

DEEP LEARNING FOR ECG INTERPRETATION: A SYSTEMATIC REVIEW OF MODEL ARCHITECTURES, TRANSFER LEARNING, AND POPULATION BIAS IN AFRICAN CONTEXTS

Zainab Hussain Jibril^{*1}, Bilkisu Larai Muhammad-Bello², Aliyu Sulaiman Muhammed¹

¹Department of Computer Science, Nile University of Nigeria, Abuja, Nigeria

²Department of Software Engineering, Nile University of Nigeria, Abuja, Nigeria

*Corresponding Author Email Address: zainab.hussain@nileuniversity.edu.ng

ABSTRACT

Automated ECG interpretation has been substantially advanced by deep learning, with several contemporary models now approaching the diagnostic accuracy of trained cardiologists. Despite this progress, a critical gap remains: virtually all existing models have been developed using data drawn from European, North American, or Asian patient populations. When these systems are deployed in African clinical settings, questions of generalizability and embedded population bias carry urgent clinical consequences. This systematic review pursued three aims: to compare the performance of major deep learning architectures on ECG interpretation tasks; to evaluate the practical utility of transfer learning in data-scarce environments; and to examine how thoroughly the existing literature accounts for population-level demographic differences, with particular attention to African patients. The review adhered to PRISMA 2020 guidelines. Searches were conducted across five databases: PubMed/MEDLINE, IEEE Xplore, Scopus, Web of Science, and Google Scholar, covering peer-reviewed work published between 2020 and 2025. From an initial pool of 1,847 records, 42 studies met all inclusion criteria following rigorous two-stage screening. Eligible studies were required to apply deep learning to ECG classification and report quantitative performance metrics. Extracted data encompassed model architecture, dataset origin and composition, preprocessing pipelines, validation methodology, and demographic reporting practices. Study quality was evaluated using a modified QUADAS-2 framework. Among the architectures examined, classical convolutional neural networks (CNNs) were most frequently used, followed by hybrid CNN-RNN models, ResNet variants, transfer learning pipelines, and attention-based designs. ResNet-based architectures consistently achieved the strongest performance on benchmark datasets such as MIT-BIH and PTB-XL. Transfer learning produced meaningful accuracy improvements, particularly where target datasets were small, suggesting it may offer a viable route forward for resource-limited settings. However, a number of methodological limitations were identified across the literature: limited external validation, inconsistent signal preprocessing, inadequate demographic reporting, and a complete absence of race- or ethnicity-stratified performance analyses. No included study used an exclusively African ECG dataset, and none evaluated model performance on African patients. This is especially concerning given well-documented differences in ECG parameters across ethnic groups. Black populations demonstrate higher rates of T-wave inversion and elevated Sokolow-Lyon voltages relative to standard reference ranges, differences that directly affect how AI models interpret cardiac tracings. In summary, while deep learning, particularly

ResNet architectures augmented by transfer learning, delivers strong classification performance on benchmark data, the total exclusion of African populations raises genuine concerns about clinical fairness and real-world applicability. This review maps those gaps and offers a concrete methodological framework for developing ECG interpretation systems that are both population-aware and fit for use in African healthcare settings.

Keywords: Electrocardiogram; Deep Learning; Convolutional Neural Networks; Transfer Learning; Systematic Review; Population Bias; African Populations.

INTRODUCTION

Global Burden of Cardiovascular Disease and the Role of ECG

Cardiovascular disease remains, by every available measure, the leading cause of death worldwide. Each year it claims approximately 17.9 million lives, nearly one in three deaths globally (World Health Organization, 2021). The burden is not evenly shared: low- and middle-income countries account for more than three-quarters of all CVD-related deaths, yet these are the settings with the fewest diagnostic and therapeutic resources (Yuyun et al., 2020). Across sub-Saharan Africa, including Nigeria, the situation is deteriorating. Rapid urbanization, dietary transitions, and increasingly sedentary lifestyles are driving a sharp rise in CVD prevalence that existing health systems are poorly equipped to manage (Androulakis & Fielder, 2023).

At the centre of cardiac diagnosis sits the electrocardiogram (ECG). As the most widely deployed non-invasive tool in cardiology, the ECG can detect arrhythmias, myocardial ischaemia, conduction abnormalities, and structural heart disease quickly, affordably, and without radiation exposure (Rafie et al., 2021). In principle, the ECG is ideally suited to resource-limited settings. In practice, however, the technology is only as useful as the clinician reading it.

This is where Africa faces a profound structural deficit. Accurate ECG interpretation requires years of specialist training, yet cardiologist density across much of the continent is extremely low. In Nigeria, there is fewer than one cardiologist per 100,000 people compared to more than 50 per 100,000 in high-income countries (Mulure et al., 2024). That gap, compounded by the vast distances between urban centres and rural communities where the majority of the population lives, means that cardiac abnormalities routinely go undetected or are diagnosed too late. The human cost is substantial: preventable deaths, avoidable complications, and entrenched cycles of poor cardiovascular outcomes.

Deep Learning for Automated ECG Interpretation

Over the past decade, deep learning has fundamentally altered what automated ECG analysis can achieve. Earlier computer-aided tools depended on manually engineered signal features and hand-crafted rules. Deep learning models, by contrast, learn meaningful representations directly from raw or minimally processed ECG signals, eliminating the need for expert-designed feature extraction (Sun et al., 2021). This shift has raised a genuine possibility: that accurate, specialist-level cardiac diagnosis might one day be delivered at scale, even in settings where specialist cardiologists are scarce.

Convolutional Neural Networks (CNNs) are well suited to ECG waveforms because they detect local morphological patterns the shape of a P wave, the slope of an ST segment, the morphology of a QRS complex by sliding learned filters across the signal (Ansari et al., 2023).

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, approach the ECG differently, following the signal through time and modelling how successive beats relate to one another. This makes them especially effective for rhythm abnormalities and temporal dependencies that unfold over longer recording windows (Xu et al., 2020).

Residual Networks (ResNets) address a fundamental engineering challenge: as neural networks grow deeper, they become progressively harder to train and prone to performance degradation. ResNets overcome this by introducing skip connections shortcut pathways that carry information across layers, enabling networks with dozens of layers to be trained reliably (Mavaddati, 2025).

Hybrid CNN-RNN models combine the strengths of both approaches: CNNs handle morphological feature extraction, while RNNs model sequential dynamics of the rhythm. This combination has proven particularly effective for multi-label classification tasks where both waveform shape and rhythm pattern are diagnostically relevant (Obeidat & Alqudah, 2021).

Transfer learning takes a different starting point. Rather than training a model from scratch on a target dataset, transfer learning begins with a model already pre-trained on a large source dataset such as PTB-XL and fine-tunes it on a smaller local dataset. This reduces the volume of labelled data required to achieve strong performance, which is precisely why it holds promise for data-scarce settings like sub-Saharan Africa (Weimann & Conrad, 2021).

