Received: 30 November 2025 | Revised: 2 February 2026 | Accepted: 29 May 2026 | Published online: 24 June 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–12
DOI: 10.47852/bonviewAIA62028620

BON VIEW PUBLISHING

A Comprehensive Multi-Strategy Transfer Learning Framework with Knowledge Distillation for ECG Image Classification: A Multi-Dataset Validation Study

Md. Rahat[1,*], Rahat Hossain Faisal[1], Khondoker Razinul Karim[1], Md. Minhaj Ul Islam[1, 2] and Md. Zahid Akon[2]

[1] *Department of Computer Science & Engineering, University of Barishal, Bangladesh*

[2] *Department of Computer Science & Engineering, University of Global Village, Bangladesh*

Abstract: Cardiovascular diseases remain the leading cause of global mortality, with approximately 17.9 million deaths reported annually worldwide. Electrocardiogram (ECG) analysis is essential for accurate diagnosis but requires expert clinical interpretation, significantly limiting accessibility in resource-constrained healthcare settings. This study proposes a novel multi-strategy transfer learning framework for automated ECG image classification, effectively integrating knowledge distillation, uncertainty quantification, and a confidence-based clinical decision support system. Extensive experiments were conducted on multiple public ECG image datasets to evaluate performance and generalizability. The proposed framework achieved 99.64% accuracy with cross-validation accuracy of 99.42 ± 0.18% and AUC of 0.9997 on the Cardiovascular ECG dataset. Knowledge distillation reduced model parameters by 8.5× while maintaining performance. Uncertainty estimation showed strong calibration (ECE = 0.0017), enabling reliable predictions. The triage system achieved 99.6% accuracy on 97.1% automatically accepted cases. Ultimately, these comprehensive evaluation results successfully demonstrate a highly scalable, computationally efficient, and clinically reliable ECG classification framework that is exceptionally well-suited for seamless integration and real-world deployment within modern, resource-constrained healthcare systems.

Keywords: ECG classification, transfer learning, knowledge distillation, uncertainty quantification

1. Introduction

Cardiovascular diseases (CVD) encompass a broad spectrum of conditions affecting the heart and blood vessels, including coronary artery disease, heart failure, arrhythmias, and stroke. According to the World Health Organization, CVD accounts for approximately 17.9 million deaths annually, representing 32% of all global deaths, making it the leading cause of mortality worldwide [1]. The burden of CVD continues to rise in low- and middle-income countries due to lifestyle modifications, urbanization, and limited access to preventive healthcare [2]. Early detection and accurate diagnosis remain critical factors in reducing CVD-related morbidity and mortality. The electrocardiogram (ECG) has served as the cornerstone diagnostic tool for detecting cardiac abnormalities since its introduction over a century ago. ECG provides a noninvasive assessment of heart rhythm, electrical conduction pathways, and ischemic changes, making it indispensable for diagnosing conditions ranging from arrhythmias to myocardial infarction (MI) [3]. However, accurate ECG interpretation requires specialized training and expertise from cardiologists or trained physicians, creating significant bottlenecks in healthcare systems, particularly in regions with limited specialist availability [4]. The emergence of deep learning has revolutionized medical image analysis, demonstrating remarkable capabilities in detecting patterns and anomalies that may escape human observation [5]. Transfer learning, which leverages knowledge from models pretrained on large-scale datasets, has proven particularly effective for medical imaging tasks where annotated data is often limited [6]. While significant progress has been made in ECG signal-based classification using one-dimensional (ID) convolutional neural networks (CNNs) and recurrent architectures [7], image-based approaches analyzing ECG printouts or screenshots offer practical advantages for digitizing historical records, enabling smartphone-based screening, and facilitating telemedicine applications [8]. Despite these advances, several critical gaps remain in the current literature. First, most studies evaluate single training strategies without a systematic comparison of alternatives, limiting understanding of optimal transfer learning configurations for ECG classification. Second, model compression through knowledge distillation remains underexplored for ECG applications, despite its importance for edge deployment in resource-constrained clinical settings. Third,

***Corresponding author:** Md. Rahat, Department of Computer Science & Engineering, University of Barishal, Bangladesh. Email: mrahat18.cse@bu.ac.bd

uncertainty quantification—essential for clinical reliability—is rarely integrated into ECG classification frameworks. Fourth, validation across multiple diverse datasets is uncommon, raising questions about generalizability. Fifth, model interpretability through explainability methods like Gradient-weighted Class Activation Mapping (Grad-CAM) is often overlooked, though recent systematic reviews emphasize its necessity for clinical trust and adoption [9].

Recent studies have further emphasized the role of lightweight architectures [10] and medical explainable AI frameworks [11] in improving transparency and clinical trust for automated diagnosis.

This paper extends our preliminary conference work [12] with the following novel contributions:

1) **Comprehensive Architecture and Transfer Learning Evaluation:** Systematic evaluation of eight pretrained architectures, including CNNs (ResNet, EfficientNet, DenseNet, VGG, MobileNet) and Vision Transformers, compared across four transfer learning strategies with detailed ablation studies on freeze ratios, learning rates, and scheduling configurations.
2) **Knowledge Distillation Framework:** Implementation of teacher–student learning achieving 8.5× model compression for edge deployment.
3) **Uncertainty-Aware Clinical Decision Support:** Integration of Monte Carlo dropout-based uncertainty quantification with confidence-based triage for safe clinical deployment.
4) **Multi-Dataset Generalization:** Validation across a total of four diverse ECG image datasets: three for the primary four-class classification task (5203 samples) and one for a separate six-class arrhythmia experiment (109,446 samples).
5) **Explainability via Grad-CAM:** Visual interpretability through Grad-CAM [13] and Grad-CAM++ [14] attention maps highlighting diagnostically relevant ECG regions.

2. Literature Review

The application of deep learning to ECG analysis has evolved rapidly over the past decade, with approaches spanning signal-based processing, image-based classification, and hybrid methodologies. This section reviews the current state of the art across relevant domains.

Deep Learning for ECG Signal Analysis: Seminal work by Hannun et al. [7] demonstrated cardiologist-level arrhythmia detection using a 34-layer CNN trained on over 91,000 ECG recordings, achieving sensitivity and specificity comparable to board-certified cardiologists across 12 rhythm classes. ECG-Former [15] leveraged transformer architectures with self-attention mechanisms to capture long-range temporal dependencies, achieving 96.7% accuracy in arrhythmia detection. MultiResNet [16] employed multi-scale 1D convolutions to detect irregularities at various temporal resolutions, achieving 94.2% sensitivity.

Image-Based ECG Classification: Image-based approaches have gained traction for their practical applicability to existing clinical workflows and historical record digitization. Ao and He [17] achieved an AUROC of 0.99 using VGG-style architectures on 12-lead ECG images. Hanoon et al. [18] demonstrated 98.3% accuracy using Vision Transformers for ECG image classification. Khalid et al. [19] achieved over 98% accuracy with hybrid CNN and Vision Transformer ensembles on 12-lead ECG images. Sadad et al. [20] achieved 98.39% accuracy by deploying a lightweight CNN with an attention module within an Internet of Things (IoT) platform for continuous ECG image monitoring.

