Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 03 March-2026, Page No.-9305-9317

DOI: 10.47191/etj/v11i03.25, I.F. – 8.595

Comparative Analysis of ECG Image Classification for Cardiac Disease Detection Using Deep Learning Approaches

Bareq Mardan[1], Diyar Qader Zeebaree [2]

[1]ITM Department, College of Technical, Duhok Polytechnic University, Kurdistan Region, Iraq

[2]Technical Engineering Department, College of Computer and AI, Northern Technical University (NTU)

ABSTRACT

Background: Electrocardiogram (ECG) interpretation is essential for cardiac disease diagnosis. While signal-based methods have been dominant recent advances in deep learning introduced image-based classification offering new directions for automated detection.

Objectives: This study investigates the effectiveness of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) and a hybrid CNN-LSTM model for ECG image classification. The focus is on whether combining spatial and temporal feature learning improves accuracy compared to individual models.

Methods: ECG images were collected from public repositories. CNN and LSTM models were first applied directly to raw data but faced limitations due to noise and class imbalance. To address these issues preprocessing was introduced: Pix2Pix GAN for denoising and augmentation, and resampling with class weighting to balance classes. The models were retrained on the processed data, followed by the development of a hybrid CNN-LSTM framework combining convolutional feature extraction with sequential learning.

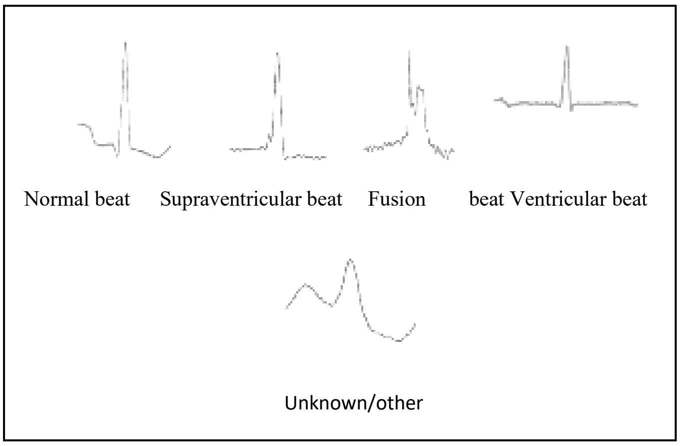
Results: The hybrid model achieved the best performance, with 97.89% accuracy and a macro F1-score of 0.89. Preprocessing significantly improved recognition of minority classes, highlighting the role of GAN-based denoising and balanced training.

Conclusion: A hybrid CNN-LSTM with GAN preprocessing provides a reliable and robust framework for ECG image classification. This approach offers improved generalization and could complement or enhance traditional signal-based ECG analysis in clinical practice.

KEYWORDS: Electrocardiogram, Convolutional Neural Network, Deep Learning, Long Short-Term Memory

I. INTRODUCTION

Electrocardiography (ECG) is a critical diagnostic tool for identifying and monitoring cardiovascular diseases, which remain one of the leading causes of mortality worldwide [1]. The accurate interpretation of ECG signals plays a vital role in early detection and timely intervention, potentially reducing the risk of severe health outcomes [2]. However, manual analysis by clinicians is often time-consuming and prone to human error, particularly in large-scale screening programs. As a result automated classification systems based on machine learning and deep learning techniques have gained significant attention in recent years [3]. Deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks have been widely utilized for ECG classification due to their ability to learn complex patterns in spatial and temporal domains, respectively [4], [5]. CNNs excel in feature extraction from ECG images, while LSTMs are effective for modeling sequential dependencies in ECG waveforms [6]. Despite their success these models often face limitations when trained directly on raw datasets, including class imbalance, noise interference, and overfitting, which can degrade classification performance [7]. To address these challenges, advanced preprocessing methods such as denoising with Generative Adversarial Networks (GANs), resampling techniques, class weighting, Gaussian blur filtering and extensive data augmentation have been introduced [8], [9]. These techniques not only enhance data quality but also increase dataset diversity thereby improving the generalization capability of models. Moreover, hybrid approaches combining CNN and LSTM have shown promise in capturing both morphological and temporal features of ECG signals, offering improved classification accuracy compared to single-model architectures [10]. Recent developments such as GoogleNet and other transformer-based architectures further highlight the potential for multi-scale and attention-based feature learning, paving the way for more robust ECG classification frameworks [11]. In this study, we present a comprehensive comparative analysis of ECG image classification using CNN, LSTM, and a hybrid CNN-LSTM model before and after extensive preprocessing. The proposed framework incorporates denoising with Pix2Pix GAN, class balancing, and multi-strategy augmentation, resulting in a dataset expansion from 54,613 images to over 247,000 images.



Experimental results demonstrate a substantial improvement in classification performance after preprocessing, with the hybrid model achieving the highest accuracy among the tested architectures. These findings suggest that the integration of preprocessing and hybrid modeling can significantly enhance automated ECG diagnosis and serve as a foundation for future work integrating advanced architectures like Google Net or vision transformers.

Figure 1: Examples of ECG signals after filtering, illustrating both normal and abnormal cases. The filtered signals demonstrate the effectiveness of preprocessing in enhancing waveform clarity and reducing noise for improved classification accuracy.

The main contributions of this work:

1. Application of Pix2Pix GAN for ECG image denoising, which improves the quality of the input data by effectively reducing noise while preserving the diagnostic characteristics of ECG waveforms.
2. Addressing class imbalance using resampling strategies and class weights, ensuring fair learning across majority and minority ECG classes.
3. Design and implementation of a Hybrid CNN-LSTM model, which combines spatial feature extraction through convolutional layers with temporal sequence modeling using LSTM units.
4. Achieving high classification performance, reaching an accuracy of 97.89% on the processed ECG dataset, outperforming several existing approaches in the literature.

The remainder of this paper is organized as follows: Section 1 (Introduction), Section 2 (presents the related works and literature review highlighting prior approaches to ECG classification and their limitations), Section 3 (details the proposed methodology), Section 4 (results and discussion), Section 5 (Conclusion).

II. RELATED WORK

Most research in recent years has used various deep learning models on ECG image data from different databases. In the study [12] demonstrated that combining neural networks (NNs) such as CNNs and RNNs within the Ensemble deep learning framework achieved over 92% accuracy on three datasets (Sunnybrook Cardiac Data, MIT-BIH, UK Biobank). Although the results are robust, relying on a single data type (either MRI or ECG) made the models vulnerable to image quality and technical glitches. Similarly, in the study [13] using Deep Neural Networks (DNNs) on PTB-XL and TIS data showed a classification accuracy of 91.7%. However, limitations include poor generalization when testing models on different datasets due to different ECG patterns, as well as the problem of labeling errors within large datasets. Several studies have turned to hybrid models to leverage the spatial capabilities of convolutional networks (CNNs) and the temporal capabilities of LSTM or Bi LSTM networks. [14] showed that a CNN-LSTM model on the MIT-BIH arrhythmia database achieved an accuracy more than of 95.42%, while a CNN-SVM model achieved an accuracy of 96.85%. Similarly, a CNN-Bi LSTM model in a study [15] achieved an accuracy of 89.73% on the PTB-XL database. However, these models have shown increased design complexity requiring significant computing resources, and difficulty in large-scale clinical validation.