Taken together, these architectures represent a powerful and expanding toolkit. However, a critical limitation runs across all of them. The public ECG datasets on which these models are trained are drawn almost exclusively from European, North American, and Asian populations (Zheng et al., 2025). There is a widespread and largely untested assumption that models performing well on these benchmarks will generalise equally well to other ethnic groups. That assumption is now under scrutiny, challenged by a growing body of evidence documenting meaningful racial and ethnic variation in ECG parameters.

Evidence for Racial Differences in ECG Patterns

A substantial body of evidence documents that ECG characteristics

are not uniform across ethnic groups. Multiple studies have identified meaningful differences in ECG parameters between African and Black populations compared to Caucasian populations differences are not attributable to underlying heart disease and likely reflect a combination of genetic, anatomical, and physiological factors (Ozo & Sharma, 2020).

Some of the most persuasive evidence comes from studies of athletes who are rigorously screened, generally healthy, and free from the confounding influence of illness or medication. Davis et al. (2022) synthesised data from 14 studies covering 11,745 athletes and documented striking interethnic differences in T-wave patterns. T-wave inversion was observed in 2.6–22.8% of Black athletes, compared to just 0–5% of Caucasian athletes. Anterior T-wave inversions were even more pronounced, at 4.3–18.7% in Black athletes versus fewer than 1% in Caucasians, while lateral T-wave inversions showed a similarly wide disparity.

The clinical implications are not abstract. Standard interpretation criteria, calibrated against Caucasian reference populations, would flag many of these healthy Black athletes as potentially harbouring cardiomyopathy, generating unnecessary investigations, unwarranted anxiety, and, in some cases, disqualification from sport, in individuals with entirely normal hearts.

Population-based studies tell the same story. The TAHES study examined 997 rural African adults in Benin to establish normal ECG values in an African community (Amidou et al., 2023). The findings diverged substantially from widely applied reference ranges. The upper limit of normal for Sokolow-Lyon voltage in African men was 47 mm, compared to the standard cut-off of 35 mm; in women, 38 mm, still above the standard threshold. QTc intervals were correspondingly longer. ST-segment elevation was present in more than 60% of Black individuals in this cohort, compared to roughly 25% of Caucasians (Ozo & Sharma, 2020). If Caucasian-derived voltage criteria are applied unchanged to African patients, up to 15–20% of healthy African men would be incorrectly diagnosed with left ventricular hypertrophy, a false-positive rate with real clinical costs (Amidou et al., 2023).

The Problem of Population Bias in Deep Learning ECG Models

The evidence reviewed above carries a direct and uncomfortable implication for how current deep learning ECG models will behave in African clinical settings. A model trained exclusively on non-African data does not simply lack knowledge about African ECG patterns; it actively internalises the wrong baselines. It encodes what is normal for European or Asian populations and treats those distributions as universal. When that model encounters an African patient, the mismatch between what it has learned and what it sees generates systematic error.

Two distinct failure modes follow, both clinically consequential. The first is the misclassification of normal variants as pathological: healthy African patients commonly present with high QRS voltages, T-wave inversions in anterior leads, and ST-segment elevation, all of which fall within normal range for their population but exceed the thresholds a non-African-trained model would recognise as benign. Such a model would flag these findings as abnormal, triggering unnecessary investigations and unwarranted diagnoses. The second failure mode is the inverse: genuine abnormalities that manifest differently in African populations may go undetected,

resulting in false negatives with potentially serious consequences for patient care.

Both errors feed into a broader structural problem: if an AI diagnostic tool performs reliably for European patients but poorly for African patients, deploying it in African healthcare settings does not merely fail to help; it actively entrenches health disparities by delivering worse care to those already facing the greatest diagnostic barriers.

Rationale, Objectives, and Research Questions

Cardiovascular disease is rising sharply across sub-Saharan Africa. The diagnostic infrastructure to meet that rise is seriously overstretched. Deep learning offers a genuine pathway to scalable, affordable ECG interpretation. However, every model currently in existence has been built on data that excludes African patients, and none has been tested on an African cohort. This mismatch between where the technology is developed and where it is most urgently needed is not a minor oversight; it is a structural problem with direct consequences for clinical equity.

This systematic review was designed to address that problem. Five objectives guided the work: (1) mapping the performance landscape of current deep learning architectures for ECG classification; (2) examining how much performance gain transfer learning provides, particularly in data-scarce settings; (3) scrutinising the literature for how well or how poorly it accounts for population diversity; (4) translating independent epidemiological evidence of racial ECG differences into a concrete assessment of risk when non-African tools are applied to African patients; and (5) offering a practical methodological framework for developing ECG interpretation systems genuinely fit for Nigerian and African healthcare settings.

Four research questions anchor the review:

- i. **RQ1:** Among the deep learning architectures applied to ECG classification, which achieve the highest accuracy and F1 scores, and what are their relative strengths?
- ii. **RQ2:** How much does transfer learning improve model performance compared to training from scratch, particularly when the target dataset is small?
- iii. **RQ3:** What does the existing evidence reveal about population-specific ECG patterns, and to what degree have deep learning studies engaged with or ignored this issue?
- iv. **RQ4:** What methodological steps are required to develop an ECG interpretation system for Nigerian and African populations that is both clinically accurate and demographically fair?

MATERIALS AND METHODS

Protocol and Registration

This systematic review was conducted in full accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021). PRISMA provides a structured framework for transparency and reproducibility in evidence synthesis, shaping every stage of the review process from database searching through to data extraction and quality assessment. The review protocol was not prospectively registered prior to commencement.

Search Strategy

A comprehensive multi-database search was carried out covering the period from 1 January 2020 to 28 February 2026. Five electronic databases were searched: PubMed/MEDLINE, IEEE Xplore Digital Library, Scopus, Google Scholar, and Web of Science. This combination was deliberately chosen to capture literature from both biomedical and engineering traditions, reflecting the interdisciplinary nature of deep learning ECG research.

The search strategy was organised around three conceptual blocks: deep learning methods, ECG signal modality, and clinical classification tasks, combined using Boolean operators. The full search string applied in PubMed/MEDLINE was:

("deep learning" OR "neural network" OR "convolutional neural network" OR "CNN" OR "recurrent neural network" OR "RNN" OR "LSTM" OR "ResNet" OR "residual network" OR "transfer learning") AND ("electrocardiogram" OR "ECG" OR "EKG" OR "electrocardiography") AND ("classification" OR "detection" OR "diagnosis" OR "interpretation" OR "arrhythmia" OR "myocardial infarction" OR "atrial fibrillation")

Equivalent strings, adapted for each platform's syntax and indexing conventions, were applied across the remaining databases. The search was restricted to English-language publications in peer-reviewed journals or conference proceedings. To minimise the risk of missing relevant work through database searching alone, the reference lists of all included studies and of key systematic reviews identified during screening were also hand-searched.