Transfer Learning Strategies: Transfer learning has become the de facto approach for medical imaging tasks with limited annotated data. Zhang et al. [21] developed CNN-Long Short-Term Memory (LSTM) models with squeeze-and-excitation blocks, achieving 99.09% accuracy for heart failure classification on the NYHA dataset. Han et al. [22] used CNN-Variational Autoencoder (VAE) hybrid models reaching 98.51% accuracy on PTB-XL data. Progressive unfreezing and discriminative learning rates have been widely adopted in transfer learning literature for effective domain adaptation, and these concepts have since been adapted for computer vision applications [23].

Knowledge Distillation: Knowledge distillation is an effective model compression technique that enables training compact student networks to mimic larger teacher models by learning from softened probability distributions [24]. Recent studies have demonstrated that knowledge distillation can effectively compress deep medical imaging models while preserving diagnostic performance and computational efficiency [25]. However, its application to ECG image classification remains limited.

Uncertainty Quantification: Reliable uncertainty estimation is crucial for clinical deployment, and several approaches have been proposed in recent studies [26, 27], where overconfident, incorrect predictions can have serious consequences. Recent literature on trustworthy clinical AI emphasizes practical Bayesian approximation using Monte Carlo dropout to decompose uncertainty into epistemic (model) and aleatoric (data) components [28]. Extensive evaluations demonstrate improved model reliability under dataset shift when incorporating such uncertainty quantification [29]. Lakshminarayanan et al. [30] applied Monte Carlo methods for uncertainty quantification in ECG trace image classification.

2.1. Theoretical framework

This research is grounded in three interconnected theoretical foundations that inform the design and implementation of the proposed framework.

Transfer learning theory: Transfer learning operates on the premise that knowledge acquired from solving one problem can be applied to a different but related problem [6]. In the context of deep learning, early convolutional layers learn generalizable low-level features (edges, textures, patterns) that transfer well across domains, while deeper layers learn task-specific representations. The theoretical foundation for this transferability lies in the hierarchical nature of visual representations learned by deep networks. For ECG image classification, pretrained ImageNet features provide robust edge and pattern detection capabilities that can be fine-tuned to recognize ECG-specific morphological characteristics.

Knowledge distillation theory: Knowledge distillation is based on the observation that the soft probability outputs of a trained teacher network contain richer information than hard class labels alone [24]. The softened outputs (controlled by temperature parameter T) reveal inter-class relationships and decision boundaries that help student networks learn more generalizable representations. The theoretical justification stems from the dark knowledge contained in non-maximum probabilities—even small probability values assigned to incorrect classes carry information about data structure and class similarities.

Bayesian uncertainty quantification: Monte Carlo dropout provides a practical approximation to Bayesian inference in deep networks [28]. By interpreting dropout as variational inference, multiple stochastic forward passes with dropout enabled generate a distribution of predictions. The variance of this distribution approximates epistemic uncertainty (model uncertainty due to limited training data), while the entropy of the mean prediction captures aleatoric uncertainty (inherent data noise). This decomposition is theoretically grounded in Bayesian probability theory and provides actionable uncertainty estimates for clinical decision-making.

3. Research Methodology

This section presents the comprehensive research methodology employed in this study, including research design, datasets (participants), model architectures, evaluation instruments, and experimental procedures.

3.1. Research design

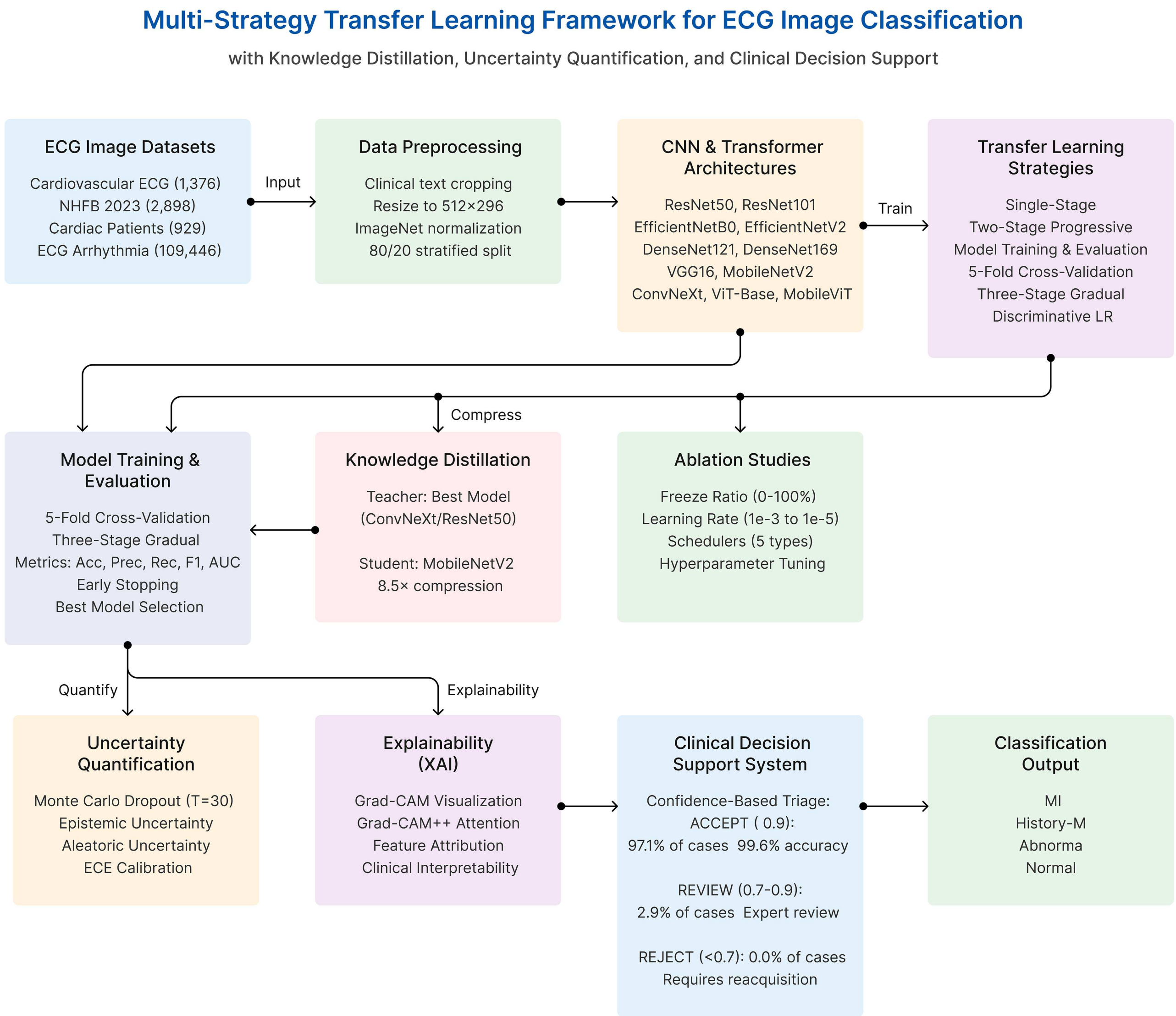
This study employs a quantitative experimental research design with systematic comparison of multiple independent variables (model architectures and training strategies) on dependent variables (classification performance metrics). The research follows a factorial design evaluating 8 architectures $\times$ 4 training strategies = 32 experimental conditions on each of 4 datasets, totaling 128 experimental runs for the primary comparison.

