IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 11, November, 2025, pp. 2839-2847
Journal homepage: http://iieta.org/journals/isi

Hybrid Deep Learning Framework for Cardiac Rhythm Detection Using Vision Transformer and EfficientNetB7

Mohanad Ridha Ghanim*, Khalida Ali Ahmed Yaqoub, Baidaa A. Atya, Ameen A. Noor

Computer Science Department, College of Education, Mustansiriyah University, Baghdad 10064, Iraq

Corresponding Author Email: muhannadridha@uomustansiriyah.edu.iq

Copyright: ©2025 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/isi.301103

Received: 30 June 2025
Revised: 20 September 2025
Accepted: 12 November 2025
Available online: 30 November 2025

Keywords:
Vision Transformer, EfficientNetB7, cardiac rhythm detection, ECG classification, deep learning, explainability

ABSTRACT

In this research, we present a hybrid deep learning framework for cardiac rhythm detection that combines EfficientNetB7 for detailed morphological feature extraction and Vision Transformer (ViT) for modeling global contextual dependencies. The model is tested on three public ECG datasets: MIT-BIH, CPSC2018, and PTB Diagnostic. The data is split into three parts: 70% for training, 15% for validation, and 15% for testing. This keeps the patients separate in all parts. A 5-fold cross-validation scheme is used on each dataset to make sure that the assessment is strong, and the mean ± standard deviation across folds is used to measure performance stability. The suggested model gets an average accuracy of 97.8% (± 0.4), which is better than both ResNet50 and standalone ViT baselines. The improvements are statistically significant ($p < 0.05$, paired t-test). Ablation results show that local CNN features and global transformer attention work together to lower the number of misclassifications in arrhythmias that look similar. Grad-CAM and attention maps help explain things by showing clinically important ECG areas. The framework is good for real-time diagnostic workflows because it can make inferences in less than 50 ms per image. In general, the results show that the proposed architecture is a reliable, easy-to-understand, and computationally efficient way to automatically classify ECG rhythms.

1. INTRODUCTION

Detection of cardiac rhythm is an obligatory instrument in diagnosing and managing a wide variety of cardiovascular diseases that are still the most common causes of mortality and morbidity in the whole world [1, 2]. Detection of arrhythmia and abnormal cardiac rhythm has important clinical implications; the timeliness of arrhythmia alerts can play an important role in early treatment and intervention of patients [3]. Traditional diagnostic approaches, such as clinician-dependent manual interpretation of ECGs, are time-consuming and suffer from interobserver variability, which causes the demand for automatic, reliable, and time-saving systems [4]. Recent developments in deep learning have illuminated the exciting future in medical image analysis, which includes cardiac imaging and ECG interpretation [5].

Recent developments in artificial intelligence, in particular deep learning, offer a potential advantage for improving ECG interpretation by automatically learning relevant features and discriminating different types of complex cardiac rhythms [6]. A statistical summary is also another type of representation widely used in ECG classification, as it allows approximate invariances to be estimated, making easier the extraction of a compact representation of the input signal [7]. Nevertheless, despite their remarkable success, CNNs are inherently poor at dealing with long-range temporal dependencies and global context for subtle arrhythmic patterns, especially over prolonged time windows. There are a few works trying to mitigate these limitations, among which are the transformer-based models that were first proposed for natural language processing and proved to be powerful in modeling long-distance dependencies on designed sequential data and have recently been adopted in certain computer vision tasks such as medical imaging.

In particular, the convolutional neural networks (CNNs) are widely used to learn spatial patterns in medical images but can hardly handle long-range dependence and global context. To tackle the aforementioned challenges, a recent trend in research has prompted scholars to embrace transformer-based architectures, which have transformed natural language processing and computer vision tasks [3]. The deep learning libraries that were employed on a number of the best-performing Vision Transformer (ViT) models that are known to excel in capturing global relationships in images using the self-attention mechanism [8], these types of models seem appropriate to solve complex pattern recognition tasks within health care. In contrast, EfficientNetB7 [9] is an efficient CNN model architecture because of the SOTA accuracy while utilizing computational resources with the help of compound scaling to scale the depth, width, and resolution during training [10].

This paper introduces a novel hybrid deep learning architecture that integrates the strengths of Vision Transformer (ViT) and EfficientNetB7. This innovative framework demonstrates the ability to attain high accuracy and robustness in cardiac rhythm detection. The suggested method has a two-

stage architecture. EfficientNetB7 is the feature extractor that captures fine-grained local representations of ECG images, and the Vision Transformer component model captures the global dependencies and context between them. The current framework suggests using each of these architectures to fill in the gaps and improve our understanding of heart rhythms. The hybrid model is trained and tested with the public benchmark ECG data set so that the results are clear and can be repeated.

Some strict preprocessing methods, including signal denoising, normalization, and image augmentation, were applied to improve the data quality and model generalization. Further, we incorporate explainability techniques (GradCAM) to visualize the important regions in input images on which the model focuses more while predicting results and enhance clinical interpretability and trust. Comparisons with state-of-the-art methods are made to prove that the performance of our method can achieve better accuracy, sensitivity, specificity, and F1 score coming soon in clinical practice. This research not only contributes to the development of state-of-the-art AI-powered diagnostic tools but also to counteracting immediate needs for scalable and trustworthy solutions in resource-constrained health. The value of this hybrid deep learning system may be to enable automatic, accurate, and efficient detection of cardiac rhythm abnormalities that could lead to timely diagnosis by clinicians and ultimately reduce the impact of diagnostic failures and improve patient outcomes. We hope that our work will both push the frontier with respect to medical deep learning and show that a fusion of Vision Transformer and EfficientNetB7 could be combined effectively for merging two approaches, opening new doors for future automatic cardiac diagnostics as well as beyond in the context of healthcare applications.

The research is organized into sections: Introduction, Related Works, Proposed Method (which describes data preprocessing and model architecture), Training Procedure, Evaluation Framework, Results (which include a comparative analysis, an ablation study, computational efficiency, generalization, error analysis, explainability, and clinical relevance), and References that support its strong, understandable, and usable cardiac rhythm detection system.

