/s1, PRISMA 2020 Checklist.

Author Contributions: Conceptualization, H.N., R.A. and M.E.; methodology, H.N. and R.A.; software, H.N.; validation, H.N. and R.A.; formal analysis, H.N.; investigation, H.N. and R.A.; resources, H.N.; data curation, H.N. and R.A.; writing—original draft preparation, H.N. and R.A.; writing—review and editing, H.N., R.A. and M.E.; visualization, H.N.; supervision, M.E.; project administration, M.E.; funding acquisition, H.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC was funded by the authors.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article as it is based on a systematic review of previously published literature. All referenced sources are publicly available and cited within the manuscript.

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

Appendix A. Summary of Included Studies

Table A1. Summary of included studies on AI-ECG, multi-modal learning, and explainability ($n = 65$).

Ref	First Author, Year	Objective	Model & Data Modalities	Explainability Method	Key Finding/Performance
[1]	Yu et al., 2022	Predict stroke in real time using multi-modal biosignals.	CNN–LSTM; multi-modal	None	Developed a real-time stroke prediction system.
[20]	Khurshid et al., 2021	Predict incident AF using AI-ECG combined with clinical risk factors.	CNN; ECG + clinical data	Saliency map	Multi-modal model improved risk stratification over single-modality models.
[2]	Yao et al., 2021	Evaluate whether AI-ECG alerts increase low LVEF diagnosis rates.	CNN; ECG only	None (pragmatic trial)	AI alerts significantly increased new low LVEF diagnoses.
[3]	Chen et al., 2025	Develop physiological patch for fall risk.	ML; ECG + IMU	Not specified	Improved fall-risk prediction via sensor fusion.
[7]	Jin et al., 2021	Detect arrhythmia using interpretable attention mechanisms.	DLA-CLSTM; ECG only	Attention mechanism	Highlighted ECG regions most relevant to model decisions.
[8]	Aashiq et al., 2025	Monitor CVD using explainable AI.	ML classifiers; ECG only	Gini index features	Transparent ECG-based monitoring model.
[9]	Lin et al., 2024	Assess mortality impact of AI-ECG alerts in a pragmatic trial.	AI-ECG; ECG only	None (pragmatic trial)	AI-ECG alerts associated with reduced all-cause mortality.
[10]	Wan et al., 2025	Generate automated ECG reports using LLMs.	GPT-2; ECG report text	None (generative)	Demonstrated automated clinical reporting feasibility.
[13]	Krishna et al., 2025	Real-time stroke symptom prediction using wearable biosignals.	CNN–LSTM; ECG + PPG	Semantic interpretation	Proposed a real-time system for elderly monitoring.
[11]	Butler et al., 2024	Predict fatal coronary heart disease using ECG and demographics.	AI-ECG; ECG + demographics	Not specified	Time-dependent AI-ECG showed strong risk stratification.
[39]	Attia et al., 2019	Predict future atrial fibrillation from ECGs in sinus rhythm.	CNN; ECG only	Saliency map	Achieved AUC of 0.87 for identifying patients at risk of future AF.
[54]	Huang et al., 2022	Predict and localize coronary artery disease.	AI-ECG; ECG only	Saliency map	Predicted angiography-proven CAD from ECGs.
[55]	Jung J., 2024	Develop ECG-based user recognition with multi-layered XAI.	Deep learning; ECG only	Multi-layered XAI	Improved interpretability and recognition accuracy.
[56]	Jung J., 2024	Optimize ECG-based user recognition.	Multilayered CNN; ECG only	Multilayered XAI	Improved stability and recognition accuracy.
[21]	Baghersalimi et al., 2023	Decentralized seizure detection on wearables.	Federated learning; ECG + EEG	Privacy-preserving	Enabled cross-institution training without data sharing.
[57]	Hannun et al., 2019	Cardiologist-level arrhythmia detection using deep neural network.	Deep CNN; ECG only	None	Achieved cardiologist-level performance on 12-rhythm classification—AUC 0.97.
[58]	Shashikumar et al., 2018	Detect AF using wearable ECG via deep learning.	Deep learning; ECG only	None	Demonstrated feasibility of wearable-based AF detection.
[29]	Kim et al., 2025	Classify multiple arrhythmias using temporal fusion networks.	Temporal fusion network; ECG only	Attention mechanism	Effectively fused global and local temporal features.
[59]	Wagner et al., 2024	Explaining deep learning for ECG analysis: auditing and knowledge discovery.	CNN; ECG only	Post-hoc XAI (saliency maps, feature attribution, concept-based methods)	Proposed a unified framework for auditing and knowledge discovery in ECG models; covers global and local explainability methods.
[60]	Xie et al., 2024	Design lightweight CNN for ectopic beat classification.	Pruned CNN; ECG only	Structural pruning	High accuracy on resource-constrained devices.

