RESEARCH ARTICLE

An Explainability-Validated Deep Learning Pipeline for 12-Lead ECG Image Classification: A ResNet-50 Case Study with Clinical Lead-Alignment Analysis on ECG Image Dataset

Ganesh Gulabrao Bhadane[1,2*], Rakesh Kumar Yadav[3], Hrishikesh Vanjari[4], Nandakishor C Patil[5], Shriganesh Mane[6], Vijay B Nerkar[7]

[1]Department of Electronics & Communication Engineering, Vikrant University, Gwalior,Madya Pradesh, India*
[2]Department of Electronics & Telecommunication Engineering, Dr.D. Y. Patil Institute of Technology, Pimpri, Pune, India, corresponding author: ganeshbhadane88@gmail.com, orcid id: 0009-0008-6225-9874
[3]Electronics & Communications Engineering, Vikrant University Gwalior, MP, India
[4]School of Engineering and Technology, DES Pune University, Pune, India
[5]Department of Electronics and Telecommunication Engineering, D N Patel College of Engineering, Shahada, Maharashtra, India,
[6]Department of AI&DS, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India,
[7]Electronics & Telecommunications Engineering.D.N.Patel College of Engineering Shahada, Maharashtra, India

Received: 25th June 2026; **Revised:** 28th July 2026; **Accepted:** 10th October 2026; **Available Online:** 15th October 2026

Abstract

Automated interpretation of 12-lead electrocardiogram (ECG) images with deep learning has repeatedly been reported to approach cardiologist-level performance, yet a persistent gap in the literature is that post-hoc explainability is typically presented as illustrative saliency imagery rather than as a mechanism that is systematically checked against the ECG's known anatomical structure. This paper reports an empirical, reproducible pipeline built specifically to close that gap. A ResNet-50 convolutional network, initialised with ImageNet weights and fine-tuned using a two-phase (frozen-head, then discriminative-learning-rate) transfer-learning schedule, was trained and evaluated on the four-class Mendeley ECG Images dataset (929 patients; Normal, Myocardial Infarction, Abnormal Heartbeat, History of Myocardial Infarction). On a held-out, stratified test set (n = 140) the model attained 49.3% accuracy and a macro-averaged F1 of 0.44 (95% bootstrap confidence interval: 0.367-0.508), with one-vs-rest per-class AUC ranging from 0.83 to 0.97 - evidence that the network's continuous class scores separate the four conditions considerably better than its arg-max decisions realise, a pattern consistent with a classifier that has learned informative representations but suffers a threshold and class-collapse problem compounded by measurable overfitting (training F1 0.81 versus validation F1 0.43). A five-method explainability battery - Gradient-weighted Class Activation Mapping (Grad-CAM), Grad-CAM++, a SHAP gradient explainer, occlusion sensitivity, and a purpose-built clinical lead-alignment analysis - was implemented and applied to correctly classified test examples. Grad-CAM and Grad-CAM++ saliency was observed to concentrate on the drawn ECG trace rather than on background grid lines or report margins, whereas the SHAP gradient explainer produced degenerate, spatially uninformative maps under the memory-constrained consumer-grade hardware used in this study. These results are reported transparently as a partial, in-progress benchmark: two further planned backbones (EfficientNet-B3 and a Vision Transformer, ViT-B/16) had not completed a training run at the time of writing, and the quantitative lead-alignment scoring is presented as a validated methodology rather than a completed numerical result, because the specific Mendeley image variant used is a single-lead (Lead II) rhythm-strip report rather than the full 3x4 twelve-lead grid the alignment method was designed to score. The paper documents the pipeline architecture, experimental protocol, diagnosed causes of the observed accuracy ceiling, and a concrete completion roadmap, positioning the work as a reproducible foundation for clinically grounded explainability research in deep-learning-based cardiac diagnostics.

Keywords: Explainable AI; Grad-CAM; SHAP; Electrocardiogram Image Classification; Myocardial Infarction Detection; Transfer Learning; ResNet-50; Clinical Lead Alignment; Deep Learning in Cardiology.

How to cite this article: Bhadane GG, Yadav RK, Vanjari H, Patil NC, Mane S, Nerkar VB., An Explainability-Validated Deep Learning Pipeline for 12-Lead ECG Image Classification: A ResNet-50 Case Study with Clinical Lead-Alignment Analysis on ECG Image Dataset. Int J Drug Deliv Technol. 2026;16(89s): 599-610; DOI: 10.25258/ijddt.16.89s.63

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, accounting for approximately 17.9 million deaths annually and nearly one-third of all global mortality [1]. The twelve-lead electrocardiogram (ECG) is the most widely used non-invasive cardiac diagnostic instrument, offering a multi-axis electrical map of the heart that clinicians use to identify ischaemia, conduction disease, arrhythmia, and structural pathology. However, accurate ECG interpretation requires years of clinical training, is subject to well-documented inter-observer variability, and is a scarce specialist resource in primary care, emergency, and low-resource settings. These pressures motivate automated, deep-learning-based ECG interpretation as a scalable complement to human expertise.

A large and rapidly growing body of work has shown that convolutional neural networks (CNNs), trained either on raw one-dimensional ECG signals [2], [3], [4] or on the two-dimensional printed/scanned images that constitute the majority of ECG records in real clinical archives [10], [11], can match or approach cardiologist-level diagnostic accuracy. The image-based paradigm is of particular practical interest because it allows direct reuse of ImageNet-pretrained computer-vision backbones and can be deployed without access to proprietary raw-signal acquisition interfaces. Transfer learning from natural-image pretraining has been shown repeatedly to outperform training from random initialisation on the comparatively small, expensive-to-annotate ECG datasets that are typically available [16].