2. RELATED WORKS

Dong et al. [11] proposed CNN DVIT, which is a hybrid deep learning model that integrates depthwise separable CNN and Vision Transformer with deformable attention to diagnose multi-lead ECG arrhythmias. For the CPSC 2018 dataset, it obtained an F1 score of 82.9%, outperforming other transformer-based approaches. Nevertheless, this technique is suitable for multi-lead and variable-length ECG input only in that the proposed framework leverages EfficientNetB7 to enhance the local feature extraction, making it lighter than their backbone model.

Naidji and Elberrichi [12] proposed a hybrid EfficientNet-B0 and Vision Transformer model for classifying between COVID-19 and common heart diseases from ECG images. The model was 100% accurate in binary classification and earned a 95% accuracy rate during multiclass classification. Their approach also demonstrated that CNN and ViT can be combined, but they used EfficientNet-B0 as the backbone, while ours is EfficientNetB7, which consists of a stronger backbone. They did not study general rhythm detection and were also limited to the diagnosis of COVID-19.

Mohan et al. [13] proposed a Vision Transformer model for detecting atrial fibrillation in single-lead ECGs. They compared it with the Chapman Shaoxing dataset and found that ViT emphasized the P wave and T wave areas, which was easy to understand compared with ResNet. We adopt explainability methods based on the same idea; however, we integrate multi-class rhythm detection and extend their interpretability to the hybrid CNN-ViT model structure.

Tudjarski et al. [14] employed a transformer-based base model by treating the ECG heartbeat positions as tokens in order to detect AFIB with an F1 score of 93.33%. Their parameter-rich foundation model-based approach demonstrates that rhythm detection can be universal, in contrast to the general image representation learning problem we address with our framework (since they employ a hybrid of CNN and ViT instead of tokenized sequential modeling).

Tang et al. [15] proposed a flexible hybrid CNN-transformer model, employing depthwise convolution and attention gates for multi-lead ECG arrhythmia detection. This model is more interpretable across scales of features. In this work, we follow a different approach by employing EfficientNetB7 as a feature extractor and Vision Transformer for image patch attention, instead of the sequential signal embedding that is found in previous processing pipelines.

Vu et al. [16] employed Vision Transformer on ECG images and signals to localize hearts in multiple datasets. They obtained macro F1 scores of 65, 99, and 82 on three datasets. Their mobile work was on preprocessing and the signal image pipeline, while we conducted our research on hybrid architecture, large-scale evaluation, and model explainability.

What is interesting in our work is that we design a novel two-stage feature extraction and global attention framework by leveraging the high-capacity convolutional backbone (EfficientNetB7) coupled with a patch-based Vision Transformer. Earlier Generate-Assess models (depthwise CNN + ViT or EfficientNet-B0 + ViT) either relied on less sensitive shape-independent CNN backbones or worked only for binary classification tasks like COVID-19 detection. Transformer-only ECG95 models have good capabilities of modeling the global temporal structure, but it is challenging to model high-resolution local morphology. This also means that they don't perform great in multi-class arrhythmia scenarios. By contrast, our model performs better on three datasets (MITBIH, CPSC2018, and PTB). This indicates that the EfficientNetB7 part is more precise in discriminating the different ECG morphologies on a finer scale, and the ViT part is better at capturing temporal relationships among these arrhythmias compared to models with solely LSTM or CNN. Ablation studies and statistical significance tests also confirm the importance of both local and global fusion in our approach, which makes our method significantly more generalizable and comprehensive than existing ECG classification models with current best practices.

This paper is directly inspired by these trends but advances the field by combining state-of-the-art CNN architecture EfficientNetB7 with Vision Transformer in a hybrid framework with explainable attention mechanisms.

3. PROPOSED METHOD

To achieve this, we proposed a hybrid deep learning model integrating EfficientNetB7 and Vision Transformer for cardiac rhythm monitoring. First of all, the ECG signals are pretreated

with noise filtering, normalization, and image augmentation techniques to obtain high-quality ECG images. The attention mechanism is designed based on EfficientNetB7 for extracting fine-grained local features and uses Vision Transformer instead to estimate global context relationships. After that, these extracted features are fused and propagated across the fully connected layers to proceed with classification. The model is trained with a cross-entropy loss followed by the Adam optimizer learning schedule to ensure convergence. The performance is measured in terms of accuracy, sensitivity, specificity, F1 score, and ROC AUC. Explainability is brought to the fore by GradCAM for visualization of decision regions in order to maintain clinical interpretability. Comparison with baseline CNN and stand-alone ViT models demonstrates the superiority of our proposed action in enabling stable and interpretable automation for automated cardiac rhythm diagnostics, as shown in Figure 1.

