Biological Data 2026, 1(1), 104-135.
https://doi.org/????

Review Article

Hybrid CNN-Transformer Models for ECG-Based Cardiac Diagnosis

Samad Najjar-Ghabel[1*], Shamim Yousefi[1], Soheila Yaghobi Niari[1]

[1] Computer Engineering Department, University of Mohaghegh Ardabili, Ardabil, Iran

***** Corresponding Author: Samad Najjar-Ghabel
Email: S.Najjar@uma.ac.ir

CITATION

Najjar-Ghabel, S., Yousefi, S., & Yaghobi Niari, S. (2026). Hybrid CNN-Transformer models for ECG-based cardiac diagnosis. Biological Data Journal, 1(1), 104–132. https://doi.org/xxxxxxxxx.

ARTICLE INFO

Received: May 2, 2026
Revised: May 30, 2026
Accepted: June 18, 2026
Available online: June 22, 2026

COPYRIGHT

© The author(s) retain the copyright. Publisher: University of Kurdistan Press.
This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Abstract: This review evaluates hybrid Convolutional Neural Network (CNN) and Transformer architectures applied to ECG-based cardiac monitoring. We analyze attention-based and multimodal frameworks using reported metrics from benchmark datasets, including MIT-BIH and PhysioNet. Because current research often restricts its scope to isolated architectures or single diagnostic tasks, this paper categorizes established methods for model selection, data preprocessing, and performance evaluation in single-lead and multi-lead ECG analysis. Published results show that hybrid configurations achieve peak accuracies of 98.6%, exceeding the ~95.0% typical of standalone CNNs and the ~96.0% of standalone Transformers. We conclude by examining limitations in data volume and cross-cohort generalization, alongside the requirements for multi-center clinical deployment.

Keywords: Deep Learning, Cardiac Disease Diagnosis, Convolutional Neural Networks (CNNs), Transformers, Hybrid Models, Electro Cardio Gram (ECG)

1. Introduction

Cardiovascular diseases are a primary global health concern, accounting for high mortality rates annually and requiring accurate diagnostic methods to improve patient outcomes (Alkhodari et al., 2024; Busia et al., 2024; Wang et al., 2024). Electrocardiogram (ECG) signals are standard tools for screening and diagnosing various conditions, including congenital valvular murmurs, arrhythmias, and stenosis. However, interpreting these signals requires advanced computational methods to maintain diagnostic reliability and clinical efficiency (Avola et al., 2024; Ayano et al., 2024; Wu et al., 2024). While conventional manual ECG interpretation is susceptible to inter-observer variability and human error (Chen et al., 2024; Chen et al., 2024; Din et al., 2024), standalone deep learning models also present specific structural limitations. Convolutional Neural Networks (CNNs) are constrained in capturing long-range sequential dependencies, whereas standalone Transformers lack the inductive bias necessary for fine-grained local feature extraction (Fredj et al., 2025; Kim et al., 2025; Wieneke & Voigt, 2024). These independent limitations restrict diagnostic accuracy, particularly when models process noisy signals or evaluate complex pathologies (Zhang et al., 2026; Aburass et al., 2025; Liu et al., 2024). Hybrid CNN-Transformer models address these constraints by integrating the spatial feature extraction capabilities of CNNs with the global contextual modeling of Transformers (Chen et al., 2024; Chen et al., 2024; Din et al., 2024). Empirical studies indicate that these hybrid architectures report accuracies up to 98.6% on benchmark datasets, including MIT-BIH and PhysioNet, demonstrating high efficacy in applications such as murmur detection, arrhythmia classification, and stenosis recognition (Fredj et al., 2025; Kim et al., 2025; Alkhodari et al., 2024; Avola et al., 2024; Ayano et al., 2024; Busia et al., 2024; Wang et al., 2024; Wieneke & Voigt, 2024; Wu et al., 2024). This paper presents a systematic literature review of hybrid CNN-Transformer models applied to cardiac disease classification. Existing review papers examine broader deep learning applications in cardiology, typically isolating their focus on standalone CNNs or general attention mechanisms, but lack a focused synthesis of architectures that explicitly hybridize CNNs and Transformers for ECG analysis. To address this gap, our work synthesizes reported empirical findings based on standard evaluation metrics, including accuracy, sensitivity, specificity, and F1-score. Rather than proposing a novel algorithmic architecture, this paper constructs a conceptual framework that consolidates standard practices in model selection, data preprocessing, and performance assessment. This structured approach provides guidelines for optimizing diagnostic accuracy, identifies persistent limitations such as data scarcity, and outlines requirements for clinical integration. The proposed solution to these structural shortcomings relies on hybrid CNN-Transformer architectures, which integrate the localized spatial feature extraction of CNNs with the long-range dependency modeling of Transformers (Wang et al., 2025; Chen et al., 2024; Chen et al., 2024; Din et al., 2024; Sakhavi et al., 2018). Published findings show theses hybrid paradigms achieve peak classification accuracies of up to 98.6% on benchmark datasets, including MIT-BIH and PhysioNet. Consequently, these models have been adopted for diagnostic applications such as murmur detection, arrhythmia classification, and stenosis recognition using single-lead and multi-lead ECG signals (Fredj et al., 2025; Kim et al., 2025; Alkhodari et al., 2024; Avola et al., 2024; Ayano et al., 2024; Busia et al., 2024; Wang et al., 2024; Wieneke & Voigt, 2024; Wu et al., 2024). Unlike previous reviews that focus on isolated architectures or distinct clinical applications (Guo et al., 2022; Zhang et al., 2021; Li et al., 2019; Lawhern et al., 2018; Zhou et al., 2018), this work introduces a methodological framework. This blueprint synthesizes practices in model selection, data preprocessing, and performance evaluation to provide a roadmap for optimizing diagnostic pipelines. By comparing hybrid models against standalone CNNs and Transformers, this study analyzes the structural advantages of hybrid paradigms while addressing fundamental limitations, particularly data scarcity and computational complexity.

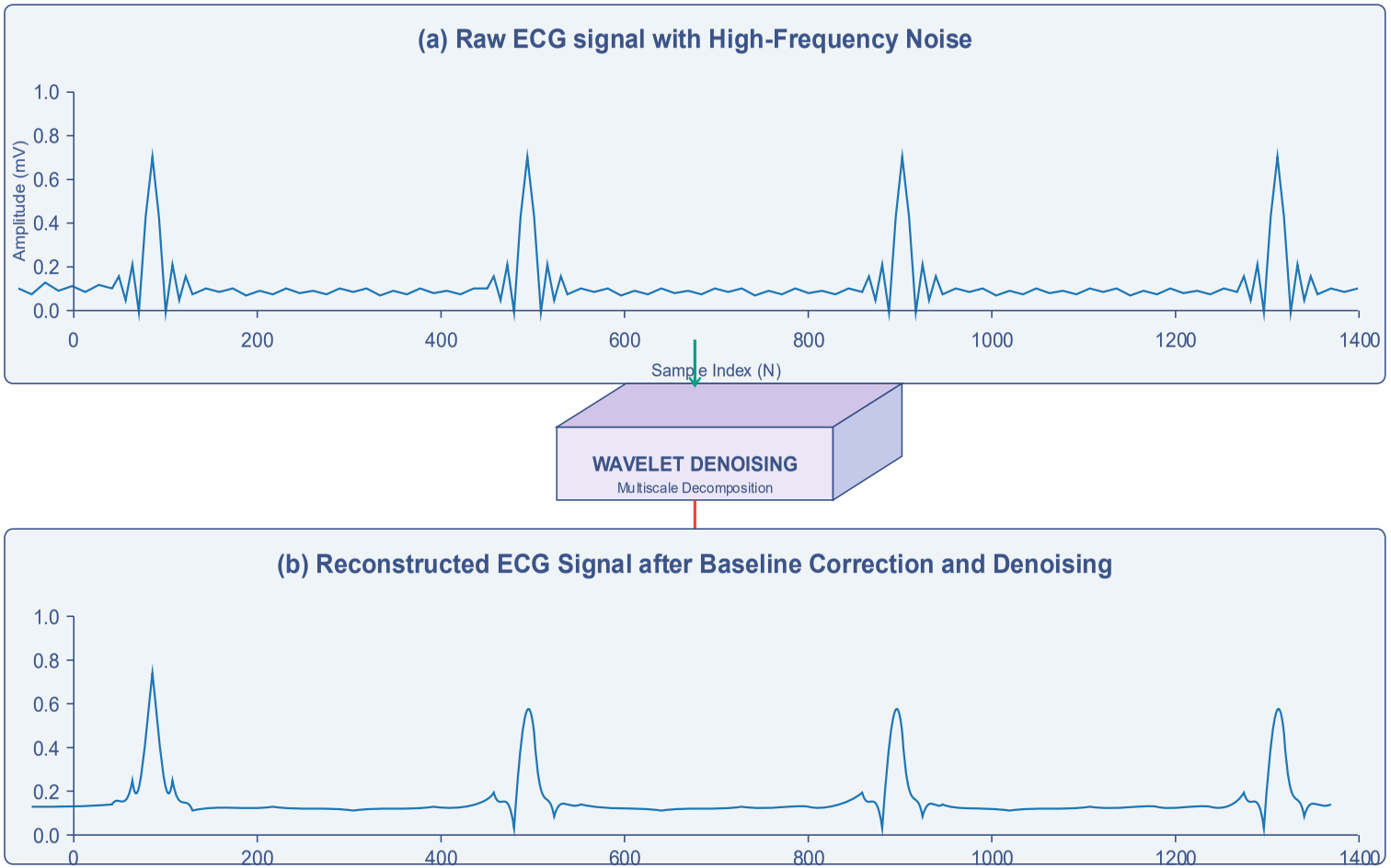


Figure 1. Comparison of the denoising effect of ECG signals (Wang et al., 2025).

To ensure structural clarity, the remainder of this review is organized as follows: Section 2 details the systematic literature review methodology and inclusion criteria. Section 3 analyzes the constituent technologies (CNNs and Transformers) and synthesizes the proposed conceptual blueprint. Section 4 examines persistent challenges, emphasizing data scarcity and the necessity for Uncertainty Quantification (UQ). Finally, Section 5 concludes the manuscript by outlining future research trajectories required for the clinical deployment of hybrid cardiac diagnostic models.

2. Background and Literature review

To understand the trajectory of automated cardiovascular diagnostics, existing deep learning methodologies must be evaluated not as isolated historical milestones, but through a critical comparative synthesis. This section systematically evaluates the structural limitations of standalone models, analyzes the underlying mechanics of hybrid CNN-Transformer superiority, evaluates the impact of dataset complexity, and examines cross-domain adaptations spanning electrophysiological signals and structural imaging. To ensure a rigorous, transparent, and reproducible methodology, the literature search and selection pipeline was conducted in strict adherence to the PRISMA guidelines. As illustrated in Figure 2, the systematic workflow—encompassing initial identification across primary databases, deduplication, and full-text eligibility assessment—ultimately distilled the initial records down to a highly relevant corpus of 34 core studies for architectural evaluation and qualitative synthesis.

