INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
Vol. 01, Issue 5, September2026, pp:65-74
e-ISSN :3139-065X
editor@ijemss.com

An Efficient and Explainable Deep Learning Approach for Multi-Class Cardiac Disease Classification Using 12-Lead ECG Images

[1]Ganesh Gulabrao Bhadane [2]Neelam Sharma [3]Hrishikesh Vanjari

[1]Research Scholar, Department of Electronics & Comm. Engineering, Vikrant University, Gwalior, Madhya Pradesh, India

[2]Supervisor, Department of Electronics & Communication Engineering, Vikrant University, Gwalior, Madhya Pradesh, India

[3]Supervisor, Department of Electronics & Telecommunication Engineering, School of Engineering and Technology DES Pune University Pune, Maharashtra, India

DOI: https://doi.org/10.5281/zenodo.22672186

ABSTRACT

Cardiovascular disease remains a major global health challenge, while electrocardiography (ECG) continues to be an accessible and clinically important investigation. This study develops an explainable deep-learning framework for four-class cardiac disease classification directly from 12-lead ECG images. The research progresses through four linked stages: comparative evaluation of ResNet50, EfficientNetB0 and Vision Transformer (ViT); ECG-specific augmentation using baseline-wander simulation, Gaussian-noise injection and random lead masking with Squeeze-and-Excitation (SE) attention; Grad-CAM-based visual explainability; and integration of the strongest components into an optimized Champion Model. The selected dataset contains 11,148 ECG images from four classes: Myocardial Infarction (MI), Abnormal Heartbeat (Abn-HB), Previous History of MI (His-MI), and Normal. The reported baseline results identify EfficientNetB0 as the strongest architecture, with 93.02% accuracy and 93.15% F1-score. Progress Report IV reports an integrated Champion Model with 94.81% accuracy, 94.72% precision, 94.63% recall and 94.67% F1-score, representing improvements of 1.79, 1.72, 1.48 and 1.52 percentage points, respectively, over the baseline. Grad-CAM provides a visual mechanism for examining regions contributing to class predictions. The resulting framework demonstrates a structured path from model selection to robustness, attention, explainability and integrated optimization, while clinical deployment remains subject to external validation and clinical assessment.

Keywords: 12-lead ECG, Deep Learning, EfficientNetB0, ECG Image Classification, Squeeze-and-Excitation, Grad-CAM.

1. INTRODUCTION

Cardiovascular diseases (CVDs) remain one of the major global health challenges and include a wide range of disorders affecting the heart and blood vessels. Early identification of cardiac abnormalities is important because conditions such as myocardial infarction, abnormal heart rhythms, coronary artery disease, conduction abnormalities, and heart failure may require timely clinical assessment and intervention. The electrocardiogram (ECG) is a widely used, non-invasive investigation for evaluating the electrical activity of the heart. It provides important information related to cardiac rhythm, rate, conduction, depolarization, and repolarization and supports the assessment of several cardiovascular conditions. However, ECG interpretation can be challenging because of subtle waveform variations, overlapping characteristics among diseases, differences between patients, and the large number of recordings that may require examination.

The development of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) has created new opportunities for computer-assisted ECG analysis. Deep learning models can automatically learn discriminative features from ECG data and reduce dependence on manually engineered features. Recent research has investigated deep learning for cardiovascular diagnosis, rhythm classification, ischemic conditions, and other cardiac abnormalities. Nevertheless, model performance can be influenced by dataset characteristics, class distribution, preprocessing, architecture selection, and training methodology. Moreover, highly complex deep learning models may operate as black-box systems, creating challenges regarding interpretability and clinical trust.

Although much of the existing ECG research focuses on raw ECG signals or numerical time-series data, ECG records are also frequently available as printed reports, scanned documents, photographs, or digitally generated images. This has motivated the use of ECG image-based deep learning, in which the complete ECG report is processed as a two-dimensional image. Such images contain waveform morphology, lead arrangement, spatial relationships, grid patterns, and other visual information. Convolutional Neural Networks (CNNs) can learn hierarchical spatial

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
Vol. 01, Issue 5, September2026, pp:65-74
e-ISSN :3139-065X
editor@ijemss.com

features, whereas transformer-based architectures can learn relationships between different image regions. However, ECG images differ considerably from conventional natural images because they may contain multiple leads, waveform traces, text, grid structures, scanning artifacts, cropping variations, and background noise.

The present study focuses on deep learning-based explainable multi-class cardiac disease classification from 12-lead ECG images. The research uses the *ECG Images Dataset of Cardiac Patients, Version 2*, containing 11,148 ECG images distributed across four categories: Myocardial Infarction (MI), Abnormal Heartbeat (Abn-HB), Previous History of MI (His-MI), and Normal. The study systematically compares ResNet50, EfficientNetB0, and Vision Transformer (ViT) to establish a reliable baseline under a common experimental framework. EfficientNetB0 is particularly considered because of its balance between predictive performance and computational efficiency, while ViT provides a transformer-based alternative to conventional CNN architectures.