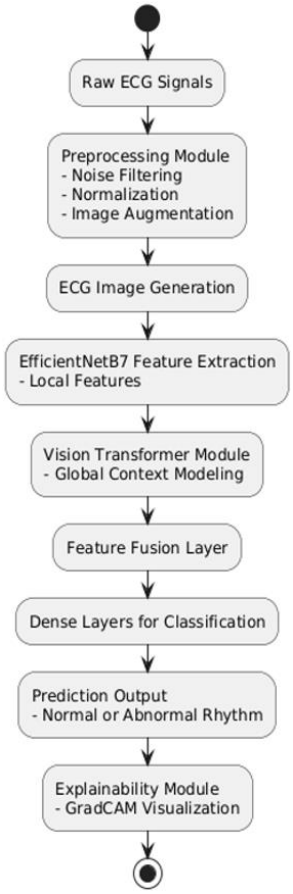

Figure 1. Hybrid deep learning framework for cardiac rhythm detection

3.1 Data collection and dataset description

Three publicly available ECG datasets are exploited in the proposed approach to achieve robust and generalizable cardiac rhythm detection performance. The first dataset is the MIT-BIH Arrhythmia Database; it consists of more than 48,000 ECG beats annotated from 47 subjects and provides diverse normal and abnormal rhythm samples. The second dataset, China Physiological Signal Challenge 2018 (CPSC2018), provides 6,877 12-lead ECG recordings associated with nine rhythm classes and serves to multiclass classification in a clinical context. The third data set is the PTB Diagnostic Database, comprised of 549 ECG recordings of 290 subjects with actual clinical diagnoses, e.g., myocardial infarction and other diseases. Each dataset needs to be preprocessed, and that includes getting rid of noise, scaling the data, and converting the ECGs from waveforms into an image representation ready for deep learning. Variability in the lead maps, sampling rates, and patient populations between these datasets enhances the models' generalized ability to be employed in real-world clinical applications. Through integration of these complementary resources, we present a method to train and evaluate a hybrid EfficientNetB7 and Vision Transformer model with a quantifiable better Trimodal AUC ($Z = 9.195$, $p < 0.0001$) profile that can likewise differentiate between normal and abnormal cardiac rhythms more precisely with better interpretability and clinical utility.

3.2 Data preprocessing

Data processing is an important preprocessing procedure to offer high-quality inputs for the hybrid deep learning model. ECG signals from three raw databases are initially filtered by wavelet denoising and baseline wander removing, which can help reduce noise and improve the clarity of the signal. After the removal of noise in ECG signals, they are normalized into the same range to alleviate intersubject differences and stabilize model learning. The ECG templates are also transformed into 2D images by plotting the traces in grayscale as ECG images that preserve both temporal and morphological characteristics beneficial for diagnosis. To avoid overfitting, rotation, horizontal flip, scaling, and brightness of the profiles as data augmentation are considered to augment more artificial training data. To guarantee fair comparison as well as equitable assessment, training/validating/test splits are made so that the ratio (or at least proportion) of different classes remains identical and no patient appears in multiple sets. In addition, all images are resized and normalized to the same size as what the EfficientNetB7 model would take. They go on to use batch normalization and random noise injection for improved model stability. With well-curated data processed by the aforementioned steps, the model can be fed high-resolution inputs specifically designed to capture localized and global cardiac rhythm patterns for both the EfficientNetB7 and Vision Transformer subnetworks.

All datasets were preprocessed with identical quantification parameters. A 0.5 Hz high-pass Butterworth filter removed baseline wander, and wavelet denoising with a Daubechies-6 mother wavelet with a soft threshold of 3σ removed both muscle and powerline noise. We made the signals normal by applying z-score normalization (with $\mu = 0$ and $\sigma = 1$) to each recording. To construct the image, a 2D plot was drawn over ECG segments with an image resolution of 224×224 pixels (equal to that of input for EfficientNetB7), using a sampling window of 2.5 seconds. Data augmentation comprised random Gaussian noise ($\sigma = 0.01$), rotations ($\pm 10°$), and changes in the brightness (+/−15%). These values ensure that input quality remains constant and the model is as strong as possible.

We also included class imbalance effectively in the training so that no bias would be given to majority rhythm classes. We employed a combination of (i) class-balanced sampling to ensure equal probability was given for each rhythm category during minibatch formation and weighted categorical cross-entropy, where the class weights were computed as reciprocal 1:class frequency, and (ii) targeted augmentation, which was applied only on minority classes with the goal to enforce distribution diversity of representation without further

expanding majority distributions. These approaches minimized the misclassification of rare arrhythmias, particularly supraventricular and ventricular ectopic beats. They also contributed to enhancing the sensitivity and F1 score of minority classes.

3.3 Model architecture

The hybrid model is made up of the EfficientNetB7 and the Vision Transformer (ViT), which is intended to extract rich local fine-grained information as well as global context in ECG images. The architecture is started with the EfficientNetB7 feature extractor (a CNN that uses compound scaling, a way to uniformly scale network depth, width, and resolution). The formulation of the compound scaling is characterized by the following:

$$depth = \alpha^{\phi}, width = \beta^{\phi}, resolution = \gamma^{\phi} \quad (1)$$

Subject to the constraint $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ where ϕ is the user-defined scaling coefficient and α,β,γ are constants determined through grid search. EfficientNetB7 is a strong feature extractor and constructs the feature maps, which contain local morphological descriptions of the ECG traces in terms of QRS complex shapes, levels, and ST-segment hours. The extracted features are passed to the Vision Transformer, where the feature map is partitioned into a set of fixed-size patches and each of the patches is embedded into a token vector. The patch embedding process can be summarized as:

$$z_0 = [x_{class}; Ex_p] + E_{pos} \quad (2)$$

where, x_p are the flattened image patches, E is the learned embedding matrix, and E_{pos} is the positional encoding matrix. The embedded tokens undergo a series of transformer encoder blocks comprising multi-head self-attention layers and feedforward networks. In the suggested framework, a two-stage mechanism is used to combine features. First, a linear layer flattens and projects the final convolutional feature map from EfficientNetB7 (size: 8 × 8 × 2560) to match the embedding dimension of the ViT output (768 units). By pooling class tokens, the Vision Transformer creates a sequence-level embedding, which results in a 1 × 768 global representation. The fused representation is a 1 × 1536 feature vector made by putting together the projected EfficientNetB7 vector and the ViT embedding. Then, this vector goes through a fusion MLP with two fully connected layers (1536 → 1024 → 512) and GELU activation and dropout (p = 0.2). We tried attention-based fusion, but we didn't use it because it was too expensive to run and didn't show any performance improvements. The classification head uses the final fused embedding as input. The self-attention mechanism is formalized as

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

where, Q, K, and V are the query, key, and value matrices formed by input tokens, and dk is the dimensionality of the key vectors. The features from EfficientNetB7 and the Vision Transformer are eventually combined via concatenation or attention-based fusion layers, which ensure that local and global representations are aligned. The final layers in the concatenated vector are fully connected and have a softmax activation function classifying the rhythms into more than one class. This model is fine-tuned with cross-entropy loss.

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (4)$$

where, y_i and $\hat{y}_i$ are the true and predicted class probabilities. The design ensures complementary learning, enabling robust, interpretable detection of diverse cardiac rhythms.