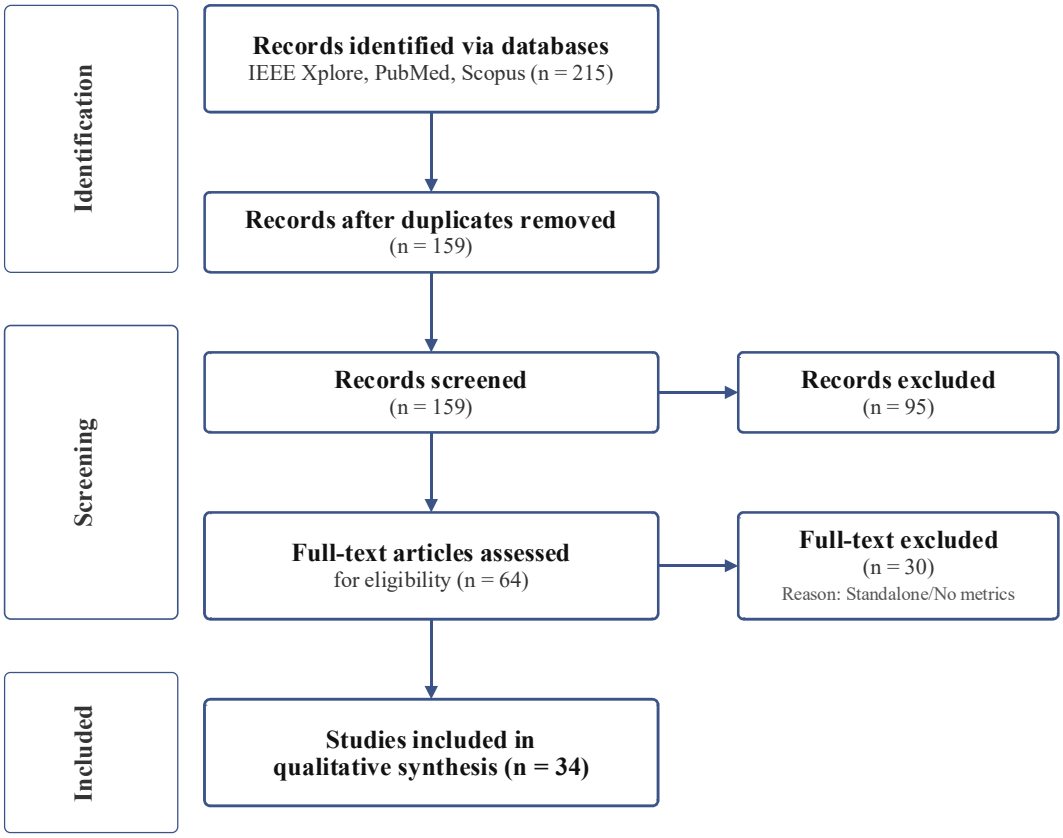


Figure 2. PRISMA flow diagram detailing the systematic literature search, screening, and rigorous eligibility assessment that yielded the final 34 core studies included in the qualitative synthesis.

2.1. Standalone Architectures and Their Structural Limitations

Convolutional Neural Networks (CNNs) serve as standard models for automated cardiac diagnostics because they derive hierarchical feature representations from raw input data. Researchers use them to detect arrhythmias, classify morphological cardiac anomalies, segment the left ventricle in echocardiograms, and identify multiscale stenosis. In electrophysiological wave analysis, localized convolutional kernels detect short-duration morphological alterations, such as QRS complex widening or ST-segment elevations. Similarly, segmentation frameworks like U-Net and U-Net++ improve spatial localization and gradient propagation across structural boundaries in echocardiography.

2.2. Critical Synthesis of Hybrid CNN-Transformer Architectures: Mechanisms of Superiority

Transformers use self-attention mechanisms to model long-range dependencies across time-series medical data (Chen et al., 2024; Chen et al., 2024; Din et al., 2024). Studies show that hybrid CNN-Transformer models outperform standalone CNN and Transformer baselines in various cardiovascular tasks (Alkhodari et al., 2024; Busia et al., 2024; J. Chen et al., 2024; W. Chen et al., 2024; Din et al., 2024; Wang et al., 2024). This performance improvement relies on specific mechanisms:

- **Complementary Feature Extraction:** CNN modules operate as inductive feature extractors that capture localized, high-frequency spatial or morphological patterns, such as individual P-QRS-T wave morphologies or valvular boundary definitions. After encoding these features into condensed representations, the self-attention layers of the Transformer compute global inter-beat relationships. This enables the system to detect long-term rhythm irregularities across thousands of sequential time steps without structural decay (Chen et al., 2024; Chen et al., 2024; Din et al., 2024).
- **Conditional Superiority:** Transformers offer diagnostic advantages when processing complex, extended-duration, or multi-channel data. In long-term ambulatory ECG monitoring, such as 24-hour Holter data, or multi-lead tasks where anomalies depend on cross-lead synchronization and temporal evolution, self-attention mechanisms dynamically weight the diagnostic importance of distant cardiac events. This approach outperforms fixed-window convolutions (Fredj et al., 2025; Kim et al., 2025; Wang et al., 2025; Wieneke & Voigt, 2024; Sakhavi et al., 2018).
- **Versatility Across Diagnostic Modalities:** These hybrid models adapt to various data types. They identify congenital valvular murmurs from phonocardiograms (Alkhodari et al., 2024), classify arrhythmias on edge devices using lightweight attention (Chen et al., 2024; Chen et al., 2024; Din et al., 2024), and detect multiscale structural stenosis in echocardiographic video sequences (Zhang et al., 2026; Aburass et al., 2025; Liu et al., 2024). Reported classification accuracies reach up to 98.6% on established benchmarks (Alkhodari et al., 2024; Wang et al., 2024).

2.3. Dataset Complexity, Performance Disparities, and Open Limitations

Existing benchmark datasets present different levels of diagnostic difficulty:

- **Curated Benchmarks vs. Clinical Complexity:** High performance metrics, often exceeding 98%, typically occur on curated, single-lead datasets like the MIT-BIH Arrhythmia Database (Alkhodari et al., 2024; Wang et al., 2024). MIT-BIH presents severe inter-patient overlap when improperly split and lacks the complex noise profiles of clinical environments. Multi-

lead clinical datasets, such as PTB-XL or ambulatory wearable recordings, introduce higher difficulty due to baseline wander, motion artifacts, muscle electromyographic (EMG) interference, and complex co-morbidities. On these datasets, existing architectures show lower sensitivity and F1-scores, particularly for underrepresented pathological classes (Zhang et al., 2026; Aburass et al., 2025; Fredj et al., 2025; Kim et al., 2025; Liu et al., 2024; Wieneke & Voigt, 2024).

- **Unresolved Methodological Limitations:** The computational complexity of standard self-attention scales quadratically, denoted as $O(N^2)$, with sequence length N. Deploying full-scale hybrid architectures on continuous-monitoring edge devices requires pruning or quantization. Additionally, data scarcity and class imbalance limit generalization for rare cardiac arrhythmias. Current models also lack clinical explainability. While self-attention maps show which time segments influenced a prediction, they do not map directly to the established electrophysiological criteria required for clinical adoption.

2.4. The Multi-Modal Continuum: Bridging Electrophysiological Signals and Structural Imaging

A review of automated cardiac diagnostics includes both 1D physiological signals and 2D or 3D clinical imaging. While 1D electrophysiological signals like ECG and phonocardiograms (Alkhodari et al., 2024) differ from structural imaging like echocardiography (Zhang et al., 2026; Aburass et al., 2025; Liu et al., 2024), they form a multi-modal clinical continuum. In clinical cardiology, 1D ECG signals capture electrophysiological conduction anomalies, and echocardiographic imaging visualizes the resulting mechanical and structural manifestations, such as left ventricular dynamics, wall motion abnormalities, and valvular stenosis. Computationally, both modalities share structural mandates. Diagnostic models must extract fine-grained local features while preserving global temporal continuity across the cardiac cycle, whether processing sequential time-series voltages in ECG or spatio-temporal frames in echocardiographic video streams. Hybrid CNN-Transformer architectures, originally designed for inter-beat dependencies in ECG (Chen et al., 2024; Chen et al., 2024; Din et al., 2024), directly apply to the spatio-temporal attention mechanisms needed to track multiscale stenosis and left-ventricular boundary deformations in sequential echocardiographic frames (Zhang et al., 2026; Aburass et al., 2025; Liu et al., 2024).

2.5. Cross-Domain Architectural Transfer: Insights from EEG Processing

Methodological developments in electroencephalography (EEG) classification provide theoretical cross-domain applications for ECG processing. Both modalities are high-dimensional, non-stationary physiological time series subject to biological artifacts and inter-channel spatial dependencies. Architectural designs for EEG complexities apply to multi-lead ECG challenges. For example, the multi-branch inception topologies in EEG-Inception CNNs show how multi-scale temporal extraction mitigates severe non-stationarity. Researchers can apply this technique directly to polymorphic arrhythmia detection. To systematically understand the architectural advancements in automated ECG analysis, it is essential to trace the methodological evolution in this domain. As illustrated in Figure 3, the field has undergone a significant paradigm shift through four distinct phases. Initially dominated by standalone CNNs and RNNs designed for local feature extraction, the landscape shifted towards pure Transformer architectures to capture long-range global dependencies, albeit at a high computational cost of $\mathcal{O}(N^2)$. Subsequently, early sequential hybrids (CNN-Transformers) emerged to reintroduce local inductive biases. Ultimately, this trajectory has led to the current state-of-the-art: advanced parallel and cross-attention hybrid models that achieve an optimal equilibrium between local morphological precision and global contextual awareness for cardiac diagnosis.

Table 1. Summary of advanced deep learning techniques in ECG and EEG signal analysis.

Study	Methodology	Application	Results
(Zhang et al., 2020)	Combined CNN and RNN with spatio-temporal attention	ECG Classification	F-score of 83.5%
(Shashikumar et al., 2018)	Combined CNN and LSTM for temporal pattern capture in ECG signals	ECG Classification	Accuracy of 96%
(Wang et al., 2023)	Two-level hierarchical GAN architecture; first level identifies anomalies; second level classifies arrhythmias	ECG Anomaly Detection and Classification	Addressed data imbalance issues
(Peng et al., 2022)	Feature extraction improved with denoising and reconstruction using discrete wavelet transform	ECG Signal Processing	Enhanced feature extraction process
(Ronneberger et al., 2015)	CNN with symmetric expanding path for segmentation, enabling precise localization	Biomedical Image Segmentation	Effective in biomedical segmentation
(Zhou et al., 2018)	Improved U-Net architecture with U-Net++; added nested and dense skip connections	Biomedical Image Segmentation	Better gradient flow and performance
(Guo et al., 2022)	Proposed SegNeXt, a convolutional network for semantic segmentation	Static Image Segmentation	Efficient convolutional attention
(Zhang et al., 2021)	EEG-inception CNN architecture based on inception-time backbone	MI-EEG Classification	High accuracy, processes raw EEG signals
(Li et al., 2019)	FBCSP with voting mechanism for three-class MI classification	Multi-Class Motor Imagery	68.6% accuracy with BCI Dataset IV 2a
(Lawhern et al., 2018)	Developed EEGNet, a CNN for EEG-based BCIs with cross-paradigm classification ability	EEG-based BCI Classification	Robust, high performance
(Sakhavi et al., 2018)	CNNs with customized parameters for subject-specific EEG temporal feature extraction	EEG Signal Processing	Improved BCI performance

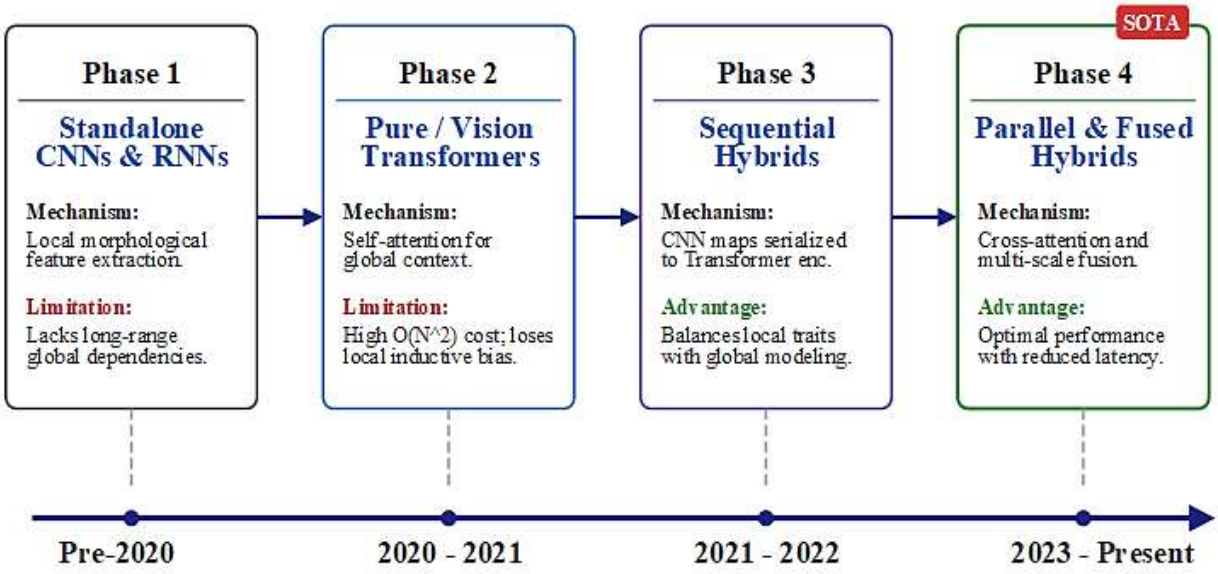


Figure 3. Evolutionary roadmap of deep learning architectures for ECG-based cardiac diagnosis. The progression illustrates the paradigm shift from standalone local extractors (CNNs) to global self-attention models (Transformers), culminating in advanced parallel and cross-attention hybrid networks.

