Benchmarking Deep Learning Architectures for ECG-Based Multi-label Heart Disease Prediction using MIMIC-IV Database

Eyiara Oladipo
Engineering and Computer Science
University of Detroit Mercy
Detroit MI, USA
oladipea@udmercy.edu

Sarwar Nazrul
Engineering and Computer Science
University of Detroit Mercy
Detroit MI, USA
nazrulsa@udmercy.edu

Mohamed Nafea
Electrical and Computer Engineering
Missouri University of Science and Technology
Rolla MO, USA
mnafea@mst.edu

2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS) | 979-8-3315-2610-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/CBMS65348.2025.00163

***Abstract*—Cardiovascular disease (CVD) is a leading cause of global mortality, accounting for an estimated 17.9 million deaths annually. CVD is broadly defined as a group of medical conditions influenced by modifiable or non-modifiable risk factors that affect the heart's ability to function properly. Machine learning (ML) has emerged as a powerful tool for analyzing complex medical data, aiding in early detection and accurate diagnosis of CVD and improving patient outcomes. Recent studies proposed various deep learning (DL) architectures for detecting CVD, yet there is a lack of robust benchmarks for comparing their performance on large-scale databases. In this work, we benchmark six state-of-the-art DL architectures for multi-label heart disease classification using 12-lead electrocardiogram (ECG) data from the large-scale publicly available Medical Information Mart for Intensive Care (MIMIC) database. Specifically, we evaluate a 1-dimensional convolutional neural network (CNN) with residual blocks (1D-CNN-ResNet); bidirectional long-short-term-memory neural network with convolutional layers (CNN-Bi-LSTM); spectrogram-based CNN (SpG-CNN); convolution-attention-transformer network (CAT-Net); hierarchical attention network (HAN), and structured state space sequence (S4) model; on a multi-label heart disease classification task with seven diagnostic targets. Model accuracy is assessed using the Hamming distance and its complexity is measured by number of model parameters. By contrasting models' accuracies versus their complexity, we establish a reliable benchmark providing constructive insights for advancing automated cardiovascular diagnostics.**

***Index Terms*—Cardiovascular disease prediction, ECG-based multi-label classification, MIMIC-IV database, deep learning architectures for healthcare diagnostics.**

I. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, highlighting the need for accurate and efficient diagnostic tools. Among the various clinical instruments available, the 12-lead ECG provides a comprehensive view of the heart's electrical activity by capturing signals from multiple anatomical perspectives. Electrodes placed on the chest, arms, and legs enhance the ability to detect subtle or localized abnormalities that might be missed with fewer leads [1], improving the diagnostic accuracy of arrhythmias, ischemic events, and other cardiac conditions. Despite its diagnostic value, 12-lead ECG interpretation remains a complex challenge, often requiring expert cardiologists to distinguish subtle patterns indicative of disease. The difficulty in interpreting overlapping ECG signals [2] has spurred interest in developing ML and DL models for automated ECG analysis.

In recent years, ML and DL techniques have demonstrated significant potential in ECG-based heart disease classification. However, several critical limitations persist. For instance, a major drawback of existing ECG-based ML classification studies is their reliance on small, private, or proprietary datasets, hindering their reproducibility as well as fair comparison of distinct model architectures [3]–[5]. Moreover, many studies focus on single-lead ECG data, neglecting the richer diagnostic information available in 12-lead ECGs [6]. Absence of standardized benchmarks for model evaluation has led to inconsistent performance reporting, making it difficult to fairly compare distinct architectures and assess/validate their clinical applicability. These challenges highlight the critical need for a unified, publicly available, large-scale evaluation benchmark, yielding fair and reproducible comparisons across DL models.