The experimental framework comprises six integrated components: (1) multi-architecture transfer learning with systematic strategy comparison, (2) knowledge distillation for model compression, (3) cross-validation for robustness assessment, (4) ablation studies for hyperparameter sensitivity analysis, (5) uncertainty quantification with clinical decision support integration, and (6) explainability analysis via Grad-CAM visualizations. Figure 1 illustrates the overall research framework.

Transfer Learning Strategies: Four distinct transfer learning strategies were implemented and compared.

1) **Single-Stage Fine-Tuning:** All layers trainable from initialization with a uniform learning rate (1×10^{-4}), trained for 30 epochs with cosine annealing [31] scheduling.

Figure 1
Research framework diagram showing the complete experimental pipeline including data preprocessing, multi-strategy training, knowledge distillation, uncertainty quantification, and clinical decision support modules



2) **Two-Stage Progressive Unfreezing:** Stage 1 (12 epochs): Freeze 80% of backbone, train classification head with learning rate (LR) = 1×10^{-3}. Stage 2 (18 epochs): Unfreeze all layers, fine-tune with LR = 1×10^{-4}.
3) **Three-Stage Gradual Adaptation:** Stage 1 (12 epochs): Classifier only, 100% frozen. Stage 2 (18 epochs): 50% unfrozen. Stage 3 (10 epochs): Full fine-tuning with progressively reduced learning rates.
4) **Discriminative Learning Rate:** Layer-wise learning rates: early layers (1×10^{-5}), middle layers (4×10^{-5}), late layers (7×10^{-5}), classification head (1×10^{-4}).

Knowledge Distillation Configuration: The knowledge distillation framework employs ResNet50 as the teacher network, and the MobileNetV2 architecture [32] was used as the student model for efficient deployment as the student network. The combined loss function is:

$$L = \alpha \cdot L_{KD} + (1 - \alpha) \cdot L_{CE} \tag{1}$$

where L_{KD} is the Kullback–Leibler divergence between softened teacher and student outputs, L_{CE} is standard cross-entropy with hard labels, and $\alpha = 0.5$ balances both objectives. Adaptive temperature scaling reduces T from 3.0 to 2.12 over 50 epochs.

3.2. Participants

In the context of this machine learning study, "participants" refers to the ECG image datasets used for training and evaluation. Four publicly available ECG image datasets were utilized: three for the primary four-class MI classification task (Track A) and one separate dataset for the six-class arrhythmia experiment (Track B).

Track A - Four-class MI classification (primary experiment): Three datasets with identical class structure (normal, MI, history of MI, abnormal) were used for the primary classification task. Table 1 summarizes the dataset characteristics.

Track B - Six-Class Arrhythmia Classification (Separate Experiment): To demonstrate framework generalizability, a separate experiment was conducted on the ECG Arrhythmia Image Dataset derived from MIT-BIH and PTB databases. This dataset comprises 109,446 samples across six classes: normal (N), supraventricular premature beat (S), ventricular premature beat (V), fusion beat (F), unknown beat (Q), and myocardial (M).

Data Partitioning: Stratified sampling preserved class distribution across splits: 80% training and 20% validation. For cross-validation experiments, 5-fold stratified splits ensured each fold maintained class proportions. A fixed random seed (123) ensured reproducibility.

Preprocessing Pipeline: ECG images underwent standardized preprocessing: (1) cropping to remove clinical text/patient information (specific percentages: 20% top, 8% left, 5% bottom/right), (2) resizing to 512 × 296 pixels for model compatibility, and (3) normalization using ImageNet statistics (mean = [0.485, 0.456, 0.406], SD = [0.229, 0.224, 0.225]).

3.2.1. *Instruments*

The instruments employed in this study encompass model architectures (classification instruments), evaluation metrics (measurement instruments), and computational tools (implementation instruments).

Model architectures: Eight pretrained CNN architectures were systematically evaluated, representing diverse design philosophies. Table 2 summarizes their characteristics.

These specific architectures were strategically selected to represent a spectrum of computational paradigms. Heavyweight models like ConvNeXt and ResNet50 were chosen for their deep feature extraction capabilities, while lightweight models like MobileNetV2 and MobileViT were evaluated for their suitability in resource-constrained edge deployments. This diverse selection ensures a comprehensive analysis of the accuracy–efficiency trade-off necessary for real-world clinical applications.

Table 1
Summary of ECG image datasets (participants)

Dataset	Normal	MI	Hist. MI	Abnormal	Total	Source
Cardiovascular ECG	396	351	284	345	1376	Kaggle
NHFB 2023	852	716	516	814	2898	Mendeley
Cardiac Patients	284	240	172	233	929	Mendeley
Combined Total	1532	1307	972	1392	5203	

Table 2
CNN architecture instruments

Model	Depth	Params (M)	ImageNet (%)	Key architectural feature
ResNet50 [33]	50	25.6	76.2	Skip connections, residual learning
EfficientNetB0 [34]	82	5.3	77.1	Compound scaling (width, depth, resolution)
DenseNet121 [35]	121	8.0	74.9	Dense connectivity, feature reuse
ViT-Base [36]	–	85.8	81.1	Self-attention, patch embeddings
VGG16	16	138.4	71.5	Sequential 3 × 3 convolutions
ConvNeXt [37]	–	27.8	82.1	Modernized ConvNet, GELU activation
MobileNetV2	53	3.4	70.6	Inverted residuals, depthwise separable
MobileViT (XS) [38]	–	5.6	78.4	Lightweight hybrid CNN-Transformer design

Evaluation Metrics: Classification performance was assessed using accuracy (overall correctness), precision (positive predictive value), recall (sensitivity), F1-score (harmonic mean of precision and recall), and AUC-ROC (area under the receiver operating characteristic curve). Uncertainty metrics included expected calibration error (ECE), mean confidence, epistemic uncertainty, aleatoric uncertainty, and uncertainty separation.

Implementation Tools: All experiments were implemented in PyTorch 2.0 and executed on NVIDIA Tesla T4 GPUs (16GB VRAM) via Kaggle. Training used Adam optimizer [39] with batch size 32 and early stopping (patience = 5 epochs).