3. Methods

This section systematically examines the architectural paradigms of hybrid CNN-Transformer models, critically evaluating their respective advantages and inherent limitations within the context of automated ECG diagnostics. In the next part, we will examine hybrid CNN-Transformer models with that in mind regarding their advantages and limitations. CNNs frequently neglect holistic trends required in the diagnosis of the heart because they are plagued with long-range temporal dependencies. With great processing efficiency, CNNs are especially good in extracting little spatial information—for example, QRS complexes in electrocardiograms). Transformers can circumvent this problem by using self-attention to collect global dependencies (for example, periodic anomalies scattered across an ECG sequence). But they need a large dataset and a lot of processing capability to get optimal performance (Chen et al., 2024; Chen et al., 2024; Din et al., 2024). Hybrid models mitigate these weaknesses by integrating CNNs local feature extraction with Transformers contextual modeling, achieving superior accuracy and robustness to noise, albeit at the cost of increased complexity. As depicted in Figure 4, the hybrid CNN-Transformer architecture (Li et al., 2022) synergistically integrates local morphological feature extraction with global sequence modeling. The network leverages shallow convolutional layers to hierarchically isolate transient cardiac anomalies, while the subsequent Transformer block utilizes self-attention mechanisms to capture long-range spatiotemporal dependencies inherent in complex arrhythmias.

3.1. Convolutional Neural Networks (CNNs) in Cardiac Diagnostics

Convolutional Neural Networks (CNNs) originated in spatial image analysis and now serve automated cardiovascular diagnostics. They process 1D electrophysiological signals like ECG and 2D or 3D structural imaging like echocardiography. Instead of depending on handcrafted clinical features, CNNs act as inductive feature extractors. They autonomously learn hierarchical representations of cardiac abnormalities from raw input data. CNNs form the foundational structure for local feature extraction in cardiac signal analysis. By using hierarchical convolutional kernels and down-sampling mechanisms like max-pooling,

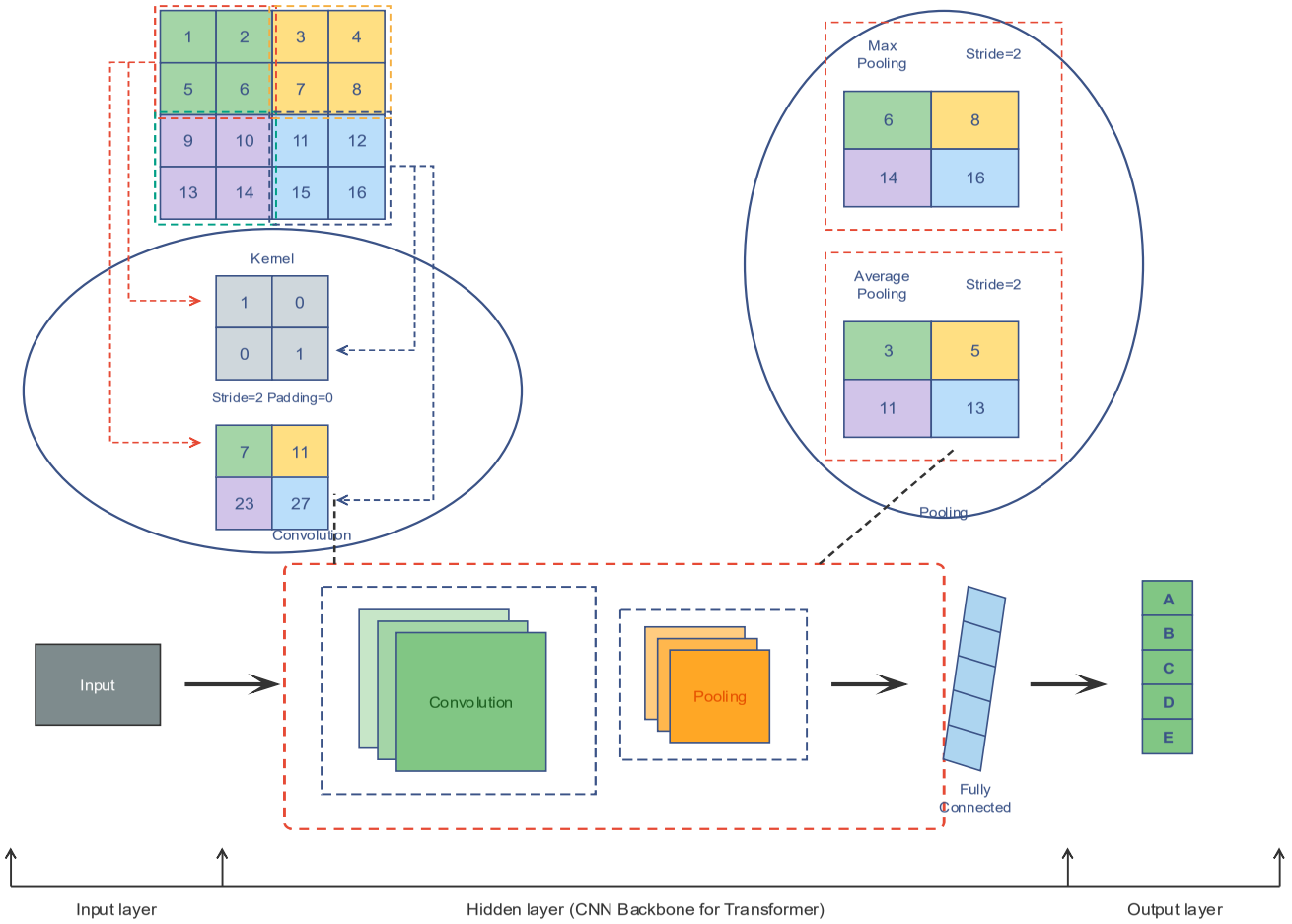


Figure 4. Schematic representation of a hybrid CNN-Transformer architecture (Li et al., 2022).

CNNs capture localized morphological anomalies in ECG signals, such as anomalous QRS complexes or ST-segment deviations. Modern hybrid architectures then pass these localized, high-frequency spatial representations to sequential modules like Transformers. This process contextualizes temporal dynamics without the computational overhead of processing raw, uncompressed signals. Figure 5 delineates the structural blueprint of a dedicated convolutional neural network (Wang et al., 2025) optimized for one-dimensional cardiac time-series data. The computational pipeline cascades sequential 1D convolutions and max-pooling operations to automatically extract shift-invariant local features—such as P-wave dispersion and QRS widening—culminating in fully connected layers for robust diagnostic classification.

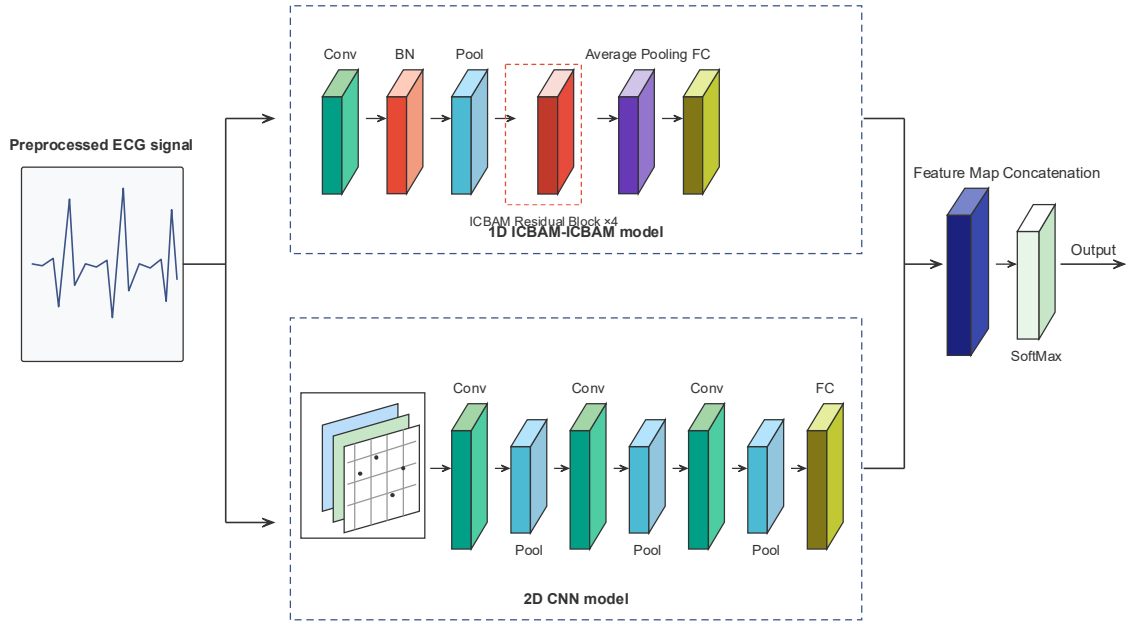


Figure 5. A schematic diagram of the structure of the CNN model (Wang et al., 2025).

- **Mathematical Mechanics and Clinical Relevance in ECG Processing**

For electrophysiological time-series data such as ECG, the foundational operation is the 1D temporal convolution. Given a standardized ECG signal sequence X and a learnable filter or kernel W, the convolution operation generates an output feature map Y at time step n. The mathematical definition is (Equation 1):

$$\mathrm{Y[n]} = \sum_{\mathrm{k=0}}^{\mathrm{K-1}} \mathrm{W\,[k]} \cdot \mathrm{X[n-k]} + \mathrm{b} \tag{1}$$

where K represents the kernel size and b denotes the bias term. In ECG waveform analysis, these mathematical parameters carry direct clinical significance. The kernel size (K) defines the local receptive field of the network. Shallow convolutional layers use small kernel sizes to capture high-frequency, localized morphological aberrations, such as sharp R-peaks, QRS complex widening, or pathological Q-waves (Chen et al., 2024; Chen et al., 2024; Din et al., 2024). As the signal passes through successive convolutional and pooling layers, the network builds a feature hierarchy. Deep layers have a broader receptive field. This allows the aggregation of spatial information over longer

temporal windows. This macroscopic view enables the network to identify low-frequency, inter-wave morphological abnormalities, such as ST-segment elevations or depressions and P-R interval prolongations. Physicians use these indicators to detect ischemia or conduction blocks (Houssein et al., 2024). The translation-invariant nature of convolution means CNNs reliably detect these morphological anomalies despite slight phase shifts or baseline wander in the patient's heart rate.

- **Strengths and Inherent Architectural Limitations**

The primary advantage of CNNs in cardiology involves their capacity for morphological extraction and spatial segmentation. In structural diagnostics, multi-dimensional CNN variants like U-Net architectures achieve exceptional precision in left ventricular boundary delineation (Fredj et al., 2025; Kim et al., 2025; Wieneke & Voigt, 2024) and multiscale stenosis recognition from echocardiographic video frames (Zhang et al., 2026; Aburass et al., 2025; Liu et al., 2024). However, when applied to complex arrhythmia classification across extended continuous monitoring, such as 24-hour ambulatory Holter data, standalone CNNs exhibit structural limitations. The localized nature of the convolutional receptive field restricts the ability of the network to model long-range beat-to-beat temporal dependencies. Expanding kernel sizes or increasing network depth marginally enlarges the receptive field, but it heavily inflates computational overhead and the risk of overfitting. Consequently, CNNs effectively determine what morphological anomaly occurs within a specific heartbeat, but they struggle to contextualize when and how these anomalies evolve over thousands of sequential beats. This structural deficit requires the integration of sequential modeling paradigms, such as Recurrent architectures or self-attention mechanisms, to capture global cardiac rhythms (Avola et al., 2024; Ayano et al., 2024; Wu et al., 2024).