To address these gaps, this work leverages the MIMIC-IV database, one of the largest publicly available repositories of 12-lead ECG recordings [7]. The MIMIC-IV database provides a publicly available large-scale resource for benchmarking DL models under standardized and clinically relevant conditions, overcoming issues related to dataset variability and evaluation inconsistencies. By *proposing a consistent preprocessing pipeline and standard evaluation metrics for various model architectures, our work aims to establish a comprehensive benchmark for 12-lead ECG-based multi-label heart disease classification.* Specifically, we evaluate a diverse range of state-of-the-art DL architectures, including a 1D-CNN-ResNet [8], CNN-Bi-LSTM [9], SpG-CNN [10], CAT-Net [11], HAN [12], and S4 [13]. *These models are chosen to capture both short-term local dependencies, using convolutional approaches, and long-range temporal patterns, using recurrent, attention-based, and/or state space architectures.*

Overall, this study aims to furnish a ground for similar large-scale benchmarks, improving the reliability and trust of automated ECG-based diagnostic tools and advancing data-driven cardiovascular research. The remainder of the paper

2372-9198/25/$31.00 ©2025 IEEE
DOI 10.1109/CBMS65348.2025.00163

is organized as follows. Section II discusses related work. Section III presents the MIMIC-IV dataset. Section IV outlines the data preprocessing steps. Section V introduces the DL architectures of interest. Section VI describes our experimental methodology, including hyperparameter tuning and evaluation metrics. Section VII presents and discusses the results.

II. Related Work

Recent studies explored using ECG data for DL models in cardiovascular as well as general medical diagnostics. For instance, [13] trained a DL model on MIMIC-IV-ECG data linked to hospital discharge diagnoses, predicting a wide range of cardiac and non-cardiac conditions and demonstrating the potential of AI-enhanced ECG analysis as a unified screening tool. This model achieved high predictive performance across 253 International Classification Diseases (ICD) codes covering 81 cardiac and 172 non-cardiac conditions. [14] investigated various DL architectures including CNNs, LSTMs, and self-supervised learning (SSL) models with auto-encoders, for cardiac arrhythmia classification using digitized ECG datasets. CNN-based models achieved the highest accuracy ($\sim$92%), yet, [14] focused primarily on Lead II heartbeats, limiting its ability to capture spatial dependencies across multiple leads. *In contrast, our work evaluates models trained on full 12-lead ECGs to fully leverage the rich diagnostic information available in multi-lead ECG data.*

Another study [15] introduced a constrained transformer network for ECG signal processing and arrhythmia classification, proposing an end-to-end DL framework that integrates CNNs with the transformer network to enhance spatial and temporal feature extraction. The study incorporates a link constraint in the loss function to improve classification accuracy by making the embedding vectors of similar ECG classes more alike. Extensive experiments on real-world ECG datasets show that this model outperforms traditional architectures like CNNs and recurrent neural networks (RNNs) by capturing long-range dependencies. However, it lacks benchmarking against a standardized dataset like MIMIC-IV-ECG, limiting its generalizability assessment. *In contrast, our work ensures standardized evaluation by testing all models under the same conditions, facilitating direct performance comparison*.

Other studies have revisited various DL architectures for ECG classification. For instance, [8] developed a deep neural network (DNN) for multi-label classification of cardiac arrhythmias using 12-lead ECG recordings, achieving an average AUC exceeding 0.95 and an average F1-score of 0.813, outperforming traditional ML methods. [8] also found that using all 12 leads simultaneously outperforms single-lead models. However, their study lacked a comparison of multiple DL models on the same benchmark dataset, limiting the assessment of which architectures generalize best. *In contrast, our work systematically evaluates various DL model architectures under uniform conditions and data pre-processing steps to establish a fair and reliable benchmark for ECG classification.*

Another study [16] developed a hybrid DL model combining AlexNet and a custom CNN for ECG classification, achieving an average accuracy of 99.58% across three heart diseases and outperforming traditional methods. This demonstrates the effectiveness of hybrid architectures, yet it is limited in scope, focusing only on a small subset of cardiac diseases rather than a broader range of ECG abnormalities. *In contrast, our work provides a more extensive evaluation by benchmarking models across seven diagnostic categories, ensuring a more comprehensive understanding of their strengths and limitations.*

III. Database

The MIMIC-IV database version-3.1 (latest release) offers a large-scale publicly available resource providing de-identified health-related data from over 65,000 ICU and 200,000 emergency department (ED) admissions at Beth Israel Deaconess Medical Center. The MIMIC-IV-ECG module, a key component of the database, includes approximately 800,000 12-lead diagnostic ECGs from nearly 160,000 distinct patients. These ECG readings are essential for analyzing heart function and *used in this work to benchmark various DL architectures in assessing a wide range of cardiovascular conditions [7].*

A. ECG Data and Format

The MIMIC-IV's ECG data is obtained using 12 standard leads, namely, leads I, II, III, aVF, aVR, aVL, and V1-V6. Each ECG recording is 10 seconds long and sampled at a frequency of 500 Hz, ensuring high-resolution data by capturing rapid heart rate fluctuations and fine electrical patterns.