Table A1. *Cont.*

Ref	First Author, Year	Objective	Model & Data Modalities	Explainability Method	Key Finding/Performance
[40]	Jabbour et al., 2024	Predict AF using ECG, clinical variables, and genomics.	Deep learning; ECG + clinical + genomics	Saliency map	Multi-modal approach outperformed single-modality models.
[14]	Dayal et al., 2025	Measure cognitive load using wearable sensors.	ML classifiers; multi-modal	SHAP, LIME	Identified key physiological contributors to cognitive load.
[19]	Ayano et al., 2022	Detect cardiac abnormalities using interpretable ML features.	SVM, RF; ECG only	Gini index features	Achieved 94.4% accuracy without black-box models.
[61]	Frassineti et al., 2024	Forecast neurodevelopmental scores in newborns with sepsis.	SVR; ECG (HRV) only	None	HRV features predicted BAYLEY-III scores at 6 and 12 months.
[48]	Ojha et al., 2024	Compare GradCAM, SHAP, and LIME for ECG interpretability.	CNN; ECG only	GradCAM, SHAP, LIME	Provided comparative evaluation of XAI techniques.
[62]	Yang et al., 2022	Improve seizure detection by fusing ECG and EEG.	Deep learning fusion; ECG + EEG	None	Fusion improved AUC by 6.71% over EEG alone.
[63]	Chakraborty et al., 2023	Estimate blood pressure from PPG signals.	Mathematical model; PPG only	None	Proposed a lightweight BP estimation method for IoT.
[64]	Najafi et al., 2024	Develop VersaSens Edge-AI wearable platform.	Edge-AI; ECG, PPG, IMU	Not specified	Proposed efficient on-device processing for wearables.
[15]	Vinay et al., 2024	Recognize CVD using heart sound signals.	GAN-based BiLSTM; PCG	Feature optimization	Demonstrated feasibility of audio-based CVD detection.
[16]	Degachi et al., 2025	Develop robust heartbeat classification framework.	BiLSTM with attention; ECG only	Self-attention	Improved classification by focusing on salient ECG segments.
[17]	Leicht et al., 2022	Estimate driver state using unobtrusive sensing.	Hybrid sensing; ECG + imaging	Not specified	Validated feasibility of capacitive ECG sensing.
[65]	Harmon et al., 2024	Detect hyperkalemia using AI-ECG.	AI-ECG; ECG only	Not specified	High diagnostic accuracy in emergency settings.
[66]	Kuttala et al., 2023	Detect psychological stress via signal fusion.	Deep learning; ECG, GSR, RESP	Not specified	Multi-modal fusion outperformed single-signal models.
[67]	Zhao et al., 2023	Detect stress using multi-temporal fusion.	Deep learning fusion; multi-modal	Not specified	Improved stress detection accuracy.
[68]	Fu et al., 2024	Optimize ECG network topology using MARL.	Reinforcement learning; ECG topology	None	Introduced novel MARL optimization framework.
[33]	Johnson et al., 2025	Automate ambulatory ECG reporting.	Deep learning; ECG only	Not specified	Reduced physician workload for Holter analysis.
[69]	Perzhilla et al., 2025	Monitor hydration using explainable AI.	ML; ECG only	SHAP	Identified ECG features correlated with hydration.
[30]	Fan et al., 2025	Detect LQTS and AV block using AI-ECG.	AI-ECG; ECG only	Not specified	Validated performance across multiple centers.
[70]	Al-Zaiti et al., 2025	Generate narrative ECG reports from findings.	GPT-2; text data	None (generative)	Translated structured findings into clinical narratives.
[36]	Saraswat et al., 2022	Review XAI opportunities in Healthcare 5.0.	Review; N/A	XAI survey	Outlined future XAI research directions.
[52]	Baker & Xiang, 2023	Survey AIoT for healthcare.	Review; N/A	Not applicable	Outlined architectures integrating ECG sensors.
[71]	Joy et al., 2023	Review AI-ECG for cardiac and extra-cardiac diseases.	Review; N/A	Not applicable	Summarized broad AI-ECG applications.
[72]	Tripathi et al., 2022	Detect insomnia sleep stages from ECG.	Ensemble ML; ECG only	Not specified	Accurate sleep-stage classification from ECG.
[73]	Wang & Leonhardt, 2024	Discuss camera-based health monitoring.	Editorial; N/A	Not applicable	Highlighted fusion with ECG signals.