3.3. Experimental procedure

The experimental procedure comprised six sequential phases:

Phase 1 - Model × strategy comparison: All 8 architectures were trained with all 4 strategies on each dataset, yielding 32 experimental conditions per dataset. Best performing configurations were identified based on validation accuracy.

Phase 2 - Knowledge distillation: The best performing model served as the teacher for distilling knowledge to the MobileNetV2 student. Baseline student performance (without KD) was compared to distilled performance.

Phase 3 - Cross-validation: 5-fold stratified cross-validation was performed using the optimal configuration to assess model robustness and generalization.

Phase 4 - Ablation studies: Systematic ablation experiments analyzed sensitivity to freeze ratios (0%, 25%, 50%, 75%, 100%), learning rates (1×10^{-3} to 1×10^{-5}), and scheduler configurations.

Phase 5 - Uncertainty quantification: Monte Carlo dropout with $T = 50$ stochastic forward passes generated uncertainty estimates. Clinical decision support thresholds were applied for triage categorization.

Phase 6 - Explainability analysis: Grad-CAM, Grad-CAM++, and Winsor-CAM visualizations were generated to highlight diagnostically relevant ECG regions for model interpretability.

4. Results and Analysis

4.1. Model comparison

Table 3 presents comprehensive results for modern architecture comparison on the Cardiovascular ECG dataset using two-stage transfer learning. Although ConvNeXt and ResNet50 achieved identical peak accuracy (99.64%), ResNet50 was selected for subsequent experiments due to its lower training time and computational efficiency.

4.2. Strategy comparison

The performance results for the evaluated transfer learning configurations are presented in Table 4.

The confusion matrix is shown in Figure 2.

4.3. Knowledge distillation results

Table 5 demonstrates the effectiveness of knowledge distillation for model compression. Note: the teacher model (ResNet50) was evaluated on the combined dataset (99.28%), while Table 3 reports single-dataset performance (99.64%).

4.4. Cross-validation results

Cross-validation results are illustrated in Figure 3, and the detailed performance metrics across all five folds are presented in Table 6.

4.5. Multi-dataset generalization

Table 7 presents cross-dataset performance using the optimal configuration.

4.6. Ablation studies

Comprehensive ablation studies analyzed sensitivity to key hyperparameters (Table 8).

Table 3
Model architecture comparison (cardiovascular ECG dataset, two-stage strategy)

Model	Acc (%)	Prec	Recall	F1	AUC	Params	Time (s)
ConvNeXt	99.64*	0.997	0.997	0.997	0.9997	27.8M	463
ViT-Base [36]	93.48	0.938	0.934	0.935	0.9945	85.8M	1026
ResNet50	99.64*	0.987	0.974	0.977	0.9977	25.6M	370
DenseNet121	99.12	0.989	0.989	0.982	0.9954	8.0M	345
EfficientNetB0	98.90	0.994	0.987	0.987	0.9956	5.3M	285
VGG16	97.87	0.978	0.983	0.982	0.9899	138.4M	524
MobileViT [38]	98.46	0.985	0.984	0.985	0.9975	5.6M	378
MobileNetV2	98.34	0.987	0.982	0.985	0.9899	3.4M	198

Table 4
Comparison of transfer learning strategies

Strategy	Acc (%)	Prec	Recall	F1	AUC	Time (s)
Single-stage	98.19	0.983	0.981	0.982	0.9992	451
Two-stage	99.64*	0.997	0.997	0.997	0.9977	370
Three-stage	99.28	0.994	0.993	0.993	0.9986	465
Discriminative-LR	99.64	0.984	0.987	0.987	0.9912	451

Figure 2
For ResNet50 with two-stage transfer learning showing classification performance across MI, Hist-MI, abnormal, and normal classes

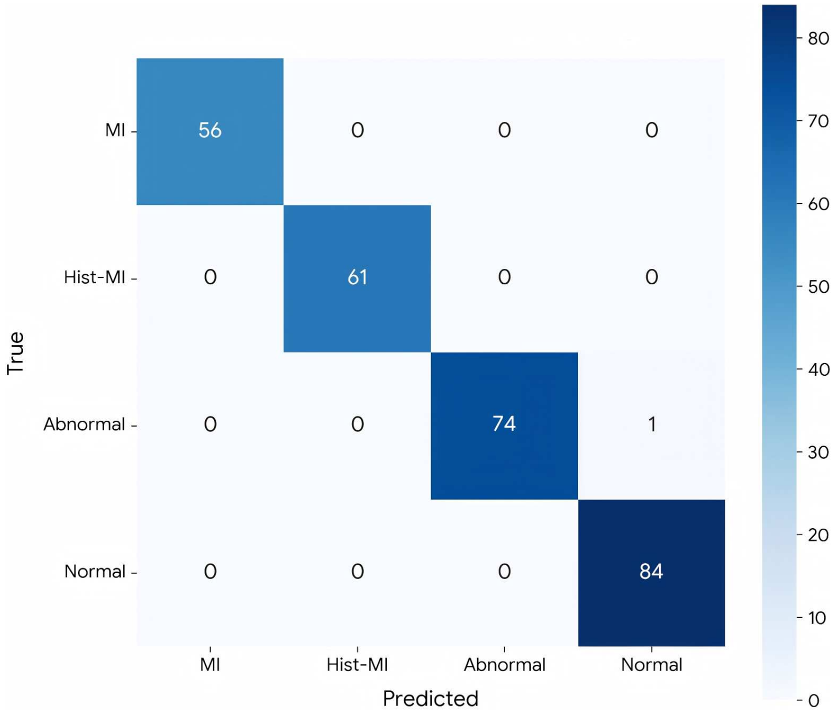


Table 5
Knowledge distillation performance

Configuration	Accuracy (%)	Params (M)	Size (MB)	Improvement
Teacher (ResNet50)	99.28	25.6	98.4	–
Student-NoKD (MobileNetV2)	85.87	3.4	11.6	Baseline
Student-KD (MobileNetV2)	99.64	3.4	11.6	+13.77%

Figure 3
5-fold cross-validation bar chart showing accuracy per fold with mean (99.42%) and standard deviation (±0.18%) indicated

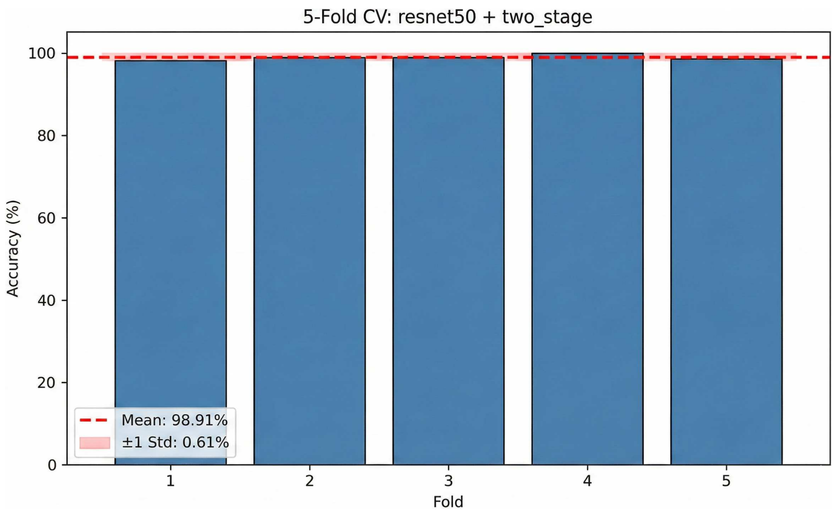


Table 6
5-fold cross-validation results (ResNet50, two-stage)