B. Linking ECG Data to Clinical Diagnoses

The MIMIC-IV database assigns each patient a unique subject ID, allowing data and diagnoses to be linked across modules, but does not directly associate ECG recordings with specific clinical diagnoses. *In this work, we utilize the MIMIC-IV-ECG-Ext-ICD extension dataset [13]*, which links ECG recordings to clinically validated diagnoses from ED and hospital discharge records, and enhances ECG data by aligning recording times with patient admission and discharge events. This facilitates the development of accurate ML models for ECG-based CVD classification. *We leverage this alignment to retrieve ICD-10 codes for each patient's discharge diagnoses, linking ECGs recorded in the ED or hospital to relevant diagnoses based on the timing of ECG recording and patient's admission/discharge.* Each patient can have up to nine ICD-10 codes for ED visits and 39 codes for hospital admissions, providing a comprehensive representation of their clinical diagnoses. These ICD-10 codes serve as critical labels, offering essential ground truth for evaluating and benchmarking ML models. *By incorporating these structured diagnostic labels, we enhance the ability of DL models to predict CVD with greater reliability and clinical relevance.*

IV. Data Preprocessing

Data preprocessing ensures the quality and consistency of input data in ML models. We employ a standard preprocessing pipeline for the 12-lead MIMIC-IV-ECG-Ext-ICD dataset to address common ECG signal processing challenges, including noise, missing data, and complex diagnostic labels.

A. Data Cleaning and Formatting

The MIMIC-IV-ECG-Ext-ICD dataset includes a wide range of medical conditions, many unrelated to circulatory system diseases [13]. To focus on heart diseases, we conduct extensive data cleaning and filtering. We filter the dataset to include only patients with hospital discharge or ED diagnoses within ICD-10 Chapter IX (codes I00-I99) [17], covering a broad spectrum of cardiovascular conditions such as coronary artery disease, arrhythmias, heart failure, and other cardiac and vascular disorders. This process excludes ECG recordings linked to non-circulatory system diagnoses, resulting in a refined dataset of approximately $\sim 361,000$ ECG recordings. We also apply "complete-case analysis" by excluding rows with missing ECG values, which are minimal and unlikely to impact the results significantly. Removing incomplete records ensures that models are trained on high-quality data, minimizing the risk of introducing biases or errors during training.

In this work, we address a multi-label classification problem, where each patient can have one or more cardiovascular conditions. To represent the seven CVD categories of interest (chronic ischemic, atrial fibrillation, heart failure, hypertensive, acute myocardial infarction, valve disorders, and others), we use a 7×1 binary vector, where '1' indicates the presence and '0' indicates the absence of a specific disease category. By employing multiple classification heads in all DL models in our benchmark, we enable the simultaneous prediction of multiple cardiovascular conditions, reflecting the realistic scenario where a patient may have more than one condition.

B. Denoising ECG Signals

We apply a denoising procedure to reduce noise and artifacts from raw ECG recordings, arising from sources like muscle activity, electrode movement, or electrical interference [18]. This noise can obscure critical ECG features such as P-waves, QRS complexes, and T-waves, essential for accurate heart rhythm detection. Denoising improves accuracy of downstream ML models by enhancing the signal-to-noise ratio, allowing them to focus on clinically relevant patterns in the ECG data. This step is vital for ensuring that models can reliably distinguish between different cardiac conditions [19].

As standard, we denoise the ECG signals using the *discrete wavelet transform (DWT)*, which decomposes the signal into multiple frequency components across different scales. DWT effectively captures both time and frequency information, making it suitable for non-stationary signals like ECG. The signal is filtered through high-pass and low-pass filters, producing high-frequency (detailed) and low-frequency (approximation) components. This process is iterated, further decomposing the approximation component into multiple levels. High-frequency components, typically containing noise, are thresholded, with coefficients below the threshold set to zero. This removes noise while preserving key signal characteristics. Finally, the denoised signal is reconstructed by combining the processed high-frequency components with the approximation component, reversing the decomposition and yielding a cleaner version of the original signal, as shown in Figures 1 and 2.