Table A1. *Cont.*

Ref	First Author, Year	Objective	Model & Data Modalities	Explainability Method	Key Finding/Performance
[74]	Solís-Lemus et al., 2023	Identify barriers and facilitators to AI adoption in cardiology.	Systematic scoping review; N/A	Not applicable	Identified trust, accountability, and workflow integration as key barriers.
[32]	Noseworthy et al., 2022	Test AI-guided AF screening.	AI-ECG; ECG only	None (interventional trial)	Increased new AF diagnoses.
[75]	Al-Zaiti et al., 2022	Review AI methods for ACS detection.	Review; N/A	Not applicable	Identified AI potential in early ACS detection.
[76]	Posan & Richie, 2024	Review AI-ECG in insurance medicine.	Review; N/A	Not applicable	Discussed detection of subclinical conditions.
[77]	Khera R., 2024	Commentary on AI reading scanned ECGs.	Commentary; N/A	Not applicable	Advocated bridging legacy ECG data gaps.
[31]	Bibi et al., 2026	Privacy-preserving multi-institution ECG-based cardiovascular prediction.	Federated learning; ECG multi-modal	Not specified	Demonstrated robust cross-institution generalizability.
[37]	Yisimitila et al., 2025	Interpretable transformer architecture aligned with clinical ECG knowledge.	Transformer; ECG only	Attention + clinical alignment	Improved explainability while maintaining high AUC.
[78]	Gaikwad et al., 2025	Design FPGA SoC for IoMT AI deployment.	Hardware design; N/A	Not applicable	Enabled efficient ECG AI deployment.
[38]	Kehkashan et al., 2026	Clinical validation of SHAP and LIME for cardiovascular prediction.	ML; ECG only	SHAP, LIME	Demonstrated clinical actionability of XAI explanations.
[35]	Alasmari et al., 2025	Federated multi-modal approach for early cardiovascular detection.	Federated learning; ECG + EHR	Not specified	Improved early detection across institutions.
[79]	Celard et al., 2023	Survey generative AI for medical imaging.	Review; N/A	Not applicable	Discussed synthetic data generation potential.
[80]	Gutta & Cheever, 2020	Overview of AI algorithms in biomedicine.	Review; N/A	Not applicable	Summarized foundational AI methods.
[81]	Sweeney-Fanelli et al., 2025	Introduce CUADS multi-modal dataset.	Dataset descriptor; ECG + GSR	Not applicable	Provided public dataset for affective research.
[82]	Site et al., 2023	Predict diabetes using multi-sensor data.	ML; ECG + PPG	Not specified	Demonstrated ECG utility beyond cardiology.
[18]	Ansari et al., 2025	Survey of transformers and LLMs for ECG analysis.	Review; N/A	Not applicable	Comprehensive overview of LLM applications to ECG.
[44]	Khan et al., 2025	Multi-modal AI integrating genomics, imaging, and ECG.	Deep learning; ECG + genomics + imaging	Attention mechanism	Improved cardiovascular risk prediction via data fusion.
[83]	Veeranki et al., 2024	Review affective computing using ECG.	Review; ECG + EDA	Not specified	Summarized emotion detection trends.
[84]	Gliner et al., 2025	Clinically meaningful interpretability of an AI model for ECG classification.	AI; ECG images (scanned 12-lead)	Jacobian-based sensitivity (pixel-level gradient interpretability)	Demonstrated a generic interpretability tool for AI models analyzing digitized 12-lead ECG images; validated on 79,226 labeled scanned ECG images; provides medical-relevance interpretability for both morphological and arrhythmogenic conditions, highlighting features in terms understandable to physicians; high correlation with gold standard interpretations of 3 electrophysiologists.

Table A1. *Cont.*

Ref	First Author, Year	Objective	Model & Data Modalities	Explainability Method	Key Finding/Performance
[34]	Lai et al., 2025	Multimodal AI to forecast arrhythmic death in hypertrophic cardiomyopathy.	Deep learning; ECG + imaging + clinical	Attention mechanism	High AUC for sudden cardiac death prediction—*Nature* paper.
[50]	Taleban et al., 2025	Explainable artificial intelligence in electrocardiography: A systematic review.	Review; ECG XAI	Systematic review of XAI methods	Comprehensive review of ECG XAI methods and evaluation families; covers saliency maps, SHAP, LIME, attention mechanisms, and concept-based approaches; identifies gaps in clinical validation.
[85]	Zubair et al., 2024	Improve imbalanced ECG classification.	Deep learning; ECG only	Augmented attention	Improved performance on imbalanced datasets.