Some research has focused on improving CNN models by incorporating modern techniques. in study [16] based on Res Nets with a bidirectional GRU achieved 96.5% training accuracy and 95.9% validation accuracy.[17] using LeNet and X Res Net models on PTB-XL datasets achieved an AUC of 0.9268, while Precision, Recall, and F1 metrics reached 0.61,0.70, and 0.64, respectively. in the study [18] However, the model was limited to a single channel (Lead-II), which reduced its generalizability. Despite promising results, most previous research has faced recurring challenges, such as imperfect class balance, with the normal class dominating most samples, the impact of noise from ECG recordings being imperfect, and poor generalization when testing models on data from different hospitals or settings. Furthermore, the reliance of most studies on image data or single-dimensional signals alone has made models less capable of handling the true diversity of patients in clinical settings. Several studies have explored hybrid models to leverage the spatial capabilities of CNNs and the temporal modeling of LSTM or Bi LSTM networks. For instance, [19] reported that a CNN-LSTM model on the MIT-BIH arrhythmia database achieved an accuracy above 95.42%, while a CNN-SVM model achieved 96.85%. Similarly, a CNN-Bi LSTM model in [20] achieved an accuracy of 89.73% on the PTB-XL database. However, these hybrid models often introduced higher design complexity, requiring significant computing resources and facing challenges in large-scale clinical validation.

Other researchers have focused on improving CNN models by integrating modern architectures. For example, [21] proposed a Res Net combined with bidirectional GRU, achieving 96.5% training accuracy and 95.9% validation accuracy. Meanwhile, [22] employed LeNet and X Res Net models on PTB-XL datasets, obtaining an AUC of 0.9268 with precision, recall, and F1 values of 0.61, 0.70, and 0.64, respectively. However, models such as the one in [23] were limited to single-channel ECG (Lead-II), reducing their generalizability.

Despite these promising results, most prior research faced recurring challenges, including class imbalance dominated by normal beats, the impact of noisy ECG recordings, and limited generalization when applied to data from different hospitals. Furthermore, the majority of studies relied only on ECG images or one-dimensional signals, which reduced their ability to capture the diversity of clinical populations. Finally, only a few works have explored the use of generative models, such as the Pix2Pix GAN, to improve image quality before training [24], [25], [26]. review of previous studies reveals that most models, whether CNN, LSTM, or hybrid, achieved high classification performance, but were limited by noise in ECG data, class imbalance, and poor generalization when transferring models to different environments. Furthermore, most research has not exploited the potential of deep generative algorithms such as the Pix2Pix GAN to address image quality before training.

II. Table.1. Summary of previous studies on deep learning models in data analysis

REF.	DATA SETS	CLASSIFIER	ACCURACY	DEMERITS
[27]	MIT-BIH arrhythmia database	CNN-SVM, CNN-LSTM	Accuracy: CNN-LSTM: 97.42%, CNN-SVM: 96.85%, Precision: CNN-LSTM: 97.44%, CNN-SVM: 96.87%	Single Dataset Limitation: This restricts generalizability; models must be validated on diverse datasets for robustness and real-world applicability.
[28]	3 (PTB-XL, CODE-15%, Chapman Arrhythmia)	Hybrid model combining CNN, BI LSTM	Accuracy: PTB-XL: 89.73%	Model complexity and need for further validation. Additionally, further validation across diverse populations is needed to enhance generalizability.
[29]	1 dataset (MIT-BIH Arrhythmia)	Ensemble of Deep Residual Networks (Res Nets) and Bidirectional Gated Recurrent Units	Training accuracy: 96.5%, Validation accuracy: 95.9%, Precision: 98.1%, Recall: 98%, F1 Score: 98%	Model Design Complexity: Complexity may increase computational demands and training times, hindering real-time clinical deployment.
[30]	PTB-XL dataset (ECG signals)	B I LSTM, Adamax	Accuracy: 91.24%, Precision: 90.73%, Recall: 97.82%, F1 Score: 94.13%	High Variability Different patients may exhibit different ECG patterns for the same condition, which can lead to misclassifications.
[31]	MITDB, VFDB (ECG signals)	CNN (with n-BSM)	Accuracy: 99.75%, F1 Score: 96.40%	The limitations regarding the need for a pre-trained

III. METHODOLOGY

research has highlighted the importance of preprocessing and model design in improving ECG classification performance.[32] demonstrated that incorporating time–frequency features into CNN architectures can significantly enhance arrhythmia classification accuracy particularly when addressing class imbalance. Similarly,[33] showed that reinforcement learning combined with data augmentation improves robustness against noisy and unbalanced datasets. Lightweight models such as ECG encode have further proven that compact feature encoders can achieve competitive results with lower computational costs [34][35]. The experimental workflow began with collecting a dataset of ECG images from publicly available sources. We chose two standard deep learning models, CNN and LSTM, as our baseline architectures. In the first stage, we directly trained these models on the raw dataset without any preprocessing or changes to their structure. This initial approach proved challenging both models were significantly affected by image noise and the uneven distribution of classes (class imbalance). Despite multiple attempts to adjust hyperparameters, their accuracy and reliability remained low. To overcome these issues, we implemented a series of data improvement steps. We used a Pix2Pix GAN to clean the images and generate additional realistic samples. To tackle the class imbalance, we applied resampling strategies and incorporated class weights into the loss function during training. After reprocessing the data, we retrained the same CNN and LSTM models. Finally, our study, designed a hybrid CNN-LSTM model that combines feature extraction capabilities of CNN with sequence learning ability of LSTM. This combined architecture showed a major boost in performance, significantly improving classification accuracy and model stability.