- **Transformers in Cardiac Sequence Modeling**

The Transformer architecture introduces a different paradigm in sequence modeling. It replaces the sequential processing loops of Recurrent Neural Networks (RNNs) and the localized receptive fields of CNNs with a global self-attention mechanism (Chen et al., 2024). Initially developed for natural language processing, Transformers demonstrate high efficacy in cardiovascular diagnostics. They specifically model extended electrophysiological ECG sequences where capturing long-range temporal correlations improves the detection of complex arrhythmias (Alkhodari et al., 2024; Busia et al., 2024; Chen et al., 2024; Chen et al., 2024; Din et al., 2024; Wang et al., 2024).

- **Overcoming RNN Limitations through Self-Attention**

In traditional cardiac sequence analysis, RNNs, including LSTMs, process data sequentially. When applied to extended continuous ECG recordings from ambulatory Holter monitors, RNNs suffer from the vanishing gradient problem. They fail to maintain contextual memory over thousands of cardiac cycles. The self-attention mechanism of the Transformer resolves this structural deficit. Given an input sequence of ECG features $X = [x_1, x_2, .., .. x_n]$, self-attention projects the data into three distinct matrices: Queries (Q), Keys (K), and Values (V). The model computes the attention scores mathematically as (Equation 2):

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}_\text{k}}}\right)\text{V} \quad (2)$$

where $d_k$is the scaling dimension. Clinically, this equation dictates that every single time step, such as a specific QRS complex, directly calculates its correlation score with every other time step in the entire recording, regardless of temporal distance. This benefits ECG analysis. It enables the model to

link distant irregular heartbeat patterns, such as intermittent premature ventricular contractions (PVCs) or paroxysmal atrial fibrillation episodes separated by prolonged periods of normal sinus rhythm, without the sequential information degradation seen in RNNs.

- **Multi-Head Attention for Dual Morphological and Rhythmic Analysis**

To capture the multifaceted nature of cardiac diseases, Transformers utilize Multi-Head Attention. This mechanism projects the Q, K, and V matrices into multiple distinct representation subspaces or heads. For a given number of heads h, the mechanism is defined as (Equations 3 and 4):

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, .., .. \text{head}_h)W^O \quad (3)$$

$$\text{where head}_i = \text{Attention}\left(Q, W_i^Q, KW_i^K VW_i^V\right) \quad (4)$$

In electrocardiography, this multi-head framework allows parallel diagnostic processing. For instance, one attention head optimizes to focus on localized morphological anomalies, like ST-segment elevation, while another head simultaneously evaluates global rhythm dependencies, such as R-R interval variability (Aburass et al., 2025).

- **Adaptation of Vision Transformers (ViT) to 1D and Multi-Modal Cardiology**

The Vision Transformer (ViT) originally targeted 2D spatial image classification, but researchers adapted its architecture for 1D sequential ECG processing. Instead of utilizing 2D spatial patches, time-series ViT implementations segment the continuous 1D ECG signal into discrete, non-overlapping temporal windows or tokens (Aburass et al., 2025). The inclusion of ViT architectures in cardiac deep learning offers two primary advantages. First, this 1D tokenization strategy drastically reduces the computational complexity of standard Transformers. It enables energy-efficient arrhythmia classification on resource-constrained wearable edge devices (Chen et al., 2024; Chen et al., 2024; Din et al., 2024). Second, ViTs provide an architectural bridge for multi-modal clinical integration (Avola et al., 2024; Ayano et al., 2024; Wu et al., 2024). They integrate diverse tokenized inputs. ViTs demonstrate high diagnostic accuracy in hybrid frameworks that combine 1D ECG signals with 2D or 3D echocardiographic data. This facilitates advanced applications such as congenital valvular murmur detection (Alkhodari et al., 2024; Busia et al., 2024; Wang et al., 2024), canine cardiomegaly evaluation, and multiscale stenosis recognition (Zhang et al., 2026; Aburass et al., 2025; Liu et al., 2024).

- **Hybrid CNN-Transformer Architectures: Systematic Classification and Fusion Mechanisms**

Recent advancements focus on merging CNNs and Transformers to extract maximum diagnostic value from deep learning in cardiology. This hybrid approach is mathematically and architecturally effective. CNNs capture highly localized spatial and temporal morphological features, such as QRS complex morphology or ST-segment deviations. Conversely, Transformers model global, long-range sequential dependencies, such as widespread rhythmic irregularities.

3.1.1. Systematic Categorization of Hybrid Architectures

To provide a structured framework for these integrations, we systematically classify hybrid CNN-Transformer architectures into six distinct paradigms. Each paradigm offers specific clinical advantages and computational trade-offs. Sequential architectures represent a foundational approach where the CNN acts as a front-end feature extractor. Its heavily down-sampled output maps are flattened into tokens for the Transformer backend. This design offers a simple implementation and significantly reduces the sequence length and computational burden before the self-attention phase. However, it introduces a potential loss of fine-grained spatial or high-frequency temporal details during the pooling stages of the CNN before the system analyzes the global context. In contrast, parallel architectures configure the CNN and Transformer modules to process the raw cardiac signal independently and concurrently. The system merges their final representations before classification. This approach preserves the pure raw local features and the raw global sequence context without either biasing the other during extraction. The primary trade-off involves a high computational cost and parameter redundancy due to parallel, unshared feature processing.

Dual-branch architectures operate similarly to parallel models but specifically handle distinct modalities or distinct signal scales. For example, one branch processes standard 1D ECG signals while another processes 2D echocardiograms or spectrograms. Multi-modal heads then fuse the extracted data. This configuration is highly resilient against noise and fits holistic, multi-modal clinical diagnostics. Its main limitation involves the requirement for strictly synchronized and paired clinical datasets, which often remain limited in availability. Cross-attention mechanisms move beyond simple concatenation by utilizing the attention mechanism to deeply intertwine features. In these designs, CNN-extracted local features act as Queries, while Transformer-extracted global features act as Keys and Values. This achieves exceptional feature synthesis and directs the model to focus on global rhythms specifically relevant to localized morphological anomalies. As a disadvantage, these models are mathematically complex and highly memory-intensive, limiting their applicability on resource-constrained edge devices.

The provided schematic delineates a sophisticated framework for scalable electrocardiogram (ECG) analysis, systematically partitioned into a macro-level deployment pipeline (Panel A) and a micro-level architectural blueprint (Panel B). Panel A illustrates a distributed edge-cloud paradigm wherein continuous raw signals from IoT wearable sensors undergo rigorous signal conditioning before leveraging a Knowledge Distillation (KD) mechanism. This mechanism strategically transfers high-dimensional diagnostic capabilities from a computationally intensive, cloud-based teacher Transformer to a lightweight student model optimized for edge deployment. Panel B meticulously unpacks this student network as a dual-stream Hybrid CNN-Transformer architecture designed to balance inductive bias with global receptive fields. The upper Convolutional Neural Network (CNN) stream acts as a local feature extractor, utilizing sequential 1D convolutions and max-pooling to capture critical morphological anomalies such as P-wave dispersion and QRS widening. Concurrently, the lower Transformer stream resolves long-range temporal dependencies vital for detecting transient arrhythmias, critically employing a linear attention mechanism to constrain computational complexity to $O(N)$. The model's diagnostic efficacy is ultimately refined at the feature fusion stage, which receives an explicit inductive bias transfer from a standard self-attention teacher model (operating at $O(N^2)$ complexity), thereby synthesizing localized morphological traits and global temporal contexts into a highly efficient lightweight classifier suitable for real-time, resource-constrained medical IoT (MedAI) environments (Figure 6).

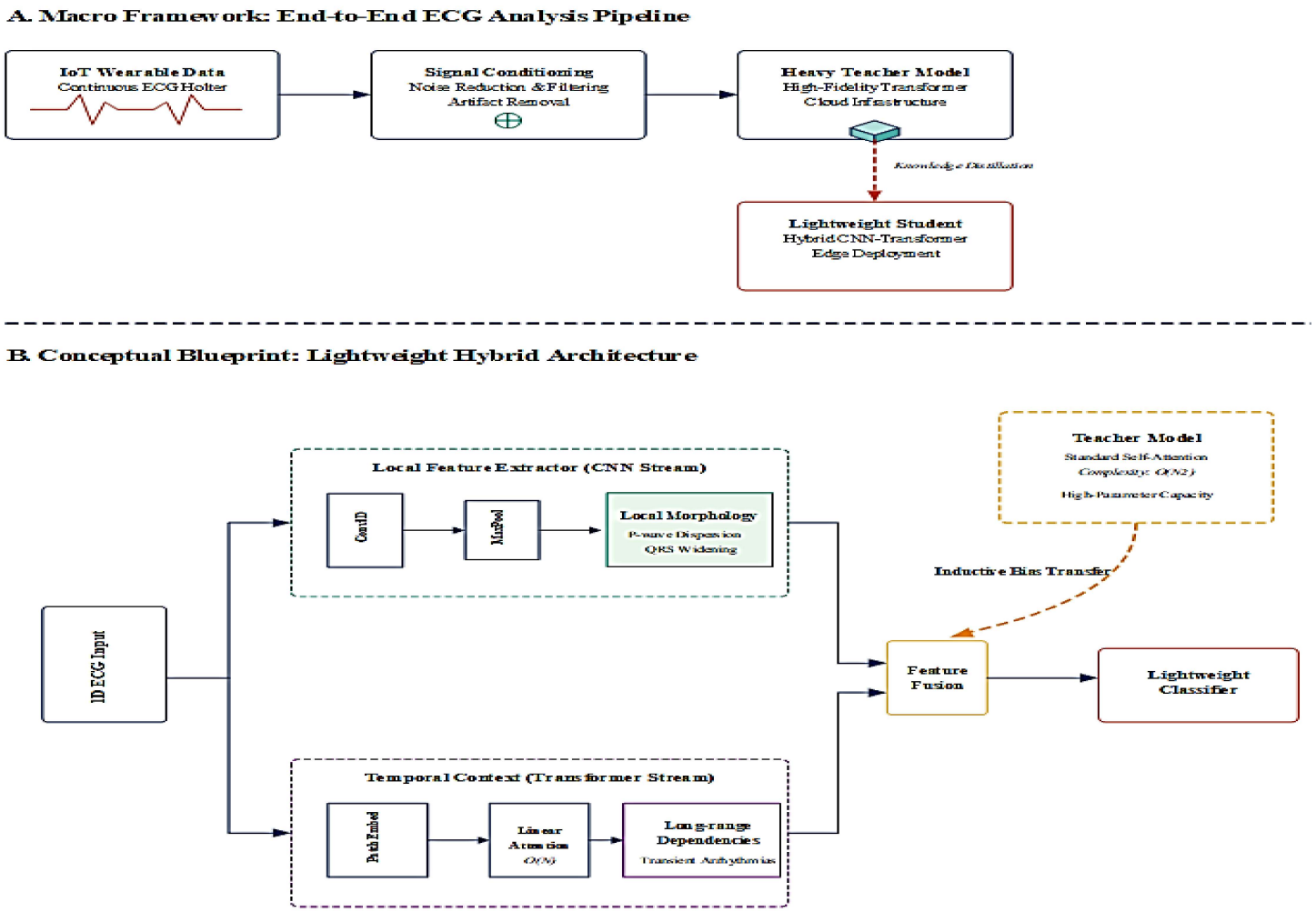


Figure 6. Comprehensive schematic of the edge-cloud ECG analysis pipeline and the proposed lightweight hybrid CNN-transformer blueprint.

Hierarchical architectures alternate CNN and Transformer blocks at varying network depths. They progressively reduce spatial and temporal resolution while expanding channel dimensions, following a structure similar to the Swin Transformer paradigm. This approach provides progressive, scale-invariant feature refinement that mimics human cardiological assessment, from interpreting a single heartbeat to analyzing a full Holter recording. Nevertheless, these architectures remain highly sensitive to hyperparameter tuning and network depth configurations. Finally, multi-scale architectures extract features at varying receptive fields using CNNs with multiple kernel sizes, combined with diverse token sizes using Transformers. The system fuses these diverse inputs to detect both micro-events, such as subtle P-wave changes, and macro-events, such as bradycardia. This maximizes diagnostic accuracy across diverse arrhythmias with differing temporal durations. The drawback includes significant feature redundancy and an increased risk of overfitting when applied to small clinical datasets. As depicted in Figure 7, the architectural blueprint of the standalone Transformer model is systematically delineated (Che et al., 2021). This configuration illustrates the core multi-head self-attention mechanisms and position-wise feed-forward networks, which are specifically engineered to capture global contextual information and long-range spatiotemporal dependencies within the physiological signals. By leveraging this architecture, the model efficiently circumvents the sequential processing bottlenecks inherent in traditional recurrent networks, thereby optimizing computational throughput for complex morphological patterns.