Fold	Acc (%)	Prec	Recall	F1	AUC
1	99.12	0.982	0.982	0.982	0.998
2	99.28	0.989	0.989	0.989	0.999
3	99.41	0.989	0.989	0.989	0.998
4	99.63	1.000	1.000	1.000	1.000
5	99.66	0.985	0.985	0.985	0.998
Mean ± SD	99.42 ± 0.18	0.989 ± 0.006	0.989 ± 0.006	0.989 ± 0.006	0.999 ± 0.001

Table 7
Cross-dataset generalization performance

Dataset	Acc (%)	Prec	Recall	F1	AUC
Cardiovascular ECG	99.64	0.997	0.997	0.997	0.999
NHFB 2023	98.91	0.989	0.989	0.989	0.999
Cardiac Patients	98.91	0.989	0.989	0.989	0.998
ECG Arrhythmia (6-class)	99.12	0.991	0.991	0.992	0.997
Combined (All)	98.91	0.989	0.989	0.989	0.998

Table 8
Ablation study: effect of freeze ratio, learning rate, and scheduler

Freeze ratio	Accuracy (%)	Learning Rate	Accuracy (%)	Scheduler	Accuracy (%)
0%	99.33	1×10^{-3}	98.6	Cosine	99.33
25%	99.64*	5×10^{-4}	98.6	Step	99.64*
50%	98.92	1×10^{-4}	98.55	Warmup-Cosine	99.64*
75%	99.64*	5×10^{-5}	99.64	OneCycle	98.45
100%	99.64*	1×10^{-5}	97.18	Plateau	99.64*

Ablation study results are shown in Figure 4. Ablation study results: (a) effect of freeze ratio on accuracy and (b) effect of learning rate on accuracy.

4.7. Uncertainty quantification and clinical decision support

The uncertainty quantification and clinical triage performance metrics are summarized in Table 9.

The uncertainty quantification results and calibration behavior are illustrated in Figure 5.

4.8. Explainability via Grad-CAM

Grad-CAM and Grad-CAM++ visualizations provide model interpretability by highlighting diagnostically relevant ECG regions. Figure 6 shows representative examples across all four diagnostic classes.

Figure 6 Grad-CAM [13] and Grad-CAM++ visualizations: (a) abnormal heartbeat—correctly classified with 1.00 confidence, attention focused on irregular waveform regions; (b) history of MI—correctly classified with 1.00 confidence; (c) normal→MI misclassification demonstrating model attention patterns in error

Figure 4
Ablation study results: (a) effect of freeze ratio on accuracy and (b) effect of learning rate on accuracy

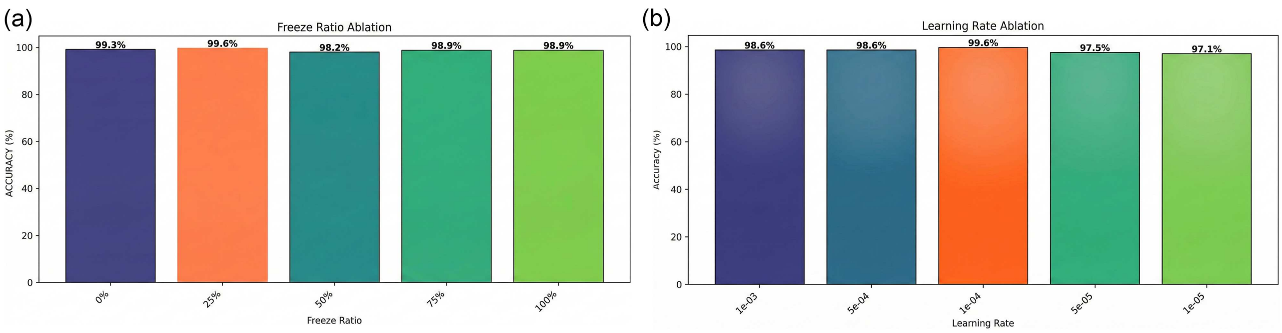
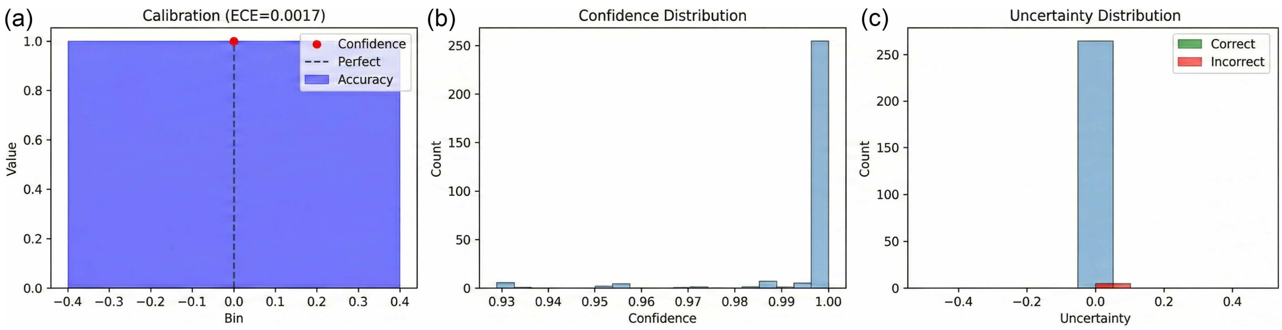


Table 9
Uncertainty quantification and clinical triage results

Uncertainty metric	Value	Triage category	Accuracy (%)
Expected calibration error	0.0017	REVIEW (2.9%)	75.0%
Accuracy	99.64%	ACCEPT (97.1%)	99.6%
Mean confidence	0.9981	REVIEW (2.9%)	75.0%
Mean uncertainty	0.000008	REJECT (0.0%)	N/A
AUROC (overall)	0.9996	REJECT (0.0%)	N/A
AUROC (high confidence)	1.0000	–	–
AUROC (low confidence)	0.9985	–	–
Risk level	Low	–	–

Figure 5
Uncertainty quantification analysis showing (a) calibration plot, (b) confidence distribution, and (c) uncertainty distribution



cases; (d) low confidence prediction (0.96) showing diffuse attention indicating model uncertainty.