Despite these advances, the deployment of deep learning in a diagnostic setting is inseparable from the question of interpretability. A cardiologist can point to ST-segment elevation in specific inferior leads and reciprocal changes in specific lateral leads as the basis for a diagnosis; a deep network's decision must be similarly auditable if it is to be trusted, regulated, or safely integrated into clinical workflows. Gradient-weighted Class Activation Mapping (Grad-CAM) [5] and Shapley-value-based attribution (SHAP) [6] are the dominant post-hoc explanation tools used in the medical imaging literature, and have been applied to ECG classifiers with some success [4]. However, a companion literature review conducted alongside this study - surveying the deep-learning-for-ECG field from traditional handcrafted-feature methods through one-dimensional signal networks to modern CNN, Vision Transformer (ViT) [7], and hybrid architectures - identifies explainability integration and validation as one of its most consequential open gaps: explanations are overwhelmingly presented as qualitative heat-map illustrations rather than being checked, systematically and quantitatively, against the specific anatomical leads and waveform regions that are clinically diagnostic for a given condition.

This paper is the empirical companion to that review. Rather than presenting a single polished accuracy figure, it documents a complete, modular, and reproducible pipeline built to (i) train and evaluate CNN and transformer backbones for four-class ECG image classification, and (ii) subject the resulting models to a battery of attribution methods that includes a purpose-built clinical lead-alignment check. The specific contributions of this paper are: An open, five-stage modular pipeline (data preparation, model factory, two-phase transfer-learning trainer, evaluator, and explainability suite) for ECG image classification, designed from the outset around explainability rather than treating it as an afterthought.

A rigorously evaluated ResNet-50 baseline on the four-class Mendeley ECG Images dataset, including bootstrapped confidence intervals, per-class one-vs-rest ROC-AUC, and a full confusion and misclassification analysis.

Implementation and qualitative evaluation of a five-method saliency battery - Grad-CAM, Grad-CAM++, a SHAP gradient explainer, occlusion sensitivity, and a clinical lead-alignment analysis that segments an ECG report image into its anatomical lead regions - together with an honest account of which of these methods produced usable output under the computational constraints of this study.

A transparent diagnosis of the specific, fixable causes of the observed accuracy ceiling (aggressive input down-scaling, class-weighting interacting with a small training set, and consumer-grade hardware memory constraints), together with a concrete completion roadmap for the two backbones (EfficientNet-B3, ViT-B/16) that had not yet completed training at the time of writing.

Section 2 reviews related work in deep-learning-based ECG classification and explainable AI. Section 3 describes the system architecture and methodology. Section 4 describes the dataset and preprocessing pipeline. Section 5 details the experimental setup. Section 6 reports quantitative and qualitative results. Section 7 discusses the findings. Section 8 states limitations. Section 9 proposes future work, and Section 10 concludes.

2. Related Work

2.1 From Handcrafted Features to End-to-End Deep Learning

Early automated ECG analysis relied on pipelines combining signal conditioning (baseline-wander removal, notch filtering, R-peak detection via the Pan-Tompkins algorithm) with handcrafted temporal, amplitude, and frequency-domain features fed into classical classifiers such as support vector machines. These approaches achieved good accuracy on constrained benchmarks such as MIT-BIH but were fundamentally bounded by the completeness of the manually designed feature set and were sensitive to distributional shift across devices and populations. The transition to end-to-end deep learning was driven by the observation that CNNs and recurrent networks could learn diagnostically relevant representations directly from raw signal or image data, without requiring an explicit feature-engineering step.

2.2 One-Dimensional and Image-Based ECG Deep Learning

Rajpurkar et al. [2] demonstrated that a 34-layer CNN trained on over 64,000 single-lead ambulatory recordings could match a panel of board-certified cardiologists on twelve arrhythmia types, and Hannun et al. [3] extended this to ten rhythm categories at larger scale. Ribeiro et al. [4] trained a residual network on 2.3 million 12-lead recordings and additionally used Grad-CAM to provide saliency explanations, reporting preliminary qualitative alignment between the model's attention and clinically established diagnostic features. In parallel, image-based approaches emerged to address the practical reality that most ECGs in clinical archives exist as printed or scanned images rather than raw digital signals: Wieclawek et al. [10] applied a fine-tuned VGG-16 to binary myocardial-infarction detection from ECG images (92.5% accuracy), and Al-Zaiti et al. [11] applied CNN-based image analysis to acute coronary occlusion triage in the emergency department. Vision Transformers have also been explored for ECG classification: Choi et al. [12] applied ViT to atrial fibrillation detection from ECG spectrograms, reporting competitive performance alongside interpretable attention maps, though transformer architectures remain constrained in this domain by their comparatively large labelled-data requirements [7].

2.3 Transfer Learning and Backbone Selection

Transfer learning from ImageNet-pretrained backbones is now the dominant strategy for image-based ECG classification, consistently outperforming training from random initialisation on datasets of typical clinical scale [16]. VGGNet [14] established that network depth alone is a strong driver of feature quality but is parameter-inefficient; ResNet [8] introduced residual connections that permit stable training of substantially deeper networks and has become the most widely adopted transfer-learning backbone in medical imaging; and EfficientNet [9] introduced principled compound scaling of depth, width, and resolution, achieving strong accuracy with markedly fewer parameters. This paper adopts ResNet-50 as its first completed backbone precisely because of this well-established reliability, while EfficientNet-B3 is planned as a parameter-efficient comparison point.