References

1. Yu, J.; Park, S.; Kwon, S.H.; Cho, K.H.; Lee, H. AI-based stroke disease prediction system using ECG and PPG bio-signals. *IEEE Access* **2022**, *10*, 44128–44138. [CrossRef]
2. Yao, X.; Rushlow, D.R.; Inselman, J.W.; McCoy, R.G.; Thacher, T.D.; Behnken, E.M.; Bernard, M.E.; Rosas, S.L.; Akfaly, A.; Misra, A.; et al. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: A pragmatic, randomized clinical trial. *Nat. Med.* **2021**, *27*, 815–819. [CrossRef] [PubMed]
3. Chen, Y.-C.; Hsiao, W.-T.; Hwang, C.-H. Bridging traditional and AI-driven fall risk assessment: A flexible physiological patch for elderly safety. *IEEE Access* **2025**, *13*, 183359–183367. [CrossRef]
4. Siontis, K.C.; Noseworthy, P.A.; Attia, Z.I.; Friedman, P.A. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat. Rev. Cardiol.* **2021**, *18*, 465–478. [CrossRef] [PubMed]
5. Xu, Q.; Li, Y.; Zhu, M.M.; Cai, Y.; Cheng, X.; Wang, W.; Ju, J.; Xu, Y.; Liu, Y.; Liu, Y. Precision cardiovascular medicine with big data and AI. *npj Digit. Med.* **2026**, *9*, 45. [CrossRef] [PubMed]
6. Jandoubi, B.; Akhloufi, M.A. Multimodal artificial intelligence in medical diagnostics. *Information* **2025**, *16*, 591. [CrossRef]
7. Jin, Y.; Liu, J.; Liu, Y.; Qin, C.; Li, Z.; Xiao, D. A novel interpretable method based on dual-level attentional deep neural network for actual multilabel arrhythmia detection. *IEEE Trans. Instrum. Meas.* **2021**, *71*, 2500311. [CrossRef]
8. Aashiq, M.; Hashim, S.J.; Rokhani, F.Z.; Hanafi, M.; Hussein, A.F. TFDGiniXML: A Novel Explainable Machine Learning Framework for Early Detection of Cardiac Abnormalities Based on Nonlinear Time-Frequency Distribution Gini Index Features. *IEEE Access* **2025**, *13*, 86666–86685. [CrossRef]
9. Lin, C.-S.; Liu, W.-T.; Tsai, D.-J.; Lou, Y.-S.; Chang, C.-H.; Lee, C.-C.; Fang, W.-H.; Wang, C.-C.; Chen, Y.-Y.; Lin, W.-S.; et al. AI-enabled electrocardiography alert intervention and all-cause mortality: A pragmatic randomized clinical trial. *Nat. Med.* **2024**, *30*, 1461–1470. [CrossRef] [PubMed]
10. Wan, Z.; Liu, C.; Wang, X.; Tao, C.; Shen, H.; Peng, Z.; Fu, J.; Arcucci, R.; Yao, H.; Zhang, M. MEIT: Multimodal electrocardiogram instruction tuning on large language models for report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 14510–14527. [CrossRef]
11. Butler, L.; Ivanov, A.; Celik, T.; Karabayir, I.; Chinthala, L.; Tootooni, M.S.; Jaeger, B.C.; Patterson, L.T.; Doerr, A.J.; McManus, D.D.; et al. Time-dependent ECG-AI prediction of fatal coronary heart disease: A retrospective study. *J. Cardiovasc. Dev. Dis.* **2024**, *11*, 395. [CrossRef] [PubMed]
12. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef] [PubMed]
13. Krishna, S.; Natarajan, R.; Mahadev, N.; Saravanan, P.; Elavarasi, J.; Thangarasu, N. Edge-AI enabled IoT framework for real-time stroke risk prediction using multimodal biosignals. In *2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA)*; IEEE: New York, NY, USA, 2025; pp. 1–7. [CrossRef]
14. Dayal, A.; Khan, Z.; Ullah, K.; Kim, H.-C. A novel approach to cognitive load measurement using wearable sensors, empirical mode decomposition with machine learning, and explainable AI. *IEEE Access* **2025**, *13*, 17083–17098. [CrossRef]
15. Vinay, N.A.; Vidyasagar, K.N.; Rohith, S.; Pruthviraja, D.; Supreeth, S.; Bharathi, S.H. An RNN-Bi LSTM Based Multi Decision GAN Approach for the Recognition of Cardiovascular Disease (CVD) From Heart Beat Sound: A Feature Optimization Process. *IEEE Access* **2024**, *12*, 65482–65502. [CrossRef]