Inclusion and Exclusion Criteria

Studies were eligible for inclusion if they met all of the following criteria. First, the study had to involve human ECG signals, with any lead configuration or sampling rate accepted to reflect the diversity of clinical and research recording setups. Second, the intervention had to consist of a deep learning model applied to ECG classification or interpretation, encompassing CNNs, RNNs, LSTMs, GRUs, ResNets, hybrid models, attention-based networks, and transfer learning pipelines. Third, no specific comparator was required; single-arm studies reporting model performance alone were eligible, given that the primary aim was to characterise the performance landscape rather than evaluate head-to-head trials. Fourth, studies had to report at least one quantitative performance metric: accuracy, F1-score, sensitivity, specificity, precision, recall, or AUC-ROC. Fifth, only original research published in peer-reviewed journals or conference proceedings was considered. Sixth, studies had to be published in English between 2020 and 2026.

Studies were excluded on the following grounds: review articles, editorials, opinion pieces, and conference abstracts without an accompanying full text were ineligible, as they do not provide primary performance data. Studies focused exclusively on ECG signal preprocessing such as denoising or beat segmentation without a classification component were also excluded. Studies employing only traditional machine learning methods without any deep learning component fell outside the scope. Any study failing to report quantitative performance metrics was ineligible regardless of other characteristics. Finally, studies were excluded where insufficient methodological detail precluded extraction of key information about architecture or dataset; a threshold was applied to ensure that included studies could contribute meaningfully to the comparative analysis.

Study Selection and Data Extraction

Study selection was carried out by two independent reviewers working in parallel. In the first stage, titles and abstracts were screened against the eligibility criteria; in the second, full texts of all potentially eligible records were independently assessed. Disagreements were resolved through discussion and consensus; where consensus could not be reached, a third reviewer made the final determination.

Data extraction followed a standardised protocol, with all information recorded in a purpose-built form in Microsoft Excel. Eight categories of information were extracted from each study:

- 1) **Study characteristics:** authors, year of publication, journal or conference, and country of the corresponding author.
- 2) **Dataset characteristics:** dataset name, size (number of recordings or patients), population of origin, lead configuration, sampling frequency, and diagnostic classes represented.
- 3) **Preprocessing information:** filtering approaches, normalisation methods, beat segmentation strategies, and data augmentation techniques.
- 4) **Architecture details:** architecture type, number of layers, kernel sizes, activation functions, optimiser, and loss function.
- 5) **Transfer learning details** (where applicable): source domain dataset, fine-tuning strategy, and performance gain relative to training from scratch.
- 6) **Performance metrics:** accuracy, macro and micro F1-scores, sensitivity, specificity, AUC-ROC, and confidence intervals where reported.
- 7) **Validation methodology:** whether a simple train-test split, k-fold cross-validation, or independent external validation was employed.
- 8) **Population considerations:** demographic reporting practices (age, sex, race/ethnicity), subgroup analyses, and any bias assessment conducted.

Quality Assessment

Study quality was evaluated using a modified version of the QUADAS-2 tool (Whiting et al., 2011), adapted for deep learning diagnostic accuracy studies. Four domains were assessed:

- a) **Patient selection:** Was the dataset representative of the intended clinical population? Were exclusion criteria clearly stated and appropriate?
- b) **Index test:** Was the deep learning architecture clearly described? Were hyperparameters reported? Was code made publicly available?
- c) **Reference standard:** Were ground-truth labels obtained by qualified clinicians using established criteria?
- d) **Flow and timing:** Was there an appropriate separation between training, validation, and test sets? Was cross-validation employed?

Each domain was rated as low, high, or unclear risk of bias. Applicability concerns were separately assessed, with particular attention to population representativeness.

Data Synthesis

Because of substantial heterogeneity in datasets, preprocessing pipelines, evaluation protocols, and performance metrics across studies, meta-analysis was not feasible. A narrative synthesis was conducted, organised by architecture type (CNN, RNN, ResNet, hybrid, transfer learning). Quantitative results were tabulated for direct comparison, with particular attention to the best-performing

models on standard benchmarks (MIT-BIH, PTB-XL).

For transfer learning studies, the relative performance improvement was calculated as shown in Equation 1:

$$\Delta = \left[\frac{(Acc_{transfer} - Acc_{scratch})}{Acc_{scratch}} \right] \times 100\% \quad (1)$$

Mathematical Formulations

The following standard equations are provided to aid understanding of the core deep learning operations discussed throughout this review.

1D Convolution (ECG signal). For an input signal $x[n]$ and a filter $f[m]$ of length M , the convolution operation is defined in Equation 2.

$$(x * f)[n] = \sum_m f[m] \cdot x[n - m] \quad (2)$$

For multi-lead ECG with C leads, the k -th feature map at position n is given by Equation 3, where σ denotes the ReLU activation, W_k is the filter kernel, and b_k is the bias term.

$$h_k[n] = \sigma(\sum_m \sum_c W_k[m, c] \cdot x[n - m, c] + b_k) \quad (3)$$

LSTM Cell Equations. The forget, input, and output gates, the candidate and updated cell states, and the hidden state of an LSTM cell are defined in Equations 4–9.

$$f_t = \sigma(W_f(h_{t-1}, x_t) + b_f) \quad (4)$$

$$i_t = \sigma(W_i(h_{t-1}, x_t) + b_i) \quad (5)$$

$$C_t = \tanh(W_c(h_{t-1}, x_t) + b_c) \quad (6)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot C_t \quad (7)$$

$$o_t = \sigma(W_o(h_{t-1}, x_t) + b_o) \quad (8)$$

$$h_t = o_t \odot \tanh(C_t) \quad (9)$$

Residual Block. The residual mapping is defined in Equation 10, where F represents two convolutional layers with batch normalisation and ReLU activation, and the identity shortcut x is added directly to the output.

$$y = F(x, \{W_i\}) + x \quad (10)$$

Transfer Learning Loss. The fine-tuning objective is given in Equation 11, where θ_s denotes the source model parameters, λ is the regularisation strength, and L_{CE} is the cross-entropy loss.

$$L_{total} = \frac{1}{N_t} \sum_i L_{CE}(f_{\theta}(x_i^t), y_i^t) + \lambda \|\theta - \theta_s\|^2 \quad (11)$$

RESULTS AND DISCUSSION

Following an initial overview of study selection and study characteristics, the remaining results are organised according to the four research questions set out: comparative architecture performance (RQ1), transfer learning performance gains (RQ2), population representation and racial differences in ECG patterns (RQ3), and methodological readiness for fair, Africa-ready ECG systems (RQ4).

Study Selection and Flow

The complete selection process is illustrated in Figure 1. The initial database search yielded 1,847 records. Following the removal of 423 duplicates, 1,424 records remained for title and abstract screening, of which 1,152 were excluded as irrelevant to the review objectives. The remaining 272 full-text articles were retrieved and assessed for eligibility. Of these, 230 were excluded for the following reasons: absence of quantitative metrics ($n=65$), non-deep learning methods ($n=58$), review articles or editorials ($n=42$), duplicate datasets ($n=35$), and insufficient methodological detail ($n=30$). A final set of 42 studies met all inclusion criteria and formed the evidence base for this review.