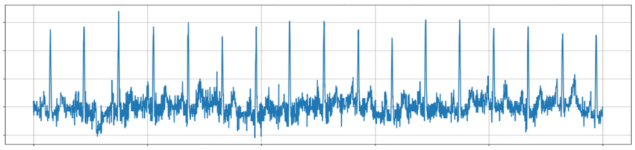

Fig. 1: Raw ECG Signal.

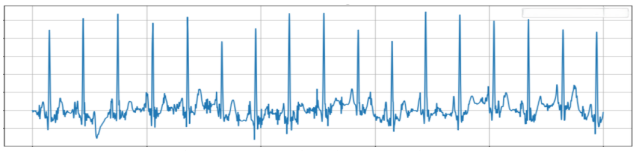

Fig. 2: Denoised ECG signal

C. Diagnostic Labeling and ICD-10 Codes

To link ECG recordings to clinically validated diagnoses, we focus on a subset of ICD-10 Chapter IX codes relevant to CVDs, building on previous research that demonstrates the feasibility of predicting these conditions using deep DL techniques [20]. DL models have been shown to effectively learn patterns from ECG signals associated with diseases like ischemic heart disease, heart failure, and arrhythmias, including atrial fibrillation and myocardial infarction [21]. By focusing on these established categories, we align our study with current advancements in ML for CVD diagnosis, leveraging existing research to apply DL models to a broad range of clinically significant heart diseases. As in [13], to manage the hierarchical complexity of ICD-10 codes, we group diagnoses by the first two digits, ensuring each category represents a distinct class of CVDs. This approach simplifies the classification problem while retaining specificity for accurate disease prediction. The diagnostic labels of interest are:

- *Chronic Ischemic Heart Disease (I25)*: Characterized by reduced blood flow to the heart due to narrowed arteries, this condition is a leading cause of morbidity and mortality worldwide.
- *Atrial Fibrillation (I48)*: A common arrhythmia associated with an irregular and often rapid heart rate, increasing the risk of stroke and heart failure.
- *Heart Failure (I50)*: A condition where the heart is unable to pump blood effectively, leading to symptoms such as shortness of breath and fatigue.
- *Hypertensive Heart Diseases (I11, I12, I13, I15, I16)*: Conditions resulting from high blood pressure, including hypertensive heart failure and hypertrophy.
- *Acute Myocardial Infarction (I21)*: Commonly known as a heart attack, this condition occurs when blood flow to a part of the heart is blocked, causing tissue damage.
- *Valve Disorders (I07, I08, I34, I35)*: Conditions affecting the heart valves, such as stenosis or regurgitation, which can impair blood flow within the heart.
- *Other*: A category for other cardiovascular conditions not covered by the above categories.

While many previous studies have used resampling techniques to address class imbalances in ECG datasets, the MIMIC-IV-ECG data is relatively well-balanced (see Table I) and reflects the typical distribution of CVDs in clinical settings. Therefore, we did not apply any resampling methods.

TABLE I: Distribution of diagnostic labels

CIHD	AF	HF	HHD	VD	AMI	Other
21.1%	18.1%	17.9%	15.3%	7.4%	6.6%	13.6%

V. Preliminaries

The DL architectures in our benchmark can be categorized into four broader classes: (1) CNN-based, (2) Hybrid CNN-RNN, (3) Attention-based and (4) State space architectures, representing distinct methodological approaches to processing ECG signals and extracting relevant features for classification.

A. CNN-based Architectures

CNNs, or pure-convolutional feed-forward models, utilize convolutional layers to extract spatial and short-term temporal features from ECG signals. These models are effective in capturing local patterns and invariant features in physiological signals [22]. Our benchmark includes two CNN-based architectures, adapted from notable recent research:

1. One-Dimensional Convolutional Neural Network with Residual Blocks (1D-CNN-ResNet): We utilize this model, adapted from [8], which employs a deep ResNet CNN architecture for classifying 12-lead ECG data across 9 diagnostic labels. The residual connections help mitigate the vanishing gradient problem, allowing for deeper architectures that can learn more complex feature maps. This model achieves an average F1-score of 0.813 on the China Physiological Signal Challenge 2018 (CPSC2018) dataset [23]. See Table II below for a description of the model's layers.