2.4 Explainable AI in ECG Diagnosis

Grad-CAM [5] remains the dominant post-hoc explanation technique for CNN-based medical image classifiers, generating spatial saliency maps from the gradients of the predicted class score with respect to the final convolutional feature maps. Grad-CAM++ refines this by using a weighted combination of positive partial derivatives to improve localisation of multiple instances of a class within an image. SHAP [6] provides a complementary, game-theoretic attribution framework based on Shapley values. Despite the availability of these tools, the ECG deep-learning literature has been criticised for validating explanations only qualitatively, without systematic quantitative comparison against the specific leads and waveform regions that are clinically diagnostic for a given condition, and without faithfulness checks (such as deletion/insertion curves) that verify an explanation actually reflects the pixels driving the prediction rather than merely looking plausible. This is the specific research gap this paper's pipeline is designed to address.

3. Methodology and System Architecture

The pipeline is organised as a modular Python package (src/ecg_xai/) with five functional stages, driven either individually via numbered CLI scripts (00-04) or end-to-end via a single driver script that is designed to continue past a per-model failure rather than aborting the entire run. Table 1 summarises the module inventory and its completion state at the time of writing.

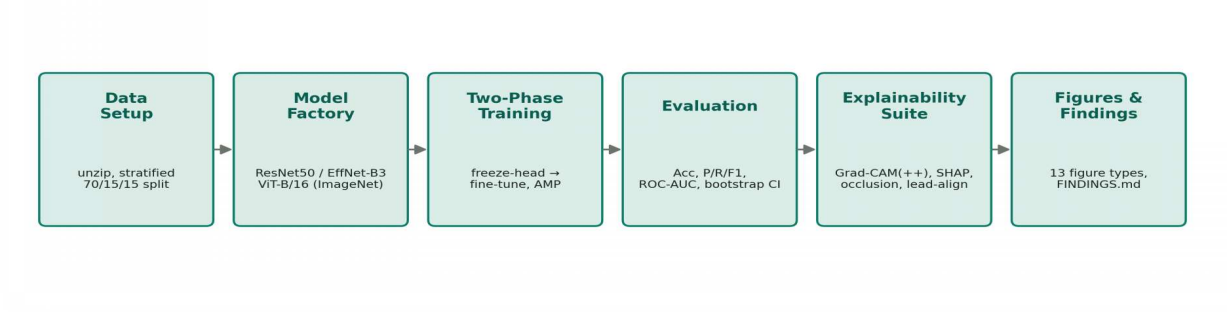


Figure 1. High-level stage architecture of the ECG-XAI pipeline: data preparation, model factory, two-phase transfer-learning training, quantitative evaluation, and the explainability suite, culminating in an automatically generated figures and findings report.

Table 1. Pipeline Module Inventory and Completion State

Module	Responsibility	State
config.py	Typed experiment configuration; YAML load/merge; project paths.	Complete
data.py	Dataset extraction; stratified 70/15/15 split with patient-level leakage checks; letterbox transform; ECG-safe augmentation; data loaders.	Complete
models.py	Model factory - ResNet-50, EfficientNet-B3, ViT-B/16 (ImageNet-pretrained via timm) and a custom CNN; declares Grad-CAM target layers.	Complete
train.py	Two-phase transfer learning (frozen head, then fine-tune); automatic mixed precision; class-weighted loss; cosine LR schedule; early stopping.	Complete (ResNet-50 run)
evaluate.py	Accuracy, per-class precision/recall/F1, confusion matrix, one-vs-rest ROC/AUC, 1000-resample bootstrap 95% CIs.	Complete (ResNet-50 run)
explain/	Grad-CAM, Grad-CAM++, ViT attention rollout, SHAP, occlusion sensitivity, lead-alignment scoring, deletion/insertion faithfulness.	Implemented; partial run
viz.py / figures.py	Thirteen figure types and an automatically generated FINDINGS.md summary report.	Implemented; partial run

3.1 Model Architecture

The completed backbone reported in this paper is ResNet-50 [8],a 50-layer residual convolutional network (approximately 25.6 million parameters) initialised from ImageNet weights, with its final fully connected layer replaced by a four-way linear classification head matching the Normal / Myocardial Infarction (MI) / Abnormal Heartbeat / History-of-MI label set. The final residual block (layer4) is registered as the Grad-CAM and Grad-CAM++ target layer, providing the spatial resolution needed for lead-level attribution. The model factory additionally implements EfficientNet-B3 [9], ViT-B/16 [7] (via the timm library), and a lightweight custom CNN baseline, enabling like-for-like comparison once all training runs are complete; only the ResNet-50 configuration had completed a full train-evaluate-explain cycle at the time of writing.

3.2 Two-Phase Transfer-Learning Protocol

Training follows a two-phase discriminative fine-tuning protocol. In Phase 1 (head warm-up), the ImageNet-pretrained backbone is frozen and only the newly initialised classification head is trained, allowing the head to adapt to the four-class ECG task without corrupting the pretrained feature hierarchy under a small dataset. In Phase 2 (fine-tuning), the backbone is unfrozen and trained jointly with the head using discriminative learning rates - a lower learning rate for the backbone and a higher learning rate for the head - together with a cosine-annealed learning-rate schedule. Automatic mixed precision (AMP) is used throughout to reduce memory footprint on the 6 GB training GPU, and training is stopped early on a plateau in validation macro-F1, with the best checkpoint by that criterion retained for evaluation.

To address the class imbalance inherent in the dataset (284 / 240 / 233 / 172 images across the four classes), the cross-entropy loss is class-weighted, where the loss for a sample with true class y is

$$L = - w_y \cdot log(softmax(z)_y)$$

and w_y is set inversely proportional to the frequency of class y in the training split. As discussed in Section 7, this weighting is implicated - alongside overfitting and input-resolution loss - in the specific failure mode observed for the Normal class.

