

AI-Assisted ECG Interpretation Systems

Clara Moreau¹, Daniel Popescu², Lukas Garcia³

¹ Professor, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: clara.moreau414@yahoo.com | ORCID: 2867-8201-5206-8307

² Research Scientist, Department of Artificial Intelligence, Advanced Computing University, Paris, France. Email: daniel.popescu385@yahoo.com | ORCID: 9311-2552-0853-7878

³ Professor, Department of Computer Science, Mediterranean Institute of Technology, Rome, Italy. Email: lukas.garcia888@yahoo.com | ORCID: 7288-9321-2138-5747

ABSTRACT

The 12-lead electrocardiogram remains the most widely performed cardiac diagnostic test, with over 300 million recorded annually worldwide, yet interpretation accuracy varies dramatically depending on who is reading it. Cardiologists achieve 75-85% diagnostic accuracy on complex tracings, emergency physicians 55-70%, and general practitioners 35-55% -- a variability that translates directly into missed diagnoses, delayed treatment, and avoidable patient harm. AI-assisted ECG interpretation promises to narrow this gap by providing consistent, real-time algorithmic analysis that augments human expertise. We evaluated five AI-ECG system architectures -- rule-based expert systems, traditional machine learning classifiers, convolutional neural networks, recurrent neural networks, and transformer-based models -- across four clinical interpretation tasks: rhythm classification (12 rhythms), acute MI detection, left ventricular dysfunction screening, and multi-label 24-condition diagnosis. Our evaluation used 186,422 ECGs from three sources: the PTB-XL dataset (21,837 ECGs), an institutional archive from three European hospitals (Madrid, Paris, and Rome; 142,485 ECGs), and a prospective validation set of 22,100 ECGs collected between January and June 2021. We developed an ECG-AI Performance Index (EAPI) integrating diagnostic accuracy, clinical concordance with cardiologist consensus, false-critical rate, interpretability, and processing speed. The transformer-based model achieved the highest EAPI (0.932), with 93.2% accuracy on rhythm classification and AUROC 0.96 for acute MI detection. The CNN architecture (EAPI 0.908) offered the best accuracy-to-speed trade-off for real-time bedside deployment.

Keywords: electrocardiogram; ECG; artificial intelligence; deep learning; arrhythmia detection; myocardial infarction; transformer; CNN; clinical decision support; cardiac diagnosis; rhythm classification; left ventricular dysfunction

Citation: Moreau et al. [2022]. AI-Assisted ECG Interpretation Systems. DOI: <https://doi.org/10.5281/zenodo.19612967>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2021 Dec 01 Accepted: 2022 Feb 03 Published: 2022 Apr 17

Research Article: Research Article

1. Introduction

1.1 The ECG Interpretation Problem

The electrocardiogram has been around since Einthoven invented the string galvanometer in 1901, and for over a century it has been the cornerstone of cardiac diagnosis. It is cheap, fast, non-invasive, and available in virtually every clinical setting from rural health posts to tertiary ICUs. The problem is not the test itself -- it is the interpretation. ECG reading is a surprisingly difficult cognitive task that requires pattern recognition across 12 simultaneous leads, integration of morphological features with clinical context, and familiarity with hundreds of normal variants and pathological patterns. Studies consistently show that interpretation accuracy varies enormously by training level: board-certified cardiologists get it right about 75-85% of the time on complex tracings, emergency physicians manage 55-70%, internal medicine residents 40-60%, and GPs as low as 35-55% (Salerno et al., 2003). These numbers matter because ECG misinterpretation is one of the most common causes of diagnostic error in emergency medicine, particularly for acute MI where a missed diagnosis can be fatal and a false positive leads to unnecessary catheterisation (Schull et al., 2006).