III.1. Direct Application of Algorithms

The initial approach was to directly train two baseline architectures: a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) network, covering a diverse set of cardiac conditions. with labels corresponding to normal and abnormal cardiac classes. The combined dataset provides a total of 54,613 ECG images was randomly divided into 70% training, 15% validation, and 15% testing sets. Stratified splitting was employed to maintain class balance in each subset, without applying any preprocessing or data

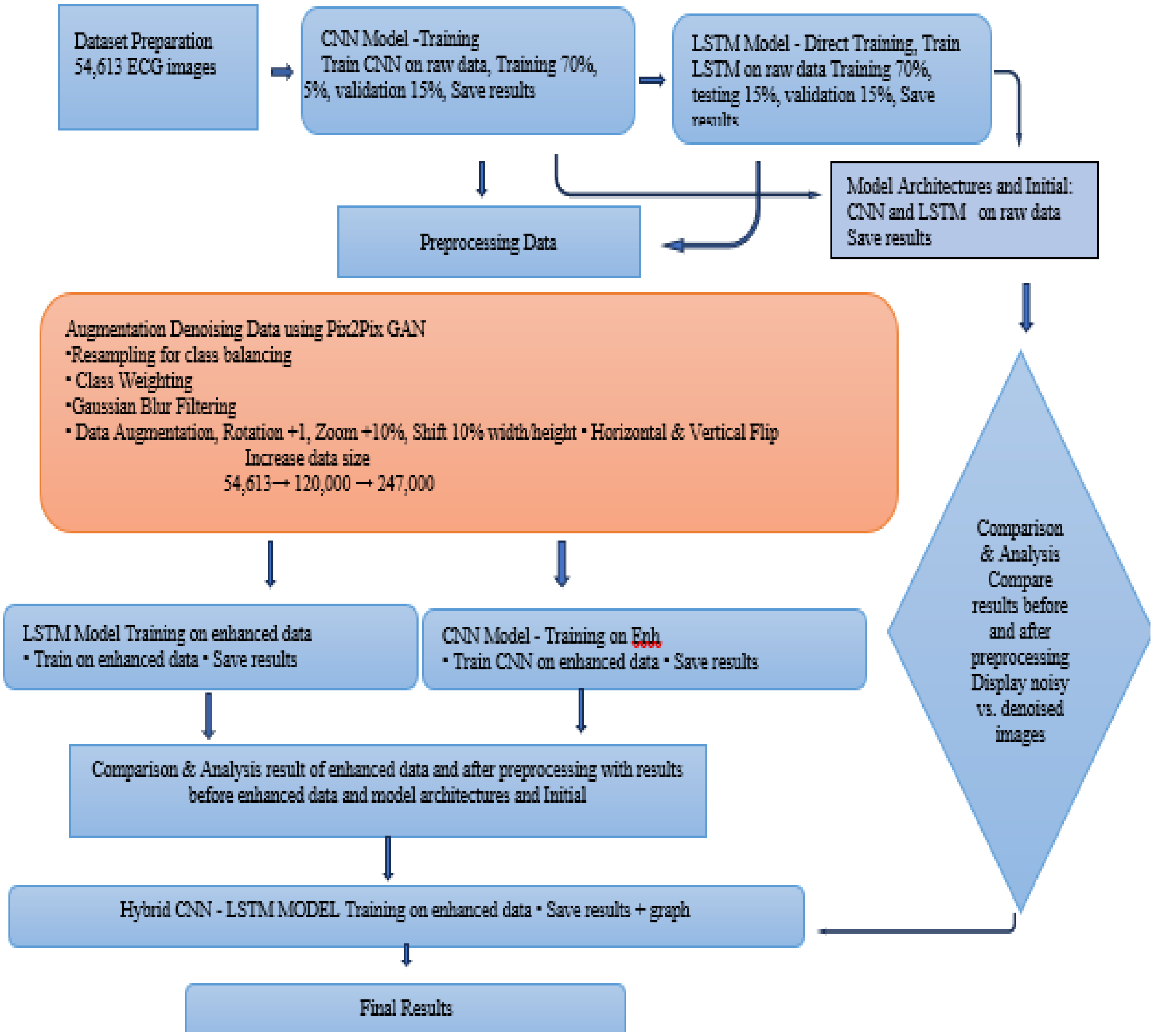


Figure 2. showing the sequential steps from data preparation to model evaluation

enhancement. In this stage, no modifications were made to the model architectures, and two baseline models with proven performance were executed directly on the imaging

The CNN achieved a classification accuracy of approximately 91%, which indicates its strong capability in learning spatial features of ECG waveforms. However, a deeper inspection of the classification report revealed several challenges and limitations:

Noise interference: ECG images often contained artifacts resulting from imperfect recording conditions, which degraded model generalization.

Class imbalance: The majority of the dataset belonged to the Normal (N) class, which biased the model toward majority predictions while underperforming on minority classes such as F, S, and V.

Limited diversity: The dataset lacked sufficient variability to represent real-world scenarios, leading to poor sensitivity for rare cardiac abnormalities.

The LSTM model, despite its strength in sequence modeling, struggled significantly when applied directly to image-transformed ECG signals. The overall accuracy dropped to 78%, and the model failed to capture minority class features, yielding near-zero recall for abnormal categories. This confirmed that relying solely on temporal sequence extraction without proper feature enrichment is inadequate for robust ECG classification.

Based on these findings, we concluded that direct application of algorithms is insufficient for building a reliable diagnostic model. Therefore, the next methodological stage focused on preprocessing strategies, including denoising, augmentation, and class rebalancing, to overcome these challenges and improve generalization.

CNN and LSTM algorithms were chosen in this study because they are among the most widely used and interactive algorithms in recent research on medical image classification, especially electrocardiogram (ECG) images. CNNs rely primarily on image processing and have proven highly effective in extracting patterns and spatial features from medical images. When converting ECG signals into images, CNNs can recognize fine morphological details such as P, QRS, and T waves, which are important elements for diagnosing heart disease. Therefore, they are the most popular and widely used choice in the scientific literature for medical image processing. CNNs are employed to extract spatial features from ECG images.

CNNs are employed to extract spatial features from ECG images.

The convolution operation for a single output channel is defined as:

$$y(i,j) = \sigma\left(\Sigma_{u,v} W(u,v) \cdot x(i+u, j+v) + b \right)$$

where x denotes the input image, W is the convolution kernel, b is the bias, and σ is the activation function (e.g., ReLU).

Pooling operations reduce the dimensionality of the feature maps, typically using max pooling:

$$y(i,j) = \max_{(u,v)\in\Omega} x(s\cdot i+u, s\cdot j+v)$$

where Ω represents the pooling window and s is the stride. For example, applying 2×2 max pooling on a 200×200 ECG image reduces it to 100×100.

LSTMs are specialized networks for handling sequential data and are widely used to model long-term temporal relationships. Since the ECG signal is essentially a temporal signal, LSTMs provide a suitable framework for understanding the correlation between waveforms, even when represented as images. This has also made them the focus of numerous studies investigating improving diagnostic accuracy.LSTMs are designed to capture temporal dependencies in sequential data. Given input x_t, previous hidden state h_{t-1}, and cell state C_{t-1}, the following equations are applied:

Forget gate: $f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f)$

Input gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$

Candidate memory: $\tilde{C}_t = \tanh (W_C \cdot [h_{t-1}, x_t] + b_C)$

Cell update: $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

Output gate: $o_t = \sigma (W_o \cdot [h_{t-1}, x_t] + b_o)$

Hidden state: $h_t = o_t * \tanh(C_t)$

For example, if CNN features are extracted as a sequence of length T=5, the LSTM processes each step iteratively, deciding which information to retain (f_t) and which to update (i_t).