3.3 Evaluation Protocol

Model evaluation computes overall accuracy; per-class precision, recall, and F1; a full four-by-four confusion matrix in both raw-count and row-normalised form; one-vs-rest ROC curves and AUC per class together with micro- and macro-averaged AUC; and 1000-resample bootstrap 95% confidence intervals for both overall accuracy and macro-F1, providing an estimate of the sampling uncertainty attributable to the modest test-set size (n = 140).

3.4 Explainability Suite

Five attribution methods are implemented, each reduced to a directly comparable spatial saliency map on correctly classified test examples: Grad-CAM and Grad-CAM++ for the CNN backbones; attention rollout for the ViT backbone; a SHAP gradient explainer; and model-agnostic occlusion sensitivity, used as a cross-check that does not depend on network internals. Beyond generating maps, three analyses

are implemented that go beyond the qualitative presentation typical of the reviewed literature: (i) clinical lead-alignment, in which an ECG report image is segmented into its anatomical lead grid and the fraction of saliency mass falling in each lead region is scored against the leads that are clinically informative for the predicted class (for example, the inferior leads II, III, and aVF for inferior myocardial infarction); (ii) an on-grid / ink-overlap check, which tests whether saliency concentrates on the drawn waveform rather than on background grid lines, margins, or calibration marks; and (iii) faithfulness evaluation via deletion and insertion curves, which measure whether the highlighted pixels causally drive the prediction, independent of whether they look clinically plausible.

4. Dataset and Preprocessing

Experiments use the Mendeley ECG Images dataset [13], comprising ECG report images collected from 929 patients and organised into four diagnostic categories verified by cardiologist review: Normal (284 images), Myocardial Infarction (240), Abnormal Heartbeat (233), and History of Myocardial Infarction (172), for a total of approximately 929 report images. Each image is a scanned or digitally rendered single-lead (Lead II) ECG report, presented as four sequential rhythm-strip rows beneath a printed header block containing patient and device metadata.

A stratified 70/15/15 patient-level split is used for training, validation, and testing respectively, with an explicit leakage assertion to guarantee no patient's images appear in more than one split. This yields a held-out test set of n = 140 images, with per-class support of 43 (Normal), 36 (MI), 35 (Abnormal Heartbeat), and 26 (History of MI), matching the class proportions of the full dataset.

4.1 Preprocessing Pipeline

Source images (approximately 2200 pixels along the longer edge) are first downscaled to 1024 pixels and then letterboxed to the fixed 224x224 input resolution required by the ImageNet-pretrained backbones. As discussed in Section 7, this two-stage downscaling is identified as a likely contributor to the model's accuracy ceiling, since it can attenuate the 1-2 pixel-wide ECG trace lines that carry the diagnostically relevant waveform morphology. Data augmentation is deliberately restricted to operations considered physiologically safe for ECG imagery - avoiding, in particular, horizontal or vertical flips and aggressive cropping, which would reverse temporal ordering or truncate diagnostically essential waveform segments, consistent with cautions raised in the ECG-specific augmentation literature [18].

5. Experimental Setup

All experiments were run on a single consumer-grade workstation GPU (NVIDIA RTX 4050, 6 GB VRAM) with 16 GB of system RAM, using PyTorch 2.4 and the timm model library. Given the constrained VRAM budget, training used a batch size of 12-16 images with gradient accumulation and automatic mixed precision. Of the four architectures implemented in the model factory, only ResNet-50 completed a full train-evaluate-explain cycle at the time of writing; EfficientNet-B3 and ViT-B/16 runs were planned but not completed, and are reported here as pending rather than with placeholder or estimated numbers, consistent with this paper's aim of transparent, non-fabricated reporting. The custom CNN baseline is available in the pipeline but was not part of the current experimental run.

The principal computational constraint encountered was host (system, not GPU) memory: with common development tooling and background processes running concurrently, only approximately 0.7 GB of system RAM was available during training, which caused a background explainability job to be terminated by the operating system's out-of-memory handler partway through the SHAP computation stage, and caused the SHAP gradient explainer to fall back to CPU execution before failing to allocate the required arrays. This is reported directly in Section 6.3 as an observed failure mode rather than omitted.

6. Results and Evaluation

6.1 Quantitative Classification Performance

Table 2 reports per-class precision, recall, F1-score, one-vs-rest AUC, and support for the ResNet-50 model on the held-out test set (n = 140). Table 3 reports overall accuracy and macro/weighted aggregate metrics with bootstrap confidence intervals.

Table 2. Per-Class Test-Set Performance - ResNet-50

Class	Precision	Recall	F1-score	AUC	n
Normal	1.00	0.02	0.05	0.85	43
Myocardial Infarction	0.94	0.47	0.63	0.94	36
Abnormal Heartbeat	0.46	0.97	0.62	0.97	35
History of MI	0.36	0.65	0.47	0.82	26

Table 3. Overall Test-Set Performance -ResNet-50 (n = 140)

Metric	Value	95% Bootstrap CI
Accuracy	0.493	[0.407, 0.579]
Macro Precision	0.691	-
Macro Recall	0.530	-
Macro F1	0.441	[0.367, 0.508]
Weighted Precision	0.732	-
Weighted F1	0.418	-
Macro-averaged AUC	0.895	-
Micro-averaged AUC	0.739	-
Training F1 (reference)	0.81	-