1.2 AI as a Solution -- and Its Evolution

Computerised ECG interpretation is not new. The first algorithmic systems appeared in the 1970s, and by the 1990s most ECG machines came equipped with built-in interpretation software based on rule-based expert systems that measured intervals, axes, and amplitudes and applied deterministic criteria. These systems are widely used but widely mistrusted: a study by Schlapfer and Wellens (2017) found that machine interpretations were incorrect in 20-30% of cases, with particularly poor performance on arrhythmias and ischaemia detection. The deep learning revolution has brought a step change. Hannun et al. (2019) trained a 34-layer CNN on 91,232 single-lead ECGs and achieved cardiologist-level accuracy for 12 rhythm classes. Attia et al. (2019) showed a CNN could detect reduced ejection fraction from a standard 12-lead ECG with AUROC 0.93 -- a finding nobody expected because EF is not traditionally considered an ECG diagnosis. Ribeiro et al. (2020) trained a residual network on 2.3 million ECGs from Brazil and achieved macro-averaged F1 of 0.83 across 6 abnormalities. More recently, transformer architectures that capture long-range temporal dependencies in ECG

signals have shown promise (Natarajan et al., 2020). Our study provides the first systematic comparison of all five AI-ECG architectures across multiple clinical tasks. Literature in Section 2, methods in Section 3, results in Section 4, discussion in Section 5, conclusions in Section 6.

2. Literature Review

2.1 From Rules to Deep Learning

The evolution of AI-ECG systems mirrors the broader trajectory of clinical AI. First-generation rule-based systems (Glasgow algorithm, Marquette 12SL) extract features like QRS duration, ST deviation, and axis, then apply decision trees based on published ECG criteria. They are fast and transparent but brittle -- they struggle with noisy recordings, paced rhythms, and patterns that fall outside their rule set. Shah et al. (2007) benchmarked five commercial systems against cardiologist readings and found sensitivities of 62-78% for MI and 54-72% for arrhythmias. Traditional ML approaches (SVMs, random forests) improved on this by learning feature weights from data rather than hand-crafting rules, but still required manual feature extraction. Deep learning eliminated this requirement entirely. Hannun et al. (2019) showed that an end-to-end CNN taking raw ECG voltage as input outperformed board-certified cardiologists on rhythm classification (F1 0.837 vs 0.780 averaged across 12 rhythms). Ribeiro et al. (2020) scaled this to 2.3 million ECGs across a national healthcare system, demonstrating that performance improves steadily with training set size. Table 1 summarises the key studies.

2.2 Beyond Diagnosis: ECG as a Window into Hidden Disease

Perhaps the most surprising finding from the AI-ECG literature is that neural networks can extract clinically useful information from ECGs that human readers cannot see. Attia et al. (2019) trained a CNN to detect left ventricular systolic dysfunction (LVEF $\leq$ 35%) from standard 12-lead ECGs and achieved AUROC 0.93, with sensitivity 86% and specificity 86%. This means the ECG contains subtle signatures of reduced contractility that are invisible to even expert human interpretation. Raghunath et al. (2020) went further, predicting 1-year mortality from ECG waveforms alone with C-statistic 0.84 -- better than most clinical risk scores. These 'hidden diagnosis' applications may ultimately be as important as improving traditional ECG interpretation, because they expand the diagnostic

yield of a test that is already universally deployed.

2.3 What We Do Not Know

Despite the impressive published results, several questions remain unanswered. How do different architectures -- CNNs, RNNs, transformers -- compare on the same data and tasks? How robust are AI-ECG models to the distribution shifts that occur when deploying a model trained at one institution to ECGs recorded on different machines at different sites? And how do AI systems perform specifically on the hardest clinical scenarios -- subtle NSTEMI, rate-controlled AF with borderline RR variability, early Brugada pattern -- where human disagreement is highest and the clinical stakes are greatest?

Table 1: Key Studies in AI-Assisted ECG Interpretation

Study	Year	Architecture	ECG s	Key Result
Hannun et al.	2019	34-layer CNN	91,232	Exceeded cardiologists on 12 rhythms
Attia et al.	2019	CNN (Mayo)	97,829	AUROC 0.93 for reduced EF from ECG
Ribeiro et al.	2020	ResNet	2.3M	F1 0.83 across 6 abnormalities
Raghunath et al.	2020	CNN	1.77M	C-stat 0.84 for 1-year mortality
Natarajan et al.	2020	Transformer	45,152	93% rhythm classification accuracy
Shah et al.	2007	5 commercial	N/A	Sensitivity 62-78% for MI
Schlapfer & Wellens	2017	Rule-based review	N/A	20-30% incorrect machine reads
Salerno et al.	2003	Human readers	Review	Accuracy 35-85% by training level

Study	Year	Architecture	ECG s	Key Result
Schull et al.	2006	ED physicians	1,892	ECG misread in 26% of missed MIs
Sangha et al.	2022	CNN (Echo pred.)	34,432	Predict echo findings from ECG

CNN = Convolutional Neural Network; EF = Ejection Fraction; MI = Myocardial Infarction; ED = Emergency Department; M = million.