TABLE II: 1D-CNN-ResNet architecture

Layer	Description
Input	ECG signals
Conv1D	1D Convolution layer
BN1D	Batch normalization
ReLU	Activation function
Pooling	Downsampling layer
Residual Block ×4	4 stacked residual blocks
Pooling	Downsampling layer
Dense	Fully connected layer
Sigmoid (9)	Sigmoid activation with 9 output classes
Output	Final output layer

2. Spectrogram-based CNN (SpG-CNN): This model is adapted from the GitHub repository[1] guided by [10]. It transforms raw ECG signals into logarithmic spectrograms, representing time-frequency distributions, and uses a stack of convolutional layers for feature extraction. Due to the variable-length input data, temporal aggregation is performed by averaging feature maps across time. By leveraging the spectral representation of ECG waveforms, the model captures frequency-domain features useful for distinguishing cardiac conditions. *We adapt this model to accept 12-lead data sampled at 500 Hz, and modify the output layer and loss function to accommodate the multi-label classification problem.*

B. Hybrid CNN-RNN Architecture

Hybrid deep models combine convolutional layers with recurrent architectures, such as bidirectional long short-term memory (Bi-LSTM) networks, to capture both spatial and sequential dependencies in ECG signals. We benchmark the following hybrid CNN-RNN architecture:

3. Bi-LSTM with Convolutional Layers (CNN-Bi-LSTM): This model, introduced by [9], uses convolutional layers to extract local spatial features from ECG waveforms, followed by BiLSTM layers that model temporal dependencies. BiLSTMs capture both forward and backward dependencies, which is critical for identifying patterns across time. The model achieves a weighted F1-score of 0.82 on a multi-class problem with three classes: Atrial Fibrillation, Normal, and Other. *We adapted this model to accept 12 leads and updated its final layer to match our multi-label classification problem.*

Fig. 3: CNN Bi-LSTM architecture

C. Attention-based Architectures.

Attention-based models, including the transformer [24], are designed to handle long-range dependencies and efficiently process long ECG sequences, capturing complex temporal relationships. For our benchmark, we consider the following two state-of-the-art attention-based architecture:

4. Hierarchical Attention Network (HAN): The HAN, adapted from [25], applies hierarchical attention mechanisms to focus on different levels of ECG features, enabling multi-scale feature aggregation. It captures both short and long-term dependencies through its hierarchical structure. Designed for single-lead ECG analysis, the HAN in [12], [26] uses R-peak fusion to break down ECG signals into a hierarchical structure, capturing local and global dependencies across multiple scales. *In this work, we extend the model to handle multi-lead ECG signals using multi-R-peak fusion across all 12 leads, enhancing its ability to capture spatial relationships between leads, and modify the final layer for multi-label classification.*

5. Convolution, Attention, and Transformer Network (CAT-Net): CAT-Net, introduced by [11], combines convolutional feature extraction with attention mechanisms and transformer-based processing to enhance detecting long-range dependencies in ECG signals; see Table III. CNN layers extract local features, while attention modules focus the model on the most relevant signal regions. *We adapt this model to accept 12-lead data and the output layer for multi-label classification*. Unlike the MIT-BIH dataset used in [11] which labels individual heartbeats (doesn't require R-peak detection), the MIMIC-IV-ECG readings do not contain R-peak labels.

TABLE III: CAT-Net architecture: Convolution, attention, and transformer based network

Layer	Description
Input	ECG signal (500 × 1)
1st Conv Block	Conv, BN, Attn, Pool (500 × 16) → (150 × 16)
2nd Conv Block	Conv, BN, Attn, Pool (150 × 16) → (75 × 32)
3rd Conv Block	Conv, BN, Attn, Pool (75 × 32) → (38 × 64)
4th Conv Block	Conv, BN, Attn (38 × 64) → (38 × 128)
Positional Encoding	Applied to features
Transformer Encoder	MHA, Add & Norm, FFN, Add & Norm
Flatten Layer	Converts 2D to 1D (38 × 128) → 4864
Dense Layer 1	Fully connected (128 neurons)
Dense Layer 2	Fully connected (5 output classes)
Output	Predicted classes

[1]https://github.com/awerdich/physionet.