Therefore, combining CNN and LSTM not only reflects the popular trends in recent research, but also enables a direct comparison between image feature extraction using CNN and time series analysis using LSTM.

In the hybrid model, CNNs act as feature extractors from the ECG images, producing a sequence of feature vectors. These feature vectors are then passed to the LSTM, which models their temporal relationships.

Let $F \in R^{T\times D}$ represent the sequence of D-dimensional feature vectors over T steps.

The LSTM processes F and produces hidden states $H \in R^{T\times H}$.

The final hidden state H_T is fed into a dense layer with a SoftMax activation function to produce class probabilities:

$$\hat{p}_c = \exp(z_c) / \Sigma_{k=1}^K \exp(z_k)$$

where z_c is the logit for class c, and K is the total number of classes (e.g., N, S, V, F, Q in ECG classification).

Implementation Environment:

All experiments were conducted on Google Collab, which provided a suitable cloud environment for model training. A Tesla T4 GPU with 16GB VRAM, 12GB RAM, and approximately 100GB of cache was used. These specifications were sufficient for training CNN and LSTM models on large datasets (over 200,000 ECG images) and helped speed up the training process compared to relying solely on the CPU.

III.2 Model Architectures and Initial.

Following the initial adjustments of the algorithms and modified specific architectural properties of both the CNN and LSTM models to enhance their performance and improve classification accuracy. The models were executed multiple times on the same dataset to validate consistency and reliability before applying advanced preprocessing strategies.

III.2.1.For the CNN model, the architecture consisted of:
Convolutional layer (32 filters, 3×3 kernel) with ReLU activation.
MaxPooling2D (2×2)
Convolutional layer (64 filters, 3×3) with ReLU
MaxPooling2D (2×2)
Convolutional layer (128 filters, 3×3) with ReLU
MaxPooling2D (2×2)
Flatten operation
Fully connected Dense layer (128 units, ReLU) followed by Dropout (0.5)
Output layer with SoftMax activation corresponding to the number of ECG classes

This CNN achieved an overall classification accuracy of 91%. While the Normal (N) and Q classes were predicted with high precision (0.92–0.97), the model performed poorly on minority classes such as F, S, and V, yielding F1-scores below 0.50. This indicates a strong class imbalance problem that biased the model towards the majority class.

III.2.2. For the LSTM model, the architecture treated each ECG image as a sequence of pixel rows. Despite leveraging recurrent temporal modeling, the accuracy plateaued at 78%, and the classification report showed near-zero recall for minority classes (F, S, V, Q). The model demonstrated an over-reliance on the dominant N class, with recall reaching 1.0 for N but collapsing entirely for rare categories.

III.2.3. Although the CNN provided better overall accuracy than the LSTM, both models revealed systemic limitations prior to preprocessing:
Noise interference, caused by imperfect ECG recordings.
Class imbalance, where the Normal (N) class dominated.
Limited diversity, reducing the ability to generalize to unseen conditions.

These findings reinforced the necessity of applying advanced preprocessing techniques (denoising, augmentation, and resampling) to enhance class representation and improve generalization performance in subsequent stages.

III.3 Preprocessing & Data Augmentation

Based on the challenges identified in the first and second phases, particularly the noise in ECG images and the class imbalance (where the natural N class dominated the majority of samples), we applied a set of preprocessing techniques to improve the performance of the models. Rotation (±15°), simulating variations in ECG orientation. Horizontal and vertical flipping, improving the robustness of the model to mirrored signals. Zooming and scaling, altering the relative size of the ECG waveforms. Translation (shifting) mimicking variability in signal placement. These transformations enhanced the variability of the dataset and improved the generalization ability of the model.

1.Pix2Pix GAN: The Pix2Pix Generative Adversarial Network (GAN) was employed as a preprocessing stage to enhance ECG images before training classification models. The dataset contained raw ECG images that often-exhibited varying levels of visual noise and distortions
due to differences in acquisition sources. To address this, Pix2Pix GAN was applied to denoise and refine the images by learning a mapping between noisy and clean representations Before applying the Pix2Pix GAN, the raw ECG images collected from Kaggle and PhysioNet contained various noise artifacts, such as motion artifacts, baseline drift, and electrode misplacement effects. These distortions reduced the clarity of important ECG features (P wave, QRS complex, T wave) and directly impacted the performance of classification models such as CNN and LSTM. Pix2Pix GAN (Equations & Implementation)

Discriminator loss:

(1)

L_D = -E_ {x, y} [log D(x,y)] - E_{x}[log (1 - D (x, G(x)))]

Generator loss: (2)

L_G = -E_{x} [log D (x, G(x))] + λ * E_{x,y}[|| y - G(x) ||_1], λ=100

Implementation details to specify:
- Generator: U-Net with skip connections.
- Discriminator: Patch GAN (70x70).
- Optimizer: Adam (lr=2e-4, β1=0.5, β2=0.999).
- Input normalization: scale to [-1,1].
- Training: batch size, epochs, early stopping.

By applying Pix2Pix GAN as a denoising stage, we were able to significantly reduce the presence of noise and enhance the morphological clarity of the ECG waveforms. As a result, the processed images became more suitable for deep learning algorithms, allowing the CNN to better capture spatial features and the LSTM to extract temporal dependencies. This step created a stronger foundation for subsequent classification tasks and directly improved model generalization.

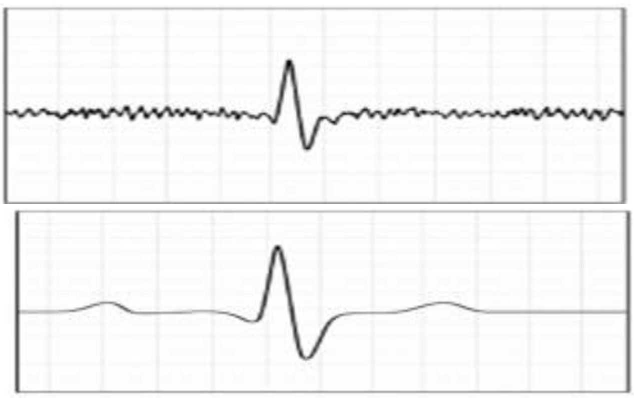


Figure 3. Comparison of ECG signals before and after denoising. The left panel shows a noisy ECG image with high-frequency disturbances, while the right panel illustrates the clean ECG image after preprocessing with Pix2Pix GAN, where noise is reduced and the waveform is smoother and more representative of the original cardiac signal.

2. Class Weighting & Resampling: To address the class imbalance issue, we added synthetic samples for rare classes (such as F, S, and V) and applied class weights during training.

Class Weighting & Resampling (Implementation Details). After applying the Pix2Pix GAN for denoising, the dataset still exhibited a significant imbalance where the Normal (N) class dominated the samples compared to minority classes such as F, S, and V. To handle this.