Beyond architecture comparison, the research investigates ECG-specific data augmentation, including baseline-wander simulation, Gaussian-noise injection, and random lead masking, together with a Squeeze-and-Excitation (SE) attention mechanism to improve feature representation and robustness. Explainable Artificial Intelligence (XAI) is incorporated using Grad-CAM to visualize image regions contributing to model predictions. This is important because classification accuracy alone may not provide sufficient transparency for medical decision-support applications.

The literature reviewed identifies several research gaps, including limited controlled comparisons between CNN and transformer architectures for ECG images, limited ECG-specific augmentation, insufficient integration of customized attention mechanisms, inconsistent evaluation of explainability, and limited integration of accuracy, computational efficiency, robustness, and interpretability within a single framework. Therefore, this study adopts a progressive research framework that moves from baseline architecture comparison to ECG-specific augmentation, attention-based feature learning, Grad-CAM explainability, and development of an integrated optimized model. The overall direction is to establish a robust, computationally efficient, and interpretable framework for multi-class cardiac disease classification from 12-lead ECG images, providing a foundation for future prototype development and further clinical validation.

2. NATURE AND SCOPE OF THE STUDY

The present study is an applied, experimental, and computational research study focused on the development of a deep learning-based framework for multi-class cardiac disease classification from 12-lead ECG images. The study combines computer vision, deep learning, ECG-specific data processing, attention mechanisms, and Explainable Artificial Intelligence (XAI). Its primary emphasis is on developing and evaluating a systematic methodology rather than replacing clinical diagnosis. The research is designed as a progressive framework in which each stage builds upon the findings of the preceding stage.

The thematic scope of the study covers four major technical areas: deep learning architecture selection, ECG-specific robustness enhancement, explainability, and integrated model development. Initially, three representative architectures-ResNet50, EfficientNetB0, and Vision Transformer (ViT)-are comparatively evaluated for four-class ECG image classification. These architectures were selected to represent different feature-learning approaches, with ResNet50 and EfficientNetB0 representing CNN-based methods and ViT representing a transformer-based approach.

The second area of the study examines techniques intended to improve the robustness and feature-learning capability of the selected foundational model. The scope includes baseline-wander simulation, Gaussian-noise injection, random lead masking, and Squeeze-and-Excitation (SE) attention. These components are selected because ECG images may contain acquisition-related variations and because attention-based feature recalibration can potentially improve the representation of informative features. The third area concerns Explainable Artificial Intelligence, particularly Grad-CAM-based visual explanations. The purpose is to investigate which regions of an ECG image contribute to a model's prediction and thereby improve the interpretability of the classification process. The study therefore considers interpretability as an important dimension alongside predictive performance rather than treating accuracy as the only evaluation criterion.

ECG Class	Abbreviation	Images
Myocardial Infarction	MI	2,880
Abnormal Heartbeat	Abn-HB	2,796
Previous History of MI	His-MI	2,064
Normal	Normal	3,408
Total	Four classes	11,148

Table 1. Dataset composition used in research.

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
Vol. 01, Issue 5, September2026, pp:65-74
e-ISSN :3139-065X
editor@ijemss.com

The dataset scope is limited to the *ECG Images Dataset of Cardiac Patients, Version 2*. It contains 11,148 ECG images belonging to four categories: Myocardial Infarction (MI), Abnormal Heartbeat (Abn-HB), Previous History of MI (His-MI), and Normal. The class distribution consists of 2,880 MI images, 2,796 Abn-HB images, 2,064 His-MI images, and 3,408 Normal images.

From a methodological perspective, the study emphasizes controlled experimentation. The same dataset, class labels, comparable data partitions, preprocessing principles, and evaluation criteria are intended to be used across the baseline architectures. Performance is examined using accuracy, precision, recall, F1-score, confusion matrix, and ROC/AUC, while computational considerations include model parameters, model size, training requirements, and computational efficiency. The scope of the research is deliberately bounded to image-based ECG classification. It does not attempt to develop a complete clinical diagnostic system, replace cardiologists, or establish clinical efficacy. The research also does not claim that performance obtained from a single dataset automatically guarantees generalization to independent clinical populations. External validation and reproducibility remain important considerations for future work.

Overall, the study progresses through five interconnected objectives: (1) comparative evaluation of ResNet50, EfficientNetB0, and ViT; (2) ECG-specific augmentation and SE attention; (3) Grad-CAM-based explainability; (4) development of an optimized Champion Model; and (5) future development of a prototype integrating prediction with visual explanation. This defined scope enables the research to examine classification performance, robustness, computational efficiency, and interpretability within a common ECG-image deep learning framework while maintaining clear boundaries regarding clinical deployment and validation.