16. Degachi, O.; El Amraoui, L.; Ouni, K. ECG Heartbeat Classification Using CNN Autoencoder Feature Extraction and Attention-Augmented BiLSTM Classifier. *IEEE Access* **2025**, *13*, 171241–171254. [CrossRef]
17. Leicht, L.; Walter, M.; Mathissen, M.; Hoog Antink, C.; Teichmann, D.; Leonhardt, S. Unobtrusive measurement of physiological features under simulated and real driving conditions. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 4767–4777. [CrossRef]
18. Ansari, M.Y.; Yaqoob, M.; Ishaq, M.; Flushing, E.F.; Mangalote, I.A.C.; Dakua, S.P.; Aboumarzouk, O.; Righetti, R.; Qaraqe, M. A survey of transformers and large language models for ECG diagnosis: Advances, challenges, and future directions. *Artif. Intell. Rev.* **2025**, *58*, 261. [CrossRef]
19. Ayano, Y.M.; Schwenker, F.; Dufera, B.D.; Debelee, T.G. Interpretable machine learning techniques in ECG-based heart disease classification: A systematic review. *Diagnostics* **2022**, *12*, 2154. [CrossRef] [PubMed]
20. Khurshid, S.; Friedman, S.; Reeder, C.; Di Achille, P.; Diamant, N.; Singh, P.; Harrington, L.X.; Wang, X.; Al-Alusi, M.A.; Sarma, G.; et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation* **2022**, *145*, 122–133. [CrossRef] [PubMed]
21. Baghersalimi, S.; Teijeiro, T.; Aminifar, A.; Atienza, D. Decentralized federated learning for epileptic seizure detection in wearables. *IEEE Trans. Mob. Comput.* **2024**, *23*, 6392–6407. [CrossRef]
22. Oke, O.A.; Cavus, N. A systematic review on the impact of artificial intelligence on electrocardiograms in cardiology. *Int. J. Med. Inform.* **2025**, *195*, 105753. [CrossRef] [PubMed]
23. Rahman, M.M.; Rivolta, M.W.; Badilini, F.; Sassi, R. A systematic survey of data augmentation of ECG signals for AI applications. *Sensors* **2023**, *23*, 5237. [CrossRef] [PubMed]
24. Manimaran, G.; Peimankar, A.; Puthusserypady, S.; Momeni, M.; Asyari, R.A.I.; Jahan, M.S.; Moll, J.; Wiil, U.K.; Ebrahimi, A. Explainable deep learning based techniques for ECG-based heart disease classification: A systematic literature review and future direction. *Comput. Biol. Med.* **2025**, *199*, 111324. [CrossRef] [PubMed]
25. Abbas, Q.; Jeong, W.; Lee, S.W. Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. *Healthcare* **2025**, *13*, 2154. [CrossRef] [PubMed]
26. Nazar, M.; Unar, S.; Ahmed, A.; Su'Ud, M.M.; Alam, M.M.; Rahmat, A. Requirements driven explainable artificial intelligence framework for secure and transparent clinical decision support systems. *IEEE Access* **2026**, *14*, 29132–29144. [CrossRef]
27. Banerjee, T.; Paçal, İ. A systematic review of machine learning in heart disease prediction. *Turk. J. Biol.* **2025**, *49*, 600–634. [CrossRef] [PubMed]