D. State Space Architectures.

Much like attention-based models, state space architectures are intended to handle very long-range dependencies and efficiently process long ECG sequences. For our benchmark, we consider the following state-of-the-art state space architecture:

6. Structured State Space Sequence (S4) Model: S4, introduced in [27], employs a state space representation to efficiently model long-range dependencies in ECG signals. Unlike recurrent models, S4 avoids vanishing gradients by using structured state space layers, making it highly effective for capturing long-term relationships in sequential data. The S4 architecture begins with a Conv1D layer, followed by S4 blocks with LayerNorm and skip connections, and concludes with an average pooling layer and a linear classification head. The code was sourced from the implementation in [13].

VI. Methodology and Experiments

A. Experimental Setup and Hyperparameter Tuning

We partition the MIMIC-IV-ECG-Ext-ICD dataset into training, validation, and test sets using an 80%–10%–10% split, respectively. To ensure robust model performance, mitigate overfitting, and enhance generalizability, each DL architecture in our benchmark underwent extensive systematic hyperparameter tuning for ECG classification. The tuning process involves optimizing the following hyperparameters:

- Learning Rates ranging from 0.0001 to 0.1 were tested, with StepLR decreasing the rate by a factor γ (ranging from 0.1 to 0.3) and at fixed intervals (every 10 to 20 epochs).

- Batch Size, evaluated within a range of 16–512 to optimize computational efficiency and generalization.

- Network-depth and hidden units. LSTM-based models are examined with 1–3 layers, while CNN layers are examined with varying number of hidden units (ranging from 64 – 1024) to assess their feature extraction capacity.

- Dropout-rate, configured between 0.1–0.5 to minimize overfitting while maintaining effective learning.

- Regularization weight. L1 & L2 regularization are applied to reduce model complexity and improve generalization (tested values ranged from 0.01 – 0.0001).

B. Optimal hyperparameter configurations

Table IV summarizes the best-performing hyperparameters for each model, based on validation error performance. Additional model-specific configurations include: *1D-CNN-ResNet* utilizes binary cross-entropy (BCE) loss, Adam optimizer and a dropout-rate of 0.2 in residual blocks. *CNN-BiLSTM* employs step-LR scheduler (step size 10, gamma 0.1), 256 LSTM hidden units, and dropout rates of 0.1 for CNN layers and 0.2 for LSTM layers. *HAN* uses channel attention with 32 filters and segment attention with 64 units. *CAT-Net* features 4 attention heads, a model dimension of 64 and a feed-forward dimension of 64.[2] *S4:* uses SGD optimizer with weight decay of 1×10^{-4} and a dropout rate of 0.1.

[2]Model dimension is defined as the size of feature vectors for each ECG time step after convolution and attention. The feedforward dimension is the size of the hidden layer in the transformer's feed-forward network

TABLE IV: Optimal hyperparameter configuration for benchmark models.

Model	Learning Rate	Batch Size	Epochs
1D-CNN-ResNet	0.001	32	50
SpG-CNN	0.001	32	40
CNN-BiLSTM	0.01	32	50
HAN	0.01	512	20
CAT-NET	0.001	64	50
S4	0.01	32	50

C. Evaluation metrics

For our multi-label benchmark, we use the Hamming loss metric to quantify classification accuracy over 5 seeds of the 7 diagnostic labels (Table I). Table V summarizes the results.

D. Reproducibility

To ensure transparency and reproducibility, we have made all code, data preprocessing scripts, model architectures, training configurations, and evaluation metrics publicly available. The full implementation can be accessed at: **https://github.com/MIMIC-Benchmarking**

VII. Discussion

Our benchmarking results show that DL models can effectively distinguish between different cardiac conditions using 12-lead ECG data. Evaluating multiple architectures provided insights into the trade-offs between model complexity, computational requirements, and classification performance. Table VI summarizes model complexity in terms of total, trainable, and non-trainable parameters.