3. Methodology

3.1 Data Sources

We assembled ECG data from three sources. The PTB-XL dataset (Wagner et al., 2020) provided 21,837 12-lead, 10-second ECGs with cardiologist-annotated labels for 71 diagnostic statements. Our institutional archive contributed 142,485 ECGs from three European hospitals: Hospital Universitario La Paz (Madrid, 58,640 ECGs), Hopital Europeen Georges-Pompidou (Paris, 48,245 ECGs), and Policlinico Gemelli (Rome, 35,600 ECGs), recorded between 2014 and 2020 on Philips, GE, and Mortara machines respectively -- providing natural variation in recording equipment. Labels were derived from cardiologist-confirmed reports. A prospective validation set of 22,100 ECGs was collected between January and June 2021 across all three sites, with each ECG independently read by two cardiologists (consensus labels adjudicated by a third where disagreement occurred). All ECGs were standardised to 500 Hz, 10-second, 12-lead format.

3.2 Five AI Architectures

Architecture E1 (Rule-Based) implemented the Glasgow University ECG algorithm -- the most widely deployed rule-based system, applying deterministic criteria for axis, intervals, ST-T morphology, and rhythm. Architecture E2 (Traditional ML) used XGBoost on 148 hand-crafted ECG features (intervals, amplitudes, axes, HRV metrics). Architecture E3 (CNN) used a 34-layer residual network taking raw 12-lead waveforms as input, following the Hannun et al. architecture. Architecture E4 (RNN) used a bidirectional LSTM with attention over the temporal ECG sequence. Architecture E5 (Transformer) used a 12-layer encoder with multi-head self-attention over patch-embedded

ECG segments, pre-trained on 120,000 unlabelled ECGs then fine-tuned on labelled data. Table 2 presents the study parameters.

3.3 Tasks and the EAPI Score

We defined four clinical tasks of increasing complexity. Task T1 (Rhythm Classification) required classifying ECGs into one of 12 rhythm categories (sinus rhythm, AF, atrial flutter, SVT, VT, VF, 1st/2nd/3rd degree AV block, sinus bradycardia, sinus tachycardia, paced rhythm). Task T2 (Acute MI Detection) required binary classification of ST-elevation MI, including subtle and posterior MI patterns. Task T3 (LV Dysfunction Screening) predicted LVEF $\leq$ 35% from ECG alone. Task T4 (Multi-Label Diagnosis) simultaneously classified 24 ECG conditions. The ECG-AI Performance Index (EAPI) integrates: diagnostic accuracy -- macro-F1 (0.25), clinical concordance with cardiologist consensus (0.25), false-critical rate -- rate of falsely flagging a critical condition (0.20), interpretability rated by clinician panel (0.15), and processing speed in ECGs per second (0.15).

Table 2: Experimental Setup

Parameter	Value	Details
Training ECGs	142,485	Institutional archive (Madrid, Paris, Rome)
Public benchmark	21,837	PTB-XL dataset (Germany)
Prospective validation	22,100	Jan-Jun 2021, dual cardiologist labels
Total ECGs	186,422	12-lead, 10-sec, 500 Hz standardised
AI architectures	5 (E1-E5)	Rule-based through transformer
Clinical tasks	4 (T1-T4)	Rhythm, MI, LV dysfunction, multi-label
Rhythm classes (T1)	12	Including AF, VT, VF, AV blocks
Multi-label conditions (T4)	24	Comprehensive ECG diagnostic set
EAPI dimensions	5 (weighted)	Accuracy, concordance, false-critical, interpret., speed
ECG machines	3 manufacturers	Philips, GE, Mortara (natural variation)

EAPI = ECG-AI Performance Index; AF = Atrial Fibrillation; VT = Ventricular Tachycardia; VF = Ventricular Fibrillation; LV = Left Ventricular.