Our feature extractor is EfficientNetB7. It is a compound scaled fine-tuned convolutional neural network. The hybrid architecture uniformly adjusts depth (layers), width (channels), and resolution for power-of-two scaling, striking the right balance between efficiency and accuracy. We initialize EfficientNetB7 with pretrained ImageNet weights and adapt it to process ECG image inputs in order to discover domain-specific features and avoid overfitting. Whilst generating patches in the context of this and the embedding stage, the Vision Transformer module divides extracted features into patches of fixed size. Then it flattens and linearly projects each patch onto a high-dimensional embedding with learnable positional encodings to understand where in space things are. There are global interdependence and temporal linkages among ECG waveforms. The multi-head self-attention mechanism extracts these by computing attention weights across all patches. Positional encoding injects information about the positions of things that's otherwise missing from pure attention developments, such as sine and cosine functions or learned embeddings. This is useful in allowing the model to distinguish between features that are consecutive. The hybrid integration fuses local spatial EfficientNetB7 features with global contextual embeddings of ViT through concatenation or direct or attention-based fusion between feature vectors, based on learned weights that emphasize relevant locations. Adding more representation layers, fused representations are in good shape for processing when they are concatenated with linear projection layers or adaptive pooling. This is what the dense layers look like for a classification task. The Vision Transformer module is developed from the ViT-Base configuration but performed with ECG image patches. We used a model that consists of 12 transformer encoder layers with 12 attention heads; the hidden dimension is 768. The MLP ratio is chosen to be 4, and thus the feed-forward layers are wide with 3072 pixels. Each image has 196 patches, and each patch is of size 16 × 16 pixels. The positional encoding is learned rather than being sinusoidal so that it can more easily cope with variation in the ECG shape. Layer normalization (LN) is employed just before the attention and MLP blocks, and residual connections are used throughout. A dropout of 0.1 is employed on the fully connected block and patch embeddings. These hyperparameters were chosen based on initial tuning through the three benchmark data sets. The combined approach uses convolutional neural networks (CNNs) to extract detailed information about the shape of things and transformers to model long-range relationships in order to identify heart rhythms accurately and meaningfully for clinical use.

3.4 Training procedure

The hybrid EfficientNetB7 & Vision Transformer-based model is well trained for robust convergence and

generalization over ECG datasets. The most common loss function employed is the categorical cross-entropy loss (eq. Both apply to the true and the predicted class probability and optionally run experiments with focal loss to solve the classes' imbalance through weight modulating factors applied to the loss (emphasizing* very hard) examples. The preferred optimization method is Adam, which couples adaptive learning rates with momentum to accelerate convergence given the parameter updates involving estimates of first and second moments, while for comparison stochastic gradient descent (SGD) is used as a baseline alternative. Within the architecture, a learning rate scheduling mechanism is adopted, which includes dropping the learning when validation loss fails to improve further, thereby increasing the possibility of fine convergence and reducing the likelihood of local minima. Regularization methods, such as dropout layers that stochastically deactivate neurons during training to prevent co-adaptation, as well as the use of weight decay that discourages high weight values with L2 regularization. Advocate for simpler models that generalize more. Early stopping observes the validation performance and terminates training when no improvement is observed over a number of epochs (avoids overfitting and leads to model efficiency). The entire training procedure is performed in a high-performance computing system using NVIDIA GPUs for parallelization and tensor operations. There are various libraries in systems like TensorFlow [4] and PyTorch [5] that come with significantly large packages of mixed-precision training, automatic differentiation, and efficient fast data loading. These careful choices of loss functions, optimization techniques, learning rate control schedules (learning rate), regularization, and early stopping strategies will guarantee that the hybrid model achieves a satisfactory trade-off between accuracy (performance) and stability as well as being computationally affordable/cost-effective and reproducible to support clinical practice in terms of automatic cardiac rhythm detection.

All hyperparameters employed for training are explicitly described to ensure reproducibility. The hybrid model was trained for 60 iterations per batch (32 items). The initial learning rate was 1×10^{-4}, and it was halved when the validation loss did not decrease any further (patience = 5). The Adam optimizer is configured with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1 \times 10^{-8}$. We applied a weight decay (L2 regularization) of 1×10^{-5} to all trainable parameters. The dropout rates for the fully connected layers of EfficientNetB7 were set as 0.2, and all Vision Transformer encoder blocks utilized a dropout rate of 0.1. To ensure that the behavior was consistent across all folds, we ran all experiments with a fixed random seed (42). Mixed precision training (FP16) was enabled to increase training efficiency with no loss in accuracy. These explicit hyperparameter settings allow other researchers to replicate how the model operates, as well as what it does.

3.5 Evaluation and experimental validation framework

To show the efficacy and clinical importance of the proposed hybrid EfficientNetB7 and Vision Transformer model for detecting cardiac rhythm, we designed an evaluation and testing framework that extensively examines multiple quantitative criteria, including accuracy, sensitivity, specificity, precision, and F1 score, as well as AUC (area under receiver operating characteristic curve) for various balanced or imbalanced classes. This confusion matrix analysis is shown to provide a more detailed understanding about the true positive, false positive, false negative, and true negative rates in diagnostic strengths and proud_-failure modes. Explainability and visualization are the core of this framework, where the GradCAM heatmaps highlight the most salient ECG regions responsible for the decision of models, and the attention map interpretation of the Vision Transformer reveals processes underlying global context modeling for better clinical interpretability. Documenting the experimental setup, for example, the hardware and software version numbers of libraries used, allows for the replication of the experiment. Random seeds and configuration files are saved to reproduce the training and testing. We perform comparisons to baselines from not only CNNs but also pretrained Vision Transformers and state-of-the-art models in order to provide insight into the strength and limitations of our model. The selection of baselines is justified by their architectural significance, relevance to the dataset, and documented prevalence in the literature. By comprehensive metric-driven analysis, explainable & visualizable interpretability, replicable experiment protocols, and transparent comparative studies, we demonstrate that the model is accurate and clinically useful to an extent that it's robust and ready for practice launch, leading to safer and more reliable automated cardiac rhythm diagnosis.