The results indicate that larger models, such as 1D-CNN-ResNet and CNN-BiLSTM, require substantial computational resources due to their high number of trainable parameters. These models demonstrated strong classification performance by leveraging their large capacity to capture intricate ECG signal patterns. However, their increased complexity results in higher memory requirements and longer inference times, making them less suitable for real-time applications in resource-constrained environments. In contrast, smaller models like HAN and CAT-Net showed significantly lower parameter counts while maintaining competitive classification performance. HAN, with only 84K total parameters, offers an efficient alternative for real-time applications where computational efficiency is crucial [26]. Additionally, CAT-Net and SpG-CNN demonstrated the ability to balance accuracy and efficiency, making them viable candidates for deployment in mobile and edge computing scenarios.

TABLE VI: Model Complexity Analysis

Model	Total Param	Trainable Param
1D-CNN-ResNet	139.6M	139.6M
CNN-BiLSTM	38.07M	38.07M
S4	6.4M	6.4M
CAT-NET	2.66M	2.66M
SpG-CNN	3.56M	3.55M
HAN	84K	83.9K

Examining the trade-off between model complexity and predictive performance is a crucial aspect of this study. Models with higher parameter counts capture richer features but require more extensive computational resources for training and

TABLE V: Model Performance

ResNet	BiLSTM	SpG-CNN	HAN	CAT	S4
0.2 ± 0.002	0.22 ± 0.002	0.204 ± 0.002	0.2 ± 0.01	0.196 ± 0.001	0.20 ± 0.002

inference. In contrast, lightweight models are computationally efficient but may need architectural enhancements, such as attention mechanisms or feature selection techniques, to achieve comparable accuracy. These findings have important implications for ECG-based disease diagnosis. In clinical settings with access to high-performance computing, models like 1D-CNN-ResNet and CNN-BiLSTM can be deployed to maximize diagnostic accuracy. Conversely, in remote or resource-limited environments, more compact models such as HAN and CAT-Net offer feasible alternatives with reduced energy consumption and faster inference speeds. Future research should explore hybrid approaches that integrate model compression techniques and transfer learning to optimize both performance and computational efficiency.

Overall, our benchmarking framework provides valuable insights into the selection of DL models for ECG classification, facilitating informed decisions based on specific deployment constraints and performance objectives.

REFERENCES

[1] S. Meek and F. Morris, "Introduction. i—leads, rate, rhythm, and cardiac axis," *BMJ*, vol. 324, no. 7334, pp. 415–418, Feb 2002.

[2] N. Rafie and K. Krishnan, "ECG interpretation: Clinical relevance, challenges, and advances," *Hearts*, vol. 2, no. 4, pp. 505–513, 2021. [Online]. Available: https://doi.org/10.3390/hearts2040039

[3] R. Sabay and M. Harris, "Overcoming small data limitations in heart disease prediction by using surrogate data," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52950491

[4] H. El-Sofany, B. Bouallegue, and Y. M. Abd El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable ai method," *Scientific Reports*, vol. 14, no. 1, p. 23277, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-74656-2

[5] P. Iacobescu, V. Marina, C. Anghel, and A.-D. Anghele, "Evaluating binary classifiers for cardiovascular disease prediction: Enhancing early diagnostic capabilities," *J. Cardiovasc. Dev. Dis.*, vol. 11, no. 12, p. 396, 2024. [Online]. Available: https://doi.org/10.3390/jcdd11120396

[6] M. Ezz, "Deep learning-driven single-lead ecg classification: A rapid approach for comprehensive cardiac diagnostics," *Diagnostics*, vol. 15, no. 3, p. 384, 2025. [Online]. Available: https://doi.org/10.3390/diagnostics15030384

[7] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023. [Online]. Available: https://doi.org/10.1038/s41597-022-01899-x

[8] D. Zhang, S. Yang, X. Yuan, and P. Zhang, "Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram," *iScience*, vol. 24, no. 4, p. 102373, 2021. [Online]. Available: https://doi.org/10.1016/j.isci.2021.102373

[9] J. Wang and W. Li, "Atrial fibrillation detection and ecg classification based on cnn-bilstm," *arXiv*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2011.06187

[10] M. Zihlmann, D. Perekrestenko, and M. Tschannen, "Convolutional recurrent neural networks for electrocardiogram classification," *arXiv*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1710.06122

[11] M. R. Islam, M. Qaraqe, K. Qaraqe, and E. Serpedin, "Cat-net: Convolution, attention, and transformer based network for single-lead ecg arrhythmia classification," *Biomed. Signal Process. Control*, vol. 93, p. 106211, 2024. [Online]. Available: https://doi.org/10.1016/j.bspc.2024.106211