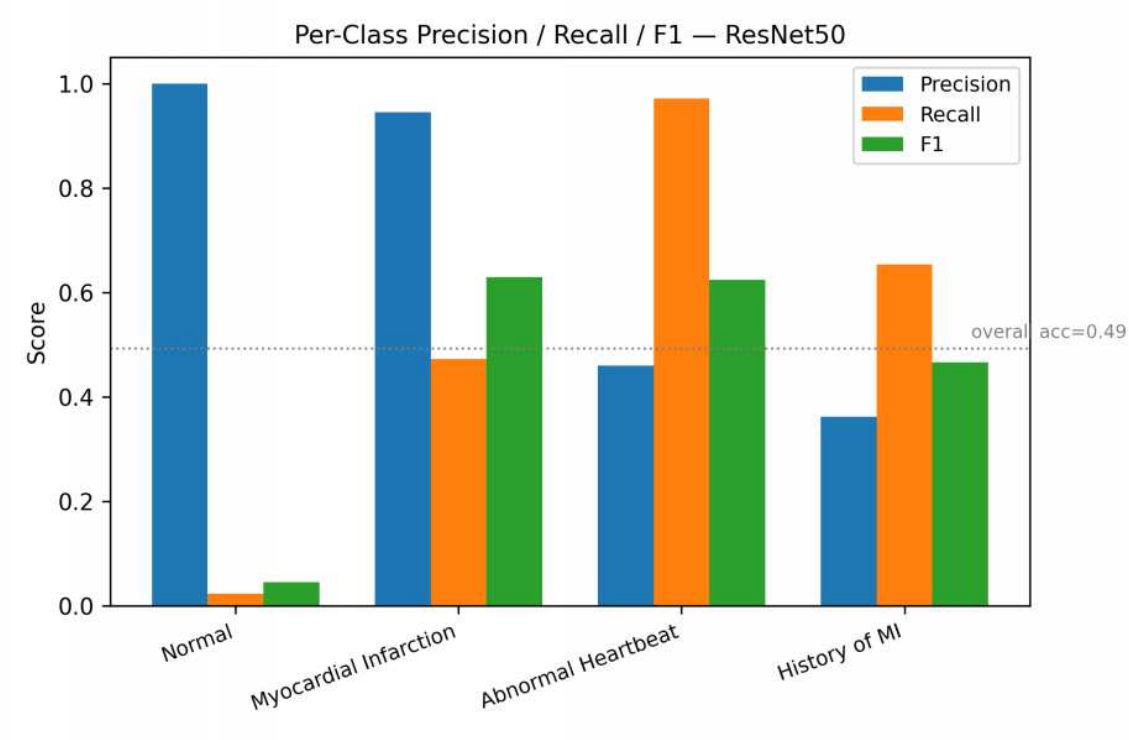


Figure 2. Per-class precision, recall, and F1-score for ResNet-50 on the held-out test set, with overall accuracy shown as *a reference line.*

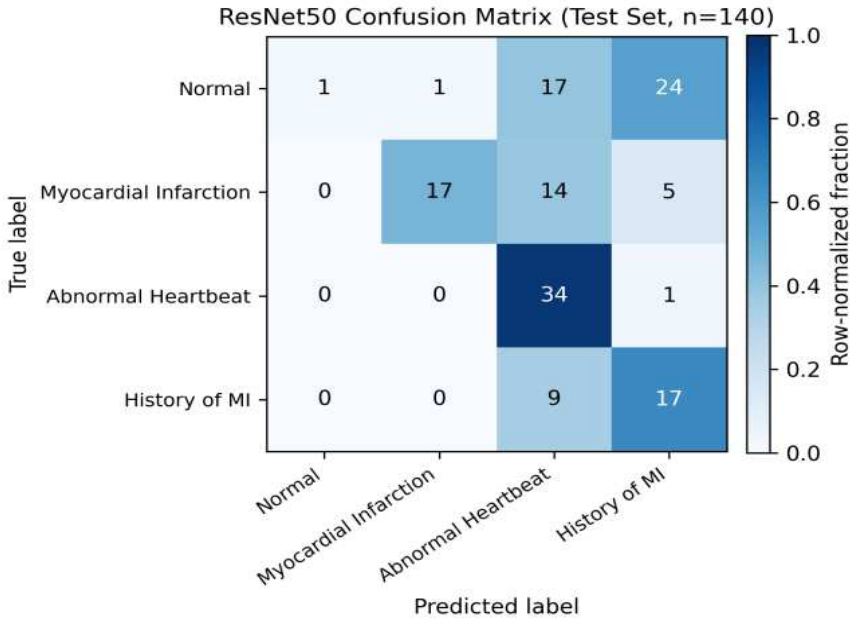


Figure 3. Row-normalised confusion matrix for ResNet-50 on the held-out test set (n = 140). Cell values show raw counts; shading shows the row-normalised fraction.

The confusion matrix reveals a pronounced asymmetry: 24 of 43 Normal test images (55.8%) are misclassified as History of MI and a further 17 (39.5%) as Abnormal Heartbeat, leaving only a single correctly classified Normal example. Table 4 lists all observed misclassification pairs and their frequency.