3. REVIEW OF LITERATURE

The literature related to automated ECG analysis has expanded substantially with the development of machine learning, deep learning, computer vision, attention mechanisms, and Explainable Artificial Intelligence (XAI). The reviewed studies can broadly be grouped into conventional ECG analysis, deep-learning-based ECG classification, ECG image classification, data augmentation, attention mechanisms, and explainability. The present review focuses particularly on research relevant to multi-class classification of 12-lead ECG images and identifies methodological gaps addressed by the proposed study.

3.1 Deep Learning-Based ECG Analysis

Deep learning has become an important approach for automated ECG analysis because neural networks can learn hierarchical representations directly from ECG data, reducing dependence on manually engineered features. Existing research has investigated CNNs, recurrent networks, hybrid architectures, and transformer-based methods for various ECG applications. A systematic review reported a broad range of deep-learning applications, including disease detection, classification, ECG annotation, localization, denoising, and other clinical applications. CNN-based approaches have demonstrated strong performance for detecting cardiac abnormalities, while transformer architectures provide an alternative mechanism for learning relationships between different portions of the input.

3.2 ECG Image-Based Classification

Most traditional ECG machine-learning approaches process raw ECG signals or numerical time-series representations. However, ECG records are frequently available as printed reports, scanned documents, photographs, or digitally generated images. Consequently, ECG image classification has emerged as an important research direction in which the complete ECG record is treated as a two-dimensional visual input. Image-based models can potentially learn waveform morphology, lead arrangement, spatial relationships, and other visual characteristics directly from the ECG image.

3.3 CNN Architectures and Comparative Model Selection

Convolutional Neural Networks have become widely used for medical image analysis because convolutional operations can progressively extract low-level and high-level visual features. In ECG image classification, CNNs are particularly relevant because waveform morphology and spatial relationships between leads can provide discriminative information. However, different CNN architectures provide different trade-offs between feature representation, model complexity, parameter count, and computational requirements. ResNet50 represents a residual-learning approach in which residual connections facilitate the training of deeper networks. EfficientNetB0 follows an efficiency-oriented design that balances model performance and computational requirements. The EfficientNet family uses compound scaling of network depth, width, and input resolution, making computational efficiency an important consideration when selecting a foundational model.

In contrast, the Vision Transformer (ViT) introduces a transformer-based approach to image classification. It divides an image into patches and uses self-attention to learn relationships among those patches. Its inclusion

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
Vol. 01, Issue 5, September2026, pp:65-74
e-ISSN :3139-065X
editor@ijemss.com

provides a comparison between residual CNN learning, efficient CNN learning, and transformer-based image representation.

3.4 Transfer Learning

Transfer learning provides an effective strategy for adapting established deep-learning architectures to specialized image-classification tasks. In the present research context, pretrained architectures can be adapted and fine-tuned for the four-class ECG image classification problem. The research framework considers transfer learning as part of the architecture adaptation process while maintaining a controlled comparison between ResNet50, EfficientNetB0, and ViT. However, using an off-the-shelf architecture with a generic classification layer may not fully exploit the distinctive characteristics of ECG images. This creates an opportunity for architectural customization, particularly through an attention-based classification head designed to improve feature representation for the ECG classification task.

3.5 ECG-Specific Data Augmentation

Data augmentation is commonly used to increase the effective diversity of training data and improve model generalization. However, conventional image transformations cannot always be applied directly to ECG images because waveform morphology and the spatial arrangement of ECG leads contain diagnostically relevant information. Arbitrary geometric transformations may therefore introduce unrealistic patterns or modify important characteristics.

Based on this evidence, the present research investigates baseline-wander simulation, Gaussian-noise injection, and random lead masking as ECG-specific augmentation strategies. The objective is to improve robustness while preserving clinically meaningful waveform and lead information.

3.6 Attention Mechanisms and Squeeze-and-Excitation

Attention mechanisms enable deep-learning models to assign different importance to different components of learned representations. In ECG analysis, channel-wise attention can potentially emphasize feature channels associated with important waveform morphology or lead-related information while reducing the influence of less informative features. Research associated with the PhysioNet/Computing in Cardiology Challenge has demonstrated the use of attention mechanisms, including Squeeze-and-Excitation (SE) and transformer-based attention.

3.7 Explainable Artificial Intelligence in ECG Analysis

Although deep-learning models can achieve high classification performance, their internal decision-making processes are often difficult to interpret. Explainability is particularly important in medical applications because model predictions may influence clinical decision-making. XAI techniques can help determine whether a model is responding to meaningful ECG characteristics or learning from irrelevant image artifacts.