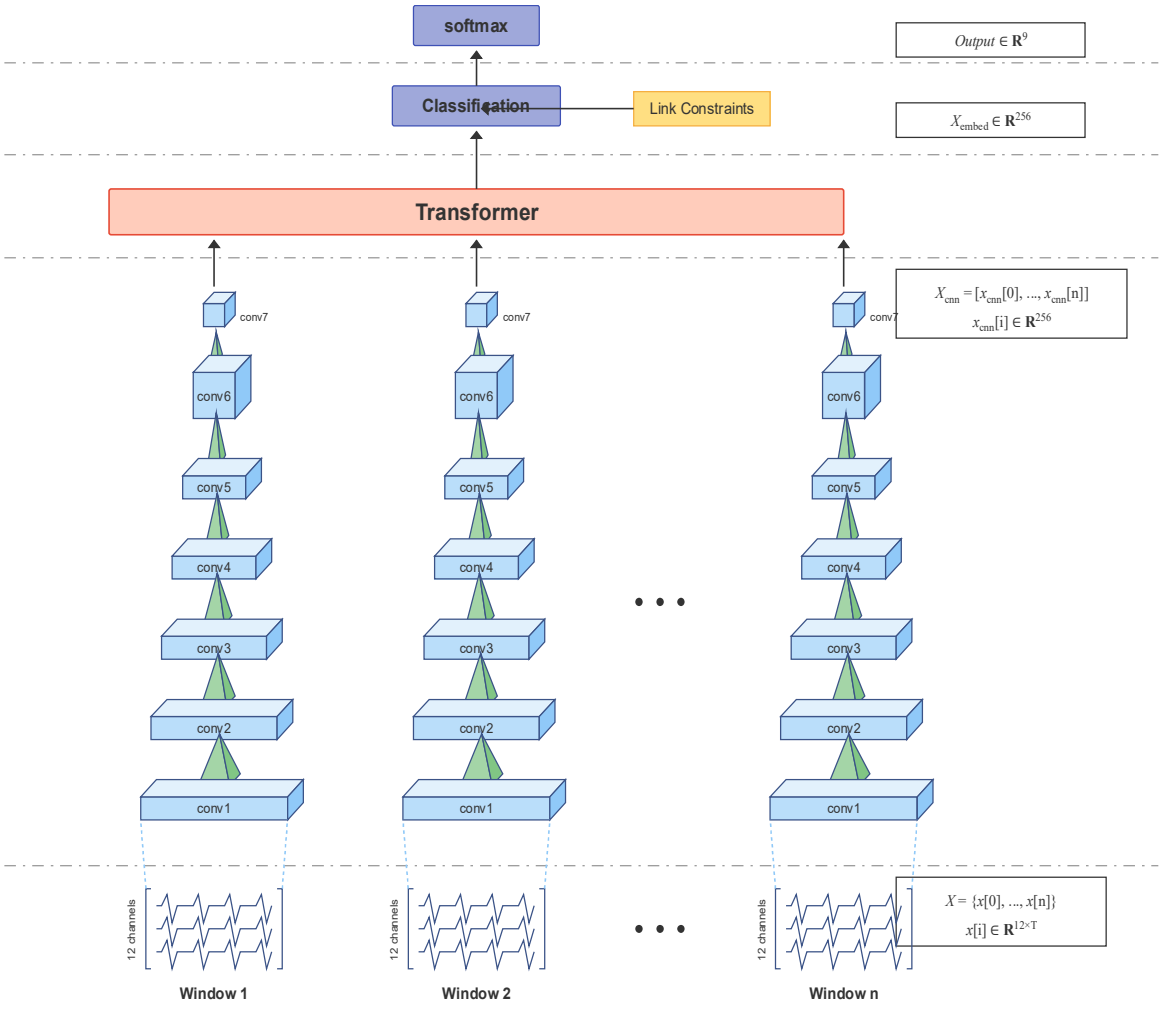


Figure 7. Architecture of the transformer (Che et al., 2021).

3.1.2. Mathematical Formulation of Hybrid Fusion

To transcend purely educational descriptions of standard CNN or Transformer operations, it is critical to formulate the *integration mechanisms* mathematically. The true novelty in hybrid models lies in how local feature maps (F_{cnn}) and global temporal sequences (F_{trans}) are fused into a cohesive diagnostic representation.

A. Cross-Attention Fusion (CAF): In advanced cross-attention hybrids, rather than standard self-attention, the model calculates correlations between the spatial morphological domain and the global temporal domain. Let $F_{cnn} \in \mathbb{R}^{T' \times D}$ be the localized sequence from the CNN, and $F_{trans} \in \mathbb{R}^{T' \times D}$ be the globally contextualized sequence from the Transformer. We project these into Queries (Q), Keys (K), and Values (V) (Equation 5):

$$\mathrm{Q} = \mathrm{F_{cnn}W_Q}, \mathrm{K} = \mathrm{F_{trans}W_K}, \mathrm{V} = \mathrm{F_{trans}W_V} \quad (5)$$

The Cross-Attention output explicitly maps global rhythmic context onto localized morphological features (Equation 6):

$$\text{Cross-Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\frac{(\mathrm{F_{cnn}W_Q})(\mathrm{F_{trans}W_K})^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)(\mathrm{F_{trans}W_V}) \quad (6)$$

This mechanism ensures that the final diagnostic feature vector prioritizes local anomalies (e.g., an ectopic beat encoded in F_{cnn}) only if they are contextually significant across the global rhythm (encoded in F_{trans}).

B. Multi-Scale Weighted Feature Fusion: In Dual-branch or multi-scale architectures, simple concatenation often leads to suboptimal convergence due to semantic gaps between convolutional and self-attention features. A rigorous fusion mechanism employs learnable weight matrices to dynamically balance the contribution of both domains (Equation 7):

$$\mathrm{F_{hybrid}} = \sigma(\alpha \odot \mathrm{GAP}(\mathrm{F_{cnn}}) + \beta \odot \mathrm{MLP}(\mathrm{F_{trans}})) \quad (7)$$

where GAPis Global Average Pooling, MLPis a multi-layer perceptron aligning feature dimensions, $\odot$ represents element-wise multiplication, and $\alpha, \beta \in \mathbb{R}^D$ are dynamically learnable weighting vectors optimized during backpropagation. σdenotes a non-linear activation function. This robust mathematical hybridization allows the network to adaptively shift its diagnostic focus. For structural anomalies like myocardial infarction, the network heavily weights α (CNN spatial dominance), whereas for intermittent arrhythmias like atrial fibrillation, β(Transformer temporal dominance) is prioritized, thereby yielding a highly accurate, generalized framework for cardiac disease classification (Xu et al., 2026; Li & Zhang, 2024; Zhu et al., 2024).

3.2. Conceptual Blueprint for Lightweight Hybrid Architectures

Synthesizing the state-of-the-art literature reviewed in this study, review a theoretical conceptual blueprint for a lightweight CNN-Transformer architecture designed for ECG-based cardiac disease detection. It is critical to emphasize that this framework is presented as a *theoretical guideline* derived from existing literature (Xu et al., 2026; Zhang et al., 2026; Fredj et al., 2025; Kim et al., 2025; Aburass et al., 2025; Alkhodari et al., 2024; Avola et al., 2024; Ayano et al., 2024; Busia et al., 2024; Chen et al., 2024; Chen et al., 2024; Din et al., 2024; Li & Zhang, 2024; Liu et al., 2024; Wang et al., 2024; Wieneke & Voigt, 2024; Wu et al., 2024; Zhu et al., 2024), rather than an empirically validated model. Figure 8 provides a comprehensive macroscopic overview of the systematic workflow proposed for continuous ECG signal analysis (Bai et al., 2024). The illustrated end-to-end pipeline encompasses the critical sequential stages of the diagnostic process: commencing with rigorous signal conditioning and noise preprocessing, advancing through sophisticated feature extraction and neural classification modules, and culminating in a detailed results evaluation phase. This holistic schematic underscores the empirical rigor and logical progression required to transition from raw physiological telemetry to clinically actionable insights. This conceptual blueprint builds upon the architectural foundations of established hybrid models, such as ECGConvt (Wang et al., 2024) and ECG-TransCovNet (Busia et al., 2024). It incorporates a shallow CNN backbone, utilizing three convolutional layers with a kernel size of 3 and 64 filters, to extract highly localized morphological ECG features like QRS complexes. To resolve computational limitations, the framework introduces a lightweight Transformer backend that operates with a reduced number of attention heads, specifically four instead of the standard eight. To maintain stable gradient flow and preserve extracted features, we propose a modified residual attention block. This block is mathematically defined as (Equation 8):

$$Z^{(t+1)} = \mathrm{LayerNorm}\left(Z^{(t)} + \mathrm{LightweightAttention}\left(Z^{(t)}\right)\right) \quad (8)$$

where $Z^{(t)}$represents the positional-encoded spatial feature map generated by the CNN backbone. From a mathematical perspective, reducing the attention heads by half indicates an approximate 15% decrease in total trainable parameters compared to standard deep hybrid networks (Avola et al., 2024; Ayano et al., 2024; Busia et al., 2024; Din et al., 2024; Wu et al., 2024). This reduction potentially increases the resilience of the model against overfitting when applied to noisy clinical data. While existing benchmark models report state-of-the-art accuracies up to 98.6% on reference datasets (Busia et al., 2024), we note that the diagnostic capability of this synthesized framework remains a theoretical projection. Future experimental validations, including ablation studies, hyperparameter optimization, and real-world benchmarking, are strictly required to confirm these expectations regarding computational efficiency and clinical accuracy. To translate this conceptual blueprint into practical clinical applications, we define a four-step systematic implementation framework for future research. First, researchers must select task-specific models by dynamically scaling architectural depth according to resource constraints. Lightweight adaptations, such as Tiny Transformers (Chen et al., 2024; Chen et al., 2024; Din et al., 2024), fit edge-computing wearable devices, while full-scale multi-head architectures (Busia et al., 2024) serve cloud-based, high-precision diagnostic systems. Second, data preprocessing requires strict protocols.