[12] S. Mousavi, F. Afghah, and U. R. Acharya, "Han-ecg: An interpretable atrial fibrillation detection model using hierarchical attention networks," *Comput. Biol. Med.*, vol. 127, p. 104057, 2020. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2020.104057

[13] N. Strodthoff, J. M. Lopez Alcaraz, and W. Haverkamp, "Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: an explorative study in emergency care," *Eur. Heart J. Digit. Health*, vol. 5, no. 4, pp. 454–460, 2024. [Online]. Available: https://doi.org/10.1093/ehjdh/ztae039

[14] S. Sattar, R. Mumtaz, M. Qadir, S. Mumtaz, M. A. Khan, T. De Waele, E. De Poorter, I. Moerman, and A. Shahid, "Cardiac arrhythmia classification using advanced deep learning techniques on digitized ecg datasets," *Sensors*, vol. 24, no. 8, p. 2484, 2024. [Online]. Available: https://doi.org/10.3390/s24082484

[15] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin, "Constrained transformer network for ecg signal processing and arrhythmia classification," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, p. 184, 2021. [Online]. Available: https://doi.org/10.1186/s12911-021-01546-2

[16] M. Kolhar and A. M. Al Rajeh, "Deep learning hybrid model ecg classification using alexnet and parallel dual branch fusion network model," *Scientific Reports*, vol. 14, no. 1, p. 26919, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-78028-8

[17] ICD-10 Data, "Diseases of the circulatory system I00-I99," https://www.icd10data.com/ICD10CM/Codes/I00-I99, 2025.

[18] P. Kumar and V. Sharma, "Detection and classification of ecg noises using decomposition on mixed codebook for quality analysis," *Healthc. Technol. Lett.*, vol. 7, no. 1, pp. 18–24, 2020. [Online]. Available: https://doi.org/10.1049/htl.2019.0096

[19] S. Murray, N. McCord, J. Diven, M. P. Fitzpatrick, H. Easlea, A. Gibbs, and A. R. J. Mitchell, "Evaluating the impacts of digital ecg denoising on the interpretive capabilities of healthcare professionals," *Eur. Heart J. Digit. Health*, vol. 5, no. 5, pp. 601–610, 2024. [Online]. Available: https://doi.org/10.1093/ehjdh/ztae063

[20] S. V. Kalmady, A. Salimi, W. Sun, N. Sepehrvand, Y. Nademi, K. Bainey, J. Ezekowitz, A. Hindle, F. McAlister, R. Greiner, R. Sandhu, and P. Kaul, "Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level," *npj Digital Medicine*, vol. 7, no. 1, pp. 1–11, 2024. [Online]. Available: https://doi.org/10.1038/s41746-024-01130-8

[21] C.-H. Lin, Z.-Y. Liu, P.-H. Chu, J.-S. Chen, H.-H. Wu, M.-S. Wen, C.-F. Kuo, and T.-Y. Chang, "A multitask deep learning model utilizing electrocardiograms for major cardiovascular adverse events prediction," *npj Digital Medicine*, vol. 8, no. 1, pp. 1–11, 2025. [Online]. Available: https://www.nature.com/articles/s41746-024-01410-3

[22] F. Khan, X. Yu, Z. Yuan, and A. u. Rehman, "Ecg classification using 1-d convolutional deep residual neural network," *PLoS ONE*, vol. 18, no. 4, p. e0284791, 2023. [Online]. Available: https://doi.org/10.1371/journal.pone.0284791

[23] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, J. Li, and E. Y. K. Ng, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *J. Med. Imaging Health Informatics*, vol. 8, no. 7, pp. 1368–1373, 2018. [Online]. Available: https://doi.org/10.1166/jmihi.2018.2442

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS 2017)*, 2017, pp. 6000–6010. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295349

[25] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. NAACL-HLT*, 2016, pp. 1480–1489. [Online]. Available: https://aclanthology.org/N16-1174/

[26] M. P. Rodriguez and M. Nafea, "Hierarchical attention network for interpretable ecg-based heart disease classification," *ArXiv preprint*, 2025.

[27] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2022. [Online]. Available: https://arxiv.org/abs/2111.00396