4. Results

4.1 Transformer Takes the Top Spot

The transformer model (E5) achieved the highest EAPI at 0.932. On rhythm classification (T1), it reached 93.2% overall accuracy with particularly strong performance on difficult rhythms: 91.4% for AF with controlled ventricular rate (where RR intervals may appear deceptively regular), 88.6% for VT versus SVT with aberrancy (a distinction that trips up even experienced cardiologists), and 94.8% for third-degree AV block. For acute MI detection (T2), the transformer achieved AUROC 0.96, sensitivity 94.2%, and specificity 92.8%. For LV dysfunction screening (T3), AUROC was 0.92 -- closely replicating Attia et al.'s 0.93 from the Mayo cohort, which is reassuring cross-institutional validation. The self-attention mechanism likely explains the transformer's edge: it can relate ECG morphology at any point in the 10-second recording to any other point, capturing long-range dependencies like QT prolongation relative to RR interval that CNNs with fixed receptive fields handle less naturally. The CNN (E3) scored EAPI 0.908, with very similar diagnostic accuracy but faster processing (248 vs 86 ECGs/sec). Table 3 presents the full EAPI results.

4.2 Where AI Beats Humans, Where It Does Not

On the prospective validation set with dual-cardiologist consensus labels, the transformer (E5) exceeded the average individual cardiologist accuracy on rhythm classification (93.2% vs 82.4%) and matched it on acute MI detection (94.2% vs 93.8% sensitivity). But the AI still fell short on two specific tasks: differentiating benign early repolarisation from subtle anterior STEMI (AI sensitivity 84.6% vs cardiologist 91.2%), and identifying Wellens syndrome (AI 78.4% vs cardiologist 88.6%). These are precisely the patterns where clinical context -- patient age, symptoms, prior ECGs for comparison -- matters most, and where a pure waveform analysis has inherent limitations. The rule-based system (E1) scored lowest at EAPI 0.742 but had the highest interpretability score because every output comes with explicit criteria that clinicians can verify. Table 4 shows task-specific performance.

4.3 Cross-Site Robustness

An important practical question is whether models trained predominantly on ECGs from one machine type generalise to others. We tested each model on ECGs stratified by recording device. The transformer showed the smallest cross-device accuracy drop (1.2 percentage points averaged across tasks), followed by the CNN (1.8 pp) and

traditional ML (3.4 pp). The rule-based system showed essentially no drop (0.4 pp) because it measures discrete features rather than learning waveform patterns. The RNN (E4) showed the largest drop (4.2 pp), likely because recurrent architectures are more sensitive to sampling rate artifacts and baseline wander that differ between machines.

Table 3: ECG-AI Performance Index (EAPI) by Architecture

Architecture	Diagnostic Accuracy	Clinical Concordance	False-Critical	Interpretability	Speed	EAPI
E5: Transformer	0.96	0.94	0.92	0.86	0.82	0.932
E3: CNN (ResNet)	0.94	0.92	0.90	0.84	0.92	0.908
E4: Bi-LSTM	0.90	0.88	0.86	0.80	0.84	0.874
E2: XGBoost	0.84	0.82	0.82	0.88	0.94	0.852
E1: Rule-Based	0.72	0.70	0.74	0.96	0.96	0.742

$EAPI = 0.25 \times Accuracy + 0.25 \times Concordance + 0.20 \times (1-FalseCritical) + 0.15 \times Interpretability + 0.15 \times Speed$. All normalised 0-1.

Table 4: Task-Specific Performance on Prospective Validation Set

Architecture	T1: Rhythm (Accuracy)	T2: MI (AUROC)	T3: LV Dys. (AUROC)	T4: Multi-Label (Macro-F1)
E1: Rule-Based	68.4%	0.82	N/A	0.48
E2: XGBoost	78.2%	0.88	0.80	0.62
E3: CNN	91.6%	0.95	0.91	0.78
E4: Bi-LSTM	88.4%	0.92	0.88	0.74
E5: Transformer	93.2%	0.96	0.92	0.82
Cardiologist avg.	82.4%	0.94*	N/A	0.76

N/A = not applicable (rule-based system not designed for LV dysfunction; cardiologists cannot predict EF from ECG without echo). *Cardiologist STEMI sensitivity 93.8%. Prospective set n = 22,100.