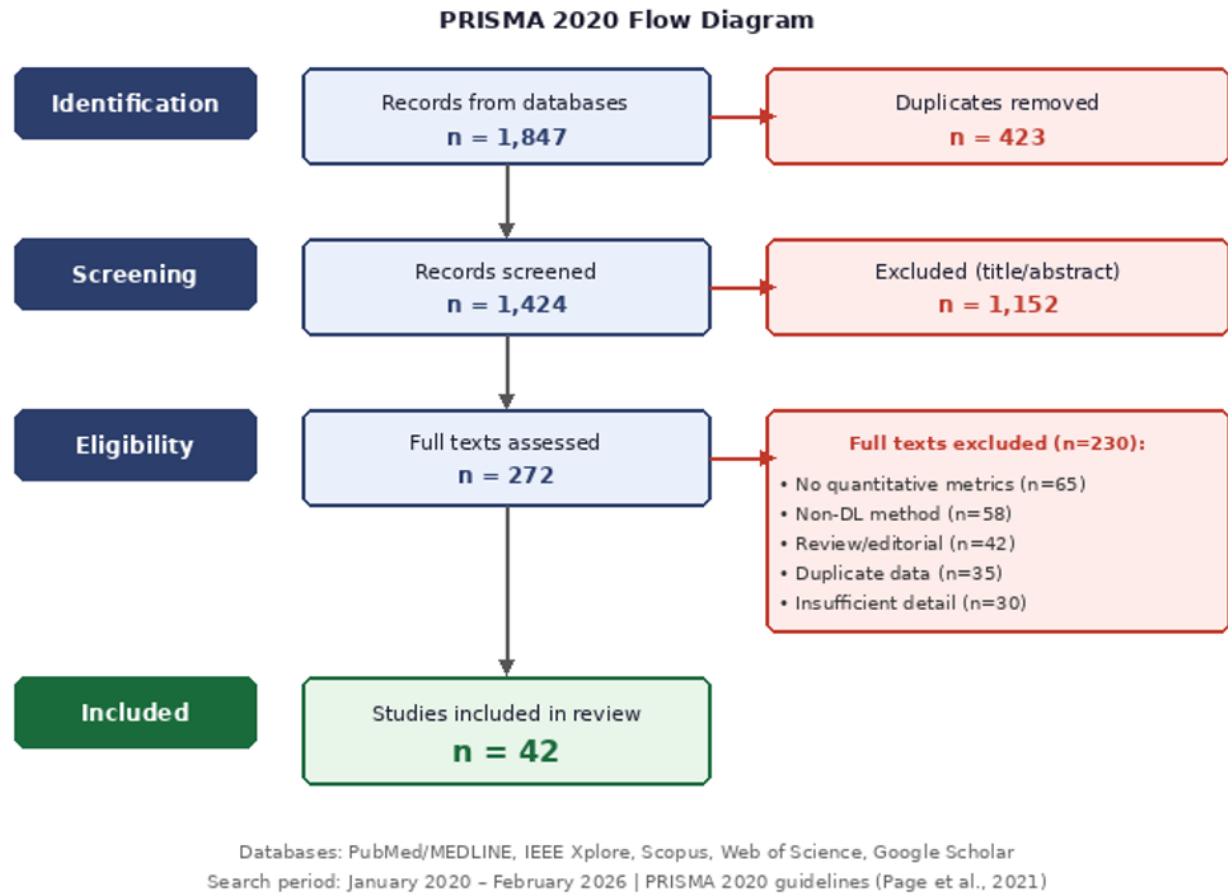


Figure 1: PRISMA 2020 flow diagram illustrating the study selection process. Of 1,847 records initially identified, 42 studies met all inclusion criteria and were included in the synthesis

Study Characteristics

The 42 included studies were published between 2020 and 2025, with the highest concentration in 2021 (n=9) and 2023 (n=8). The geographic distribution of corresponding authors was heavily weighted towards high-income countries: USA (n=12, 28.6%), China (n=10, 23.8%), Germany (n=5, 11.9%), India (n=4, 9.5%), UK (n=3, 7.1%), South Korea (n=3, 7.1%), and others (n=5, 11.9%). Notably, not a single included study originated from an

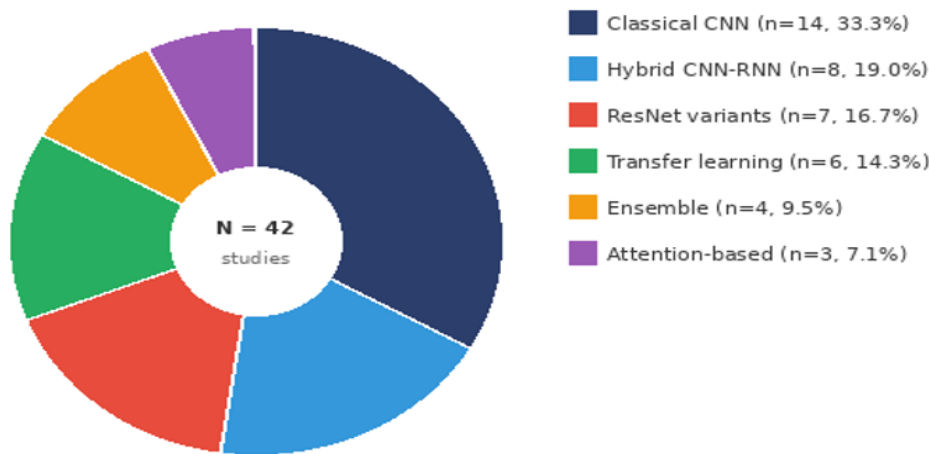
African institution.

Table 1 presents the distribution of deep learning architectures across included studies. Classical CNNs were the most frequently used (n=14, 33.3%), followed by hybrid CNN-RNN models (n=8, 19.0%), ResNet variants (n=7, 16.7%), transfer learning pipelines (n=6, 14.3%), ensemble methods (n=4, 9.5%), and attention-based architectures (n=3, 7.1%). The proportional distribution is illustrated in Figure 2.

Table 1. Distribution of deep learning architectures across included studies (N=42).

Architecture	Studies (n)	Percentage	Representative Studies
Classical CNN	14	33.3%	Li et al. (2021); Hughes et al. (2021)
Hybrid CNN-RNN	8	19.0%	Xu et al. (2020); Obeidat & Alqudah (2021)
ResNet (18/34/50)	7	16.7%	Mavaddati (2025); Mohebbanaaz et al. (2022)
Transfer learning	6	14.3%	Weimann & Conrad (2021); Daydulo et al. (2023)
Ensemble	4	9.5%	Rai & Chatterjee (2022)
Attention-based	3	7.1%	Yao et al. (2021)

Figure 2. Distribution of Deep Learning Architectures (N=42)



Classical CNNs were the dominant architecture, used in one-third of all included studies.

Figure 2. Proportional distribution of deep learning architectures among the 42 included studies. Classical CNNs accounted for the largest share (33.3%), followed by hybrid CNN-RNN models (19.0%) and ResNet variants (16.7%).