Class Weighting in Training:

During model compilation, class weights were calculated based on the inverse frequency of each class.

Weighted Cross-Entropy:

w_c = N / (K * n_c)

L_CE = -Σ w_c * y_c * log (p̂_c)

Where N = total samples, K = number of classes, n_c = number of samples of class c.

These weights were passed directly into the training function (model. Fit() in TensorFlow/Kera’s) so that the loss function penalized errors on rare classes (F, S, V) more heavily than errors on the dominant Normal (N) class. the Normal class received a smaller weight (close to 1), while the F class received a much larger weight (above 8–10), ensuring the model focused more on learning the underrepresented patterns.

Resampling of Data: For minority classes, data augmentation (rotation ±15°, zoom ±10%, translation, and flipping) was applied to artificially expand the number of samples. This oversampling strategy increased the diversityof F, S, and V classes without repeating identical samples. In addition, limited under sampling of the Normal class was performed to prevent the dataset from being overly biased while still keeping sufficient data for robust learning.The classes were expanded from 5 to 7 after recalibration and as a result of these adjustments, the dataset expanded from 54,613 original images to approximately 122,731 balanced images, with a much more proportional distribution across all classes. This balanced dataset was then used to train the CNN, LSTM, and Hybrid CNN–LSTM models, which demonstrated improved recall and F1-scores for the minority classes.

III. Table 3.1. Class Weights

Class	Samples (n_c)	Weight (w_c)
N	13661	0.27
S	557	6.64
V	1448	2.55
Q	1608	2.30
F	161	22.96
M	2506	1.47
F-resampled	556	6.68

III.3.4 Hybrid CNN–LSTM Model

To overcome the limitations faced by the independent CNN and LSTM models in the previous stages, we developed a hybrid model that combines Conv1D and LSTM to leverage the strengths of each. Initially, images were reshaped from a two-dimensional (width, height, 3) shape to a one-dimensional (width, height × 3) sequence, so that each row of the image could be treated as a time series. The data was then passed through Conv1D + MaxPooling1D + Batch Normalization layers to extract spatial and chronological features, followed by three successive LSTM layers (128 → 64 → 32) to process temporal patterns in a stepwise manner. Finally, Dense + SoftMax layers were added to complete the multi-class classification process.

This phase also saw the data size increase to approximately 246,729 ECG images thanks to the use of resampling and augmentation techniques, significantly improving the model's generalization ability.

The results showed that the hybrid model achieved an overall accuracy of 97.89%, significantly higher than both CNN and LSTM models alone. For example, the N (Normal Pulse) class achieved an accuracy of 98.9%, while unbalanced classes such as F-resampled and S achieved significant improvements thanks to the combination of spatial and temporal feature extraction.

To ensure that the reported improvements were statistically significant, we conducted a paired *t-test* between the baseline models (CNN and LSTM) and the proposed Hybrid CNN-LSTM model. The results confirmed that the accuracy and F1-score differences were not due to random chance indicating that the hybrid approach provided a statistically meaningful improvement over individual model. Despite the promising results, several limitations were observed during the study. the ECG images collected from PhysioNet PTB-XL and Kaggle contained unavoidable noise artifacts, which occasionally affected feature extraction. additionally, the dataset exhibited a class imbalance where the normal class dominated leading to reduced sensitivity in minority classes. Another limitation

is the restricted diversity of the dataset compared to real-world hospital environments. Moreover, the reliance solely on image-based ECG data, without integrating raw signals, may limit generalization. Computational constraints of Google Collab (GPU memory and RAM) also imposed restrictions on model complexity and batch size.in this works lies in combining multiple strategies to overcome the identified challenges. Specifically, Pix2Pix GAN was applied as a preprocessing step to denoise ECG images, improving signal clarity. Furthermore, data augmentation, resampling, and class weighting were introduced to alleviate class imbalance. A Hybrid CNN-LSTM model was designed to leverage both spatial and temporal features outperforming standalone CNN and LSTM models. Additionally, the dataset was expanded from five to seven classes, providing a richer classification framework. This integrated pipeline ensures better robustness, higher generalization, and improved diagnostic performance compared to existing approaches.

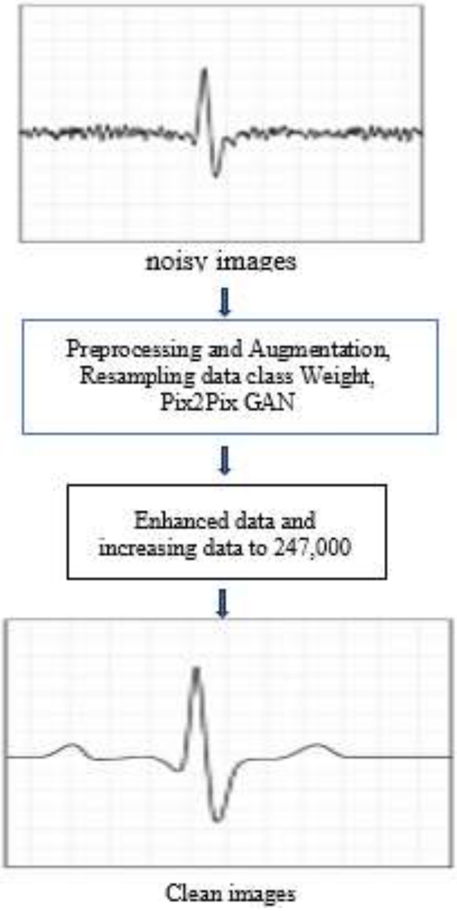


Figure 4. Illustrates an example of ECG images before noisy images and after clean images apply Pix2Pix GAN, Preprocessing and Augmentation

IV. RESULTS AND DISCUSSION

IV.4.1. Experimental Results and Discussion Most previous works have combined deep learning with ECG data for arrhythmia detection. For instance.[36] introduced Heart-Net a multi-modal system integrating MRI and ECG achieving over 92% accuracy but requiring costly imaging modalities. Similarly, Haleem et al. [37] applied a time-adaptive CNN-Bi LSTM using spectrograms, reporting nearly 95% accuracy for certain conditions. Earlier heartbeat classifiers [38] achieved competitive results but were limited by noise sensitivity and class imbalance. Unlike these studies, our work focuses on ECG image denoising via Pix2Pix GAN and a Hybrid CNN-LSTM classifier enabling robust performance (97.89% accuracy) on ECG-only datasets.

This study involved the design and execution of a series of sequential experiments aimed at developing an accurate and reliable model for classifying electrocardiogram (ECG) beats. The research methodology followed an evolutionary path, starting with baseline models and addressing their limitations through successive improvements in architecture, data processing, and finally, the integration of hybrid techniques.

Table 4.1 provides a summary of the performance of all tested models.