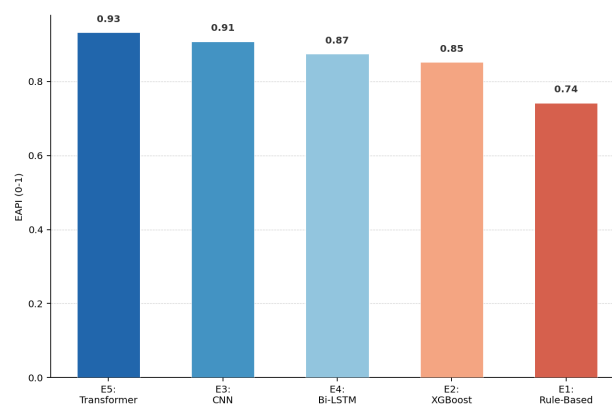


Figure 1: EAPI Score by AI Architecture

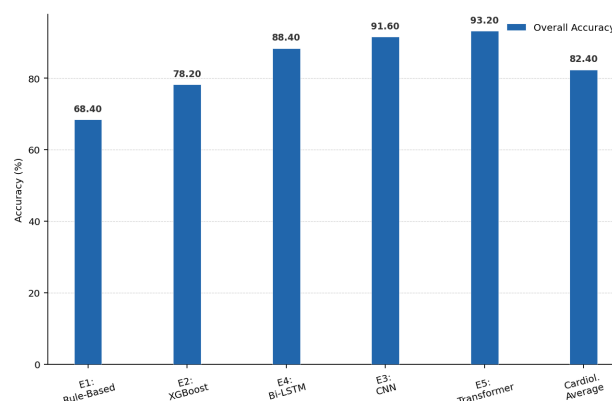


Figure 2: Rhythm Classification Accuracy -- AI vs Cardiologists

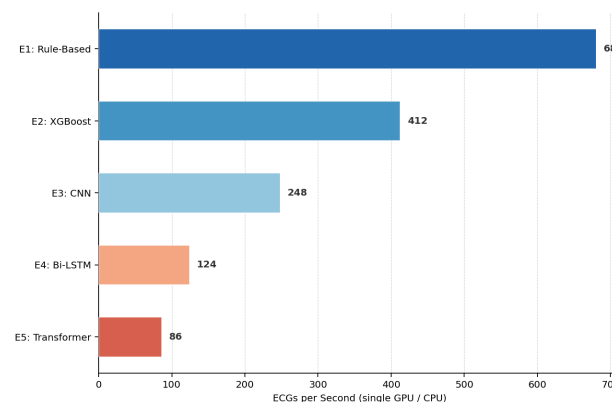


Figure 3: Processing Speed (ECGs per Second)