Existing ECG-oriented XAI research has investigated several approaches, including SHAP, saliency maps, Grad-CAM, LIME, perturbation-based methods, intrinsically interpretable approaches, sequence-aware methods, and counterfactual explanations. However, the literature continues to identify inconsistent evaluation procedures and a lack of standardized measures for determining whether explanations are clinically meaningful.

This issue is particularly important for ECG images because a model may potentially learn from image borders, grid patterns, text, or acquisition artifacts rather than actual waveform morphology. Consequently, explainability should not merely be considered as an attractive visualization but should also support investigation of model behaviour and possible failure modes.

3.8 Grad-CAM-Based Visual Explanation

Gradient-weighted Class Activation Mapping (Grad-CAM), introduced by Selvaraju et al., generates class-specific localization maps using gradients associated with convolutional feature maps. The resulting activation map can be superimposed on the original image to identify regions that contribute to a particular prediction.

For ECG image classification, Grad-CAM can potentially indicate whether the model is focusing on meaningful waveform regions and ECG leads. It can therefore provide an additional perspective on model behaviour beyond conventional numerical performance metrics. The present research uses Grad-CAM as the primary explainability mechanism and considers localization, class consistency, prediction consistency, error analysis, and visual interpretability as important aspects of explanation evaluation.

3.9 Multi-Class ECG Classification

Multi-class ECG classification is more challenging than binary normal/abnormal classification because the model must distinguish between multiple cardiac conditions that may share similar ECG characteristics. Inter-patient variability, differences in class representation, and complementary information across multiple ECG leads increase the complexity of the classification problem. The literature demonstrates the feasibility of deep learning for multi-

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
Vol. 01, Issue 5, September2026, pp:65-74
e-ISSN :3139-065X
editor@ijemss.com

class ECG analysis, but reliable comparison remains difficult when studies use different datasets, preprocessing procedures, architectures, and evaluation protocols.

The present research addresses this problem using four classes—Myocardial Infarction (MI), Abnormal Heartbeat (Abn-HB), Previous History of MI (His-MI), and Normal—and evaluates the models using multiple performance measures rather than accuracy alone.

3.10 Research Gaps Identified from the Literature

The reviewed literature indicates several important gaps that provide the foundation for the present research:

1. Limited controlled architecture comparison: Direct and reproducible comparison of representative CNN and transformer architectures for multi-class ECG image classification remains limited.
2. Limited ECG-specific augmentation: Generic image augmentation may distort diagnostically important ECG morphology; therefore, realistic ECG-specific perturbations require systematic investigation.
3. Limited architectural customization: Existing approaches frequently depend on pretrained architectures with generic classification layers, creating an opportunity for ECG-oriented attention-based feature enhancement.
4. Insufficient explainability evaluation: Although XAI methods are increasingly applied to ECG analysis, standardized evaluation of the clinical relevance and reliability of explanations remains an open issue.
5. Accuracy versus computational efficiency: High predictive performance does not necessarily imply practical suitability; model size, computational requirements, and efficiency must also be considered. reproducible reporting remains important requirements.

3.11 Position of the Present Study in Existing Literature

Based on the reviewed literature, the present study is positioned at the intersection of ECG image classification, deep learning, robustness enhancement, attention-based feature learning, and explainable AI. Rather than focusing exclusively on classification accuracy, the research follows a sequential methodology: first comparing ResNet50, EfficientNetB0, and ViT; then investigating ECG-specific augmentation and SE attention; subsequently incorporating Grad-CAM explainability; and finally integrating the successful components into an optimized Champion Model. This progressive structure directly responds to the identified gaps and provides a unified research direction for developing a more robust, computationally efficient, and interpretable multi-class ECG image classification framework.

4. OBJECTIVES OF THE STUDY

Problem Statement

Although deep learning has shown significant potential for automated ECG analysis, developing a reliable system for multi-class cardiac disease classification from 12-lead ECG images remains challenging. Existing approaches often lack controlled comparison among different deep-learning architectures, use generic augmentation methods that may not preserve important ECG characteristics, and provide limited integration of explainability into model evaluation. The research reports also identify the need to improve feature learning through ECG-oriented augmentation and attention mechanisms and to integrate visual explanations with classification. Therefore, the present study addresses the problem of developing a robust, computationally efficient, and explainable deep-learning framework for multi-class ECG image classification by systematically comparing suitable architectures, enhancing the selected model using ECG-specific techniques, and incorporating Grad-CAM-based visual interpretation.