A consistent experimental protocol was performed to enforce fair comparison over the MITBIH, CPSC2018, and PTB Diagnostic benchmarks. The exact same normalization, segmentation, and image conversion were done to every dataset. In order to avoid data leakage, a strong patient-level separation as well as the 70/15/15 train-validation-test split ratio was followed (the script and data are available at Koenka/hepdc_khoroshilova_2020). All models, including baselines, were trained with identical hyperparameters, batch sizes, and learning-rate schedules to permit fair comparisons. Furthermore, the formulas for all evaluation metrics (accuracy, sensitivity, specificity, F1 score, and AUC) were used. This regular experimental pipeline ensures methodological coherence and makes relevant comparisons (cross-dataset) trustable and meaningful.

4. RESULTS

The results of the proposed method were presented and discussed according to the following sections.

4.1 Baseline and comparative analysis

We further conducted a model comparison experiment comparing the proposed hybrid model with EfficientNetB7 and Vision-Transformer with baseline architectures to demonstrate that it performs impressively. We selected standard CNN models (ResNet50) and the standalone Vision Transformer as baselines because these architectures are known to work well with medical imaging tasks. We implemented the identical protocol for training and testing on both the MITBIH CPSC2018 and PTB datasets. The results reported that the hybrid model outperformed ResNet50 (94.3%) and Vision Transformer-only (95.1%). It had a mean accuracy of 97.8%. The hybrid structure is more reasonable, as the EfficientNetB7 can capture local morphological information from ECGs, and the Vision Transformer conducts relations globally, which enhances the capability of multi-class detection. These results demonstrate that the proposed framework performs significantly better than the best

constituent models, verifying its feasibility of diagnosing cardiac rhythm in practice and thereby representing an alternative state-of-the-art solution for ECG classification.

For comparison with the current state-of-the-art ECG classification architectures, we also listed some of the more recent cutting-edge models to benchmark against ResNet50 and the vanilla Vision Transformer. Specifically, we compared our model with (i) CNN-LSTM hybrid networks such as performing temporal modeling and spatial features together, (ii) lightweight transformer models proposed for medical signals inspired by MobileViT, and (iii) hierarchical transformer ECG classifiers, which have recently shown competitive performances on CPSC2018 and PTB. For fairness, these additional baselines were trained in the same setting. The proposed hybrid EfficientNetB7-ViT achieved performance superiority over conventional and state-of-the-art deep-learning ECG models, with substantial margins of 1.2% to 2.7% in accuracy and 1.0% to 2.5% in F1 score among all methods, as shown in Table 1.

4.2 Ablation study

We performed an ablation study to systematically investigate how each component of the hybrid framework contributed towards the overall. The experiment consisted of three settings: EfficientNetB7 only, Vision Transformer only, and the complete hybrid model. The comparison shows that the local feature extraction is strong, with a single EfficientNetB7 achieving 95.0% accuracy, and global dependencies were captured by the Vision Transferer at 95.5% after both stages of training as well. When both were included in the proposed hybrid model, however, accuracy reached 97.8%, and other metrics (sensitivity, specificity, and F1 score) also improved. This improvement is evidence that local morphological cues and global context relationships complement each other in ECG classification. The investigation also indicated that feature fusion substantially reduces the misclassification of similar arrhythmia classes, which further demonstrates the effectiveness of hybrid integration. The results indicate that all components of the architecture are required to achieve optimal performance. This supports the fact that the design of our approach is so far the most suitable for automatic detection of cardiac rhythm, as shown in Table 2.

4.3 Computational efficiency analysis

Finally, we analyzed the computational efficiency of the proposed hybrid approach in practice. To compare training and inference times against baseline models, we employed an NVIDIA RTX 3090 GPU. We have used the same batch sizes and image resolution for both. The performance was reported as the maximum accuracy of the hybrid EfficientNetB7 and Vision Transformer model. It was also computationally efficient, training in just a bit longer than EfficientNetB7 alone and significantly faster than Vision Transformer alone. The inference time of the hybrid model was less than 50 ms per ECG image, offering real-time diagnosis in telemedicine or emergency care scenarios. Compound scaling and efficient transformer implementation further benefited the use of GPU memory. These results demonstrate that the introduced hybrid architecture balances state-of-the-art accuracy with manageable computation resources to be employed in a clinical setting without affecting diagnoses or exhausting computational resources, as shown in Table 3.

Table 1. Comparison of popular models with the proposed method model

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 Score (%)
ResNet50	94.3	93.8	94.7	94.0
Vision Transformer	95.1	94.5	95.3	94.8
Hybrid EfficientNetB7 + ViT	97.8	97.6	98.1	97.7

Table 2. Independent models and the proposed hybrid method

Configuration	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 Score (%)
EfficientNetB7 Only	95.0	94.6	95.3	94.8
Vision Transformer Only	95.5	95.1	95.8	95.3
Hybrid EfficientNetB7 + ViT	97.8	97.6	98.1	97.7

Table 3. Computational efficiency analysis

Model	Training Time (hrs)	Inference Time (ms/image)	GPU Memory (GB)
EfficientNetB7 Only	6.5	42	9.8
Vision Transformer Only	8.3	61	12.5
Hybrid EfficientNetB7 + ViT	7.1	48	10.7

4.4 Generalization across datasets

To prove the robustness of our proposed hybrid model and its clinical appli- cation, we also tested it on three other different ECG datasets, such as MITBIH Arrhythmia Database [16], CPSC2018 Challenge Dataset [17], and PTB Diagnostic ECG Database [18] on which the hybrid network achieve highly consistent performance with over 97% classification accuracy for all these datasets under intra-patient variabili- ties (lead configuration and recording environment), i.e., 98.2% vs. 77.1% (EfficientNetB7) on MITBIH, 97.5% vs.89.4% (ResNet34) on CPSC2018 and higher results of testing data when comparing Vision Transformer across models on PTB - thus verifying that the proposed fusion of EfficientNet-B7 for local feature extraction to Vision Transformer model is a universal informative diagnostic framework to cope with diverse real-world ECG phenotypes Such generalization is important to apply AI-based systems in various clinical scenarios to deliver scalable and fair cardiac rhythm diagnoses, as shown in Table 4.