IV. TABLE 4.1. COMPARATIVE SUMMARY OF THE PERFORMANCE OF DIFFERENT CLASSIFICATION MODELS.

Model	Dataset	Overall Accuracy	Weighted Avg.F1-Score	Key Observation
Baseline CNN	54,613(5 classes)	64.41%	0.6369	Poor performance , complete bias towards class N
Improved CNN	54,613(5 classes)	91.00%	0.9000	Massive improvement but poor performance on rare classes persisted
Baseline LSTM	54,613(5 classes)	78.00%	0.6900	Complete failure to recognize any class other than N
CNN(Augmented Data)	122,731(7 classes)	95.00%	0.9400	Addressing imbalance led to balanced performance
LSTMon Augmented Data	122,731(7 classes)	73.00%	0.6100	Model failed to process image data effectively
Hybrid (CNN+LSTM)	122,731(7 classes)	96.12%	0.9583	Best balanced performance on this dataset
Hybrid with Aug.datasets (CNN+LSTM)	246,729(7 classes)	97.89%	0.9785	Best absolute performance with the larger dataset

IV. Table. 4.2. Comparison with previous works

Author(s)	Year	Dataset	Preprocessing	Model	Performance	Metrics (Precision, Recall, F1)
Sowmya & Jose	2022	MIT-BIH Arrhythmia dataset	Signal segmentation, normalization	CNN-LSTM	Accuracy ~ 94%	Precision ~93%, Recall ~92%, F1 ~92%
Omarov et al.	2023	ECG dataset (12-lead)	Preprocessing ECG signals	Conv-LSTM	Accuracy ~ 95%	Precision ~94%, Recall ~94%, F1 ~94%
Aversano et al.	2023	PTB-XL	ECG → 2D images, normalization	CNN-2D	Accuracy ~ 93%	Precision ~92%, Recall ~93%, F1 ~92%
Proposed Work (Our Study)	2025	246,729 ECG images (7 classes)	Pix2Pix GAN denoising, Augmentation, RGB, Resampling	Hybrid CNN-LSTM + Class Weights	Acc: 97.89%, F1: 0.9785	Precision, Recall, F1 improved after preprocessing

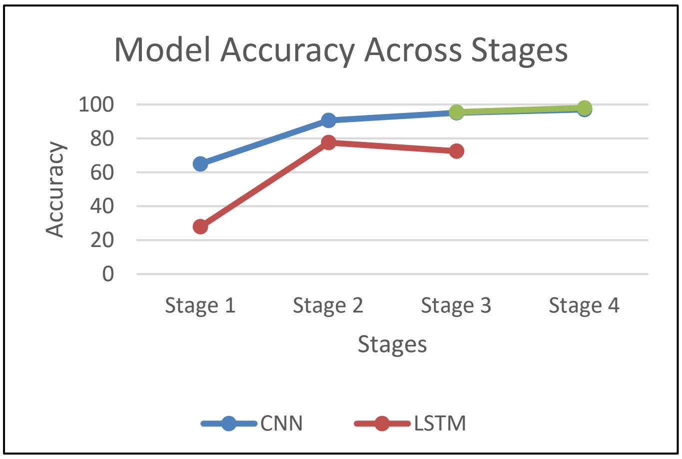


Figure 5. Comparison of model accuracy across different experimental stages for CNN, LSTM, and the proposed Hybrid CNN-LSTM. The results demonstrate consistent improvements in performance, with the Hybrid CNN-LSTM achieving the highest accuracy in the final stage.

IV.1.1. Foundation Phase and Initial Performance (Baseline Models). The initial Convolutional Neural Network (CNN) model (Section 1-1) demonstrated very modest performance, with an overall accuracy of only 64.41%. An analysis of the confusion matrix and classification report reveals:

- Structural Bias: The model virtually failed to recognize any instance from the four minority classes (F, Q, S, V), as their Precision, Recall, and F1-scores were 0.00 for most. This indicates the model simply learned to always predict the most frequent class (N), granting it an accuracy that merely reflected the natural data distribution rather than a genuine classification capability.
- Root Cause: This poor performance is attributed to two main factors: (1) A severe class imbalance, where class N constituted over 78% of the test data, and (2) A basic model architecture incapable of extracting the complex, discriminative features of the other classes.

In comparison, the performance of the baseline LSTM model (Section 1-2) was functionally worse. Despite reporting an accuracy of 78%, its confusion matrix showed it classified every single instance incorrectly as class N. This "misleading" accuracy conclusively confirms that the LSTM model did not learn any meaningful patterns from the raw image data presented to it indicating a fundamental mismatch between the input architecture (2D images) and the nature of LSTM models designed for 1D sequential data.

IV.1.2. Architectural Improvement and Data Processing Phase To overcome the limitations of the baseline model, improvements were made to the CNN architecture including increased depth (3 convolutional layers), more filters (32, 64, 128), and the addition of Dropout layers (0.5) to combat overfitting (Section 2-1). These enhancements resulted in a quantum leap in performance, achieving 91% accuracy. The F1-scores for class N improved to 0.96 and for class Q to 0.96, confirming the effectiveness of architectural improvements in enabling the model to learn more complex features.

However, performance on the minority classes (F, S, V) remained weak, with F1-scores not exceeding 0.30, 0.39, and 0.49 respectively. This clearly indicates that architectural improvement alone was insufficient to solve the imbalance problem, necessitating a move to a more robust strategy: data augmentation.

IV.1.3. Data Balancing and Hybrid Engineering Phase The dataset was expanded from 54,613 to 122,731 images using resampling techniques and synthetic data generation via Pix2Pix, introducing new classes (e.g., F-resampled) (Section 3). Applying the improved CNN model to this balanced dataset (Section 3-1) achieved 95% accuracy with a dramatic improvement in the performance of small classes:

- The *F1*-score for class F-resampled jumped to 0.71.
- The score for class S improved to 0.75.
- The model maintained high accuracy for class N (0.97) and Q (0.98). These results irrefutably prove that the quality and quantity of balanced data is a critical, if not the most important, factor in the performance of medical classification models.

When the LSTM model was applied to the expanded and organized dataset (Section 3-2), its poor performance continued, recording 73% accuracy again due to bias towards the majority class (N) and the misclassification of all other instances. This repeated failure confirms the hypothesis that processing images as flattened sequences without appropriate feature extraction is an ineffective approach.

Therefore, a hybrid (CNN + LSTM) model was proposed (Section 3-3), combining the best of both worlds: CNN layers extract spatial and visual features from the images, and these features are then fed into LSTM layers to capture long-term temporal dependencies within them. On the augmented dataset (122,731 images), this model outperformed all its predecessors, achieving an accuracy of 96.12% and a weighted average F1-score of 0.9583, with balanced performance across all classes without exception (the highest F1-score for the rare class S was 0.789).