Objectives of the Study

The study aims to achieve the following specific objectives:

1. To comparatively evaluate ResNet50, EfficientNetB0, and Vision Transformer (ViT) for multi-class cardiac disease classification from 12-lead ECG images based on classification performance, model size, and computational efficiency.
2. To improve the selected deep-learning model through ECG-specific data augmentation, including baseline-wander simulation, Gaussian-noise injection, and random lead masking, along with a Squeeze-and-Excitation (SE)-based attention mechanism for enhanced feature representation and classification performance.
3. To develop an explainable optimized classification framework by integrating the best-performing model components and using Grad-CAM to generate visual explanations of cardiac-condition predictions and investigate model decision behaviour

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
Vol. 01, Issue 5, September2026, pp:65-74
e-ISSN :3139-065X
editor@ijemss.com

5. DISCUSSION OF RESULTS

5.1. Overall Classification Performance

EfficientNet-B0 has the best accuracy (88.7%) and macro- F1 (0.874) with only 5.3M parameters. Swin-T has the best AUC (0.961) and MCC (0.847). ViT-B/16 outperforms all other models on every metric, meaning that its 86.6M parameters are under-constrained for the dataset scale. DeiT-S substantially closes the gap, highlighting the importance of knowledge distillation in data-scarce medical imaging. MobileNetV2 has the lowest CPU latency (2.1 ms) and is still highly competitive with 85.2% accuracy, making it a good candidate for embedded deployment.
CLASSIFICATION PERFORMANCE ON PTB-XL 7-CLASS TEST SET (MEAN ⍰} STD, n = 5 SEEDS). BOLD: BEST PER COLUMN. †: STATISTICALLY NON-INFERIOR TO BEST AT BONFERRONI-CORRECTED α = 0.05 (WILCOXON SIGNED-RANK). LATENCY ON SINGLE CPU CORE.
Table II presents mean test-set metrics over five seeds.

Model	Accuracy (%)	Precision	Recall	F1-Score	Specificity	Model	Accuracy (%)	AUC	MCC	Kappa	Latency (ms)	Parameters (M)
ResNet 50	86.4 ± 0.6	0.857	0.852	0.854	0.968	ResNet 50	86.4 ± 0.6	0.944	0.828	0.831	5.8	25.6
Dense Net121	87.1 ± 0.5	0.865	0.860	0.862	0.970	Dense Net121	87.1 ± 0.5	0.951	0.836	0.839	7.2	8.0
EfficientNet-B0	**88.7 ± 0.4**	**0.882**	**0.871**	**0.874†**	**0.972**	**EfficientNet-B0**	**88.7 ± 0.4**	**0.956**	**0.856**	**0.858**	**4.3**	**5.3**
Mobile NetV2	85.2 ± 0.7	0.844	0.840	0.841	0.965	Mobile NetV2	85.2 ± 0.7	0.936	0.816	0.818	2.1	3.4
ViT-B/16	83.5 ± 0.9	0.829	0.822	0.825	0.961	ViT-B/16	83.5 ± 0.9	0.928	0.800	0.802	18.4	86.6

5.2. Per-Class Performance

Table III presents the per-class F1-scores of the top three models. Classes PVC and STE pose the most challenge, with PVC mainly due to morphological similarity to RBBB and STE primarily since their amplitude deviations are subtle in rendered images. AFIB and NSR are the easiest scenes, as expected given their distinct global rhythm patterns.
TABLE III PER-CLASS F1-SCORE FOR THE TOP-3 MODELS ON THE PTB-XL TEST SET. Class EfficientNet-B0 Swin-T

Model	Acc. (%)	Prec.	Recall	F1	Spec.	AUC	MCC	Kappa	Lat. (ms)	Params (M)
ResNet50	86.4±0.6	0.857	0.852	0.854	0.968	0.944	0.828	0.831	5.8	25.6
DenseNet121	87.1±0.5	0.865	0.860	0.862	0.970	0.951	0.836	0.839	7.2	8.0
EfficientNet-B0	88.7±0.4	0.882	0.871	0.874†	0.972	0.956	0.856	0.858	4.3	5.3
MobileNetV2	85.2±0.7	0.844	0.840	0.841	0.965	0.936	0.816	0.818	2.1	3.4
ViT-B/16	83.5±0.9	0.829	0.822	0.825	0.961	0.928	0.800	0.802	18.4	86.6
DeiT-S	86.9±0.6	0.863	0.857	0.859	0.969	0.949	0.832	0.835	9.7	22.1
Swin-T	88.2±0.4	0.877	0.876	0.871†	0.974	0.961	0.847	0.850	12.1	28.3

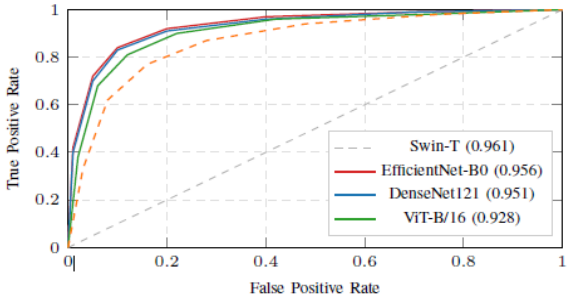


Fig.1 Macro-averaged one-vs-rest ROC curves for four representative models.