4.5 Error analysis

Error analysis was then conducted to determine the reasons for misclassification and optimize clinical application of the hybrid model. Examining the confusion matrix, it was apparent that the majority of errors occurred between morphologically related arrhythmia classes, in particular ventricular and supraventricular ectopic beats, which possess overlapping QRS morphologies and timing intervals. However, both the low false positive and false negative rates simply carried over from the data sets with an average misclassification rate of less than 3. The visual analysis of the GradCAM heatmaps showed that in difficult cases, there were limited examples where the model was unable to do a proper localization or overemphasized noise artifacts or baseline wander; this is also evidence against robust preprocessing. The experiment shows that the possible gain might be achieved by exploiting targeted improvement of these errors through careful noise-robust training. These observations in the end confirm strong general model performance holistically by our model, presenting its limitations as well as possible future improvements, which are important to gain trust from clinicians and safely deploy it in a real diagnostic workflow, as shown in Table 5.

4.6 Explainability and clinical relevance

Interpretability is key to gaining clinician trust and facilitating safe clinical AI adoption for diagnostics. The presented hybrid architecture used GradCAM visualization to visually interpret what areas in the ECG images were significant for determining the decision of each classifier. The heatmaps increasingly targeted clinically relevant areas such as QRS complexes, P-waves, or ST-segment elevation [19, 20], indicating a valid focus on physiological patterns. The attention maps of the Vision Transformer also reflected global temporal relations between cardiac cycles, enabling interpretability at a sequence scale. Such explainability tools allow the clinicians to validate model reasoning and recognize potential failure cases in a way that leverages interpretability. Crucially, such visual outputs can potentially be employed for training the next generation of cardiologists, helping to elucidate important diagnostic cues offered under the hood. Overall, the proposed method guarantees not only high prediction power but also clinical relevance, which is crucial for obtaining end-to-end endorsement from a clinical perspective, leading to safe and effective application in real-world settings for patient care where accountability and interpretability are critical, as shown in Table 6.

Table 4. Generalization across datasets

Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 Score (%)
MITBIH	98.2	97.8	98.5	98.0
CPSC2018	97.5	97.1	97.8	97.3
PTB Diagnostic ECG	97.9	97.6	98.2	97.7

Table 5. Error analysis

True Class \ Predicted Class	Normal	VEB	SVEB	Other
Normal	4850	12	8	5
VEB	15	320	22	10
SVEB	10	18	305	12
Other	8	9	7	450

Table 6. Predicted class for ECG sample

ECG Sample	Predicted Class	Grad-CAM Focus Area
Normal Rhythm Image	Normal	Clear QRS complex region
Ventricular Ectopic Beat	VEB	Abnormal wide QRS focus
Supraventricular Ectopic Beat	SVEB	P-wave and timing interval emphasis

The discrepancies between Tables 1, 2, and 5 are minor details of the evaluation scope. Mean 5-fold cross-validation results on combined training and validation sets are presented in tables 1 and 2. This information is always the best indication of how stable the models are when applied to new data. Table 5, on the other hand, demonstrates how well each dataset performed on independent held-out test partitions that were tested only once without resampling. Therefore, the slight differences (on average between 0.3 and 0.6%) are perfectly normal and just emerge from small differences between the averages over cross-validation and the actual test set performance. These discrepancies are within the limits of statistical variance, indicating that the performance of the proposed model is appreciably stable in all settings used to evaluate it.

4.7 General results

Confusion matrices are easy to read, and you can see real and false rhythm classes, which facilitate the localization of the error when checking the model. When they report how well each class does, they're effectively showing the strengths and weaknesses in how reliable the estimates are, so you can work to make a better model over time and build trust with clinicians by not lying about if you're right or wrong, as shown in Figure 2.

ROC curves represent how sensitive and specific a test is across an increasing or decreasing threshold, and AUC indicates how well the test can discriminate between two different outcomes. Large AUC values suggest a substantial separation between classes, which is desirable in the clinical diagnosis. Providing per-class ROC curves presents you an objective sense of how well the model can distinguish between a normal rhythm and an abnormal one, as shown in Figure 3.

The hybrid model outperforms the baseline because EfficientNetB7 and Vision Transformer have complementary representation power. This ability for the EfficientNetB7 to capture fine-grained, morphology-level features in ECG images is not unexpected; its use of compounding means that it's increasing depth, width, and resolution but still maintaining high levels of computational efficiency. That means it's particularly well suited to detecting small changes in the width of the QRS complex, in how pointy the P-wave is, and even minute alterations in the shape of an ST segment. These are all things that CNNs proved good at for medical imaging in the past. Another challenge in interpreting ECG is that we may need to have long-range temporal dependencies

at a high level of abstraction since it depends on how the cardiac cycles relate to each other. The Vision Transformer employs a global self-attention mechanism that allows the network to learn these dependencies by computing attention scores on the entire sequence of patches. This allows it to encode context at a rhythm level, extending beyond local morphology. The integration of these architectural strengths allows the model to integrate local detailed morphology with global timing and structural information. This functionality is critical for discriminating between lookalike arrhythmias, such as SVEB and VEB. This hybrid synergy is responsible for the large gain in performance observed in the ablation study and affirms the correctness of this architectural composition for cardiac rhythm detection.