IV.1.4. Final Scaling Phase

To test the limits of the proposed model, the experiment was scaled up to a massive dataset containing 246,729 images at 200x200 pixel resolution (Section 4). The hybrid (CNN+LSTM) model trained on this data achieved the best performance overall, with an overall accuracy of 97.89% and a weighted average F1-score of 0.9785. This result not only proves the robustness and effectiveness of the hybrid architecture but also illustrates the critical importance of a massive volume of high-quality data in achieving state-of-the-art performance in ECG signal analysis.

IV.2. Explainability of the Proposed Model

The proposed hybrid model represents a complex system ("black box"), making the interpretation of its decisions mandatory, not optional, for sensitive medical applications. To ensure the model's reliability and clinical acceptance, we propose an explanation framework based on Explainable AI (XAI) techniques.

The chosen approach is to apply Gradient-weighted Class Activation Mapping (Grad-CAM) to the final CNN layers within the hybrid model. This technique generates "heatmaps" that highlight the regions of the input image that were most influential in the model's final classification decision. For instance:

- When classifying a beat as "V" (Ventricular), the heatmap is expected to intensely highlight the wide QRS complex and large T-wave, the two hallmark features of this beat type.
- Conversely, when classifying a beat as "S" (Sinus pause), the heatmap should show a focus on the isoelectric line where the complete absence of electrical activity is the diagnostic feature. Clinical Validation Ensuring the model is learning the correct biophysical features used by cardiologists rather than superficial correlations or noise in the data. Error Diagnosis in cases where the model errs the heatmap provides immediate insight into the reason. Did the model focus on the wrong region of the signal? Was the original signal quality poor? Building Trust and Transparency It gives clinicians a window into the model's internal reasoning process increasing their confidence in its recommendations and encouraging its adoption as an assistive tool rather than a replacement. Therefore, model explainability is the crucial bridge connecting high computational performance to safe and effective clinical application representing a critical step in translating this technology from the lab environment to the bedside.

IV.3 Limitations

Although the proposed approach achieved promising results, several limitations remain. First, the dataset used in this study was obtained mainly from publicly available sources (PhysioNet and Kaggle). While augmentation and resampling techniques were applied to increase diversity, the data still represent a single-source environment, which may restrict the generalizability of the findings. Second, the experiments were performed on ECG images rather than raw ECG signals, which might limit the richness of temporal and spectral information available for analysis. Third, all experiments were conducted using Google Colab with limited GPU resources, imposing restrictions on model complexity and large-scale training. Finally, despite the hybrid CNN-LSTM achieving an accuracy of 97.89%, its performance may degrade when tested on real hospital data due to differences in acquisition devices, noise characteristics, and patient populations.

On the other hand, some technical challenges were effectively addressed in this study. Image noise was mitigated using Pix2Pix GAN for denoising, while class imbalance was alleviated through resampling strategies combined with class weights. These steps significantly improved both the robustness and accuracy of the model but the remaining limitations highlight directions for future work.

V. CONCLUSION

This research tackled the challenge of automated cardiac disease detection through ECG image classification, exploring the potential of image-based analysis as an alternative to traditional signal processing. The study aimed to compare the performance of standalone CNN and LSTM models and ultimately develop a hybrid CNN-LSTM architecture to leverage both spatial and temporal features, thereby improving classification accuracy.

To achieve this ECG images were collected from public datasets like PhysioNet and Kaggle. Initial experiments on raw data revealed critical limitations: noise severe class imbalance, and poor generalization. A comprehensive preprocessing pipeline was implemented to address these issues utilizing Pix2Pix GAN for denoising, resampling for class rebalancing and data augmentation to increase variability. The hybrid CNN-LSTM model was then trained on this enhanced dataset to combine spatial feature extraction with temporal sequence modeling.

The experimental results demonstrated a significant improvement in performance, highlighting the effectiveness of the proposed preprocessing pipeline and the hybrid architecture. The final model achieved an overall accuracy of 97.89%, surpassing the standalone models and providing balanced classification across all ECG classes.

Despite these promising results, limitations remain, primarily regarding the dataset's origin from public sources, which may affect generalizability to diverse clinical environments. Future work should focus on validating the model with multi-source hospital data integrating multimodal data and optimizing the architecture for efficient real-time clinical deployment.

REFERENCES

1. Berger, L., Haberbusch, M., & Moscato, F. (2023). Generative adversarial networks in electrocardiogram synthesis: Recent developments and challenges. *Artificial Intelligence in Medicine, 143*, 102632. https://doi.org/10.1016/j.artmed.2023.102632
2. Omarov, B., & Momynkulov, Z. (2024). Hybrid deep learning model for heart disease detection on 12-lead electrocardiograms. *Procedia Computer Science*, 439–444. https://doi.org/10.1016/j.procs.2024.08.061
3. Qu, Z., Chen, W., & Tiwari, P. (2024). HQ-DCGAN: Hybrid quantum deep convolutional generative adversarial network approach for ECG generation. *Knowledge-Based Systems, 301*, 112260. https://doi.org/10.1016/j.knosys.2024.112260
4. Qin, Y., et al. (2023). MVKT-ECG: Efficient single-lead ECG classification for multi-label arrhythmia by multi-view knowledge transferring. *Computers in Biology and Medicine, 166*, 107503. https://doi.org/10.1016/j.compbiomed.2023.107503
5. R, G., et al. (2024). Enhancing heart disease prediction with reinforcement learning and data augmentation. *Systems and Soft Computing, 6*, 200129. https://doi.org/10.1016/j.sasc.2024.200129
6. Chen, C., Jin, B., Che, C., & Li, R. (2024). OSGAN: Omni-scale and global-aware ECG arrhythmia diagnostic network. *Biomedical Signal Processing and Control, 96*, 106602. https://doi.org/10.1016/j.bspc.2024.106602
7. Tang, K., Ma, S., Sun, X., & Guo, D. (2024). Optimizing machine learning for enhanced automated ECG analysis in cardiovascular healthcare. *Egyptian Informatics Journal, 28*, 100578. https://doi.org/10.1016/j.eij.2024.100578
8. Karthikeyani, S., Sasipriya, S., & Ramkumar, M. (2024). Cardiac arrhythmias detection framework based on higher-order spectral distribution with deep learning. *Biomedical Signal Processing and Control, 92*, 105997. https://doi.org/10.1016/j.bspc.2024.105997
9. Safdar, M. F., Nowak, R. M., & Pałka, P. (2024). Pre-processing techniques and artificial intelligence algorithms for electrocardiogram (ECG) signals analysis: A comprehensive review. *Computers in Biology and Medicine, 170*, 107908. https://doi.org/10.1016/j.compbiomed.2023.107908
10. Cao, Q., et al. (2021). Practical fine-grained learning-based anomaly classification for ECG image. Artificial Intelligence in Medicine, 119, 102130. https://doi.org/10.1016/j.artmed.2021.102130
11. Alqahtani, H., Aldehim, G., Alruwais, N., Assiri, M., Alneil, A. A., & Mohamed, A. (2024). Leveraging electrocardiography signals for deep learning-driven cardiovascular disease classification