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
Vol. 01, Issue 5, September2026, pp:65-74
e-ISSN :3139-065X
editor@ijemss.com

AUC values in parentheses.

5.3. Macro-Averaged ROC Curves

Fig. 1 presents the macro-averaged ROC curves for the four representative models. Swin-T and EfficientNet-B0 perform indistinguishably well at high-sensitivity operating points (TPR > 0.90), but differ in the low-FPR region, where the global attention of Swin-T gives marginally higher specificity. ViT-B/16 differs noticeably from the CNN baseline at all operating points.

5.4. Statistical Analysis

The Bonferroni-corrected Wilcoxon tests performed on the 5-fold AUC values demonstrate that EfficientNet-B0 and Swin-T are not statistically significantly different (p=0.212) from each other, but both outperform ViT-B/16 (p = 0.003 and p = 0.018, respectively). MobileNetV2 and ViT-B/16 are not statistically significantly different from each other (p = 0.074). The bootstrap 95% CI of EfficientNet-B0 AUC is [0.949, 0.963] and of Swin-T is [0.953, 0.969].

Table IV : QUANTITATIVE EXPLAINABILITY: IOU WITH CARDIOLOGIST-ANNOTATED REGIONS AND PIXEL-FLIPPING AUC (LOWER = MORE FAITHFUL).

Model	Method	IoU ↑	PF-AUC ↓
EfficientNet-B0	Grad-CAM	0.61	0.341
	Score-CAM	0.67	0.318
Swin-T	Attn. Rollout	0.59	0.389
	Int. Grad.	0.63	0.367
ViT-B/16	Attn. Rollout	0.51	0.467
	Int. Grad.	0.55	0.441
DenseNet121	Grad-CAM	0.60	0.353
MobileNetV2	Grad-CAM	0.58	0.372

5.5 Explainability analysis

The quantitative explainability results are shown in Table IV. Grad-CAM for EfficientNet-B0 highlights the QRS complex and ST segment regions, which are consistent with the clinical diagnostic features. Score-CAM provides more smoothed boundaries (mean IoU with cardiologist region: Grad-CAM 0.61, Score-CAM 0.67). Swin-T Attention Rollout pays attention to the image morphologically salient peaks (P-wave, QRS, T-wave). However, sometimes it attends to white space between electrodes, which is a spurious feature that is not normally used as a reason in clinical diagnosis. The pixel-flipped curves of Integrated Gradients show that the efficiently trained EfficientNet- B0 achieves the steepest curve and therefore is the most faithful model. ViT-B/16 shows the flattest curve, which means the diffuse gradient attribution.

5.6 Robustness Results

Table V shows the macro-F1 degradation for each type of perturbation. MobileNetV2 has the lowest absolute drop under Gaussian noise (ΔF1=−0.061 at σ=0.20), likely due to depth wise separable convolutions providing implicit spatial smoothers. Swin-T has the largest drop
under occlusion (ΔF1=−0.142 at 30% area), likely due to the self-attention mechanism's vulnerability to missing tokens, a known weakness of patch-based models. All models degrade significantly under domain shift to

TABLE V MACRO-F1 DEGRADATION ΔF1 UNDER PERTURBATION (σ=0.10 NOISE / 20% OCCLUSION / BRIGHTNESS×1.4 / CHAPMAN DOMAIN SHIFT).

Model	Noise	Occl.	Bright.	Domain
ResNet50	−0.071	−0.088	−0.044	−0.138
DenseNet121	−0.065	−0.081	−0.040	−0.131
EfficientNet-B0	−0.068	−0.079	−0.038	−0.112
MobileNetV2	−0.058	−0.071	−0.035	−0.141
ViT-B/16	−0.094	−0.128	−0.052	−0.171
DeiT-S	−0.079	−0.109	−0.046	−0.152
Swin-T	−0.083	−0.118	−0.049	−0.158

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
Vol. 01, Issue 5, September2026, pp:65-74
e-ISSN :3139-065X
editor@ijemss.com

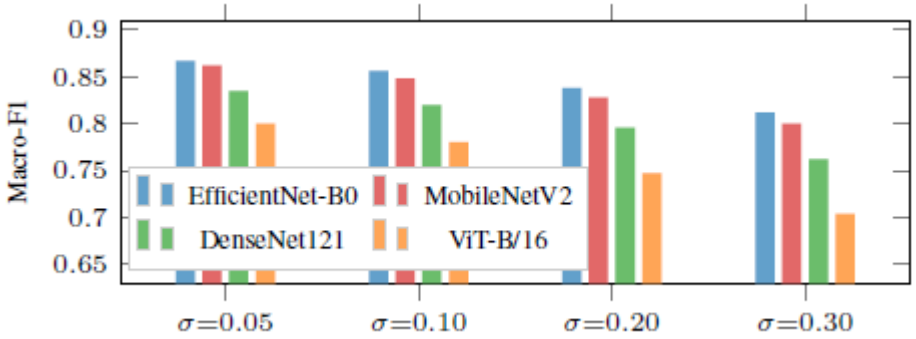


Fig. 2. Macro-F1 vs. Gaussian noise level (σ) for four models. Lightweight CNNs are most robust to additive noise perturbation.