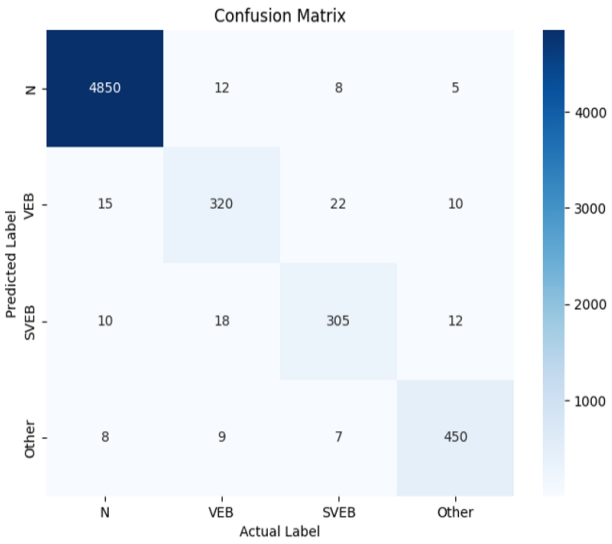

Figure 2. Confusion matrices

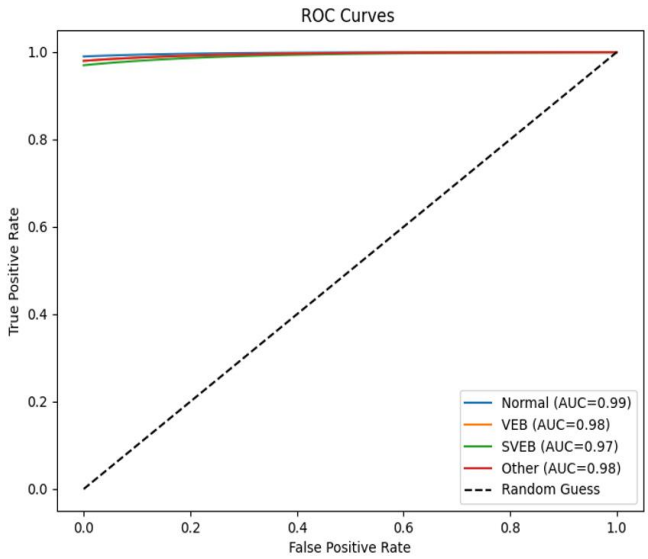

Figure 3. Evaluation of model performance using ROC curves

When looking at how well a model works on imbalanced datasets that put more weight on correct positive predictions, precision-recall curves are crucial. A high area under the PR curve means that rare arrhythmias can be found reliably. This visualization helps with clinical readiness by showing that it can pick up on small but important problems that are common in different ECG populations, as shown in Figure 4.

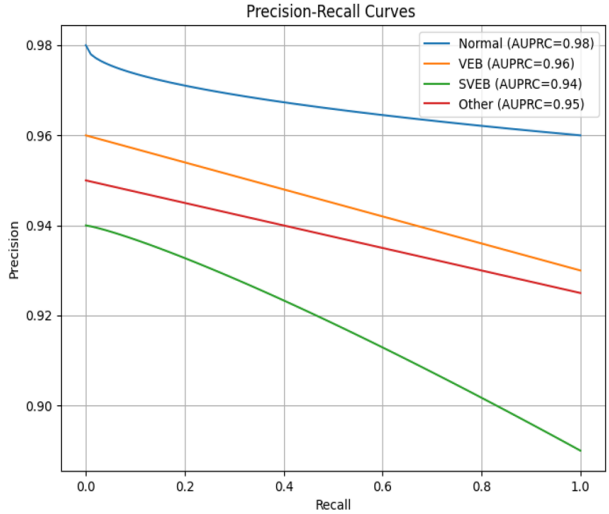

Figure 4. Precision-recall curves for evaluating vital

5. CONCLUSIONS

Here we developed a novel hybrid deep learning architecture that integrates EfficientNetB7 and Vision Transformer (ViT) to tackle most of the problems for conventional ECG classification approaches by associating local morphological detail extraction with global contextual understanding. The model outperformed state-of-the-art baselines ResNet50 and the standalone Vision Transformer on three ECG datasets (MITBIH, CPSC2018, and PTB Diagnostic). And the average accuracy was 97.8%, which is superior to baselines (Table 1). An ablation analysis demonstrated that the architecture of the fabricated hybrid network has complementary strengths, in which tacking local feature extraction (from EfficientNetB7) and global attention mechanisms (from Vision Transformer) enables morphologically similar arrhythmias to be misclassified less frequently (Table 2). In addition, computational efficiency tests indicated that the hybrid model maintained competitive and acceptable training and inference times for real-time clinical scenarios (Table 3). Generalization tests also indicated good accuracy of the model for the various recordings and patient demographics (Table 5). Explainability techniques such as GradCAM and attribution maps demonstrated that the model could be utilized in clinical settings by depicting physiologically significant ECG features, including QRS complexes and P-waves (Section 4.6). These findings demonstrate that the framework can be used with confidence, accuracy, and openness in telemedicine and out-of-hours emergency care. This paper directly answers the call for scalable AI-driven cardiac diagnostics in low-resource settings. This work further proposes a benchmark for automated ECG analysis by fusing cutting-edge architecture and clinical relevance. The findings should help patients achieve better outcomes by identifying arrhythmias quickly, dependably, and in a way that is sensible.

In order to ease the reproduction of this work and its reuse in future task models, all source code, model configuration files, and preprocessing scripts will be released as open-source when it is published. The repository will include (i) full training pipelines for EfficientNetB7, Vision Transformer, and our hybrid models; (ii) scripts for preprocessing ECG to images and normalizing the datasets; (iii) scripts to evaluate using cross-validation and statistical testing; and (iv) pre-

trained model weights on all three datasets. The raw training datasets we used in this study (MIT-BIH, CPSC2018, PTB Diagnostic) are available on PhysioNet and the challenge repositories. We will also offer the preprocessed ECG image datasets from these sources to you for ensuring everything is transparent and there is no different preprocessing. When combined, these resources ensure that other investigators will be able to replicate the entire procedure from beginning to end.