28. Grün, D.; Rudolph, F.; Gumpfer, N.; Hannig, J.; Elsner, L.K.; von Jeinsen, B.; Hamm, C.W.; Rieth, A.; Guckert, M.; Keller, T. Identifying heart failure in ECG data with artificial intelligence—A meta-analysis. *Front. Digit. Health* **2020**, *2*, 584555. [CrossRef] [PubMed]
29. Kim, Y.K.; Lee, M.; Song, H.S.; Lee, S.-W. Local-global temporal fusion network with an attention mechanism for multiple and multiclass arrhythmia classification. *IEEE Trans. Syst. Man Cybern. Syst.* **2025**, *55*, 6569–6584. [CrossRef]
30. Fan, S.; Zhang, D.; Wang, Y.; Geng, S.; Lu, K.; Sang, M.; Hong, S. Detecting Long QT Syndrome and First-Degree AV Block using single-lead AI-ECG: A multi-center real-world study. *Heart Rhythm* **2025**, *22*, 1–9.
31. Bibi, M.; Ahmad, R.; Rizwan, A.; Khan, A.N.; Khan, Q.W.; Kim, D.-H. Explainable multi-modal fusion-based federated learning for mortality prediction in energy-constrained healthcare systems. *IEEE Access* **2026**, *14*, 6146–6166. [CrossRef]
32. Noseworthy, P.A.; Attia, Z.I.; Behnken, E.M.; Giblon, R.E.; Bews, K.A.; Liu, S.; Gosse, T.A.; Linn, Z.D.; Deng, Y.; Yin, J.; et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: A prospective non-randomised interventional trial. *Lancet* **2022**, *400*, 1206–1212. [CrossRef] [PubMed]
33. Johnson, L.S.; Zadrozniak, P.; Jasina, G.; Grotek-Cuprjak, A.; Andrade, J.G.; Svennberg, E.; Diederichsen, S.Z.; McIntyre, W.F.; Stavrakis, S.; Benezet-Mazuecos, J.; et al. Artificial intelligence for direct-to-physician reporting of ambulatory electrocardiography. *Nat. Med.* **2025**, *31*, 925–931. [CrossRef] [PubMed]
34. Lai, C.; Yin, M.; Kholmovski, E.G.; Popescu, D.M.; Lu, D.Y.; Scherer, E.; Binka, E.; Zimmerman, S.L.; Chrispin, J.; Hays, A.G.; et al. Multimodal AI to forecast arrhythmic death in hypertrophic cardiomyopathy. *Nat. Cardiovasc. Res.* **2025**, *4*, 891–903. [CrossRef] [PubMed]
35. Alasmari, S.; AlGhamdi, R.; Tejani, G.G.; Kumar Sharma, S.; Mousavirad, S.J. Federated learning-based multimodal approach for early detection of cardiovascular conditions. *Front. Physiol.* **2025**, *16*, 1563185. [CrossRef] [PubMed]
36. Saraswat, D.; Bhattacharya, P.; Verma, A.; Prasad, V.K.; Tanwar, S.; Sharma, G.; Bokoro, P.N.; Sharma, R. Explainable AI for healthcare 5.0: Opportunities and challenges. *IEEE Access* **2022**, *10*, 84486–84517. [CrossRef]
37. Yisimitila, T.; Wang, C.; Hou, M.; Maimaitiniyazi, M.; Aili, Z.; Liu, T.; Tan, H.; Chellamani, N.; Shanmuganathan, M.; Liu, M.; et al. Bridging clinical knowledge and AI: An interpretable transformer framework for ECG diagnosis. *npj Digit. Med.* **2026**, *9*, 41. [CrossRef] [PubMed]
38. Kehkashan, T.; Abdelhaq, M.; Al-Shamayleh, A.S.; Riaz, R.A.; Abdullah, M.; Ahmed, A.I.A.; Akhunzada, A. Sustainable and interpretable heart disease prediction: A clinical decision support approach for biomedical healthcare applications. *Sci. Rep.* **2026**, *16*, 7213. [CrossRef] [PubMed]