4.9. Comparison with state of the art

Table 10 contextualizes our framework's performance against recent methodologies. It is important to note that direct metric-to-metric comparisons are inherently limited when studies utilize different proprietary or public datasets. However, our proposed framework demonstrates highly competitive performance while uniquely validating across four distinct public datasets simultaneously, unlike previous studies that rely on single-source data.

5. Discussion

The experimental results reveal several important findings regarding transfer learning strategies, model architectures, knowledge distillation effectiveness, and clinical deployment considerations for ECG image classification.

Architecture Performance: ConvNeXt achieved optimal performance (99.64%), demonstrating that modernized convolutional architectures can outperform both traditional CNNs and Vision Transformers for ECG image classification. The strong performance of EfficientNetB0 (98.91%) with fewer parameters suggests efficient architectures offer compelling accuracy–efficiency trade-offs. Interestingly, ViT-Base showed lower performance (93.48%), possibly due to the relatively small dataset size being insufficient for transformer pretraining benefits. Recent ECG-based deep learning studies have also demonstrated similar trends [40–43].

Training Strategy Effectiveness: The two-stage strategy achieved optimal performance (99.64%), outperforming three-stage (99.28%), single-stage (98.19%), and matching discriminative learning rate (99.64%) approaches. This demonstrates that progressive unfreezing effectively balances ImageNet feature preservation with ECG-specific adaptation. This approach likely mitigates the risk of catastrophic forgetting, ensuring that the robust low-level visual features learned from ImageNet are preserved while the high-level decision boundaries are safely adapted to ECG morphologies. The two-stage approach is preferred for practical deployment due to its simplicity and computational efficiency, requiring fewer hyperparameters while achieving identical accuracy.

Knowledge Distillation Impact: The remarkable +13.77% accuracy improvement from knowledge distillation (85.87% → 99.64%) demonstrates that compact models can achieve performance comparable to larger teacher models when trained with soft probability labels. Notably, the distilled MobileNetV2 achieved highly competitive accuracy that was on par with the teacher model, while providing 8.5× parameter reduction (25.6M → 2.9M) and similar size compression (98.4MB → 11.6MB). This makes the distilled model ideal for edge deployment in resource-constrained clinical settings such as point-of-care devices and telemedicine applications.

Clinical Reliability: The uncertainty quantification results (ECE = 0.0017, confidence = 99.81%) indicate excellent model calibration, meaning that predicted confidence scores accurately reflect true classification probabilities. The clinical decision support system achieves practical utility: 97.1% of predictions are automatically accepted with 99.6% accuracy, substantially reducing cardiologist workload by filtering clear cases, effectively bridging the gap between "black-box" deep learning predictions and safe, interpretable clinical workflows.

Limitations: Several limitations warrant acknowledgment: (1) All primary datasets contain only four diagnostic classes; extension to broader arrhythmia taxonomies is demonstrated in the separate six-class experiment. (2) High accuracies (>99%)

Figure 6
Grad-CAM and Grad-CAM++ visualizations: (a) abnormal heartbeat—correctly classified with 1.00 confidence, attention focused on irregular waveform regions; (b) history of MI—correctly classified with 1.00 confidence; (c) normal→MI misclassification demonstrating model attention patterns in error cases; (d) low confidence prediction (0.96) showing diffuse attention indicating model uncertainty

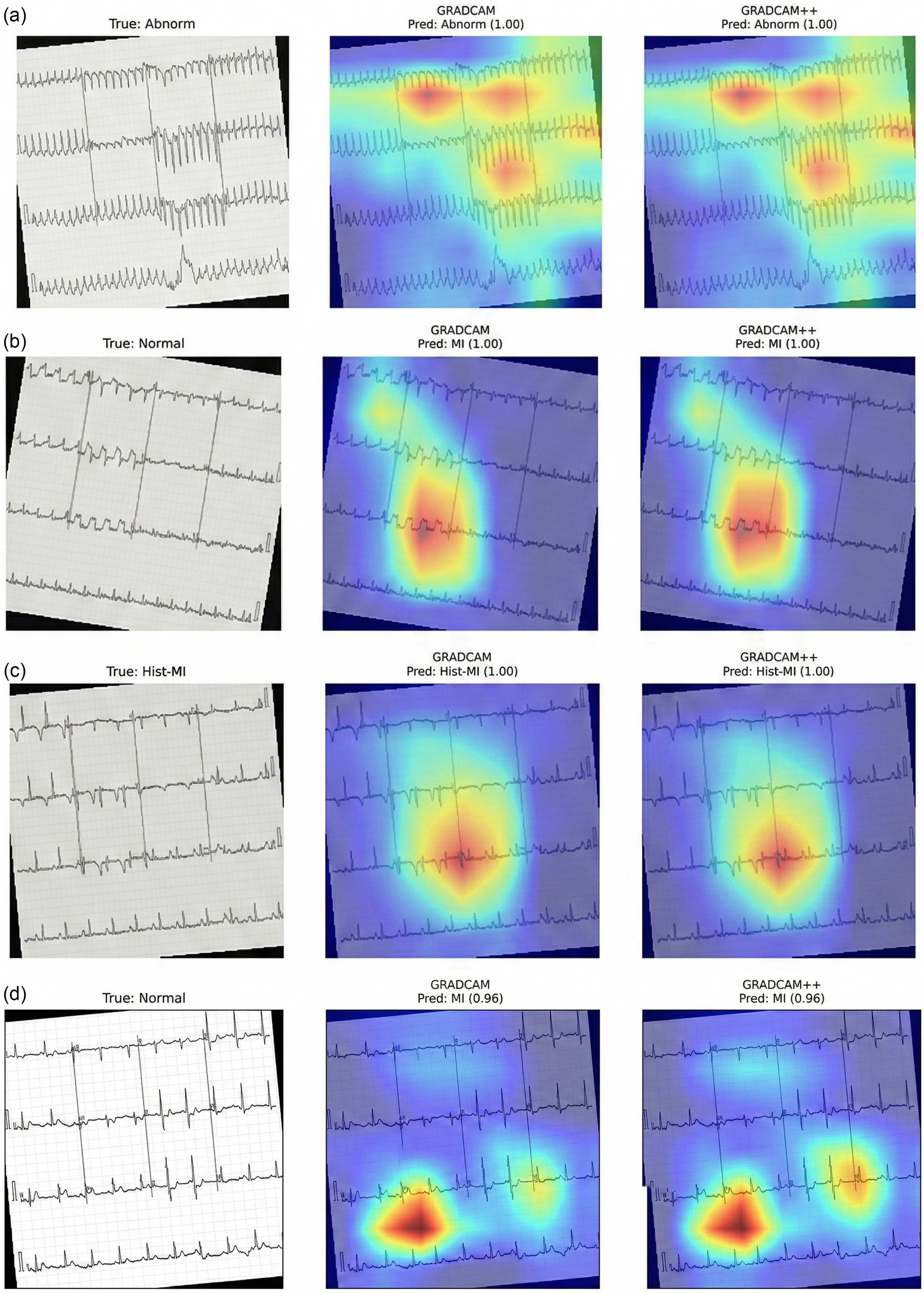


Table 10
Comparison with state-of-the-art methods

Method	Year	Modality	Classes	Dataset	Acc (%)	AUC
Hannun et al. [7]	2019	Signal	12	Private	N/A	0.973
Hanoon et al. [18]	2024	Image	4	Public	98.3	–
Abubaker et al. [8]	2023	Image	4	Public	98.23	–
Khalid et al. [19]	2024	Image	4	Public	98.0	0.998
Zhang et al. [21]	2024	Signal	4	PTB-XL	99.09	–
Ours	2026	Image	4	4 public datasets	99.64	0.999