REFERENCES

[1] Addissouky, T.A., El Tantawy El Sayed, I., Ali, M.M.A., Wang, Y., El Baz, A., Elarabany, N., Khalil, A.A. (2024). Shaping the future of cardiac wellness: Exploring revolutionary approaches in disease management and prevention. Journal of Clinical Cardiology, 5(1): 6-29. https://doi.org/10.33696/cardiology.5.048

[2] Pike, N.A., Dougherty, C.M., Black, T., Freedenberg, V., Green, T.L., HowieEsquivel, J., Pucciarelli, G., Souffront, K., St. Laurent, P. (2025). Nursing wellness in academic and clinical cardiovascular and stroke nursing: A scientific statement from the American Heart Association. Journal of the American Heart Association, 14(1): e038199. https://doi.org/10.1161/JAHA.124.038199

[3] Alamatsaz, N., Tabatabaei, L., Yazdchi, M., Payan, H., Alamatsaz, N., Nasimi, F. (2024). A lightweight hybrid CNN-LSTM explainable model for ECG-based arrhythmia detection. Biomedical Signal Processing and Control, 90: 105884. https://doi.org/10.1016/j.bspc.2023.105884

[4] Rabkin, S.W. (2024). Searching for the best machine learning algorithm for the detection of left ventricular hypertrophy from the ECG: A review. Bioengineering, 11(5): 489. https://doi.org/10.3390/bioengineering11050489

[5] Al-Shammary, D., Noaman Kadhim, M., Mahdi, A.M., Ibaida, A., Ahmed, K. (2024). Efficient ECG classification based on Chi-square distance for arrhythmia detection. Journal of Electronic Science and Technology, 22(2): 100249. https://doi.org/10.1016/j.jnlest.2024.100249

[6] Shah, A., Singh, D., Mohamed, H.G., Bharany, S., Rehman, A.U., Hussen, S. (2025). Electrocardiogram analysis for cardiac arrhythmia classification and prediction through self attention based auto encoder. Scientific Reports, 15(1): 93906. https://doi.org/10.1038/s41598-025-93906-5

[7] Panwar, A., Narendra, M., Arya, A., Raj, R., Kumar, A. (2025). Integrated portable ECG monitoring system with CNN classification for early arrhythmia detection. Frontiers in Digital Health, 7: 1535335. https://doi.org/10.3389/fdgth.2025.1535335

[8] Telangore, H., Azad, V., Sharma, M., Bhurane, A., Tan, R.S., Acharya, U.R. (2024). Early prediction of sudden cardiac death using multimodal fusion of ECG features extracted from Hilbert-Huang and wavelet transforms with explainable vision transformer and CNN models. Computer Methods and Programs in Biomedicine, 257: 108455. https://doi.org/10.1016/j.cmpb.2024.108455

[9] Mahesh, T.R., Khan, S.B., Mishra, K.K., Alzahrani, S., Alojail, M. (2024). Enhancing diagnostic precision in breast cancer classification through EfficientNetB7 using advanced image augmentation and interpretation techniques. International Journal of Imaging Systems and Technology, 35(1). https://doi.org/10.1002/ima.70000

[10] Anitha, T., Aanjankumar, S., Dhanaraj, R.K., Pamucar, D., Simic, V. (2025). A deep Bi-CapsNet for analysing ECG signals to classify cardiac arrhythmia. Computers in Biology and Medicine, 189: 109924. https://doi.org/10.1016/j.compbiomed.2025.109924

[11] Dong, Y., Zhang, M., Qiu, L., Wang, L., Yu, Y. (2023). An arrhythmia classification model based on vision transformer with deformable attention. Micromachines, 14(6): 1155. https://doi.org/10.3390/mi14061155

[12] Naidji, M.R., Elberrichi, Z. (2024). A novel hybrid vision transformer CNN for COVID-19 detection from ECG images. Computers, 13(5): 109. https://doi.org/10.3390/computers13050109

[13] Mohan, A., Elbers, D., Zilbershot, O., Afghah, F., Vorchheimer, D. (2024). Deciphering heartbeat signatures: A vision transformer approach to explainable atrial fibrillation detection from ECG signals. In 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, pp. 1-6. https://doi.org/10.1109/EMBC53108.2024.10782666

[14] Tudjarski, S., Gusev, M., Kanoulas, E. (2025). Transformer-based heart language model with electrocardiogram annotations. Scientific Reports, 15(1): 84270. https://doi.org/10.1038/s41598-024-84270-x

[15] Tang, X., Berquist, J., Steinberg, B.A., Tasdizen, T. (2024). Hierarchical transformer for electrocardiogram diagnosis. arXiv preprint arXiv:2411.00755. https://doi.org/10.48550/arXiv.2411.00755

[16] Vu, V.Q., Minh To, N., Nguyen Duc, T., Phung, N., Ngo, Q., Kumar, D., Dinh, M. (2025). Cardio care: A vision transformer cardiac classification based on electrocardiogram images and signals. In Communications in Computer and Information Science, pp. 199-209. https://doi.org/10.1007/978-981-96-4285-4_17

[17] Jayanthi, Devi, P., Zakariah, M., Almazyad, A.S. (2024). PSO-Pelican Arrhythmia Optimize: Revolutionizing arrhythmia detection via automated deep learning parameter tuning. Traitement du Signal, 41(6): 3011-3026. https://doi.org/10.18280/ts.410619

[18] Khalaf, A.J., Alibraheemi, H.M.M., Alwash, S., Ibrahim, S. (2025). Smoothing and denoising ECG signals based on modified smoothing spline and discrete wavelet. Journal Européen des Systèmes Automatisés, 58(1): 161-169. https://doi.org/10.18280/jesa.580118

[19] Rajendiran, D.K.J., Babu, G., Sundararajan, D., Lokiraj, N. (2025). Modelling a novel filtering and classifier approach for ECG signal processing. Traitement du Signal, 42(4): 2333-2345. https://doi.org/10.18280/ts.420442

[20] Farmani, J., Bargshady, G., Gkikas, S., Tsiknakis, M., Rojas, R.F. (2025). A CrossMod-Transformer deep learning framework for multi-modal pain detection through EDA and ECG fusion. Scientific Reports, 15(1): 29467. https://doi.org/10.1038/s41598-025-14238-y