39. Attia, Z.I.; Noseworthy, P.A.; Lopez-Jimenez, F.; Asirvatham, S.J.; Deshmukh, A.J.; Gersh, B.J.; Carter, R.E.; Yao, X.; Rabinstein, A.A.; Erickson, B.J.; et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. *Lancet* **2019**, *394*, 861–867. [CrossRef] [PubMed]
40. Jabbour, G.; Nolin-Lapalme, A.; Tastet, O.; Corbin, D.; Jordà, P.; Sowa, A.; Delfrate, J.; Busseuil, D.; Hussin, J.G.; Dubé, M.P.; et al. Prediction of incident atrial fibrillation using deep learning, clinical models, and polygenic scores. *Eur. Heart J.* **2024**, *45*, 481–491. [CrossRef] [PubMed]
41. Noor, N.; Bilal, M.; Abbasi, S.F.; Pournik, O.; Arvanitis, T.N. A novel transformer-based approach for cardiovascular disease detection. *Front. Digit. Health* **2025**, *7*, 1548448. [CrossRef] [PubMed]
42. Ikram, S.; Ikram, A.; Singh, H.; Awan, M.D.A.; Naveed, S.; De la Torre Díez, I.; Gongora, H.F.; Candelaria Chio Montero, T. Transformer-based ECG classification for early detection of cardiac arrhythmias. *Front. Cardiovasc. Med.* **2025**, *12*, 1600855. [CrossRef] [PubMed]
43. Alghieth, M. DeepECG-Net: A hybrid transformer-based deep learning model for real-time ECG anomaly detection. *Sci. Rep.* **2025**, *15*, 20714. [CrossRef] [PubMed]
44. Khan, S.N.; Danishuddin; Khan, M.W.A.; Guarnera, L.; Akhtar, S.M.F. Multi-modal AI in precision medicine: Integrating genomics, imaging, and EHR data for clinical insights. *Front. Artif. Intell.* **2025**, *8*, 1743921. [CrossRef] [PubMed]
45. Zhou, Z.; Wang, J. A narrative review of multimodal data fusion strategies for precision risk prediction in coronary artery disease: Advances, challenges, and future directions. *Rambam Maimonides Med. J.* **2025**, *16*, e0023. [CrossRef] [PubMed]
46. Jain, S.; Wallace, B.C. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1, pp. 3543–3556. [CrossRef]
47. Wiegreffe, A.; Pinter, Y. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 11–20. [CrossRef]
48. Ojha, J.; Haugerud, H.; Yazidi, A.; Lind, P. Exploring interpretable AI methods for ECG data classification. In Proceedings of the ACM Workshop on Intelligent Cross-Data Analysis and Retrieval, Phuket, Thailand, 10–14 June 2024; pp. 11–18. [CrossRef]
49. Khan, M.; Zhihao, Z.; Ayyob, F.; Hussain, A. A privacy-preserving IoMT digital twin: Integrating wearable multimodal sensing and edge-DRL for precision geriatric cardiology. *Spectr. Eng. Manag. Sci.* **2026**, *4*, 1204–1241.
50. Taleban, A.; Alimohammadi, H.; Ghezelbash, F.; Zarei, R.; Mohammadi, A. Explainable artificial intelligence in electrocardiography: A systematic review of methods, applications, and evaluation frameworks. *Biomed. Signal Process. Control* **2026**, *18*, 109325. [CrossRef] [PubMed]
51. Sathi, T.A.; Jany, R.; Azad, A.K.M.; Alyami, S.A.; Alotaibi, N.; Hussain, I.; Hossain, M.A. Explainable federated learning for multi-class heart disease diagnosis via ECG fiducial features. *Diagnostics* **2025**, *15*, 3110. [CrossRef] [PubMed]
52. Baker, S.; Xiang, W. AI of Things for healthcare: A survey of advancements. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 1261–1293. [CrossRef]
53. Martínez-Sellés, M.; Marina-Breysse, M. Cognitive alignment in cardiovascular AI: Designing predictive models that clinicians can understand, trust, and act upon. *Front. Cardiovasc. Med.* **2025**, *12*, 1651324. [CrossRef] [PubMed]
54. Huang, P.S.; Tseng, Y.H.; Tsai, C.F.; Chen, J.; Yang, S.C.; Chiu, F.; Tsai, C.T. An AI-enabled ECG algorithm for the prediction and localization of angiography-proven coronary artery disease. *Diagnostics* **2022**, *10*, 394. [CrossRef] [PubMed]
55. Jung, J. ECG-based user recognition with multi-layered explainable AI. *IEEE Access* **2024**, *12*, 1–10.
56. Jung, J. Optimizing ECG-based user recognition using multilayered CNN and XAI. *IEEE Sens. J.* **2024**, *24*, 1–9.
57. Hannun, A.Y.; Rajpurkar, P.; Haghpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **2019**, *25*, 65–69. [CrossRef] [PubMed]
58. Shashikumar, S.P.; Shah, A.J.; Li, Q.; Clifford, G.D.; Nemati, S. A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. In Proceedings of the 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), Orlando, FL, USA, 16–19 February 2017; pp. 141–144. [CrossRef]
59. Wagner, P.; Strodthoff, N.; Bousseljot, R.-D.; Samek, W.; Schaeffter, T. Explaining deep learning for ECG analysis: Building blocks for auditing and knowledge discovery. *Comput. Biol. Med.* **2024**, *176*, 108525. [CrossRef] [PubMed]
60. Xie, Y.-L.; Lin, X.-R.; Lee, C.-Y.; Lin, C.-W. ARM-based AI module for ectopic beat classification using pruned CNN. *IEEE Sens. J.* **2024**, *24*, 19834–19844. [CrossRef]
61. Frassineti, L.; Bertini, G.; Calà, F.; Dani, C.; Parente, A.; Reyes García, C.A.; Manfredi, C.; Lori, S.; Gabbanini, S.; Lunardi, C.; et al. HRV-Based Regression Analysis in Newborns with Sepsis: Forecasting BAYLEY-III Scores at 6 and 12 Months. *IEEE Access* **2024**, *12*, 99058–99070. [CrossRef]
62. Yang, Y.; Truong, N.D.; Eshraghian, J.K.; Maher, C.; Nikpour, A.; Kavehei, O. A multimodal AI system for out-of-distribution generalization of seizure identification. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 3529–3538. [CrossRef] [PubMed]