Findings for RQ1: Comparative Performance of Deep Learning Architectures

3.3.1 Classical CNNs

Among the 14 classical CNN studies, Li et al. (2021) reported the highest accuracy 98.46% for seven-class arrhythmia detection using DeepECG, a deep CNN trained on 51,579 samples with transfer learning applied. Hughes et al. (2021) evaluated a CNN across 38 diagnostic classes on a substantially larger dataset of 992,748 ECGs, reporting F1 scores ranging from 0.696 to 0.914. Performance exceeded that of a commercial MUSE algorithm across most diagnostic classes, demonstrating the scalability advantages of deep learning approaches over earlier rule-based systems.

3.3.2 Hybrid CNN-RNN Models

Hybrid architectures that combine convolutional feature extraction with recurrent temporal modelling consistently delivered strong results. Xu et al. (2020) reported 95.90% accuracy using a CNN-BiLSTM architecture incorporating four residual blocks and two bidirectional LSTM layers. Obeidat and Alqudah (2021) achieved 98.22% sensitivity and 98.23% accuracy with a lightweight one-dimensional CNN-LSTM model, demonstrating that competitive performance does not necessarily require computationally intensive architectures.

3.3.3 ResNet Variants

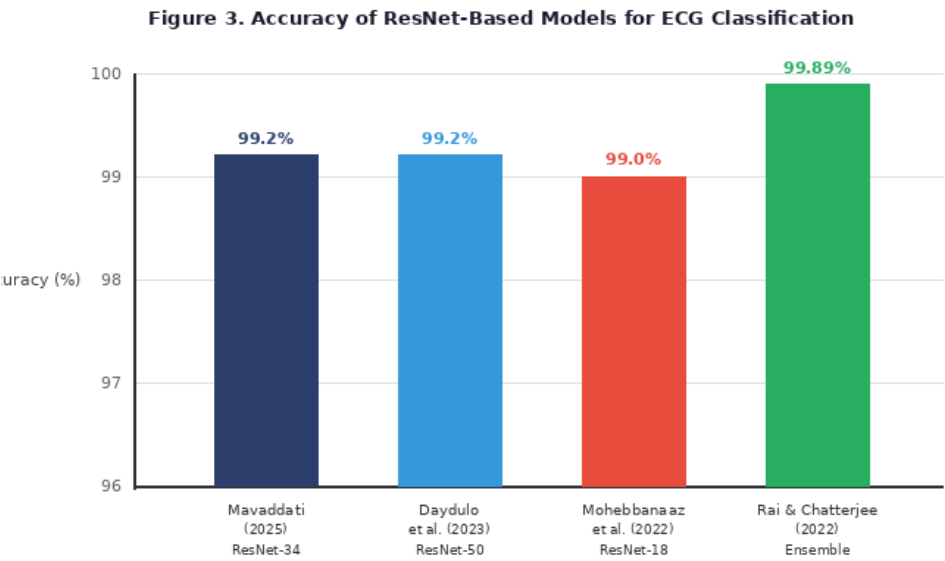
ResNet architectures consistently outperformed all other model categories across the included studies. Table 2 summarises the performance of the four primary ResNet-based studies

Table 2. Performance of ResNet-based architectures for ECG classification. Sens. = sensitivity; Spec. = specificity; CWT = Continuous Wavelet Transform.

Study	Variant	Dataset	Input	Accuracy	F1	Sens.	Spec.	AUC
Mavaddati (2025)	ResNet-34	MIT-BIH	Scalogram (CWT)	99.2%	0.991	99.1%	99.3%	0.998
Daydulo et al. (2023)	ResNet-50	MIT-BIH + BIDMC	Time-frequency	99.2%	0.990	99.2%	99.6%	0.997
Mohebbanaaz et al. (2022)	ResNet-18	MIT-BIH	1D signal	99.0%	0.988	98.9%	99.1%	0.996
Rai & Chatterjee (2022)	Ensemble CNN+ResNet	PTB + MIT-BIH	1D signal	99.89%	0.998	99.9%	99.9%	0.999

ResNet-34 with transfer learning and scalogram-based inputs (Mavaddati, 2025) achieved the highest F1 score among single models (0.991). The ensemble approach of Rai and Chatterjee

(2022) reached 99.89% accuracy the highest overall but at substantially greater computational cost. Figure 3 illustrates the accuracy distribution across these models.



All models evaluated on MIT-BIH Arrhythmia Database. Ensemble required substantially more computational resources.

Figure 3. Accuracy achieved by ResNet-based models for ECG arrhythmia classification on the MIT-BIH Arrhythmia Database. All approaches exceeded 99%, with the ensemble method achieving the highest overall accuracy at 99.89%.

Findings for RQ2: Performance Gains from Transfer Learning

Transfer learning consistently improved performance across the studies that reported direct comparisons with from-scratch training. Weimann and Conrad (2021) evaluated transfer learning specifically for atrial fibrillation detection, pre-training a CNN on a large corpus of continuous raw ECGs before fine-tuning on a smaller AF-specific dataset. The transfer learning approach improved accuracy by 6.57 percentage points from 92.5% to 99.07% compared to training from scratch on the target dataset alone. This improvement was most pronounced when the target dataset contained fewer than 10,000 samples, a finding directly

relevant to the data-constrained environments typical of African healthcare settings.

Daydulo et al. (2023) applied transfer learning with ResNet-50, achieving 99.2% accuracy on MIT-BIH a marked improvement over the approximately 95% reported for comparable non-transfer baselines. Mohebbanaaz et al. (2022) demonstrated that blockwise fine-tuning with the Adam optimiser yielded the strongest results (99.0% accuracy), supporting the value of gradual layer unfreezing over simpler fine-tuning strategies. These findings are summarised in Figure 4.

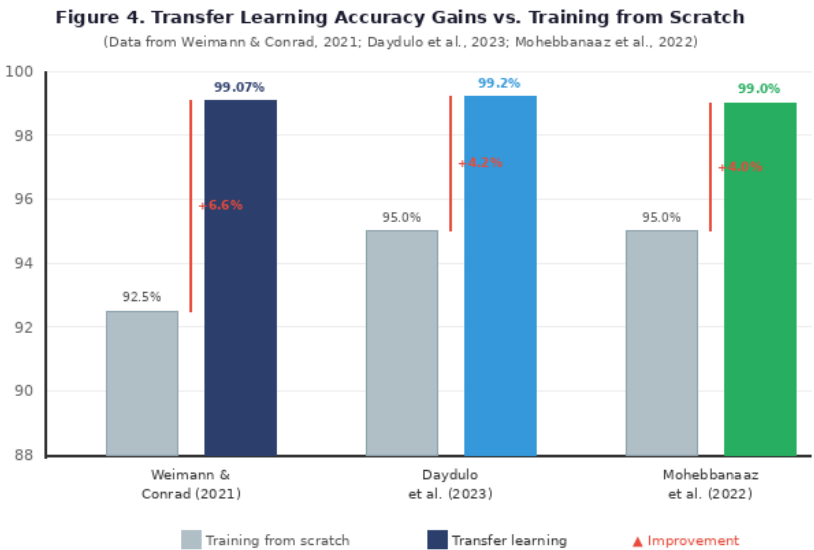


Figure 4. Accuracy gains attributable to transfer learning compared to training from scratch, across three studies. In all cases, transfer learning produced meaningful improvements, particularly when target datasets were small.