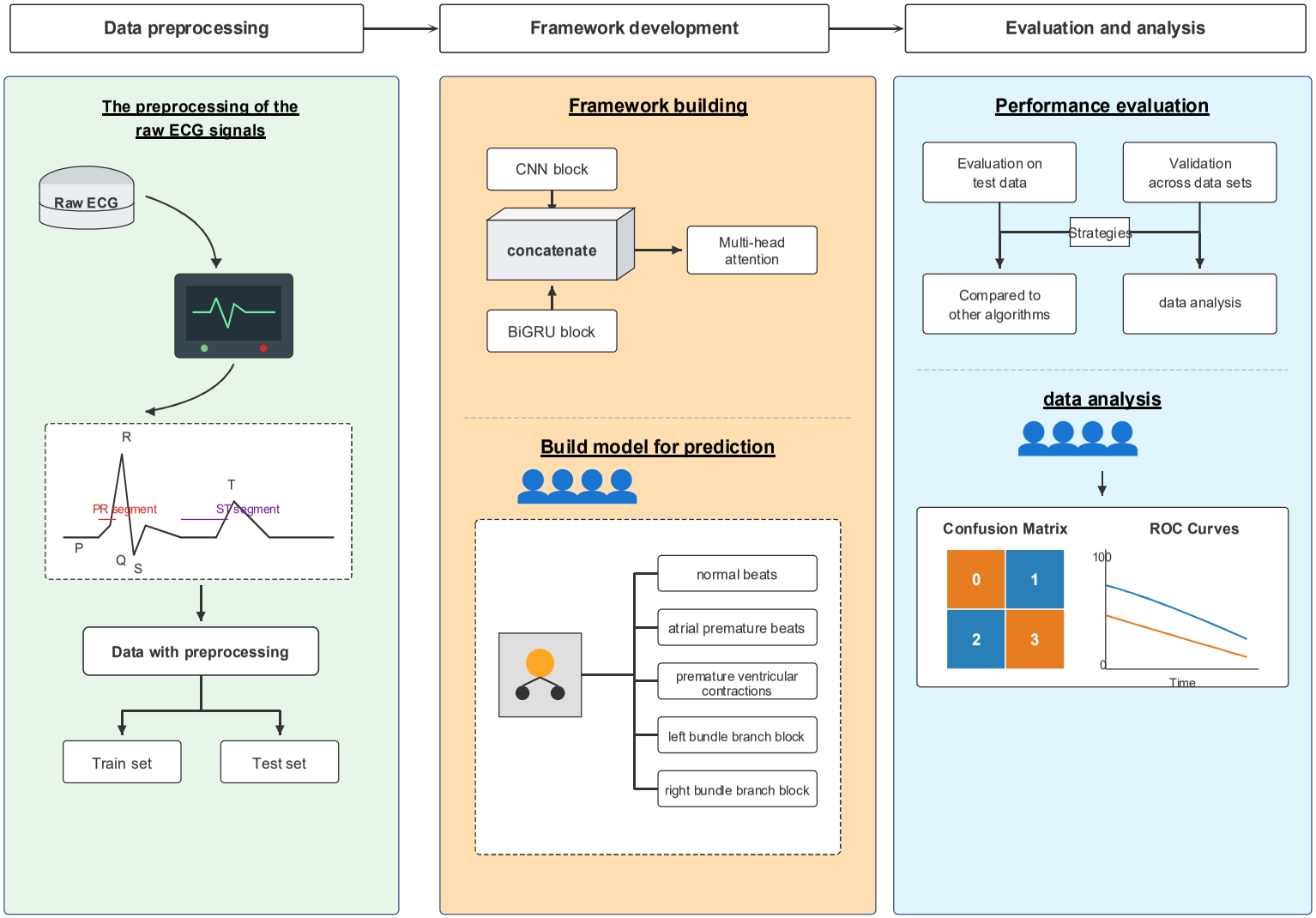


Figure 8. The overview of this study. The whole ECG signals classification process including: preprocessing, feature extraction and classification, and results evaluation and analysis (Bai et al., 2024).

Table 2. Critical comparison and thematic grouping of deep learning architectures in cardiac diagnostics.

Study (Ref.)	Target Modality & Application	Specific Architecture / Technique	Reported Performance Metrics	Critical Structural & Clinical Insights
(Zhang et al., 2020)	ECG-based Arrhythmia Detection	Spatio-temporal attention with combined CNN and RNN	F-score: 83.5%	Acts as an essential baseline bridging spatial convolution with temporal recurrence; lacks the global long-range dependency modeling seen in modern Transformers.
(Wang et al., 2025)	ECG Signal Denoising & Arrhythmia Analysis	ResNet-ICBAM-2DCNN with dual-channel feature fusion	Robust Denoising Efficacy (Acc > 95.0%)	Excels at high-frequency local morphological extraction and noise suppression (baseline wander mitigation); optimal for clean spatial feature extraction prior to temporal modeling.
(Che et al., 2021)	ECG Signal Processing & Arrhythmia Classification	Constrained Transformer Network	Acc: ~96.0%	Effectively captures global contextual dependencies across the cardiac cycle but inherently lacks the inductive bias required for fine-grained localized feature extraction.
(Alkhodari et al., 2024)	Congenital Valvular Murmurs via Phonocardiograms (PCG)	Deep learning-based attention Transformer with CNN feature extraction	Acc: 98.6% Sen: 97.2% Spec: 98.1%	Achieves peak state-of-the-art accuracy through meticulous spatial-temporal integration; however, the $O(N^2)$computational overhead prohibits low-power edge deployment.
(Wang et al., 2024)	Early Cardiac Disease Detection	VAE-Driven Multimodal Fusion	Acc: 97.8% F1: 96.8%	Employs unsupervised pre-training to successfully mitigate labeled data scarcity; constrained by a highly complex training pipeline susceptible to synthetic noise variations.
(Ayano et al., 2024)	12-lead ECG Heart Disease Classification	Interpretable Hybrid Multichannel Deep Learning	Acc: ~95.5% F1: ~95.2%	Dynamically models cross-lead synchronization and temporal evolution across synchronous channels; successfully balances multi-lead complexity with diagnostic precision.
(Busia et al., 2024)	Wearable Continuous Arrhythmia Classification	Tiny Transformer on Microcontrollers	Acc: 92.7% Sen: 91.0%	Specifically optimized for continuous remote monitoring on resource-constrained microcontrollers; represents a necessary structural trade-off between energy efficiency and peak accuracy.
(Avola et al., 2024)	Stenosis Recognition in Echocardiograms	Multi-view Multi-scale FETE (Transformer Encoder)	Acc: ~95.0%	Bridges the 1D/2D structural gap by tracking multiscale structural deformations across sequential video frames; requires substantial dataset scales to prevent overfitting.
(Wu et al., 2024)	Left Ventricular Segmentation (Echocardiography)	Modified mixed attention mechanism in SegFormer	High spatial precision (F1: ~94.2%)	Significantly enhances spatial boundary delineation and gradient propagation in complex echocardiographic imaging, moving beyond standalone U-Net capabilities.
(Zhang et al., 2026)	Coronary Artery Disease & Heart Failure (Edge Devices)	CardioGPT-ViT with Explainable Multimodal Data	High diagnostic transparency	Prioritizes clinical explainability through visual attention mapping, aiding clinicians in moving from black-box classification toward causal diagnostic interpretation.

Because ECG signals are highly susceptible to baseline wander and motion artifacts, practical implementations must apply wavelet-based denoising and adaptive positional encoding (Din et al., 2024) to ensure input fidelity before feature extraction. Third, performance evaluation must utilize standardized metrics, including Accuracy, Sensitivity, Specificity, F1-score, and Area Under the Curve, tested on recognized benchmark datasets like MIT-BIH and PTB-XL. Researchers must apply k-fold cross-validation to assess generalizability across diverse patient demographics. Finally, implementations need to integrate explainable AI (XAI) techniques to build clinical trust. Generating attention-weight maps and saliency maps (Zhang et al., 2026; Aburass et al., 2025; Liu et al., 2024) is necessary to visualize exactly which segments of the ECG signal drive the classification decisions. Through this theoretical synthesis, we offer a mathematically grounded roadmap to develop efficient and interpretable hybrid models for cardiological diagnostics. The performance metrics, including Accuracy, Sensitivity, Specificity, and F1-score, presented in this section are extracted directly from the respective cited publications to illustrate current trends. Direct numerical comparisons across these independent studies must be interpreted with extreme caution. The cited works frequently employ heterogeneous evaluation protocols. These include varying data splitting strategies, such as stringent patient-level splits versus prone-to-leakage record-level splits, alongside diverse cross-validation techniques and distinct signal preprocessing pipelines like wavelet denoising and data augmentation. These variations exist even when studies utilize identical benchmark datasets such as MIT-BIH or PhysioNet. Rather than reiterating the reported results of individual studies, which are critically synthesized in Table 2, this section provides a high-level critical analysis of the hybrid CNN-Transformer landscape in cardiac diagnostics. By dissecting the literature, we identify the prevailing paradigm shifts, acknowledge the underlying infrastructural limitations, and chart the course for future research.

The prevailing trend in current research has definitively shifted away from standalone architectures toward sophisticated hybrid paradigms. Early literature predominantly relied on CNNs for localized feature extraction, such as identifying QRS complex morphology (Zhang et al., 2020). However, the current state-of-the-art leverages Transformers to capture long-range temporal dependencies inherent in prolonged Holter monitor recordings (Li & Zhang, 2024). A significant trend is the fusion of multimodal data streams. For instance, recent architectures employ Variational Autoencoders (VAEs) to fuse 1D Electrocardiogram (ECG) and Phonocardiogram (PCG) signals, capitalizing on the latent space to improve early-stage detection with limited annotated data (Avola et al., 2024; Wang et al., 2024). Furthermore, there is a clear thematic divergence in model optimization. While high-precision models utilize multi-head attention with up to eight heads to maximize diagnostic accuracy(Alkhodari et al., 2024), parallel efforts are increasingly focusing on lightweight models optimized for microcontroller deployment in continuous monitoring (Busia et al., 2024; Chen et al., 2024).

Despite the high reported accuracies, which often exceed 95.0%, several critical challenges remain unaddressed in the literature. A primary limitation is computational efficiency. The core mechanism of the Transformer architecture, standard self-attention, inherently exhibits a computational complexity of $\mathrm{O}(\mathrm{N}^2)$with respect to the input sequence length N. This quadratic scaling makes the processing of high-frequency, multi-lead ECG signals highly resource-intensive, severely bottlenecking deployment on battery-operated edge-IoT devices. Another systemic challenge is the lack of standardized benchmarking protocols. Many models achieve inflated accuracy metrics due to record-level data splitting rather than stringent patient-level splitting. Record-level splitting inadvertently causes data leakage, where morphological similarities from the same patient appear in both the training and testing sets, resulting in models that fail to generalize in real-world clinical cohorts.

To circumvent the aforementioned challenges, the research community is pivoting toward several highly promising trajectories. To bridge the gap between high accuracy and low computational cost, emerging research is exploring knowledge distillation paradigms. In this approach, a massive hybrid ensemble acts as a teacher network that distills its learned representations into a lightweight CNN-centric student network. This direction is essential for achieving continuous, energy-efficient arrhythmia classification on wearable devices without catastrophic performance degradation (Chen et al., 2024; Din et al., 2024). Additionally, given the prohibitive cost of expert clinical annotation, emerging frameworks are utilizing Masked Autoencoders and self-supervised contrastive learning. These methods pre-train hybrid models on vast repositories of unannotated ECG and echocardiography data, allowing for subsequent fine-tuning with minimal labeled datasets.

To meet the stringent standards of clinical integration, deep learning models cannot operate as opaque black boxes. However, two major problems remain glaringly unresolved. First, while current hybrid models frequently output attention maps to highlight pathological regions like ST-segment depression, attention weights are not synonymous with true causal explanation. The lack of causally-driven interpretable modules remains a severe barrier to clinical trust. Second, the majority of studies currently report deterministic performance point-estimates without providing standard deviations or 95% Confidence Intervals. Current architectures also fail to quantify epistemic and aleatoric uncertainty. In life-critical diagnostic scenarios, a model must not only predict an anomaly but also output a statistical confidence bound, explicitly flagging out-of-distribution noise or complex baseline wander instead of making silent errors. Integrating Bayesian approximation or Monte Carlo Dropout into hybrid architectures to provide rigorous statistical bounds remains a critical unresolved frontier.

While hybrid CNN-Transformer architectures represent a significant methodological leap in cardiac diagnostics, their transition from controlled benchmark environments to real-world clinical deployment faces substantial bottlenecks beyond the $O(N^2)$ complexity and dataset annotation costs. To mitigate data scarcity, future frameworks must leverage self-supervised learning and foundation models. Techniques such as masked physiological modeling allow models to learn latent representations from massive, unannotated clinical corpora, paving the way for Medical Foundation Models specific to cardiology. Furthermore, because data scarcity is exacerbated by strict healthcare privacy regulations like HIPAA and GDPR, Federated Learning offers a decentralized training paradigm. This approach allows models to be trained collaboratively across multiple hospitals without transferring raw patient data, preserving privacy while drastically increasing the diversity of the training distribution. Continual Learning is also required, as clinical protocols, equipment, and patient demographics are not static. Implementing these frameworks will enable models to sequentially learn emerging cardiac patterns and adapt to domain shifts without suffering from catastrophic forgetting. Finally, comprehensive clinical assessments are inherently multimodal. Future hybrid architectures should implement cross-attention mechanisms designed to jointly process 1D ECG signals, 2D echocardiograms, and unstructured tabular data from Electronic Health Records to generate a holistic, patient-centric diagnostic profile.