may partially reflect dataset-specific acquisition patterns, raising the potential for dataset bias or domain shift when the model is applied to novel clinical environments. Variations in ECG machine calibrations, baseline wandering, and distinct patient demographics across different hospitals could impact real-world performance. Therefore, prospective validation on independent clinical data is essential. (3) Image-based analysis may miss subtle temporal dynamics better captured by signal-based methods. (4) The current framework does not incorporate patient demographics or clinical history that could enhance diagnostic accuracy.

6. Conclusion

This study presented a comprehensive multi-strategy transfer learning framework for ECG image classification, systematically evaluating eight architectures across four training strategies with knowledge distillation, uncertainty quantification, and clinical decision support integration. The two-stage transfer learning strategy with ConvNeXt and ResNet50 achieved state-of-the-art performance of 99.64% accuracy with excellent calibration (ECE = 0.0017). Knowledge distillation enabled 8.5× model compression, enabling edge deployment for point-of-care applications. The clinical triage system automatically accepts 97.1% of predictions with 99.6% accuracy, substantially reducing specialist workload while maintaining diagnostic safety. Cross-validation (99.42 ± 0.18%) and multi-dataset validation confirm framework robustness. Grad-CAM visualizations provide model interpretability essential for clinical trust. This work establishes foundations for scalable, reliable, and clinically deployable automated ECG classification systems. Future work will explore multi-modal fusion with patient metadata, extension to broader arrhythmia classes, and prospective clinical validation studies.

Acknowledgment

The authors gratefully acknowledge the Department of Computer Science and Engineering, University of Barishal, Bangladesh, and the University of Global Village, Barishal, Bangladesh, for providing computational resources and institutional support for this research. Special thanks to Dr. Rahat Hossain Faisal for invaluable guidance and mentorship throughout this project. The authors also thank the anonymous reviewers for their constructive feedback that significantly improved the manuscript quality.

Ethical Statement

The authors declare that this study did not require formal ethical approval because the University of Barishal does not require Institutional Review Board or ethics committee approval for research utilizing exclusively publicly available, anonymized datasets. This exemption is based on the standard research policies issued by the institutional authority of the University of Barishal.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Cardiovascular ECG Images at https://www.kaggle.com/datasets/jayaprakashpondy/ecgimages, National Heart Foundation 2023 ECG Dataset at https://doi.org/10.17632/gwycg64pfx.1, ECG Images Dataset of Cardiac Patients at https://data.mendeley.com/datasets/gwbz3fsgp8/2, and ECG Image Data at https://www.kaggle.com/datasets/erhmrai/ecg-image-data.

Author Contribution Statement

Md. Rahat: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data curation, Writing – original draft, Visualization. **Rahat Hossain Faisal:** Supervision, Project administration, Writing – review & editing. **Khondoker Razinul Karim:** Validation, Data curation, Writing – review & editing. **Md. Minhaj Ul Islam:** Methodology, Validation, Writing – review & editing. **Md Zahid Akon:** Validation, Writing – review & editing.

References

[1] Global Burden of Cardiovascular Diseases and Risks 2023 Collaborators. (2025). Global, regional, and national burden of cardiovascular diseases and risk factors in 204 countries and territories, 1990-2023. *Journal of the American College of Cardiology*, *86*(22), 2167–2243. https://doi.org/10.1016/j.jacc.2025.08.015

[2] Martin, S. S., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., . . . , & Palaniappan, L. P. (2024). 2024 heart disease and stroke statistics: A report of us and global data from the American heart association. *Circulation*, *149*(8), e347–e913. https://doi.org/10.1161/CIR.0000000000001209

[3] Kim, D., Lee, K. R., Lim, D. S., Lee, K. H., Lee, J. S., Kim, D.-Y., & Sohn, C.-B. (2025). A novel hybrid CNN-transformer model for arrhythmia detection without R-peak identification using stockwell transform. *Scientific Reports*, *15*(1), 7817. https://doi.org/10.1038/s41598-025-92582-9

[4] Almansouri, N. E., Awe, M., Rajavelu, S., Jahnavi, K., Shastry, R., Hasan, A., . . . , & Haider, A. (2024). Early diagnosis of cardiovascular diseases in the era of artificial intelligence: An in-depth review. *Cureus*, *16*(3), e55869. https://doi.org/10.7759/cureus.55869

[5] Li, J., Guan, Z., Wang, J., Cheung, C. Y., Zheng, Y., Lim, L.-L., ..., & Wong, T. Y. (2024). Integrated image-based deep learning and language models for primary diabetes care. *Nature Medicine*, *30*(10), 2886–2896. https://doi.org/10.1038/s41591-024-03139-8

[6] Singh, V. (2025). Transfer learning in deep learning models for medical imaging: Utilizing pretrained models to improve performance in medical image analysis. *Journal of Image Processing & Pattern Recognition Progress*, *12*(1), 67–85.

[7] Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, *25*(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3

[8] Abubaker, M. B., & Babayiğit, B. (2023). Detection of cardiovascular diseases in ECG images using machine learning and deep learning methods. *IEEE Transactions on Artificial Intelligence*, *4*(2), 373–382. https://doi.org/10.1109/TAI.2022.3159505

[9] Cobo, M., Fontecha, D. C., Silva, W., & Iglesias, L. L. (2025). Physical foundations for trustworthy medical imaging: A survey for artificial intelligence researchers. *Artificial Intelligence in Medicine*, *169*, 103251. https://doi.org/10.1016/j.artmed.2025.103251

[10] Gracy, G. A., & Pravin, S. C. (2026). A lightweight residual dilated temporal transformer block for ECG classification on edge devices. *Scientific Reports*, *16*, 8834. https://doi.org/10.1038/s41598-026-35531-4

[11] Bhati, D., Neha, F., & Amiruzzaman, M. (2024). A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging. *Journal of Imaging*, *10*(10), 239. https://doi.org/10.3390/jimaging10100239

[12] Rahat, M., Islam, M. M. U., Akon, M. Z., & Faisal, D. R. H. (2025). A deep feature extraction and optimization framework for ECG classification via pretrained vision networks. In *International Conference on Data Science, AI and Applications*, 374–389. https://doi.org/10.1007/978-3-032-11352-8_26

[13] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision*, 618–626. https://doi.org/10.1109/ICCV.2017.74

[14] Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision*, 839–847. https://doi.org/10.1109/WACV.2018.00097

[15] Akan, T., Alp, S., & Bhuiyan, M. A. N. (2023). ECGformer: Leveraging transformer for ECG heartbeat arrhythmia classification. In *2023 International Conference on Computational Science and Computational Intelligence*, 1412–1417. https://doi.org/10.1109/CSCI62032.2023.00231

[16] Shi, J., Wang, K. A., & Fox, E. (2023). Sequence modeling with multiresolution convolutional memory. In *International Conference on Machine Learning*, 31312–31327.