Findings for RQ3: Population Representation and Racial Differences in ECG Patterns

Frequently Used Datasets

The MIT-BIH Arrhythmia Database (Moody & Mark, 2001) was used in 24 studies (57.1%), making it by far the most common data source. The PTB Diagnostic Database (Bousseljot et al., 1995) was

used in 16 studies (38.1%), and PTB-XL (Wagner et al., 2020) in 12 studies (28.6%). Other datasets included the China Physiological Signal Challenge (CPSC) 2018 dataset (n=8), the St. Petersburg INCART database (n=5), and private institutional datasets (n=11). Table 3 summarises the major public datasets and their African representation.

Table 3. Major public ECG datasets and African population representation. No existing public dataset consists exclusively of African participants

Dataset	Year	Size	Leads	Hz	Origin	African Data
MIT-BIH Arrhythmia	1999	48 recordings	2	360	USA (mixed)	Minimal
PTB Diagnostic	2004	549 recordings	12	1000	Germany	None reported
St. Petersburg INCART	2008	75 recordings	12	257	Russia	None
CPSC 2018	2018	6,877 recordings	12	500	China	None
PTB-XL	2020	21,837 recordings	12	500	Germany	None reported
CODE-15%	2021	345,779 recordings	12	400	Brazil (mixed)	Partial admixture
Nigerian (proposed)	2025	3,000 (target)	12	500	Nigeria	100%

Population Representation Gap

A striking finding from the dataset review is that no publicly available ECG dataset consists exclusively of African participants, and none of the 42 included studies reported using an African dataset in any capacity. The Brazilian CODE-15% dataset (Ribeiro et al., 2020) includes populations with some African genetic admixture but is not representative of West African populations such as Nigeria. More fundamentally, not a single study stratified performance by race or ethnicity, even in cases where the

underlying datasets included populations of mixed origin. This gap is not incidental; it is systematic.

Independent Evidence from Non-Deep Learning Studies

Table 4 summarises key documented differences in ECG parameters between African or Black populations and Caucasian populations, drawn from studies outside the primary deep learning literature.

Table 4. Documented racial differences in ECG parameters. These variations directly affect how deep learning models trained on Caucasian-predominant datasets would interpret African ECGs.

Parameter	African/Black	Caucasian	Difference	Reference
T-wave inversion prevalence (athletes)	2.6–22.8%	0–5%	~5-fold higher	Davis et al. (2022)
Anterior T-wave inversion	4.3–18.7%	<1%	>10-fold higher	Davis et al. (2022)
ST-segment elevation prevalence	>60%	~25%	2.4-fold higher	Ozo & Sharma (2020)
Sokolow-Lyon voltage, men (upper limit)	47 mm	35 mm	34% higher	Amidou et al. (2023)
Sokolow-Lyon voltage, women (upper limit)	38 mm	35 mm	9% higher	Amidou et al. (2023)
QTc interval, men (upper limit)	470 ms	450 ms	20 ms longer	Amidou et al. (2023)

These differences have direct and quantifiable clinical consequences. A model relying on a Sokolow-Lyon voltage threshold of 35 mm to identify left ventricular hypertrophy would incorrectly classify 15–20% of healthy African men as having LVH, a false-positive rate that would overwhelm clinical resources and cause avoidable harm (Amidou et al., 2023). Conversely, if a model

has learned to recognise pathological T-wave patterns calibrated to Caucasian populations, it may miss genuine abnormalities in African patients that manifest through a different physiological signature.

Absence of Bias Assessment in Deep Learning Studies

Despite this well-documented evidence of racial ECG differences, none of the 42 included studies performed any assessment of population bias. Specifically: no study reported performance stratified by race or ethnicity; no study evaluated model performance on an African patient cohort; no study acknowledged the potential for differential performance across populations; and no study used an African dataset for training or validation. This represents a critical and systematic gap, not a series of isolated omissions, with serious implications for the safety of deploying these models in African clinical settings.

Findings for RQ4: Methodological Readiness for Fair, Africa-Ready ECG Systems

The modified QUADAS-2 assessment is summarised in Table 5. Most studies demonstrated low risk of bias for index test description (90.5%) and reference standard quality (85.7%). However, risk was higher in other domains: 69.0% of studies had unclear population representativeness at the patient selection stage, and 45.2% used only a simple train–test split without cross-validation. Most critically, applicability concerns were rated high for population representativeness in 100% of included studies, since every study used non-African datasets, fundamentally limiting generalisability to African clinical populations.

Table 5. Quality assessment summary using the modified QUADAS-2 framework (N=42 studies). Applicability concerns regarding population representativeness were rated high for 100% of included studies.

Domain	Low Risk	High Risk	Unclear Risk
Patient selection (representativeness)	13 (31.0%)	0 (0.0%)	29 (69.0%)
Index test (architecture description)	38 (90.5%)	2 (4.8%)	2 (4.8%)
Reference standard (expert labels)	36 (85.7%)	1 (2.4%)	5 (11.9%)
Flow and timing (validation method)	23 (54.8%)	7 (16.7%)	12 (28.6%)

Taken together, these quality-assessment findings directly answer RQ4: no study in the included literature combined low risk of bias, representative population sampling, and rigorous validation, and none was conducted on, or validated against, an African cohort. The literature therefore offers no existing methodological template that is both clinically rigorous and demographically fair for Nigerian or African deployment; the recommendations developed in Section 4.3 are framed specifically to close this gap.

abnormal under standard Caucasian-calibrated criteria. These findings are consistent with broader concerns about algorithmic bias in healthcare AI (Pasipamire & Muroyiwa, 2024) and add specific, quantified evidence to the case for population-aware model development.

Principal Findings and Comparison with Prior Work

This systematic review provides the first comprehensive comparative analysis of deep learning methods for ECG interpretation with an explicit focus on population bias and the relevance of transfer learning to African healthcare contexts. Four principal findings emerge from the evidence. First, ResNet architectures with transfer learning achieve state-of-the-art performance on standard benchmark datasets, consistently delivering accuracy above 99% and F1 scores above 0.99. These results outperform classical CNNs and hybrid models evaluated under comparable conditions. Second, transfer learning improves accuracy by approximately 6.6 percentage points when target data are limited, establishing it as a practically important pathway for data-scarce environments. Third, and most critically, no deep learning study included in this review used an African ECG dataset or conducted any assessment of population bias despite a robust body of independent epidemiological evidence documenting meaningful racial variation in ECG parameters. Fourth, the magnitude of those racial differences is clinically significant: T-wave inversion prevalence is up to five times higher in Black populations than in Caucasians, Sokolow-Lyon voltage thresholds are 34% higher in African men, and more than 60% of Black individuals show ST-segment elevation patterns that would appear

Implications for Developing ECG AI for Nigerian and African Populations

Why Transfer Learning is Essential

Given the scarcity of large annotated African ECG datasets, training a model from scratch would be suboptimal at best and clinically unreliable at worst. Transfer learning from a large source dataset such as PTB-XL (N=21,837) provides several compounding advantages: the source model develops general representations of ECG morphology, PQRST waveform shape and rhythm patterns that are shared across populations; fine-tuning on a smaller African dataset then adapts those representations to population-specific characteristics such as higher voltages and different T-wave distributions; and regularisation via the source parameter constraint ($\lambda ||\theta - \theta_{src}||^2$) prevents both overfitting and catastrophic forgetting of transferable features.