63. Chakraborty, A.; Goswami, D.; Mukhopadhyay, J.; Chakrabarti, S. Blood pressure estimation based on photoplethysmography for personalized healthcare. *IEEE Trans. Consum. Electron.* **2023**, *69*, 1195–1203. [CrossRef]
64. Najafi, T.A.; Miranda Calero, J.A.; Thevenot, J.; Duc, B.; Albini, S.; Amirshahi, A.; Taji, H.; Belda Beneyto, M.J.; Affanni, A.; Atienza, D. VersaSens: An Extendable Multimodal Platform for Next-Generation Edge-AI Wearables. *IEEE Trans. Circuits Syst. Artif. Intell.* **2024**, *1*, 83–96. [CrossRef]
65. Harmon, D.M.; Liu, K.; Dugan, J.; Jentzer, J.C.; Attia, Z.I.; Friedman, P.A.; Dillon, J.J. Validation of Noninvasive Detection of Hyperkalemia by Artificial Intelligence–Enhanced Electrocardiography in High Acuity Settings. *Clin. J. Am. Soc. Nephrol.* **2024**, *19*, 952–958. [CrossRef] [PubMed]
66. Kuttala, R.; Subramanian, R.; Oruganti, V.R.M. Multimodal hierarchical CNN feature fusion for stress detection. *IEEE Access* **2023**, *11*, 6867–6878. [CrossRef]
67. Zhao, L.; Niu, X.; Wang, L.; Niu, J.; Zhu, X.; Dai, Z. Stress detection via multimodal multitemporal-scale fusion: A hybrid of deep learning and handcrafted feature approach. *IEEE Sens. J.* **2023**, *23*, 27817–27827. [CrossRef]
68. Fu, Q.; Qiu, T.; Yi, J.; Pu, Z.; Ai, X. Self-clustering hierarchical multi-agent reinforcement learning with extensible cooperation graph. *IEEE Trans. Emerg. Top. Comput. Intell.* **2025**, *9*, 1688–1698. [CrossRef]
69. Perzhilla, L.; Siyoucef, S.; Al-Aslani, R.; Rahman, M.M.U.; Al-Naffouri, T.Y. In-situ dehydration monitoring via a stable diffusion-aided single-lead ECG IoMT: ML/DL models shine while LLMs hallucinate. *IEEE Internet Things J.* **2025**, *12*, 36617–36633. [CrossRef]
70. Al-Zaiti, S.; Macleod, R.; Dam, P.V.; Smith, S.W.; Birnbaum, Y. Narrative ECG report generation from structured findings using GPT-2. *IEEE J. Transl. Eng. Health Med.* **2025**, *13*, 1–8.
71. Joy, S.I.; Kumar, K.S.; Palanivelan, M.; Lakshmi, D. Review on AI in ECG for detection of extra-cardiac and cardiovascular disease. *IEEE Can. J. Electr. Comput. Eng.* **2023**, *46*, 99–106. [CrossRef]
72. Tripathi, P.; Ansari, M.A.; Gandhi, T.K.; Mehrotra, R.; Heyat, M.B.B.; Akhtar, F.; Ukwuoma, C.C.; Muaad, A.Y.; Kadah, Y.M.; Al-Antari, M.A.; et al. Ensemble computational intelligent for insomnia sleep stage detection via the sleep ECG signal. *IEEE Access* **2022**, *10*, 108710–108721. [CrossRef]
73. Wang, W.; Shan, C.; Leonhardt, S.; Mukkamala, R.; Nowara, E. Guest editorial camera-based health monitoring in real-world scenarios. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 595–597. [CrossRef]
74. Solís-Lemus, J.A.; Stramer, B.; Slabaugh, G.; Bhatt, D.L. Barriers and facilitators to the adoption of artificial intelligence in cardiology: A systematic scoping review. *Heart* **2023**, *109*, 1858–1865.
75. Al-Zaiti, S.; Macleod, R.; Dam, P.V.; Smith, S.W.; Birnbaum, Y. Emerging ECG methods for acute coronary syndrome detection: Recommendations & future opportunities. *J. Electrocardiol.* **2022**, *74*, 65–72. [CrossRef] [PubMed]
76. Posan, E.; Richie, R. Unlocking hidden risks: Harnessing AI to detect subclinical conditions from an ECG. *J. Insur. Med.* **2024**, *51*, 64–76. [CrossRef] [PubMed]
77. Khera, R. Commentary: AI reading scanned ECGs—Bridging legacy data gaps. *J. Am. Coll. Cardiol.* **2024**, *83*, 13. [CrossRef] [PubMed]
78. Gaikwad, N.B.; Khare, S.K.; Mendhe, D.; Mir, H.; Kosta, S.; Acharya, U.R. FPGA SoC implementation of adaptive deep neural network based multimodal edge intelligence for Internet of Medical Things. *IEEE Access* **2025**, *13*, 34041–134056. [CrossRef]
79. Celard, P.; Iglesias, E.L.; Sorribes-Fdez, J.M.; Romero, R.; Vieira, A.S.; Borrajo, L. A survey on deep learning applied to medical images: From simple artificial neural networks to generative models. *Neural Comput. Appl.* **2023**, *35*, 2291–2323. [CrossRef] [PubMed]
80. Gutta, S.; Cheever, E. AI algorithms for biomedical applications. *IEEE Potentials* **2020**, *39*, 29–34.
81. Sweeney-Fanelli, T.C.; Ahmed, A.; Imtiaz, M.H. Descriptor: Clarkson University Affective Research Dataset (CUADS). *IEEE Data Descr.* **2025**, *2*, 41–49. [CrossRef]
82. Site, A.; Nurmi, J.; Lohan, E.S. Machine-learning-based diabetes prediction using multisensor data. *IEEE Sens. J.* **2023**, *23*, 28370–28377. [CrossRef]
83. Veeranki, Y.R.; Diaz, L.R.M.; Swaminathan, R.; Posada-Quintero, H.F. Nonlinear signal processing methods for automatic emotion recognition using electrodermal activity. *IEEE Sens. J.* **2024**, *24*, 8079–8093. [CrossRef]
84. Gliner, V.; Levy, I.; Tsutsui, K.; Rav Acha, M.; Schliamser, J.; Schuster, A.; Yaniv, Y. Clinically meaningful interpretability of an AI model for ECG classification. *npj Digit. Med.* **2025**, *8*, 109. [CrossRef] [PubMed]
85. Zubair, M.; Woo, S.; Lim, S.; Kim, D. Deep representation learning with sample generation and augmented attention module for imbalanced ECG classification. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2461–2472. [CrossRef] [PubMed]