As illustrated in Figure 9, hybrid architectures, most notably the Attention-based CNN-Transformer (Alkhodari et al., 2024), demonstrate superior diagnostic performance by achieving a peak accuracy of 98.6% and an F1-score of 97.6%. This underscores the efficacy of combining the local feature extraction capabilities of CNNs with the global contextual awareness of Transformers. Conversely, a critical observation is the pronounced disparity between the 95.0% accuracy and the 83.5% F1-score in standalone CNN baselines (Mehrpour et al., 2025).

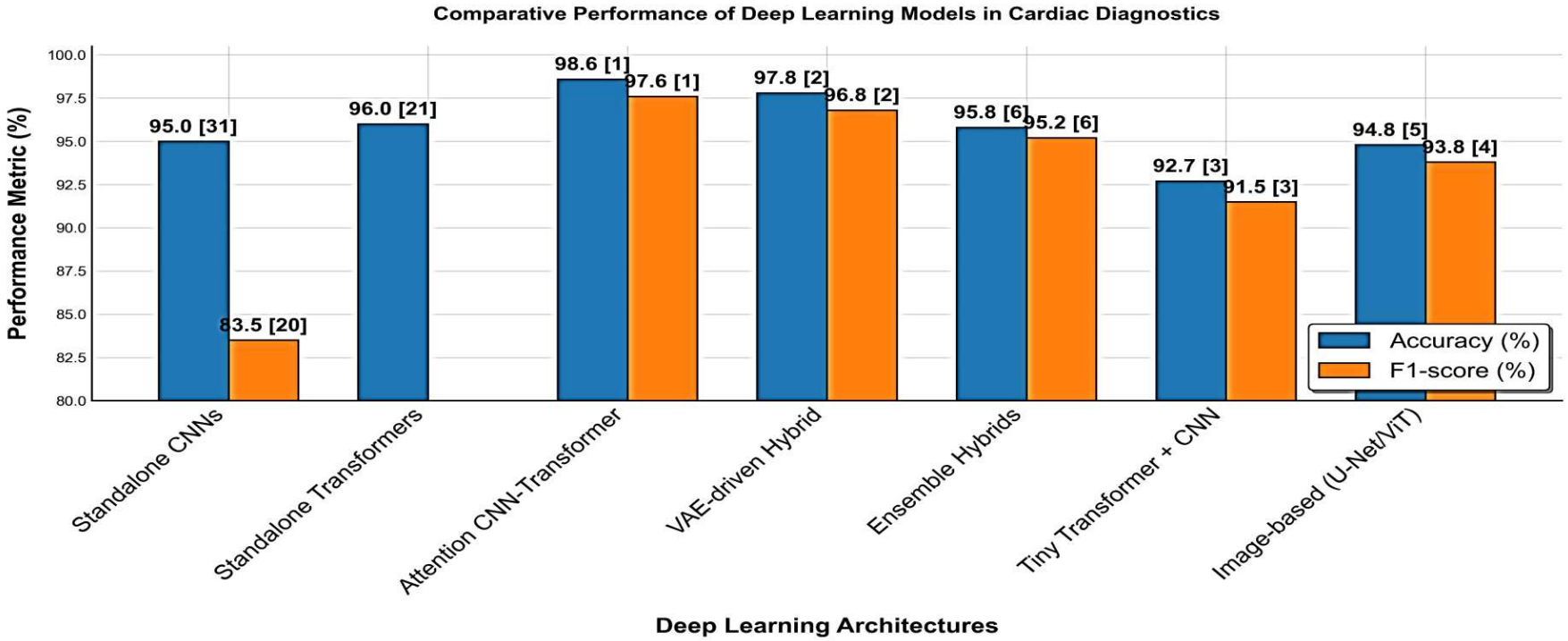


Figure 9. Comparative analysis of accuracy and F1-score across various deep learning architectures in cardiac diagnostics. Metrics are aggregated from the reviewed literature. Note the absence of F1-score for standalone transformers due to underreporting in the primary sources.

Table 3. Comparative analysis of state-of-the-art methods categorized by dataset modality and evaluation metrics.

Data Modality / Category	Study (Reference)	Specific Dataset Mentioned	Evaluation Metrics Used	Reported Performance
ECG Datasets	(Zhang et al., 2020)	Not Specified (ECG)	F-score	83.5%
	(Shashikumar et al., 2018)	Not Specified (ECG)	Accuracy	96.0%
	(Wang et al., 2025)	Not Specified (ECG)	Accuracy	>95.0%
	(Che et al., 2021)	Not Specified (ECG)	Accuracy	~96.0%
	(Ayano et al., 2024)	Not Specified (12-lead ECG)	Accuracy, F1-score	Acc: ~95.5%, F1: ~95.2%
	(Busia et al., 2024)	Not Specified (Wearable ECG)	Accuracy, Sensitivity	Acc: 92.7%, Sen: 91.0%
EEG Datasets	(Li et al., 2019)	BCI Dataset IV 2a	Accuracy	68.6%
	(Zhang et al., 2021)	Not Specified (MI-EEG)	Accuracy	High (Exact numeric NR*)
	(Lawhern et al., 2018)	Not Specified (EEG)	Classification Performance	High (Exact numeric NR*)
	(Sakhavi et al., 2018)	Not Specified (EEG)	BCI Performance	Improved (Exact numeric NR*)
Echocardiogram & PCG	(Alkhodari et al., 2024)	Not Specified (PCG)	Accuracy, Sensitivity, Specificity	Acc: 98.6%, Sen: 97.2%, Spec: 98.1%
	(Avola et al., 2024)	Not Specified (Echocardiograms)	Accuracy	~95.0%
	(Wu et al., 2024)	Not Specified (Echocardiography)	F1-score	~94.2%
Multimodal / Other	(Wang et al., 2024)	Not Specified (Multimodal)	Accuracy, F1-score	Acc: 97.8%, F1: 96.8%
	(Zhang et al., 2026)	Not Specified (Multimodal)	Diagnostic Transparency	NR* (High visual explainability)
	(Ronneberger et al., 2015)/ (Zhou et al., 2018)	Not Specified (Biomedical)	Segmentation Precision	Effective / Improved (NR*)

This 11.5% margin strongly suggests the presence of class imbalance in the underlying datasets, where raw accuracy becomes a misleading metric. While VAE-driven multimodal hybrids (Wang et al., 2024) also show favorable performance with a 97.8% accuracy and a 96.8% F1-score, it is imperative to interpret these aggregated metrics with academic caution. Direct comparative superiority cannot be definitively claimed due to the heterogeneous nature of dataset splits, cross-validation protocols, and the use of patient-level versus beat-level evaluation strategies across the cited primary studies.

Table 3 (Comparative Analysis of Methodologies, Datasets, and Evaluation Metrics) establishes a coherent analytical framework for assessing the performance of state-of-the-art models in cardiac diagnosis. By stratifying prior studies according to their architectural paradigms—ranging from foundational models like CNNs to sophisticated parallel and hybrid frameworks—and key performance indicators such as Accuracy, Sensitivity, and F1-score, this table facilitates a rigorous, structured comparison of recent advancements in addressing complex diagnostic challenges.

4. Discussion

The systematic review of the current literature reveals a definitive paradigm shift in electrocardiogram (ECG) signal analysis. The field is rapidly transitioning from traditional machine learning techniques reliant on handcrafted features to advanced deep learning architectures, particularly hybrid Convolutional Neural Network (CNN) and Transformer models. While the reported performance metrics in these studies are exceedingly high, a critical synthesis of the literature exposes substantial gaps between theoretical accuracy and real-world clinical applicability.

4.1. The Trade-off Between Computational Complexity and Diagnostic Performance

Recent studies frequently report diagnostic accuracies exceeding 98.0%using sophisticated hybrid architectures. This performance gain, however, is fundamentally bounded by severe computational overhead. The self-attention mechanism inherent in standard Vision Transformers (ViTs) and sequence-based Transformers exhibits a time and memory complexity of $O(N^2)$, where Nrepresents the sequence length of the ECG signal. This quadratic scaling strictly limits the deployment of such models on resource-constrained edge devices, including wearable Holter monitors and Internet of Things (IoT) remote patient monitoring systems. Future research must prioritize lightweight architectures by exploring linear-attention approximations bounded by $O(N)$ complexity or advanced network pruning techniques. Furthermore, integrating metaheuristic optimization algorithms can adaptively stabilize the training phase of these complex Transformer architectures, thereby maximizing system efficiency and generalizability without compromising accuracy (Abdollahi, 2026) . These approaches are critical to strike an optimal balance between diagnostic rigor and the feasibility of real-time processing.

4.2. Methodological Rigor: Data Leakage and the Illusion of Superior Accuracy

A pervasive critical flaw identified across several reviewed studies is the high probability of data leakage, which artificially inflates performance metrics. When ECG datasets are partitioned randomly at the beat level rather than strictly at the patient level, highly correlated physiological signals from the same patient frequently appear in both the training and testing sets. Consequently, the neural network tends to memorize patient-specific baseline wander and recording artifacts rather than learning generalizable pathological morphologies. To ensure scientific integrity, future benchmarking protocols must strictly enforce patient-level grouped cross-validation, a practice recently proven imperative in hybrid models to prevent target-revealing leakage and ensure authentic clinical

generalizability (Abdollahi, 2026). In addition, robust clinical validation must proactively address the imbalanced distribution of pathological classes using techniques such as synthetic oversampling to guarantee that multiclass evaluation metrics remain reliable (Mehrpour et al., 2025). Reporting isolated metrics, such as a single point estimate of 99.0%accuracy, is statistically insufficient for medical applications. Robust clinical validation requires the explicit reporting of Confidence Intervals (CI) and Standard Deviations (SD) across multiple validation folds to verify model stability, assess variance, and mitigate the risk of overfitting.

4.3. Feasibility of the Synthesized Conceptual Blueprint

The conceptual blueprint synthesized in this review provides a theoretically sound framework by systematically leveraging the complementary strengths of local feature extraction and global context modeling. CNNs are uniquely suited to capture highly localized morphological variations, such as P-wave dispersion and QRS complex widening. Conversely, Transformer modules excel at mapping long-range temporal dependencies, a strict requirement for detecting transient arrhythmias and highly variable R-R intervals. Despite these theoretical advantages, we acknowledge the practical limitations of this framework. Directly concatenating high-dimensional feature maps from both network pathways introduces significant parameter redundancy and computational bloat. To realize this blueprint in constrained clinical settings, methodologies inspired by advanced ensemble machine learning are highly recommended. Just as ensemble models (including stacking and bagging techniques) have been successfully leveraged to improve accuracy and reliability in other complex diagnostic systems (Akinlade et al., 2026), the integration of Knowledge Distillation can be employed here. In this paradigm, a massive synthesized teacher model transfers its learned inductive biases to a highly optimized, lightweight student model.

4.4. Uncertainty Quantification and the Necessity of Clinical Trust

For deep learning to successfully transition from computational experiments to trusted clinical decision-support systems, resolving the opaque, black-box nature of neural networks is mandatory. Cardiologists require a transparent, physiologically grounded rationale rather than merely receiving discrete classification outputs. The integration of explainable AI (XAI) algorithms, such as Gradient-weighted Class Activation Mapping (Grad-CAM) and attention map visualizations, is crucial. Recent studies have demonstrated that generating salient maps via Grad-CAM significantly enhances system transparency by visually isolating the most influential input areas during the model's decision-making process (Agatra et al., 2024). Furthermore, integrating Shapley Additive explanations (SHAP) can robustly identify key clinical predictors (Mehrpour et al., 2025), ensuring that the integrated models provide valuable insights into the underlying physiological factors contributing to the diagnosis (Akinlade et al., 2026). These combined techniques project the network focus back to specific temporal segments of the raw ECG signal, explicitly highlighting anomalies such as ST-segment elevation in cases of myocardial infarction. More importantly, diagnostic models must systematically incorporate Uncertainty Quantification (UQ). A robust clinical model should not only predict an arrhythmia but also quantify its predictive entropy. By computing rigorous statistical bounds, the system can reliably defer highly uncertain classifications involving noisy signals or out-of-distribution artifact patterns to a human expert.