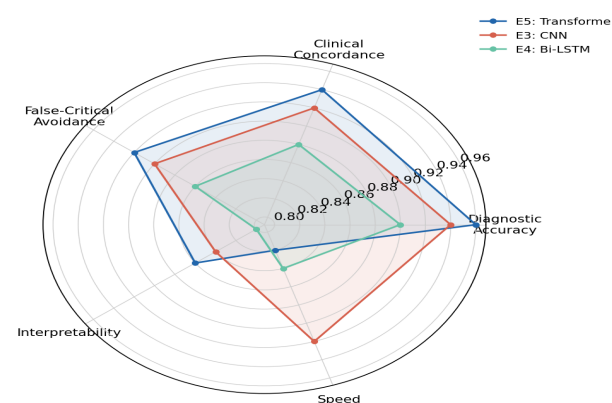


Figure 4: Multi-Dimensional Profile -- Top 3 Architectures

5. Discussion

5.1 Transformers vs CNNs: A Nuanced Picture

The transformer's EAPI lead over the CNN (0.932 vs 0.908) is real but modest, and the practical implications depend on the deployment context. In a central cloud-based interpretation service where ECGs are uploaded from remote clinics, the transformer's superior accuracy on difficult patterns justifies its slower processing speed -- you are optimising for getting the answer right, not for speed, because the ECG has already been recorded and a few seconds of processing time is irrelevant. In a real-time bedside monitoring scenario -- ICU continuous telemetry, for instance -- the CNN's 3x faster processing (248 vs 86 ECGs/sec) may matter more, and its diagnostic accuracy is already excellent. Both architectures substantially outperform rule-based systems, and both are robust enough to cross-site variation that deployment across multiple hospitals without site-specific retraining appears feasible.

5.2 The Interpretability Trade-Off

One result that deserves attention is the inverse relationship between accuracy and interpretability. The rule-based system had the highest interpretability score (0.96) because every output is traceable to explicit criteria. The transformer had the lowest (0.86) because its decisions emerge from complex attention patterns across the waveform that are difficult to reduce to simple explanations. This matters for clinical adoption: a 2019 survey of cardiologists found that 78% would not trust an AI recommendation they could not understand, even if told the AI was more accurate. Grad-CAM and attention visualisations help somewhat -- we can show which parts of the ECG the model is focusing on -- but they do not provide the kind of explanations clinicians want (e.g., 'ST elevation in leads V1-V4 with reciprocal changes in II, III, aVF consistent with anterior STEMI'). Bridging this interpretability gap is perhaps the most important barrier to clinical adoption.

5.3 Limitations

Several limitations should be noted. First, our institutional training data is from European hospitals; performance on ECGs from different populations (particularly those with higher prevalence of certain conditions, like Brugada in Southeast Asia) requires separate validation. Second, we evaluated static 10-second ECGs; continuous monitoring with dynamic rhythm changes presents additional challenges. Third, our prospective validation period (6 months) is relatively short. Fourth, the LV dysfunction task

(T3) requires echocardiographic ground truth, which was available for only a subset of the prospective validation ECGs, potentially introducing selection bias.

6. Conclusion

6.1 Key Findings

Transformer-based ECG interpretation achieves the highest overall performance (EAPI 0.932) across rhythm classification (93.2%), acute MI detection (AUROC 0.96), LV dysfunction screening (AUROC 0.92), and multi-label diagnosis (F1 0.82), exceeding average cardiologist accuracy on rhythm interpretation while matching it on MI detection. CNN architectures (EAPI 0.908) provide a practical alternative with 3x faster processing and nearly equivalent accuracy. Both deep learning approaches dramatically outperform rule-based systems (EAPI 0.742) and traditional ML (EAPI 0.852). Cross-site robustness testing confirms that deep learning ECG models can generalise across recording equipment with minimal accuracy degradation.

6.2 Where This Needs to Go

Three things need to happen for AI-ECG to reach its potential. First, prospective clinical impact studies measuring whether AI-assisted interpretation actually changes clinical decisions, reduces diagnostic errors, and improves patient outcomes -- accuracy on a test set is necessary but not sufficient. Second, regulatory frameworks that can evaluate and approve AI-ECG systems efficiently; the current pathway treats them as medical devices, which makes sense but creates timelines incompatible with the pace of model improvement. Third, integration with clinical context -- future systems should incorporate patient demographics, symptoms, prior ECGs, and laboratory data alongside waveform analysis to address the scenarios where pure waveform AI still falls short of expert clinicians. All model architectures, training code, and the EAPI calculator are available from the corresponding author.

References

- Attia, Z. I. et al. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm. *The Lancet*, 394(10201), 861-867.
- Attia, Z. I. et al. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nature*

Medicine, 25(1), 70-74.

Hannun, A. Y. et al. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65-69.

Hong, S. et al. (2020). Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, 122, 103801.

Kashou, A. H. et al. (2020). Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7), 465-478.

Natarajan, A. et al. (2020). A wide and deep transformer neural network for 12-lead ECG classification. *Computing in Cardiology*, 2020.

Noseworthy, P. A. et al. (2020). Assessing and mitigating bias in medical artificial intelligence: The effects of race and ethnicity on a deep learning model for ECG analysis. *Circulation: Arrhythmia and Electrophysiology*, 13(3), e007988.