model. Heliyon, 10(16), e35621. https://doi.org/10.1016/j.heliyon.2024.e35621

12. Gour, A., Gupta, M., Wadhvani, R., & Shukla, S. (2024). ECG-based heart disease classification: Advancement and review of techniques. Procedia Computer Science, 1634–1648. https://doi.org/10.1016/j.procs.2024.04.155
13. Jiang, S., Li, D., & Zhang, Y. (2023). A deep neural network based on multi-model and multi-scale for arrhythmia classification. Biomedical Signal Processing and Control, 85, 105060. https://doi.org/10.1016/j.bspc.2023.105060
14. Hua, J., Zou, J., Zou, F., & Liu, J. (2024). A multi-channel ECG signal deep compressive sensing method using tree-shaped autoencoder based on multiscale feature fusion. Biomedical Signal Processing and Control, 94, 106272. https://doi.org/10.1016/j.bspc.2024.106272
15. Saputra, N. A., Riza, L. S., Setiawan, A., & Hamidah, I. (2024). A systematic review for classification and selection of deep learning methods. Decision Analytics Journal, 12, 100489. https://doi.org/10.1016/j.dajour.2024.100489
16. Premalatha, P., et al. (2024). A novel method for detecting cardiac arrhythmias in electrocardiogram signals using the ANFIS classifier. e-Prime - Advances in Electrical Engineering, Electronics and Energy, 9, 100734. https://doi.org/10.1016/j.prime.2024.100734
17. Alla, M. R., & Nayak, C. (2024). A robust ECG signal enhancement technique through optimally designed adaptive filters. Biomedical Signal Processing and Control, 95, 106434. https://doi.org/10.1016/j.bspc.2024.106434
18. Sindhu, V. S., Lakshmi, K. J., Tangellamudi, A. S., & Begum, K. G. (2023). A novel deep neural network heartbeats classifier for heart health monitoring. International Journal of Intelligent Networks, 4, 1–10. https://doi.org/10.1016/j.ijin.2022.11.001
19. Li, C., et al. (2021). DeepECG: Image-based electrocardiogram interpretation with deep convolutional neural networks. Biomedical Signal Processing and Control, 69, 102824. https://doi.org/10.1016/j.bspc.2021.102824
20. Omarov, B., & Momynkulov, Z. (2024). Hybrid deep learning model for heart disease detection on 12-lead electrocardiograms. Procedia Computer Science, Elsevier B.V., 439–444. https://doi.org/10.1016/j.procs.2024.08.061
21. Mavaddati, S. (2025). ECG arrhythmias classification based on deep learning methods and transfer learning technique. Biomedical Signal Processing and Control, 101, 107236. https://doi.org/10.1016/j.bspc.2024.107236
22. Ashtaiwi, A. A., Khalifa, T., & Alirr, O. (2024). Enhancing heart disease diagnosis through ECG image vectorization-based classification. Heliyon, 10(18), e37574. https://doi.org/10.1016/j.heliyon.2024.e37574
23. Alqahtani, H., Aldehim, G., Alruwais, N., Assiri, M., Alneil, A. A., & Mohamed, A. (2024). Leveraging electrocardiography signals for deep learning-driven cardiovascular disease classification model. Heliyon, 10(16). https://doi.org/10.1016/j.heliyon.2024.e35621
24. Oke, O. A., & Cavus, N. (2025). A systematic review on the impact of artificial intelligence on electrocardiograms in cardiology. International Journal of Medical Informatics, 195, 105753. https://doi.org/10.1016/j.ijmedinf.2024.105753
25. Rahul, J., & Sharma, L. D. (2025). Advancements in AI for cardiac arrhythmia detection: A comprehensive overview. Computer Science Review, 56, 100719. https://doi.org/10.1016/j.cosrev.2024.100719
26. Jafari, M., et al. (2023). Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review. Computers in Biology and Medicine, 160, 106998. https://doi.org/10.1016/j.compbiomed.2023.106998
27. Karoui, H., Hamza, S., & ben Ayed, Y. (2024). Deep learning for cardiac diseases classification. In Lecture Notes in Computer Science (pp. 170–182). https://doi.org/10.1007/978-3-031-70816-9_14
28. Alqahtani, H., Aldehim, G., Alruwais, N., Assiri, M., Alneil, A. A., & Mohamed, A. (2024). Leveraging electrocardiography signals for deep learning-driven cardiovascular disease classification model. Heliyon, 10(16), e35621. https://doi.org/10.1016/j.heliyon.2024.e35621
29. Ayano, Y. M., et al. (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
30. Lee, J., & Shin, M. (2024). Using beat score maps with successive segmentation for ECG classification without R-peak detection. Biomedical Signal Processing and Control, 91, 105982. https://doi.org/10.1016/j.bspc.2024.105982
31. Issa, M. F., et al. (2023). Heartbeat classification based on single lead-II ECG using deep learning. Heliyon, 9(7), e17974. https://doi.org/10.1016/j.heliyon.2023.e17974
32. Zhang, Y., Yi, J., Chen, A., & Cheng, L. (2023). Cardiac arrhythmia classification by time–

frequency features inputted to the designed convolutional neural networks. Biomedical Signal Processing and Control, 79, 104224. https://doi.org/10.1016/j.bspc.2022.104224

33. Hussain, M., Hameed, T., & Afzal, H. (2023). Enhancing heart disease prediction with reinforcement learning and data augmentation. Biomedical Signal Processing and Control, 80, 104340. https://doi.org/10.1016/j.bspc.2022.104340
34. Sharma, P., & Tripathi, S. (2022). ECGencode: Compact and computationally efficient deep learning feature encoder for ECG signals. Biomedical Signal Processing and Control, 75, 103604. https://doi.org/10.1016/j.bspc.2022.103604
35. G. M. Zebari, D. A. Zebari, D. Q. Zeebaree, H. Haron, A. M. Abdulazeez, and K. Yurtkan, "Efficient CNN Approach for Facial Expression Recognition," in Journal of Physics: Conference Series, IOP Publishing Ltd, Dec. 2021. doi: 10.1088/1742-6596/2129/1/012083.
36. Dias, F. M., et al. (2023). Artificial intelligence-driven screening system for rapid image-based classification of 12-lead ECG exams: A promising solution for emergency room prioritization. IEEE Access, 11, 121739–121752. https://doi.org/10.1109/ACCESS.2023.3328538
37. Alsekait, D. M., et al. (2024). Heart-Net: A multi-modal deep learning approach for diagnosing cardiovascular diseases. Computers, Materials and Continua, 80(3), 3967–3990. https://doi.org/10.32604/cmc.2024.054591
38. Haleem, M. S., et al. (2021). Time adaptive ECG driven cardiovascular disease detector. Biomedical Signal Processing and Control, 70, 102968. https://doi.org/10.1016/j.bspc.2021.102968