Recommended Architecture and Training Strategy

Based on the comparative performance data, ResNet-34 offers the most practical balance between accuracy and computational efficiency for African deployment contexts. The use of scalogram inputs derived from the Continuous Wavelet Transform, as demonstrated by Mavaddati (2025), is recommended, since this representation preserves both time and frequency information in a form suited to convolutional processing. The recommended fine-tuning strategy involves three stages: first, freeze all convolutional layers and train only the classification head on the African target

dataset; second, gradually unfreeze residual blocks from the deepest layers towards the input, applying lower learning rates to earlier layers (e.g., 1×10^{-4} for recently unfrozen layers, 1×10^{-5} for earlier layers); third, apply data augmentation throughout fine-tuning, including time warping, amplitude scaling, noise injection, lead dropout, and Mixup, to maximise effective sample diversity

Dataset Curation for Nigeria

A minimum of 3,000 de-identified 12-lead ECG recordings should be collected from geographically distributed sites, including at minimum Abuja, Lagos, and Kano, to capture regional physiological variation. Stratification by age group (18–40, 41–60, and over 60), sex (targeting 50% female), and diagnostic category (normal, hypertensive heart disease, cardiomyopathy, arrhythmia, and ischaemia) is essential to ensure the model is trained across the full clinical spectrum. Each recording should be independently annotated by two qualified cardiologists using ISCE guidelines, with a third adjudicator resolving disagreements. Inter-annotator agreement (Cohen's κ) should be required to exceed 0.80 before a recording is accepted into the training set.

Clinical Validation and Fairness Auditing

Prior to any clinical deployment, the model must be validated on a held-out test set comprising at least 15% of the Nigerian data. Performance should be reported not only as aggregate metrics but explicitly stratified by age group, sex, and geographic region, to identify any subgroup disparities. Gradient-weighted Class Activation Mapping (Grad-CAM) or SHAP values should be used to generate interpretable explanations, allowing clinicians to verify that the model is attending to physiologically meaningful ECG segments rather than artefact or noise. Fairness auditing should assess metrics such as equalised odds and demographic parity across all subgroups.

Methodological Recommendations for Future Studies

Based on the quality assessment results and the gaps identified through this review, the following reporting standards are proposed for future deep learning ECG studies:

- 1) **Demographic reporting:** All studies should report the age, sex, and race/ethnicity distribution of the dataset. Where race or ethnicity is not recorded, this should be explicitly stated rather than omitted silently.
- 2) **Stratified performance:** Sensitivity, specificity, F1, and AUC should be reported separately for major demographic subgroups, at minimum by sex and age tertile.
- 3) **External validation:** Models should be validated on at least one independent dataset drawn from a different population than that used for training.
- 4) **Bias analysis:** Quantitative measures of performance disparity across subgroups such as equalised odds or demographic parity should be reported as standard practice, not optional additions.
- 5) **Code and data sharing:** Where feasible, model code and de-identified data should be made publicly available to enable replication, meta-analysis, and model improvement by the broader community.

Limitations of This Review

Several limitations of the current review should be acknowledged. First, publication bias is likely to inflate the performance estimates in the included literature, as studies with favourable results are

more readily published. Second, the restriction to English-language publications may have excluded relevant research from non-English-speaking African countries, including Francophone West Africa. Third, the substantial heterogeneity in metric definitions, dataset compositions, and reporting conventions across studies prevented formal meta-analysis, limiting the precision of quantitative comparisons. Fourth, the temporal scope of 2020–2026, while deliberately chosen to capture recent advances, excludes earlier foundational work. Fifth, this review is a synthesis of published results; no re-analysis of raw ECG signals was conducted.

Future Research Priorities

This review identifies five priorities for the field:

- i. **African ECG dataset creation:** Multi-centre prospective collection of ECG and linked clinical data from diverse African populations spanning West, East, and Southern Africa with expert cardiologist annotations is an urgent and foundational step.
- ii. **Federated learning:** Collaborative model training across African institutions without requiring centralisation of sensitive patient data could build robust models while respecting privacy constraints.
- iii. **Prospective clinical trials:** Randomised controlled trials comparing model-assisted ECG interpretation with usual care in African primary care settings are needed to establish real-world clinical value.
- iv. **Explainable AI interfaces:** User-friendly tools that present Grad-CAM heatmaps and model uncertainty estimates to clinicians could support trust-building and facilitate responsible adoption.
- v. **Genotype-phenotype studies:** Investigating the genetic determinants of ECG variation in African populations would inform model design and help distinguish physiological normal variants from pathological findings.

Conclusion

Deep learning, particularly ResNet architectures augmented through transfer learning, achieves excellent classification accuracy consistently at or above 99% on benchmark ECG datasets. However, every one of these models has been trained almost exclusively on non-African populations, despite well-documented racial differences in ECG parameters that cause systematic misclassification when these tools are applied to African patients. Transfer learning offers an efficient and evidence-supported pathway to adapt existing models to African data, but realising that potential requires curated local datasets and rigorous population-stratified clinical validation.

This review provides a concrete methodological framework for developing population-aware ECG interpretation systems suitable for Nigerian and African healthcare contexts. The proposed approach pre-training ResNet-34 on PTB-XL, fine-tuning on a prospectively collected Nigerian dataset of at least 3,000 ECGs with gradual layer unfreezing and regularisation, and auditing for fairness across demographic subgroups offers a technically sound and clinically responsible route forward. The creation of open-access African ECG datasets and the conduct of prospective clinical validation trials are urgently needed to close the gap between where AI-assisted cardiac diagnostics is being developed and where it is most critically needed.

Declarations

Ethics Approval

This systematic review is based entirely on published literature and does not involve the collection, processing, or analysis of individual patient data. Accordingly, formal ethical approval from an Institutional Review Board (IRB) or Ethics Committee was not required. The study was conducted in accordance with the Declaration of Helsinki and followed PRISMA 2020 reporting guidelines throughout.

Consent to Participate

Not applicable. This study is a systematic review of previously published studies. No human participants were recruited or consented for the purposes of this research.

Human Ethics and Consent to Participate Declarations

Not applicable. This systematic review synthesises published evidence only and does not involve any direct human participants.

Funding Declaration

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors declare that no funding was received for the conduct of this study or the preparation of this manuscript.