Table 4. Observed Misclassification Pairs - ResNet-50 Test Set

True Class	Predicted Class	Count
Normal	History of MI	24
Normal	Abnormal Heartbeat	17
Myocardial Infarction	Abnormal Heartbeat	14
History of MI	Abnormal Heartbeat	9
Myocardial Infarction	History of MI	5
Normal	Myocardial Infarction	1
Abnormal Heartbeat	History of MI	1

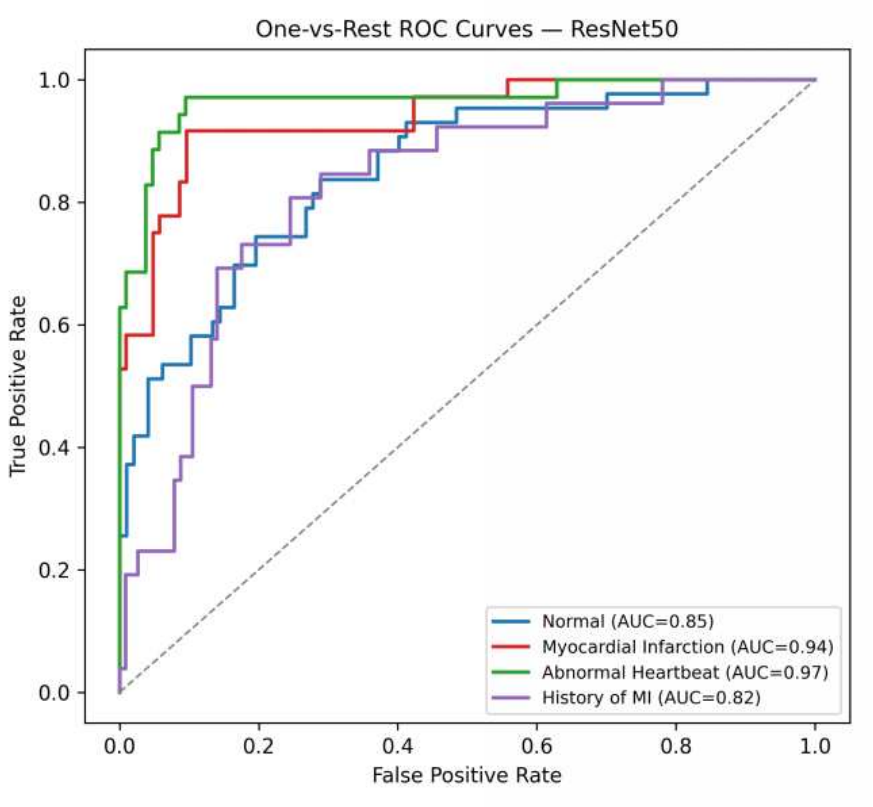


Figure 4. One-vs-rest ROC curves per class for ResNet-50 on the held-out test set. Per-class AUC ranges from 0.82 to 0.97, indicating that the model's continuous class scores separate the four conditions considerably better than its arg-max decisions.

6.2 Qualitative Explainability Results

Figure 5 and Figure 6 show representative attribution panels for a correctly classified Myocardial Infarction example and a correctly classified Normal example respectively, each comparing the original image against Grad-CAM, Grad-CAM++, SHAP, and occlusion-sensitivity output. In both cases, Grad-CAM and Grad-CAM++ concentrate saliency mass on the central portion of the drawn ECG waveform trace - specifically around the region containing the QRS complex and adjoining ST-segment/T-wave morphology in the printed rhythm strip - rather than on the surrounding grid paper, header text, or the solid-colour margin bands introduced by the letterboxing step. Occlusion sensitivity produces a coarser, patch-resolution map that broadly agrees with the Grad-CAM family but also attributes some importance to header-region patches, consistent with occlusion sensitivity's known sensitivity to patch size and its comparatively lower spatial precision as a cross-check method rather than a primary attribution technique.

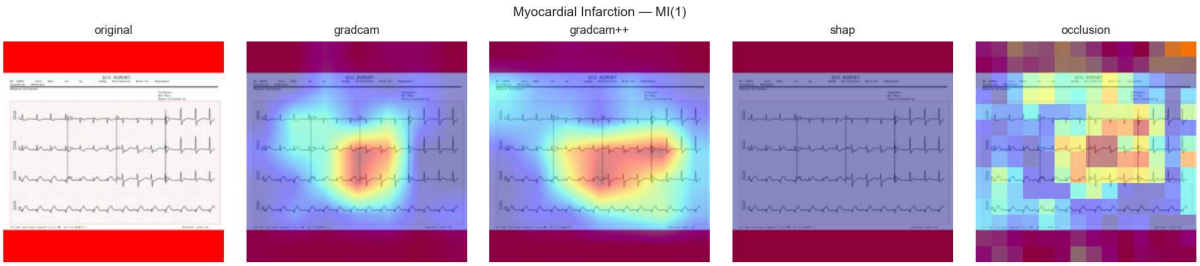


Figure 5. Attribution panel for a correctly classified Myocardial Infarction test example. From left to right: original image, Grad-CAM, Grad-CAM++, SHAP, occlusion sensitivity.

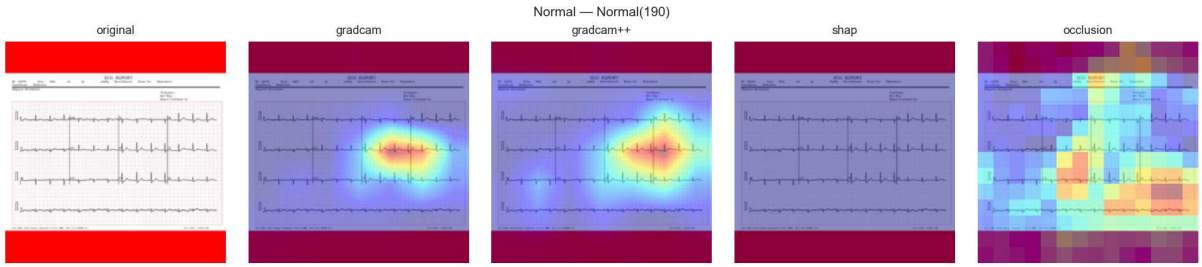


Figure 6. Attribution panel for a correctly classified Normal test example, structured identically to Figure 5.