[17] Ao, R., & He, G. (2023). Image based deep learning in 12-lead ECG diagnosis. *Frontiers in Artificial Intelligence*, *5*, 1087370. https://doi.org/10.3389/frai.2022.1087370

[18] Hanoon, F. M., & Ali, K. H. (2024). Accurate ECG images classification using Vision Transformer. *Journal of Basrah Research Sciences*, *50*(1), 328–328. https://doi.org/10.56714/bjrs.50.1.26

[19] Khalid, M., Pluempitiwiriyawej, C., Abdulkadhem, A. A., Afzal, I., & Truong, T. (2024). ECGConVT: A hybrid CNN and vision transformer model for enhanced 12-lead ECG images classification. *IEEE Access*, *12*, 193043–193056. https://doi.org/10.1109/ACCESS.2024.3516495

[20] Sadad, T., Safran, M., Khan, I., Alfarhood, S., Khan, R., & Ashraf, I. (2023). Efficient classification of ECG images using a lightweight CNN with attention module and IoT. *Sensors*, *23*(18), 7697. https://doi.org/10.3390/s23187697

[21] Zhang, C. J., Tang, F. Q., Cai, H. P., & Qian, Y. F. (2024). Heart failure classification using deep learning to extract spatiotemporal features from ECG. *BMC Medical Informatics and Decision Making*, *24*(1), 17. https://doi.org/10.1186/s12911-024-02415-4

[22] Zhao, B., Gao, Z., Liu, X., Zhang, Z., Xiao, W., & Zhang, S. (2025). DRL-ECG-HF: Deep reinforcement learning for enhanced automated diagnosis of heart failure with imbalanced ECG data. *Biomedical Signal Processing and Control*, *107*, 107680. https://doi.org/10.1016/j.bspc.2025.107680

[23] Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., . . . , & Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, *109*(5), 820–838. https://doi.org/10.1109/JPROC.2021.3054390

[24] Burhan, A., & Purwono, P. (2024). A comprehensive review of knowledge distillation for lightweight medical image segmentation. *Journal of Advanced Health Informatics Research*, *2*(2), 95–101. https://doi.org/10.59247/jahir.v2i2.294

[25] Moslemi, A., Briskina, A., Dang, Z., & Li, J. (2024). A survey on knowledge distillation:. *Recent advancements Machine Learning with Applications*, *18*, 100605. https://doi.org/10.1016/j.mlwa.2024.100605

[26] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., . . . , & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, *76*, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

[27] Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., . . . , & Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. *Artificial intelligence review*, *56*(Suppl 1), 1513–1589. https://doi.org/10.1007/s10462-023-10562-9

[28] Faghani, S., Moassefi, M., Rouzrokh, P., Khosravi, B., Baffour, F. I., Ringler, M. D., & Erickson, B. J. (2023). Quantifying uncertainty in deep learning of radiologic images. *Radiology*, *308*(2), e222217. https://doi.org/10.1148/radiol.222217

[29] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, *70*, 1321–1330.

[30] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv Preprint: 1612.01474*

[31] Vaid, A., Jiang, J., Sawant, A., Lerakis, S., Argulian, E., Ahuja, Y., . . . , & Nadkarni, G. N. (2023). A foundational

vision transformer improves diagnostic performance for electrocardiograms. *NPJ Digital Medicine*, *6*(1), 108. https://doi.org/10.1038/s41746-023-00840-9

[32] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[33] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

[34] Deepti, H G., & Latha, K B. (2026). ECG image classification for heart disease detection using a hybrid efficientNet-ViT model. *International Journal of Engineering Research & Technology*, *15*(3), 1–7. https://doi.org/10.5281/zenodo.19440105

[35] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708. https://doi.org/10.1109/CVPR.2017.243

[36] Randazzo, V., Caligari, S., Pasero, E., Giustetto, C., Saglietto, A., Bertarello, W., . . . , & Gaita, F. (2025). A vision transformer model for the prediction of fatal arrhythmic events in patients with Brugada syndrome. *Sensors*, *25*(3), 824. https://doi.org/10.3390/s25030824

[37] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167

[38] Mahel, A. S. B., Al-Gaashani, M. S., Alkanhel, R. I., Hassan, D. S., Muthanna, M. S. A., Muthanna, A., & Aziz, A. (2025). A multi-scale deep learning framework combining MobileViT-ECA and LSTM for accurate ECG analysis. *IEEE Access*, *85473*, 85492. https://doi.org/10.1109/ACCESS.2025.3568026

[39] Sun, H., Zhou, W., Yang, J., Shao, Y., Xing, L., Zhao, Q., & Zhang, L. (2024). An improved medical image classification algorithm based on Adam optimizer. *Mathematics*, *12*(16), 2509. https://doi.org/10.3390/math12162509

[40] Sawano, S., Kodera, S., Setoguchi, N., Tanabe, K., Kushida, S., Kanda, J., . . . , & Komuro, I. (2024). Applying masked autoencoder-based self-supervised learning for high-capability vision transformers of electrocardiographies. *Plos One*, *19*(8), e0307978. https://doi.org/10.1371/journal.pone.0307978

[41] Zhang, M., Jin, H., & Yang, Y. (2024). ECG classification efficient modeling with artificial bee colony optimization data augmentation and attention mechanism. *Mathematical Biosciences and Engineering*, *21*(3), 4626–4647. https://doi.org/10.3934/mbe.2024203

[42] Xiong, N., Wei, Z., Wang, X., Wang, Y., & Wang, Z. (2025). HCTG-Net: A hybrid CNN–transformer network with gated fusion for automatic ECG arrhythmia diagnosis. *Bioengineering*, *12*(11), 1268. https://doi.org/10.3390/bioengineering12111268

[43] Wu, Y., Chen, Y., Zhang, B., Chen, X., Peng, C., Jiang, J., . . . , & Wu, G. (2026). Deep learning-based early prediction of life-threatening ventricular arrhythmias using long-term Holter ECG signals. *Frontiers in Cardiovascular Medicine*, *13*, 1714589. https://doi.org/10.3389/fcvm.2026.1714589

How to Cite: Rahat, M., Faisal, R. H., Karim, K. R., Islam, M. M. U., & Akon, M. Z. (2026). A Comprehensive Multi-Strategy Transfer Learning Framework with Knowledge Distillation for ECG Image Classification: A Multi-Dataset Validation Study. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62028620