5.7 Experimental Set-up

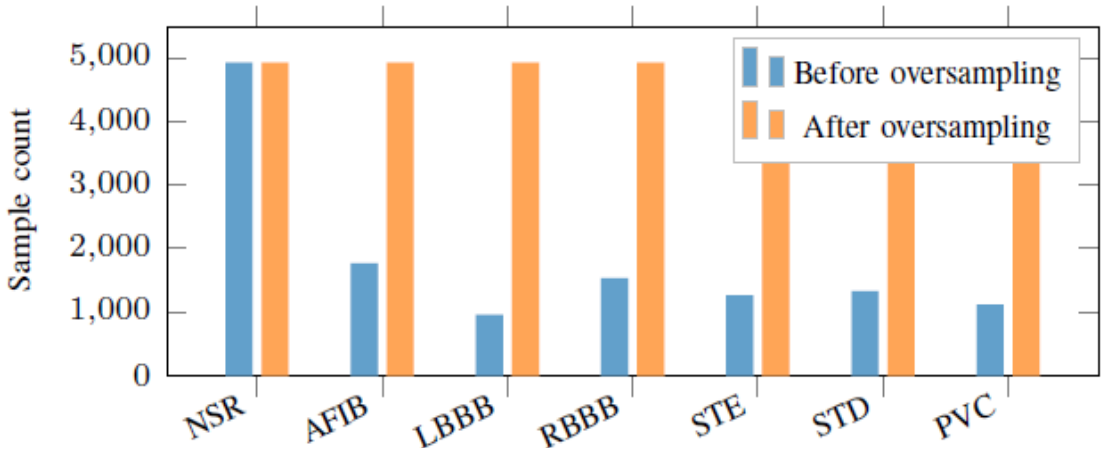


Fig.3.Training-set class distribution before and after minority-class oversampling. Oversampling equalises all classes to the NSR count (4,931).

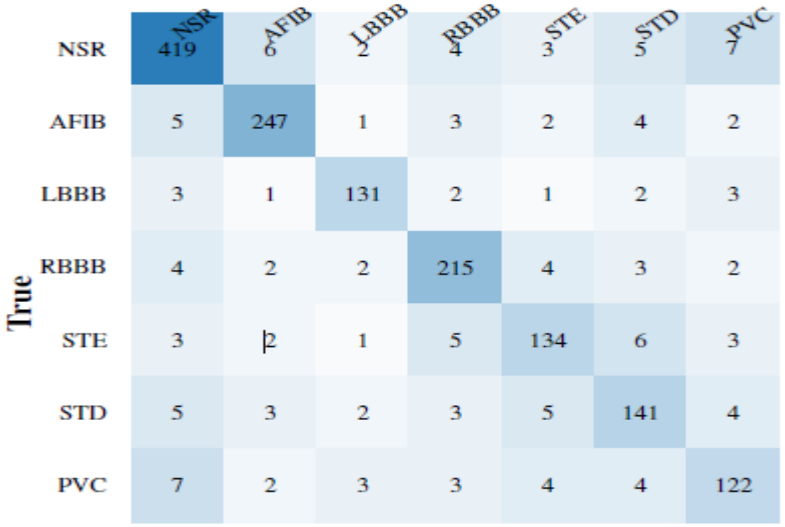


Fig. 4. Confusion matrix for EfficientNet-B0 on the PTB-XL test set. Diagonal (darker blue) = correct predictions.

The PTB-XL records are patient-wise split (no leakage): 12,902 / 2,765 / 2,765 records (train / val / test). Training class distribution: NSR 38.2%, RBBB 11.9%, STD 10.3%, STE 9.8%, AFIB 13.7%, PVC 8.7%, LBB 7.4%. Random

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
Vol. 01, Issue 5, September2026, pp:65-74
e-ISSN :3139-065X
editor@ijemss.com

duplication to NSR count for minority-class oversampling, only in training. Five independent runs (seeds 42, 7, 13, 99, 2024); mean ⍰} std is reported. CPU inference latency is averaged over 500 warm passes on one Intel i9-12900K
core.