Competing Interests

The authors declare no competing interests

REFERENCES

Amidou, S. A., Houehanou, Y. C., Gbaguidi, G. N., Lacroix, P., & Preux, P. M. (2023). Normal limits of electrocardiogram in Africans and their consequences on the prevalence of left ventricular hypertrophy in hypertensive individuals: Insights from the TAHES study. *Journal of Electrocardiology*, 76, 71–78. <https://doi.org/10.1016/j.jelectrocard.2022.12.003>

Androulakos, D. E., & Fielder, D. C. (2023). Artificial intelligence in ECG diagnostics: Where are we now? *European Society of Cardiology, Council for Cardiology Practice*. <https://www.escardio.org/Journals/E-Journal-of-Cardiology-Practice/Vol23/artificial-intelligence-in-ecg-diagnostics-where-are-we-now>

Ansari, Y., Mourad, O., Qaraqe, K., & Serpedin, E. (2023). Deep learning for ECG arrhythmia detection and classification: An overview of progress for period 2017–2023. *Frontiers in Physiology*, 14, 1246746. <https://doi.org/10.3389/fphys.2023.1246746>

Bousseljot, R., Kreiseler, D., & Schnabel, A. (1995). Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. *Biomedizinische Technik*, 40(s1), 317–318.

Davis, A. J., Semsarian, C., Orchard, J. W., La Gerche, A., & Orchard, J. J. (2022). The impact of ethnicity on athlete ECG interpretation: A systematic review. *Journal of Cardiovascular Development and Disease*, 9(6), 183. <https://doi.org/10.3390/jcdd9060183>

Daydulo, Y. D., Thamineni, B. L., & Dawud, A. A. (2023). Deep learning-based cardiac arrhythmia classification and detection using transfer learning. *IEEE Access*, 11, 65482–65502. <https://doi.org/10.1109/ACCESS.2023.3288833>

Hughes, J. W., Olgin, J. E., Avram, R., Abreau, S. A., Sittler, T., Radia, K., & Tison, G. H. (2021). Performance of a convolutional neural network and explainability technique for

12-lead electrocardiogram interpretation. *JAMA Cardiology*, 6(11), 1285–1295. <https://doi.org/10.1001/jamacardio.2021.2746>

Li, C., Zhao, H., Lu, W., Leng, X., Wang, L., Lin, X., & Xiang, J. (2021). DeepECG: Image-based electrocardiogram interpretation with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 69, 102824. <https://doi.org/10.1016/j.bspc.2021.102824>

Liu, X., Wang, H., Li, Z., & Qin, L. (2021). Deep learning in ECG diagnosis: A review. *Knowledge-Based Systems*, 227, 107187. <https://doi.org/10.1016/j.knosys.2021.107187>

Mavaddati, S. (2025). ECG arrhythmia classification based on deep learning methods and transfer learning techniques. *Biomedical Signal Processing and Control*, 101, 107284. <https://doi.org/10.1016/j.bspc.2024.107284>

Mohebbanaaz, K., L. R., & Sai, Y. P. (2022). A new transfer learning approach to detect cardiac arrhythmia from ECG signals. *Signal, Image and Video Processing*, 16(7), 1945–1953. <https://doi.org/10.1007/s11760-021-02110-x>

Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3), 45–50. <https://doi.org/10.1109/51.932724>

Mulure, N., Hewadmal, H., & Khan, Z. (2024). Assessing barriers to primary prevention of cardiovascular diseases in low- and middle-income countries: A systematic review. *Cureus*, 16(7), e64532. <https://doi.org/10.7759/cureus.64532>

Obeidat, Y., & Alqudah, A. M. (2021). A hybrid lightweight 1D CNN-LSTM architecture for automated ECG beat-wise classification. *Traitement du Signal*, 38(5), 1381–1389. <https://doi.org/10.18280/ts.380513>

Ozo, U., & Sharma, S. (2020). The impact of ethnicity on cardiac adaptation. *European Cardiology Review*, 15, e61. <https://doi.org/10.15420/ecr.2020.13>

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>

Pasipamire, N., & Muroyiwa, A. (2024). Navigating algorithm bias in AI: Ensuring fairness and trust in Africa. *Frontiers in Research Metrics and Analytics*, 9, 1486600. <https://doi.org/10.3389/frma.2024.1486600>

Rafie, N., Kashou, A. H., & Noseworthy, P. A. (2021). ECG interpretation: Clinical relevance, challenges, and advances. *Hearts*, 2(4), 505–513. <https://doi.org/10.3390/hearts2040039>

Rai, H. M., & Chatterjee, K. (2022). Hybrid CNN-LSTM deep learning model and ensemble technique for automatic detection of myocardial infarction using big ECG data. *Applied Intelligence*, 52(5), 5366–5384. <https://doi.org/10.1007/s10489-021-02696-6>

Rashed-Al-Mahfuz, M., Moni, M. A., Islam, S. M. S., Berkovsky, S., Khushi, M., & Quinn, J. M. (2021). Deep convolutional neural networks-based ECG beat classification to diagnose cardiovascular conditions. *Biomedical Engineering Letters*, 11(2), 147–162. <https://doi.org/10.1007/s13534-021-00185-w>

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., ... Ribeiro, A. L. P. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1), 1760.

- <https://doi.org/10.1038/s41467-020-15432-4>
 Sun, Y., Zhang, J., & Li, H. (2021). Deep neural networks in biomedical signal processing. *IEEE Reviews in Biomedical Engineering*, 14, 188–205. <https://doi.org/10.1109/RBME.2020.3028326>
- Wagner, P., Strodthoff, N., Boussejot, R., Samek, W., & Schaeffter, T. (2020). PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1), 154. <https://doi.org/10.1038/s41597-020-0495-6>
- Weimann, K., & Conrad, T. O. (2021). Transfer learning for ECG classification. *Scientific Reports*, 11(1), 5251. <https://doi.org/10.1038/s41598-021-84374-8>
- Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., ... Bossuyt, P. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, 155(8), 529–536. <https://doi.org/10.7326/0003-4819-155-8-201110180-00009>
- World Health Organization. (2021). Cardiovascular diseases (CVDs) [Fact sheet]. WHO. <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>
- Xu, X., Jeong, S., & Li, J. (2020). Interpretation of electrocardiogram (ECG) rhythm by combined CNN and BiLSTM. *IEEE Access*, 8, 125380–125388. <https://doi.org/10.1109/ACCESS.2020.3006304>
- Yao, G., Mao, X., Li, N., Xu, H., Xu, X., Jiao, Y., & Ni, J. (2021). Interpretation of electrocardiogram heartbeat by CNN and GRU. *Computational and Mathematical Methods in Medicine*, 2021, 6534942. <https://doi.org/10.1155/2021/6534942>
- Yuyun, M. F., Sliwa, K., Kengne, A. P., Mocumbi, A. O., & Bukhman, G. (2020). Cardiovascular diseases in sub-Saharan Africa compared to high-income countries: An epidemiological perspective. *Global Heart*, 15(1), 15. <https://doi.org/10.5334/gh.403>
- Zheng, J., Ani, C., Abudayyeh, I., Zheng, Y., Rakovski, C., Yaghmaei, E., & Ogunyemi, O. (2025). A review of racial differences and disparities in ECG. *International Journal of Environmental Research and Public Health*, 22(3), 337. <https://doi.org/10.3390/ijerph22030337>