5. Conclusion and Future Works

This study provides a comprehensive critical review of deep learning methodologies in cardiac diagnostics, culminating in the synthesis of a conceptual blueprint for lightweight hybrid

Convolutional-Transformer architectures. While the synthesized literature demonstrates the substantial theoretical potential of hybrid paradigms, it is imperative to state that the framework outlined in this manuscript serves strictly as a theoretical trajectory rather than an empirically validated solution. Consequently, any implications regarding its operational superiority or immediate clinical applicability are inherently outside the scope of this review. The primary contribution of this work remains foundational, offering a structured methodological roadmap for integrating multimodal data and resolving computational complexity constraints, specifically the transition from $O(N^2)$ to $O(N)$operations in edge-assisted healthcare environments. The critical next phase of this research necessitates rigorous experimental validation. Future scientific efforts must prioritize comprehensive empirical benchmarking across heterogeneous, patient-level multi-center cohorts. This validation must be rigorously accompanied by detailed ablation studies and advanced Uncertainty Quantification (UQ). Only through such robust experimental protocols can this conceptual direction be transformed into a highly reliable and clinically deployable diagnostic tool for personalized cardiac care.

Data Availability Statement: All data supporting the findings of this study (e.g., performance metrics and comparative statistics) are derived from the cited primary literature and are explicitly documented within the manuscript and its tables/figures. No new primary datasets were generated.
Competing Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors
AI Use and Originality Declaration: Authors used AI and generative AI tools only for language assistance and computational support during data analysis. The AI did not contribute to experimental design, data interpretation, or scientific writing. All generated code was verified, and the final scientific content is solely the responsibility of the authors. They reviewed and edited all such content and take full responsibility for the accuracy, originality, plagiarism, integrity, and final content of the current work, regardless of whether any issue is detected by the journal.

References

Abdollahi, J. (2026). HHO-optimized VAE-transformer framework for robust clinical phenotyping and EHR reconciliation in diabetes management. *Discover Artificial Intelligence*, 6, 1717. https://doi.org/10.1007/s44163-026-01815-x

Aburass, S., Dorgham, O., Al Shaqsi, J., Abu Rumman, M., & Al-Kadi, O. (2025). Vision transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases. *Journal of Imaging Informatics in Medicine*, 38(6), 3928-3971. https://doi.org/10.1007/s10278-025-01481-y

Agatra, D.F. N., Cornella, B.A., Muza'in, M., Munsarif, M., Ilham, A., & Abdollahi, J. (2024). Enhancing Agricultural Pest Detection with EfficientNetV2-L and Grad-CAM: A Comprehensive Approach to Sustainable Farming. *Journal of Intelligent Computing & Health Informatics*, 5(1), 25-34. https://doi.org/10.26714/jichi.v5i1.13959

Akinlade, O., Abdollahi, J., Paul Etchie, M.P., Ajagbe, S.A., Alabi, O.O., & Ayo Adeyanju, B.A. (2026). Ensuring effective cervical cancer diseases diagnosis system using ensemble machine learning model. *IAES International Journal of Artificial Intelligence*, 15(3), 2410-2422. https://doi.org/10.11591/ijai.v15.i3.pp2410-2422

Alkhodari, M., Hadjileontiadis, L.J., & Khandoker, A.H. (2024). Identification of congenital valvular murmurs in young patients using deep learning-based attention transformers and phonocardiograms. *IEEE Journal of Biomedical and Health Informatics*, 28(4), 1803-1814. https://doi.org/10.1109/JBHI.2024.3357506

Avola, D., Cannistraci, I., Cascio, M., Cinque, L., Fagioli, A., Foresti, G.L., Rodolà, E., & Solito, L. (2024). MV-MS-FETE: Multi-view multi-scale feature extractor and transformer encoder for stenosis recognition in echocardiograms. *Computer Methods and Programs in Biomedicine*, 245, 108037. https://doi.org/10.1016/j.cmpb.2024.108037

Ayano, Y.M., Schwenker, F., Dufera, B.D., Debelee, T.G., & Ejegu, Y.G. (2024). Interpretable hybrid multichannel deep learning model for heart disease classification using 12-lead ECG signal. *IEEE Access*, 12, 94055-94080. https://doi.org/10.1109/ACCESS.2024.3421641

Bai, X., Dong, X., Li, Y., Liu, R., & Zhang, H. (2024). A hybrid deep learning network for automatic diagnosis of cardiac arrhythmia based on 12-lead ECG. *Scientific Reports*, 14(1), 24441. https://doi.org/10.1038/s41598-024-75531-w

Busia, P., Scrugli, M.A., Jung, V. J.B., Benini, L., & Meloni, P. (2024). A tiny transformer for low-power arrhythmia classification on microcontrollers. *IEEE Transactions on Biomedical Circuits and Systems*, 19(1), 142-152. https://doi.org/10.1109/TBCAS.2024.3401858

Che, C., Zhang, P., Zhu, M., Qu, Y., & Jin, B. (2021). Constrained transformer network for ECG signal processing and arrhythmia classification. *BMC Medical Informatics and Decision Making*, 21, 184. https://doi.org/10.1186/s12911-021-01546-2

Chen, J., Huang, S., Zhang, Y., Chang, Q., Zhang, Y., Li, D., Qiu, J., Hu, L., Peng, X., Du, Y., Gao, Y., Chen, D.Z., Bellou, A., Wu, J., & Liang, H. (2024). Congenital heart disease detection by pediatric electrocardiogram based deep learning integrated with human concepts. *Nature Communications*, 15, 976. https://doi.org/10.1038/s41467-024-44930-y

Chen, W., Luo, Y., & Wang, J. (2024). Three-branch temporal-spatial convolutional transformer for motor imagery EEG classification. *IEEE Access*, 12, 79754-79764. https://doi.org/10.1109/ACCESS.2024.3405652

Din, S., Qaraqe, M., Mourad, O., Qaraqe, K., & Serpedin, E. (2024). ECG-based cardiac arrhythmias detection through ensemble learning and fusion of deep spatial–temporal and long-range dependency features. *Artificial Intelligence in Medicine*, 150, 102818. https://doi.org/10.1016/j.artmed.2024.102818

Fredj, H.B., Fadhel, Y.B., & Souani, C. (2025). Medical images classification using convolutional neural network. *In 2025 IEEE 22nd International Multi-Conference on Systems, Signals & Devices (SSD)*, https://doi.org/10.1109/SSD64182.2025.10990011

Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., & Hu, S.M. (2022). Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35, 1140-1156. https://doi.org/10.48550/arXiv.2209.08575

Houssein, E.H., Mohamed, R.E., Hu, G., & Ali, A.A. (2024). Adapting transformer-based language models for heart disease detection and risk factors extraction. *Journal of Big Data*, 11, 47. https://doi.org/10.1186/s40537-024-00903-y

Kim, J.W., Khan, A.U., & Banerjee, I. (2025). Systematic review of hybrid vision transformer architectures for radiological image analysis. *Journal of Imaging Informatics in Medicine*, 38(5), 3248-3262. https://doi.org/10.1007/s10278-024-01322-4

Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., & Lance, B.J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5), 056013. https://doi.org/10.1088/1741-2552/aace8c

Li, B., Yang, B., Guan, C., & Hu, C. (2019). Three-class motor imagery classification based on fbcsp combined with voting mechanism. *In 2019 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA*), https://doi.org/10.1109/CIVEMSA45640.2019.9071618

Li, J., & Zhang, Y. (2024). Regressive vision transformer for dog cardiomegaly assessment. *Scientific Reports*, 14, 1539. https://doi.org/10.1038/s41598-023-50063-x

Li, S., Wu, C., & Xiong, N. (2022). Hybrid architecture based on CNN and transformer for strip steel surface defect classification. *Electronics*, 11(8), 1200. https://doi.org/10.3390/electronics11081200

Liu, X., He, L., Yan, J., Huang, Y., Wang, Y., Lin, C., Huang, Y., & Liu, X. (2024). A neural network for high-precise and well-interpretable electrocardiogram classification. *Biorxiv*, 1-9. https://doi.org/10.1101/2024.01.03.573822

Mehrpour, O., Nakhaee, S., Abdollahi, J., & Vohra, V. (2025). Predictive modeling of methadone poisoning outcomes in children≤ 5 years: utilizing machine learning and the National Poison Data System for improved clinical decision-making. *European Journal of Pediatrics*, 184(2), 186. https://doi.org/10.1007/s00431-024-05957-x

Peng, X., Shu, W., Pan, C., Ke, Z., Zhu, H., Zhou, X., & Song, W.W. (2022). DSCSSA: A classification framework for spatiotemporal features extraction of arrhythmia based on the Seq2Seq model with attention mechanism. *IEEE Transactions on Instrumentation and Measurement*, 71, 1-12. https://doi.org/10.1109/TIM.2022.3194906

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-assisted Intervention – MICCAI 2015*, 234-241. https://doi.org/10.1007/978-3-319-24574-4-28

Sakhavi, S., Guan, C., & Yan, S. (2018). Learning temporal information for brain-computer interface using convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11), 5619-5629. https://doi.org/10.1109/TNNLS.2018.2789927

Shashikumar, S.P., Shah, A.J., Clifford, G.D., & Nemati, S. (2018). Detection of paroxysmal atrial fibrillation using attention-based bidirectional recurrent neural networks. Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, https://doi.org/10.48550/arXiv.1805.09133

Wang, C., Ma, J., Wei, G., & Sun, X. (2025). Analysis of cardiac arrhythmias based on ResNet-ICBAM-2DCNN dual-channel feature fusion. *Sensors*, 25(3), 661. https://doi.org/10.3390/s25030661

Wang, J., Li, J., Wang, R., & Zhou, X. (2024). VAE-driven multimodal fusion for early cardiac disease detection. *IEEE Access*, 12, 90535-90551. https://doi.org/10.1109/ACCESS.2024.3420444

Wang, Z., Stavrakis, S., & Yao, B. (2023). Hierarchical deep learning with generative adversarial network for automatic cardiac diagnosis from ECG signals. *Computers in Biology and Medicine*, 155, 106641. https://doi.org/10.1016/j.compbiomed.2023.106641

Wieneke, H., & Voigt, I. (2024). Principles of artificial intelligence and its application in cardiovascular medicine. *Clinical Cardiology*, 47(1), e24148. https://doi.org/10.1002/clc.24148

Wu, H., Qu, G., Xiao, Z., & Chunyu, F. (2024). Enhancing left ventricular segmentation in echocardiography with a modified mixed attention mechanism in SegFormer architecture. *Heliyon*, 10(15), e34845. https://doi.org/10.1016/j.heliyon.2024.e34845

Xu, Y., Kong, Z., Wu, Y., Wang, Y., Liu, X., Lu, Y., Wu, J., & Xu, H. (2026). MambaCapsule: Toward Transparent cardiac disease diagnosis with electrocardiography using mamba capsule network. *IEEE Transactions on Computational Social Systems*, 14(8), 1-15. https://doi.org/10.48550/arXiv.2407.20893

Zhang, C., Kim, Y.K., & Eskandarian, A. (2021). EEG-inception: an accurate and robust end-to-end neural network for EEG-based motor imagery classification. *Journal of Neural Engineering*, 18(4), 046014. https://doi.org/10.1088/1741-2552/abed81

Zhang, J., Liu, A., Gao, M., Chen, X., Zhang, X., & Chen, X. (2020). ECG-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network. *Artificial Intelligence in Medicine*, 106, 101856. https://doi.org/10.1016/j.artmed.2020.101856

Zhang, Z., Govindarajan, V., Ouyang, X., Yang, J., Alabdullah, B., Alammar, S., Por, L. Y., Guo, Q., & Li, M. (2026). CardioGPT-ViT: Explainable deep learning models with multimodal data for coronary artery disease and heart failure on edge consumer devices. *IEEE Transactions on Consumer Electronics*, 72(3), 6951-6963. https://doi.org/10.1109/TCE.2026.3667507

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support,* 3-11. https://doi.org/10.1007/978-3-030-00889-5-1

Zhu, J., Feng, Y., Liu, Q., Xu, H., Miao, Y., Lin, Z., Li, J., Liu, H., Xu, Y., & Li, F. (2024). An improved convnext with multimodal transformer for physiological signal classification. *IEEE Access*, 12, 11217-11229. https://doi.org/10.1109/ACCESS.2024.3355273