6.3 Explainability Failure Modes (Reported Transparently)

Two explainability results are reported here as negative or incomplete findings rather than omitted. First, the SHAP gradient explainer produced visually degenerate, near-uniform maps carrying no discernible spatial structure in both inspected examples (Figures 5 and 6). This is consistent with the host-memory exhaustion described in Section 5: the progress log for this run records that SHAP fell back from GPU to CPU execution and subsequently failed to allocate the arrays required for a complete explanation, so the SHAP panels in this paper should be read as evidence of a compute-constrained failure mode rather than as a genuine attribution result. Second, the clinical lead-alignment analysis - designed to segment an ECG image into its 3x4 twelve-lead grid plus rhythm strip and score attention mass per anatomical lead - was found on visual inspection (Figure 7) to be a methodological mismatch for the specific Mendeley image variant used in this experiment: the images are single-lead (Lead II) rhythm-strip reports presented as four repeated rows of the same lead, not a full twelve-lead grid. Consequently, while the grid-segmentation code itself was verified against a genuine multi-lead printout layout (Figure 7), no quantitative per-lead alignment score is reported for the present dataset, since scoring a single-lead image against a twelve-lead anatomical template would not be a meaningful measurement. We flag this explicitly as a limitation in Section 8 and describe the corrective path in Section 9.



Figure 7. Verification of the twelve-lead anatomical grid segmentation (I, aVR, V1, V4 / II, aVL, V2, V5 / III, aVF, V3, V6 / rhythm strip) against a full multi-lead ECG printout layout. This segmentation is correct for a genuine 3x4 grid layout but is not applicable, as noted in Section 6.3, to the single-lead report images used in the present experiment.

We additionally note, as an audit observation rather than a demonstrated failure, that the current letterboxing implementation pads images with a saturated red fill (visible as the top and bottom bands in Figures 5 and 6) rather than a neutral black, white, or mean-pixel fill. Grad-CAM and Grad-CAM++ attention was not observed to concentrate on these margins in the inspected examples, but an unconventional, high-saturation padding colour is a preprocessing detail worth auditing in future work, since colour-based artefacts have in other imaging contexts been shown to introduce shortcut features that a classifier can exploit unintentionally.

7. Discussion

The central quantitative finding of this study is the divergence between ranking quality and decision quality. Per-class AUC of 0.82-0.97 indicates that the model's continuous output scores order the four classes in a manner that would support strong performance under an appropriately calibrated decision rule, yet arg-max accuracy

is only 49.3% and Normal-class recall collapses to 0.02. This pattern is diagnostic rather than merely disappointing: it points to a threshold and class-collapse problem layered on top of measurable overfitting (training F1 0.81 versus validation F1 0.43 on only approximately 650 training images), rather than to a model that has failed to learn any useful signal.

Three interacting causes are identified. First, class-weighted cross-entropy, introduced to counteract the moderate class imbalance in the training data, appears in this run to have pushed the decision boundary away from the Normal class, causing the great majority of Normal test images to be pushed into the Abnormal Heartbeat or History-of-MI categories (41 of 43 cases). Second, the two-stage downscale-then-letterbox preprocessing pipeline reduces source images from roughly 2200 pixels to a 224x224 input, a process that can attenuate the 1-2 pixel-wide trace lines that carry the ECG's diagnostic morphology, plausibly limiting the ceiling on achievable accuracy regardless of architecture or training schedule. Third, the training set size of approximately 650 images across four classes is small relative to the capacity of a 25.6-million-parameter ResNet-50, a mismatch reflected directly in the observed train-validation F1 gap.

On the explainability side, the qualitative concentration of Grad-CAM and Grad-CAM++ saliency on the drawn waveform rather than on background grid or margin regions is an encouraging sign that the model, despite its calibration problems, has learned to attend to genuinely ECG-relevant image content rather than to spurious correlates such as header text or scan artefacts. However, the honest reporting of the SHAP failure and the lead-alignment mismatch is, in the authors' view, as important a contribution of this paper as the classification numbers themselves: the companion review's Gap 3 specifically criticises the field for presenting explainability as decorative rather than validated, and a paper that reproduces that same pattern - by showing pretty heat-maps without disclosing that one of five methods failed and that a purpose-built validation check does not apply to the dataset at hand - would not meaningfully close that gap. Reporting these findings transparently is itself evidence that the pipeline's evaluation discipline is functioning as intended.

8. Limitations

Single completed architecture. Only ResNet-50 has completed a full train-evaluate-explain cycle; EfficientNet-B3 and ViT-B/16 results are not yet available, so no cross-architecture comparison can be reported at this time, and the taxonomy-level claims common in the broader literature (e.g., attention-based models outperforming CNNs) cannot be tested against this pipeline's own data yet.

Small held-out test set. With n = 140 test images across four classes, per-class metrics (particularly for the smallest class, History of MI, with only 26 test examples) carry wide sampling uncertainty, reflected in the reported bootstrap confidence intervals.

Single-institution, single-lead dataset. The Mendeley dataset used here is a single-lead (Lead II) report format from one data source; findings may not generalise to genuine twelve-lead printouts, different acquisition hardware, or other patient populations.

SHAP explanation failure. The SHAP gradient explainer did not produce usable output under the memory-constrained hardware used in this study; no SHAP-based quantitative or qualitative claim can be supported by the current run.