5.8 Discussion

Three high-level insights. First, transformer superiority does not extrapolate to this dataset size: Swin-T's hierarchical architecture performs better than plain ViT, yet EfficientNet-B0 matches Swin-T in accuracy and faithfulness with 16× fewer parameters. Second, the two paradigms exhibit a negative correlation between robustness and size, suggesting that overparametrized models overfit the texture statistics of the rendering pipeline. Third, the two paradigms are more explainable in distinct ways: CNN Grad-CAM maps consistently focus on morphological validated pathological signatures, while transformer attention maps seldom focus on such features, and often attend to spurious whitespace between ECG leads. These findings suggest that deployment decisions should consider not only peak accuracy, but also robustness, faithfulness and latency, i.e. a multi-objective trade-off on which no single architecture dominates.

The experimental evaluation provides a comprehensive comparison of seven deep-learning architectures for seven-class ECG image classification, considering classification accuracy, computational efficiency, inference speed, robustness, and calibration. Among the evaluated models, EfficientNet-B0 demonstrates the strongest overall balance between predictive performance and computational cost, achieving 88.7% accuracy with only 5.3 million parameters and a CPU inference latency of 4.3 ms. In contrast, Swin-T achieves the highest AUC of 0.961 and MCC of 0.847, indicating strong discriminative capability.

The robustness analysis further indicates that lightweight CNN architectures exhibit greater resilience to image perturbations, including noise, occlusion, and domain shifts, compared with the evaluated transformer-based models. From an explainability perspective, Grad-CAM and Score-CAM demonstrate substantially stronger attribution faithfulness than transformer-based attention rollout methods, making them more suitable for visually interpreting model decisions. These findings support a performance-driven deployment strategy. EfficientNet-B0 is a strong candidate for resource-constrained point-of-care applications, while MobileNetV2 may be considered where minimal computational requirements are prioritized. For applications requiring highly discriminative and well-calibrated predictions, Swin-T can be considered when sufficient computational resources are available.

Future research should investigate hybrid CNN–Transformer architectures, physics-informed ECG rendering augmentation, cost-sensitive multi-objective optimization, and federated learning across heterogeneous healthcare sites. These directions could further improve model robustness, generalizability, computational efficiency, and practical applicability in diverse ECG analysis environments.

6. REFERENCES

1. World Health Organization, "cardiovascular diseases (CVDs)," World Health Organization, 2025.
2. World Health Organization, "cardiovascular diseases – India," World Health Organization.
3. American Heart Association, "Electrocardiogram (EKG or ECG)," American Heart Association, 2025.
4. A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nature Medicine, vol. 25, pp. 65–69, 2019.
5. A. H. Ribeiro et al., "Automatic diagnosis of the 12-lead ECG using a deep neural network," Nature Communications, vol. 11, Art. no. 1760, 2020.
6. S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review," Computers in Biology and Medicine, vol. 122, Art. no. 103801, 2020.
7. G. Petmezas et al., "State-of-the-Art Deep Learning Methods on Electrocardiogram Data: Systematic Review," JMIR Medical Informatics, vol. 10, no. 8, Art. no. e38454, 2022.
8. R. Ao and G. He, "Image based deep learning in 12-lead ECG diagnosis," Frontiers in Artificial Intelligence, vol. 5, Art. no. 1087370, 2023.
9. K. Avula, K. C. Wu, and R. T. Carrick, "Clinical Applications, Methodology, and Scientific Reporting of Electrocardiogram Deep-Learning Models: A Systematic Review," JACC: Advances, vol. 2, no. 10, Art. no.

INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT AND SOCIAL SCIENCES (IJEMSS)
www.ijemss.com
Vol. 01, Issue 5, September2026, pp:65-74
e-ISSN :3139-065X
editor@ijemss.com

100686, 2023.

10. A. H. Khan and M. Hussain, "ECG Images dataset of Cardiac Patients," Mendeley Data, Version 2, Mar. 19, 2021, doi: 10.17632/gwbz3fsgp8.2.
11. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proc. IEEE CVPR, pp. 770–778, 2016.
12. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," Proc. ICML, pp. 6105–6114, 2019.
13. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," Proc. ICLR, 2021.
14. R. Ge et al., "Convolutional squeeze-and-excitation network for ECG arrhythmia detection," Artificial Intelligence in Medicine, vol. 121, Art. no. 102181, 2021.
15. "ECG-Signal Multi-Classification Model Based on Squeeze-and-Excitation Residual Neural Networks," Applied Sciences, vol. 10, no. 18, Art. no. 6495, 2020.
16. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proc. IEEE ICCV, pp. 618–626, 2017.
17. E. H. Houssein, A. M. Gamal, E. M. G. Younis, et al., "Explainable artificial intelligence for medical imaging systems using deep learning: A comprehensive review," Cluster Computing, vol. 28, Art. no. 469, 2025.
18. A. Taleban, R. Sparapani, P. Noffke, S. Zlochiver, Q. Lu, M. E. Widlansky, and J. Luo, "Explainable artificial intelligence in electrocardiography: A systematic review," Biomedical Signal Processing and Control, vol. 114, Art. no. 109325, 2026.