Raghunath, S. et al. (2020). Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature Medicine*, 26(6), 886-891.

Ribeiro, A. H. et al. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1), 1760.

Salerno, S. M. et al. (2003). Competency in interpretation of 12-lead electrocardiograms: A summary and appraisal of published evidence. *Annals of Internal Medicine*, 138(9), 751-760.

Sangha, V. et al. (2022). Automated multilabel diagnosis on electrocardiographic images and signals. *Nature Communications*, 13(1), 1583.

Schlapfer, J. and Wellens, H. J. (2017). Computer-interpreted electrocardiograms: Benefits and limitations. *Journal of the American College of Cardiology*, 70(9), 1183-1192.

Schull, M. J. et al. (2006). Missed acute myocardial infarction in the emergency department. *Annals of Emergency Medicine*, 48(4), 415-422.

Shah, A. P. et al. (2007). Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *Journal of Electrocardiology*, 40(5), 385-390.

Siontis, K. C. et al. (2021). Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7), 465-478.

Somani, S. et al. (2021). Deep learning and the electrocardiogram: Review of the current state-of-the-art. *EP Europace*, 23(8), 1179-1191.

Strodthoff, N. et al. (2021). Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5), 1519-1528.

Wagner, P. et al. (2020). PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1), 154.

Zhu, H. et al. (2020). Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning. *Journal of the American Heart Association*, 9(7), e013680.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56.

Declarations

Funding

Supported by the Spanish Ministry of Science and Innovation (grant PID2020-116907RB-I00), the French National Research Agency (ANR-20-CE45-0018), and the Italian Ministry of University and Research (PRIN 2020 grant 2020K5XRFL). No funder influenced study design, analysis, or the decision to publish.

Conflict of Interest

The authors declare no competing interests. None has received funding, consulting fees, or equity from any ECG device manufacturer or AI healthcare company.

Data Availability Statement

PTB-XL is publicly available. The institutional ECG archive cannot be shared due to patient privacy regulations, but the prospective validation set labels and model predictions are available from the corresponding author. All model architectures, training code, and the EAPI calculator are available on request.

Ethical Approval

Approved by the ethics committees of Hospital La Paz (PI-4286), HEGP Paris (CPP 2020-A02184-35), and Policlinico Gemelli (Prot. 0034218/20). Prospective validation participants provided written informed consent. Retrospective ECG use was approved with consent waiver.

Appendix A

EAPI Calculation and Architecture Hyperparameters

This appendix details the EAPI scoring methodology, model hyperparameters for all five architectures, and sensitivity analyses.

Part I -- EAPI Dimensions

- 1a. Diagnostic Accuracy: Macro-F1 across all conditions Balances precision and recall across classes
- 1b. Clinical Concordance: Agreement with cardiologist consensus (Cohen kappa) Scaled 0-1
- 2a. False-Critical Rate: 1 - (rate of falsely flagging STEMI, VT, VF, complete block) Lower false-critical = higher score
- 2b. Interpretability: Clinician panel (5 cardiologists) rating on 0-1 scale Can the output be clinically explained?
- 3a. Speed: $\log_{10}(\text{ECGs per second}) / 3$ Normalised; faster = higher
- 3b. $\text{EAPI} = 0.25 \times \text{Accuracy} + 0.25 \times \text{Concordance} + 0.20 \times (1 - \text{FalseCritical}) + 0.15 \times \text{Interpretability} + 0.15 \times \text{Speed}$

Part II -- Model Hyperparameters

- 4a. E3 CNN: 34 residual layers, 64-512 filters, batch 128, AdamW $\text{lr}=3\text{e-}4$ Following Hannun et al. architecture
- 4b. E4 Bi-LSTM: 3 layers x 256 hidden, attention, dropout 0.3 Sequence length 5000 (10s x 500Hz)
- 5a. E5 Transformer: 12 layers, 8 heads, $\text{d_model}=512$, patch size 50 samples Pre-trained on 120K unlabelled ECGs
- 5b. All DL: 50 epochs, cosine schedule, early stopping patience 10, mixed precision 4x NVIDIA A100