Lead-alignment methodology not yet quantitatively applied. The clinical lead-alignment scoring - arguably the paper's most direct response to the companion review's identified explainability gap - has been implemented and its grid segmentation verified against a genuine multi-lead layout, but has not yet produced a quantitative per-lead alignment score on this single-lead dataset, and would require either a genuine multi-lead image source or a redefinition of the alignment target region for a single-lead report.

No faithfulness quantification reported. Deletion/insertion faithfulness curves are implemented in the pipeline but are not reported numerically in this paper, pending a complete run across all correctly classified test examples.

No prospective or multi-reader clinical validation. As with the great majority of the reviewed literature, this study reports retrospective, single-dataset performance; no comparison against cardiologist readers or prospective clinical outcome data is available.

9. Future Scope

Complete the EfficientNet-B3 and ViT-B/16 training, evaluation, and explanation runs, enabling a genuine cross-architecture comparison under identical data splits, preprocessing, and evaluation protocol - directly addressing the companion review's observation that the field lacks standardised, controlled architectural comparisons.

Diagnose and correct the accuracy ceiling by (i) raising input resolution and reducing the aggressiveness of the pre-downscale step so that fine ECG trace lines are preserved, (ii) revisiting class weighting to address the Normal-class collapse, for example via calibrated decision thresholds or focal loss, and (iii) applying stronger regularisation and a longer frozen-head warm-up phase to reduce the observed train-validation F1 gap.

Acquire or construct genuine full twelve-lead grid images (or adapt the alignment target to a validated single-lead region definition) so that the clinical lead-alignment score can be computed and reported quantitatively, and can, in turn, be correlated with the deletion/insertion faithfulness curves already implemented in the pipeline.

Resolve the SHAP memory failure - for example via a lighter-weight SHAP explainer variant, reduced background-sample count, or execution on hardware with a larger memory budget - and re-run the full explainability battery to completion.

Audit and, if necessary, replace the current red-fill letterboxing implementation with a neutral padding scheme, and verify through an ablation that padding colour does not measurably influence classification or attribution.

Extend evaluation to external, multi-institution ECG image datasets to test generalisation, and, in the longer term, pursue prospective validation against cardiologist reference readings, consistent with the companion review's identification of prospective clinical validation as the field's most consequential open gap.

10. Conclusion

This paper has presented the empirical, experimental companion to a review of deep-learning architectures for automated cardiac disease detection using 12-lead ECG images, focusing specifically on the review's identified gap in validated explainability. A modular, five-stage pipeline was built covering data preparation, model factory, two-phase transfer-learning training, rigorous quantitative evaluation, and a five-method explainability suite including a purpose-built clinical lead-alignment check. A ResNet-50 baseline, trained and evaluated on the four-class Mendeley ECG Images dataset, achieved 49.3% test accuracy and a macro-F1 of 0.44, with per-class AUC of 0.82-0.97 indicating that the model has learned meaningful class-separating representations despite a currently poorly calibrated decision boundary and observable overfitting. Grad-CAM and Grad-CAM++ attribution concentrated on the ECG waveform trace itself, a qualitatively encouraging sign, while the SHAP explainer failed under the study's memory-constrained hardware and the lead-alignment check was found, on honest inspection, to be a methodological mismatch for the single-lead image variant used. Rather than presenting a polished but incomplete benchmark, this paper has aimed to document these results - both the encouraging and the unresolved - with the same rigour the companion review calls for from the field at large, and to lay out a concrete, specific roadmap for completing the multi-architecture comparison and the quantitative explainability validation that this pipeline was built to support.

References

[1] World Health Organization, "cardiovascular diseases (CVDs): Key facts," WHO, 2021.

[2] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," arXiv:1707.01836, 2017.

[3] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nature Medicine, vol. 25, no. 1, pp. 65-69, 2019.

[4] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixao, D. M. Oliveira, P. R. Gomes, J. A. Canazart, and A. L. P. Ribeiro, "Automatic diagnosis of the 12-lead ECG using a deep neural network," Nature Communications, vol. 11, no. 1, pp. 1-9, 2020.

[5] 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," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618-626.

[6] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765-4774.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv:2010.11929, 2020.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[9] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Machine Learning (ICML), 2019, pp. 6105-6114.

[10] W. Wieclawek, T. Piatkowski, and P. L. Porwik, "Classification of ECG images using a fine-tuned VGG-16 network," Applied Sciences, vol. 10, no. 18, p. 6525, 2020.

[11] S. S. Al-Zaiti, L. Besomi, Z. Bouzid, Z. Faramand, S. Frisch, C. Martin-Gill, and G. Clermont, "Machine learning-based prediction of acute coronary artery occlusion following emergency department evaluation for chest pain," Nature Communications, vol. 11, no. 1, pp. 1-10, 2020.

[12] H. Choi, S. Lee, and S. Lee, "Applying vision transformer for detecting atrial fibrillation from ECG spectrograms," Applied Sciences, vol. 11, no. 16, p. 7545, 2021.

[13] A. H. Khan, M. Hussain, and M. K. Malik, "ECG images dataset of cardiac patients," Mendeley Data, V2, 2021.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[15] P. Wagner, N. Strodthoff, R. D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," Scientific Data, vol. 7, no. 1, pp. 1-15, 2020.

[16] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299-1312, 2016.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, vol. 25, 2012, pp. 1097-1105.

[18] T. T. Um, F. M. J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, and D. Kulic, "Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks," in Proc. 19th ACM Int. Conf. Multimodal Interaction, 2017, pp. 216-220.

[19] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA: MIT Press, 2016.

[20] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," Nature Reviews Cardiology, vol. 16, no. 11, pp. 658-670, 2019.